SLAM-based Localization of 3D Gaze using a Mobile Eye Tracker

Haofei Wang*
Hong Kong University of Science and Technology
Hong Kong
hwangba@connect.ust.hk

Jimin Pi*
Hong Kong University of Science and Technology
Hong Kong
jpi@connect.ust.hk

Tong Qin
Hong Kong University of Science and Technology
Hong Kong
tong.qin@connect.ust.hk

Shaojie Shen
Hong Kong University of Science and Technology
Hong Kong
eeshaojie@ust.hk

Bertram E. Shi
Hong Kong University of Science and Technology
Hong Kong
eebert@ust.hk

ABSTRACT

Past work in eye tracking has focused on estimating gaze targets in two dimensions (2D), e.g. on a computer screen or scene camera image. Three-dimensional (3D) gaze estimates would be extremely useful when humans are mobile and interacting with the real 3D environment. We describe a system for estimating the 3D locations of gaze using a mobile eye tracker. The system integrates estimates of the user's gaze vector from a mobile eye tracker, estimates of the eye tracker pose from a visual-inertial simultaneous localization and mapping (SLAM) algorithm, a 3D point cloud map of the environment from a RGB-D sensor. Experimental results indicate that our system produces accurate estimates of 3D gaze over a much larger range than remote eye trackers. Our system will enable applications, such as the analysis of 3D human attention and more anticipative human robot interfaces.

CCS CONCEPTS

• **Human-centered computing** → *Human computer interaction (HCI)*; *Interaction devices*;

KEYWORDS

3D Gaze Estimation, SLAM, Point Cloud, Eye Tracker, RGB-D camera, Human-Robot Interaction

ACM Reference Format:
Haofei Wang, Jimin Pi, Tong Qin, Shaojie Shen, and Bertram E. Shi. 2018. SLAM-based Localization of 3D Gaze using a Mobile Eye Tracker. In *ETRA '18: 2018 Symposium on Eye Tracking Research and Applications, June 14–17, 2018, Warsaw, Poland*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3204493.3204584

*These authors contributed equally to this work.

ACM ISBN 978-1-4503-5706-7/18/06…$15.00
https://doi.org/10.1145/3204493.3204584

1 INTRODUCTION

Eye gaze tracking has been used successfully as an input modality for interacting with computers, e.g. gaze based typing [Majaranta et al. 2009; Pi and Shi 2017; Räihä and Ovaska 2012] or cursor control [Dong et al. 2015], and for controlling assistive devices, such as wheelchairs [Ktena et al. 2015; Wästlund et al. 2010; Yu et al. 2014], robot arms [Admoni and Srinivasa 2016; Dziemian et al. 2016; Wang et al. 2015], exoskeletons [Katyal et al. 2013; McMullen et al. 2015, 2014], and drones [Hansen et al. 2014; Yu et al. 2014].

Gaze estimation depends upon eye tracking, the estimation of the eye position and orientation in some coordinate frame, usually associated with the measurement apparatus. If the measurement apparatus is fixed in the world, as with the scleral contact lens/search coil or a remote eye tracker, then eye position/orientation is provided in world-centric coordinate systems. If the measurement apparatus is attached to the head, as with Electro-OculoGraphy or in head-mounted eye trackers, then the eye tracker returns the eye position/orientation in head-centric coordinates. In this case, the position and orientation of the head must be estimated in order to map the eye position/orientation to gaze direction. When head position is fixed in world coordinates, e.g. by a bite bar or chin rest as common with tower mounted systems, the two coordinate systems are related by a fixed transformation.

Most applications to date have used 2D gaze estimates provided either in world-centric coordinates or in head-centric coordinates. Remote eye trackers typically provide 2D gaze estimates on a screen located at a fixed position in front of the user and above or below the eye tracker. Head-mounted eye trackers typically provide 2D gaze estimates on an image taken by a scene camera which is mounted on the same set of goggles holding the eye tracker.

