Accuracy of Monocular Gaze Tracking on 3D Geometry

Xi Wang, David Lindlbauer, Christian Lessig and Marc Alexa

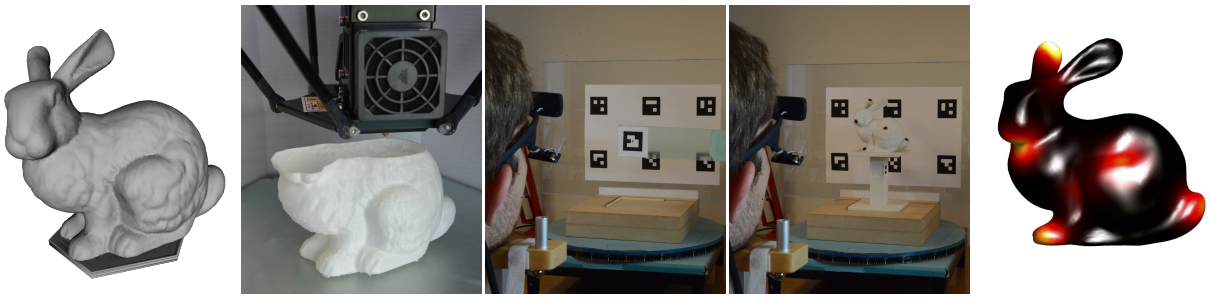

Fig. 1: We accurately estimate 3D gaze positions by combining digital manufacturing, marker tracking and monocular eye tracking. With a simple calibration procedure we attain an angular accuracy of 0.8°.

Abstract— Many applications in visualization benefit from accurate knowledge of where a person is looking at. We present a system for accurately tracking gaze positions on a three dimensional object using a monocular head mounted eye tracker. We accomplish this by 1) using digital manufacturing to create stimuli with accurately known geometry, 2) embedding fiducial markers directly into the manufactured objects to reliably estimate the rigid transformation of the object, and, 3) using a perspective model to relate pupil positions to 3D locations. This combination enables the efficient and accurate computation of gaze position on an object from measured pupil positions. We validate the accuracy of our system experimentally, achieving an angular resolution of 0.8° and a 1.5% depth error using a simple calibration procedure with 11 points.

Index Terms—eye tracking, calibration, accuracy

1 Introduction

Understanding the viewing behaviour of humans when they look at objects plays an important role in applications such as data visualization, scene analysis, object recognition, and image generation [30]. The viewing behaviour can be analyzed by measuring fixations using eye tracking. In the past, such experiments, especially for object exploration tasks, were performed with flat 2D stimuli presented on a screen [13]. Since the human visual attention mechanism has developed in 3D environments, depth may have an important effect on the viewing behaviour [20]. To understand the role of depth information, some studies [16, 21, 9] recently employed stereoscopic displays. However, these displays fail to provide natural depth cues; for example they suffer from stereoscopic decoupling, the mismatch of accommodation and vergence with the displayed depth [14]. Since our research objective is to investigate the viewing behavior of humans for stimuli that are genuinely three-dimensional, we need to be able to track 3D gaze positions with high accuracy.

Standard eye tracking setups only determine the viewing direction. The most common approach for determining viewing depth is to employ a binocular eye tracker and measure eye vergence, that is the orientation difference between the left and the right eye that ensures both are focused on the same point in space. However, as exemplified in Fig. 2, experimentally determining depth from binocular vergence is inherently ill-conditioned, since even for an object at a modest distance the eyes and the object form a highly acute triangle so that the inevitable inaccuracies in measuring pupil positions [13] lead to large errors in the estimated depth values. Although nonlinear mappings can be employed to reduce the error [8, 23, 1, 12, 19, 22, 25], these require more complex calibration and training while still leading to relatively large inaccuracies.

We base our approach on the idea of relating the viewing direction gathered by an eye tracker to the physical world. This is done similar to EyeSee3D [26] by tracking fiducial markers in physical space with a camera mounted on the eye tracker. Our goal in this setup is to understand if this setup can be made accurate enough to enable trakcing visual attention on three-dimensional objects. The main ingredients to achieve accurate tracking are:

1. 3D stimuli are generated by digital manufacturing so that their geometry is known to high accuracy and also available in digital form without imposing restrictions on the geometry that is represented.
2. Fiducial markers are integrated into the 3D stimuli in order to reliably and accurately estimate the stimuli's 3D position relative to the head.
3. A careful calibration allows accurately computating the perspective mapping from 3D positions to monocular pupil positions.
4. An error model for the mapping allows computing plausible positions on the 3D stimulus.

