3D Gaze Estimation for Head-Mounted Eye Tracking System with Auto-Calibration Method

Liu, Meng; Li, Youfu; Liu, Hai

Published in:
IEEE Access

Published: 01/01/2020

Document Version:
Final Published version, also known as Publisher’s PDF, Publisher’s Final version or Version of Record

License:
CC BY

Publication record in CityU Scholars:
Go to record

Published version (DOI):
10.1109/ACCESS.2020.2999633

Publication details:
Liu, M., Li, Y., & Liu, H. (2020). 3D Gaze Estimation for Head-Mounted Eye Tracking System with Auto-Calibration Method. IEEE Access, 8, 104207-104215. Article 9107144. https://doi.org/10.1109/ACCESS.2020.2999633

Citing this paper
Please note that where the full-text provided on CityU Scholars is the Post-print version (also known as Accepted Author Manuscript, Peer-reviewed or Author Final version), it may differ from the Final Published version. When citing, ensure that you check and use the publisher's definitive version for pagination and other details.

General rights
Copyright for the publications made accessible via the CityU Scholars portal is retained by the author(s) and/or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights. Users may not further distribute the material or use it for any profit-making activity or commercial gain.

Publisher permission
Permission for previously published items are in accordance with publisher's copyright policies sourced from the SHERPA RoMEO database. Links to full text versions (either Published or Post-print) are only available if corresponding publishers allow open access.

Take down policy
Contact lbscholars@cityu.edu.hk if you believe that this document breaches copyright and provide us with details. We will remove access to the work immediately and investigate your claim.

Download date: 08/06/2024
3D Gaze Estimation for Head-Mounted Eye Tracking System With Auto-Calibration Method

MENG LIU1, YOUFU LI1, (Senior Member, IEEE), AND HAI LIU1,2, (Member, IEEE)
1Department of Mechanical Engineering, City University of Hong Kong, Hong Kong
2National Engineering Research Center for E-Learning, Central China Normal University, Wuhan 430079, China
Corresponding author: Youfu Li (meyfli@cityu.edu.hk)

This work was supported in part by the Research Grants Council of Hong Kong Project under Grant CityU 11203619, and in part by the National Natural Science Foundation of China under Grant 61873220 and Grant 61875068.

ABSTRACT The general challenges of 3D gaze estimation for head-mounted eye tracking systems are inflexible marker-based calibration procedure and significant errors of depth estimation. In this paper, we propose a 3D gaze estimation with an auto-calibration method. To acquire the accurate 3D structure of the environment, an RGBD camera is applied as the scene camera of our system. By adopting the saliency detection method, saliency maps can be acquired through scene images, and 3D salient pixels in the scene are considered potential 3D calibration targets. The 3D eye model is built on the basis of eye images to determine gaze vectors. By combining 3D salient pixels and gaze vectors, the auto-calibration can be achieved with our calibration method. Finally, the 3D gaze point is obtained through the calibrated gaze vectors, and the point cloud is generated from the RGBD camera. The experimental result shows that the proposed system can achieve an average accuracy of 3.7° in the range of 1 m to 4 m indoors and 4.0° outdoors. The proposed system also presents a great improvement in depth measurement, which is sufficient for tracking users’ visual attention in real scenes.

INDEX TERMS Head-mounted gaze tracking system, saliency maps, auto-calibration, 3D gaze estimation.

I. INTRODUCTION Gaze estimation is the process of predicting where someone is looking, either as gaze directions or as points of regard (PoR) in space. Vision is one of the primary resources of collecting surrounding information; our eyes will turn toward a person or an object we are looking. According to the analysis of researchers in the United States, the degree to which we rely on each sense in the performance of everyday activities is approximately: 1% for taste, 1.5% for touch, 3.5% for smell, 11% for hearing, and 83% for sight [1]. Therefore, estimating users’ gaze vectors or gaze points will be of great help to understand human activities. Nowadays, gaze estimation systems have been applied in many fields, such as human-machine interactions [2], assisted driving [3], and surgery assistance [4].

Gaze estimation systems can be generally classified into remote devices and head-mounted devices (HMD) [5]. The remote device is a screen-based interaction system that works at a distance from the subject [6]. This device contains at least one user-facing camera to capture images of the subject’s face so that the PoR can be estimated according to the extracted features from face images. Based on the remote device’s working principle, it needs the participant’s head to be in the field of the camera all the time, which limits the participant’s head-body mobility. In contrast, HMD is designed as a head equipment that can acquire clear eye images and allows users to move their heads freely. HMD usually consist of a scene camera and two eye cameras. The scene camera is used to obtain scene images that the user sees, and eye cameras are used to record eye movement while the user is looking at a scene. In this way, HMD estimate humans’ PoR in the scene camera’s coordinate on the basis of eye images. Recently, lightweight HMD have become popular eye tracking systems for gaze researchers due to their flexibility and mobility. Such devices extends a user’s gaze estimation field from desktop or computer screen into other scenes that greatly enriches the collection of human gaze data.

