Towards Gaze Interaction in Immersive Virtual Reality: Evaluation of a Monocular Eye Tracking Set-Up

Thies Pfeiffer[1]

[1] tpfeiffe@techfak.uni-bielefeld.de
AI Group, Faculty of Technology
Bielefeld University, Germany

Abstract: Of all senses, it is visual perception that is predominantly deluded in Virtual Realities. Yet, the eyes of the observer, despite the fact that they are the fastest perceivable moving body part, have gotten relatively little attention as an interaction modality. A solid integration of gaze, however, provides great opportunities for implicit and explicit human-computer interaction. We present our work on integrating a lightweight head-mounted eye tracking system in a CAVE-like Virtual Reality Set-Up and provide promising data from a user study on the achieved accuracy and latency.

Keywords: human-computer interaction, virtual reality, eye tracking, monocular

1 Introduction

Computer Graphics and Virtual Reality (VR) primarily target the eyes of the user as a sensory organ. However, the human eyes are by no means a passive receptor, but rather are actively used to explore visual scenes in a rapid manner. Moreover, they are also considerably faster than speech or gestures.

Gaze has been used in human-computer interaction since 1979 [TKFW+79]. Since then, gaze-typing applications allow physically challenged people to interact with computers. Supported by a boost in desktop processing power, customer video-based eye tracking units started to provide near real-time access to gaze direction in the late 1980s. Today, eye trackers have evolved to feasible input devices. Today, portable head-mounted eye trackers are available and the user is neither required to remain stable, i.e., seated on a chair, nor, as in some cases, bound to use a chin rest. Eye tracking therefore has also become an attractive input methodology for Augmented and Virtual Reality. In fact, gaze has played a role in VR research at least since 1981, if we consider Bolt's concept of Gaze-Orchestrated Dynamic Windows [Bol81]. However, today most gaze input approaches in VR use Head-Mounted-Displays as projection units.

Of the many interesting applications for gaze-based interaction technology, we at the AI group at Bielefeld University are especially interested in communication, both computer mediated human-human interactions and human-agent interactions. Here, information about

the gaze of the user allows for establishing eye contact, for ensuring mutual understanding, and for recognizing turn-taking signals to control interaction in dialogues.

After a short overview of related work, we present our work on integrating a head-mounted eye tracking system in a CAVE-like VR Set-Up. In this context, we describe relevant hardware and software, as well as the required procedures. Finally, we present data on latency and accuracy from a small user study to demonstrate the applicability of our approach.

2 Related Work

Knowledge about the direction of gaze can inform off-line and online optimizations of VR systems. If saliently done, the results are either computationally less expensive but indistinguishable for the human observer or constitute a noticeably visual improvement. Examples have been demonstrated by Luebke et al. [LHNW00] who reduce the complexity of geometry models in the periphery of the gaze direction without noticeable effects to the observer. More recently, Hillaire et al. [HLCC08] demonstrated how online gaze-tracking can be used to improve visual effects such as depth-of-field and camera motions.

These approaches cover more indirect interactions of a human observer with the machine. However, gaze can also be used in direct interaction. Tanriverdi and Jacob [TJ00] attest a significantly faster object selection for their gaze based approach than compared to a pointing gesture. Their algorithm combines the picking algorithm provided by SGI Performer with a histogram-based approach, counting the relative frequencies of fixations per object and selecting the most frequently fixated object within a time window.

Tanriverdi and Jacob use the gaze position projected onto a 2D plane as basis for the picking ray. Similar approaches are followed by Duchowski et al. [DMC+02] and Barabas et al. [BGA+04]. They anchor the ray in the position of the eye or the head and project it through a fixation on a 2D plane, which is defined by the plane of projection.

Knowledge about the user's visual attention can also be used to facilitate mediated human-to-human interaction in VR environments. Examples have been demonstrated by Duchowski et al. [DCC+04], who apply the eye movements of a user onto a virtual avatar and show advantages of a visible line of sight for the communication of references to objects.

