2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) | 978-1-7281-6212-6/20/$31.00 ©2020 IEEE | DOI: 10.1109/IROS45743.2020.9341755

3D Gaze Estimation for Head-Mounted Devices based on Visual Saliency

Meng Liu[1], You Fu Li[1] and Hai Liu[2]

Abstract— **Compared with the maturity of 2D gaze tracking technology, 3D gaze tracking has gradually become a research hotspot in recent years. The head-mounted gaze tracker has shown great potential for gaze estimation in 3D space due to its appealing flexibility and portability. The general challenge for 3D gaze tracking algorithms is that calibration is necessary before the usage, and calibration targets cannot be easily applied in some situations or might be blocked by moving human and objects. Besides, the accuracy on depth direction has always come to be a crucial problem. Regarding the issues mentioned above, a 3D gaze estimation with auto-calibration method is proposed in this study. We use an RGBD camera as the scene camera to acquire the accurate 3D structure of the environment. The automatic calibration is achieved by uniting gaze vectors with saliency maps of the scene which aligned depth information. Finally, we determine the 3D gaze point through a point cloud generated from the RGBD camera. The experiment result demonstrates that our proposed method achieves 4.34° of average angle error in the field from 0.5m to 3m and the average depth error is 23.22mm, which is sufficient for 3D gaze estimation in the real scene.**

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. As human beings, vision is the primary resource of collecting the surrounding information in our daily lives. While we observe surroundings, our eyes will turn towards the person or the 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: taste 1%, touch 1.5%, smell 3.5%, hearing 11%, and visual 83%. Therefore, estimating users' gaze vectors or gaze points will be of great help to understand human activities. Nowadays, gaze estimation techniques have been applied in many fields, such as human-computer interactions, assisted driving and surgery assistance.

Gaze estimation systems can be generally classified into remote devices and head-mounted devices (HMD) [1]. The remote device is a screen-based interaction system which works at a distance from the subject [2], [3]. It 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. Even some remote devices have been applied in open interaction settings, it still 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 head equipment that can acquire clearer eye images and allows users to move their head freely. HMD usually consists 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 eyes' movement while the user is looking at the scene. In this way, HMD estimate human's PoR in the scene camera's coordinate based on eye images. Recently, lightweight HMD has become a popular topic for gaze researchers due to its flexibility and mobility. It extends the user's gaze estimation field from desktop or computer screen into other scenes which greatly rich the collection of human gaze data.

After years of studies that predict the PoR on the scene image plane or the screen, there is an increasing interest in estimating human's gaze location in 3D coordinate nowadays. 3D gaze estimation not only predicts PoR in the field of view with the depth information but also prove the connection between the scene's saliency and human-related motion. To our knowledge, one of the earliest methods that is able to predict 3D PoR requires the subject's head fixed to the camera [4]. Kwon et al. introduce a innovative 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 [5], [6], research on HMD is still limited. For both gaze estimation systems, most approaches determine their 3D gaze points by calculating the midpoint of the shortest segment between both eyes' visual vectors [7]–[9]. However, deviations in the eye gaze vector's calculation are likely causing significant variance in the PoR's depth direction. To address this problem, Ji Woo Lee et al. use multi-layered perception to obtain the depth gaze position [10]. But the method employs dual Purkinje images as the input, which is hard to be detected in practice.

Another challenge for 3D gaze estimation is that the gaze tracker needs to be calibrated for each user before the estimation. During the calibration step, the subject needs to stare at the specific reference marker, yet sometimes such active personal calibration interrupts user-scene interactions. Although the calibration procedure has become much simpler, the number of calibration markers has been reduced

* This work was supported by the National Natural Science Foundation of China under Grant 61873220 and Grant 61875068, the Research Grants Council of Hong Kong under Project CityU 11255716, and the Fundamental Research Funds for the Central Universities under Grant CCNU20ZT017 and Grant CCNU2020ZN008.