Our results demonstrate that for typical geometries we are able to obtain reliable depth values within 1.5% range and 0.8° angular resolution, including around silhouettes where the geometry has a large slope. We accomplish this with only a monocular eye tracker and an 11 points calibration procedure.

In the next section, we discuss related work on 3D gaze tracking. Subsequently, in Sec. 3, we detail our setup and explain how 3D positions can be related to pupil coordinates. This is followed by a discussion of how 3D viewing positions can be obtained from pupil positions in Sec. 4. Experimental results verifying the accuracy of our approach are presented in Sec. 5. We conclude the paper with a discussion of directions for future work in Sec. 6.

Xi Wang, David Lindlbauer, Christian Lessig and Marc Alexa are with Technical University of Berlin, E-mail: {xi.wang | david.lindlbauer | christian.lessig | marc.alexa@tu-berlin.de}

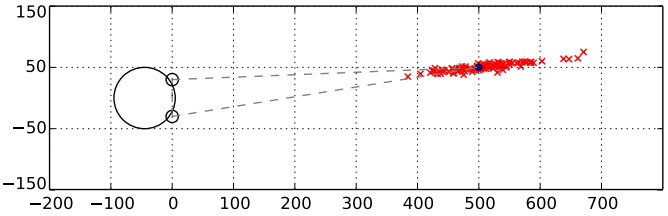

Fig. 2: Inherent error of vergence based depth estimation for an object at a distance of 500 mm away from the eyes. The red crosses mark estimated 3D positions for normally distributed gaze directions with mean equal to the correct angle for the object (black dot) and a variance of 0.5°. The highly acute triangle that leads to the ill-conditioning of the depth calculation is shown as dashed lines. The worst case relative error is almost 50%.

2 Related Work

The viewing behaviour of humans is typically analyzed using eye tracking by measuring a subject's fixations. However, usually only flat 2D stimuli on a screen are employed, e.g. [5, 17, 24, 27], even when one is interested in 3D objects. Only recently the first studies considering the effect of depth were performed. Lang et al. [21] collected a large eye fixation database for still images with depth information presented on a stereoscopic display. Their results show that depth can have a significant influence on a subject's fixations. Jansen et al. [16] also employed a stereoscopic display to analyze the effect of depth, demonstrating that depth information leads to an overall increase in spatial exploration. Both Lang et al. [21] and Jansen et al. [16] also report that visual attention shifts over time from objects closer to the viewer to those farther away. Differences in fixations between 2D and 3D stimuli were recently investigated for stereoscopic video [9, 10, 15, 28]. For these stimuli, discrepancies were mainly observed for scenes that lack on obvious (high-level) center of attention, with fixations having a larger spatial distribution when depth information is present.

Existing work investigating the role of depth information on fixation locations hence demonstrates that, at least under certain circumstances, depth has a significant effect on a subject's viewing behaviour. However, so far only stereoscopic displays were employed, which do not provide all depth cues and suffer from stereoscopic decoupling [14]. Moreover, Duchowski et al. [7] showed that for stereoscopic displays the gaze depth of subjects does not fully correspond to the presented depth. Therefore, we believe that to understand real-world viewing behaviour for 3D objects one should study stimuli that are genuinely three-dimensional. This provides the principal motivation for our work.

With 3D stimuli also the depth values of fixation points have to be determined. The most common approach for obtaining fixation depth is to measure the vergence using a binocular eye tracker. However, computing depth values from binocular vergence is ill-conditioned since already for modest distances minuscule measurement errors in the pupil positions lead to large depth errors, cf. Fig. 2. To improve the accuracy, Essig et al. [8] trained a neural network that maps from eye vergence to depth values. Maggia, Guyader and Guérin-Dugué [23] proposed a somewhat simpler but also nonlinear model for the mapping from measured disparity to depth. Building on these works, current techniques [1, 12, 19, 22, 25] that employ binocular vergence to determine fixation depth obtain an error that is within 10% of the correct value.

The approaches that inspired our work are taking an alternative approach by relating the view direction with the *known* geometry of physical reality. This can be conveniently in virtual reality [29, 6]. Pfeiffer and Renner have used fiducial markers to align physical world to camera space [26]. By using vergence of the eyes, they have achieved an angular accuracy of 2.25 degrees, which gives correctly classified fixation targets on the scale of whole objects. However, for investigating human viewing behaviour on the surface of 3D objects, more accurate gaze tracking is required. Consequently, we try to adjust the setup with the goal of accurate tracking of visual attention on 3D objects in mind.

