Gaze and Gestures in Telepresence: multimodality, embodiment, and roles of collaboration

Mauro Cherubini, Rodrigo de Oliveira, Nuria Oliver, and Christian Ferran
Telefonica Research
via Augusta, 177 – ES-08021 Barcelona, Spain
{mauro, oliveira, nuriao, icfb}@tid.es

ABSTRACT
This paper proposes a controlled experiment to further investigate the usefulness of *gaze awareness* and *gesture recognition* in the support of collaborative work at a distance. We propose to redesign experiments conducted several years ago with more recent technology that would: a) enable to better study of the integration of communication modalities, b) allow users to freely move while collaborating at a distance and c) avoid asymmetries of communication between collaborators.

Author Keywords
Deixis, eye-tracking, focus of attention, gesture interfaces, natural interaction.

INTRODUCTION
Despite many years of research on Media Spaces, we are still far from developing technologies that would allow people to collaborate at distance with the same efficiency and ease than when face-to-face. Ethnographical observations from real work settings show that many solutions developed to support collaborative work at a distance are flawed as they *"fracture the relation between action and the relevant environment"* [14]. For example, using many video cameras to capture and share different points of view between two remote locations might seem to be an improvement over the use of a single camera. However, users might feel lost in the attempt to understand which view is the partner currently looking at or how to adapt common communication strategies to this multitude of perspectives. As Luff *et al.* explain [14, p. 73]: *"Ironically, the more we attempt to enhance the environment, the more we may exacerbate difficulties for the participants themselves in the production and coordination of action"*.

Similarly, our argument is that we need to find more subtle technological solutions to translate communication mechanisms which are effective in presence but not available when collaborators are not co-located. These solutions should allow to recreate the same functions using different but equivalent strategies. In this position paper we focus on two of these mechanisms, namely the awareness of the *focus of attention* and the use of *gestures*, particularly deictic gestures, to disambiguate references used during the interaction (*e.g.,* discussing blood test reports from different patients) or to better support comparisons between various information media (*e.g.,* combining a broken leg x-ray result with a plastic leg miniature so that the physician can point specific articulations in the former and manipulate the latter while explaining the injury cause).

We concentrate on these two elements because linguistic theories have shown that when we communicate the production of our elocution is inextricably linked to the responses of our audience. It is crucial for the speaker to monitor his/her audience for evidence of continued attention and understanding [2]. Clark [3] explained that communication is ordinarily anchored to the material world and that one way it gets anchored is through *pointing*. Pointing-to can be achieved through the linguistic channel using specific terms like "this", "he", or "here". Also, deixis can be produced through gestures. Furthermore, it has been suggested that deixis is intertwined with gaze awareness because face-to-face gestures can be perceived and acknowledged by recipients using gaze [4].

Research in this area has been extremely active in these last few years. As we will detail in the next section, scholars have designed and tested interfaces that support gaze awareness (*e.g.*, [16, 17]) and gesturing (*e.g.*, [8, 9, 10], etc.). However, we believe that there are three important aspects that still require further consideration and research. First, researchers should develop a careful *(a) integration between the communication modalities* because human communication is intrinsically multimodal. When we are face-to-face we tend to use both linguistic and non-linguistic channels to minimize the communicative effort and maximize the outcome of the interaction. Little research so far has compared different solutions to combine communication modes, and few studies have focused on the effect of different combinations on collaboration.

Second, designers should enable collaboration environments that *(b) ease users transition between digital and physical workspaces*. When we are co-located, we tend to use seamlessly (*i.e.*, embody) the space around us. We can point to digital artifacts on the screens of our computational devices and at the same time to physical objects located nearby. However, many telepresence prototypes designed in the past have been somewhat limited in letting users transition from

International Workshop New Frontiers in Telepresence 2010,
part of CSCW2010, Savannah, GA, USA, 7th of February, 2010.
http://research.microsoft.com/en-us/events/nft2010/

one modality to the other. Therefore, future work should focus in enhancing the capabilities of telepresence environments so to enable easy transition from digital environments to physical workspaces.

