Evaluation of Gaze Depth Estimation from Eye Tracking in Augmented Reality

Seyda Z. Öney
st144066@stud.uni-stuttgart.de
University of Stuttgart

Nils Rodrigues
Michael Becher
Thomas Ertl
{first}.{last}@visus.uni-stuttgart.de
University of Stuttgart

Guido Reina
Michael Sedlmair
Daniel Weiskopf
{first}.{last}@visus.uni-stuttgart.de
University of Stuttgart

ABSTRACT

Head-mounted displays for augmented reality can place objects at any distance from the viewer in the real world. Gaze tracking in 3D has the potential to improve interaction with objects and visualizations in augmented reality. However, previous research showed that subjective perception of distance varies between real and virtual surroundings. We wanted to determine whether objectively measured 3D gaze depth through eye tracking also exhibits differences between entirely real and augmented environments. To this end, we conducted an experiment ($N = 25$) in which we used Microsoft HoloLens with a binocular eye tracking add-on from Pupil Labs. Participants performed a task that required them to look at stationary real and virtual objects while wearing a HoloLens device. We were not able to find significant differences in the gaze depth measured by eye tracking. Finally, we discuss our findings and their implications for gaze interaction in immersive analytics, and the quality of the collected gaze data.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; **Mixed / augmented reality**; *Mobile devices.*

KEYWORDS

Augmented reality, eye tracking, depth perception, user study, visualization, immersive analytics

ACM Reference Format:
Seyda Z. Öney, Nils Rodrigues, Michael Becher, Thomas Ertl, Guido Reina, Michael Sedlmair, and Daniel Weiskopf. 2020. Evaluation of Gaze Depth Estimation from Eye Tracking in Augmented Reality. In *Symposium on Eye Tracking Research and Applications (ETRA '20 Short Papers), June 2–5, 2020, Stuttgart, Germany*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3379156.3391835

1 INTRODUCTION

In recent years, augmented reality (AR) has gained importance in many areas, such as industry and gaming. Increasing interest in this field has been leading to further development of AR technologies that enhance the fusion of real and virtual worlds. Head-mounted displays (HMDs), such as Microsoft HoloLens, have high-resolution displays that place virtual objects seemingly anywhere in the real environment of their wearer. Immersive analytics [Marriott et al. 2018] can benefit from this new hardware. For example, volumetric data for scientific visualization could appear to be floating in a real room, ready to be analyzed by domain specialists. Interactive exploration of a dataset in AR often requires hand-held devices for manipulation of viewed objects. 3D gaze has the potential to replace these hand-held devices and allows for intuitive input, e.g., 3D selection, but needs precise measurements of gaze depth with both real and virtual objects in the same environment.

Previous research has shown that depth perception varies between real and virtual environments [Drascic and Milgram 1996; Duchowski et al. 2014]. The vergence–accommodation conflict (VAC) is relevant for this disparity. Human perception of distance takes cues from the relative angle between the orientation of both eyes, as well as their focal depth. Commonly used HMDs display virtual objects at varying distances from the viewer, however, their focal plane remains constant—at 2 m in the case of HoloLens.

In this paper, we address the following research question that is especially relevant for gaze interaction in immersive analytics: does the objectively measured gaze depth vary between real and virtual objects in AR? Our main contributions are aimed at answering this question and start with the design of an experiment that allows us to measure gaze depth with an entirely real and an augmented scene[1]. We conduct the experiments by asking participants to perform a task that requires them to look at real objects or their virtual counter parts at varying distances. We use HoloLens to show virtual objects and measure 3D gaze via binocular eye tracking with an add-on device from Pupil Labs. Finally, we analyze the study data using descriptive and inferential statistics and discuss the implications for gaze interaction in immersive analytics.