3 From 3D positions to pupil coordinates

In this section we describe how points in space can be related to 2D pupil positions corresponding to the gaze directions, which can be related to the points in space a person is looking at Fig. 3. We assume a setup using a monocular head mounted eye tracking device with a front facing world camera capturing the environment and an eye facing camera capturing the pupil movement. The world camera is employed to compute the position and orientation of fiducial markers, for example fixed to objects, relative to the subject's head. A projective mapping is then used to relate these 3D coordinates to pupil positions. The mapping is calibrated by having the subject focus on markers at different locations, including varying depths. In the following we will describe these steps in more detail.

3.1 From local 3D positions to world-camera coordinates

We employ fiducial markers to determine the 3D coordinates of locations in space in the world camera coordinate system. The mapping of a position $\mathbf{x} \in \mathbb{R}^3$, for example a point on a marker, to its projection $\mathbf{m}$ in the world camera image is given by

$$\mathbf{m} = \mathbf{K}(\mathbf{R}\mathbf{x} + \mathbf{t}), \quad \mathbf{R}^\mathsf{T}\mathbf{R} = \mathbf{I} \tag{1}$$

where $\mathbf{K} : \mathbb{R}^3 \to \mathbb{R}^2$ is the intrinsic world camera matrix, modelling the perspective mapping, and $\mathbf{R}$ and $\mathbf{t}$ are the rotation and translation of the camera forming the rigid transformation. The mapping of $\mathbf{x}$ to its representation $\mathbf{w}$ in the world camera coordinate system is hence

$$\mathbf{w} = \mathbf{R}\mathbf{x} + \mathbf{t}. \tag{2}$$

We determine the intrinsic world camera matrix $\mathbf{K}$, which includes both radial and tangential distortion, in a preprocessing step using the approach proposed by Heng et al. [11]. To determine the rigid transformation given by $\mathbf{R}$ and $\mathbf{t}$ we exploit that detected marker corner points $\mathbf{m}_i \in \mathbb{R}^2$ in the camera image have known 3D locations $\mathbf{x}_i \in \mathbb{R}^3$ in the marker's local coordinate system. Given at least three such points $\mathbf{m}_i$ in the camera image, we can determine $\mathbf{R}$ and $\mathbf{t}$ by minimizing the reprojection error.

Once $\mathbf{R}$ and $\mathbf{t}$ have been estimated, we can employ Eq. 2 to determine the position of the center of the marker in the world camera coordinate system, as required for calibration, or to map an object with a fixed relative position to a marker into the space, as is needed to determine gaze positions.

3.2 From world camera coordinates to pupil positions

Given positions $\mathbf{w} \in \mathbb{R}^3$ in the world camera coordinate system, obtained as described in the last section, we have to relate these to a person's gaze direction, described by pupil positions $\mathbf{p}$ in the eye camera image. We model the mapping as a projective transformation, because the cameras and the system of the eye (i.e. the head) are in fixed relative orientation and position. In homogeneous coordinates we hence have

$$s \begin{pmatrix} \mathbf{p} \\ 1 \end{pmatrix} = \mathbf{Q} \begin{pmatrix} \mathbf{w} \\ 1 \end{pmatrix} \tag{3}$$

where $\mathbf{Q} \in \mathbb{R}^{3\times4}$ is a projection matrix that is unique up to scale. Given a set of correspondences $\{(\mathbf{w}_i, \mathbf{p}_i)\}$ between 3D points $\mathbf{w}_i$ in the world camera coordinate system and pupil positions $\mathbf{p}_i$ describing a gaze direction at $\mathbf{w}_i$, we can determine $\mathbf{Q}$ by minimizing

$$E(\mathbf{Q}) = \sum_i \left\| s_i \begin{pmatrix} \mathbf{p}_i \\ 1 \end{pmatrix} - \mathbf{Q} \begin{pmatrix} \mathbf{w}_i \\ 1 \end{pmatrix} \right\|_2^2. \tag{4}$$

Fixing one coefficient of $\mathbf{Q}$ to eliminate the freedom on scale (we choose $\mathbf{Q}_{3,4} = 1$), this is a standard linear least squares problem. In practice, we solve this problem using correspondences $\{(\mathbf{w}_i, \mathbf{p}_i)\}$ obtained during calibration, as described in Sec. 5.

