Model-based Acquisition and Analysis of Multimodal Interactions for Improving Human-Robot Interaction

Patrick Renner*
Artificial Intelligence Group
Faculty of Technology, Bielefeld University, Germany

Thies Pfeiffer†
CITEC - Cognitive Interaction Technology
Faculty of Technology, Bielefeld University, Germany

Figure 1: *The scenario: The task is to plan routes on floor plans of a larger building. Human communicative behaviour including gaze and gestures are recorded using mobile eye tracking glasses and gloves with optical marker tracking.*

Abstract

For solving complex tasks cooperatively in close interaction with robots, they need to understand natural human communication. To achieve this, robots could benefit from a deeper understanding of the processes that humans use for successful communication. Such skills can be studied by investigating human face-to-face interactions in complex tasks. In our work the focus lies on shared-space interactions in a path planning task and thus 3D gaze directions and hand movements are of particular interest.

However, the analysis of gaze and gestures is a time-consuming task: Usually, manual annotation of the eye tracker's scene camera video is necessary in a frame-by-frame manner. To tackle this issue, based on the EyeSee3D method, an automatic approach for annotating interactions is presented: A combination of geometric modeling and 3D marker tracking serves to align real world stimuli with virtual proxies. This is done based on the scene camera images of the mobile eye tracker alone. In addition to the EyeSee3D approach, face detection is used to automatically detect fixations on the interlocutor. For the acquisition of the gestures, an optical marker tracking system is integrated and fused in the multimodal representation of the communicative situation.

CR Categories: I.4.8 [Image Processing and Computer Vision]: Scene Analysis—Tracking; H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems—Artificial, augmented, and virtual realities;

Keywords: 3D gaze analysis, eye tracking, motion tracking, marker tracking, geometric modelling

*e-mail: prenner@techfak.uni-bielefeld.de
†e-mail: Thies.Pfeiffer@uni-bielefeld.de

1 Introduction

More and more, robotics approaches to our life: As of today, the first consumer-tailored roboters are available. In the near future, robots will have skills for helping humans in different situations. It won't be possible anymore to just operate these robots by buttons or a small display: For solving complex tasks cooperatively in close interaction with robots, they need to understand natural human communication.

This raises completely new requirements regarding a robot's skills in interacting. In cooperation, a robot has to solve a task jointly with a human partner in a shared space. We are able to reach our goals quickly and precisely in face-to-face interaction. Especially gaze behaviour is important for implicitly communicating our attention and thus intention, so that a common goal can be reached. The mechanism of joint attention [Tomasello et al. 2005], e.g., is well-studied in human-human interaction.

To identify appropriate communication strategies which enable robots to communicate more smoothly requires studying humans in free interaction. In our work, the focus lies on interactions in shared space. The experiment which motivated the presented methodical developments investigates face-to-face interaction in a route planning scenario: Participants are to plan paths to rooms on three different floors on floor plans located on a table between them (Fig. 1, bottom).

However, there is one general problem with studying gaze behaviour in dynamic interactions: To learn about the timing and sequences of visual attention on the objects of interest, mobile eye tracking is used to allow study participants to move their head freely. But this comes along with time-consuming manual annotation of gaze videos. The EyeSee3D tool [Pfeiffer and Renner 2014] tackles this problem by combining geometric modeling of the stimuli (floor plans, hands and faces) and 3D marker tracking using only the scene camera image of the eye tracking device. This way, the virtual model can be used to identify the object of interest in real-time without further manual annotations. In our work we extend the EyeSee3D approach to more complex scenarios: Our target stimuli on the floor plans are more densely packed, resulting in a high number of stimuli. In addition to that, we are also interested in gaze on

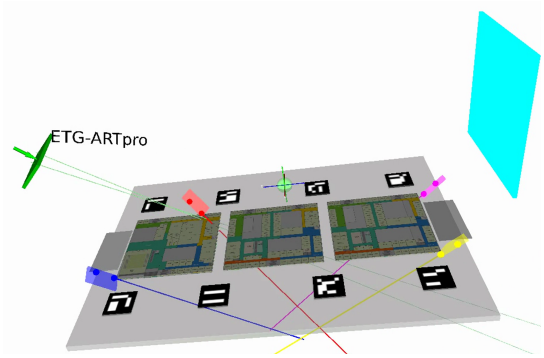

Figure 2: *The 3D representation of the scenario including three floor plans, the fiducial markers, participants' hands with highlighted pointing direction as well as the positions of the interlocutors' faces.*

Figure 3: *Overlay of the virtual representation over the scene camera image of the eye tracking device. The position of the participant's face is calculated by a face detection algorithm.*

the hands and the face of the interlocutor. We thus have to fuse several modalities in a shared multimodal representation.