After several studies have attempted to predict PoR on a scene image plane or a screen, interest in estimating humans’ gaze location in 3D coordinate has been increasing. 3D gaze estimation not only predicts PoR in the field of view with
depth information, but also proves the connection between a scene’s saliency and human-related motion [7]. To our knowledge, one of the earliest systems that is capable of estimating 3D PoR requires a fixed head-to-camera setting [8]. Kwon et al. introduce a novel binocular technique, in which gaze direction is computed by using glints on the corneal and then depth is inferred by interpupillary distance. Though 3D gaze estimation has been widely studied in remote gaze estimation systems [9], research on HMD remains limited. For both gaze estimation systems, the most common approach is to estimate the 3D gaze point as the midpoint of the shortest segment between both eyes’ visual vectors [10]. However, deviations in the eye gaze vector’s calculation are likely to cause considerable variance in the PoR’s depth direction. To address this problem, Lee et al. use multi-layered perception to obtain depth gaze position [11]. However, the method employs dual Purkinje images as the input, which are hard to be detected in practice.

Another challenge for 3D gaze estimation is that eye tracking systems need to be calibrated for each user before an estimation. During calibration, the subject needs to stare at the specific reference marker and sometimes, such active personal calibration interrupts user-scene interactions. Although the calibration procedure has become much simpler and the number of calibration markers has been reduced down to one in some works [12]; it still requires the user to participate in the calibration task actively.

Several approaches have been proposed to avoid the person-specific calibration process. Alnajar et al. claim that the gaze patterns of several viewers provide important cues for the auto-calibration in [13]. By making use of the topology of pre-recorded gaze pattern, a transformation is computed to map the initial gaze points to match the gaze pattern. Sugano treats mouse clicks on a computer screen as gaze points to train the mapping function between the eye features and PoR [14]. Similar to [14], the algorithm in [15] detects the user’s hand and finger tip which indicate the user’s point of interest. This method can easily collect calibration samples in different environments quickly, and the proposed method achieves comparable accuracy to standard marker-based calibration. Moreover, Lander and his co-workers combine the pupil center position and the scene reflection on the corneal surface to predict the actual PoR in the scene [16]. All these approaches attempt to avoid using calibration markers. Nonetheless, they all rely on observations of a specific person and environment, which limits their applicability.

Visual salience has also been considered as the cue for predicting PoRs, as Parkhurst et al. claim that there is a connection between visual saliency and gaze positions [17]. Biologically, a human is more likely to focus on a part in the scene with high saliency, i.e., a region with peculiar characters compared with surrounding features. After Koch and Ullman came up with the concept of saliency [18], different kinds of saliency algorithms have been proposed [19], [20]. In contrast to the methods mentioned above, visual saliency can be applied to a variety of scenes and does not require any restrictions on the user, which shows its high flexibility and wide adaptability.

In this paper, we present a head-mounted eye tracking system to predict users’ PoR in 3D environment (as shown in Fig. 1). This system consists of a simple setup and can achieve good estimation accuracy. We also propose a novel auto-calibrating 3D gaze estimation method named 3DGAC. This method combines with the saliency-based algorithm to calculate possible 3D calibration targets in the scene instead of asking the subject to fixate on a certain point, thereby achieving a fully automatic calibration procedure without any restrictions and human intervention enforced on the subject. For the system’s architecture, we replace the regular RGB camera with an RGBD camera, which provides accurate 3D data of the scene. During the usage, participants can change their location and head pose with no constraint. In the calibration step, we integrate saliency maps with gaze vectors to search for extrinsic parameters between the scene camera and two eye cameras. In the gaze estimation section, we firstly rotate gaze vectors from both eye camera coordinate systems into the scene camera coordinate system. Then, we use the point cloud generated by the RGBD camera to find the optimal PoR in a 3D environment. The main contributions of this paper are threefold:

1) We combine RGBD images with the saliency-based method into auto-calibration to achieve 3D gaze points estimation for our head-mounted eye tracking system. To the best of our knowledge, this is the first work to do that;

2) We propose a novel method to determine the rotation matrix that transfers gaze vectors into the scene camera coordinate system;

3) We perform experiments indoors and outdoors. The performance of PoR estimation is encouraging, and we also achieve the remarkably improvement in the accuracy of PoR’s depth estimation.