Since $\mathbf{Q}$ is a projective transformation we can factor it into an upper triangular intrinsic camera matrix $\mathbf{A}_Q$ and a rigid transformation matrix $\mathbf{T}_Q = (\mathbf{R}_Q, \mathbf{t}_Q)$. The factorization is given by

$$\mathbf{Q} = \mathbf{A}_Q\mathbf{T}_Q = (\mathbf{A}_Q\mathbf{R}_Q, \mathbf{A}_Q\mathbf{t}_Q) \tag{5}$$

and hence by the RQ decomposition of the left 3×3 block $\mathbf{A}_Q\mathbf{R}_Q$ of $\mathbf{Q}$. It can be computed using the QR decomposition as

$$\mathbf{J}(\mathbf{A}_Q\mathbf{R}_Q)^T\mathbf{J} = (\mathbf{J}\mathbf{A}_Q^T\mathbf{J})(\mathbf{J}\mathbf{R}_Q^T\mathbf{J}) \tag{6}$$

where $\mathbf{J}$ is the exchange matrix, which in our case is the column inversed version of the identity matrix.

3.3 From pupil positions to angular accuracy

So far we related 3D locations to pupil positions. To determine a gaze point on an object we also have to relate pupil positions to a cone of positions in space. This also corresponds to the angular accuracy of our setup.

With the intrinsic eye camera matrix $\mathbf{A}_Q$, as determined in the last section, we can relate a homogeneous pupil position $\hat{\mathbf{p}} = (\mathbf{p}, 1)^T$ to an associated ray $\mathbf{r}$ in 3D world camera space:

$$\hat{\mathbf{p}} = \mathbf{A}_Q\mathbf{r}; \tag{7}$$

the depth along $\mathbf{r}$ is indeterminate since $\mathbf{A}_Q$ is a projection matrix. The angle between two rays $\mathbf{r}_i, \mathbf{r}_j$, represented by pupil coordinates $\mathbf{p}_i, \mathbf{p}_j$, is hence given by

$$\cos\eta_{ij} = \frac{\mathbf{r}_i^T\mathbf{r}_j}{\|\mathbf{r}_i\|\|\mathbf{r}_j\|} = \frac{\hat{\mathbf{p}}_i^\mathsf{T}\mathbf{A}_Q^{-\mathsf{T}}\mathbf{A}_Q^{-1}\hat{\mathbf{p}}_j}{\|\mathbf{A}_Q^{-1}\hat{\mathbf{p}}_i\|\|\mathbf{A}_Q^{-1}\hat{\mathbf{p}}_j\|}. \tag{8}$$

This suggests to interpret the matrix $\mathbf{A}_Q^{-\mathsf{T}}\mathbf{A}_Q^{-1}$ as an induced inner product $\mathbf{M}_Q = (\mathbf{A}_Q\mathbf{A}_Q^\mathsf{T})^{-1}$ on homogeneous pupil coordinates. The angle η_{ij} then becomes

$$\cos\eta_{ij} = \frac{\hat{\mathbf{p}}_i^\mathsf{T}\mathbf{M}_Q\hat{\mathbf{p}}_j}{(\hat{\mathbf{p}}_i^\mathsf{T}\mathbf{M}_Q\hat{\mathbf{p}}_i)^{1/2}(\hat{\mathbf{p}}_j^\mathsf{T}\mathbf{M}_Q\hat{\mathbf{p}}_j)^{1/2}}. \tag{9}$$

For multiple pairs $\mathbf{p}_i$, $\mathbf{p}_j$, Eq. 9 can be solved efficiently when the involved matrices are precomputed.

4 From pupil coordinates to locations on an object

Our objective is to determine a gaze position $\bar{\mathbf{w}} \in \mathbb{R}^3$ in space from a pupil position $\bar{\mathbf{p}}$ describing a gaze direction. Central to our approach for determining $\bar{\mathbf{w}}$ is that the geometry of the observed object is known to high accuracy. This is ensured by 3D printing the object $\mathcal{M}$ from its digital representation as a triangulated surface M. The printed object also includes a fiducial marker, which allows us to determine the rigid transformation of the object in space as described in Sec. 3.1.

To simplify the problem, we do not determine the exact gaze location $\bar{\mathbf{w}}$ on the object corresponding to $\bar{\mathbf{p}}$ but instead consider the vertices $\bar{\mathbf{v}}_i$ in the digital model M that map to pupil positions close to $\bar{\mathbf{p}}$. Let

$$\hat{\mathbf{p}}_i = \mathbf{Q}(\mathbf{R}\mathbf{v}_i + \mathbf{t}) \tag{10}$$

be the homogeneous pupil position $\mathbf{p}_i = (p_{i_1}, p_{i_2}, p_{i_3})^\mathsf{T}$ corresponding to vertex $\mathbf{v}_i$. Then we consider the set of vertices

