Localizing Teleoperator Gaze in 360° Hosted Telepresence

Jingxin Zhang*
Human-Computer Interaction
University of Hamburg

Nikolaos Katzakis†
Human-Computer Interaction
University of Hamburg

Fariba Mostajeran‡
Human-Computer Interaction
University of Hamburg

Frank Steinicke§
Human-Computer Interaction
University of Hamburg

(a) The tablet variant of the telepresence robot.

(b) AR display: The guest is superimposed on top of the robot. Image captured through HTC Vive Pro HMD.

(c) Experiment in virtual reality space: AR display and tablet display.

Figure 1: Depiction of our telepresence setup and experimental environment in Unity.

Abstract

We evaluate the ability of locally present participants to localize an avatar head's gaze direction in 360° hosted telepresence. We performed a controlled user study to test two potential solutions to indicate a remote user's gaze. We analyze the influence of the user's distance to the avatar and display technique on localization accuracy. Our experimental results suggest that all these factors have a significant effect on the localization accuracy with varying effect sizes.

1 Introduction

Traditional telepresence platforms consist of a display that depicts the face of the teleoperator so that bystanders feel an enhanced sense of teleoperator presence. In some cases, the teleoperator is accompanied by a host. e.g. A telepresent doctor is making a house call accompanied by a family member; a telepresent building site inspector or industrial safety inspector accompanied by the locally present foreman; an apartment viewing, where a real-estate agent is attending to show the property to a remotely telepresent client. We refer to these as *hosted telepresence*. With a traditional telepresence platform, because the camera is mounted facing forward, when teleoperators wish to change their view direction, they rotate the platform manually. This change of heading by the platform is an important gaze cue for the bystanders. However, this means that in hosted telepresence there are occasions when the display is facing away from the host, and as such, the head or face of the guest[1] is not visible.

At the same time 360° video telepresence platforms are increasingly prescribed for remote teleoperation [1]. In 360° video teleoperation, the operator wears a head-mounted display (HMD) and the remote environment is displayed as a spherical surface. Because of the HMD, it is technically complicated to capture an accurate reconstruction of the teleoperator's head and face. In addition to this, since the teleoperator has a 360° video sphere surrounding them, the robotic base does not need to rotate when he or she looks around thereby conserving battery. The platform would rotate only to move to a new heading and if the locomotion mechanism is holonomic/omnidirectional rotation is not necessary at all. If the base is not rotating, the host has no cues for the gaze direction of the guest and that impedes gaze localization by the host, hinders his or her ability to offer proactive information and devalues the hosted telepresence experience for both parties.

2 Evaluated Techniques

This first solution (Figure 1a) is designed so that it can be used with existing tablet-equipped telepresence platforms, with the simple addition of a 360° camera.

The second solution is to show the avatar in AR, superimposed on the robot platform. Again, the avatar is consistent with the teleoperator's heading and its heading is, similar to the tablet solution, completely detached from the heading of the robot. In this solution, because the display would be an AR HMD, the robotic platform does not need to rotate at all to keep the avatar in view for the host (Figure 1b). We prototyped this with an HTC Vive Pro HMD in AR mode to display the avatar together with the real environment. The avatar is essentially a puppet, controlled by the teleoperator's HMD orientation. A Vive tracker was attached on the robotic platform and used for rendering the AR avatar at the correct location.

*e-mail: jxzhang@informatik.uni-hamburg.de
†e-mail: nikolaos.katzakis@uni-hamburg.de
‡e-mail: mostajeran@informatik.uni-hamburg.de
§e-mail: frank.steinicke@uni-hamburg.de

[1] the teleoperator will be occasionally referred to as "guest" throughout the paper depending on the context