[1]Meng Liu and You Fu Li are with the Department of Mechanical Engineering, City University of Hong Kong. Kowloon, Hong Kong. Email:mliu68-c@my.cityu.edu.hk, meyfli@cityu.edu.hk

[2]Hai Liu is with National Engineering Research Center for E-Learning, Central China Normal University. Wuhan, Hubei Province, China. Email:hailiu0204@ccnu.edu.cn

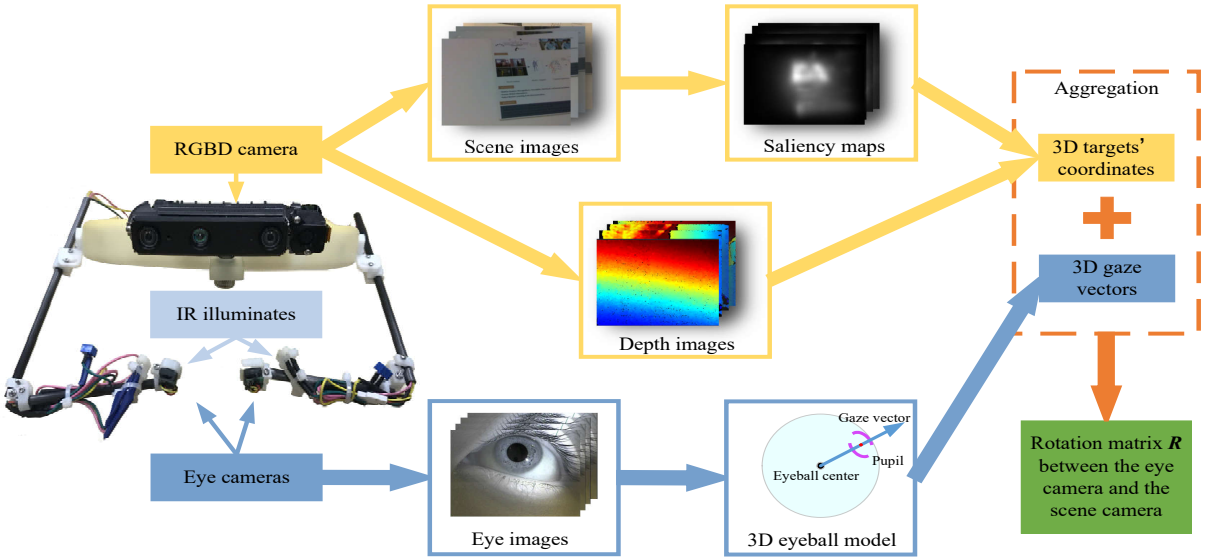

Fig. 1. The calibration procedure in our method

down to one in some works, 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. find out that some subjects' gaze patterns can provide important cues to achieve auto-calibration in [11]. By making use of the topology of the pre-recorded gaze pattern, a mapping function is calculated to transform the initial fixate location to match the subject's gaze pattern. Sugano treats mouse clicks on the computer screen as gaze points to train the mapping function between the eye features and PoR [12]. Similar to [12], the algorithm in [13] detects the user's hand and fingertip 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 actual PoR in real world [14]. All these approaches have tried to avoid using calibration markers. Nonetheless, they all rely on observations of a specific person or environment, which limits their applicability. Apart from these studies, visual saliency also has been taken into consideration for the gaze tracker's auto-calibration. As there are experiments show the correlation between bottom-up saliency and gaze points' position [15]. Several works have applied the visual saliency to calibrate their gaze trackers [16]–[18], yet they are all designed for remote devices.

In this paper, we propose a novel method to estimate 3D fixation for HMD, which combines with the saliency-based algorithm thereby achieving auto-calibration without pre-setting markers and any external assistance. For HMD's architecture, we replace the regular RGB camera with an RGBD camera, which provides accurate 3D data in the scene. During the usage, participants can change their location and head pose with no constraint. In the calibration section, we utilize a salience algorithm to generate saliency maps of scene images. By merging saliency maps and gaze vectors, we can determine two rotation matrices that convert gaze vectors from both eye camera coordinates into the scene camera coordinate. To the best of our knowledge, this is the first work to apply the saliency-based method into automatic calibration for HMD's 3D gaze estimation. In the gaze estimation section, we use the scene image and its depth data to generate a point cloud of the scene, and the PoR is obtained as the point from the point cloud which is closest to both visual vectors.

