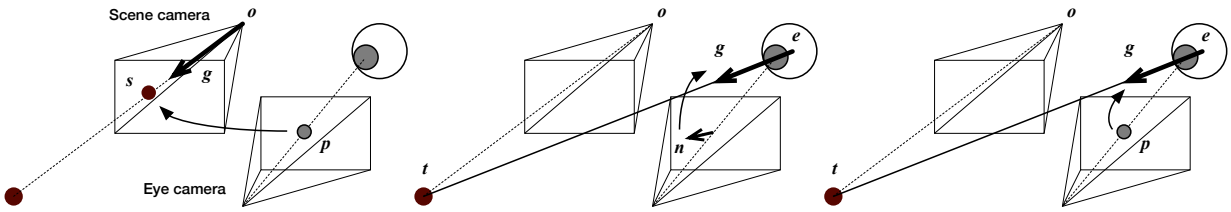
3D Gaze Estimation from 2D Pupil Positions on Monocular Head-Mounted Eye Trackers

Mohsen Mansouryar Julian Steil Yusuke Sugano Andreas Bulling
Perceptual User Interfaces Group
Max Planck Institute for Informatics, Saarbrücken, Germany
{mohsen,jsteil,sugano,bulling}@mpi-inf.mpg.de



(a) *2D-to-2D mapping* **(b)** *3D-to-3D mapping* **(c)** *2D-to-3D mapping*

Figure 1: *Illustration of the (a) 2D-to-2D, (b) 3D-to-3D, and (c) 2D-to-3D mapping approaches. 3D gaze estimation in wearable settings is a task of inferring 3D gaze vectors in the scene camera coordinate system.*

Abstract

3D gaze information is important for scene-centric attention analysis, but accurate estimation and analysis of 3D gaze in real-world environments remains challenging. We present a novel 3D gaze estimation method for monocular head-mounted eye trackers. In contrast to previous work, our method does not aim to infer 3D eyeball poses, but directly maps 2D pupil positions to 3D gaze directions in scene camera coordinate space. We first provide a detailed discussion of the 3D gaze estimation task and summarize different methods, including our own. We then evaluate the performance of different 3D gaze estimation approaches using both simulated and real data. Through experimental validation, we demonstrate the effectiveness of our method in reducing parallax error, and we identify research challenges for the design of 3D calibration procedures.

Keywords: Head-mounted eye tracking; 3D gaze estimation; Parallax error

1 Introduction

Research on head-mounted eye tracking has traditionally focused on estimating gaze in screen coordinate space, e.g. of a public display. Estimating gaze in scene or world coordinates enables gaze analysis on 3D objects and scenes and has the potential for new applications, such as real-world attention analysis [Bulling 2016]. This approach requires two key components: 3D scene reconstruction and 3D gaze estimation.

ETRA '16,, March 14 - 17, 2016, Charlston, SC, USA
ISBN: 978-1-4503-4125-7/16/03...$15.00
DOI: http://dx.doi.org/10.1145/2857491.2857530

In prior work, 3D gaze estimation was approximately addressed as a projection from estimated 2D gaze positions in the scene camera image to the corresponding 3D scene [Munn and Pelz 2008; Takemura et al. 2014; Pfeiffer and Renner 2014]. However, without proper 3D gaze estimation, gaze mapping suffers from parallax error caused by the offset between the scene camera origin and eyeball position [Mardanbegi and Hansen 2012; Duchowski et al. 2014]. To fully utilize the 3D scene information it is essential to estimate 3D gaze vectors in the scene coordinate system.

While 3D gaze estimation has been widely studied in remote gaze estimation, there have been very few studies in head-mounted eye tracking. This is mainly because 3D gaze estimation typically requires model-based approaches with special hardware, such as multiple IR light sources and/or stereo cameras [Beymer and Flickner 2003; Nagamatsu et al. 2010]. Hence, it remains unclear whether 3D gaze estimation can be done properly only with a lightweight head-mounted eye tracker. Świrski and Dodgson proposed a method to recover 3D eyeball poses from a monocular eye camera [Świrski and Dodgson 2013]. While it can be applied to lightweight mobile eye trackers, their method has been only evaluated with synthetic eye images, and its realistic performance including the eye-to-scene camera mapping accuracy has never been quantified.

