Accuracy of Deictic Gestures to Support Telepresence on Wall-sized Displays

Ignacio Avellino
Inria, Univ Paris-Sud & CNRS
avellino@lri.fr

Cédric Fleury
Univ Paris-Sud & CNRS, Inria
cfleury@lri.fr

Michel Beaudouin-Lafon
Univ Paris-Sud & CNRS, Inria
mbl@lri.fr

ABSTRACT

We present a controlled experiment assessing how accurately a user can interpret the video feed of a remote user showing a shared object on a large wall-sized display by looking at it or by looking and pointing at it. We analyze distance and angle errors and how sensitive they are to the relative position between the remote viewer and the video feed. We show that users can accurately determine the target, that eye gaze alone is more accurate than when combined with the hand, and that the relative position between the viewer and the video feed has little effect on accuracy. These findings can inform the design of future telepresence systems for wall-sized displays.

Author Keywords

Pointing; Telepresence; Wall-sized display;
Remote collaboration

ACM Classification Keywords

H.5.3 [Group and Organization Interfaces]: Collaborative Computing; Computer-supported cooperative work

INTRODUCTION

Large interactive rooms with wall-sized displays help users manage the increasing size and complexity of data in science, industry and business. They naturally support co-located collaboration among small groups but can also be interconnected to support remote collaborative work. Video is critical to such remote collaboration as it supports non-verbal cues, turn-taking and shared understanding of the situation [3]. However, current telepresence systems are designed for meetings where users sit around a conference table and do not support spaces where users move around and work on shared data.

Our goal is to study telepresence systems that support remote collaboration across wall-sized displays by combining the shared task space with the shared person space [2]. The former refers to the task at hand and involves actions such as making changes, annotating and referencing objects; the latter refers to the collective sense of co-presence and involves facial expressions, voice, gaze and body language. Buxton [1] defines the overlap between these spaces as the *reference space*, where *"the remote party can use body language to reference the work"*. Our goal is therefore to study the reference space in the context of wall-sized displays.

We focus on telepresence systems linking two distant rooms with wall-sized displays showing the same content (Figure 1). Live video feeds of the users are captured by an array of cameras at eye level and displayed on the remote display at the corresponding position. Users can thus see the face of remote users and interact in a consistent way with the shared content.

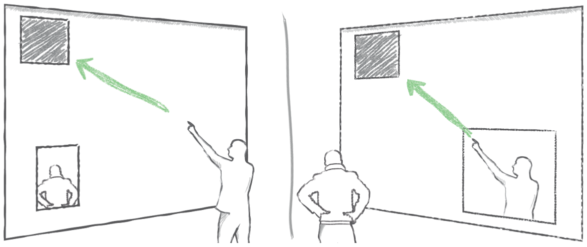

Figure 1. Users working with shared objects using a wall-sized display.

Referencing shared objects to support collaboration and mutual understanding is common when working together [6]. This paper investigates users' ability to accurately determine which shared object is referenced by a remote user without the need for dedicated technology such as telepointers.

We conducted a controlled experiment to study (a) how accurately a local user perceives a reference to a shared object performed by a remote user either by looking at it or by pointing their hand at it, and (b) whether the position of the local user in front of the wall-sized display affects this accuracy.

RELATED WORK

A number of systems, from the early VideoWhiteboard [9] and ClearBoard [4] to the more recent Connectboard [8] and Holoport [5], have explored how to combine person and task spaces with vertical or slanted displays, based on a glass metaphor where both spaces are overlaid. All these systems provide gaze awareness, the ability to notice direct eye contact and to notice what the remote person is looking at, and perception of which object the remote person is drawing or manipulating. However they are meant to be used by participants who do not move much in front of the display, and the accuracy of remote pointing has not been evaluated.