2 RELATED WORK

Previous work on depth perception in virtual reality (VR) mostly relied on subjective feedback. One of the first studies used an optical see-through HMD and asked participants to decide if a real or a virtual object was further away [Rolland et al. 1995]. Cubes and cylinders served as targets at distances of 0.8 m to 1.2 m from the observers. Findings suggested that virtual objects seemed farther away than real ones and that the difference between actual and perceived distance was unstable. The authors mention VAC as a possible cause for their results. Later, the method of adjustments replaced constant stimuli to get more precise results [Rolland et al.

[1] https://github.com/UniStuttgart-VISUS/ar-depth-comparison

ETRA '20 Short Papers, June 2–5, 2020, Stuttgart, Germany

ACM ISBN 978-1-4503-7134-6/20/06…$15.00
https://doi.org/10.1145/3379156.3391835

(a) (b) (c) (d)

Figure 1: (a) General layout of the experimental environment, with books (B1, ... B5) placed on a table. (b) Books were covered with visual search puzzles: "How many nines are there?" (c) Black cloth was put on the table in the real test environment; this avoided light shining through when we augmented the scene with virtual objects (d), as seen through the HMD.

2002]. A more recent study employed the perceptual matching method [Swan et al. 2015]. Participants indicated the distance of a target with the help of a pointing device. They observed an overestimation of the distance of virtual target objects that were placed at a distance of 34 to 50 cm. Again, VAC was identified as possible cause. Multiple methods determined that subjective depth perception of real environments within a VR headset (HTC Vive) also results in underestimation [Kelly et al. 2017].

The aforementioned papers aimed at determining the *subjectively perceived* distance of objects in VR. However, we want to *objectively measure* gaze depth. Duchowski et al. [2011] used a Wheatstone stereoscope and an eye tracker as measurement device. However, the results were noisy and required data processing and filtering techniques to yield more reliable values. In later work, eye trackers served again to compare errors in measured gaze depth between real and stereoscopic virtual scenes [Duchowski et al. 2014]. Another approach used dynamic lenses in HMDs to shift the focal plane and avoid the VAC [Johnson et al. 2016]. This worked well when the gaze point was known, however, gaze tracking was not accurate enough to apply this technique to arbitrary virtual scenes. In an effort to improve 3D gaze measurement, Lee et al. [2017] processed the eye tracking data with a machine-learning approach and got an error rate of 10% with an average error distance of 42 cm.

Accurate 2D eye tracking can serve to evaluate visualization techniques [Kurzhals et al. 2014] and has large potential for visual analytics [Büschel et al. 2018; Silva et al. 2019]. Volumetric gaze interaction in AR might also provide great benefits for immersive analytics. However, to the best of our knowledge, no previous research has tried to determine whether the objectively measured gaze depth from eye trackers varies between genuine and augmented reality.

3 EXPERIMENT

Previous research has shown that depth perception differs between real and virtual environments (see Section 2). We assumed that there might also be differences in objectively measured gaze depth between real and augmented reality. Therefore, we conducted an experiment to investigate our hypothesis

> **H**: Gaze depth differs between viewing real and augmented objects, as measured by eye tracking.

3.1 Apparatus

We opted to use Microsoft HoloLens together with an eye tracker from Pupil Labs for multiple reasons. In this way, we were able to keep the gaze measurement device identical between the real and augmented scenario, avoiding device dependency as a confounding variable. Using an optical see-through device, such as HoloLens, allowed us to keep the HMD and the eye tracker on the participant's head and continue measuring without the need for switching hardware or re-calibration. Video see-through was not an option for our experiment, as it presents images of the real environment on the same focal plane as the virtual augmentations. The hardware of the Pupil Labs HoloLens Binocular Add-on integrates well with the HMD. It is small, unintrusive, and does not add much weight. Its right camera stopped functioning during initial work. We replaced it with a compatible camera from a Pupil Core setup. While the original camera supported capturing at up to 200 Hz, the replacement part was only capable of 120 Hz. Therefore, the Pupil Capture software defaulted to the lower frame rate, providing gaze data at up to 120 Hz. We selected low capturing resolutions of 192 × 192 (left) and 320 × 320 (right) because they yielded the highest accuracy in preliminary tests.

