Enhancing Human Cooperation with Multimodal Augmented Reality

Christian Mertes, Angelika Dierker,
Thomas Hermann, Marc Hanheide, and Gerhard Sagerer

Bielefeld University, P.O. Box 10 01 31, 33501 Bielefeld, Germany,
{cmertes, adierker, thermann, mhanheid,
sagerer}@TechFak.Uni-Bielefeld.DE,
http://www.cit-ec.de/

Abstract. Humans naturally use an impressive variety of ways to communicate. In this work, we investigate the possibilities of complementing these natural communication channels with artificial ones. For this, augmented reality is used as a technique to add synthetic visual and auditory stimuli to people's perception. A system for the mutual display of the gaze direction of two interactants is presented and its acceptance is shown through a study. Finally, future possibilities of promoting this novel concept of artificial communication channels are explored.

Keywords: Augmented Reality, Joint Attention, Sonification, Artificial Communication Channels, Gaze Direction

1 Introduction

Humans use not only speech to communicate very effectively but also a large variety of non-verbal communication channels (a good overview is given by Knapp and Hall [1]). However, technology has advanced to the point where it becomes interesting to think about complementing these natural communication skills with what we call *artificial communication channels*. In the following, we will present ways to use gaze direction as such an artificial channel in face to face communication using augmented reality (AR) as a display device. Firstly, a brief overview on possible ways to convey gaze direction will be given. In Section 2 we will present the hardware we use and our implementation of these channels. In Section 3 we will briefly look at the results of a preliminary user study and finally in Section 4 we will discuss further possible developments.

1.1 Gaze Direction Displays

There are many possible communication cues that can be measured or detected and then displayed in one way or another, especially considering AR as a very immediate display mode [2]. We chose gaze direction because it conveys very useful information on spacial attention that can be displayed in intuitive ways on head-mounted displays. Furthermore, a simplified form of gaze direction is

easy to measure when already using video see-through AR goggles, assuming the optical axis of the camera to be the center of visual attention. Pointing gestures are a very similar communication cue in many regards but their recognition is not a typical by-product of AR systems.

Leaving picture-in-picture displays aside, we propose two main criteria to classify gaze direction displays. The first one distinguishes between the two types *discrete* (with regard to the objects being seen) and *continuous*. Thereby a continuous display is one that shows the field of view itself, while a discrete display is only concerned with certain objects of interest within this field of view (FOV). This distinction is applicable to visual as well as auditory displays, although the area or volume that represents the FOV in the continuous type is difficult to convey in sound so only its center will typically be used.

The second criterion is the presence or absence of a temporal dynamic. On the one hand, the addition of a kind of memory (or *decay envelope*) to the display adds utility in cases where an interactant A is not looking at B's focus of attention at a given point in time but is drawn to it by some reference B makes. Now if B already directed his or her attention to another spot (or—in the related case of pointing gesture display—ended the gesture) when A looks there eventually, some kind of history might be helpful. On the other hand, an *attack envelope* (i.e. a gradual rise of highlight intensity) smoothes out very short saccade-like movements or can otherwise convey a sense of the duration of the attention. As with the first criterion, the application of some kind of temporal dynamic is not bound to a modality but in most cases the auditory domain will be more prone to becoming cluttered than the visual one.

2 System Overview

From the above-mentioned possibilities we implemented a discrete visual display with and without memory, a continuous visual display without memory and both a discrete and a continuous auditory display without memory. As a framework we use the *Interception Interface* detailed in Dierker et al. [3]. It uses the Augmented Reality Toolkit [4] to show virtual objects on top of fiduciary markers. This enables us to highlight the objects easily, for example by changing their color. These fiduciary markers also provide all the information needed for the discrete displays while the continuous displays need a head tracking system.

The discrete visualization highlights the virtual objects that are in the interaction partner's FOV by changing their color. An optional temporal dynamic can be configured by choosing three different envelope types and arbitrary durations for both the attack and the decay phase. The discrete sonification simply plays two different sounds when an object enters or leaves the partner's FOV respectively.

The continuous visualization consists of a projection of the partner's FOV onto the table surface on which the interaction takes place in our current scenario (i.e. a quadrilateral on the table top). The continuous sonification is a one-to-

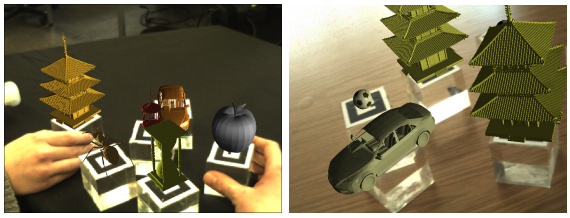

Fig. 1. Screenshots of the image a user of the system might see

one mapping of certain attributes of the line of sight to parameters of a sound synthesis[1] (a so-called *parameter mapping sonification* [5]).