Nguyen & Canny [7] developed a telepresence system that uses a camera and projector per participant to create spatial faithfulness with multiparty conferencing. This system avoids the so-called *Mona Lisa effect*, where the image of

CHI 2015, April 18–23, 2015, Seoul, Republic of Korea
Copyright © 2015 ACM 978-1-4503-3145-6/15/04$15.00
http://dx.doi.org/10.1145/2702123.2702448

a subject looking into the camera is seen by remote participants as looking at them, irrespective of their position. Wong & Gutwin [10] assessed pointing accuracy but used a Collaborative Virtual Environment, where users are represented by avatars instead of live video feeds.

In summary, while a number of telepresence systems have been proposed that enable remote collaboration on shared objects, very few studies have assessed the accuracy of designating such objects remotely through pointing or gazing.

EXPERIMENT

We conducted a controlled experiment to assess how accurately an observer can determine which object a remote user is showing on a wall-sized display. The remote user shows a target and the observer must determine which one it is. We use pre-recorded videos of the remote user and display them on the wall-sized display in front of the observer at the same position where the recording camera was placed.

We control three factors: how the remote user specifies the target (by turning the head or by turning the head and pointing at it), the position of the target relative to the video displayed on the wall (19 positions) and the position of the observer in front of the wall-sized display (5 positions).

Experimental Setup

The 19 targets are displayed on a $5.5m \times 1.8m$ wall-sized display made of a grid of 8×4 30” monitors. Each target is a black letter on a white background surrounded by a blue circle. We exclude letters that could be confused, such as O and Q. We use a concentric radial distribution of targets in order to control for both distance and angle to the video. Three rings of targets surround a central one (*ring0*), where the video is displayed (Figure 2). The first two rings (*ring1* and *ring2*) have 8 targets, one for each cardinal and diagonal direction. Due to the aspect ratio of the wall-sized display, the third ring (*ring3*) has only two targets. *Ring0, ring1* and *ring2* are 11.5° apart when measured from the viewing position, while *ring2* and *ring3* are 23° appart.

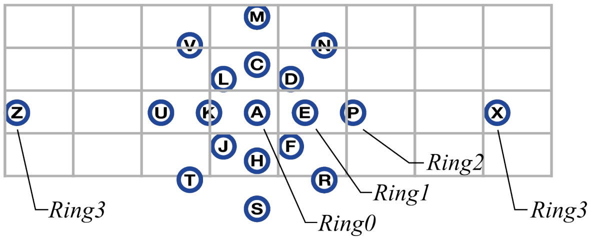

Figure 2. The 19 targets laid out in 4 rings on the wall-sized display.

Video Recording

We recorded 114 10-second videos of three different actors showing the 19 targets on a wall-sized display in both conditions (*head* and *pointing*): a 29 year-old woman with pulled back hair and brown eyes, a 29 year-old man with brown medium-length hair and brown eyes and a 27 year-old man with short brown hair and hazel eyes. The camera was set in front of the wall-sized display, at the position of the central target. The actors were placed 230cm away from the camera so that the pointing hand was within the recorded frame in all pointing directions. The actors were instructed to point or look successively at each target.

Video Playback

The participants in the experiment sat in front of the same display used for video recording, 230cm away from the display. The recorded videos were displayed on top of the central target, at the same position as the recording camera. Based on the focal length used at recording time (43mm in 35mm equivalent focal length), we adjusted the size of the video so that the remote users appeared life-size, as if they were sitting 230cm behind the display, i.e. 460cm away from the participants. The height of the chair was adjusted for each participant so that the video was at eye-level.

Participants

12 right-handed participants (8 male), aged 21 to 33 (median 27), all computer science graduates, participated in the study. All participants had normal or corrected to normal vision. 10 had a right dominant eye, 2 a left one. 2 participants used conferencing systems every day, 6 more than once a week, 3 once a week, and 2 almost never. All participants received sweets as compensation for their time.

Task

Participants watch each video playing in an infinite loop. When they are ready to answer, they tap a large “Stop” button on an iPad 3 tablet and answer which target was being shown, by tapping the target on a replica of the display layout.