There is increasing interest in estimating gaze target locations in 3D world coordinates. For example, 3D gaze estimates hold promise as an excellent cue for human robot interaction. By knowing where in the environment a user is looking, a robot may be able to infer something about the user's intent and provide anticipative assistance or to suggest possible types of assistance it might render.

The most common approach to estimate 3D gaze position is by triangulating gaze vectors from the left and right eyes [Abbott and Faisal 2012; Li et al. 2017; Pirri et al. 2011]. However, this approach results in large variance along the line of sight, due to errors in estimating eye gaze as well as fixational eye movements [Jacob

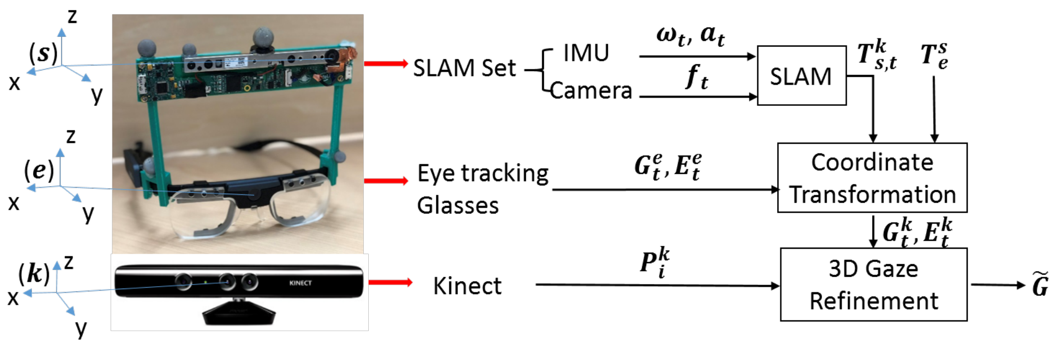

Figure 1: System architecture.

1993]. Hennessey et al. addressed this problem by filtering the raw gaze data [Hennessey and Lawrence 2009]. Unfortunately, this introduces a system delay depending on the length of filters. Wang et al. improved initial 3D gaze estimates obtained by triangulation using a point cloud representation of the environment by assuming that gaze targets are located on object surfaces [Wang et al. 2017].

There is typically a trade-off between the allowable user movement and the accuracy of gaze estimates. Tower-mounted eye trackers have the least tolerance for head movement, but produce more accurate estimates than remote eye trackers, which tolerate a small amount of head movement [Holmqvist et al. 2012]. Popular remote eye trackers currently available on the market claim head box ranges of 40×40 cm^2 (SR Research EyeLink 1000Plus), 50×30 cm^2 (SMI REDn) and 44×22 cm^2 (Tobii T60XL) [Niehorster et al. 2018]. The allowable range of head movements can be increased through the use of multiple cameras (SmartEye, http://smarteye.se/) or by mounting the eye-tracker camera on motorized gimbals controlled by a wide range camera that tracks user's head position (Eyefollower, http://www.eyegaze.com). The range of orientations allowed by remote eye tracking systems is limited, as they require the user to face towards the eye tracker cameras, which are typically fixed. Head-mounted eye trackers provide gaze estimates allow for the widest range of head movements and orientations. However, they provide the gaze estimates in head-centric coordinates. In order to translate these gaze estimates to world-centric coordinates, the position and orientation of the eye tracker must be estimated dynamically as the user moves. This problem has only begun to attract investigation [Paletta et al. 2013].

In this paper, we propose an algorithm for estimating gaze targets in 3D world-centric coordinates while also allowing for a wide range of head movements. Our algorithm combines head-centric gaze estimates from a head-mounted eye tracker with estimates of head position and orientation obtained by visual-inertial simultaneous localization and mapping (SLAM) to estimate 3D gaze vectors in world-centric coordinates. These initial gaze vectors are refined using a point cloud representation of the environment. Our experimental results with this system show that it tolerates larger head displacements than a system based upon a remote eye tracker. The system has a mean angular error of only 2.9±1.0 degrees over a testing range covering 13 square meters (400×335 cm^2).

