Investigating Workspace Awareness in 3D Face-to-Face Remote Collaboration

Maurício Sousa, Daniel Mendes, Rafael Kuffner Dos Anjos, Daniel Simões Lopes, Joaquim Jorge
INESC-ID Lisboa, Insituto Superior Técnico, Universidade de Lisboa
{antonio.sousa, danielmendes, rafaelkuffner, daniel.s.lopes, jorgej}@tecnico.ulisboa.pt

***Abstract*—Face-to-face telepresence promotes the sense of "being there" and can improve collaboration by allowing immediate understanding of remote people's nonverbal cues. Several approaches successfully explored interactions with 2D content using a see-through whiteboard metaphor. However, with 3D content there is a decrease in awareness due to ambiguities originated by participants' opposing points-of-view. We investigate how people and content should be presented for discussing 3D renderings within face-to-face collaborative sessions. To this end, we performed a user evaluation to compare four different conditions, in which we varied reflections of both workspace and remote person's representation. Results suggest potentially more benefits to remote collaboration from workspace consistency rather than people's representation fidelity.**

***Index Terms*—Telepresence, 3D Workspace Awareness, Face-to-face Communication**

I. Introduction

Videoconferencing and telepresence allow for virtual encounters to take place and expedite the communication between geographically separated people. Videoconferencing systems using real size portrayal of people become closer to a co-located experience and, in fact, full-body face-to-face communication has been shown to improve task completion time, presence, and efficiency of communication [3], while enabling non-verbal visual cues including deictic gestures. Hence, people should rely on natural communication, verbal and non verbal, to convey the focus of the collaboration and pinpoint details on shared content as if they were physically co-located.

When designing for face-to-face collaboration it is necessary to take into account how to address interactions in a shared task space. Despite being typically considered separated from the person space, it has been suggested that both task and person spaces should be integrated when considering face-to-face meetings [1]. Indeed, with transparent displays two participants are able to see one another and share digital content, rendered between them, that can be jointly manipulated. Yet, in plain face-to-face interactions mediated by displays, people have no common orientation of right or left. This can be addressed by mirror-reversing the remote person's video stream, producing gaze and pointing awareness, since 2D graphics and text can be corrected to the participant's point-of-view.

However, 3D digital content gives rise to detracting issues that affect and impair workspace awareness. Participants do not share the same *forward-backwards* orientation, occlusions can affect the understanding of where or what the remote person is pointing at. Also, contrary points-of-view can result in different perceptions or even serious communication missteps.

II. Evaluating Workspace Awareness

We set out to assess if different manipulations of person and task spaces can enhance workspace awareness and the way people communicate when collaborating in a face-to-face setting with 3D content. We developed a full body telepresence prototype and implemented four different workspace conditions. For this, we designed a collaborative 3D assembly task where an *Instructor* guides a remote *Assembler* to reach the correct solution of a toy problem using cubes. Our goal was to study the participants' *point-of-view*, remote participant's *embodiment* and *workspace* rendering. For *point-of-view* we considered that participants could observe workspace in usual opposing points-of-view or simulating an identical viewing experience. Also, similarly to Ishii et al. [2], *embodiment* and *workspace* variables could both be horizontally inverted or not. Therefore, our evaluation followed a within subjects design with four conditions:

1) *Real Life Face-to-face (RL):* Derived from the real world face-to-face scenario, both participants can see each other and the workspace as if they were in opposite ends. As such, the reference space should be natural, but participants have contrary points-of-view and cannot observe the workspace's opposite side.
2) *Simulated Side-by-side (SS):* While remaining face-to-face in regard to the embodied representation, participants share the same point-of-view of the workspace, simulating a side-by-side approach. Participants perceive the workspace from the same side and can use verbal relative directions, but pointing gestures do not match the reference space.
3) *Mirrored Person (MP):* Participants share the same point-of-view, yet the instructor's embodied representation is horizontally inverted to match the reference space. Despite the assembler perceives a mirror embodiment of the instructor, both deictic gestures and verbal relative directions match.
4) *Mirrored Workspace (MW):* With an identical point-of-view, participants also share faithful face-to-face embodiment representations of each other. However, assembler's workspace is horizontally inverted, so that deictic

gestures can be used to reference a point. Yet, any verbal relative direction is in reverse.

A total of 16 participants were grouped in pairs, separated into two different rooms equipped with similar setups comprising a 55 inch display in portrait mode and a Microsoft Kinect v2, and were asked to perform four tasks, one with each condition. All tasks consisted in solving a block-based puzzle with five colored cubes on top of a checkerboard, where the instructor guides the assembler to complete the puzzle using verbal and non-verbal communication cues. Also, only the instructor could see the instructions and the colors of the cubes. For the assembler all cubes were rendered in gray. The instructor's duty was to make it clear to the assembler which cube to pick up next and where to place it.