The rest of this paper is organized as follows. Section II presents details of 3DGAC. Section III shows the experiment result and discussion. Section IV concludes this study.
II. METHODOLOGY

A. ARCHITECTURE FOR HMD

A head-mounted eye tracking system consists of two parts. An Intel RealSense D435 RGBD camera is used as the scene camera to provide scene images with the depth measurement range of over 10m. Two IR cameras are leveraged as eye cameras to capture clear IR eye streams [21], [22]. Besides, there is an IR light illuminating eye regions. All the capture devices are connected to a laptop, and they are set to be triggered at the same time so that we will capture two IR eye images, one RGB scene image, and one depth image simultaneously.

As the scene camera, Intel RealSense D435 RGBD camera contains two modules. The RGB module captures RGB scene images; the depth module has two IR cameras and an IR projector for obtaining depth images of the scene. With the software module in librealsense libraries, the depth image can be easily aligned to the RGB image. Thus, all the scene image pixels’ 3D positions can be obtained in the scene camera coordinates.

B. CALIBRATION PROCEDURE

The framework of 3DGAC is shown in Fig. 2. During calibration, the participant can scan the surrounding environment randomly.

For RGB scene images, we apply the algorithm in [23] to generate saliency maps, which represent the distinctive features of scene images. We have also considered using the RGB-D saliency detection algorithm to implement our self-calibration algorithm. However, those methods generally calculate the saliency map based on the trained neural network [24], [25]. In contrast, traditional saliency algorithms have better universality and can be used on low-cost equipment. Therefore, we finally use the traditional RGB saliency detection algorithm to calculate the saliency maps of scene images. Fig. 3 presents examples of scene images and their corresponding saliency maps. After the generation of each saliency map, we pick out pixels with higher saliency value than the pre-setting threshold. Then a clustering method based on [26] is employed to remove noise and improve the reliability of the saliency map. For the multiple salient regions in the scene, we preserve all the pixels of salient parts as shown in Fig. 4. Consequently, the target data set can be collected \( \{ t_i \} \) for all the 3D coordinates of the space points that correspond to our chosen pixels. \( t_i \) stands for the selected 3D targets of the saliency map, and \( j \) is the image’s index. The details of the data acquisition process are outlined in Fig. 5.

There are generally two types of methods establish the associations between eye features and targets in the scene: the regression-based and model-based approaches. Unlike regression-based approaches that directly create a mapping relationship between eye features and gaze points, model-based approaches first use extracted eye features to build the eyeball model. Once the model is built, the initial gaze vectors would be obtained, and they will be used to determine the real gaze vectors in the scene camera coordinates. In this paper, we adopt the 3D eyeball reconstruction method as in [27]. When modeling the camera imaging sensor as a pinhole model, the pupil contours on the eye images can be back-projected into 3D space as circles:

\[
P_{\text{pupil}} = (m, l) \tag{1}
\]
where \( m \) donates a pupil position and \( I \) is a normal passing through \( m \). After the back-projection, a 3D eye model can be built on the basis of multiple circles.

We assume the 3D eye model as a sphere and every space circle would be tangent to the sphere at the circle’s center. In this case, \( I \) is treated as a normal which is passing through the eye model center and parallel with the gaze vector. Once we have collected enough eye images, all the pupil contours on those eye images are extracted and back-projected into 3D space. The eyeball center \( e \) is determined as the point closest to each normal, which can be obtained by minimizing the function as follows:

\[
e = \left( \sum_j (I - I_j^t)^{-1} \left( \sum_j (I - I_j^t)m_j \right) \right)^{-1}(2)
\]

where \( j \) is the index of eye image. Once the eyeball center \( e \) is recovered, the 3D eyethe 3D gaze vector \( n \) can be calculated as the vector origins from \( e \) and points to \( m \) in the eye camera coordinate system.

Given that each user has the specific gaze habit in a scene, the salience part is not sufficient enough for determining the exact position as the calibration marker does. To combat this problem, we propose a novel method to find relationships among the saliency map and two gaze vectors.

Considering short distances between user’s eyes and the scene camera are within a few centimeters whereas user’s fixation usually locates at meters away, it is reasonable to assume that the user’s eyeball center coincides with the scene camera coordinate system’s origin (we perform a simulation to evaluate the angle deviation without this eye-camera distance in Section III-A). Thus, the eyes and scene camera observe the object with the approximately same viewing directions.

