An Evaluation of Binocular Eye Trackers and Algorithms for 3D Gaze Interaction in Virtual Reality Environments

Thies Pfeiffer*, Marc E. Latoschik†, Ipke Wachsmuth‡

AI & VR Lab, Faculty of Technology
Bielefeld University
Universitaetsstrasse 25, Bielefeld, Germany
email: * tpfeiffe, † marcl, ‡ ipke@techfak.uni-bielefeld.de
www: http://www.techfak.uni-bielefeld.de/ags/wbski/

Abstract

Tracking users' visual attention is a fundamental aspect in novel human-computer interaction paradigms found in Virtual Reality and ambient computing. For example, multimodal interfaces or dialogue based communications with virtual and real agents greatly benefit from the analysis of the users' visual attention as a vital source for deictic references or turn-taking signals. Current approaches to determine visual attention rely primarily on monocular eye trackers. Hence they are restricted to the interpretation of two-dimensional fixations relative to a defined area of projection.

The study presented in this article compares precision, accuracy and application performance of two binocular eye tracking devices. Two algorithms are compared which derive depth information as required for visual attention based 3D interfaces. This information is further applied to an improved VR selection task in which a binocular eye tracker and an adaptive neural network algorithm is used during the disambiguation of partly occluded objects.

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

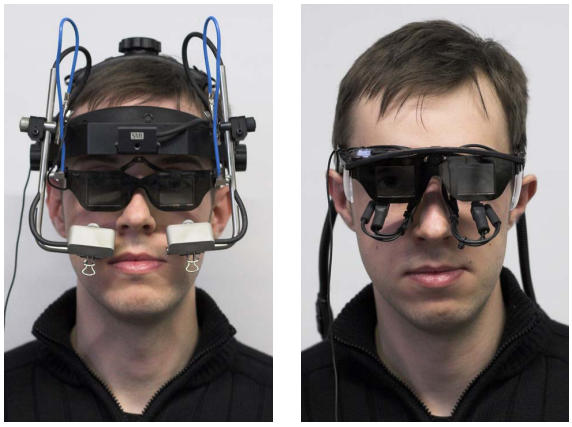
1 Introduction

Knowledge about the visual attention of users is a highly attractive benefit for information interfaces. The human eye is a powerful device for both perceiving and conveying information. It is faster than speech or gestures and it is closely coupled to cognition. Eye trackers offer a technical solution for acquiring the direction of gaze, from which the focus of attention can be derived. This has made eye tracking a powerful tool for basic research. For instance, in psycholinguistics the *visual world* paradigm [TSKES95] has gained much attention. This paradigm is used to investigate the interaction between visual context and speech processing by tracking the users' gaze on a scene while producing or interpreting spoken language. Eye tracking has become part of the standard toolkit of usability engineers in offline interface evaluation.

The most prominent examples of online applications are gaze typing systems, which provide alternative means for text input for the physically challenged, e.g., the Eye-Switch system [TKFW$^+$79]. 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. Since then eye trackers have evolved to a feasible input device.

Today, portable head-mounted eye trackers are available (see Figure 1) and the user is no longer required to remain stable, i.e., seated on a chair or, in some cases, use a chin rest. Eye tracking therefore has also become an attractive input methodology for Augmented and Virtual Reality (VR).

Relevant features of the eye movements are fixations, i.e., short moments when the eyes are resting on a specific area, and the movements in between such resting points, the saccades. While the eye tracking hardware is capable of capturing features necessary for reconstructing the fixated areas from the orientation of the eyes, as we will show later, the common approach provided by current state-of-the-art eye movement analysis software is to project the output onto a



(a) (b)

Figure 1: The head-mounted eye trackers used in the study: (a) SMI EyeLink I and (b) Arrington Research PC60. The shutter-glasses have been attached to the head-mount and the cameras are recording the eyes from below.

two dimensional plane, either a computer screen or a video image acquired by a so-called scene camera.

Thus it is not surprising that to the authors' knowledge, information about the depth of fixations are only rarely used in todays research, even though there are many lines of research that could greatly benefit from this knowledge, e.g., for the interpretation of spatial propositions (in front of vs. behind [GHW93]). And even more, it has to be questioned whether findings obtained using 2D or 2 1/2D stimuli can be automatically generalized to 3D environments (see [FPR06] for an example). A reliable algorithm for determining the depth of a fixation could therefore open new grounds for basic research.

Robust mobile 3D gaze tracking systems could also increase the capabilities of physically challenged users. Internal representations of users' surroundings could be augmented with semantic scene descriptions. By grounding the 3D gaze trajectories in a combined geometric and semantic representation of their surroundings, interaction models could utilize the additional context information to provide improved context-aware user centered interactions. Attentive computer vision systems could follow the guidance of the human gaze and selectively extract relevant information from the environment.

Knowledge about the area fixated by users in 3D space could also improve human-computer interaction in several ways. First of all, gaze plays an important role in computer mediated communication, e.g., when establishing eye contact to ensure mutual understanding or as turn-taking signals to control interaction in dialogues. Tracking gaze is therefore highly interesting for novel interfaces for teleconferencing such as Interactive Social Displays [PL07] developed in the PASION project [BMWD06]. The knowledge about the elements fixated by the users within the virtual world is also highly relevant for embodied virtual agents, such as MAX [KJLW03].

In direct interaction the 3D fixations could, e.g., be used for precise selection of entities in dense data visualizations. Using depth information, it is possible to detect fixations on objects behind transparent or sparse geometries, e.g. generated by shaders (grass, bushes). There are even applications on a technical level, e.g., in rendering technology, where the focused area is rendered in greater detail than the rest of the scene, and thus with equivalent rendering performance an increase in visual appearance is possible.

There already exist a number of successful approaches employing eye tracking technology in VR. Some of them we will describe in more detail in the following section. However, all of the approaches known to the authors rely on a single eye and therefore can only utilize the direction of the gaze and not reliably estimate the depth of the fixated area. Thus, the approaches have a lower resolution than technically possible and are subject to ambiguities. This will be elaborated in more detail in section 2.1.