We used the Unity game engine for creating and rendering virtual objects. It only called our custom code for data recording once for each rendered frame. Thus, the logging frequency was variable and depended on the rendering performance of HoloLens. An open source Unity package from Pupil Labs facilitated communication and calibration for the eye tracker. The calibration scene was set up to include calibration points on four ellipses at various depths to improve accuracy in the third dimension. The same hardware setup measured gaze depth with real and virtual objects, i.e., the quality of the measurements depended on the alignment between both types of objects. Several ArUco markers at known positions served to align the virtual scene with the real environment.

3.2 Stimulus and Task

As stimulus, we designed a common scene for the study and built a virtual and a real version of it. It consisted of five books B1 to B5 that stood on a table (see Figure 1c and Figure 1d). They were positioned at different distances and angles, as shown in Figure 1a. All books were close to or within the recommended comfort zone of

HoloLens (1.25 m to 5 m). Books B4 and B5 were further away and needed to be larger to remain readable: 21.6 cm × 28.7 cm versus 17.4 cm × 24.1 cm.

We created puzzles that consisted of visual search tasks with letters and numbers (see example in Figure 1b) in order to engage participants and make them fixate different parts of the scene. The tasks were designed to let participants view different objects in the scene at different depths, which served as the data of interest for our study. We replaced the original book covers with these visual puzzles to make the participants look directly at the objects and asked participants to look at the books one after the other.

3.3 Study Design

Our experiment adopted a within-subject design. The independent factor was the type of presented book with two levels: real vs. virtual (within AR). The dependent variable was the error, i.e., the difference between measured gaze depth and actual distance to the currently observed object. The within-subject design led to each participant experiencing both conditions with the same hardware setup, providing measurements without the disadvantages of re-calibration. The order of conditions was counter-balanced to avoid systematic effects from learning.

3.4 Participants

We estimated the required number of participants based on power analysis. With the significance level 0.05, power 0.8, and effect size 0.8 (Cohen's d), we arrived at a minimum sample size $n \geq 25.52$ for a two-sided t-test. We recruited 26 participants (12 female, 14 male). The age ranged from 18 to 33 years (average 23.3 years). Among the participants, there were 22 students, one person with completed vocational training, and one person with a university entrance qualification. Out of these, 16 persons had their major in a field in, or related to, computer science. Twenty reported having some previous experience with AR/VR devices, including three regular AR/VR users.

We used a Snellen chart to check and confirm that all participants had normal or corrected-to-normal vision. Eleven participants wore glasses and two had contact lenses. Participants received 10 EUR as compensation. We had to abort the experiment with one participant due to a hardware failure. Therefore, valid study data is only available for 25 participants.

3.5 Procedure

First, participants were asked to read and sign a consent form—which included a short introduction—and received the monetary compensation. They drew a random number, which became their ID for data anonymization. Then, we went through a demographic questionnaire and checked their eye sight.

We asked participants to sit on a chair and fasten the HoloLens strap tightly to their heads to minimize movement during the experiment. This helped with the quality of eye tracking and positioning of virtual objects in the real environment. However, we could not record objective quality metrics—other than a general accuracy of 1°—because they were not available in the provided software. Once the HMD was strapped to a participant's head, we performed a calibration of HoloLens and the eye tracker.

We started the actual experiment with either the real or virtual scene, depending on the participant's ID that guaranteed counterbalancing. We allowed for a short break—without removing the HMD—so that the conductor of the experiment could manually set up or remove the physical books for the next scene. During the experiment, participants were supposed to perform the search task by targeting a book with the HoloLens' head-gaze cursor and pressing an external clicker to mark the beginning of their task. This hid the cursor and started gaze recording. We stopped recording the gaze as soon as the participant provided their answer to the puzzle and pressed the clicker again. At the end of the 45 to 60-minute experiment, we asked for subjective feedback in a structured questionnaire.