We present a novel 3D gaze estimation method for monocular head-mounted eye trackers. Contrary to existing approaches, we formulate 3D gaze estimation as a direct mapping task from 2D pupil positions in the eye camera image to 3D gaze directions in the scene camera. Therefore, for the calibration we collect the 2D pupil positions as well as 3D target points, and finally minimize the distance between the 3D targets and the estimated gaze rays.

The contributions of this work are threefold. First, we summarize and analyze different 3D gaze estimation approaches for a head-mounted setup. We discuss potential error sources and technical difficulties in these approaches, and provide clear guidelines for designing lightweight 3D gaze estimation systems. Second, following from this discussion, we propose a novel 3D gaze estimation method. Our method directly maps 2D pupil positions in the eye camera to 3D gaze directions, and does not require 3D observation from the eye camera. Third, we provide a detailed comparison of our method with state-of-the-art methods in terms of 3D gaze esti-

mation accuracy. The open-source simulation environment and the dataset are available at http://mpii.de/3DGazeSim.

2 3D Gaze Estimation

3D gaze estimation is the task of inferring 3D gaze vectors to the target objects in the environment. Gaze vectors in scene camera coordinates can then be intersected with the reconstructed 3D scene. There are three mapping approaches we discuss in this paper: 2D-to-2D, 3D-to-3D, and our novel 2D-to-3D mapping approach. In this section, we briefly summarize three approaches. For more details, please refer to the technical report [Mansouryar et al. 2016].

2D-to-2D Mapping

Standard 2D gaze estimation methods assume 2D pupil positions $\boldsymbol{p}$ in the eye camera images as input. The task is to find the mapping function from $\boldsymbol{p}$ to 2D gaze positions $\boldsymbol{s}$ in the scene camera images (Figure 1 (a)). Given a set of N calibration data items $(\boldsymbol{p}_i, \boldsymbol{s}_i)_{i=1}^{N}$, the mapping function is typically formulated as a polynomial regression. 2D pupil positions are first converted into their polynomial representations $\boldsymbol{q}(\boldsymbol{p})$, and the linear regression weight is obtained via linear regression methods. Following Kassner et al. [Kassner et al. 2014], we did not include cubic terms and used an anisotropic representation as $\boldsymbol{q} = (1, u, v, uv, u^2, v^2, u^2v^2)$ where $\boldsymbol{p} = (u, v)$.

In order to obtain 3D gaze vectors, most of the prior work assumes that the 3D gaze vectors are originating from the origin of the scene camera coordinate system. In this case, estimated 2D gaze positions $\boldsymbol{f}$ can be simply back-projected to 3D vectors $\boldsymbol{g}$ in the scene camera coordinate system. This is equivalent to assuming that the eyeball center position $\boldsymbol{e}$ is exactly the same as the origin $\boldsymbol{o}$ of the scene camera coordinate system. However, in practice there is always an offset between the scene camera origin and the eyeball position, and this offset causes the parallax error.

3D-to-3D Mapping

If we can estimate a 3D pupil pose (unit normal vector of the pupil disc) from the eye camera as done in Świrski and Dodgson [Świrski and Dodgson 2013], we can instead take a direct 3D-to-3D mapping approach (Figure 1 (b)). Instead of the 2D calibration targets $\boldsymbol{s}$, we assume 3D calibration targets $\boldsymbol{t}$ in this case.

With the calibration data $(\boldsymbol{n}_i, \boldsymbol{t}_i)_{i=1}^{N}$, the task is to find the rotation $\boldsymbol{R}$ and translation $\boldsymbol{T}$ between the scene and eye camera coordinate systems. This can be done by minimizing distances between 3D gaze targets $\boldsymbol{t}_i$ and the 3D gaze rays which are rotated and translated to the scene camera coordinate system. In the implementation, we further parameterize the rotation $\boldsymbol{R}$ by a 3D angle vector with the constraint that rotation angles are between $-\pi$ and π, and we initialize $\boldsymbol{R}$ assuming that the eye camera and the scene camera are facing opposite directions.