In this article we are going to tackle the following questions:

1. *How can the depth of a fixation be determined?*
2. *What algorithms are known and which of them are suitable for applications, especially in VR?*
3. *How well do different eye trackers cope with the demands of 3D fixation determination?*
4. *What are the quantified benefits for applications, exemplarily tested on a visual selection task?*

2 State of the Art

Gaze plays an important role in the design of embodied conversational agents (ECAs) [TCP97] and eye tracking technology provides a viable source of data on human gazing behavior. Vertegaal et al. [VSvdVN01] derive implications for gaze behavior of ECAs in communicative situations from eye tracking studies on human conversations. Others, such as

Lee et al. [LBB02], create computational models for gaze pattern production in virtual agents based on data on natural eye movements. Examples of online interpretation of eye tracking data are the already mentioned gaze typing systems.

Knowledge about the user's visual attention can be used to facilitate human-to-human interaction in VR environments. Duchowski et al. [DCC+04] 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. More technical approaches employ information about the focused area to optimize rendering processes [LHNW00].

Human-machine interaction within VR systems can also greatly benefit from information gained by eye tracking. Tanriverdi and Jacob [TJ00] demonstrate a significantly faster object selection when it is based on gaze as compared to gestures. 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. They use the gaze position projected onto a 2D plane as basis for the picking ray.

Using a ray along the visual axis as the basis for a gaze-based interaction model is an approach also 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.

Interpreting pointing as a ray or vector is quite common for pointing gestures [Kit03] and we have conducted a study on the performance of human pointing [KLP+06] to evaluate models for the interpretation of pointing gestures. The studies show that taking gaze into account improves the accuracy of the interpretation of pointing gestures. In these models the direction of gaze is only approximated by the direction of the face and thus we expect even better performance when considering the actual direction of the gaze measured by an eye tracker. The algorithms described in this paper will be integrated in our Interactive Augmented Data Explorer [PKL06] which serves as a platform for subsequent user studies.

2.1 Problems with Ray Based Approaches

In ray based approaches, determining the depth of a fixation is coupled with several problems (see Figure 2): it is (a) only possible if the ray of sight directly intersects a geometry; there is (b) an ambiguity whenever several geometries are intersecting the ray of sight and these approaches (c) do not respect the dominance of a specific eye when determining the fixation.

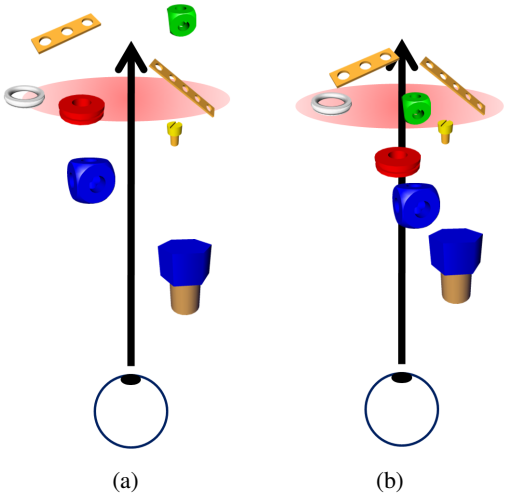


Figure 2: Selecting objects via ray based approaches suffers from ambiguities. If the ray does not hit any object, (a), the selection is underspecified. If several objects are hit, (b), the selection is overspecified. Both cases demand appropriate selection heuristics.

Problems (a) and (b) are also relevant and known for pointing/picking and there exist several approaches to improve performance. Natural interaction technologies are often employing heuristics, for instance, they take the distances of objects to the picking ray into account and thus they do not require a direct intersection with the object's geometries [OBF03]. More technical approaches either use tools [FHZ96] or visualizations as aiming aids.

Modelling human visual perception as a ray can only be a simplification. In reality, the eyes cannot see equally well all along the visual axis. In the following section we provide a brief review of depth perception, focusing on features appropriate for sensory acquisition.

2.2 3D Visual Perception

Although the retina of the human eye only samples a 2D projection of the surroundings, humans are capable of reconstructing a three-dimensional impression of their environment. In the literature (e.g. [Gol02]) several criteria for depth perception can be found:

monocular depth criteria such as *occlusion*, *relative size/height in the field of view*, *common size of ob-*

jects, *atmospherical and linear perspective*, the *gradient of texture*, or *motion parallax* convey spatial information with a single eye only.

binocular depth criteria are *disparity* (differences in the retinal picture caused by the disparity of the eyes), *vergence* (see Figure 3), or *accommodation*.

Binocular depth perception, *stereopsis*, provides means to differentiate between the depth of objects up to a distance of about 135 meters. If the depth of a fixation should be determined, only such criteria can be used which require measurable effort from the perceptual system. As most criteria do not have a sensory-motor component, from the listed criteria only vergence and accommodation remain for consideration. Both vary depending on the distance of the fixated object.

The human eyes are optimized to see very accurate only within a small area of the retina, the *fovea centralis*. The area covered by the fovea centralis is less than 1°. This implies that if an object is to be inspected, the eyes have to be oriented in such a way that the projection of the object onto the retina falls (partly) onto the fovea centralis. If this happens, the images of both eyes can be fused. The visual line is the projection of the object through the center of the eye onto the retina. Two categories of eye movements are distinguished: when the eyes follow an object horizontally or vertically, moving in the same direction, they are called *version movements* and when the eyes move locally in opposite directions, they are called *vergence movements*. The vergence movements are those associated with objects altering their depth and these vergence can be measured by binocular eye trackers. The horizontal component of the movement is the relevant movement for the stereoscopic depth perception [Whe38]. Measuring vergence angles one may differentiate fixation depths up to a distance of 1.5 m to 3 m depending on the user's visual faculty.

Accommodation can be measured with research prototypes of vision based eye trackers [SMIB07], but not with off the shelf technology. A healthy eye of a young adult has an operational range between focal lengths from 1.68 cm to 1.80 cm. Thus differences in accommodation can be measured for distances between approximately 0.25 m and 100 m.

The working range of vergence movements nicely covers typical interaction spaces within immersive setups. Whether a state-of-the-art binocular eye tracker does provide sufficient means to measure vergence angles at resolutions reasonable for human-machine interaction will be one of the questions tackled by the user study presented in section 3. Tracking accommodation would significantly increase the operational range of 3D gaze determination. However, to our knowledge the readily available head-mounted devices do not currently offer this functionality.