Finally, researchers should embrace a more *(c) flexible definition of roles within the context of collaboration*. Many environments for telepresence studied in the past were built following a helper-worker scenario where a remote expert could provide instructions to an on-site worker (*e.g.*, [10, 11, 16]). Although motivated by real situations of the use of communication technology, telepresence prototypes that assign static roles within the collaboration are unrealistic and create communication asymmetries that are generally non existing in face-to-face scenarios.

We will expand these three points in the next section. Subsequently, we will introduce a research framework that could be used to conduct further research on the above issues and we will discuss the expected outcomes of this research.

COLLABORATION IN DUAL SPACES

Effective collaboration requires participants to be able to communicate their intents, agree on a methodology to achieve their goals, share information and monitor the development of their interaction. Daly-Jones *et al.* [5] defined four pragmatic needs that must be fulfilled in human interaction: 1) the need to make contact; 2) the need to allocate turns for talking; 3) the need to monitor understanding and audience attention; and finally 4) the need to support deixis. The last two points are particularly interesting for the design of systems for remote collaboration, as we will detail in the next subsections.

Gaze and the Focus of Attention

Previous research has demonstrated how gaze is connected to attention and, in turn, to cognition []. Gaze is also used to marshal turn-taking. Therefore, the awareness of gaze is beneficial to collaboration because collaborators can use this communication modality to manage their interaction and to pinpoint the possible interpretations of a referent. A strict relation between gaze and collaborative work was demonstrated by Ishii and Kobayashi [9]. They showed that preserving the relative position of the participants and their gaze direction could be beneficial for cooperative problem solving. They used a system called ClearBoard, which allowed users to collaboratively sketch on a shared display while maintaining eye-contact. A similar setup was proposed by Monk and Gale [16], which they named GAZE system.

One of the limitations of the ClearBoard and GAZE prototypes was that of using half-silvered mirrors to merge the remote image of the user with the computer display. Unfortunately, users of these systems could see the reflection of their body and hands on top of the remote image. During the experiments, this factor emerged as bringing additional difficulties to the interaction. Additionally, users were forced to interact in front of the camera because of the technology that was used to capture the video. Therefore, their movements were constrained to the field of view of the cameras. Furthermore, using physical objects in addition to the digital objects on the display was somewhat complicated by the reflection and by the physical setups of the mirrors (see point (b) of the introduction).

Simpler techniques used in the past to provide users with proactive control over the focus of attention consisted in Media Spaces with multiple cameras. The users of these systems could operate a manual switch to choose which camera view was given to the remote user (see Gaver's *et al.* MTV system [7]). Note that in a face-to-face interaction, the control of the focus of attention is embodied and therefore it does require a minimal effort. However, when using the manual selectors in these systems users have to spend cognitive resources to keep track of which view is offered to the remote collaborator (see point (b) of the introduction). Furthermore, gaze awareness is naturally intertwined with face-to-face communication whereas at a distance these two modalities have been decoupled in many telepresence environments designed in the past (see point (a) of the introduction).

Supporting Gestures and Deixis

Gestures represent an extremely important communication mechanism that allows people to coordinate their efforts and disambiguate their contributions in the interaction. In the last twenty years, many solutions have been proposed to support remote gestures. Many of the early prototypes used video technology to capture and display the hands of the collaborators at the remote sites (*e.g.*, [9, 10]). As Luff *et al.* [14] clearly explained, video solutions suffered from a fracture of the ecology of the remote sites. In such systems, the gestures were fractured from the place where they were produced and where they were received. Restricted field of views and distortion of projection are just few examples of how video may hamper the usefulness of remote gestures.

An important limitation of these systems is due to the use of the unmediated video-capture of the hands to communicate the gestures to the remote collaborator, as remote collaborators had a hard time to infer when a particular gesture was associated to a communicative intent (see point (b) of the introduction). Other research lead by Kuzuoka [12] proposed the use of robots on the remote sites to re-embody the interaction of the remote collaborators. However, this research was conducted under the assumption that one of the collaborators was the "expert" while the other was the "novice". As highlighted in point (c) of the introduction, in face-to-face interactions these roles can switch multiple times during the task resolution. Therefore, systems that are designed around a static definition of the roles can introduce unnatural asymmetries in the interaction.