4 RESULTS

We now present the results of our experiment in the form of statistics for measured depth, and feedback from the participants.

4.1 Data Filtering

The measured gaze depth in both scenes contained outliers that extended to −150 km. This would mean that participants looked through their own heads and far behind them. The eye tracker is accurate to about 1°. Therefore, a gaze depth of more than 2 m could theoretically result in a measured distance of infinity. The room in which we conducted the experiment is not longer than 10 m and participants were seated approximately in the center. We discarded any measurements that resulted in a gaze depth of more than $[-10; 10]$ m. However, we did not remove any other data points, to allow for inaccuracies to cancel each other out.

4.2 Analysis of Gaze Data

The filtered data included over 380,000 gaze depth measurements. The number of recorded measurements varied between each person and book because we did not enforce a specific time for the completion of each puzzle and the rendering performance on HoloLens was not constant. We first averaged the error between measured gaze depth and the actual positions of the books for each participant, book, and scene. Then, we used the mean of observations between all books to get the final data by participant and scene. Through this aggregation, we obtained 25 samples (i.e., participants) for each condition (i.e., real vs. virtual). Figure 2 shows the resulting data distributions. The histograms (see Figure 2a) exhibited a tendency toward underestimation when measuring the gaze depth via eye tracking. The median depth error for real books was −1.11 m (M=−1.10, SD=0.87). For augmented books, this value was higher: −1.25 median error (M=−1.09, SD=1.16). While the values differed, the standard deviation indicated very wide distributions in both scenes. Quantiles and confidence intervals (see Figures 2b and 2c) did not exhibit significant differences. The recorded data indicated that the error increased with the book's distance. This seemed plausible, given that the measured depth scales with the tangent of the gaze angle.

The Shapiro-Wilk test confirmed the impression from the histogram in Figure 2a: the data seemed normally distributed, both for the real (W=0.97, p=0.72) and the virtual scene (W=0.96, p=0.35). A paired t-test showed that there is no statistically significant

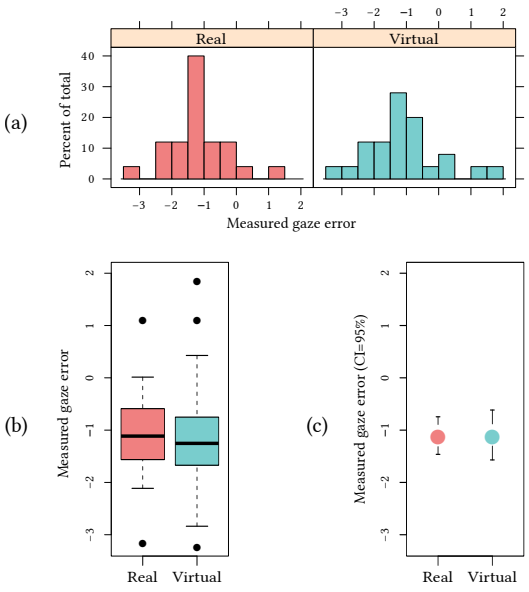

Figure 2: Error in measured gaze depth in the real and virtual scene. A histogram (a) and box plot (b) show the distribution, while the 95% confidence intervals of means is encoded as a dot-and-whisker plot (c).

difference between the two conditions (t=−0.06, df=24, p=0.95, 95%-CI=[−0.37, 0.35], mean of differences=−0.01).

4.3 Subjective Feedback

In the questionnaire, participants generally reported good visibility of all books. However, there was one person who had trouble with real B2. In the virtual scene, one participant had issues with B1, whereas two had problems with B2 and B3. Five persons reported having difficulties with focusing on the books, and one restricted the answer to only the virtual objects. Only two participants were not able to discern real and virtual books. The others mentioned that the virtual ones looked translucent and less clear. Two reported that the books seemed to be glossy. We also asked participants whether they agreed with the observation that "real and virtual books look similar". Their responses were recorded on a Likert scale from 1 (*strongly disagree*) to 5 (*strongly agree*) and suggest a good similarity (mean of 4.04).