2D-to-3D Mapping

Estimating 3D pupil pose is not a trivial task in real-world settings. Another potential approach is to directly map 2D pupil positions $\boldsymbol{p}$ to 3D gaze directions $\boldsymbol{g}$ (Figure 1 (c)).

In this case, we need to map the polynomial feature $\boldsymbol{q}$ to unit gaze vectors $\boldsymbol{g}$ originating from an eyeball center $\boldsymbol{e}$. $\boldsymbol{g}$ can be parameterized in a polar coordinate system, and we assume a linear mapping from the polynomial feature $\boldsymbol{q}$ to the angle vector. The regression weight is obtained by minimizing distances between 3D calibration targets $\boldsymbol{t}_i$ and the mapped 3D gaze rays as in the 3D-to-3D approach.

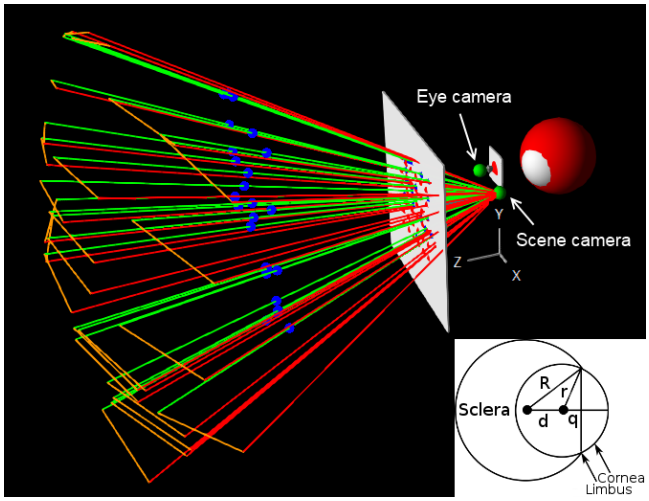


Figure 2: *3D eye model and simulation environment with 3D target points given as blue dots. The green and red rays correspond to ground truth and estimated gaze vectors, respectively.*

In the implementation, we used the same polynomial representation as the 2D-to-2D method to provide a fair comparison with the baseline.

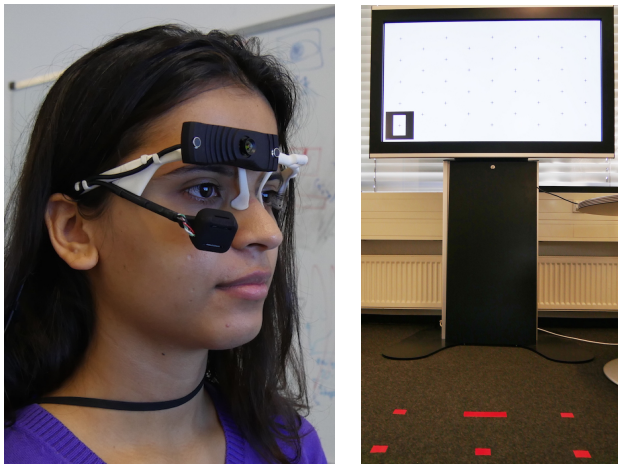
3 Data Collection

In order to evaluate the potential and limitations of the introduced mapping approaches, we conducted two studies. First, we used data we obtained from a simulation environment, whereas the second study exploited real-world data collected from 14 participants.

Simulation Data

We first analyzed the different mapping approaches in a simulation environment. Our simulation environment is based on a basic model of the human eye consisting of a pair of spheres [Lefohn et al. 2003] and the scene and eye camera models. The eye model and a screenshot of the simulation environment are illustrated in Figure 2. We used human average anatomical parameters: $R = 11.5mm$, $r = 7.8mm$, $d = 4.7mm$, and $q = 5.8mm$. The pupil is considered as the center of the circle which represents the intersection of the two spheres. For both eye and scene cameras, we used the pinhole camera model. Intrinsic parameters were set to values similar to those of the actual eye tracking headset we used in the real-world environment.