2.3 Estimating Fixation Depth

In our study presented in section 3 we want to test two different approaches to estimate the depth of a fixation. One is a straight forward approach using linear algebra: the depth is determined by the intersection of the optical axes of the two eyes converging on the target. The second approach has been proposed by Essig and colleagues [EPR06]: a parameterized self-organizing map adapts to the viewing behavior of the user and the visual context, learning the mapping from the 2D coordinates of the fixations on a display to the fixated point in depth. This approach has previously only been tested with an anaglyphic stereo projection and dot-like targets. In the study presented in this article shutter-glasses are used in a desktop VR scenario. For greater realism, small geometric models of real objects are used as targets.

2.3.1 Crosscutting the Optical Axes

In theory, the point being fixated with both eyes can be determined by intersecting two rays (see Figure 3). For the following equations we assume a coordinate system with an origin between the eyes of the observer. Given the positions of the two eyes $\vec{a}_{left}$ and $\vec{a}_{right}$, as well as the fixations of both eyes $\vec{s}_{left}$ and $\vec{s}_{right}$ on the plane of projection, we can derive the following parameterized line equations $\vec{g}_{left}$ and $\vec{g}_{right}$ for the visual axes as follows:

$$\vec{g}_{left} = \vec{a}_{left} + \mu \cdot (\vec{s}_{left} - \vec{a}_{left})$$
$$\vec{g}_{right} = \vec{a}_{right} + \eta \cdot (\vec{s}_{right} - \vec{a}_{right})$$

The points $\vec{f}_{left}$ and $\vec{f}_{right}$ on both visual axes in Figure 3 b) are the points with the lowest distance to the other axis. The point of fixation $\vec{f}$ then is the mean of $\vec{f}_{left}$ and $\vec{f}_{right}$.

This approach, though, has some disadvantages. First, the physical parameters such as the height, the

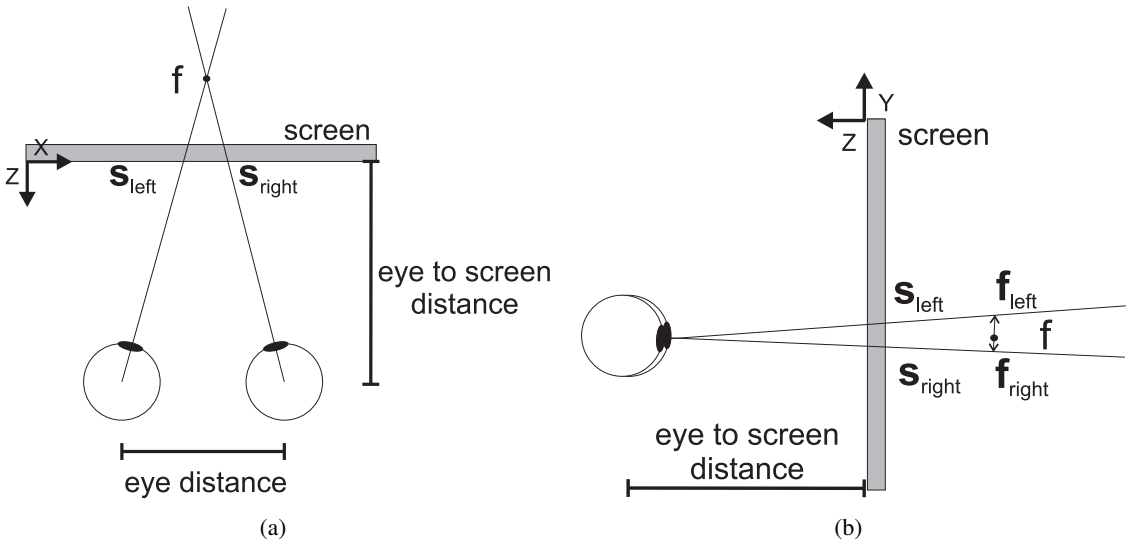


Figure 3: Calculating the depth of a fixation using linear algebra. Although the visual axis of the eyes may intersect in the focal point f when projected to a plane (a, top view), in three dimensions they may still not intersect (b, side view).

disparity and the geometry of the eyes vary between users and would have to be measured for each person. Also, one of the eyes typically dominates the other, that is, this eye's fixation are likely to be more precise and accurate than those of the other. More generally, users may have different behavioral patterns in their vergence eye movements. Together with device specific systematic errors and noise in the angles measured by the eye trackers this will lead to differences between the real and the approximated visual line. These parameters are not taken into account by this algorithm. An accurate calibration procedure could help to estimate some of the parameters. But to get reasonable data, calibration may have to be repeated several times, which would make it a tedious procedure. As the maintenance of an accurate tracking requires a recalibration every time the eye tracker slips, this would soon be tiring.

Essig et al. [EPR06] proposed an adaptive algorithm to estimate the depth of a fixation, which may be more suitable under these conditions. Their approach is summarized below.

2.3.2 Holistic Approximation Via a Parametrized Self-Organizing Map

The idea is to replace the fixed mapping provided by the linear algebra approach with a flexible mapping provided by a machine learning approach. This mapping should translate the 2D coordinates provided by the eye tracker for both eyes to a 3D coordinate describing the singular binocular fixation in depth. This mapping will have to be learned and thus will require user interaction. The 2D calibration procedure required for the 2D eye tracking software will therefore be followed by a 3D calibration procedure using a 3D grid of points. A usability-requirement is that the learning procedure is as smooth and fast as possible, as relearning will be necessary every time the eye tracking device slips.

Essig et al. [EPR06] proposed to use a Parameterized Self-Organizing Map (PSOM), a smooth high-dimensional feature-map [Rit93] for approximating the 3D fixation. The PSOM is derived from the SOM [Koh90] but needs less training to learn a non-linear mapping. It consists of neurons $a \in A$ with a reference vector $\vec{w}_a$ defining a projection into the input space $X \subseteq \mathbb{R}^d$. The reference vector is defined as $\vec{w}_a = (x_l, y_l, x_r, y_r, x_{div})$ with (x_l,y_l) and (x_r,y_r) being the fixations on the projection plane measured by the eye tracker. As the horizontal distance of the fixations has a significant contribution to the determination of the depth, it is added as an additional parameter $x_{div} = x_r - x_l$ to $\vec{w}_a$.