This work improves upon past work in several ways. First, in contrast to past work that provided gaze estimates only in head-centric coordinates [Abbott and Faisal 2012; Li et al. 2017; Pirri et al. 2011], our algorithm provides gaze estimates in world-centric coordinates. The algorithm also allows for a wider range of head movements than previously reported by [Wang et al. 2017], who requested that the subjects keep their heads still, and by [Hennessey and Lawrence 2009], who reported results for head movements only over the range of $3.2 \times 9.2 \times 14$ cm (horizontal×vertical×depth). Finally, by using SLAM to dynamically estimate head pose and map the environment based on the past trajectory and image data, we expect that our system will be more robust to changes in the environment than approaches which use SLAM to obtain a pre-computed static model of the environment and localize the camera using image matching with the eye tracker's scene camera image [Paletta et al. 2013].

2 METHOD

Fig. 1 shows the system architecture. The system contains three sensor systems: a set of sensors for the SLAM algorithm (an inertial measurement unit (IMU) and an RGB camera), a head-mounted eye tracker, and a Kinect red-green-blue-depth (RGBD) sensor. We define a coordinate system for each sensor system: SLAM (s), eye tracker (e) and Kinect (k). Because the SLAM sensors are physically fixed to the eye tracker, the transformation from eye tracker to SLAM coordinates T^s_e is constant. However, the transformation from SLAM to Kinect coordinates $T^k_{s,t}$ changes over time t.

The IMU and RGB camera of the SLAM sensor set provides gyroscope measurements ω_t, accelerometer measurements a_t and image frames f_t, where t indexes time. The eye tracking glasses provide estimates of the 3D gaze location G^e_t and the average of the left and right pupil positions E^e_t in eye tracker coordinates. The Kinect RGB-D sensor provides a description of the environment as a 3D point cloud $\{P^k_i\}^N_{i=1}$ in Kinect coordinates, where N is the number of points.

We use a SLAM algorithm [Qin et al. 2017] to estimate $T^k_{s,t}$ from the past gyroscope, accelerometer and image data. We use this to map G^e_t and E^e_t to Kinect coordinates, e.g.,

$$G^k_t = T^k_{s,t} \cdot T^s_e \cdot G^e_t \tag{1}$$

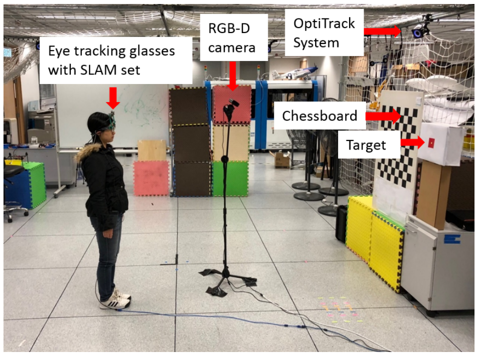

Figure 2: Experimental setup.

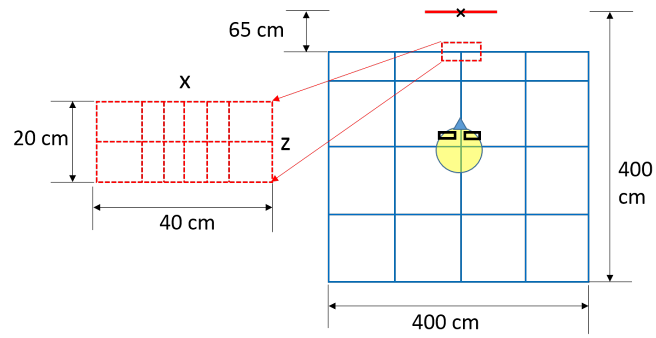

Figure 3: The locations used to evaluate the performance.

Finally, we refine the 3D gaze estimate finding the point in $\{P_i^k\}_{i=1}^N$ that is closest to the cyclopean line of sight $G_t^k - E_t^k$ [Wang et al. 2017]. The following subsections describe these operations in more detail.