Modeling human visual perception as a ray can only be a simplification. In reality, the visual field extends vertically and horizontally, the latter even more than the former, and also in depth the eyes cannot see equally well all along the visual axis. An alternative approach to ray-based modeling based on self-organizing maps has been proposed by Essig et al.[EPR06], which we successfully validated in an earlier desktop-based VR setting [PDLW07].

The approach we present here is a first step towards transferring these findings to fully immersive settings with a freely moving user. Such an environment creates several challenges:

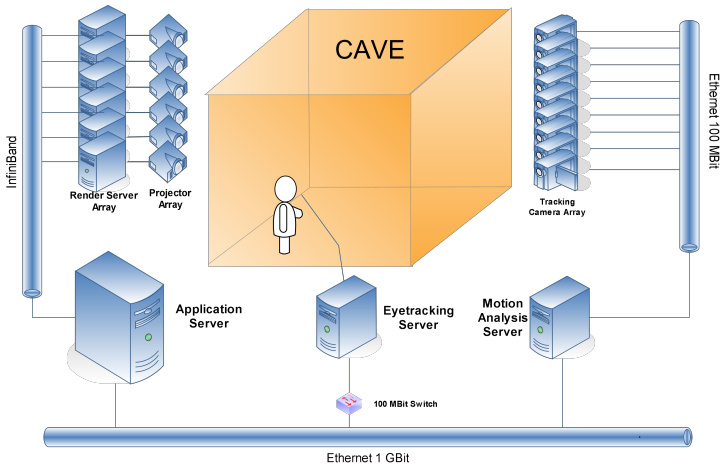


Figure 1: *The immersive VR Set-Up at the AI group at Bielefeld University.*

How is gaze tracking integrated on the software and the hardware level? How is the system calibrated in the presence of a freely moving user? How is the gaze evaluated in the application context? And how does the system perform? In the following, we will provide answers to those questions by describing our own approach and by providing data on accuracy and latency of the system from a small user study. In this first step, we will present data on monocular gaze tracking with the traditional ray-based model for gaze focus.

3 Hardware Set-Up

In the following, we will describe the hardware set-up of the CAVE-like VR installation at the AI group at Bielefeld University (see Figure 1). In view of the user study, we will provide accurate details about the hardware and software we used to allow for comparison and replication.

3.1 VR Application and Rendering

The VR application is driven by AVANGO [Tra01] on a dual AMD Opteron 248 2.2GHz machine with 3GB RAM. The views are distributed by Chromium [HHN+02] (version 1.6) to six render clients, each with AMD Athlon 64 3000+, 1GB RAM, and a NVIDIA Quadro FX 5600 card. They are running Ubuntu with a Linux Kernel 2.6.20 and a NVIDIA Kernel Module version 100.14.19. The cluster is networked by InfiniBand using Mellanox Technologies MT25204.

	ViewPoint PC-60
temp. res. (Hz)	30 / **60**
opt. res. (pixel)	640×480 / **320×240**
accuracy	$0,25\,^\circ$ - $1,0\,^\circ$
precision	$0,15\,^\circ$

(a)

(b)

Figure 2: *(a) Technical details of the ViewPoint PC-60 eye tracker from Arrington Research. The configuration used in the study is printed in bold.(b) Optical markers of the head tracking system and filters for polarized light have been mounted on the eye tracker.*

3.2 Eye Tracking

We were looking for an eye tracking device that could be easily combined with the markers for the optical tracking and the polarized filters for the projection. In the end, we settled on the ViewPoint PC-60 EyeFrame BS007 eye tracker (see Figure 2(b)) manufactured by Arrington Research [Inc08], as it fits with our requirements while being quite affordable.

The ViewPoint PC-60 offers moderate resolutions in time and space (see Table 2(a)), which should be adequate for normal interaction tasks (disregading saccades or microsaccades). The eye tracker is driven by the ViewPoint software in version 2.8.3,33 on an Intel Core2Duo 6600 machine with 2.4 GHz running Windows XP Professional SP 2. The machine is connected to the VR application via a 100 MBit Ethernet connection.