For the computation of the extrinsic parameter between the left eye camera coordinate system and the scene camera coordinate system, we acquire the corresponding gaze vector \( n_j^l \), and the target data set is acquired as \( \{t_i\}_j \) for \( j^{th} \) saliency map. Then we delete targets whose depth is less than 1m (based on the simulation result) to improve the reliability of our proposed calibration method. Each target vector can be represented as \( n_j^l = e_l \) where \( e_l \) stands for the user’s left eyeball center location. \( \alpha_l \) and \( \beta_l \) represent the vertical and horizontal angle between the gaze and target vectors, respectively. Then the specific rotation matrix \( R_{lj} \) can be represented as

\[
R_{lj} = \begin{bmatrix}
\cos \beta_l & \sin \beta_l \sin \alpha_l & \sin \beta_l \cos \alpha_l \\
0 & \cos \alpha_l & -\sin \alpha_l \\
-\sin \beta_l & \cos \beta_l \sin \alpha_l & \cos \beta_l \cos \alpha_l
\end{bmatrix}
\]

(3)

We obtain the rotation matrix set \( \{R_{lj}\} \) through

\[
R_{lj} \times n_j^l = \frac{t_l - e_l}{|t_l - e_l|}
\]

(4)

which can be represented as solving the overdetermined linear equation as follows:

\[
\begin{bmatrix}
\cos \beta_l & \sin \beta_l \sin \alpha_l & \sin \beta_l \cos \alpha_l \\
0 & \cos \alpha_l & -\sin \alpha_l \\
-\sin \beta_l & \cos \beta_l \sin \alpha_l & \cos \beta_l \cos \alpha_l
\end{bmatrix}\begin{bmatrix}
n_{lj}^1 \\
n_{lj}^2 \\
n_{lj}^3
\end{bmatrix} = \begin{bmatrix}
t_l^1 \\
t_l^2 \\
t_l^3
\end{bmatrix}
\]

(5)

Similarly, we further acquire a set of rotation matrices \( \{R_{lj}\} \) that corresponds to all the gaze vectors and all selected pixels in all the saliency maps. During the calculation, a two-dimensional cumulative array \( A(\alpha_l, \beta_l) \) is created to count the frequency of angles (as shown in Fig. 6). Those angle pairs’ frequencies keep growing as more and more saliency maps being involved. It may occur the situation that multiple angle pairs show high frequency during the calculation as Fig. 6(a), Fig. 6(b) and Fig. 6(c). Since there is only one correct extrinsic parameter between the scene camera and the left eye camera, the frequency of one angle pair will be significantly higher than others in the end like Fig. 6(f). After the calculation, the \( R_{lj} \) is determined based on the most frequent angle pair and leveraged to restore the extrinsic parameters \( R_l \) between the left eye camera and the scene camera coordinate systems. We determine the rotation matrix \( R_l \) when its maximum angle set’s frequency is two times greater than the second largest frequency.

With the same method, we can also recover the extrinsic parameters between the right eye camera and the scene camera coordinate systems, which is represented by \( R_r \).

The two rotated gaze vectors are shown as follows:

\[
V_{lj} = e_l + R_l \times n_j^l
\]

(6)

\[
V_{ri} = e_r + R_r \times n_j^r
\]

(7)

C. GAZE ESTIMATION

For the traditional model-based gaze estimation method, the gaze point in 3D is computed as the midpoint of the shortest segment between two rotated gaze vectors. However, due to the short baseline between human eyes, small angle calculation errors in rotated gaze vectors can cause large deviations in the Z direction of the gaze point estimation (as shown in Fig. 7).
FIGURE 6. Angle sets’ frequency changes as the number of scene saliency maps increases. From (a) to (f), the total number of involved scene saliency maps are: 1, 5, 10, 20, 50, 100.

FIGURE 7. Angular error of gaze estimation barely impacts the correct positioning on the facing plane, but it can cause considerably variance in the PoR’s depth direction.

To refine the raw 3D gaze estimate method, we generate a point cloud of the environment (as shown in Fig. 8). Consider a 3D vector \( N \) passing through the origin \( O = (0, 0, 0) \), the distance of space point \( p \) from \( N \) can be calculated as

\[
d(p, N) = \frac{|N \times (p - O)|}{|N|}
\]  

(8)

For two rotated gaze vectors in space, we define the gaze point by minimizing the equation as follows

\[
d = \frac{|V_{li} \times (p_i - e_l)|}{|V_{li}|} + \frac{|V_{ri} \times (p_i - e_r)|}{|V_{ri}|}
\]  

(9)

FIGURE 8. Point cloud of the environment. The gaze point is defined as the point with the minimal distance to both rotated gaze vectors.

III. EXPERIMENT RESULT AND DISCUSSION

A. SIMULATION RESULTS

We use MATLAB to evaluate the effect of ignoring eye-camera distance on the accuracy of 3D gaze point estimation at 10 different depths from 0.5m to 5m. Based on the architecture of our gaze tracker, the distances from the scene camera to the user’s left and right eyeball centers are approximately 70mm and 90mm, respectively. At each depth, 100 points are randomly generated as 3D gaze targets, and those targets cover a field of view which is 60° × 55°. For each target, the vector that starts from the eyeball center and points to the 3D target is tread as the real gaze vectors while the assumptive gaze vector starts from the scene camera’s origin. We calculate the angle between each pair vectors and the result is shown in Fig. 9.