To train the PSOM, all 27 points of a three-dimensional $3 \times 3 \times 3$ calibration grid are presented subsequently and the corresponding $\vec{w}_a$ are measured. From this one can derive a function $\vec{w}(s)$ mapping the coordinates of the 3D grid onto the reference vectors.

Thus $\vec{w}(s)$ is constructed in such a way that the co-

Figure 4: The set-up of the experiment uses shutter-glasses and a cathode-ray display. The head of the user was stabilized using a chin rest.

ordinates of the 3D grid can be mapped to the 2D positions of the fixations. To find the fixation one has then to find the solution of the inverse function numerically using gradient descent, which is done in the network's recurrent connections.

In the user study we use exactly the PSOM as specified in the paper by Essig and colleagues [EPR06].

3 Method

We conducted a user study to test accuracy, precision and application performance of the two algorithms in combination with two eye trackers available to our group. Our goal was to find a combination of software and hardware suitable for 3D gaze-based interaction in Virtual Environments.

3.1 Hypotheses

Based on the questions presented in the introduction, the following hypotheses were guiding the study:

A: PSOM is more precise and accurate than the geometric approach

Of the two algorithms presented, the PSOM should have noticeable advantages. This approach was therefore expected to provide higher precision and accuracy compared to the geometric approach, according to the reasons pointed out in section 2.1.

B: The high-end device is more precise and accurate than the low-cost device in binocular use

In this study two different head-mounted devices were tested (see Figure 1): the EyeLink I from SMI as a representative of the high-end devices ($>$ €30,000) and the system PC60 from Arrington Research as a representative for the low-cost sector ($<$ €12,000). The technical details presented in Table 1 show that the device from SMI has noticeable advantages regarding speed and accuracy.

Object Number	x	y	z
1, 2	23	8	23
3, 4, 17, 19, 22	20	24	17
5	30	30	30
6	30	10	30
7, 18, 20, 21	20	24	20
8	20	34	20
9	20	60	17
10	20	20	34
11	20	17	24
12, 15	20	20	24
13, 14, 16	20	17	24

Table 2: **Object Dimensions** (in mm) of the target set of objects used for the fixation and selection task. The numbers refer to the objects as specified in Figure 5.

C: Knowing the depth of a fixation will increase success rate when selecting objects

Exploiting knowledge about the depth of a fixation should improve the disambiguation of difficult cases where objects are partially occluded, but have significant differences in depth (see Figure 2). Therefore this approach should have a higher success rate for these object selections than traditional 2D based approaches.

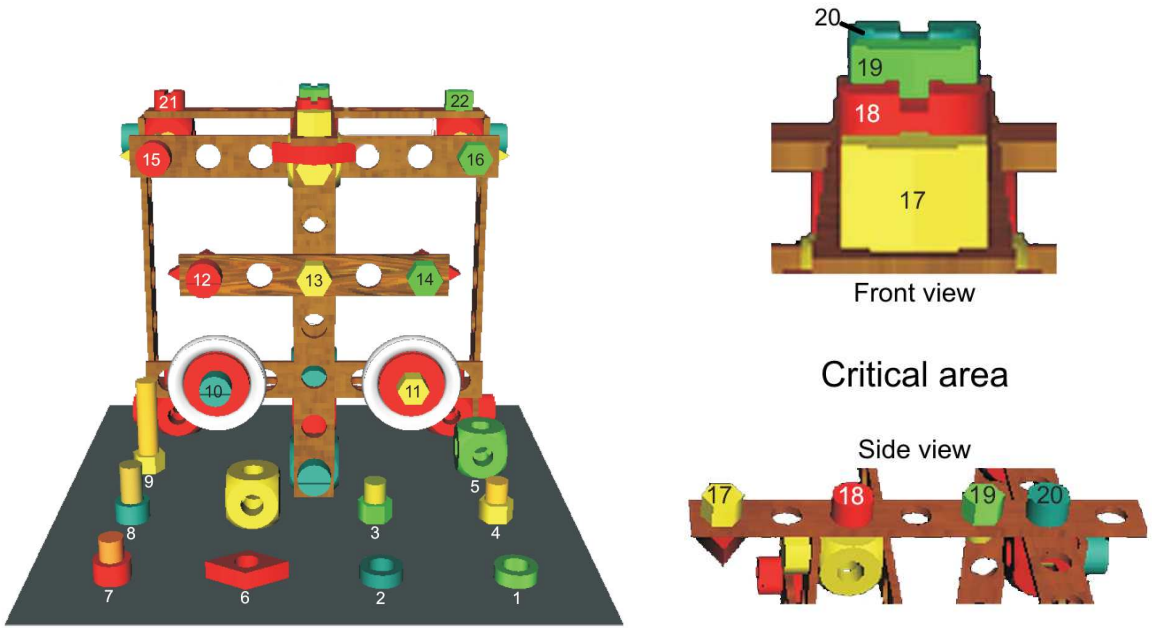
3.2 Scenario

In the study the participants looked at a 3D scene showing a structure build out of toy building blocks (see Figure 4 and Figure 5). The dimensions of the relevant target objects used in the study are provided in Table 2. A 21" Samsung SyncMaster 1100 cathode-ray monitor was used together with a NVidia Quadro4 980 XGL and Elsa Retaliator consumer class shutter-glasses for the stereoscopic projection. Both eye tracking systems are prepared to be used in monitor based settings. The implementation of the experiment is based on the 3D extension of the VDesigner software described in [FPR06].

The study had four conditions, resulting from an intra-personal covariation of two tested eye trackers and two algorithms. To stabilize external factors for

	Arrington PC60	SMI EyeLink I
temporal resolution (Hz)	30 / 60	250
optical resolution (pixel)	640×480 / 320×240	-
deviation from real eye pos	$0.25\,^\circ$ - $1.0\,^\circ$ visual angle	$< 1.0\,^\circ$ visual angle
accuracy	$0.15\,^\circ$ visual angle	$0.01\,^\circ$ visual angle
compensation of head shifts	not possible	$\pm 30\,^\circ$ horizontal, $\pm 20\,^\circ$ vertical

Table 1: **Technical details of the eye tracking systems**



(a) In the online version you may click on this image to explore a 3D view of the setting.