$$\Gamma_c(\bar{\mathbf{p}}) = \left\{ \mathbf{v}_i \in M \;\middle|\; \frac{\hat{\bar{\mathbf{p}}}^\mathsf{T}\mathbf{M}_Q\hat{\mathbf{p}}_i}{(\hat{\bar{\mathbf{p}}}^\mathsf{T}\mathbf{M}_Q\hat{\bar{\mathbf{p}}})^{1/2}(\hat{\mathbf{p}}_i^\mathsf{T}\mathbf{M}_Q\hat{\mathbf{p}}_i)^{1/2}} > c \right\}; \tag{11}$$

that is, we are determining which vertices $\mathbf{v}_i$ on the object lie within a cone centered around the eye ray corresponding to $\bar{\mathbf{p}}$ with angular size c. From these vertices, we consider the one closest to the eye as the intersection point. This point can be determined efficiently solely using p_{i_3}. Note that since the metric $\mathbf{M}_Q$ has a natural relation to eye ray angle, we can choose c based on the accuracy of our measurements. Space partitioning data structures can be used to speed up the search, however, this is not an issue in our current implementation.

5 Experiments

In the following, we will report on preliminary experimental results that validate the accuracy of our setup for tracking 3D gaze points and that demonstrate that a small number of correspondences suffices for calibration. These results were obtained using two exploratory experiments with a small number of subjects ($n = 6$).

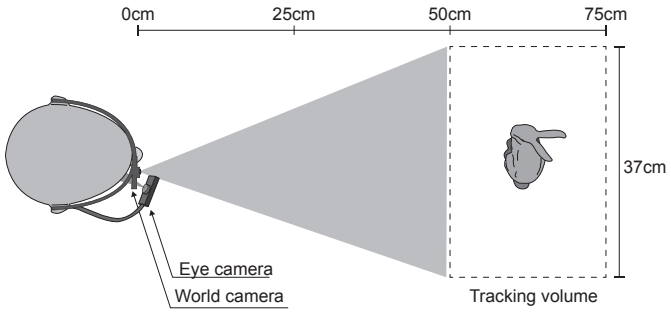

Fig. 3: Physical setup used in our experiments.

Participants We recruited 6 unpaid participants (all male), all of which were students or staff from a university. Their age ranged from 26 to 39 years and all had normal or corrected-to-normal vision, based on self-reports. Four of them had some previous experience with eye tracking.

Apparatus The physical setup of our experiments is shown in Fig. 3. For measuring fixations we employed the Pupil eye tracker [18]. Our software implementation uses OpenCV [4], which was in particular employed to solve for the rigid transformations $\mathbf{R}$,$\mathbf{t}$ as described in Sec. 3.1. We determine $\mathbf{Q}$ using Eq. 4 with the Ceres Solver [2].

5.1 Accuracy of calibration and gaze direction estimation

In Sec. 3.2 we explained how the projective mapping $\mathbf{Q}$ from world camera coordinates to pupil positions can be determined by solving a linear least squares problem. As input to the problem one requires correspondences $\{(\mathbf{w}_i, \mathbf{p}_i)\}$ between world camera coordinates $\mathbf{w}_i$ and pupil positions $\mathbf{p}_i$ describing the gaze direction of a particular subject towards $\mathbf{w}_i$. The correspondences have to be determined experimentally, and hence will be noisy. The accuracy with which $\mathbf{Q}$ is determined therefore depends on the number of correspondences that is used. In our first experiment we investigated how many correspondences are needed to obtain a robust estimate for $\mathbf{Q}$. The same data also allows us to determine the angular error of our setup.

Procedure We obtained correspondences $\{(\mathbf{w}_i, \mathbf{p}_i)\}$ by asking a subject to focus on the center of a single fiducial marker while it is presented at various locations in the desired view volume (see Fig. 1, middle); we have augmented the center of the marker with a red dot to make this task as unambiguous as possible. At each position of the marker, we estimate a single correspondence $(\mathbf{w}_i, \mathbf{p}_i)$ based on the estimation of the rigid transformation for the marker, cf. Sec. 3.1. For each participant, we recorded 100 correspondences $\{(\mathbf{w}_i, \mathbf{p}_i)\}$ for two different conditions, resulting in a total of 200 measurements per participant. In the first condition the head was fixed on a chin rest while in the second condition participants were only asked to keep facing towards the marker. For both condition the marker was moved in a volume of 0.37m (width) $\times$ 0.4m (height) $\times$ 0.25m (depth) at a distance of 0.75m from the subject (see Fig. 3).