Other researchers have proposed prototypes where gestures are represented by digital metaphors like digitalized sketches or pointers. The underlying assumption of this work was that a sketch could incorporate features of a gesture that could suffice to replace the real gesture. More research is required to define what features of gestures are most important for communication, what kinds of gestures are necessary and in what circumstances. For instance, Gutwin and Greemberg [8] have reported mixed results in supporting workspace awareness for collaborative work at a distance. One of their solutions included a *minimap* of the interface of

the remote person that was shown as part of their system's interface. This visualization reported in a schematic way the basic elements of the interface plus the information of where the other was using his/her mouse pointer (*i.e.*, *telepointer*). Other evaluations of the telepointer mechanism as a gesturing device have reported negative results because the cursor activity could not be always related to the user's intention, attention, or presence. Essentially, this type of tools present some embodiment issues that lead to users' access and control disparities (see point (b) of the introduction).

PROPOSED METHODOLOGY

In the last few years, computer vision techniques for real-time video processing have evolved quickly. We see four major advancements that can help develop the new class of telepresence environments:

(1). Several companies have released on the market *eye-tracking* solutions that can trace the point of focus of a person's gaze moving freely in a room. Xuuk Inc. recently released a long-range eye sensor, called eyebox2 which is able to detect a user's gaze from up to 10 meters of distance[1]. As discussed in the previous section, a person's gaze can be mapped onto the focus of attention. Hence, being able to automatically detect what the users are looking at is extremely valuable to support interaction at a distance.

(2). Real-time computer vision algorithms combined with sophisticated cameras enable the detection and tracking of the hands and the *recognition of the gestures* that a person is producing. For instance, Miralles *et al.* [15] have analyzed and tested a number of metaphors for the definition of a gestural language to operate an interface as part of the Spanish-funded VISION project. The camera used in their studies was originally produced by 3DV Systems Ltd[2]. Recently the company was bought by Microsoft and the technology was incorporated in the latest version of the Xbox gaming console. The advantage of using this technology is that instead of displaying the unmediated video of the hands to the remote site, the *gist* of the gestures that are mostly important can be captured and highlighted on the remote site (*i.e.*, gestures that are associated to communicative intents).

(3). 3D computer vision techniques allow to *reconstruct the three-dimensional volume of static and moving objects* in a certain scene from multiple camera-views [13]. These techniques could be used to infer the 3D position of the person in a certain environment (*i.e.*, body tracking) and the objects s/he is interacting with. This information could be combined with other sources of information about the user's activity to build models of the user's actions and intentionality, and to discern what to show or represent on the remote site.

(4). Finally, advancements of vision techniques might allow to provide *video devices that can avoid the distortion of the gaze direction*. Within the EU-funded 3DPresence project a multi-perspective auto-stereoscopic 3D display has been developed that is able to correct directional eye-contact and to render proper sense of perspective in videoconferencing [6]. Using this technology, conferees can feel simultaneously, and individually, whether they are being looked by other conference participants.

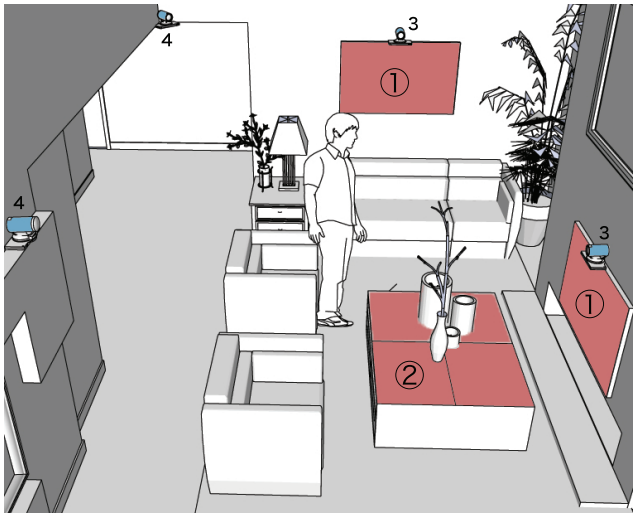

Figure 1. Possible experimental setup: (1) screens used to project remote information, (2) shared workspace, 3 primary cameras used for eye-tracking, 4 additional cameras used for 3D reconstruction of static and moving objects