3.3 Motion Tracking

For body tracking, we use the marker-based optical tracking system ARTtrack1 manufactured by A.R.T. [artGA08]. To track the user's head, we attached a 6DOF marker to the left side of the frame of the ViewPoint PC-60 (see Figure 2(b)). Our tracking set-up comprises nine ARTrack1 cameras which ensure a stable presence of the 6DOF marker in at least three cameras during the planned interactions.

4 Software Framework

Our design of the software framework is targeted at scenegraph-based VR applications. The data from the eye and body tracking nodes is sent to the application via Ethernet. The application connects to the two sensor nodes and realizes the visualization. The visualization is distributed via Chromium over InfiniBand to our six render clients (see Figure 1).

4.1 Sensor Interface

For the interaction handling, the application gets the current eye position in 2D coordinates relative to the plane the eyes have been calibrated to. The position and orientation of the head is provided relative to the coordinate system of the ARTtrack tracking system, which is mapped to the origin of the root coordinate system. Incoming sensor data is managed by special purpose nodes, e.g. the *Eye Data Client*, in the Avango framework. They deserialize the networked data and provide them in their field interface to subscribing nodes within the scenegraph, following the data-flow paradigm which is orthogonal to the scenegraph hierarchy.

4.2 Representation and Integration

A single eye is represented in the scenegraph by an *EyeNode*. This node is parented by the *HeadActuator* which represents head position and orientation in the scene as provided by the ARTtrack1 system. In normal operation mode the *EyeNode* computes a gaze-direction by projecting a ray from the position of the eye through the detected 2D eye position relative to the plane of calibration.

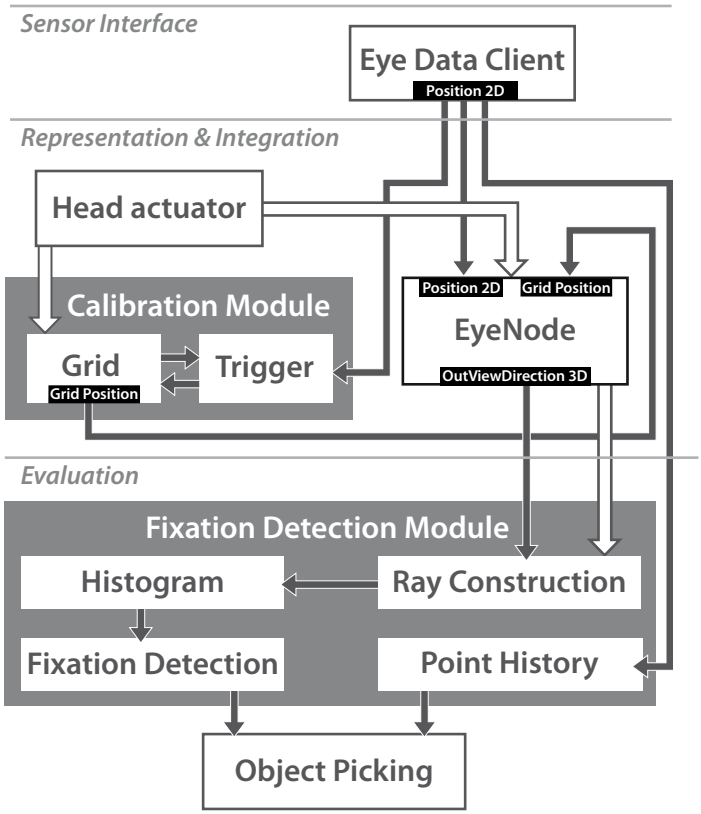


Figure 3: *The flow of data within the scenegraph.*

4.3 Calibration Module

The most interesting problem of the integration of the eye tracking in the VR Set-Up has been the calibration of the eye tracking system. In normal operation mode, i.e. when tracking gaze on the image of the scene camera, the system is calibrated using a relatively complicated procedure: a scene camera streams live video from the perspective of the user (between the eyes) to the operator's display. This display is then overlaid with a grid of calibration points, which are iteratively highlighted throughout the procedure. For each grid point the operator has to present a target to the participant at the location corresponding to the grid point on the display. The operator then waits for the participant to focus on the target and presses a button.