In Fig. 9, it emerges that the eye-camera distance can cause a considerably angle deviation when the depth is small, but the mean error drops to an acceptable level as the depth increases to 1m. Besides, when the depth exceeds 2m, mean angle errors and standard deviations for both eyes can achieve an encouraging precision. Considering the subject’s gaze distance normally beyond 1m in a mobile situation, assuming that the user’s eyeball center coincides with the scene camera coordinate system’s origin is reasonable.

B. EXPERIMENT RESULTS AND DISCUSSION

Our experimental system is based on a low-cost HMD developed in our laboratory, which applies Intel RealSense D435 RGBD camera as the scene camera and two IR cameras as eye cameras. All the cameras run at approximately 25fps during the experiment procedure with resolutions of 640*480 pixels. Before the calibration and the estimation, three cameras have been calibrated, and as a result, the lens distortion...
can be corrected by the MATLAB toolbox. The proposed gaze tracking algorithm in this paper is running on a laptop PC with an Intel i7 2.70GHz, 2.90GHz quad core CPU and an 8.00GB RAM. The computation cost for calculating the gaze vector from an eye image and generating a saliency map from a scene image are about 15ms and 18ms, respectively. To ensure the real-time gaze calculation, our proposed method is executed in a multi-thread manner. One thread runs at 30 fps to calculates gaze vectors based on eye images. In the meanwhile, the other thread runs at 10fps, which generates a saliency map from one of the three consecutive scene images and discarding the other two.

First, we conduct the 3D gaze estimation experiment in the office. 10 subjects are invited to validate the proposed method. Fig. 10 demonstrates the setup of our indoor experiment. During the calibration, subjects are encouraged to look at surroundings randomly, and there is no limitation on their head poses or standing position. The entire calibration process for each subject usually takes 20 seconds, as the algorithm takes time to establish eyeball models. In the step of gaze estimation, we apply a board with 5 concentric circle targets to test the angle and the depth accuracy of 3DGAC. Taking the RGBD camera’s depth measure range and the

room size into consideration, four tests are carried out by putting the board at 4 different depths from the subject: 1m, 2m, 3m, and 4m. For each depth, they should fixate at the 5 targets in the same order: top-left, top-right, bottom-right, bottom-left, and center, each target lasts for 2 to 3 seconds. Fig. 11(a) illustrates the PoR estimation errors in degrees for each subject at 4 different distances and the depth estimation error is shown in Fig. 11(c).

The average result in different depth is shown in Table 1, we use the average depth error (ADE) and the average angular error(AAE) to measure the accuracy. The overall ADE of indoors experiment is 55.9mm and the overall AAE is 3.7 degrees.

Table 2 demonstrates the comparison of the proposed method with several state-of-the-art remote-based gaze estimation methods, four models are included as follows.

| Methods      | 3DDEFM [28] | RGBD-GT [29] | RTGTK [30] | RGBD-EMGE [31] | 3DGAC |
|--------------|-------------|--------------|------------|----------------|-------|
| Error(deg)   | 3.5         | 4.4          | 4.0        | <10            | 3.7   |

According to Table 2, 3DGAC gives a relatively good performance in estimating the gaze vector’s direction compared with four remote-based methods that also adopt RGBD sensors. Although the angular error of 3DDEFM is smaller than the proposed method, it is evaluated in a hand-reaching distance. Besides, our system demonstrates advantages with more flexibility: the experiment is conducted in a larger range (from 1m to 4m); the participant can move and change his or
her head pose freely during the entire process, and calibration markers are unnecessary to assist the calibration process.

In Table 3, we compare the proposed method with four head-mounted-based gaze estimation methods.

a) 3DGE-GPR: 3D gaze estimation with Gaussian processes regressor method is proposed by Elmadjian et al. in [32].

b) 3DGE-3DEP: 3D gaze estimation on a 3D environment map is designed by Takemura et al. in [33].

c) GEMFR: GEMFR is a method using multiple feature regression to predict 3D gaze which is submitted by Weier et al. in [34].

d) SCGE: SCGE is an self-calibrating gaze estimation method proposed by Sugano and Bulling in [35].

From Table 3, we can figure that the average angular error of our method is 3.7°, which outperforms the accuracy of 3DGE-GPR, 3DPoRE-3DEP and SCGE (4.9°, 5.27° and 6° respectively). GEMFR reports the accuracy of 1°. However, their method requires complex hardware setups and time-consuming calibration period. For estimating 3D PoR’s depth, the average depth error of 3DGAC from 1m to 4m is 55.9mm, which has achieved significant improvement compared with other four head-mounted-based gaze estimation methods.