The rest of this paper is organized as follows. The details of our algorithm are presented in section II. The experiment result and evaluation are shown in Section III. Section IV gives the conclusion of this paper.

II. METHODOLOGY

A. Architecture for HMD

The head-mounted gaze tracker 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 10m. Two IR cameras are leveraged as eye cameras to capture IR eye streams. 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 pixel's 3D position can be obtained in the scene camera coordinates.

Fig. 2. Examples of scene images and their saliency maps.

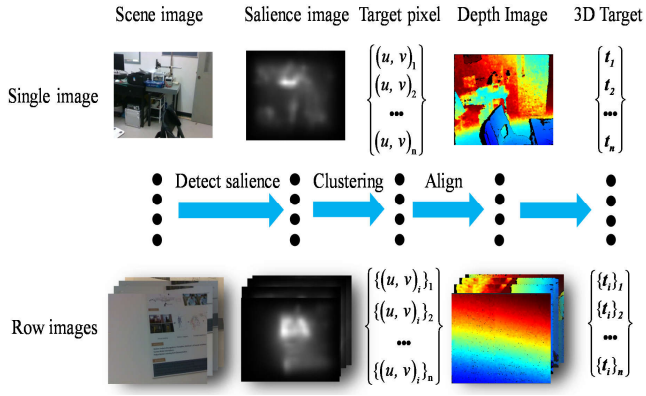

Fig. 3. The process of obtaining calibration targets' dataset. The upper row shows the operation on the single scene image and its corresponding depth image. The lower row demonstrates the operation on a set of scene images and their corresponding depth images.

B. Calibration Procedure

The calibration procedure is shown in Fig. 1. In the process of calibration, the participant can scan the surrounding environment randomly. For RGB scene images, we apply the algorithm in [19] to generate saliency maps, which represent the distinctive features of scene images. Fig. 2 presents some examples of scene images and their corresponding saliency maps. After the generation of each saliency map, we pick out pixels $\left\{\{(\boldsymbol{u},\boldsymbol{v})_i\}_j\right\}$ whose saliency value higher than the pre-setting threshold. Then we employ a clustering method based on [20] to remove noise and improve reliability of the saliency map. Consequently, the target data set can be collected $\left\{\{\boldsymbol{t_i}\}_j\right\}$ for all the 3D coordinates of the space points that correspond to the pixels we choose. $\{\boldsymbol{t_i}\}$ 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. 3.

There are generally two types of methods to 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 firstly use extracted eye features to build the eyeball model. Once the model is built, the initial gaze vectors would be obtained, and it 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 [21]. 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. Then a 3D eye model can be built based on the multiple reprojected pupil contours in 3D space. Assuming that every space circle would be tangent to the ball at the center of the circle, we can find a set of circles' normal passing through their centers. The point that closest to each normal is determined as the eyeball center. Once the eyeball center is recovered, 3D gaze vectors can be calculated in the eye camera coordinate system. Let $\boldsymbol{n_{lj}}$, $\boldsymbol{n_{rj}}$ represent the gaze vector of j^{th} left and right eye image respectively.

Since each user has the specific gaze habit in the 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 robust method to find relationships among the saliency map and two gaze vectors.