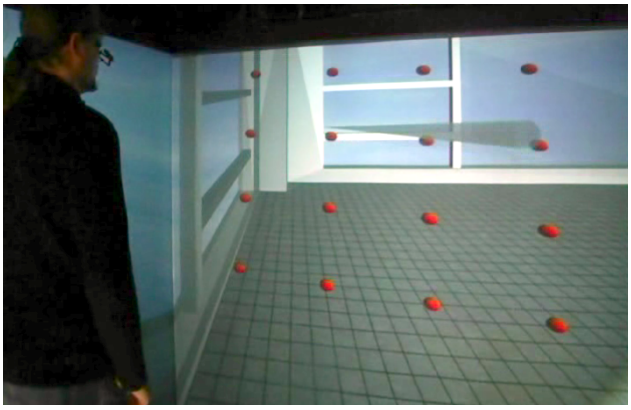


Figure 4: *For the VR calibration of the eye tracker, a grid of 16 spheres is presented one by one in a viewer local perspective. This way the grid is always aligned to the current position of the head.*

We have been able to adopt this schema to the VR and, moreover, do the calibration automatically without the need for an operator. The key idea is that a calibration grid, similar to the one on the display of the operator pc, is attached to the *HeadActuator* and relocated to an appropriate distance (see Figure 4). This way the eye and the calibration grid constitute a stable system which is independent of any head movements, as long as the framerate is adequate. The calibration module then interacts with the ViewPoint software by simulating the operator and presenting a red sphere as target at the location of each grid point.

4.4 Evaluation of Object Selections by Gaze

As a first interaction, we realized the selection of objects. For this we constructed a gaze ray in the *Ray Construction* node based on the position and direction provided by the *EyeNode*. A *Histogram* node collects the angular distances of all objects, if any, within an angle of 2.5° around the ray during an interval of 400ms. The object with the highest ranking for at least 200ms within this interval is taken as the fixated object. Longer and/or frequent fixations

of an object could then result in a selection of the object, based on an application specific threshold. In the study we are presenting, we are interested in object fixations only.

5 Measuring Accuracy and Latency with the Visual Ping

We evaluated the accuracy and latency of the system in a human-in-the-loop scenario we call the *visual ping*. The task of the participants is to fixate a single hihglighted sphere from a test-grid of 64 spheres. These spheres are placed on one of four test-grids in a plane perpendicular to the head orientation at the distances near (0.7m), normal (1.7m), far (2.7m), and very far (6.7m). All test-grids are placed in such a way that they are within angular eye movements of horizontally -35.29° to 35.36° and vertically -36.33° to 36.33°. The test-grids exceed the calibration grid in each direction about half a distance between the rows/columns. Thus the calibrated points lay in between the points of the test-grids.

The loop starts with the participant fixating the one single visible sphere. When the system detects this fixation, it hides the current sphere and highlights a new sphere from the grid. This moment defines the start time of the visual ping. The participant detects the vanishing of the fixated sphere and answers with a search movement of the eyes. The time of the first eye movement that is detected more than 2.5° away from the originally fixated position is taken as the response. The difference between this and the start time of the visual ping defines the latency of the visual ping. The deviation between the position of the detected fixation and the sphere defines the accuracy. Once the participant has found and fixated the newly highlighted sphere, the procedure continues.