Procedure

The within-subject design has the following factors:

- TECHNIQUE used to indicate the targets, with 2 conditions: *head*, the natural combination of head turning and gazing, and *pointing*, the combination of head turning, gazing and pointing the target with the arm and finger;
- POSITION of the participant in front of the display, with 5 conditions: *center*, located in front of the videos, *left* (resp. *farLeft*), located 1m (resp. 2m) to the left, and *right* (resp. *farRight*), located 1m (resp. 2m) to the right;
- ACTOR: we recorded 3 sets of videos with 3 different actors to ensure that the choice of the remote person does not have an effect;
- TARGETS: 19 targets were used, a central one surrounded along 8 directions by 3 rings of targets with only the left and right targets on the last ring ($19 = 1 + 8 \times 2 + 2$).

For each participant, the conditions were grouped by TECHNIQUE, then by ACTOR and then by POSITION. The order of presentation was counterbalanced across conditions by using Latin squares for the first three factors and a randomized order for TARGET. Each Latin square was mirrored and the result was repeated as necessary. For each TECHNIQUE×ACTOR×POSITION condition, the order of the 19 targets was randomized so that successive videos never showed targets in adjacent rings from the same direction.

For training, we used the same subset of 12 videos covering all directions and distances for each participant to practice the task and the entry of answers. 4 different random positions were used (2 for the *head* condition and 2 for the *pointing* condition) with 3 videos each. Then, the 570 videos were presented: 2 TECHNIQUEs x 3 ACTORs x 5 POSITIONs x 19 TARGETs. A mandatory break was held every 190 trials, corresponding to 2 sets of 5 positions, and a reminder of an optional break was provided every 95 trials.

For each trial, we collected the answer from the participants, i.e. which target they thought was being shown. At the end of the experiment, participants filled out a short questionnaire.

Results

We measure the accuracy of the participants in determining the target shown by the remote users by two types of errors: Distance error, the number of rings between the actual target and the target chosen by the participant; angle error, the difference between the angles of the two targets, as a multiple of 45°. For angle errors, we remove trials where the indicated target is the central one, since it has no meaning in this case.

We find a small learning effect[1] of TECHNIQUE on distance error: it significantly decreases from 0.36 for the first technique to 0.29 for the second one ($F(1, 6827) = 39.19, p < 0.0001$). Many participants learned which videos correspond to the farthest targets and used that information to assign intermediate targets to subsequent videos, based on the angle of the head or arm being smaller than that of the farthest targets. We did not find a learning effect on the angle error ($F(1, 6467) = 3.58, p = 0.059$).

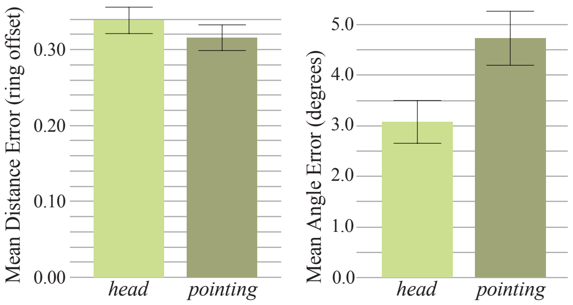

Figure 3. Distance and angle error by TECHNIQUE (bars indicate CI)

Figure 3 shows the mean and 95% Confidence Intervals (CI) of distance and angle error by TECHNIQUE. The multiway ANOVA with REML shows no significant difference in distance error between the two TECHNIQUEs ($F(1, 6817) = 0.87, p = 0.35$), but the difference in angle error is significant ($F(1, 6459) = 39.37, p < 0.0001$), with a mean of 3.10 for *head* vs. 4.72 for *pointing*. Surprisingly, using the arm led to higher angle errors than using only the head.

Figure 4 shows the mean and CI of distance and angle error by POSITION. The multiway ANOVA with REML shows no significant differences in distance error among the five POSITIONS ($F(4, 6817) = 0.95, p = 0.44$), but a significant difference for angle error ($F(4, 6459) = 3.84, p = 0.0040$), with means (from left to right) 5.00, 3.47, 3.33, 3.37, 4.31. A Tukey HSD post-hoc test reveals two groups of positions significantly different from each other: {*farLeft*, *farRight*}, {*center*, *left*, *right*, *farRight*}.

