Stress testing VR Eye-tracking System Performance

Ashima Keshava[1], Farbod Nosrat Nezami[1], Nora Maleki[1], Linus Tiemann[1], Peter König[1,2]
[1] University of Osnabrück, Germany
[2]University Medical Center Hamburg-Eppendorf, Germany

Eye-tracking experiments in virtual reality (VR) have become progressively popular in the last decade. These experiments measure human eye movement behavior in naturalistic settings that afford complex, natural head and body movements. Given the complexity, eye-tracking systems require high spatial accuracy and precision of the measured gaze in the face of natural movements, differing illumination, depth of field, and calibration decay. (Holmqvist et al., 2012) have stressed the need for assessing eye-tracking data quality in general. Furthermore, there is a lack of data quality standards when it comes to VR head-mounted displays specifically. The present study aims to introduce a standardized way of benchmarking VR eye-tracking systems to assess their feasibility for vision research in mobile settings.

We adapted a 2D screen-based eye-tracker test battery (Ehinger et al., 2019) to VR-based head-mounted displays. The test battery includes ten spatial accuracy and precision tests for standard gaze parameters like gaze position, pupil dilation, blink detection, and smooth pursuit. We then used the test battery to compare the performance of two commercially available VR head-mounted displays (HTC Vive Pro Eye and Varjo VR-2 Pro) with a built-in eye-tracker for 13 participants (Figure 1A).

Here, we report our results based on the most critical metrics, namely: **1. Spatial Accuracy** (Figure 1B)**:** we calculated the calibration error across a 5x5 grid of fixation locations. Our results show that both HTC Vive Pro Eye and Varjo VR-2 Pro have a mean calibration error greater than 1 degree without an explicit validation of the calibration accuracy. In the horizontal direction, HTC Vive had a mean error of 1.28°, IQR=[0.60, 1.17], and Varjo had a mean error of 3.29°, IQR=[1.55, 3.45]. In the vertical direction, HTC Vive had a mean error of 0.89°, IQR=[0.50, 1.19] and Varjo had a mean error of 4.93°, IQR=[2.24, 6.93]; **2. Spatial Precision** (Figure 1C): we used the median absolute deviation of the calibration error across the 5x5 grid to measure the eye trackers' spatial precisions. We found that the mean precision in the horizontal axis of the HTC Vive Pro Eye was 3.22° (SD: ±0.75) and 4.76° ± 1.48 for Varjo. In the vertical direction, HTC VIVE Pro Eye had a precision of 1.86°± 0.76, and Varjo had 4.02°±2.33; **3. Calibration Decay:** to assess the decay of calibration during the experiment, we calculated the mean difference in calibration error just after eye-tracker calibration and at the end of each test block. HTC Vive showed a mean calibration decay of 4.09° ± 1.06 in the horizontal direction and 3.21° ± 1.09 in the vertical direction. In contrast, the Varjo system showed a calibration decay of 5.86° ± 2.46 in the horizontal direction and 5.11° ±1.95 in the vertical direction; **4. Effect of Illumination on Pupil Dilation** (Figure 1D): we further investigated the pupil size detection differences between the eye trackers for different illumination levels. Our results show that the Varjo VR-2 eye tracker estimated larger normalized pupil sizes than the HTC VIVE (mean difference = 2.55 % ± 4.47); **5. Blink Detection** (Figure 1E): we investigated how well the two eye trackers detected blinks by asking subjects to voluntarily blink 10 times during a test block. We found the HTC Vive Pro Eye detected 10.49 ± 3.14 blinks, and the Varjo VR-2 Pro system detected 1.40 ± 1.82 blinks; **6. Smooth Pursuit** (Figure 1F): In the smooth pursuit task, we found that the HTC Vive

system tracks the eyes at -0.18°/s ± 4.11 slower than the stimulus velocity, whereas the Varjo system tracks the eyes at -3.84°/s ± 3.27 slower than the stimulus velocity.

Our results show that both VR eye-tracking systems are somewhat error-prone and can have high variance across different subjects. Hence, vision researchers should not take the quality of the data measured by these systems as a given. In studies that rely on high spatial accuracy or measurement of specific gaze features like blinks or pupil dilation, the eye-tracking equipment alone can make an immense difference. Our study offers an implemented test battery to evaluate and benchmark VR eye-tracking systems based on several gaze features useful for naturalistic experiments. The tests can comprehensively assess the quality of commercially available VR eye trackers beyond the values provided by the manufacturers. Furthermore, we have made the VR setup, the collected data, and the analysis pipeline available publicly to help researchers adapt this study for any VR-based eye tracker.