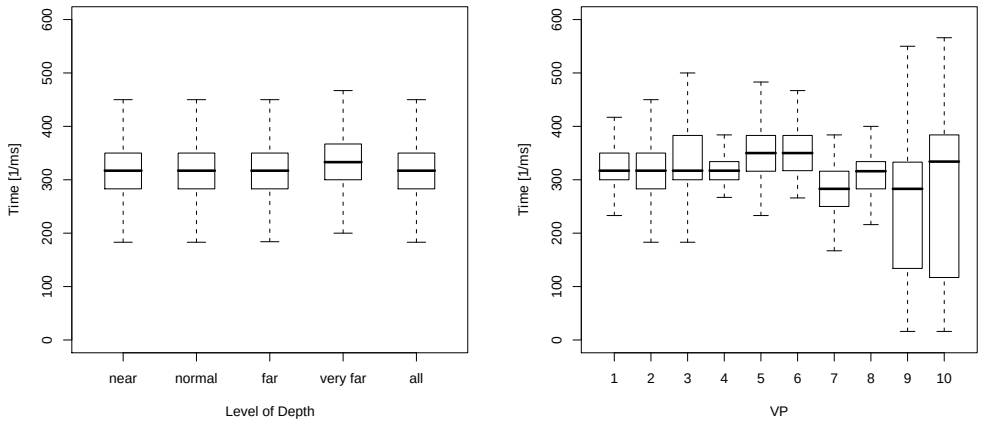
This loop is iterated over all spheres within each test-grid. In between individual grids, a calibration run with the 4x4 calibration grid is executed at normal distance.

6 Results

A total of 10 untrained people with no immersive VR experience participated in the study, 6 women and 4 men. The mean age was 27.7 years with a standard deviation (sd) of 6.17 years. The recorded data has been cleaned by removing outliers beyond 2 standard deviations. These are mostly outliers with large vertical deviations that we attribute to eye blinks. Using this procedure, alltogether 10.43% of the entries have been removed.

6.1 Latency

The results for the latency in the visual ping test are depicted in Figure 5(a) for each distance. The mean latency over all distances is 307.9ms, the median is 317ms and the sd is 99.9ms. The results for the participants are depicted in Figure 5(b).

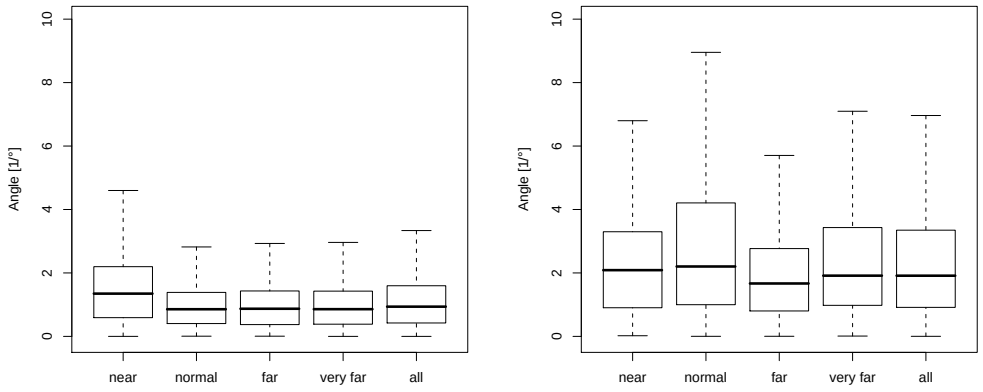


(a) (b)

Figure 5: *The latencies for the visual ping test (a) for each distance and (b) for each participant.*