One of the key questions about 3D gaze estimation is whether calibration at single depth is sufficient or not. Intuitively, obtaining calibration data at different depths from the scene camera can improve the 3D mapping performance. We set calibration and test plane depths d_c and d_t to 1m, 1.25m, 1.5m, 1.75m, and 2m. At each depth, points are selected from two grids, a 5 by 5 grid which gives us 25 calibration points (blue) and an inner 4 by 4 grid for 16 test points (red) displayed on the white plane of Figure 2. Both of the grids are symmetric with respect to the scene camera's principal axis. From the eye model used, we are able to estimate the corresponding gaze ray.



(a) *Video-based head-mounted eye tracker* **(b)** *Display and distance markers*

Figure 3: *The recording setup consisted of a Lenovo G580 laptop, a Phex Wall 55" display and a PUPIL head-mounted eye tracker.*

Real-World Data

We also present evaluation of gaze estimation approaches using a real-world dataset to show the validity of 3D gaze estimation approaches.

Procedure

The recording system consisted of a Lenovo G580 laptop and a Phex Wall 55" display (121.5cm $\times$ 68.7cm) with a resolution of 1920 $\times$ 1080. Gaze data was collected using a PUPIL head-mounted eye tracker connected to the laptop via USB [Kassner et al. 2014] (see Figure 3a). The eye tracker has two cameras: one eye camera with a resolution of 640 $\times$ 360 pixels recording a video of the right eye from close proximity, as well as an egocentric (scene) camera with a resolution of 1280 $\times$ 720 pixels. Both cameras recorded videos at 30 Hz. Pupil positions in the eye camera were detected using the PUPIL eye tracker's implementation.

We implemented remote recording software which conducts the calibration and test recordings shown on the display to the participants. As shown in Figure 3b, the target markers were designed so that their 3D positions can be obtained using the ArUco library [Garrido-Jurado et al. 2014]. Intrinsic parameters of the scene and eye cameras were calibrated before recording, and used for computing 3D fixation target positions $\boldsymbol{t}$ and 3D pupil poses $\boldsymbol{n}$.

We recruited 14 participants aged between 22 and 29 years. The majority of them had little or no previous experience with eye tracking. Every participant had to perform two recordings, a calibration and a test recording of five different distances from the display. Recording distances were marked by red stripes on the ground (see Figure 3b). They were aligned parallel to the display with an initial distance of 1 meter and the following recording distances with a spacing of 25cm (1.0, 1.25, 1.5, 1.75, 2.0). For every participant we recorded 10 videos.

As in the simulation environment, the participants were instructed to look at 25 fixation target points from the grid pattern in Figure 3b. After this step the participants had to perform the same procedure again while looking at 16 fixation targets placed on different positions than in the initial calibration to collect the test data for our evaluation part. This procedure was then repeated for the other four mentioned distances. The only restriction we imposed was that the participants should not move their head during the recording.

Error Measurement

Since the ground-truth eyeball position $\boldsymbol{e}$ is not available in the real-world study, we evaluate the estimation accuracy using an angular error observed from the scene camera. For the case where 2D gaze positions are estimated (2D-to-2D mapping), we back-projected the estimated 2D gaze position $\boldsymbol{f}$ into the scene, and directly measured the angle θ between this line and the line from the origin of the scene camera $\boldsymbol{o}$ to the measured fixation target $\boldsymbol{t}$. For the cases where 3D gaze vectors are estimated, we first determined the estimated 3D fixation target position $\boldsymbol{t}'$ assuming the same depth as the ground-truth target $\boldsymbol{t}$. Then the angle between the lines from the origin $\boldsymbol{o}$ was measured.