(b) This set of occluding objects defines the critical area for the 3D selction algorithm.

Figure 5: Position of the objects in the model (left). The objects 17 to 20 define the critical area where a selection based on 2D methods leads to ambiguities (right).

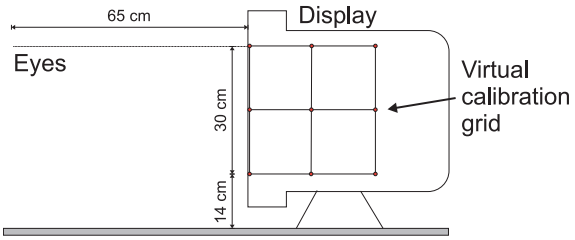


Figure 6: Sketch of the set-up: the virtual space fits exactly inside a cube of 30 cm located behind the plane of projection.

the comparison between the different algorithms, the distance from the head to the projection plane was fixated to 65 cm using a chin rest. The height of the chin rest was adjusted so that the eyes of the user were positioned on level with the upper edge of the virtual calibration grid (see Figure 6).

The two eye trackers, the SMI EyeLink I and the Arrington PC60 are both head-mounted. In addition to the eye tracker the participants also had to wear the shutter-glasses. The combination of a projection technology requiring special glasses and vision-based eye tracking systems is delicate, as the cameras of the eye tracking systems cannot see clearly through the glasses. In our case we adjusted them to a position below the glasses with a free, but very steep, perspective onto the eye. For the SMI EyeLink I we had to construct a special mounting for the glasses, as the original one interfered with the bulky head-mounted eye tracking system. This also allowed us to increase the gap between the eyes and the glasses so that orienting

the cameras of the eye tracking systems was easier.

After the standard 2D calibration procedure provided by the accompanying eye tracking software a 3D calibration procedure was run. For this the participants were presented the points of the calibration grid; for a side view see Figure 6. To fixate the leftmost calibration point on the front side of the cube the right eye of the user had to rotate $49.27°$ to the left, whereas the rightmost point was $32.19°$ to the right. To fixate all points on the back side of the cube, the right eye had to rotate $36.16°$ to the left and $22.15°$ to the right. To fixate a point in the upper center of the front side the eyes had to converge $8.99°$ and for a corresponding point on the back side $6.16°$.

A pilot study had shown that each person needed an individual timespan to acquire 3D perception with the projection technology used, so the calibration was self-paced. During the calibration procedure, all points of the grid were presented dimly lit and only the point to be fixated was highlighted. The points were traversed on a per plane basis, as has been recommended by Essig and colleagues [EPR06]. However, they only displayed the points of one plane at a time while we showed all points simultaneously, but dimly lit.

A life-sized VR model of a Baufix model was shown during the experiment (see Figure 4). The experimenter verbally referenced objects within the model which should then be fixated by the participants. As soon as they fixated the object, the participants affirmed this by pressing a key. The 3D fixation points were calculated internally for each fixation using both algorithms and the results were logged. This was performed with each participant using the 22 objects depicted in Figure 5.

4 Results

In this study we tested 10 participants (4 females and 6 males). The mean age was 26.2 years, the youngest participant was 21 years and the oldest 41 years old. Four participants were nearsighted and one farsighted. All participants had normal or corrected sight (contact lenses) during the experiment. They rated the difficulty of the experiment with 2.2 on a scale from 1 (very easy) to 6 (extreme hard).

Four participants reported difficulties in fixating the virtual calibration crosses: they had problems getting the crosses to overlap for getting the 3D impression.

4.1 Precision and Accuracy

The relative deviations of the calculated fixations from the real object positions (defined by the center of the object geometries) over all participants are shown in the bagplots for the axes y and z (depth) in Figure 7.

The Kolmogorov-Smirnov test shows that both datasets are not normally distributed. We therefore applied the Mann-Whitney-Wilcoxon test to examine whether the absolute means of both datasets are significantly different and if the means differ significantly from the nominal values. An alpha level of 0.05 was considered significant (see Table 3) in all tests.

In the test series for the two eye trackers the results for the z axis show that the means of the fixations approximated by the PSOM are significantly closer to the nominal value than those calculated by the geometric approach (7 from left to right). Still all means differ significantly from the nominal value. The means of the results for the device from Arrington Research are closer to the nominal value than those from the SMI eye tracker (7 from top to bottom). The device from Arrington Research has a higher accuracy in our study.

The SMI device achieves a higher precision, which is expressed in the lower standard deviations when compared to the device from Arrington Research. The precision using the PSOM algorithm is higher than the precision of the geometric algorithm for both devices.

4.2 Performance in the Object Selection Task

Besides the described quantitative accuracy study, qualitative implications for applications have been tested on an object selection task. We tested whether a selection algorithm based on the 3D fixation manages to successfully identify more objects than an approach based on 2D fixations only. Backed by the previous results we only considered the PSOM approach using the Arrington Research PC60 for the 3D fixations.

The 2D selection algorithm determines the Euclidean distance between the 2D coordinates on the projection plane provided by the eye tracking software and the projected screen coordinates of the 22 objects (center of object). The object with the smallest distance to at least one of the fixations of both eyes was taken as the selected object. The selection was then checked against the prompted object.

The 3D selection algorithm worked similarly using a standard 3D distance metric. Of the 22 objects, 4 were positioned in such a way that their projections partially occluded each other and thus led to an am-

device	algorithm	normally distributed	mean	difference btw. algorithms	nominal error	standard deviation
Arr.	*geom.*	no, $p < 0.001$	-195.77 mm	sig. $p < 0.001$	sig. $p < 0.001$	526.69 mm
	PSOM	yes, $p = 0.943$	-18.75 mm		sig. $p = 0.005$	96.92 mm
SMI	*geom.*	no, $p = 0.038$	-248.55 mm	sig. $p < 0.001$	sig. $p < 0.001$	149.3 mm
	PSOM	yes, $p = 0.661$	-70.57 mm		sig. $p < 0.001$	60.06 mm

Table 3: **Results** comparing the different conditions a significant difference of the means of the fixation depths show up in favor of the PSOM-algorithm.

biguous situation for the 2D selection test. This set of objects defined a critical area for the test.