Data Analysis For each dataset we perform 10 trials of 2-fold cross validation and estimate the projection matrix using $\{4, 5, 7, 9, 11, 12, 13, 14, 16, 18, 20, 25, 50\}$ point pairs. In each trial, the 100 correspondences are randomly divide into 2 bins of 50 point pairs each. One bin is used as training set and the other as testing set. Point pair correspondences from the training set are used to compute the projection matrix $\mathbf{Q}$ which is then employed to compute the error between the gaze direction given by the pupil position $\mathbf{p}_i$ and the true direction given by the marker center $\mathbf{w}_i$ for the points in the test data set. From Eq. 9 this error can be calculated as

$$\eta_i = \cos^{-1}\frac{\mathbf{p}_i^\mathsf{T}\mathbf{M}_Q\mathbf{Q}\mathbf{w}_i}{(\mathbf{p}_i^\mathsf{T}\mathbf{M}_Q\mathbf{p}_i)^{1/2}(\mathbf{w}_i^\mathsf{T}\mathbf{Q}^\mathsf{T}\mathbf{M}_Q\mathbf{Q}\mathbf{w}_i)^{1/2}}. \tag{12}$$

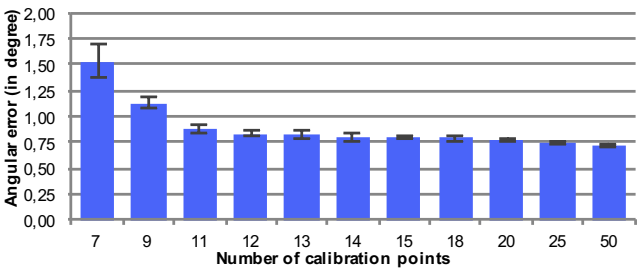

Fig. 4: Mean values and standard errors for angular error with respect to the number of calibration points. Note that 4 and 5 calibration points were omitted in the figure since they result in a more than 10 times higher angular error.

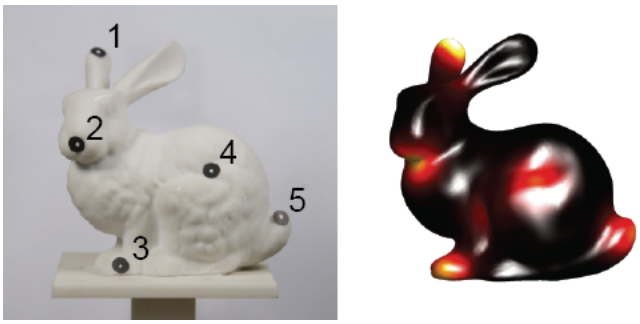

Fig. 5: *Left*: physical bunny model with target markers (numbers indicate order); *right*: heat map of obtained gaze directions.

Results In order to analyze the influence of the number of calibration points as well as the usage of the chin rest on the estimation accuracy, we performed a repeated measures ANOVA ($\alpha = .05$) on the independent variable *Chin rest* with 2 levels (with, without) and *Calibration* with 13 levels (the corresponding number of calibration points, i.e., 4, 5, 7, 9, 11, 12, 13, 14, 16, 18, 20, 25, 50). The dependent variable was the angular error in degree. We used 10 rounds of cross validation for our repeated measures, with each data point being the average angular error per round. This resulted in an overall of 260 data points per participant (2 *Chin reset* $\times$ 13 *Calibration* $\times$ 10 cross validation).

Results showed a main effect for *Calibration* ($F_{12,60} = 103.064$, $p < .001$). Post-hoc pairwise comparisons revealed that the difference between using 4 calibration points ($M = 17.38, SE = 0.9$), the minimum number required to determine the 11 parameters of the matrix $\mathbf{Q}$, and all other conditions was significantly different ($p < .05$). Comparing 5 points ($M = 12.69, SE = 1.69$) to all subsequent conditions showed significance levels of $p \approx .07$. Furthermore, using 9 calibration points compared to 20, 25, and 50 showed significantly different angular errors ($p < .05$). Mean values and standard errors are illustrated in Figure 4.

When using 11 to 50 calibrations points, the angular error averages at around 0.8, which is within the range of human visual accuracy and goes in line with the specifications of the pupil eye tracker for 2D gaze estimation [18, 3]. The results furthermore demonstrate that even for a relatively low number of calibration points, comparable to the 9 points typically used for calibration for 2D gaze estimation [13, 18], our method is sufficiently accurate.

No significant effect for *Chin rest* ($F_{1,5} = 0.217$, $p = .661$; with chin rest $M = 3.02, SE = 0.267$; without chin rest $M = 3.13, SE = 0.26$) was present, suggesting that the usage of a chin rest has negligible influence on the angular accuracy. However, it should be noted that participants, although not explicitly instructed, were mostly trying to keep their head steady, most likely due to the general setup of the experiment. We believe that a plausible conclusion is that our method is not sensitive to minor head motion. This is also supported by the setup, in that slight head motion has no effect on the relative orientation and position of eye, eye camera, and world camera. As long as this system stays fixed, the mapping is unchanged. Giving participants the ability to move their head freely is an important feature for exploring objects in a natural, unconstrained manner. However, quantifying the effect of motions at larger scales should be subject to further investigations.