From Table 1, it can be observed that our gaze estimation method gives better performance in long distance. This instance may be due to the assumption of ignoring the distance between the eyeball center and the scene camera when we calculate the rotation matrix for each eye. If the estimation distance was relatively short, then the eye-camera distance may affect the computation of rotation matrices. In this way, the gaze system’s result will be more accurate when the user gazes at a place more extensive.

Fig. 12 demonstrates the setup of our outdoor experiment. The calibration and the gaze estimation period are the same as the first experiment. The PoR estimation errors in degrees for each subject at 4 different distances is shown in Fig. 11(b), and the depth estimation error is shown in Fig. 11(d). Table 4 illustrate the average result in different depth, and the overall
and use them as possible 3D calibration targets. Through the aggregation, we obtain the rotation matrix between the eye camera coordinate system and the scene camera coordinate system. The entire calibration procedure is completed automatically without pre-setting markers and any external assistance. To improve the accuracy of depth estimation, environmental point cloud data is applied in the PoR’s estimation. Once we have identified the final gaze vector, the PoR is calculated as it is the closest point to the final gaze vector. On the basis of the experiment results and the comparison with other state-of-the-art approaches, the proposed method achieves a relatively accurate measurement in depth with encouraging angular estimation precision.

### TABLE 4. Outdoors PoR estimation errors.

| Distance (m) | Average |
|--------------|---------|
| 1            | AAE(deg) 7.1 |
|              | ADE(mm) 6.7 |
| 2            | AAE(deg) 2.9 |
|              | ADE(mm) 43.9 |
| 3            | AAE(deg) 2.5 |
|              | ADE(mm) 65.6 |
| 4            | AAE(deg) 3.6 |
|              | ADE(mm) 157.3 |

average depth error is 68.4mm and the average angular error is 4.0 degrees.

Comparing the results of two experiments, the accuracy of PoR estimation and depth measurement decrease outdoor when the subject stands at the same distance from the board. We believe that the main reason for this result is the impact of sunlight. Given that we utilize IR light sources to capture the dark pupil image and extract the pupil contour, sunlight will affect the illumination of the IR source which causes poor pupil detection. Moreover, the RGBD camera acquires depth data based on the IR projector and two IR cameras; hence, sunlight also bring bad influence on PoR’s depth measurement.

### IV. CONCLUSION

In this paper, we present a head-mounted eye tracking system to estimate users’ 3D PoR. A novel auto-calibrating 3D gaze estimation method named 3DGAC is proposed to simplify the calibration procedure and improve the system’s flexibility. For the hardware design, we replace the regular RGB camera of head-mounted eye tracking system with an RGBD camera to obtain the depth information of the scene. During the calibration procedure, the saliency algorithm is introduced to generate saliency maps based on scene images. We align those salient pixels to the depth image and use them as possible 3D calibration targets. Through the aggregation, we obtain the rotation matrix between the eye camera coordinate system and the scene camera coordinate system. The entire calibration procedure is completed automatically without pre-setting markers and any external assistance. To improve the accuracy of depth estimation, environmental point cloud data is applied in the PoR’s estimation. Once we have identified the final gaze vector, the PoR is calculated as it is the closest point to the final gaze vector. On the basis of the experiment results and the comparison with other state-of-the-art approaches, the proposed method achieves a relatively accurate measurement in depth with encouraging angular estimation precision.

### REFERENCES

[1] L. D. Rosenblum, *See What I’m Saying: The Extraordinary Powers Our Five Senses*. New York, NY, USA: WW Norton & Company, 2011.

[2] M. A. Eid, N. Giakoumis, and A. El Saddik, “A novel eye-gaze-controlled wheelchair system for navigating unknown environments: Case study with a person with ALS,” *IEEE Access*, vol. 4, pp. 558–573, 2016.

[3] S. Martin, S. Vora, K. Yuen, and M. M. Trivedi, “Dynamics of driver’s gaze: Explorations in behavior modeling and maneuver prediction,” *IEEE Trans. Intell. Veh.*, vol. 3, no. 2, pp. 141–150, Jan. 2018.

[4] P. Li, X. Hou, X. Duan, H. Yip, G. Song, and Y. Liu, “Appearance-based gaze estimator for natural interaction control of surgical robots,” *IEEE Access*, vol. 7, pp. 25095–25110, 2019.

[5] A. Kar and P. Corcoran, “A review and analysis of eye-gaze estimation systems, algorithms and performance evaluation methods in consumer platforms,” *IEEE Access*, vol. 5, pp. 16495–16519, 2017.