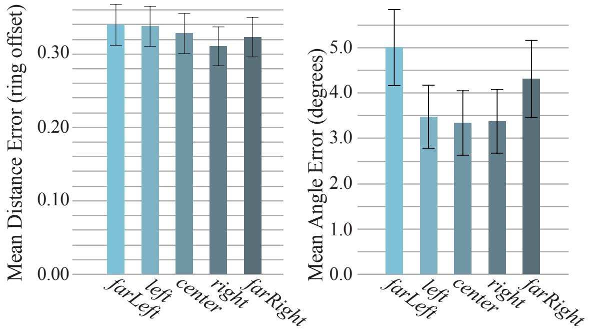

Figure 4. Distance and angle error by POSITION (bars indicate CI)

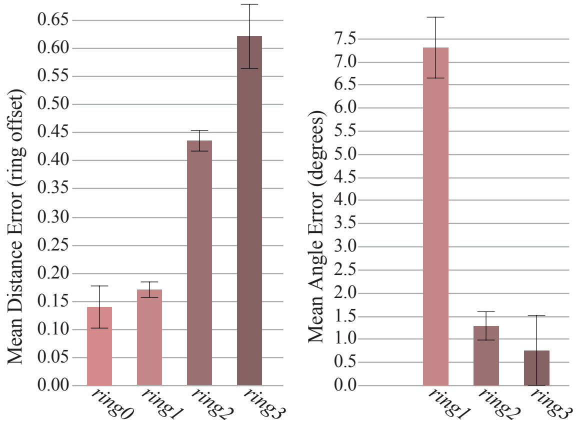

Figure 5. Distance and angle error by target ring (bars indicate CI)

Figure 5 shows the mean and CI of distance and angle error for each ring of targets. The multiway ANOVA with REML shows significant differences for distance ($F(3, 6817) = 294.14, p < 0.0001$) and angle ($F(2, 6459) = 165.01, p < 0.0001$) error (as mentioned before, *ring0* was removed for the angle error analysis). Distance error means (from *ring0* to *ring3*) are: 0.14, 0.17, 0.44, 0.62. Angle error means (from *ring1* to *ring3*) are: 7.30, 1.28, 0.75. A Tukey HSD post-hoc test reveals three significantly different groups of rings for distance error: {*ring0*, *ring1*}, {*ring2*}, {*ring3*}. For angle error, {*ring1*} is significantly different from {*ring2*, *ring3*}.

Discussion

According to our results, distance error is not significantly different when using the head or head+arm, but the angle error is larger when using the head+arm. While the difference is small, this is a surprising result. By analyzing the videos, we noticed that most of the time the direction of the arm does not indicate the target. This is because users place the tip of their finger on the line of sight between their eye and the target. In the post-hoc questionnaires, we found that 4 participants used

[1] Statistical analyses were performed with SAS JMP. We performed Mixed Model REML (Restricted Maximum Likelihood) analyses with "participant" as a random factor.

only the arm direction when determining the pointed target; 6 used first the arm and when in doubt looked at the eyes and head direction; only one participant determined the correct location by connecting the eyes with the tip of the finger. Because the arm is the most salient cue in the video, it is likely that users use it as the primary source for determining the pointed target, ignoring the geometrical interpretation that we perform in a face-to-face environment.

The fact that the position of the participant relative to the video feed has little or no effect was also surprising. We did not expect this effect, analogous to the *Mona-Lisa effect*, to be so strong. The extreme positions, *farLeft* and *farRight*, had a higher angle error, which can be explained by the fact that the observers are looking at the image with an angle of 49°, making the task harder.

We found that distance error increases when the targets are further away from the video feed. This may be due to the fact that the videos are shot from the front. In this setting, a small change of angle, e.g. 10°, when pointing near the center produces a noticeable change in distance between the shoulder and the finger in the 2D projection of the arm. The same change in angle when the arm is pointing at the last ring of targets results in a much smaller distance between the finger and the shoulder in the 2D projection of the arm, making it harder to notice. Angle error, on the other hand, decreases when the targets are further away from the video feed. This is due to the radial layout: the distance between targets on an inner ring is smaller than on an outer ring, resulting in larger variations in the direction of the hand and arm. It is interesting to note that on the farthest targets, for which it was only possible to express distance, users still made angle errors because they thought that the target laid on *ring2*.