References

Ehinger, B. V., Groß, K., Ibs, I., & König, P. (2019). A new comprehensive eye-tracking test battery concurrently evaluating the Pupil Labs glasses and the EyeLink 1000. *PeerJ*, *7*, e7086. https://doi.org/10.7717/peerj.7086

Holmqvist, K., Nyström, M., & Mulvey, F. (2012). Eye tracker data quality. *Proceedings of the Symposium on Eye Tracking Research and Applications - ETRA '12*. the Symposium, Santa Barbara, California. https://doi.org/10.1145/2168556.2168563

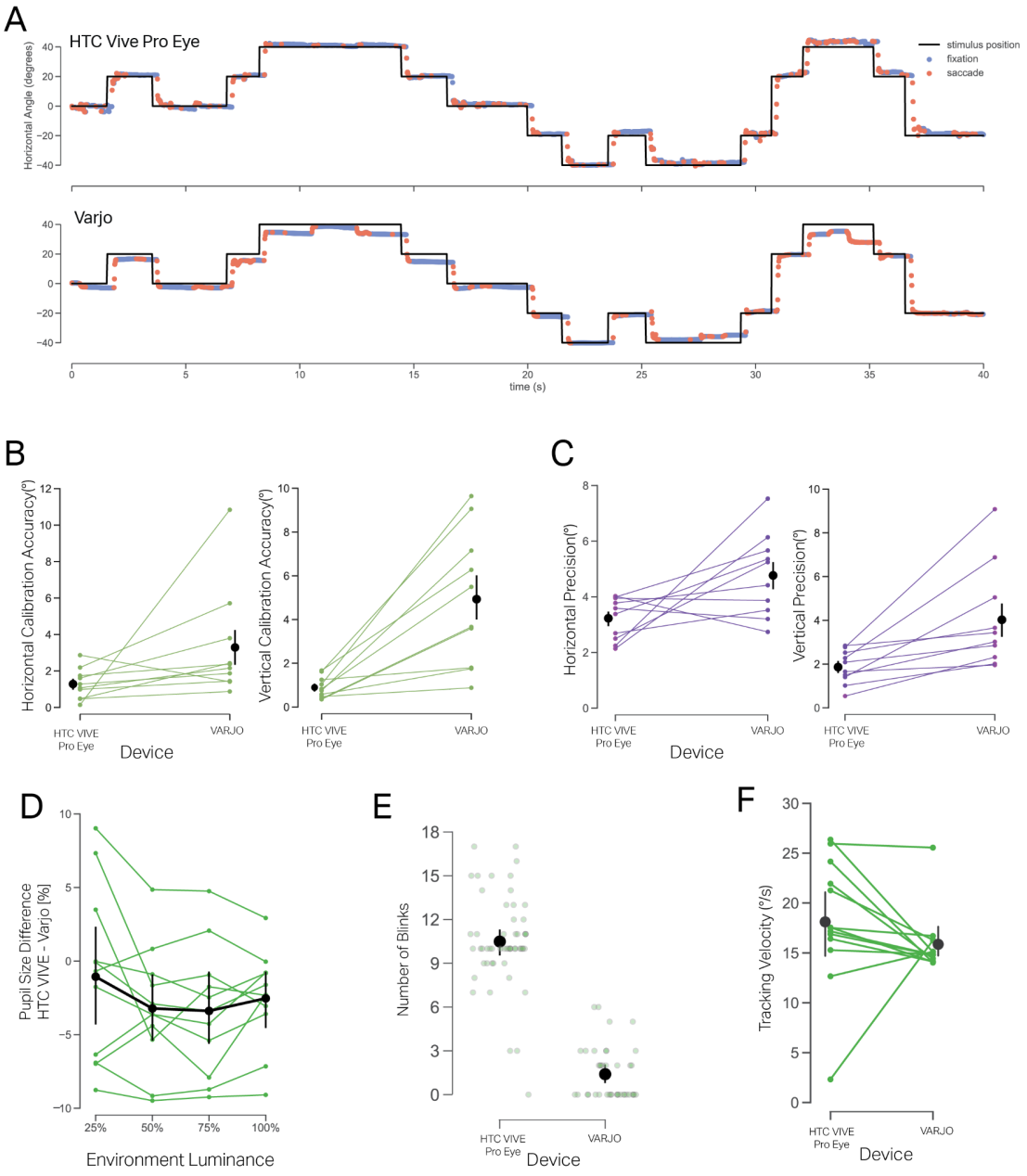

Figure 1. A) Exemplar raw data showing horizontal gaze angle for the fixation probe shown in VR for the two head-mounted displays (HMDs). The blue samples correspond to fixations and the orange to the saccades. **B)** The calibration error for the two HMDs across 10 subjects. We computed the calibration error (in visual degrees) as the 20% winsorized mean of the difference between the fixation probe position and the actual fixation position. Thus, each dot represents one subject and the calibration error for the two devices. **C)** The precision of the HMDs across subjects. Here, we used median absolute deviation as a metric of precision, where lower values correspond to high precision and vice versa. **D)** % Difference in the normalized pupil sizes measured by the two devices for the different environment luminance. Green dots indicate each subject, and the black dots represent mean difference, and the error bars represent the standard error of mean. **E)** Number of Blinks detected by the eye tracker. Each green dot represents the number of blinks per subject and test block. The filled black dots represent the mean number of blinks, and the error bars show the standard error of the mean. **F)** Velocity of the tracked gaze for a moving stimulus during smooth pursuit. Green dots indicate each subject, and the black dots represent mean difference, and the error bars represent the standard error of mean.