Considering short Euclidean distances between the user's eyes and the scene camera are within a few centimeters, while PoR locate at much longer distances in practice, it is reasonable to assume that the user's eyeball center coincides with the scene camera coordinate system's origin. 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, the target data set is acquired as $\{\boldsymbol{t_i}\}_j$ for j^{th} saliency map and the corresponding gaze vector $\boldsymbol{n_{lj}}$. Each target vector can be represented as $\boldsymbol{t_i} - \boldsymbol{e_l}$ where $\boldsymbol{e_l}$ stands for the user's left eyeball center location in the scene camera's coordinate system. We apply a pair of angle α_{li} and β_{li} to represent the vertical and horizontal angle between the gaze and target vectors, respectively. Then the specific rotation matrix $\boldsymbol{R_{li}}$ can be represented as

$$\boldsymbol{R_{li}} = \begin{bmatrix} \cos\beta_{li} & \sin\beta_{li}\sin\alpha_{li} & \sin\beta_{li}\cos\alpha_{li} \\ 0 & \cos\alpha_{li} & -\sin\alpha_{li} \\ -\sin\beta_{li} & \cos\beta_{li}\sin\alpha_{li} & \cos\beta_{li}\cos\alpha_{li} \end{bmatrix} \quad (1)$$

We obtain the rotation matrix set $\{\boldsymbol{R_{li}}\}$ through

$$\boldsymbol{R_{li}}\boldsymbol{n_{lj}} = \frac{\boldsymbol{t_i} - \boldsymbol{e_l}}{|\boldsymbol{t_i} - \boldsymbol{e_l}|} \quad (2)$$

Similarly, we can further acquire a set of rotation matrices $\left\{\{\boldsymbol{R_{li}}\}_j\right\}$ that corresponds to all the gaze vectors and all selected pixels in all the saliency maps. During the calculation, a two-dimensional cumulative array $A(\alpha_{li}, \beta_{li})$ is created to count the frequencies of angle pairs (as shown in Fig. 4). We can see from Fig. 4 that the distribution of angle pairs' frequencies keeps changing as the number of obtained saliency maps increasing. As we assume that people tend to look at salience part in the scene, the unique

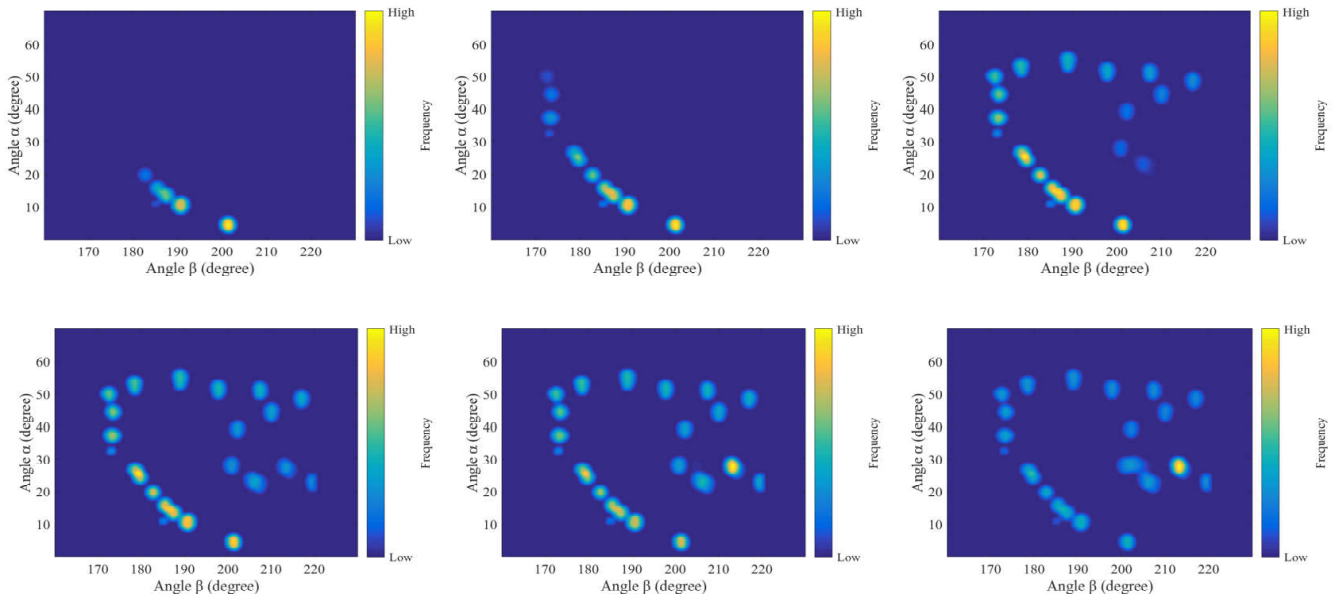

Fig. 4. The most frequent angle set changes as the number of scene saliency maps increases. From (a) to (f), the number of scene saliency maps are: 1, 10, 50, 100, 200, all the saliency maps.