Only four people reported they felt absolutely fine during the experiment. Most complained of burning eyes, uncomfortably warm HMD, headaches, or pressure on their head. Three participants reported nausea.

5 DISCUSSION

The t-test showed that the differences between real and virtual books do not seem to be sufficiently large to confirm our hypothesis **H** from Section 3. Our results suggest that, despite *subjective* differences in depth perception found in previous work [Duchowski et al. 2014; Kelly et al. 2017; Rolland et al. 1995; Swan et al. 2015], *objective* measurements from eye tracking do not seem be affected by VAC in AR. This outcome is positive with regards to our motivation: use of eye gaze for intuitive interaction in immersive analytics. However, the eye tracker consistently tended toward underestimation of gaze depth with all books.

Our obtained measurements still lacked precision. The average deviation between actual and estimated gaze depth quickly increased to more than a meter when the focused target was only 3.5 m away from the viewer. Therefore, further experiments with more precise eye trackers are necessary to search for subtle differences in measurements. To increase measurement accuracy, the HMD was firmly strapped to the participants' heads and they were seated on a chair. Despite this endeavor, we were not able to keep the eye tracker tightly attached without any movement. This led to a degradation of the quality of measurements as time progressed in a session of the experiment. Infrared radiation from the eye trackers seemed to dry out and strain the participants' eyes. Combined with discomfort from the HMD, this might have led to a loss of focus, fatigue, and an increased blink rate, which might have also decreased the precision of our measurements.

Despite the difficulties with eye tracking hardware, the visual appearance of the virtual books in AR was well accepted. Participants noticed differences to the real objects but generally found them to be a close match and were able to complete the tasks. This suggests that the study design itself was sound.

6 CONCLUSION AND FUTURE WORK

We investigated the effect of AR on objectively measured gaze depth. For this purpose, we used Microsoft HoloLens with a binocular eye tracking add-on from Pupil Labs. To analyze the gaze depth estimation, we created a real and a virtual scene of books on a table and recorded 3D gaze in an experiment with 25 participants. The resulting data did not indicate that objective measurements are affected by subjective differences in depth perception between genuine and augmented reality. This could allow for intuitive interaction in immersive analytics, e.g., volumetric 3D selection. However, eye trackers still have deficits in accurately capturing the 3D gaze positions. There were many outliers that could have an influence on the result. In addition, tightly strapping the HoloLens device caused discomfort to the users, potentially influencing their eye focus and thus the gaze data.

Microsoft's second-generation HoloLens will have integrated eye tracking hardware and provide access to a single gaze ray [Stellmach and Microsoft Corporation 2020]. The combination of built-in support for eye gaze and new calibration-free eye tracking, such as with Pupil Invisible [Pupil Labs GmbH 2020], could provide more accurate 3D eye tracking and enable easier integration into immersive analytics, realizing advances that researchers have previously envisioned [Silva et al. 2019]. In future work, we want to revisit the study to investigate the accuracy of newer eye tracking hardware and the effects of fatigue from HMDs [Wang et al. 2019].

ACKNOWLEDGMENTS

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project ID 251654672 – TRR 161 (projects A08 and B01) and – project ID 279064222 – SFB 1244 (project B05).

REFERENCES

W. Büschel, J. Chen, R. Dachselt, S. Drucker, T. Dwyer, C. Görg, T. Isenberg, A. Kerren, C. North, and W. Stuerzlinger. 2018. Interaction for immersive analytics. In *Immersive Analytics*. Springer International Publishing, 95–138.