5.2 Accuracy of 3D gaze position

In our second experiment we explored the accuracy of our approach for viewing 3D stimuli. As model we employed the Stanford bunny and marked a set of pre-defined target points on the printed bunny as shown in Fig. 5, left. After a calibration with 11 correspondences as described in the last section, the test subjects were asked to individually focus on each of the targets (between 1 and 2 seconds). A heat map of the obtained gaze positions is shown in Fig. 5, right. Fixations are calculated based on Eq. 11 where c is set to be 0.6. Table 1 shows the angular error of each target in degrees as well as the depth error in mm.

Table 1: Errors of individual markers on bunny.

Marker index	1	2	3	4	5
Degree	0.578	1.128	0.763	0.846	0.729
Depth	7.998	8.441	10.686	3.036	8.381

Angular error depends mostly on the tracking setup, however, since the intersection computation with eye ray cones is restricted to points on the surface (vertices in our case), we get smaller angular errors on silhouettes.

Depth accuracy, on the other hand, depends on the slope of the geometry. In particular, at grazing angles, that is when the normal of the geometry is orthogonal or almost orthogonal to the viewing direction, it could become arbitrarily large. For the situations of interest to us where we have some control over the model, the normal is orthogonal or almost orthogonal to the viewing direction only around the silhouettes. However, since we determine the point on the object that best corresponds to the gaze direction, we obtain accurate results also around silhouettes. This is reflected in the preliminary experimental results where we obtain an average depth error of 7.71mm at a distance of 553.97mm, which corresponds to a relative error of less than 2%, despite three of five targets being very close to a silhouette.

6 Conclusion

We presented a simple yet accurate approach for tracking 3D gaze positions on known geometry using a monocular eye tracker. This is enabled by

- generating stimuli using digital manufacturing to obtain precisely known 3D geometry without restricting its shape;
- utilizing fiducial markers in a known relative position to the geometry to reliably determine its position relative to a subject's head;
- using a projective mapping to relate 3D positions to 2D pupil coordinates.

We experimentally verified our approach using two explorative user studies. The results demonstrate that 11 correspondences suffice to reliably calibrate the mapping from pupil coordinates to 3D gaze locations with an angular accuracy of 0.8 degree, which closely matches those of 2D gaze tracking. We also achieve a depth accuracy of 8.3mm at a distance of 550mm, corresponding to a relative error of less than 2%.

We developed our approach for 3D gaze tracking to analyze viewing behaviour for genuine 3D stimuli, and to explore what differences to 2D stimuli exist. Our approach enables researchers to study visual saliency on physical objects without scarifying to accuracy. Given the substantial amount of work on saliency and related questions that employed 2D stimuli for studying 3D objects, we believe this to be a worthwhile research question that deserves further attention.

REFERENCES

[1] W. W. Abbott and A. A. Faisal. Ultra-low-cost 3D gaze estimation: an intuitive high information throughput compliment to direct brain-machine interfaces. *Journal of Neural Engineering*, 9:1–11, 2012.

[2] S. Agarwal, K. Mierle, and Others. Ceres solver. `http://ceres-solver.org`.

[3] M. Barz, A. Bulling, and F. Daiber. Computational modelling and prediction of gaze estimation error for head-mounted eye trackers. Technical report, German Research Center for Artificial Intelligence (DFKI), 2015.

[4] G. Bradski. OpenCV. *Dr. Dobb's Journal of Software Tools*, 2000.

[5] N. Bruce and J. Tsotsos. Saliency Based on Information Maximization. In *Advances in Neural Information Processing Systems*, pages 155–162, 2006.

[6] N. Cournia, J. D. Smith, and A. T. Duchowski. Gaze-vs. hand-based pointing in virtual environments. In *CHI'03 extended abstracts on Human factors in computing systems*, pages 772–773. ACM, 2003.

[7] A. T. Duchowski, B. Pelfrey, D. H. House, and R. Wang. 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 '11*, page 15, New York, New York, USA, Aug. 2011. ACM Press.

[8] K. Essig, M. Pomplun, and H. Ritter. A neural network for 3D gaze recording with binocular eye trackers. *International Journal of Parallel, Emergent and Distributed Systems*, 21(2):79–95, Apr. 2006.

[9] J. Häkkinen, T. Kawai, J. Takatalo, R. Mitsuya, and G. Nyman. What do people look at when they watch stereoscopic movies? In A. J. Woods, N. S. Holliman, and N. A. Dodgson, editors, *Stereoscopic Displays and Applications XXI*, 2010.