[6] N. M. Murar, H. Gao, and J.-P. Thiran, “A regression-based user calibration framework for real-time gaze estimation,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 12, pp. 2623–2638, Dec. 2017.

[7] K. R. Jerripothula, J. Cai, and J. Yuan, “Efficient video object co-localization with co-saliency activated tracklets,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 3, pp. 744–755, Mar. 2019.

[8] Y.-M. Kwon, K.-W. Jeon, J. Ki, Q. M. Shahab, S. Jo, and S.-K. Kim, “3D gaze estimation and interaction to stereo display,” *Int. J. Virtual Reality*, vol. 5, no. 3, pp. 41–45, Jan. 2006.

[9] F. Lu, Y. Gao, and X. Chen, “Estimating 3D gaze directions using unlabeled eye images via synthetic iris appearance fitting,” *IEEE Trans. Multimedia*, vol. 18, no. 9, pp. 1772–1782, Sep. 2016.

[10] F. Pirri, M. Pizzoli, and A. Rudi, “A general method for the point of regard estimation in 3D space,” in *Proc. CVPR*, Jun. 2011, pp. 921–928.

[11] J. W. Lee, C. W. Cho, K. Y. Shin, E. C. Lee, and K. R. Park, “3D gaze tracking method using puckinie images on eye optical model and pupil,” *Opt. Lasers Eng.*, vol. 50, no. 5, pp. 736–751, May 2012.

[12] M. Qodseya, M. Sanzari, V. Ntouskos, and F. Pirri, “3D: A device for studying gaze in 3D,” in *Proc. Eur. Conf. Comput. Vis.*, vol. 7, pp. 25095–25110, 2017.

[13] F. Alnajar, T. Gevers, R. Valenti, and S. Ghebreab, “Auto-calibrated gaze estimator for natural interaction control of surgical robots,” *IEEE Trans. Human-Mach. Syst.*, vol. 5, no. 6, pp. 750–760, Dec. 2015.

[14] A. Kar and P. Corcoran, “A review and analysis of eye-gaze estimation systems, algorithms and performance evaluation methods in consumer platforms,” *IEEE Access*, vol. 5, pp. 16495–16519, 2017.

[15] M. Bâce, S. Staal, and G. Sörös, “Wearable eye tracker calibration at your fingertips,” in *Proc. ACM Symp. Eye Tracking Res. Appl.*, Jun. 2018, pp. 1–5.

[16] C. Lander, F. Kosmalla, F. Wiehr, and S. Gehring, “Using corneal imaging for measuring a human’s visual attention,” in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput. Proc. ACM Int. Symp. Wearable Comput.*, Sep. 2017, pp. 947–952.

[17] D. Parkhurst, K. Law, and E. Niebur, “Modeling the role of salience in the allocation of overt visual attention,” *Vis. Res.*, vol. 42, no. 1, pp. 107–123, Jan. 2002.

[18] C. Koch and S. Ullman, “Shifts in selective visual attention: Towards the underlying neural circuitry,” in *Matters of Intelligence*. Dordrecht, The Netherlands: Springer, 1987, pp. 115–141.
[19] R. Cong, J. Lei, H. Fu, M.-M. Cheng, W. Lin, and Q. Huang, “Review of visual saliency detection with comprehensive information,” IEEE Trans. Circuits Syst. Video Technol., vol. 29, no. 10, pp. 2941–2959, Oct. 2019.

[20] R. Cong, J. Lei, C. Zhang, Q. Huang, X. Cao, and C. Hou, “Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion,” IEEE Signal Process. Lett., vol. 23, no. 6, pp. 819–823, Jun. 2016.

[21] T. Liu, H. Liu, Y.-F. Li, Z. Chen, Z. Zhang, and S. Liu, “Flexible FTIR spectral imaging enhancement for industrial robot infrared vision sensing,” IEEE Trans. Ind. Informat., vol. 16, no. 1, pp. 544–554, Jan. 2020.

[22] T. Liu, H. Liu, Y. Li, Z. Zhang, and S. Liu, “Efficient blind signal reconstruction with wavelet transforms regularization for educational robot infrared vision sensing,” IEEE/ASME Trans. Mechatronics, vol. 24, no. 1, pp. 384–394, Feb. 2019.

[23] J. Harel, C. Koch, and P. Perona, “Graph-based visual saliency,” in Proc. Adv. Neural Inf. Process. Syst., 2007, pp. 545–552.

[24] C. Li, R. Cong, S. Kwong, J. Hou, H. Fu, G. Zhu, D. Zhang, and Q. Huang, “ASIF-Net: Attention steered interweave fusion network for RGB-D salient object detection,” IEEE Trans. Cybern., early access, Feb. 13, 2020, doi: 10.1109/TCYB.2020.2969255.