III. Results and Observations

We logged completion times, number of wrong cube selections and wrong cube placements. Regarding time, no statistically significant differences were found. Also, no statistically significant differences were found for either wrong cube selections or placements.

After the completion of each task, participants were asked to fill up a preferences questionnaire related to the condition they just experimented. Statistical significant differences were found on three questions for the instructor (It was easy to complete the task: $\chi^2(3)$=10.892, p=.012; It was easy to explain the row of the cube to select: $\chi^2(3)$=11.598, p=.009; It was easy to explain the column of the cube to select: $\chi^2(3)$=10.102, p=.018). Participants in the instructor's role strongly agreed that *MW* was overall more difficult than *SS* (Z=-2.743, p=.006) and *MP* (Z=-2.722, p=.006). Instructors agreed that in the *MP* condition, explaining the row of the cube to select is easier than *MW* (Z=-2.967, p=.003). It was also easier for instructors to explain the column of the next cube to be selected in the *MP* condition than *MW* (Z=-2.675, p=.007). We did not find any significant statistical difference after participants experienced tasks in the role of assembler.

Throughout all conditions, verbal communication was predominant using combined spatial and temporal references (e.g. *"left to the cube you have previously moved."*). We observed that participants developed an informal shared protocol to better understand how to complete the task. This was achieved by the instructor asking several questions to the assembler. More specifically, instructors inquired if the assembler could raise a arm and/or select a cube on a specific corner of the workspace. Henceforth, instructors would communicate the commands already in the assemblers' reference frame, which justifies the existence of significant differences in the questionnaires only for instructors.

Participants that started with *RL* condition used indicative gestures much more naturally and frequently, until experiencing the *SS* where these were ambiguous. At that point, the mentioned communication style would be established, overpowering deictics, which would be only applied as a last resort. Even so, involuntary non-verbal cues such as gaze, subtle hand, finger gestures accompanying speech, or leaning the body to a certain direction was frequently picked up by assemblers, who would try to predict the next instruction according to these visual cues. Explicit line and column indications had seldom use and had a negative impact in all of its occurrences. Indications such as *"third row, second column"* were harder to disambiguate than temporal references.

The usage of non-verbal communication varied widely according to the workspace condition. In *RL*, gestures were used to disambiguate depth, given that it was the only condition where this mapping was accurate. Also, *RL* was the only condition where we had some users use non-verbal cues as their main communication method. In *SS*, all attempts of using hand gestures resulted in errors by the assemblers. *MP* allowed users to use gestures naturally as a complement to clear verbal instructions. Finally, in *MW*, gestures were used by majority of participants, but less accurately than in *RL*, due to the fact that there was not a direct mapping between pointing and verbal directions.

IV. Conclusions

We presented an evaluation of several combinations of different points-of-view, and workspace and embodiment characteristics to study remote face-to-face collaborative work on 3D shared content. Results show an absence of significant differences in task performance and, for user preferences, statistical significant differences were found on instructors' answers. This happened because it was mostly the instructor who did the calculations regarding reference frames, which rendered all conditions alike to the assembler.

Although participants established the informal shared protocol to calibrate reference frames and achieved similar performance in all conditions, a reflected workspace was clearly identified as being more difficult than an exact representation. We argue that the cognitive workload of being constantly converting coordinates between both frames is mentally demanding.

In complex scenarios, where it is imperative for both participants to observe the same details, the *RL* condition is unfit. This and the cognitive cost associated to the *MW* condition, leads us to suggest that, for this kind of scenarios, having an exact workspace with an identical point-of-view is highly desirable. The choice between *SS* or *MP* will be dependent on whether the accuracy of the remote person's representation is more relevant than the consistency between the person and task spaces, respectively.

Acknowledgements

This work was supported by FCT through grants UID/CEC/50021/2013 and IT-MEDEX PTDC/EEISII/6038/2014.

References

[1] Ishii, H., Kobayashi, M. (1992). ClearBoard: a seamless medium for shared drawing and conversation with eye contact. In Proceedings of CHI '92. ACM.

[2] Ishii, H., Kobayashi, M., Grudin, J. (1993). Integration of interpersonal space and shared workspace: ClearBoard design and experiments. ACM Transactions on Information Systems (TOIS), 11(4).

[3] Pejsa, T., Kantor, J., Benko, H., Ofek, E., Wilson, A. (2016). Room2room: Enabling life-size telepresence in a projected augmented reality environment. In Proceedings of CSCW '16. ACM.