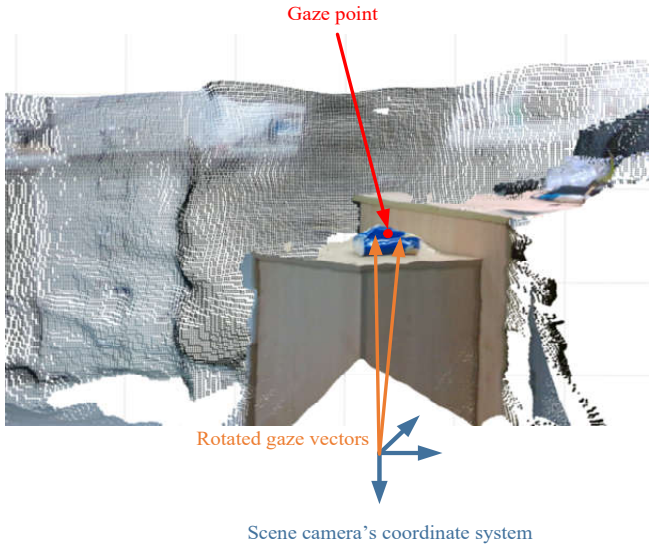

Fig. 5. The point cloud of the environment. The gaze point is defined as the point with the minimal distance to both rotated gaze vectors.

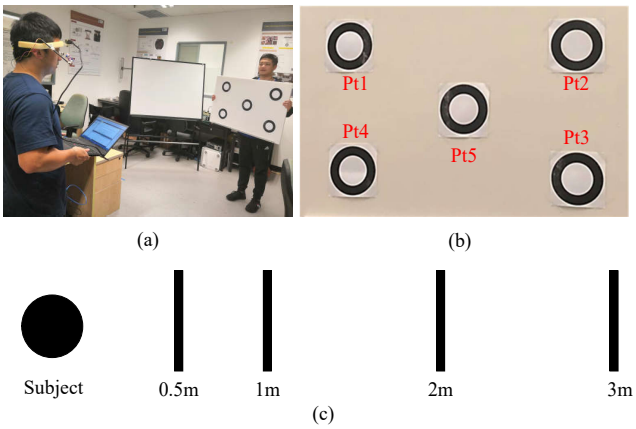

Fig. 6. The setup of our experiment indoors. (a) The gaze estimation accuracy test in the office. (b) 5 concentric circle targets on a board. (c) The 4 different depth that the board is put in this setup.

relationship between gaze vectors and targets i.e. the most frequent angle pair can be determined once we collected enough calibration data (like (f) in Fig. 4). We determine the final angle pair when its maximum frequency is greater than 2 times of the second largest frequency. After the calculation, the most frequent $\boldsymbol{R_{li}}$ is leveraged to restore the extrinsic parameters $\boldsymbol{R_l}$ between the left eye camera and the scene camera coordinate systems.

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 $\boldsymbol{R_r}$.

The two rotated gaze vectors are shown as follows:

$$\boldsymbol{V_{li}} = \boldsymbol{e_l} + \boldsymbol{R_l n_{lj}} \tag{3}$$

$$\boldsymbol{V_{ri}} = \boldsymbol{e_r} + \boldsymbol{R_r n_{rj}} \tag{4}$$

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 the human eyes, small angle calculation errors in rotated gaze vectors can cause large deviations in the Z direction of the gaze point estimation. To refine the raw 3D gaze estimate method, we generate a point cloud of the environment (as shown in Fig. 5). For two