4 Results

We compared different mapping approaches in Figure 4 using an increasing number of calibration depths in both simulation and real-world environments. Each plot corresponds to mean estimation errors of all test planes and all combinations of calibration planes. Angular error is evaluated from the ground-truth eyeball position. It can be seen that in all cases the estimation performance can be improved by taking more calibration planes. Even the 2D-to-2D mapping approach performs slightly better with multiple calibration depths overall in both environments. The 2D-to-3D mapping approach performed better than the 2D-to-2D mapping in all cases in the simulation environment. For the 3D-to-3D mapping approach a parallax error near to zero can be achieved.

Similarly to the simulation case, we first compare the 2D-to-3D mapping with the 2D-to-2D mapping in terms of the influence of different calibration depths displayed as stable lines in Figure 4. Since it turned out that the 3D-to-3D mapping on real-world data has more angular error (over 10°) than the 2D-to-3D mapping, we omit the results in the following analysis.

Contrary to the simulation result, with a lower number of calibration depths the 2D-to-2D approach performs better than the 2D-to-3D approach for real-world data. However, with an increasing number of calibration depths, the 2D-to-3D approach outperforms 2D-to-2D comparing the angular error in visual degrees. For five calibration depths we can achieve for the 2D-to-3D case an overall mean of less than 1.3 visual degrees over all test depths and all participants. A more detailed analysis and discussion with corresponding performance plots are available in the technical report [Mansouryar et al. 2016].

5 Discussion

We discussed three different approaches for 3D gaze estimation using head-mounted eye trackers. Although it was shown that the 3D-to-3D mapping is not a trivial task, the 2D-to-3D mapping approach was shown to perform better than the standard 2D-to-2D mapping approach using simulation data. One of the key observations from the simulation study is that the 2D-to-3D mapping approach requires at least two calibration depths. Given more than two calibration depths, the 2D-to-3D mapping can significantly reduce the parallax error.

On the real data, we could observe a decreasing error for the 2D-to-3D mapping with an increasing number of calibration depths, and could outperform the 2D-to-2D mapping. However, the performance of the 2D-to-3D mapping became worse than in the simulation environment. Reasons for the different performance of the

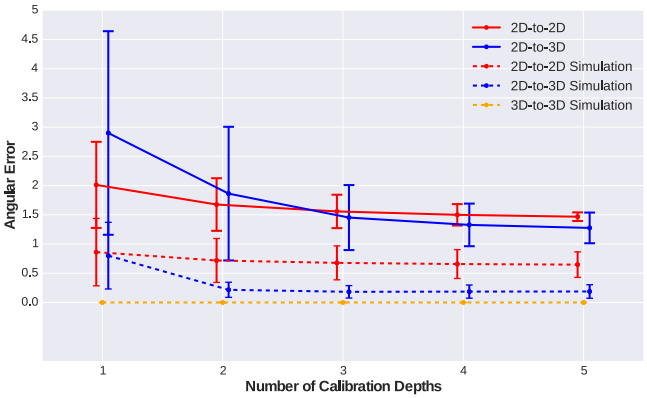


Figure 4: *Angular error performance over different numbers of calibration depths for 2D-to-2D, 2D-to-3D and 3D-to-3D mapping approaches. Each point corresponds to the mean over all angular error values for each number of calibration depths. The error bars provide the corresponding standard deviations.*

mapping approaches in the simulation and real-world environment are manifold and reveal their limitations. Our simulation environment considers an ideal setting and does not include noise that occurs in the real world. This noise is mainly produced by potential errors in the pupil and marker detection, as well as head movements of the participants.

In future work it will be important to investigate how the 3D-to-3D mapping approach can work in practice. The fundamental difference from the 2D-to-3D mapping is that the mapping function has to explicitly handle the rotation between eye and scene camera coordinate systems. In addition to the fundamental estimation inaccuracy of the 3D pupil pose estimation technique without modeling real-world factors such as corneal refraction, we did not consider the difference between optical and visual axes. A more appropriate mapping function could be a potential solution for the 3D-to-3D mapping, and another option could be to use more general regression techniques considering the 2D-to-3D results.

Throughout the experimental validation, this research also illustrated the fundamental difficulty of the 3D gaze estimation task. It has been shown that the design of the calibration procedure is also quite important, and it is essential to address the issue from the standpoint of both calibration design and mapping formulation. Since the importance of different calibration depths has been shown, the design of automatic calibration procedure, e.g., how to obtain calibration data at different depths using only digital displays, is another important HCI research issue.