In a natural face-to-face interaction, the interlocutors' heads move around and can thus not be modeled by a static proxy object. In order to detect fixations on the interlocutor's face, we thus applied a head tracking algorithm to the scene camera video and mapped the result into the virtual model of the scene.

2 Method

Normally, mobile eye trackers can detect 2D fixations relative to a fixed plane, e.g. the scene camera video. In addition, most devices calculate a 3D gaze vector relative to the scene camera.

The EyeSee3D approach allows for automatically assigning fixations to stimuli, only using the scene camera video and the 3D gaze vector as input. This is done by tracking simple fiducial markers (Fig. 1, middle left) in the video. When the scene camera is calibrated, i.e. its intrinsic parameters are determined, markers found in the 2D images can be correctly transformed to 3D with respect to position and orientation. Thus, the pose of the scene camera with respect to the stimuli can be calculated. By re-modeling the stimuli in virtual reality and representing the 3D gaze vector as rays, fixations can be assigned to the stimuli automatically by intersection testing. The generated model can be viewed from all directions (Fig. 2), or it can be aligned to the real scenario and can be overlayed over the scene-camera image (Fig. 3). This way, the experimenter can check for correctness of the data during runtime.

For the planning scenario of the human-human study, we added several extensions to EyeSee3D. In addition to using gaze directions as single input we are also interested in pointing directions. To detect these accurately, we integrated an external tracking system. Participants have to wear light gloves (Fig. 1, left) which are tracked. The link between marker tracking and external tracking is established by placing a tracking target with known position and orientation relative to the fiducial markers. This way, both inputs can be fused in a multimodal 3D representation of the interaction scenario.

The floor plans contain a high number of rooms. Instead of modelling each of these separately (which would be necessary as each room is a stimulus), a technique from web design was made use of: The plans were represented as webpages with HTML image maps, i.e. the coordinates of rooms and areas were annotated for the images. Each annotated area can be provided with a specific text, here the name of the area, that is output when the user fixates the area on the floor plan. This approach reduces modeling effort and increases system performance.

We are also interested in fixations on the interlocutor's face, but in this free kind of interaction head positions are not fixed. To overcome this issue, we detect faces in the scene camera images of the mobile eye tracker using the Viola/Jones algorithm [Viola and Jones 2004]. To complete the 3D model, the face position is approximated in 3D as well, which can be seen in Fig. 2 and 3.

3 Conclusion

We present our extensions of the EyeSee3D method: External hand tracking is integrated into the system, complex surfaces are handled by the use of image maps and faces are detected in the scene camera images. All this is fused to a multimodal 3D representation of the interaction scenario which supports an automatic annotation of objects of interest.

The system is capable of running in real time. Thus, the correctness of experiment data can be verified during runtime; the results are presented directly after an experiment is finished. Moreover, offline analysis is possible, e.g., if model or algorithms need to be changed after a study is conducted.

As a convenient side-effect, the frame-by-frame knowledge about the interlocutor's head position enables us to anonymize experiment videos automatically.

References

PFEIFFER, T., AND RENNER, P. 2014. EyeSee3D: A Low-Cost Approach for Analysing Mobile 3D Eye Tracking Data Using Computer Vision and Augmented Reality Technology. In *Proceedings of the 2014 Symposium on Eye-Tracking Research and Applications, ETRA 2014*. Accepted.

TOMASELLO, M., CARPENTER, M., CALL, J., BEHNE, T., AND MOLL, H. 2005. Understanding and sharing intentions: the origins of cultural cognition. *The Behavioral and brain sciences 28*, 5, 675–91; discussion 691–735.

VIOLA, P., AND JONES, M. J. 2004. Robust Real-Time Face Detection. *International Journal of Computer Vision 57*, 2, 137–154.