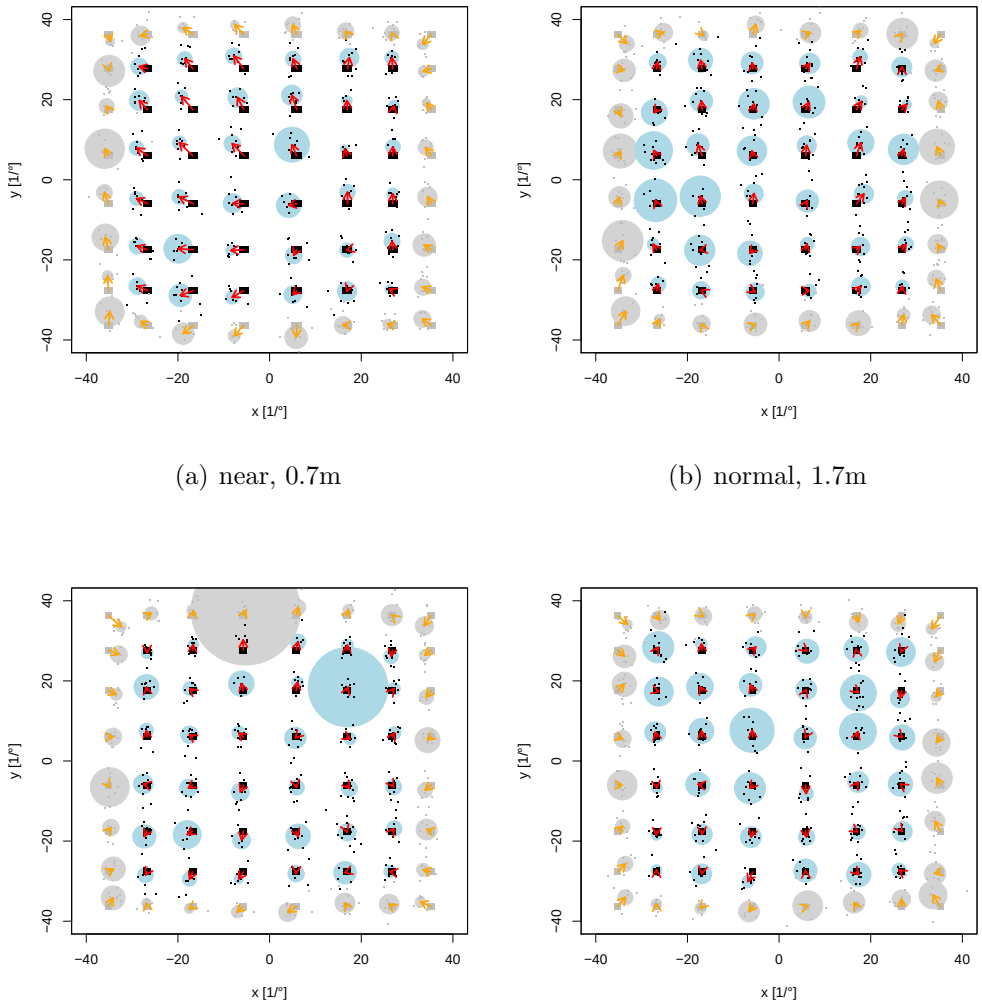
6.2 Accuracy

The accuracy of the detected fixations is depicted in Figure 6(a) for the horizontal and in Figure 6(b) for the vertical deviation. The horizontal accuracy over all distances is 1.18° (mean) or rather 0.94° (median). The precision (in sd) is 1.51°. The vertical accuracy over all distances is 2.52° (mean) or rather 1.91° (median). The precision (in sd) is 2.24°. A detailed overview is given in Figure 7 with median and sd for each grid point.



(a) (b)

Figure 6: *The accuracy of the detected fixations along the (a) horizontal and (b) vertical axis for each level of depth.*



(a) near, 0.7m

(b) normal, 1.7m

(c) far, 2.7m

(d) veryfar, 6.7m

Figure 7: *The plots show the angular accuracy of the detected fixations for each distance. The arrows show the deviation of the median and the circles highlight one standard deviation around the median. The outer grid-points exceed the calibrated area and are depicted in a lighter color.*

7 Discussion and Conclusion

We have presented technical details of an eye tracking system for our immersive Virtual Environment. Using a lightweight eye tracker, the user can freely interact in the CAVE-like system. The projection technology based on polarized light fits together well with the direct camera recordings of the eye tracking system. Systems with indirect eye recording over a semi-transparent mirror might be more difficult to handle. We observed no interference between the infrared optical tracking system and the infrared eye tracking. However, an additional infrared LED was used to cast enough light on the eye for the tracking.