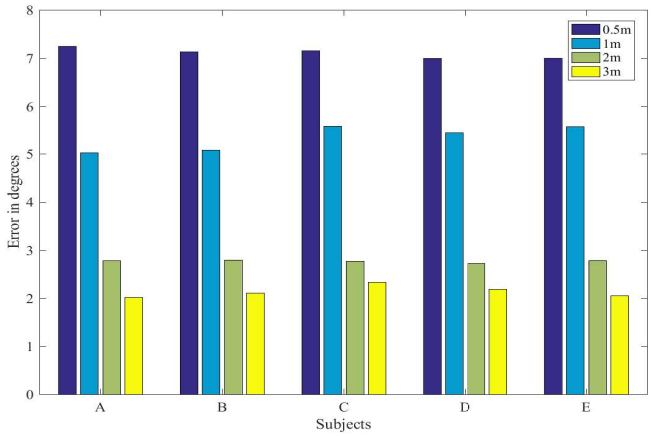

Fig. 7. The mean PoR angular estimation errors of 5 subjects over 4 different test distances.

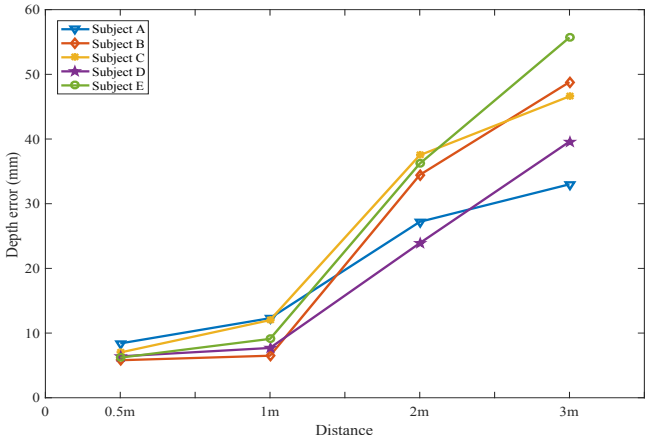

Fig. 8. The mean PoR depth estimation errors of 5 subjects over 4 different test distances.

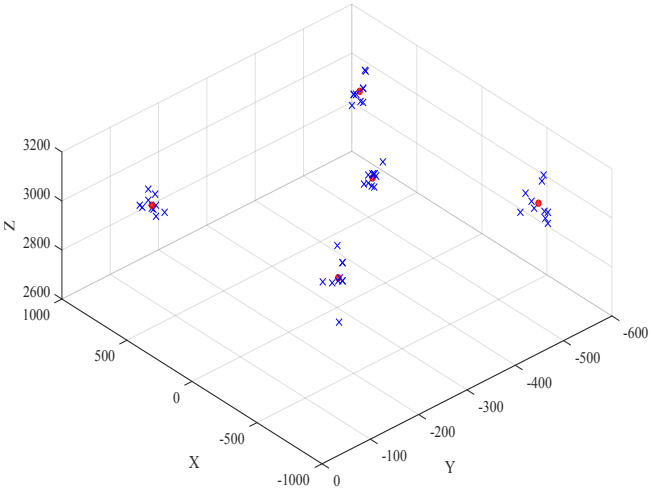

Fig. 9. Visualization of 3D estimated gaze points from subject A at 3m. Red dots indicate the ground truth, while blue crosses represent the estimated PoR.

TABLE I
POR ESTIMATION ERROR

Subject	ADE(mm)	AAE(deg)
A	20.23	4.27
B	23.89	4.28
C	25.78	4.46
D	19.40	4.34
E	26.80	4.36
Average	**23.22**	**4.34**

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

$$\boldsymbol{d} = \frac{|\boldsymbol{V_{li}} \times (\boldsymbol{p_i} - \boldsymbol{e_l})|}{|\boldsymbol{V_{li}}|} + \frac{|\boldsymbol{V_{ri}} \times (\boldsymbol{p_i} - \boldsymbol{e_r})|}{|\boldsymbol{V_{ri}}|} \quad (5)$$

where $\boldsymbol{p_i}$ is the point in the point cloud and $\boldsymbol{d}$ represents the sum distance from the point to two rotated gaze vectors.