2.1 Camera calibration

We estimate the transformations between the three coordinate systems: (s), (e), and (k) using images of a chessboard pattern taken by the SLAM camera, the eye tracker scene camera and the Kinect RGB camera simultaneously at an initial time (t_0). We use camera calibration to estimate the transformations between the three coordinate systems at time t_0 and the chessboard coordinates (c): $T_{e,t_0}^c, T_{s,t_0}^c, T_k^c$, where we assume that the chessboard pattern is stationary in Kinect coordinates. Since T_e^s is also time invariant

$$T_e^s = T_c^{s,t_0} \cdot T_{e,t_0}^c \tag{2}$$

2.2 Eye tracker pose estimation

We use VINS-Mono [Qin et al. 2017], a monocular visual-inertial SLAM algorithm to estimate the transformation $(T_{s,t}^k)$. The addition of inertial data enables us to estimate metric scale information, which cannot be obtained from a monocular camera alone. The VINS-Mono algorithm features a robust procedure for estimator initialization and failure recovery, high accuracy visual-inertial odometry by fusing pre-integrated IMU measurements and feature observations by nonlinear optimization, and a loop detection module for re-localization with minimum computation overhead.

The VINS-Mono provides estimates of the SLAM IMU/camera pose relative to the chessboard, $T_{s,t}^c$. We obtain the transformation from SLAM to Kinect coodinates by $T_{s,t}^k = T_c^k \cdot T_{s,t}^c$, where T_c^k is obtained by camera calibration.

2.3 3D gaze estimation refinement

Since the eye tracker computes 3D gaze estimates by triangulating gaze vector estimates for the two eyes, small errors in estimating the gaze direction lead to large errors in the 3D gaze estimates, especially along the depth dimension [Hennessey and Lawrence 2009]. To reduce these errors, we refined the raw 3D gaze estimates by taking into account the structure of the environment. Given initial estimates of the cyclopean eye and 3D gaze positions, E^k and G^k, we obtained a refined gaze estimate $\widetilde{G}_t^k = P_j^k$ where

$$j = \underset{i}{\operatorname{argmin}} \frac{|(P_i - E^k) \times (P_i - G^k)|}{|(P^k - G^k)|} \tag{3}$$

3 EXPERIMENTS

Our experimental system is based on a Tobii Pro Glasses 2 head-mounted eye tracker, which provides gaze estimates at 50 Hz. The SLAM camera is an Intel Real-sense camera ZR300, which has a global shutter, a fish-eye lens resulting in a $100° \times 133°$ field of view, and a frame rate of 30Hz. The sampling rate of the IMU is 350 Hz. The SLAM algorithm computes pose estimates at 10 Hz.

Fig. 2 shows the experimental setup. To provide ground truth estimates of the pose of the SLAM system, we use an OpiTrack system (http://optitrack.com/) mounted at the ceiling and markers attached to the SLAM sensors (See Fig. 1). Before each experiment, we calibrated the eye tracker and the cameras. The subject calibrated the eye tracker using the manufacturer's provided one-point calibration.

We conducted three experiments to evaluate our system.

In the first experiment, we compared the performance of our system with that of a remote eye tracker, the Tobii X60. We used the method in [Wang et al. 2017] to estimate the 3D gaze position. We collected gaze data for subjects standing at the 21 positions shown as the red grid in Fig. 3. This grid is centered at a point directly in front of and 65 cm away from the X60, and covers a range of 40×20 cm. We chose these points because the recommended operating distance of the X60 is 65 cm, and the points all lie within the 44×22×30 cm^3 (width×height×depth) headbox advertised by Tobii.

We evaluated the accuracy for six subjects. Subjects stood at each of the positions and gazed at a fixation cross located inside the red square target shown in Fig. 2. For each subject, each location and each of the two systems, we collected 100 3D gaze estimates over a period of 2 seconds.