[25] R. Cong, J. Lei, H. Fu, J. Hou, Q. Huang, and S. Kwong, “Going from RGB to RGBD saliency: A depth-guided transformation model,” IEEE Trans. Cybern., early access, Aug. 20, 2019, doi: 10.1109/TCYB.2019.2932005.

[26] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in Proc. KDD, vol. 96, 1996, pp. 226–231.

[27] L. Swirski and N. Dodgson, “A fully-automatic, temporal approach to single camera, glint-free 3D eye model fitting,” Proc. PETMEI, Aug. 2013, pp. 1–11.

[28] K. Wang and Q. Ji, “Real time eye gaze tracking with 3D deformable eye-face model,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 1003–1011.

[29] X. Xiong, Q. Cai, Z. Liu, and Z. Zhang, “Eye gaze tracking using an RGBD camera: A comparison with a RGB solution,” in Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput. Adjunct Publication Ubicomp Adjunct, 2014, pp. 1113–1121.

[30] K. Wang and Q. Ji, “Real time eye gaze tracking with Kinect,” in Proc. 23rd Int. Conf. Pattern Recognit. (ICPR), Dec. 2016, pp. 2752–2757.

[31] L. Jianfeng and L. Shigang, “Eye-model-based gaze estimation by RGB-D camera,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops, Jun. 2014, pp. 592–596.

[32] C. Elmadjian, P. Shukla, A. D. Tula, and C. H. Morimoto, “3D gaze estimation in the scene volume with a head-mounted eye tracker,” in Proc. Workshop Commun. Gaze Interact. COGAIN, 2018, p. 3.

[33] K. Takemura, K. Takahashi, J. Takamatsu, and T. Ogasawara, “Estimating 3-D point-of-regard in a real environment using a head-mounted eye-tracking system,” IEEE Trans. Human-Mach. Syst., vol. 44, no. 4, pp. 531–536, Aug. 2014.

[34] M. Weier, R. Roth, A. Hinkenjann, and P. Shusallek, “Predicting the gaze depth in head-mounted displays using multiple feature regression,” in Proc. ACM Symp. Eye Tracking Res. Appl., Jun. 2018, p. 19.

[35] Y. Sugano and A. Bulling, “Self-calibrating head-mounted eye trackers using egocentric visual saliency,” in Proc. 28th Annu. ACM Symp. User Interface Softw. Technol. UIST, 2015, pp. 363–372.

MENG LIU received the M.S. degree from the Department of Science and Engineering, Huangzho  University of Science and Technology (HUST), Wuhan, China, in 2016. He is currently pursuing the Ph.D. degree with the Department of Mechanical Engineering, City University of Hong Kong, Hong Kong.

His research interests include computer vision, human–machine interaction, and visual tracking.

YOUPU LI (Senior Member, IEEE) received the Ph.D. degree in robotics from the Department of Engineering Science, University of Oxford, U.K., in 1993.

From 1993 to 1995, he was a Research Staff with the Department of Computer Science, University of Wales, Aberystwyth, U.K. He joined the City University of Hong Kong, in 1995, where he is currently a Professor with the Department of Mechanical Engineering. His research interests include robot sensing, robot vision, and visual tracking. In these areas, he has published over 150 articles in refereed international journals. He has served as an Associate Editor for the IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING (T-ASE) and the IEEE Robotics and Automation Magazine (RAM), an Editor for CEB and the IEEE International Conference on Robotics and Automation (ICRA), and a Guest Editor for the IEEE Robotics and Automation Magazine (RAM).

HAI LIU (Member, IEEE) received the M.S. degree in applied mathematics and the Ph.D. degree in pattern recognition and artificial intelligence from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2010 and 2014, respectively.

Since June 2017, he has been an Assistant Professor with the National Engineering Research Center for E-Learning, Central China Normal University, Wuhan. He was a Hong Kong Scholar Postdoctoral Fellow with the Department of Mechanical Engineering, City University of Hong Kong, Hong Kong, where he was hosted by the Prof. Y.-Fu Li. He has held the position two years till March 2019. He has authored more than 60 peer-reviewed articles in international journals from multiple domains, such as pattern recognition, image processing, and 3D gaze estimation. More than six articles are selected as the ESI highly cited articles. His current research interests include robot vision, big data processing, artificial intelligence, spectral analysis, optical data processing, and pattern recognition.

Dr. Liu has been frequently serving as a Reviewer for more than six international journals including the IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, the IEEE TRANSACTIONS ON CYBERNETICS, the IEEE/ASME TRANSACTIONS ON MECHATRONICS, and the IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT. He is a Communication Evaluation Expert for the National Natural Science Foundation of China.