III. EXPERIMENT AND EVALUATION

Our experimental system is based on a low-cost HMD developed by our lab, 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 achieved by using hybrid programming in C++ and MATLAB on a laptop PC with an Intel i7 2.70GHz, 2.90GHz quad core CPU and an 8.00GB RAM.

Fig. 6 demonstrates the setup of our indoors experiment. We have invited 5 subjects to evaluate the accuracy of our method. During the calibration step, subjects are allowed to look randomly at the surroundings with no limitation on their head poses and standing positions. The whole calibration process averagely takes 20 seconds, as the algorithm costs some 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 our method. Taken 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 depth from the subject: 0.5 m, 1 m, 2 m, and 3 m. For each depth, subjects are told to fixate at the 5 targets in the same order: top-left, top-right, bottom-right, bottom-left, center, each target for 2 to 3 seconds.

The experiment result is shown in table I, ADE is the average depth error, and AAE represents the average angular error. The overall average depth error is 23.22mm and the average angular error is 4.34 degree. Fig. 7 illustrates the PoR estimation errors in degrees of 5 subjects at 4 different distances while the depth estimation error is shown in Fig.

TABLE II
COMPARISON WITH STATE-OF-THE-ART 3D GAZE ESTIMATION METHODS

Method	AAE	ADE
[22]	4.9°	194mm for the test distance range from 0.75m to 2.75m
[23]	6.0°	110mm for the 1m test distance
[24]	5.27°	-
[25]	1°	80mm for the distance of 1m; 500mm for the distance of 6m
Ours	**4.34°**	**23.22mm for the test distance range from 0.5m to 3m**

8. And the distribution of 3D PoR from a subject (subject A) with the test distance of 3m is presented in Fig. 9.

We can see from Fig. 7 that all 5 subjects give relatively steady performance in the experiment, and their average angular error decreases when the test distance broadens from 0.5m to 3m, we think this phenomenon may due to our assumption that ignores the Euclidean distance between the user's eye and the scene camera. When the test depth is relatively short, this eye-camera distance can cause the deviation on the calculation of rotation matrices. However, with the increase of the estimation field, it will bring less effect on the result.

In Table II, we compare the proposed method with head-mounted-based gaze estimation methods. In contrast to algorithms in [22], [23] and [24], our gaze system achieves better accuracy on average angular error. Besides, we have a significant improvement in depth estimation. The average angular error of [25] is indeed smaller. However, their approach requires complex hardware setups and time-consuming calibration period, and our method performs better in the depth estimation.

IV. CONCLUSIONS

In this paper, we propose a novel 3D gaze estimation framework with automatic calibration method. For the hardware designing, we replace the regular RGB camera of head-mounted gaze tracker with an RGBD camera to obtain the depth information of the scene. During the calibration procedure, the saliency algorithm is introduced to find salient parts in scene images. We align those pixels to the depth image and use them as 3D calibration targets. Through the aggregation, we obtain the rotation matrix between the eye camera coordinate system and the scene camera coordinate system. To improve the accuracy of depth estimation, environmental point cloud data is applied in the PoR's estimation. Once we identified the final gaze vector, the PoR is calculated as it is the closest point to the final gaze vector. Based on 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] D. W. Hansen and Q. Ji, "In the eye of the beholder: A survey of models for eyes and gaze," *IEEE. Trans. Pattern. Anal. Mach. Intell*., vol.32, no.3, pp.478-500, 2010.

[2] E. D. Guestrin and M. Eizenman, "General theory of remote gaze estimation using the pupil center and corneal reflections," *IEEE Trans. Biomed. Eng*., vol.53, no.6, pp.1124-1133, 2006.

[3] S. Park, et al, "Learning to find eye region landmarks for remote gaze estimation in unconstrained settings," *In Proc. ETRA*., 2018, ACM.

