DEPTH FROM GAZE

Tzu-Sheng Kuo, Kuang-Tsu Shih, Sheng-Lung Chung, and Homer H. Chen

National Taiwan University

ABSTRACT

Eye trackers are found on various electronic devices. In this paper, we propose to exploit the gaze information acquired by an eye tracker for depth estimation. The data collected from the eye tracker in a fixation interval are used to estimate the depth of a gazed object. The proposed method can be used to construct a sparse depth map of an augmented reality space. The resulting depth map can be applied to, for example, controlling the visual information displayed to the viewer. A mathematical model for determining whether two depths in the augmented reality space are statistically distinguishable is also developed. Experimental results show that the proposed method can estimate and distinguish different object depths effectively.

Index Terms—Gaze, depth estimation, eye tracker, human computer interaction, augmented reality.

1. INTRODUCTION

For decades, exploiting human gaze for 2D graphical user interface has been a popular research topic in human computer interaction, gaming, and psychology [1]. Recently, the rising interest in augmented reality (AR) and virtual reality (VR) has further fueled the development of effective means for interaction with a 3D visual world. This paper investigates the estimation of depth in 3D space using an eye tracker. Specifically, the 3D depth of a gazed object with respect to the viewer is estimated using the gaze information obtained from an eye tracker.

The eye tracker is considered because it has become a popular component of see-through devices to track the eye movement of a user and thereby control the presentation of images to the user. In this application scenario, besides observing how the user navigates the visual world, the gaze information can be utilized to compute the 3D depth of each scene point the user looks at during the visual navigation journey. A sparse depth map of the attended visual stimuli can be thus obtained. A critical step towards this goal is the depth estimation of a gazed object.

We believe that it is possible to address the depth estimation problem using gaze information because human eyeballs rotate when gazing objects at different depths. However, it should be noted that the depth of a gazed object in the context of this work is different from the depth of gaze at a particular time instant. The difference is due to the fact that gaze position, which is measured by the intersection of the visual axes of both eyes, varies with time despite the same point is gazed at [2], [3].

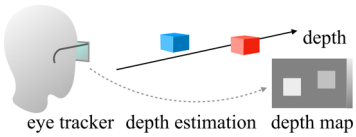


Fig. 1. An application scenario where a sparse depth map is obtained from the gaze information generated by an eye tracker.

In this paper, we propose a depth-from-gaze method applicable to typical indoor interactions ranging from 0.65m to 2m in depth. The estimation is achieved by modeling the effect of temporal variation of gaze as Gaussian noise and by processing the gaze data over a fixation time interval. Furthermore, a Gaussian model is developed to determine the minimal distance between two statistically distinguishable depths acquired by an eye tracker.

2. RELATED WORK

2.1. Vergence

Humans are capable of perceiving the 3D relative depth. Although the exact mechanism of human depth perception is not yet fully understood, the depth perception models developed so far can be classified into two categories: 1) special model for near-distance viewing and 2) general model for both near- and far-distance viewing [4].

Vergence is a near-distance depth perception model that describes how the eyeballs rotate inward (outward) when gazing at a near (far) object. The inward and outward eye movements, respectively, are called convergence and divergence. The variation of vergence is smaller when the point of gaze is at a relatively far distance [4], [5]. Therefore, the minimal distance between two distinguishable depths increases as the point of gaze moves away from the viewer.

On the other hand, vergence does not correspond to the exact distance of the object being gazed at. Research has shown that the depth of the intersection of visual axes may not be equal to the depth of the point of gaze [2]. Furthermore, fixation eye-movements may introduce variation to the visual axes during fixation [3]. Therefore, it is inappropriate to assume that the depth of gaze is equal to the depth of the gazed object.

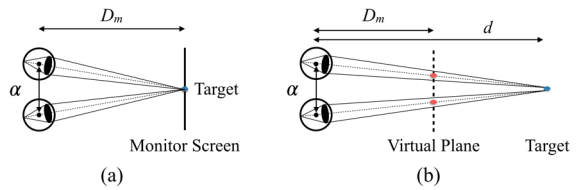


Fig. 2. Top view of (a) a typical 2D gaze estimation and (b) a 3D gaze estimation.