We use Trivisio video see-through AR goggles with a resolution of 640 × 480 pixels. For each goggle, one of its two video streams is fed through a laptop computer for the augmentation (the other channel is currently unused). For the auditory augmentation we use headphones and for the augmentation modes which require head positions instead of only the location of the objects within the subjects' FOVs we use a Vicon tracking system with passive markers on top of the goggles.

3 Evaluation

We tested the basic effectiveness with a very simple task we call the *gaze game*. One of the two players stares at a certain object. The other player has to find the same object and look at it as fast as possible. Both players are not allowed to speak or gesture during the task. We tested two conditions: One with the object highlighting and the event-triggered sounds and one without both. The goggles and headphones had to be worn in both cases. Only these discrete augmentations were tested in this first study. We measured speed and error rate and had the participants fill out a questionnaire after the experiment.

What we find most prominent is a remarkable gap between the perception of helpfulness of the augmentations as the subjects experienced it themselves and the actual measured performance gain: From the 16 subjects all stated to have used the highlighting when it was available and 93.75 % perceived it as helpful or very helpful. No one had the impression of having used the auditory augmentation and nobody perceived it to be helpful. Despite this perceived

[1] The speed of the head movement is mapped to the amplitude, the x axis intercept to the panning (where x is the axis parallel to the table edge the user is sitting at), the proximity of the partner's center of focus to the own center of focus is mapped to the consonance of the sound, and the z axis intercept to the fundamental frequency (z being orthogonal to the table top).

(a) (b)

Fig. 2. (a) Our AR setup and (b) two users of the system during the familiarization phase that preceded each trial

helpfulness of at least the visual augmentation, there was no significant decrease in search time nor a significant drop in errors being made.[2]

3.1 Discussion

Possible explanations for the gap between perceived and measured helpfulness include the possibility of learning the object positions by heart, the possibility of "cheating" by bypassing the goggles[3], a somewhat unfavorable virtual lighting that made the colors of some objects hard to discern, the large distance between the objects that played into the hands of the non-highlight condition and the fact that objects seen peripherally were highlighted as much as objects in the center of the FOV even when not consciously perceived by a subject.

An already implemented but not yet evaluated change of highlighting color depending on the laterality of an object in the partner's FOV might help with this last problem. This could enable putting the objects closer together which in turn might be a more advantageous situation for our system. Randomizing the object positions for every turn should help against the memorization. The continuous FOV display conveys a totally different set of information to the user and it will be very interesting to see how it performs.

The auditory display on the other hand probably suffered from a lack of conveyed information, the unusual mode of presentation and possibly from the subconscious workings of sound which might have led the subjects to further underestimate any small utility that might have been there. The continuous sonification conveys much more information and will be a more useful test of sound as a medium to display gaze direction.

[2] This excludes three pairs of subjects who—for a lack of prescribed methods—did not come up with the idea of looking into their partner's eyes and fell back to guessing.
[3] This "cheating" can be desirable for certain kinds of scenarios though.

4 Conclusion

We used AR to display the gaze direction of two interactants to each other in four different ways. We evaluated the discrete (object-based) visualization and sonification and found that almost everybody perceived the visual display to be helpful although no effect on the performance was found. The sonification was not perceived to be helpful. The continuous visualization and sonification have yet to be evaluated and there are some possible improvements to the already tested methods.

4.1 Outlook

The presented system provides a good basis for much further research, some of which has already been mentioned in Section 3. Moving on to more realistic scenarios will be an important next challenge. Less obtrusive hardware is very important with regard to usage outside of teleconferencing applications and scenarios where AR is already in use anyway. Recent prototypes look promising in this regard. Other promising extensions to our system include enabling it to support more than two users at once, possibly using different colors, making the users more mobile to have them move about instead of being more or less confined to a small interaction space and adding gesture display. The system might also be a useful tool for other fields such as communication research.

Finally, artificial communication channels might incorporate different kinds of sensor data such as information about movement or posture, voice or speech, vocal characterizers (as defined by Knapp and Hall [1]), physiological signals, worn objects or the environment and might also be extended to remote users or virtual agents.

Acknowledgments. We would like to thank the DFG and its Sonderforschungsbereich 673 *Alignment in Communication* for supporting this work.

References

1. L. Knapp, M., A. Hall, J.: Nonverbal Communication in Human Interaction. 5 edn. Wadsworth/Thomson Learning (2001)
2. Mertes, C.: Multimodal augmented reality to enhance human communication. Master's thesis, Bielefeld University (August 2008)
3. Dierker, A., Bovermann, T., Hanheide, M., Hermann, T., Sagerer, G.: A multimodal augmented reality system for alignment research. In: Proceedings of the 13th International Conference on Human-Computer Interaction. (2009)
4. Kato, H., Billinghurst, M.: Marker tracking and hmd calibration for a video-based augmented reality conferencing system. Proceedings of the 2nd IEEE and ACM International Workshop on Augmented Reality **99** (1999) 85–94
5. Hermann, T.: Taxonomy and definitions for sonification and auditory display. In: Proceedings of the 14th International Conference on Auditory Display. (2008)