The results from the user study exceeded our expectations. An accuracy of about 1° on the horizontal axis is nearly perfect, considering that the opening angle of the foveal high-accuracy vision is about 2°. The vertical accuracy is also good, although it is less than the horizontal. In most applications this difference should not matter, as horizontal differences are more important, e.g. in stereo vision. The performance also seems to be quite stable over all tested distances, even though we calibrated only for one distance. This stability is not initially surprising, because we arranged the test-grids to exactly overlap. However, while the normal distance was more or less presented exactly on the projection surface, and was thus not affected by ghosting or other influences due to stereo projection, the visual systems of the participants had to cope with disparity for the near, far and very far distances. Nevertheless the fixations were detected very accurately. We thus conclude that one can safely rely on a single distance for calibration.

The results for latency of the visual ping show that the performance is quite similar in every tested distance. Besides participants 9 and 10, the individual performance is also quite comparable for practical reasons. But how should a mean latency of 307.9ms be rated? For a meaningful interpretation, we would have to separate the performance of the human from the overall performance. The model for human performance of Card et al. [CMN83] might provide a rough approximation for this separation. According to this model, the perception of the missing sphere (100ms), the deliberation about the task (70ms), and the issuing of the motoric response (70ms) should sum up to 240ms. The contribution of the system to the latency would then be about 70ms, including frame grabbing with 60Hz, image processing, networking and visualization with a frame-rate of 60Hz.

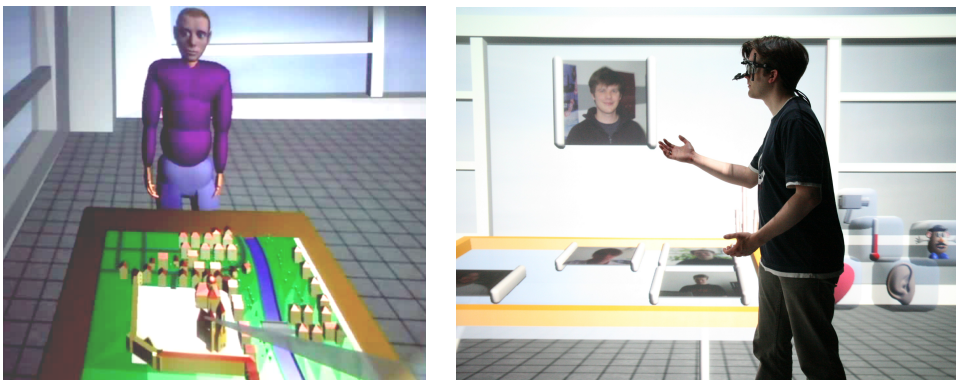
7.1 Future Work

In previous work, we successfully used a parameterized self-organizing map to reconstruct the depth of a fixation from the vergence angle of both eyes in real and desktop-based settings [PDLW07]. It is our plan to carry on this work in the immersive setting. The information about the depth of a fixation would allow for the reconstruction of the *volume of focus*, which could in turn be used to localize rendering, e.g. in terms of multi-framerate rendering, visual quality (depth-of-field), or localized perspective projection.

On the application level, we already use the eye tracking system to inform joint attention processes in the interaction with the embodied conversational agent Max ([PLW08], see Figure 8).

Acknowledgements

This work has been funded by the German Research Foundation within the Collaborative Research Center 673 *Alignment in Communication*. The author wants to thank Nikita



(a) (b)

Figure 8: *The eye tracking system is used to inform the interaction with the embodied conversational agent Max (a) and in a teleconferencing scenario (b). Relevant lines of research include turn-taking and joint attention.*

Mattar and Dennis Wiebusch for their support in implementing the software and conducting the study.

References

[artGA08] advanced realtime tracking GmbH A.R.T. WWW: http://www.ar-tracking.de, last seen 2008.