2.2. Eye Trackers and Gaze Estimation

An eye tracker is a device for measuring eye positions and movements. Commercially available 2D eye trackers come in the form of desktop or head-mounted devices [6]. For most desktop eye trackers, the fiducials to be gazed at are displayed on a monitor screen for calibration. For head-mounted eye trackers, the fiducials are placed at or moved to various positions on a 2D plane in space. Both types of eye trackers use the eye images captured in the calibration stage to estimate the 2D gaze in the evaluation stage.

3D gaze estimation methods have been developed for stereoscopic displays. A method called "parameterized self-organizing map" employs neural networks to construct a function that maps 3D gaze position to 2D on-screen gaze position in the calibration stage. The inverse of the function is then used for 3D gaze estimation in the evaluation stage [7]. Another method enhances gaze estimation by a 3D calibration process, which maps coarse estimations to gaze positions at a finer granularity [8].

3D gaze estimation methods have also been developed for applications in a real-world environment. A method using an eye tracker together with a body tracking system demonstrates that it is possible to visualize gaze positions with 3D scan paths and attention volumes [9]. Methods such as spatial triangulation or parameterized self-organizing map [7] that work for stereoscopic display can also be applied to real-world environment to achieve 3D gaze estimation [10], [11]; however, additional work is needed to overcome the limitations on working distance and portability.

3. PROPOSED METHOD

Unlike the methods discussed in Sec. 2 that target gaze estimation, we focus on the depth estimation of gazed object. Note that gaze estimation is an intermediate step of our method rather than an entire process. In this section, we describe how to obtain reliable gaze information from the eye tracker, how this information is used to estimate the depth of object, and how to estimate if two objects are distinguishable in depth judged from the gaze information.

3.1. Gaze Information from the Eye Tracker

We explain our 3D gaze estimation method using an illustration of the configuration of a desktop eye tracker shown in Fig. 2. For typical 2D gaze estimation shown in (a), both the fiducials for calibration and the targets for evaluation are displayed on a monitor, and the gaze point is simply at the intersection of the visual axes with the monitor screen. For 3D gaze estimation, the configuration for calibration is still the same, but the configuration for evaluation is different because the target is no longer on the monitor screen. Imagine that the monitor is removed and replaced by a virtual plane, as shown in Fig. 2 (b). The visual axes would intersect the virtual plane at two points. These are the two points obtained from a binocular eye tracker.

The following derivation holds regardless of which type of eye tracker is used. Consider the two intersection points described above and denote them by (x_l, y_l) and (x_r, y_r) corresponding to left and right eyes, respectively. Let $\Delta x = x_r - x_l$ and denote the depth of gaze by d. By triangular similarity, we have

$$d = \frac{\alpha D_m}{\alpha - \Delta x}, \tag{1}$$

where α is the interocular distance and D_m is the depth of the virtual plane. In practice, D_m is a known control parameter, but α varies across individuals. For simplicity, however, it is common to set $\alpha = 6.3$ cm [8], [10]. A notable property of human gaze is that the intersection of visual axes varies during fixation despite the same point is gazed at. As a result, the depth of gaze obtained from (1) varies with time.

Estimation of the depth of gaze can be enhanced by substituting Δx_n for Δx in (1), where $\Delta x_n = \Delta x - \Delta x_{D_m}$ and Δx_{D_m} denotes the mean of Δx's that are obtained when the target is located at depth D_m. This operation can be considered to be a normalization of Δx because ideally the visual axes would intersect at the virtual plane ($\Delta x = 0$) when gazing the target at depth D_m.

3.2. Depth of Target from Gaze Information

In this subsection, we describe how the depth of a target is obtained by exploiting the relation between d and Δx in (1). The main idea of our method is to calculate the probability of a target being at a possible depth, then the depth with the highest probability is selected as the final depth estimate. This process can be done by using a set of Δx's collected over a time interval, which can be set to the fixation duration in different viewing scenarios.