The 2D selection algorithm successfully identified 165 (75%) of the 220 possible object selections (22 per participant). The 3D selection algorithm identified only 92 (42%). In the critical area comprising the objects 17 to 20 (40 selections) the 3D algorithm manages to disambiguate 17 (42%) object selections compared to 12 (30%) identified by the 2D algorithm. Figure 8 shows the successful selections per object. The numbering of the objects is depicted in Figure 5.

5 Discussion

5.1 Hypotheses

The following conclusions for the three hypotheses can be derived from the results of the study :

A accepted: PSOM is more precise than the geometric approach

The fixations approximated by the PSOM are significantly more precise and accurate than the results of the geometric approach for the y and z coordinates for both eye trackers.

This result replicates the findings of Essig and colleagues [EPR06]. Compared to their results we found greater deviations of the means and of the standard errors. This was expected, as in our setting we considered objects with a distance between 65 cm and 95 cm from the observer and in their setting the objects were located in an area between 39 cm and 61 cm in front of the observer. They had already shown that the error increases with distance from the observer.

We also did not use dots or crosses (diameter: 1° of visual angle) as targets, but models of small real objects (diameter: $1^\circ - 3^\circ$ of visual angle), such as bolts and nuts. Thus the error, which is defined as the deviation of the fixation from the center of the object, will have a higher standard deviation because the participant can fixate on a larger area than with dots.

B (partly) rejected: the low-cost device has been more accurate in the study

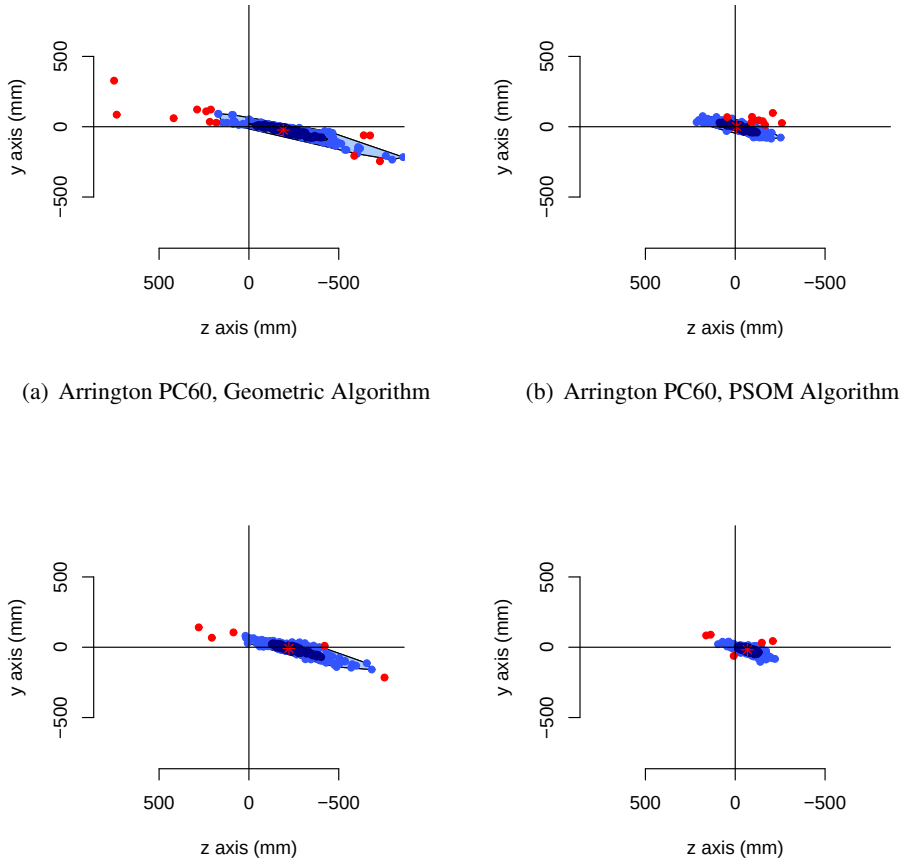
Although the EyeLink I has a higher precision, the PC60 proved to be more accurate in our setting. One possible explanation could be that the 2D calibration using the shutter-glasses is more difficult with the EyeLink I because the adjustment of the cameras of the EyeLink I system is more difficult. In the study the fixations could often only be rated *poor* by the provided software. Thus the base data was less precise. This predication only holds under limited conditions for the technical equipment used, especially for the projection technology and the shutter-glasses.

C (partly) rejected: Considering fixation depth does reduce success rate

Using the results of the algorithms to estimate the 3D fixation for object selection on the whole scene yields a lower success rate than with the original 2D selection (42% to 75%). Only in the criticial area (objects 17 to 20 in Figure 5), where the partial occlusion of objects leads to ambiguities with the 2D selection, the 3D selection method can demonstrate an improvement (42% to 30%). Comparing the coordinates estimated by the 3D approaches with the coordinates provided by the eye tracker shows that the estimated values are less precise. This can be explained by an imprecise calibration originating from problems with the projection technology, which has also been testified by some of the participants.

6 Conclusion

The results show that 3D fixations can be derived from vergence movements and the findings of Essig and col-



(a) Arrington PC60, Geometric Algorithm

(b) Arrington PC60, PSOM Algorithm

(c) SMI EyeLink I, Geometric Algorithm

(d) SMI EyeLink I, PSOM Algorithm

Figure 7: Bagplots showing the relative errors of the different conditions for the y axis and the z axis. The perspective is equal to 6, thus the user is looking towards negative z. The darker areas contain the best 50% fixations (those with the lowest deviations) and the brighter areas contain the best 75% fixations. The red dots mark outliers and the asterisks within the darker area marks the mean value.

leagues [EPR06] can be generalized to shutter-glass based projection technologies. The problem of ambiguity in the critical area can be resolved better than with 2D approaches. However, in the scenario presented, the adaptive approach cannot be recommended generally, because of the low precision in the XY-plane. Too many parameters are influencing the calculation. Great improvements have been achieved using the adaptive approach based on the PSOM, but the current implementation still does not respect all parameters. In addition, there are limitations depending on the applied VR technology: Insufficient channel separation (ghosting) of the applied stereoscopy method and the limited interaction space of the desktop based VR platform complicates the application of eye tracking methods and calls for further investigations, e.g., using full immersive displays and alternative stereoscopy methods.