In the second experiment, we evaluated the proposed system over a much wider range of positions shown as the blue grid covering a range of 400×335 cm, as shown in Fig. 3. We evaluated the accuracy of the system for six subjects. The subjects stood at the each of the locations and gazed at the fixation cross. At each location, we collected 100 3D gaze estimates over a period of 2 seconds. The SLAM algorithm was used to estimate pose continuously, both

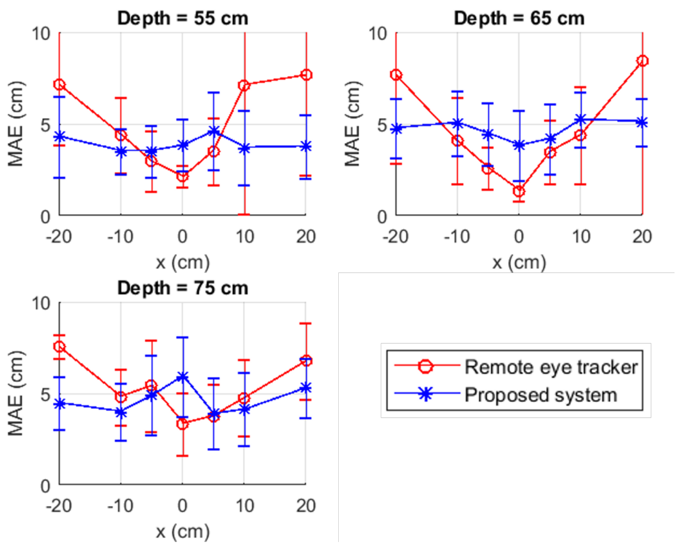

Figure 4: Performance comparison between the remote eye tracker and proposed system. Error bars show the standard deviation computed across all samples and all subjects.

while the users were standing at each position and as they were moving from point to point. Since subjects need to look down to locate the test positions, head movements included both translation and rotation.

In the final experiment, we recorded the estimated gaze as user moved in a more natural office environment.

4 RESULTS

4.1 Comparison with remote eye tracker

Fig. 4 compares performance of the proposed system with the remote eye tracker, as measured by the mean absolute error, the mean Euclidean distance from the estimated 3D gaze location to fixation cross.

For the remote eye tracker, the error was smallest when the subject was directly in front of the eye tracker, but degraded rapidly as the subject moved to the left or right (along the x axis). Consistent with [Blignaut and Wium 2014], we found that despite the advertised horizontal range of 44cm, error increased dramatically 10cm to the left or right of the calibration point.

The error of the proposed system was larger than that of the remote eye tracker when the subject was directly in front of the remote eye tracker, but remained stable over the entire testing range. It was lower than that of the remote eye tracker at the left and right endpoints.

In comparison to other work reporting results in world-centric coordinate systems, our system works over a wider range (40 cm in x and 20 cm in depth), albeit with a larger mean absolute error 4.4±1.9 cm. Hennessey et al. reported a mean absolute error of 3.9±2.8 cm over a range 3.2 cm in x, 14 cm in depth and 9.2 cm in height [Hennessey and Lawrence 2009]. Wang et al. reported a mean absolute error of 1.7±0.7 cm, but evaluated only at the calibration position [Wang et al. 2017].

Table 1: Angular error at different positions(degrees)

z \ x	-200cm	-100cm	0cm	100cm	200cm
65cm	1.8±0.9	3.0±0.8	2.8±1.3	3.5±1.2	2.4±1.3
100cm	2.3±0.9	2.6±0.9	3.6±0.6	3.2±0.8	2.9±1.2
200cm	2.2±0.8	3.1±0.5	3.1±0.7	3.2±0.8	2.7±0.9
300cm	2.3±0.8	2.8±0.8	3.3±1.0	3.4±0.7	3.0±0.7
400cm	2.6±1.0	2.6±0.7	3.1±0.8	2.7±0.5	2.7±0.7