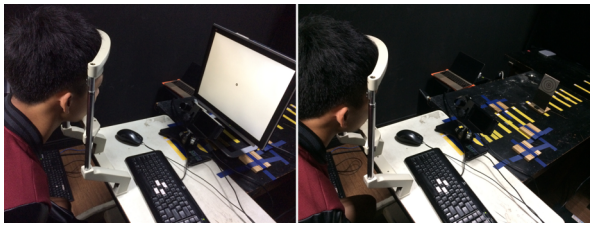
We now explain how Δx's can be used to estimate the depth of the target. Let us denote the set of estimates of the depth of gaze obtained from (1) by D, the probability of a target at depth h given D by $P(h|D)$, and the set of all possible depths of the target by H. The depth estimate for target Z is obtained by

$$Z = \underset{h \in H}{\operatorname{argmax}}\, P(h|D). \tag{2}$$

From the Bayes' theorem, $P(h|D)$ can be expressed by

$$P(h|D) = P(D|h)\frac{P(h)}{P(D)}. \tag{3}$$

Under the common uniform prior assumption (that is, $P(h)$ is the same for a different depth) and with $P(D)$ being a fixed normalization factor, we can formulate the depth estimation problem as a maximum-likelihood problem,



(a) (b)

Fig. 3. Experimental setup for (a) calibration and (b) evaluation.

$$Z = \underset{h\in H}{\operatorname{argmax}} P(D|h). \tag{4}$$

Assuming the elements in D are i.i.d. yields

$$P(D|h) = \prod_{d\in D} P(d|h). \tag{5}$$

Since d and Δx in (1) bear a one-to-one relationship, we can rewrite (5) as follows:

$$P(D|h) = \prod_{d\in D} P(\Delta x_d|\Delta x_h), \tag{6}$$

where Δx_d and Δx_h denote the lateral distance (Δx) between the intersections of the visual axes with the virtual plane when the target is located at d and h, respectively. Substituting (6) into (4) for $P(D|h)$ and taking logarithm yield

$$Z = \underset{h\in H}{\operatorname{argmax}} \sum_{d\in D} \ln P(\Delta x_d|\Delta x_h). \tag{7}$$

Using a normal distribution with standard deviation σ to approximate $P(\Delta x_d|\Delta x_h)$, we rewrite (7) as follows:

$$\begin{aligned} Z &\approx \underset{h\in H}{\operatorname{argmax}} \sum_{d\in D} \ln \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(\Delta x_d - \Delta x_h)^2}, \\ &= \underset{h\in H}{\operatorname{argmin}} \sum_{d\in D} (\Delta x_d - \Delta x_h)^2. \end{aligned} \tag{8}$$

As a result, depth estimation is formulated as a mean squared error problem. The depth estimate of the target can be obtained by searching for the depth h in H that minimizes the sum of squared errors with respect to Δx_d.

3.3. Minimal Distance in Depth

Consider the case where two objects are 5 cm apart in depth. We can easily tell which of them is closer to us when the objects are 20 cm away. However, it is hard to do so when they are 200 cm away. This phenomenon has to do with the fact that the convergent sensitivity of human eyes decreases with object distance. In the context of this work, the minimal distance between two distinguishable depths increases as the point of gaze moves away from the viewer. We propose a model to estimate the minimal distance for different depths.

Consider three independent normal distributions N_1, N_2, and N_3 with the same standard deviation σ but different means Δx_1, Δx_2, and Δx_3. Let f_1, f_2, and f_3 denote their probability density functions. Assume that Δx_i is an arithmetic progression and $\Delta x_i < \Delta x_{i+1}$. Then, the probability P of an x sampled from N_2 such that $f_2(x) < f_1(x)$ or $f_2(x) < f_3(x)$ is

$$P = 2\big(1 - F_2(\Delta x_2 + I\,/\,2)\big), \tag{9}$$

where F_2 is the cumulative distribution function of f_2 and I is the common difference of Δx's. We determine the value of I so that the probability P is below a tolerable threshold and obtain iteratively a series of Δx's given the initial value Δx_0. In this way, the minimal distance between distinguishable depths is determined by the difference between d's, where d's are obtained from Δx's using (1).

4. EXPERIMENT

4.1. Apparatus

The experimental setup is shown in Fig. 3. A chin rest was used to fix the head position of each viewer participating in the experiment. An EyeLink binocular eye tracker with sample rate 1000 Hz was placed at 45 cm away from the viewer. A ViewSonic VX912 monitor screen with resolution 1024×768 was placed at 65 cm away from the viewer in the calibration stage. Identical targets were placed at 65, 72, 85, 106, 138, and 200 cm away from the viewer for evaluation. The depths of the targets were determined from (9) with $P = 0.05$ and $\sigma^2 = 40$ pixels. Our program was implemented in MATLAB using the Psychophysics Toolbox [12].