D. Drascic and P. Milgram. 1996. Perceptual issues in augmented reality. In *Stereoscopic Displays and Virtual Reality Systems III*, Vol. 2653. International Society for Optics and Photonics, 123–134.

A. T. Duchowski, D. H. House, J. Gestring, R. Congdon, L. Świrski, N. A. Dodgson, K. Krejtz, and I. Krejtz. 2014. Comparing estimated gaze depth in virtual and physical environments. In *Proceedings of the 8th ACM Symposium on Eye Tracking Research & Applications, ETRA 2014*. ACM, 103–110.

A. T. Duchowski, B. Pelfrey, D. H. House, and R. Wang. 2011. Measuring gaze depth with an eye tracker during stereoscopic display. In *Proceedings of the ACM SIGGRAPH Symposium on Applied Perception in Graphics and Visualization, APGV 2011*. ACM, 15–22.

P. V. Johnson, J. AQ. Parnell, J. Kim, C. D. Saunter, G. D. Love, and M. S. Banks. 2016. Dynamic lens and monovision 3D displays to improve viewer comfort. *Optics Express* 24, 11 (May 2016), 11808–11827.

J. W. Kelly, L. A. Cherep, and Z. D. Siegel. 2017. Perceived space in the HTC Vive. *ACM Transactions on Applied Perception* 15, 1, Article 2 (Jul 2017), 16 pages.

K. Kurzhals, B. D. Fisher, M. Burch, and D. Weiskopf. 2014. Evaluating visual analytics with eye tracking. In *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization, BELIV 2014*. ACM, 61–69.

Y. Lee, C. Shin, A. Plopski, Y. Itoh, T. Piumsomboon, A. Dey, G. Lee, S. Kim, and M. Billinghurst. 2017. Estimating gaze depth using multi-layer perceptron. In *Proceedings of the 2017 International Symposium on Ubiquitous Virtual Reality, ISUVR 2017*. IEEE, 26–29.

K. Marriott, F. Schreiber, T. Dwyer, K. Klein, N. H. Riche, T. Itoh, W. Stuerzlinger, and B. H. Thomas. 2018. *Immersive Analytics*. Springer International Publishing.

Pupil Labs GmbH. 2020. Pupil Labs | Pupil Invisible - The world's first deep learning powered eye tracking glasses. https://pupil-labs.com/products/invisible/. Accessed: 2020-02-17.

J. P. Rolland, W. Gibson, and D. Ariely. 1995. Towards quantifying depth and size perception in virtual environments. *Presence: Teleoperators & Virtual Environments* 4, 1 (Winter 1995), 24–49.

J. P. Rolland, C. Meyer, K. Arthur, and E. Rinalducci. 2002. Method of adjustments versus method of constant stimuli in the quantification of accuracy and precision of rendered depth in head-mounted displays. *Presence: Teleoperators & Virtual Environments* 11, 6 (Dec 2002), 610–625.

N. Silva, T. Blascheck, R. Jianu, N. Rodrigues, D. Weiskopf, M. Raubal, and T. Schreck. 2019. Eye tracking support for visual analytics systems: foundations, current applications, and research challenges. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications, ETRA 2019*. ACM, 11:1–11:10.

S. Stellmach and Microsoft Corporation. 2020. Eye tracking - Mixed Reality | Microsoft Docs. https://docs.microsoft.com/en-us/windows/mixed-reality/eye-tracking. Accessed: 2020-02-17.

J. E. Swan, G. Singh, and S. R. Ellis. 2015. Matching and reaching depth judgments with real and augmented reality targets. *IEEE Transactions on Visualization and Computer Graphics* 21, 11 (Nov 2015), 1289–1298.

Y. Wang, G. Zhai, S. Chen, X. Min, Z. Gao, and X. Song. 2019. Assessment of eye fatigue caused by head-mounted displays using eye-tracking. *BioMedical Engineering OnLine* 18, 1 (Nov 2019), 111.