Finally, it is also important to combine the 3D gaze estimation approach with 3D scene reconstruction methods and evaluate the overall performance of 3D gaze mapping. In this sense, it is also necessary to evaluate performance with respect to scene reconstruction error.

6 Conclusion

In this work, we provided an extensive discussion on different approaches for 3D gaze estimation using head-mounted eye trackers. In addition to the standard 2D-to-2D mapping approach, we discussed two potential 3D mapping approaches using either 3D or 2D observation from the eye camera. We conducted a detailed analysis of 3D gaze estimation approaches using both simulation and real data.

Experimental results showed the advantage of the proposed 2D-to-3D estimation methods, but its complexity and technical challenges were also revealed. Together with the dataset and simulation environment, this study would provide a solid basis for future research on 3D gaze estimation with lightweight head-mounted devices.

Acknowledgements

We would like to thank all participants for their help with the data collection. This work was funded, in part, by the Cluster of Excellence on Multimodal Computing and Interaction (MMCI) at Saarland University, the Alexander von Humboldt Foundation, and a JST CREST research grant.

References

BEYMER, D., AND FLICKNER, M. 2003. Eye gaze tracking using an active stereo head. In *Proc. CVPR*.

BULLING, A. 2016. Pervasive attentive user interfaces. *IEEE Computer 49*, 1, 94–98.

DUCHOWSKI, A. T., HOUSE, D. H., GESTRING, J., CONGDON, R., ŚWIRSKI, L., DODGSON, N. A., KREJTZ, K., AND KREJTZ, I. 2014. Comparing estimated gaze depth in virtual and physical environments. In *Proc. ETRA*, 103–110.

GARRIDO-JURADO, S., MUÑOZ-SALINAS, R., MADRID-CUEVAS, F. J., AND MARÍN-JIMÉNEZ, M. J. 2014. Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognition 47*, 6, 2280–2292.

KASSNER, M., PATERA, W., AND BULLING, A. 2014. Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In *Adj. Proc. UbiComp*, 1151–1160.

LEFOHN, A., BUDGE, B., SHIRLEY, P., CARUSO, R., AND REINHARD, E. 2003. An ocularist's approach to human iris synthesis. *IEEE Trans. Computer Graphics and Applications 23*, 6, 70–75.

MANSOURYAR, M., STEIL, J., SUGANO, Y., AND BULLING, A. 2016. 3d gaze estimation from 2d pupil positions on monocular head-mounted eye trackers. *arXiv:1601.02644*.

MARDANBEGI, D., AND HANSEN, D. W. 2012. Parallax error in the monocular head-mounted eye trackers. In *Proc. UbiComp*, ACM, 689–694.

MUNN, S. M., AND PELZ, J. B. 2008. 3d point-of-regard, position and head orientation from a portable monocular video-based eye tracker. In *Proc. ETRA*, ACM, 181–188.

NAGAMATSU, T., SUGANO, R., IWAMOTO, Y., KAMAHARA, J., AND TANAKA, N. 2010. User-calibration-free gaze tracking with estimation of the horizontal angles between the visual and the optical axes of both eyes. In *Proc. ETRA*, ACM, 251–254.

PFEIFFER, T., AND RENNER, P. 2014. Eyesee3d: A low-cost approach for analyzing mobile 3d eye tracking data using computer vision and augmented reality technology. In *Proc. ETRA*, ACM, 369–376.

ŚWIRSKI, L., AND DODGSON, N. A. 2013. A fully-automatic, temporal approach to single camera, glint-free 3d eye model fitting [abstract]. In *Proc. PETMEI*.

TAKEMURA, K., TAKAHASHI, K., TAKAMATSU, J., AND OGASAWARA, T. 2014. Estimating 3-d point-of-regard in a real environment using a head-mounted eye-tracking system. *IEEE Trans. on Human-Machine Systems 44*, 4, 531–536.