4.2 Performance over larger range

For each 3D gaze estimate, we define the angular error as the angle of the vectors from the ground truth eye position as given by the OptiTrack system to the positions of the estimated 3D gaze and the fixation cross. The mean angular error of the gaze estimates from our system over users and all positions shown on the blue grid in Fig. 3 was 2.9 degrees. The standard deviation computed over all users and all positions was 1.0 degrees. The standard deviation of the mean angular errors of the 3D gaze estimates at the 25 positions is 0.4 degrees, suggesting that the system performance is stable at different positions. The standard deviations of the mean angular errors in the 3D gaze estimates for different subjects is 0.4 degrees. The standard deviations of the mean angular error of the orientation estimates from SLAM for different subjects was also 0.4 degrees. These results both suggest that the performance is fairly subject independent.

There are several sources of this angular error. The first is angular error in the 3D gaze estimates from the eye tracker. The second is the angular error in the estimates of the orientation of the eye tracker from the SLAM algorithm. The third is the Euclidean error in the estimates of the location of the eye tracker from SLAM. Comparing with the head pose estimates from the OptiTrack system, we estimated the mean angular error of SLAM estimates to be 1.4 ± 0.9 degrees and the mean translational error to be 8.4 ± 5.9 cm.

4.3 Performance in office environment

We tested the system in a more natural office settings, where a subject viewed different objects (a box, a book and a cup) placed on a table as the subject moved around the room. A video of the experiment can be found at (https://youtu.be/t5c88_bhRsY), demonstrating that our system provides accurate 3D gaze estimates in fairly unconstrained indoor environments, and therefore might be used during activities of daily living.

5 CONCLUSIONS

We have proposed a novel system for estimating the 3D locations of gaze targets in world-centric coordinates by integrating gaze estimates from a head-mounted eye tracker, head pose estimates from monocular visual inertial SLAM, environmental point cloud data from an RGB-D sensor. The key advantages of this work over prior approaches are the availability of accurate estimates in world-centric coordinates, a larger working range, and robustness to changes in the environment. We are currently applying our system to human robot interaction in the 3D environment.

REFERENCES

William Welby Abbott and Aldo Ahmed Faisal. 2012. Ultra-low-cost 3D gaze estimation: an intuitive high information throughput compliment to direct brain–machine interfaces. *Journal of Neural Engineering*. 9, 4 (2012), 046016.

Henny Admoni and Siddhartha Srinivasa. 2016. Predicting user intent through eye gaze for shared autonomy. In *Proceedings of the AAAI Fall Symposium Series: Shared Autonomy in Research and Practice (AAAI Fall Symposium)*. 298–303.

Pieter Blignaut and Daniël Wium. 2014. Eye-tracking data quality as affected by ethnicity and experimental design. *Behavior Research Methods*. 46, 1 (2014), 67–80.

Xujiong Dong, Haofei Wang, Zhaokang Chen, and Bertram E Shi. 2015. Hybrid brain computer interface via Bayesian integration of EEG and eye gaze. In *7th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 150–153.

Sabine Dziemian, William W Abbott, and A Aldo Faisal. 2016. Gaze-based teleprosthetic enables intuitive continuous control of complex robot arm use: Writing & drawing. In *6th IEEE International Conference on Biomedical Robotics and Biomechatronics (BioRob)*. IEEE, 1277–1282.

John Paulin Hansen, Alexandre Alapetite, I Scott MacKenzie, and Emilie Møllenbach. 2014. The use of gaze to control drones. In *Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA '14)*. ACM, 27–34.

Craig Hennessey and Peter Lawrence. 2009. Noncontact binocular eye-gaze tracking for point-of-gaze estimation in three dimensions. *IEEE Transactions on Biomedical Engineering* 56, 3 (2009), 790–799.

Kenneth Holmqvist, Marcus Nyström, and Fiona Mulvey. 2012. Eye tracker data quality: what it is and how to measure it. In *Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA '12)*. ACM, 45–52.