[4] Y. M. Kwon, et al, "3d gaze estimation and interaction to stereo dispaly," *IJVR*., vol. 5, no. 3, pp. 41-45, 2006.

[5] P. M. Tostado, W. W. Abbott, and A. A. Faisal, "3d gaze cursor: Continuous calibration and end-point grasp control of robotic actuators," *In Proc. ICRA*., 2016, IEEE.

[6] S. Mujahidin, et al, "3d gaze tracking in real world environment using orthographic projection," *In Proc. AIP*., 2016.

[7] C. Hennessey and P. Lawrence, "Noncontact binocular eye-gaze tracking for point-of-gaze estimation in three dimensions," *IEEE Trans. Biomed. Eng*., vol.56, no.3, pp.790-799, 2008.

[8] S. Li, X. Zhang, and J. D. Webb, "3-d-gaze-based robotic grasping through mimicking human visuomotor function for people with motion impairments," *IEEE Trans. Biomed. Eng*., vol.64, no.12, pp.2824-2835, 2017.

[9] F. Pirri, M. Pizzoli, and A. Rudi, "A general method for the point of regard estimation in 3d space," *In Proc. CVPR*., 2011, IEEE.

[10] J. W. Lee, et al, "3d gaze tracking method using purkinje images on eye optical model and pupil," *Optics and Lasers in Engineering*., vol.50, no.5, pp.736-751, 2012.

[11] F. Alnajar, et al, "Auto-calibrated gaze estimation using human gaze patterns," *Int. J. Computer Vision*., vol.124, no.2, pp.223-236, 2017.

[12] Y. Sugano, et al, "Appearance-based gaze estimation with online calibration from mouse operations," *IEEE Trans. Human-Machine Systems*, vol.45, no.6, pp.750-760, 2015.

[13] M. Bâce, S. Staal, and G. Sörös, "Wearable eye tracker calibration at your fingertips," *In Proc. ETRA*., 2018, ACM.

[14] C. Lander, et al, "Using corneal imaging for measuring a human?s visual attention," *In Proc. UbiComp*., 2017, ACM.

[15] D. Parkhurst, K. Law, and E. Niebur, "Modeling the role of salience in the allocation of overt visual attention," *Vision research*., vol.42, no.1, pp.107-123, 2002.

[16] Y. Sugano, Y. Matsushita, and Y. Sato, "Appearance-based gaze estimation using visual saliency," *IEEE. Trans. Pattern. Anal. Mach. Intell*., vol.35, no.2, pp.329-341, 2012.

[17] J. Chen and Q. Ji, "A probabilistic approach to online eye gaze tracking without explicit personal calibration," *IEEE. Trans. Image. Process*., vol.24, no.3, pp.1076-1086, 2015.

[18] M. Hiroe, M. Yamamoto, and T. Nagamatsu, "Implicit user calibration for gaze-tracking systems using an averaged saliency map around the optical axis of the eye," *In Proc. ETRA*., 2018, ACM.

[19] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," *In Proc. NIPS*., 2007.

[20] M. Ester, et al, "A density-based algorithm for discovering clusters in large spatial databases with noise," *in Kdd*., vol.96, no.34, pp.226?231, 1996.

[21] L. Swirski and N. Dodgson, "A fully-automatic, temporal approach to single camera, glint-free 3d eye model fitting," *Proc. PETMEI*., 2013.

[22] C. Elmadjian, et al, "3d gaze estimation in the scene volume with a head-mounted eye tracker," *In Proc. COGAIN*., 2018, ACM.

[23] Y. Sugano and A. Bulling, "Self-calibrating head-mounted eye trackers using egocentric visual saliency," *In Proc. UIST*., 2015, ACM.

[24] K. Takemura, et al, "Estimating 3-d point-of-regard in a real environment using a head-mounted eye-tracking system," *IEEE. Trans. Human-Machine Systems*., vol.44, no.4, pp.531-536, 2014.

[25] M. Weier, et al, "Predicting the gaze depth in head-mounted displays using multiple feature regression," *In Proc. ETRA*., 2018, ACM.