The presented study is part of a series of studies. In a subsequent study we plan an analogous scenario with real objects to exclude influences on the vergence movements induced by the projection technology. The results of this study will be compared to the presented study and provide a baseline. If viable, the procedure followed in the study could also be used as a benchmark for projection technology, taking similar vergence movement behavior as an indicator for life-like projections.

A following study will then test the accuracy and precision of the Arrinton Research PC60 within a 3-sided immersive VR display (two adjacent walls, one floor) using a projection technology based on polar-

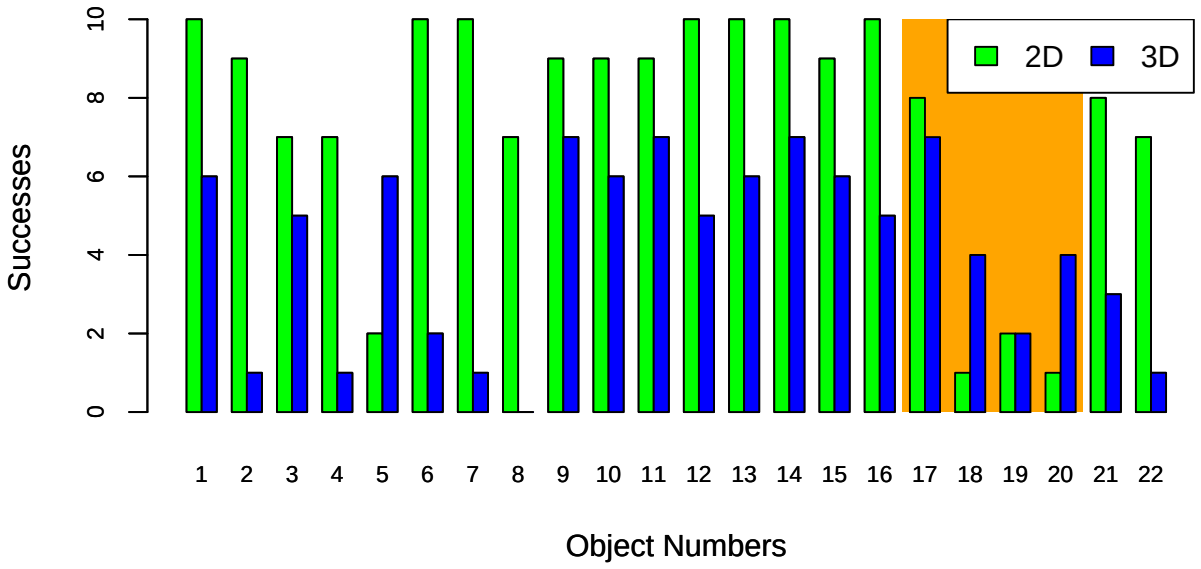


Figure 8: Histogram of the correct object selections over all 10 sessions (see Figure 5). The critical area of overlapping objects (numbers 17 to 20) is highlighted.

ized light. This setting also provides new technical challenges as the user may move freely in the VR setup of a size of $8\,m^3$.

The current implementation is based on the 2D projection of the fixations provided by most eye tracking software. In upcoming studies, we will analyze fixations' coordinates in eye space by including the users' head orientations as provided by head tracking techniques. Especially for the EyeLink I, elementary head tracking mechanism are provided to compensate for small head movements. But head tracking capabilities are commonly found in VR setups and hence could be further utilized for extensive eye space based approaches and intended application cases.

For an improved algorithm for 3D fixations in dense environments, or when disambiguating between background (walls, etc) and foreground, a hybrid approach seems viable. This algorithm should use the 2D fixation method per default and switch to the 3D fixation method when ambiguities arise. We expect that the upcoming studies will provide valuable data to further improve the PSOM approach. For the object selection task we expect that further progress can be made by adapting models created for the interpretation of pointing gestures.

7 Acknowledgments

The authors wish to thank Matthias Donner, who conducted the user study as part of his diploma thesis. This work has been partly funded by the German Research Foundation within the Collaborative Research Center 673 *Alignment in Communication* and by the EU within the project PASION (Psychologically Augmented Social Interaction Over Networks).

References

[BGA+04] James Barabas, Robert B. Goldstein, Henry Apfelbaum, Russell L. Woods, Robert G. Giorgi, and Eli Peli, *Tracking the line of primary gaze in a walking simulator: Modeling and calibration*, Behavior Research Methods, Instruments and Computers **36 (4)** (2004), 757–770, ISSN 0743-3808.

[BMWD06] Maria Christina Brugnoli, Federico Morabito, Richard Walker, and Fabrizio Davide, *The PASION Project: Psychologically Augmented Social Interaction*

Over Networks, PsychNology **4** (2006), no. 1, 103–116, ISSN 1720-7525.

[DCC+04] Andrew T. Duchowski, Nathan Cournia, Brian Cumming, Daniel McCallum, Anand Gramopadhye, Joel Greenstein, Sajay Sadasivan, and Richard A. Tyrrell, *Visual Deictic Reference in a Collaborative Virtual Environment*, Eye Tracking Research & Applications Symposium 2004 (San Antonio, TX), ACM Press, March 2004, pp. 35–40, ISBN 1-58113-825-3.

[DMC+02] Andrew T. Duchowski, Eric Medlin, Nathan Cournia, Hunter Murphy, Anand Gramopadhye, Santosh Nair, Jeenal Vorah, and Brian Melloy, *3D Eye Movement Analysis*, Behavior Research Methods, Instruments and Computers **34** (2002), no. 4, 573–591, ISSN 0743-3808.

[EPR06] Kai Essig, Marc Pomplun, and Helge Ritter, *A neural network for 3D gaze recording with binocular eye trackers*, The International Journal of Parallel, Emergent and Distributed Systems **21 Nr. 2** (2006), 79–95, ISSN 1744-5779.

[FHZ96] Andrew Forsberg, Kenneth Herndon, and Robert Zeleznik, *Aperture based selection for immersive virtual environment*, Proceedings of the 1996 ACM Symposium on User Interface Software and Tech. (UIST'96), 1996, pp. 95–96, ISBN 0-89791-798-7.