[BGA+04] J. Barabas, R. B. Goldstein, H. Apfelbaum, R. L. Woods, R. G. Giorgi, and E. Peli. Tracking the line of primary gaze in a walking simulator: modeling and calibration. *Behavior Research Methods, Instruments, & Computers*, 36(4):757–770, 2004.

[Bol81] R.A. Bolt. Gaze-orchestrated dynamic windows. *Proceedings of the 8th annual conference on Computer graphics and interactive techniques*, pages 109–119, 1981.

[CMN83] S.K. Card, T.P. Moran, and A. Newell. *The Psychology of Human-Computer Interaction*. Lawrence Erlbaum Associates, 1983.

[DCC+04] A. T. Duchowski, N. Cournia, B. Cumming, D. McCallum, A. Gramopadhye, J. Greenstein, S. Sadasivan, and R. A. Tyrrell. Visual Deictic Reference in a Collaborative Virtual Environment. In *Eye Tracking Research & Applications Symposium 2004*, pages 22–24, San Antonio, TX, March 2004. ACM Press.

[DMC+02] A. T. Duchowski, E. Medlin, N. Cournia, A. Gramopadhye, B. Melloy, and S. Nair. 3D Eye Movement Analysis for VR visual inspection training. In

Proceedings of the Symposium on Eye tracking Research & Applications, pages 103 – 110, New Orleans, Louisiana, 2002. ACM Press.

[EPR06] K. Essig, M. Pomplun, and H. Ritter. A neural network for 3D gaze recording with binocular eye trackers. *The International Journal of Parallel, Emergent and Distributed Systems*, 21(2):79–95, 2006.

[HHN+02] G. Humphreys, M. Houston, R. Ng, R. Frank, S. Ahern, P. D. Kirchner, and J. T. Klosowski. Chromium: a stream-processing framework for interactive rendering on clusters. *ACM Trans. Graph.*, 21(3):693–702, 2002.

[HLCC08] S. Hillaire, A. Lecuyer, R. Cozot, and G. Casiez. Using an Eye-Tracking System to Improve Camera Motions and Depth-of-Field Blur Effects in Virtual Environments. *Virtual Reality Conference, 2008. VR'08. IEEE*, pages 47–50, 2008.

[Inc08] Arrington Research Inc. WWW: http://www.arringtonresearch.com/, last visit 2008.

[LHNW00] D. Luebke, B. Hallen, D. Newfield, and B. Watson. Perceptually Driven Simplification Using Gaze-Directed Rendering. Technical report, University of Virginia Technical Report, University of Virginia, 2000.

[PDLW07] T. Pfeiffer, M. Donner, M. E. Latoschik, and I. Wachsmuth. Blickfixationstiefe in stereoskopischen VR-Umgebungen: Eine vergleichende Studie. In *Vierter Workshop Virtuelle und Erweiterte Realität der GI-Fachgruppe VR/AR*, pages 113–124, Aachen, 2007. Shaker. ISBN: 978-3-8322-6367-6.

[PLW08] N. Pfeiffer-Lessmann and I. Wachsmuth. Toward alignment with a virtual human - achieving joint attention. In A.R Dengel, K. Berns, and T.M. Breuel, editors, *KI 2008: Advances in Artificial Intelligence. Berlin: Springer*, 2008.

[TJ00] V. Tanriverdi and R. J. K. Jacob. Interacting with eye movements in virtual environments. In *Conference on Human Factors in Computing Systems, CHI 2000*, pages 265–272, New York, 2000. ACM Press.

[TKFW+79] J. H. Ten Kate, E. E. E. Frietman, W. Willems, B. M. Ter Haar Romeny, and E. Tenkink. Eye-Switch Controlled Communication Aids. *Proceedings of the 12th International Conference on Medical & Biological Engineering*, pages 19–20, August 1979.

[Tra01] H. Tramberend. Avango: A Distributed Virtual Reality Framework. In *Proceedings of Afrigraph '01*. ACM, 2001.