4.2. Procedure

As illustrated in Fig. 3 (a), each viewer was asked to gaze at the fiducials displayed on the monitor screen in the calibration stage. Then, the monitor was removed after the calibration was completed and the viewer was asked to gaze at each target for 5 seconds, as show in Fig. 3 (b). At each target position, 5000 pairs of (x_l, y_l) and (x_r, y_r) were obtained from the eye tracker. Those labeled "saccade" or "blink" were removed. Three males and four females were recruited to participate in the experiment. Each subject went through three trials, each including a calibration and evaluation procedure.

4.3. Results and Discussions

Results were obtained with H = {65, 72, 85, 106, 138, 200} for the seven subjects (denoted by S_i, i=1..7), each of them went through three trials (denoted by T_j, j=1..3). We show the estimated gaze position of two participants in Fig. 4. The results for the other participants are similar but omitted from the figure due to page limitation. Two important observations are made. First, Δx_n becomes larger as the target is placed farther. This proves that the eye tracker, which was originally designed for 2D gaze tracking, can be used to obtain the vergence information required for 3D gaze estimation. Second, the increment of Δx_n decreases with the target depth. For example, we can see that the difference between Δx_n's for the targets at 65 and 85 cm are larger than that for the targets at 180 and 200 cm. This result is consistent with our earlier statement in Sec. 3 that the minimal distance between two distinguishable depths increases as the point of gaze moves away from the viewer.

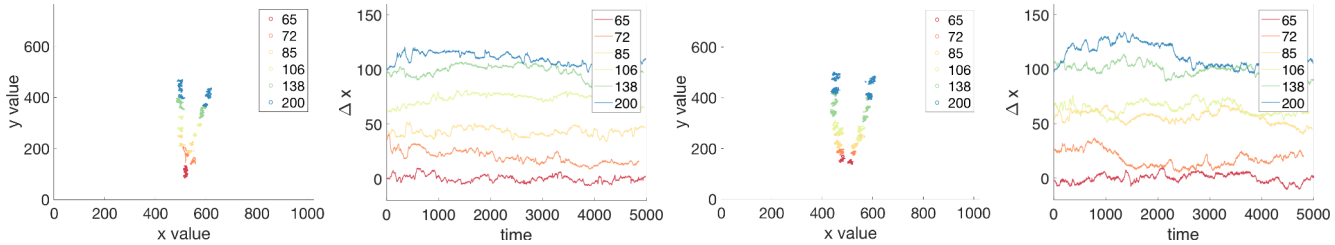


Fig. 4. Example experimental results for different target depths coded by color. The results for Subjects S_1 and S_4 are shown in (a)–(b) and (c)–(d), respectively. (a) and (c) show the scatter diagram of the gaze positions for both eyes, the left branch for (x_l, y_l) and the right branch for (x_r, y_r). (b) and (d) show the value of Δx_n over time (ms).

Table 1. Evaluation of depth estimates (with $\sigma^2 = 40$ in (9))

●: 65 cm; ●: 72 cm; ●: 85 cm; ●: 106 cm; ●: 138 cm; ●: 200 cm.

subject	trial	Proposed Method: ground truth 65	72	85	106	138	200	quality of estimation: best	2nd best	others	Baseline Method: ground truth 65	72	85	106	138	200	quality of estimation: best	2nd best	others
S_1	T_1	●	●	●	●	●	●	12	0	0	●	●	●	●	●	●	10	2	0
	T_2	●	●	●	●	●	●				●	●	●	●	●	●			
S_2	T_1	●	●	●	●	●	●	13	5	0	●	●	●	●	●	●	3	3	12
	T_2	●	●	●	●	●	●				●	●	●	●	●	●			
	T_3	●	●	●	●	●	●				●	●	●	●	●	●			
S_3	T_1	●	●	●	●	●	●	7	11	0	●	●	●	●	●	●	3	5	10
	T_2	●	●	●	●	●	●				●	●	●	●	●	●			
	T_3	●	●	●	●	●	●				●	●	●	●	●	●			
S_4	T_1	●	●	●	●	●	●	8	3	1	●	●	●	●	●	●	2	4	6
	T_2	●	●	●	●	●	●				●	●	●	●	●	●			
S_5	T_1	●	●	●	●	●	●	11	6	1	●	●	●	●	●	●	3	3	12
	T_2	●	●	●	●	●	●				●	●	●	●	●	●			
	T_3	●	●	●	●	●	●				●	●	●	●	●	●			
S_6	T_1	●	●	●	●	●	●	5	7	0	●	●	●	●	●	●	2	2	8
	T_2	●	●	●	●	●	●				●	●	●	●	●	●			
S_7	T_1	●	●	●	●	●	●	6	5	1	●	●	●	●	●	●	5	5	2
	T_2	●	●	●	●	●	●				●	●	●	●	●	●			
								61%	36%	3%							27%	24%	49%