CONCLUSION AND FUTURE WORK

We investigated the ability to accurately determine which shared object a remote user is referencing when sharing data on a wall-sized display. We found that showing objects only with the head leads to smaller angle error than with the head and arm. We also found no effect of the observer's position on accuracy, except at the farthest positions for angle error.

However, while distance accuracy decreases when the object is further away from the video of the remote user, angle accuracy increases with object distance. We attribute this effect to the fact that when capturing the remote user's image from the front, some body movements are more salient than others, conveying cues that help users determine more accurately the distance of close targets and the direction of far targets.

This study has three main implications for design:

1. Users can accurately estimate which object is being indicated only by looking at the head on the remote video, without requiring explicit pointing actions nor telepointers;
2. Therefore even when their hands are busy, users can point using their head without losing accuracy, supporting, e.g., the use of deictic instructions when holding tools;
3. The position of the video feed relative to the observer is not critical for accuracy when indicating remote objects, thus it can be moved around without loss of accuracy; this allows the system to match the video with the camera position on the remote side, maintaining spatial relationships with shared objects on both sides.

In future work we plan to investigate the body cues used to indicate an object and how to convey them over a telepresence system. We also plan to apply a similar methodology to situations that involve eye contact and understand how video distorts the perception of simple communicative acts. Finally we want to apply these findings to a functional system that supports video communication in large interactive rooms.

ACKNOWLEDGMENTS

We thank Wendy Mackay for discussions about the experimental design and statistical analysis. This work is supported by French National Research Agency grant ANR-10-EQPX-26-01 “Digiscope”.

REFERENCES

1. Buxton, W. Mediaspace–meaningspace–meetingspace. In *Media Space 20 + Years of Mediated Life*, Computer Supported Cooperative Work. Springer, 2009, 217–231.
2. Buxton, W. A. S. Telepresence: Integrating shared task and person spaces. In *Proc. Graphics Interface, GI'92* (1992), 123–129.
3. Isaacs, E. A., and Tang, J. C. What video can and cannot do for collaboration: A case study. *Multimedia Systems 2*, 2 (1994), 63–73.
4. Ishii, H., and Kobayashi, M. Clearboard: A seamless medium for shared drawing and conversation with eye contact. In *Proc. Human Factors in Computing Systems, CHI'92*, ACM (1992), 525–532.
5. Kuechler, M., and Kunz, A. HoloPort - a device for simultaneous video and data conferencing featuring gaze awareness. In *Proc. Virtual Reality, VR'06*, IEEE (2006), 81–88.
6. Mackay, W. E. Media spaces: environments for informal multimedia interaction. In *Computer Supported Co-operative Work*, M. Beaudouin-Lafon, Ed., John Wiley & Sons (1999), 55–82.
7. Nguyen, D., and Canny, J. MultiView: Spatially faithful group video conferencing. In *Proc. Human Factors in Computing Systems, CHI'05*, ACM (2005), 799–808.
8. Tan, K.-H., Robinson, I., Samadani, R., Lee, B., Gelb, D., Vorbau, A., Culbertson, B., and Apostolopoulos, J. Connectboard: A remote collaboration system that supports gaze-aware interaction and sharing. In *Workshop on MMSP '2009*, IEEE (2009), 1–6.
9. Tang, J. C., and Minneman, S. VideoWhiteboard: Video shadows to support remote collaboration. In *Proc. Human Factors in Computing Systems, CHI'91*, ACM (1991), 315–322.
10. Wong, N., and Gutwin, C. Where are you pointing?: The accuracy of deictic pointing in CVEs. In *Proc. Human Factors in Computing Systems, CHI'10*, ACM (2010), 1029–1038.