In summary, the three main points of our proposal are: (1) *Multimodality*: We intend to instrument an experimental environment where the three previously described technologies can co-exist and complement each other to capture and model the user's activity (see Figure 1); (2) *Active Focus of Attention*: Instead of limiting the communication to one modality, the telepresence system should automatically detect and direct the remote participant – by means of the right communication modality – to the part of the scene that is most relevant to the current situation. Encouraging results in this sense have been obtained by Ranjan *et al.* [18]; and (3) *Evaluation*: We plan to test different combinations of communication modalities, particularly analyzing the effects of these different mixes on efficiency and ease of use. We initially plan to measure the efficiency of the telepresence solution using task resolution time and we will measure ease of use using the NASA TLX tool for measuring task load.

Our goal is to conduct a controlled experiment using a factorial design where we manipulate the availability of *non-verbal communication* – GAZE and GESTURE – and the *way this non-verbal communication is transferred* to the remote site – UNMEDIATED, when the continuous feed of this information is provided and GIST when the information is processed to identify relevant episodes (*e.g.*, one of the participants indicates an object and says "this"). Finally, we would like to combine a factor related to the *strategy used to show the focus of the interaction* between the participants – MANUAL when it is left to the user's choice, AUTOMATIC when it is driven by the user's model, and SEMI-AUTOMATIC when it is set as a compromise between the two. Our experiment will be similar to that of Monk and Gale [16] but with more sophisticated technology and combining more factors in the experimental design.

[1] See `https://www.xuuk.com/`, last retrieved November 2009.

[2] See `http://www.3dvsystems.com/`, last retrieved November 2009.

EXPECTED OUTCOMES AND CONCLUSIONS

The experiment of Monk and Gale [16] showed that *gaze awareness* reduced the number of turns and number of words required to complete a task. However, the setup they used forced the user to sit in a certain position and at a certain distance from the half-mirrored screen. Also, subjects could see all the time the position of the eyes of the other participants, even when gaze was not associated to communicative intent. We believe that redesigning this experiment with modern technology might provide concluding evidences on the usefulness of gaze awareness on collaborative work at a distance.

Furthermore, we believe that further research is necessary to understand how gaze awareness should be combined with *gesture recognition* and the effect of the availability of both modalities on remote collaboration. While it is clear that support for gaze awareness and gestures is important in CSCW-environments, there is evidence that providing this information continuously might be detrimental to problem-solving, as it increases the amount of effort required to the participants. We believe that the experiments proposed in this paper will yield relevant implications in the design of mechanisms that could *mediate* the representation of both non-verbal communication modalities.

For instance, Cherubini *et al.* [1] demonstrated that the probability of misunderstandings between distant collaborators in problem solving task is related to the distance between the collaborators' focus of gaze over the shared workspace. Therefore, Cherubini's results support a telepresence solution where the focus of gaze is shown to the remote site *only* when the system infers that there might be a misunderstanding between the collaborators. The experiment we are proposing here will compare this mediated modality of representing gaze and gesture to the remote site with the unmediated approach.

To conclude, this paper briefly describes some of the telepresence ideas we plan to work on in the near future. We intend to contribute to the workshop with these concepts and to receive relevant feedback on our proposed approach. Finally, we hope to stimulate rich discussions on the future of telepresence.

ACKNOWLEDGMENTS

Telefónica I+D participates in Torres Quevedo subprogram (MICINN), cofinanced by the European Social Fund, for Researchers recruitment.