Robert JK Jacob. 1993. Eye movement-based human-computer interaction techniques: Toward non-command interfaces. *Advances in Human-Computer Interaction* 4 (1993), 151–190.

Kapil D Katyal, Matthew S Johannes, Timothy G McGee, Andrew J Harris, Robert S Armiger, Alex H Firpi, David McMullen, Guy Hotson, Matthew S Fifer, Nathan E Crone, et al. 2013. HARMONIE: A multimodal control framework for human assistive robotics. In *6th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 1274–1278.

Sofia Ira Ktena, William Abbott, and A Aldo Faisal. 2015. A virtual reality platform for safe evaluation and training of natural gaze-based wheelchair driving. In *7th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 236–239.

Songpo Li, Xiaoli Zhang, and Jeremy Webb. 2017. 3D-Gaze-based robotic grasping through mimicking human visuomotor function for people with motion impairments. *IEEE Transactions on Biomedical Engineering* (2017).

Päivi Majaranta, Ulla-Kaija Ahola, and Oleg Špakov. 2009. Fast gaze typing with an adjustable dwell time. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 357–360.

David P McMullen, Matthew S Fifer, Brock A Wester, Guy Hotson, Kapil D Katyal, Matthew S Johannes, Timothy G McGee, Andrew Harris, Alan D Ravitz, Michael P McLoughlin, et al. 2015. Semi-autonomous hybrid brain-machine interface. In *Brain-Computer Interface Research*. Springer, 89–104.

David P McMullen, Guy Hotson, Kapil D Katyal, Brock A Wester, Matthew S Fifer, Timothy G McGee, Andrew Harris, Matthew S Johannes, R Jacob Vogelstein, Alan D Ravitz, et al. 2014. Demonstration of a semi-autonomous hybrid brain–machine interface using human intracranial EEG, eye tracking, and computer vision to control a robotic upper limb prosthetic. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 22, 4 (2014), 784–796.

Diederick C Niehorster, Tim HW Cornelissen, Kenneth Holmqvist, Ignace TC Hooge, and Roy S Hessels. 2018. What to expect from your remote eye-tracker when participants are unrestrained. *Behavior research methods* 50, 1 (2018), 213–227.

Lucas Paletta, Katrin Santner, Gerald Fritz, Heinz Mayer, and Johann Schrammel. 2013. 3D attention: measurement of visual saliency using eye tracking glasses. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*. ACM, 199–204.

Jimin Pi and Bertram E Shi. 2017. Probabilistic adjustment of dwell time for eye typing. In *10th International Conference on Human System Interactions (HSI)*. IEEE, 251–257.

Fiora Pirri, Matia Pizzoli, and Alessandro Rudi. 2011. A general method for the point of regard estimation in 3D space. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 921–928.

Tong Qin, Peiliang Li, and Shaojie Shen. 2017. Vins-mono: A robust and versatile monocular visual-inertial state estimator. *arXiv preprint arXiv:1708.03852* (2017).

Kari-Jouko Räihä and Saila Ovaska. 2012. An exploratory study of eye typing fundamentals: dwell time, text entry rate, errors, and workload. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 3001–3010.

Haofei Wang, Marco Antonelli, and Bertram E Shi. 2017. Using point cloud data to improve three dimensional gaze estimation. In *39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 795–798.

Haofei Wang, Xujiong Dong, Zhaokang Chen, and Bertram E Shi. 2015. Hybrid gaze/EEG brain computer interface for robot arm control on a pick and place task. In *37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 1476–1479.

Erik Wästlund, Kay Sponseller, and Ola Pettersson. 2010. What you see is where you go: testing a gaze-driven power wheelchair for individuals with severe multiple disabilities. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications (ETRA '10)*. ACM, 133–136.

Mingxin Yu, Yingzi Lin, David Schmidt, Xiangzhou Wang, and Yu Wang. 2014. Human-robot interaction based on gaze gestures for the drone teleoperation. *Journal of Eye Movement Research* 7, 4 (2014), 1–14.