[FPR06] Helmut Flitter, Thies Pfeiffer, and Gert Rickheit, *Psycholinguistic experiments on spatial relations using stereoscopic presentation*, Situated Communication (Gert Rickheit and Ipke Wachsmuth, eds.), Mouton de Gruyter, Berlin, 2006, pp. 127–153, ISBN 3-11018-897-X.

[GHW93] Joachim Grabowski, Theo Herrmann, and Petra Weiss, *Wenn "'vor"' gleich "'hinter"' ist - zur multiplen Determination des Verstehens von Richtungspräpositionen*, Kognitionswissenschaft **Nr. 3** (1993), 171–183, ISSN 0938-7986.

[Gol02] E. Bruce Goldstein, *Wahrnehmungspsychologie*, Spektrum Akademischer Verlag, 2002, ISBN 3-82741-083-5.

[Kit03] Sotaro Kita, *Pointing: A Foundational Building Block of Human Communication*, Pointing: Where Language, Culture, and Cognition Meet (Sotaro Kita, ed.), Lawrence Erlbaum Associates, Inc., Mahwah, New Jersey, 2003, pp. 1–8, ISBN 0-80584-014-1.

[KJLW03] Stefan Kopp, Bernhard Jung, Nadine Lessmann, and Ipke Wachsmuth, *Max - A Multimodal Assistant in Virtual Reality Construction*, KI-Künstliche Intelligenz **4** (2003), 11–17, ISSN 0933-1875.

[KLP+06] Alfred Kranstedt, Andy Lücking, Thies Pfeiffer, Hannes Rieser, and Ipke Wachsmuth, *Deixis: How to Determine Demonstrated Objects Using a Pointing Cone*, Gesture Workshop 2005 (Berlin Heidelberg) (Sylvie Gibet, Nicolas Courty, and Jean-Franois Kamp, eds.), LNAI 3881, Springer-Verlag GmbH, 2006, pp. 300–311, ISBN 0302-9743.

[Koh90] Teuvo Kohonen, *The self-organizing map*, Proceedings of IEEE **78** (1990), no. 9, 1464–1480, ISSN 0018-9219.

[LBB02] Sooha Park Lee, Jeremy B. Badler, and Norman I. Badler, *Eyes alive*, SIGGRAPH '02: Proceedings of the 29th annual conference on Computer graphics and interactive techniques (New York, NY, USA), ACM Press, 2002, pp. 637–644, ISBN 1-58113-521-1.

[LHNW00] David Luebke, Benjamin Hallen, Dale Newfield, and Benjamin Watson, *Perceptually Driven Simplification Using Gaze-Directed Rendering*, Tech. report, University of Virginia Technical Report, University of Virginia, 2000.

[OBF03] Alex Olwal, Hrvoje Benko, and Steven Feiner, *SenseShapes: Using Statistical Geometry for Object Selection in a Multimodal Augmented Reality System*, Proceedings of The Second IEEE and ACM International Symposium on

Mixed and Augmented Reality (ISMAR 2003) (Tokyo, Japan), October 710 2003, pp. 300–301, ISBN 0-7695-2006-5.

[PKL06] Thies Pfeiffer, Alfred Kranstedt, and Andy Lücking, *Sprach-Gestik Experimente mit IADE, dem Interactive Augmented Data Explorer*, Dritter Workshop Virtuelle und Erweiterte Realität der GI-Fachgruppe VR/AR (Aachen) (Stefan Müller and Gabriel Zachmann, eds.), Shaker, 2006, accepted, ISBN: 3-8322-5474-9 ISBN-13: 978-3-8322-5474-2, pp. 61–72, ISBN 3-8322-5474-9.

[PL07] Thies Pfeiffer and Marc Erich Latoschik, *Interactive Social Displays*, IPT-EGVE 2007, Virtual Environments 2007, Short Papers and Posters (Bernd Fröhlich, Roland Blach, and Robert van Liere, eds.), Eurographics Association, 2007, pp. 41–42.

[Rit93] Helge Ritter, *Parametrized self-organizing maps*, ICANN93 Proceedings (1993), 568–577.

[SMIB07] Rajaraman Suryakumar, Jason P. Meyers, Elizabeth L. Irving, and William R. Bobier, *Application of video-based technology for the simultaneous measurement of accommodation and vergence*, Vision research(Oxford) **47** (2007), no. 2, 260–268, ISSN 0042-6989.

[TCP97] Obed Torres, Justine Cassell, and Scott Prevost, *Modeling Gaze Behavior as a Function of Discourse Structure*, Paper presented at the First International Workshop on Human-Computer Conversation (1997).

[TJ00] Vildan Tanriverdi and Robert J. K. Jacob, *Interacting with eye movements in virtual environments*, Conference on Human Factors in Computing Systems, CHI 2000 (New York), ACM Press, 2000, pp. 265–272, ISBN 1-58113-216-6.

[TKFW+79] JH Ten Kate, EEE Frietman, W. Willems, BM Ter Haar Romeny, and E. Tenkink, *Eye-Switch Controlled Communication Aids*, Proceedings of the 12th International Conference on Medical & Biological Engineering (1979), 19–20.

[TSKES95] Michael K. Tanenhaus, Michael J. Spivey-Knowlton, Kathleen M. Eberhard, and Julie C. Sedivy, *Integration of visual and linguistic information in spoken language comprehension*, Science **268** (1995), no. 5217, 1632–1634, ISSN 0036-8075.

[VSvdVN01] Roel P. H. Vertegaal, Robert Slagter, Gerrit C. van der Veer, and Anton Nijholt, *Eye Gaze Patterns in Conversations: There is More to Conversational Agents Than Meets the Eyes*, Proceedings ACM SIGCHI Conference CHI 2001: Anyone. Anywhere, Seattle, USA (New York) (Julie Jacko, Andrew Sears, Michel Beaudouin-Lafon, and Robert J. K. Jacob, eds.), ACM Press, March 2001, pp. 301–308, ISBN 1-58113-327-8.

[Whe38] Charles Wheatstone, *Contributions to the Physiology of Vision. Part the First. On some remarkable, and hitherto unobserved, Phenomena of Binocular Vision*, Philosophical Transactionsof the Royal Society of London **128** (1838), 371–394, ISSN 0261-0523.