REFERENCES

1. M. Cherubini, M.-A. Nüssli, and P. Dillenbourg. Deixis and gaze in collaborative work at a distance (over a shared map): a computational model to detect misunderstandings. In *Proc. ETRA'08)*, pages 173–180, Savannah, GA, USA, March 26-28 2008.
2. H. H. Clark and S. E. Brennan. *In L. Resnick, J. Levine and S. Teasley, editors, Perspectives on Socially Shared Cognition*, chapter Grounding in Communication, pages 127–149. American Psychological Association, Washington, 1991.
3. H. H. Clark. *Pointing: Where language, culture, and cognition meet*, chapter Pointing and placing, pages 243–268. Lawrence Erlbaum Associates, Mahwah, NJ, USA, 2003.
4. H. H. Clark and M. A. Krych. Speaking while monitoring addressees for understanding. *J. of Memory and Language*, (50):62–81, 2004.
5. O. Daly-Jones, A. F. Monk, and L. Watts. Some advantages of video conferencing over high-quality audio conferencing: fluency and awareness of attentional focus. *Int. J. Hum.-Comput. Stud.*, 49(1):21–58, 1998.
6. Ò. Divorra, W. IJsselsteijn, O. Schreer, *et al*. Towards 3D-Aware Telepresence: Working on Technologies Behind the Scene. *Paper presented at the International Workshop New Frontiers in Telepresence*, part of CSCW'10 (G. Venolia, K. Inkpen, J. Olson, and D. Nguyen, eds.), (Savannah, GA, USA), February 7th 2010.
7. W. Gaver, A. J. Sellen, C. Heath, and P. Luff. One is not enough: Multiple views in a media space. In *Proc. of INTERCHI'93*, pages 335–341, Amsterdam, The Nederlands, April 24-29 1993.
8. C. Gutwin and S. Greenberg. The importance of awareness for team cognition in distributed collaboration. In E. Salas, S. Fiore, and J. Cannon-Bowers, editors, *Team Cognition: Understanding the Factors that Drives Process and Performance*, pages 177–201, Washington, 2004. APA Press.
9. H. Ishii and M. Kobayashi. Clearboard: A seamless medium for shared drawing and conversation with eye contact. In *Proc. of CHI'92*, pages 525–532, Monterey, CA, USA, May 3-7 1992.
10. D. S. Kirk, T. Rodden, and D. S. Fraser. Turn it this way: grounding collaborative action with remote gestures. In *Proc. of CHI '07*, pages 1039–1048, New York, NY, USA, 2007.
11. H. Kuzuoka, T. Kosuge, and K. Tanaka. Gesturecam: A video communication system for sumpathetic remote collaboration. In *Proc. CSCW '94*, pages 35–43, Chapel Hill, North Carolina, United States, 1994.
12. H. Kuzuoka, K. Yamazaki, A. Yamazaki, J. Kosaka, Y. Suga, and C. Heath. Dual ecologies of robot as communication media: thoughts on coordinating orientations and projectability. In *Proc. CHI '04*, pages 183–190, Vienna, Austria, 2004.
13. J. L. Landabaso and M. Pardas. A unified framework for consistent 2D/3D foreground object detection. *Circuits and Systems for Video Technology*, 18(8):1040–1051, 2008.
14. P. Luff, C. Heath, H. Kuzuoka, J. Hindmarsh, K. Yamazaki, and S. Oyama. Fractured ecologies: Creating environments for collaboration. *Human-Computer Interaction*, 18(1&2):51–84, 2003.
15. I. Miralles, M. Jorquera, C. Botella, R. Banos, J. Montesa, C. Ferran, and D. Miralles. Analysis and testing of metaphors for the definition of a gestural language based on real users interaction: Vision project. In *Proc. of HCI International'09*, pages 67–70, San Diego, CA, USA, July 19-24 2009.
16. A. F. Monk and C. Gale. A look is worth a thousand words: Full gaze awareness in video-mediated conversation. *Discourse Processes*, 33(3):257–278, 2002.
17. P. Qvarfordt, D. Beymer, and S. Zhai. Realtourist: A study of augmenting human-human and human-computer dialogue with eye-gaze overlay. In *Proc. of INTERACT'05*, Rome, Italy, 2005.
18. A. Ranjan, J. P. Birnholtz, and R. Balakrishnan. Dynamic shared visual spaces: experimenting with automatic camera control in a remote repair task. In *Proc. CHI'07*, pages 1177–1186, New York, NY, USA, 2007.