[10] P. Hanhart and T. Ebrahimi. EYEC3D: 3D video eye tracking dataset. In *2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 55–56. IEEE, Sept. 2014.

[11] L. Heng, B. Li, and M. Pollefeys. Camodocal: Automatic intrinsic and extrinsic calibration of a rig with multiple generic cameras and odometry. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, pages 1793–1800. IEEE, 2013.

[12] C. Hennessey and P. Lawrence. Noncontact Binocular Eye-Gaze Tracking for Point-of-Gaze Estimation in Three Dimensions. *IEEE Transactions on Biomedical Engineering*, 56(3):790–799, 2009.

[13] K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. van de Weijer. *Eye Tracking: A Comprehensive Guide to Methods and Measures*. Oxford University Press, 2011.

[14] I. P. Howard. *Preceiving in Depth*. Oxford Psychology Series. Oxford University Press, 2012.

[15] Q. Huynh-Thu and L. Schiatti. Examination of 3D visual attention in stereoscopic video content. In B. E. Rogowitz and T. N. Pappas, editors, *IS&T/SPIE Electronic Imaging*, pages 78650J–78650J–15. International Society for Optics and Photonics, Feb. 2011.

[16] L. Jansen, S. Onat, and P. König. Influence of disparity on fixation and saccades in free viewing of natural scenes. *Journal of Vision*, 9(1):1–19, 2009.

[17] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *2009 IEEE 12th International Conference on Computer Vision*, pages 2106–2113. IEEE, Sept. 2009.

[18] M. Kassner, W. Patera, and A. Bulling. Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing Adjunct Publication - UbiComp '14 Adjunct*, pages 1151–1160, New York, New York, USA, Sept. 2014. ACM Press.

[19] J. Ki and Y.-M. Kwon. 3D Gaze Estimation and Interaction. In *2008 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video*, pages 373–376. IEEE, May 2008.

[20] J. J. Koenderink. Pictorial relief. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 356(1740):1071–1086, May 1998.

[21] C. Lang, T. V. Nguyen, H. Katti, K. Yadati, M. Kankanhalli, and S. Yan. Depth Matters: Influence of Depth Cues on Visual Saliency. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, editors, *Computer Vision ECCV 2012*, Lecture Notes in Computer Science, pages 101–115. Springer, 2012.

[22] J. W. Lee, C. W. Cho, K. Y. Shin, E. C. Lee, and K. R. Park. 3D gaze tracking method using Purkinje images on eye optical model and pupil. *Optics and Lasers in Engineering*, 50:736–751, 2012.

[23] C. Maggia, N. Guyader, and A. Guérin-Dugué. Using natural versus artificial stimuli to perform calibration for 3D gaze tracking. In B. E. Rogowitz, T. N. Pappas, and H. de Ridder, editors, *Human Vision and Electronic Imaging XVIII*. SPIE, 2013.

[24] S. Mathe and C. Sminchisescu. Dynamic Eye Movement Datasets and Learnt Saliency Models for Visual Action Recognition. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, editors, *Computer Vision ECCV 2012*, pages 842–856. Springer, 2012.

[25] T. Pfeiffer, M. E. Latoschik, and I. Wachsmuth. Evaluation of Binocular Eye Trackers and Algorithms for 3D Gaze Interaction in Virtual Reality Environments, 2008.

[26] T. Pfeiffer and P. Renner. Eyesee3d: A low-cost approach for analyzing mobile 3d eye tracking data using computer vision and augmented reality technology. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 369–376. ACM, 2014.

[27] S. Ramanathan, H. Katti, N. Sebe, M. Kankanhalli, and T.-S. Chua. An Eye Fixation Database for Saliency Detection in Images. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *Computer Vision ECCV 2010*, pages 30–43. Springer, 2010.

[28] C. Ramasamy, D. H. House, A. T. Duchowski, and B. Daugherty. Using eye tracking to analyze stereoscopic filmmaking. In *Posters on - SIGGRAPH '09*, page 1, New York, New York, USA, Aug. 2009. ACM Press.

[29] S. Stellmach, L. Nacke, and R. Dachselt. 3d attentional maps: aggregated gaze visualizations in three-dimensional virtual environments. In *Proceedings of the international conference on advanced visual interfaces*, pages 345–348. ACM, 2010.

[30] A. Toet. Computational versus psychophysical bottom-up image saliency: a comparative evaluation study. *IEEE transactions on pattern analysis and machine intelligence*, 33(11):2131–46, Nov. 2011.