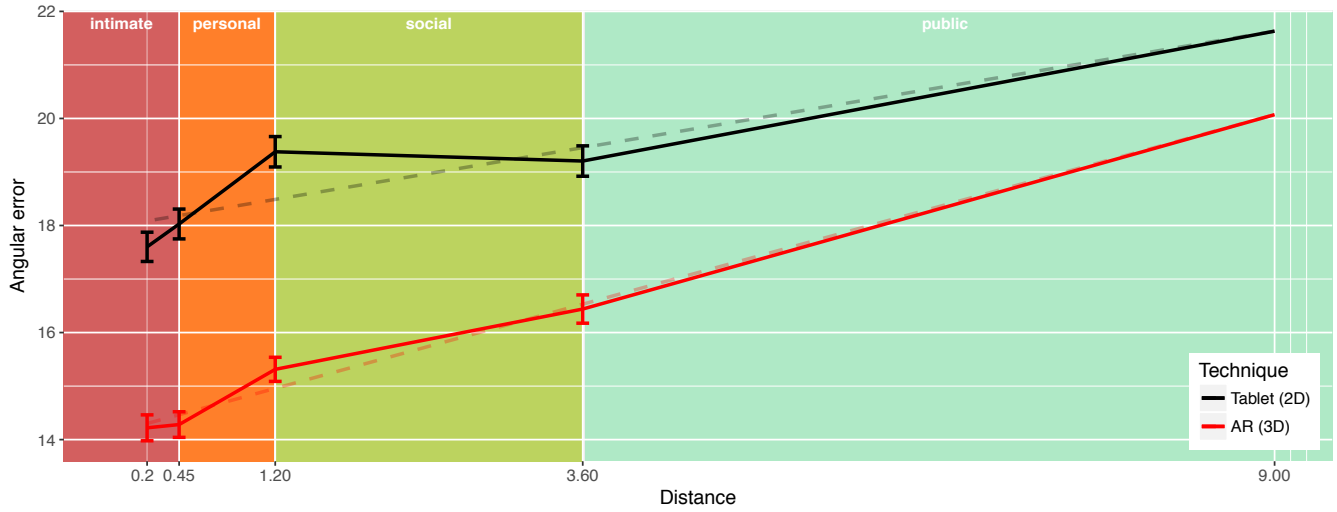

Figure 2: Effect of robot *distance* on *error* (lower is better). Dotted line is linear regression results according to the measurement data at distances of 0.2m, 0.45m, 1.2m and 3.6m. Based on this, we have a predicted extrapolation beyond distance of 3.6m, which were not tested in our experiment.

3 Experiment

The environment consisted of a spherical virtual space with a radius of 5 meters. The sphere was always anchored to the robot. A tiny polka dot texture was applied to the inside surface of the sphere to allow participants to feel the size and bounds of the space (Figure 1c). During the experiment, participants were asked to stand on a target spot. The distances between telepresence robot and participants in the spherical space were chosen based on Hall's research in proxemics. We used 0.2m as the minimum possible distance in the intimate space, while 0.45m, 1.2m and 3.6m are the boundary distances between intimate space, personal space, social space and public space respectively.

In addition the avatar attached on top of the robot were displayed with two *techniques*, which were 3D stereoscopically rendered avatar (*AR*) and 2D flattened tablet avatar (*tablet*) respectively. We manipulated the avatar's gaze by rotating about the *yaw* (axis defined by the spine) and *pitch* angles (axis defined by the ears). Angles were 0°, 30°, 60°, 90°, 120°, 150°, 180° for clockwise yaw; -45°, -30°, -15°, 0°, 15°, 30°, 45° for pitch.

When the experiment started, an avatar attached to the telepresence robot appeared directly in front of the participants in four specific distances (D = 0.2m, 0.45m, 1.2m, 3.6m) with two display techniques (AR and tablet) respectively (Figure 1c). The avatar was displayed for two seconds after which it disappeared. Participants then had to point to indicate, using their wand, which point on the sphere they thought the avatar was looking at. A pink sphere was displayed on the sphere surface as cursor at the participant's wand intersection point. Participants would click to confirm their estimation and one second later the avatar would appear for the next trial.

Data was analyzed with a repeated-measures ANOVA. Display *technique*, *distance*, *yaw*, *pitch* were the manipulated factors. We investigated their effects on angular error. i.e. the angle between the line defined by the avatar's gaze direction and an avatar gaze line to the point the participant pointed. Angular error will simply be referred to as *Error* (symbol *e*) for the remainder of the paper.

4 Results - Conclusion

For the main effects, *Distance* had a significant effect on *Error* ($F_{3,39} = 4.77, p < 0.01, \eta^2 = 0.013$). A plot for distance can be found in Figure 2; *Technique* also had a significant effect on *Error* ($F_{1,13} = 26.47, p < 0.001, \eta^2 = 0.060$); *Yaw* also had a significant effect on *Error* ($F_{6,78} = 11.89, p < 0.001, \eta^2 = 0.080$); Finally, *Pitch* had a significant effect on *Error* ($F_{6,78} = 5.00, p < 0.001, \eta^2 = 0.038$).

The AR display technique maintained a steady error throughout the intimate space and then degraded in personal and social space. Conversely, the tablet technique had the largest error at the border between personal and social space (1.2m) (Figure 2).

We presented two methods for improving gaze direction estimation of a teleoperator in 360° *hosted telepresence* systems; constantly orienting the tablet to face the host and an AR avatar method. We contributed a formal user study (simulated AR) that shed some light into the performance of these two proposed display techniques. A summary of our findings.

Telepresence "hosts" were 19% better at estimating the gaze of a stereoscopically rendered avatar over a flattened avatar image. Certain combinations of yaw, pitch, distance were better for the host to estimate the gaze direction (e.g., yaw = 180°/pitch = 0° @1.2m with the AR display has the least localization error 4.84° among all the tested conditions). This is the first time such results have been reported in the literature. Yaw of the guest avatar head can affect localization accuracy by as much as 22.05° for a tablet display and 17.04° for an AR display. Pitch of the guest avatar head can affect localization accuracy by as much as 22.08° for a tablet display and 16.42° for an AR display. Host localization error exhibits an increasing trend with distance from the guest, but not a linear one.

These results should provide reference for *hosted telepresence* system designers but also for any social VR system where participants are required to estimate the gaze of an avatar.

Acknowledgments

This work was partially funded by the German Research Foundation (DFG) and the National Science Foundation of China (NSFC) in project Crossmodal Learning, TRR-169.

References

[1] J. Zhang, E. Langbehn, D. Krupke, N. Katzakis, and F. Steinicke. A 360 video-based robot platform for telepresent redirected walking. In *Proceedings of the 1st International Workshop on Virtual, Augmented, and Mixed Reality for Human-Robot Interactions (VAM-HRI)*, pp. 58–62. ACM/IEEE, 2018.