Table 1 shows the depth estimates obtained from the results in Fig. 4. For evaluation purpose, the depth estimates are classified into different levels coded by color. Note that, in some trials (e.g., T_3 of S_1), the eye tracker misidentified nostrils as pupils. Such trials are excluded from the table. Three ratings are given to the depth estimates: best estimate, second-best estimate, and others. A best estimate means a depth estimate Z is at the ground truth level, a second-best estimate means Z is one depth level off the ground truth (for example, the estimate is 72 or 106 for a target at 85), and others means Z is more than one level off the ground truth.

Three observations are made from the results shown in Table 1. First, the depth estimate increases with the target depth for most of the trials. That is, the proposed method is able to tell the relative depth of a target in different positions. The ability to judge relative depth is useful for many applications. For example, it allows foreground object to be separated from background. Second, the best and the second-best estimates together account for 97% of the cases, which implies that it is feasible that the coarse depth estimate obtained from an eye tracker allows us to find the viewer's point of regard. This is useful for applications such as AR and makes depth-from-gaze a complementary function of a head-mounted see-through device for controlling and processing the visual information to be displayed to the viewer. Third, the quality of depth estimate based on gaze information varies with individuals. For example, all depth estimates for Subject S_1 fall in the best estimate category, but only 67% of the depth estimates belong to this category for S_4. This may have to do with the fact that the amplitude of fixation eye-movement is different between individuals, as we can see from the scatter diagrams shown in Figs. 4 (a) and (c) that the points for S_1 are more densely packed than those for S_4. Consequently, the value of σ in (9) has to be customized for each individual.

Table 2. Evaluation of depth estimates (with $\sigma^2 = 95$ in (9))

●: 65 cm; ●: 85 cm; ●: 120 cm; ●: 180 cm.

subject	trial	ground truth 65	85	120	180	quality of estimation: best	2nd best	others
S_1	T_1	●	●	●	●	8	0	0
	T_2	●	●	●	●			
S_4	T_1	●	●	●	●	7	1	0
	T_2	●	●	●	●			

To show the effectiveness of the proposed method, we compare the proposed method against a baseline method that outputs the average depth of gaze. Specifically, it uses (1) to compute the depth of gaze d and averages the resulting d values obtained over a period of time. The average depth estimate is then classified into the nearest depth level. From the experimental results shown in Table 1, we can see that the best and the second-best estimates of the baseline method account for only 27% and 24%, respectively, of the cases. The results clearly show that the proposed method is more accurate than the baseline method.

To show the importance of customization, we performed an additional experiment with $\sigma^2 = 95$ (as opposed to 40 in the first experiment) for S_1 and S_4. The results are show in Table 2. We can see that 88% of the depth estimates now fall in the best estimate category for S_4. (Of course, it is expected that the percentage for S_1 should remain the same since the distance between depth levels is increased). The granularity of depth estimation discussed so far is based on the gaze information. It is interesting to study further how the actual depth perception of a person is related to such granularity.

5. CONCLUSION

In this paper, we have demonstrated that gaze information can be used for depth estimation. We have also described a model to determine the minimal distance between distinguishable depths. The proposed method can be integrated into domain-specific solutions to construct a sparse depth map during the visual navigation journey of a viewer. We believe such 3D sensing capability is useful for applications to psychology, gaming, and human computer interaction.

6. REFERENCES

[1] A.T. Duchowski, *Eye Tracking Methodology: Theory and Practice*. Switzerland: Springer International Publishing AG, pp. 249–338, 2017.

[2] W. Jaschinski, "Fixation disparity and accommodation as a function of viewing distance and prism load," *Ophthalmic and Physiological Optics*, vol. 17, pp. 324–339, 1997.

[3] J. Otero-Millan, S. L. Macknik, and S. Martinez-Conde, "Fixation eye movements in binocular vision," *Frontiers in Integrative Neuroscience*, vol. 8, Jul. 2014.

[4] M.R. Watson and J.T. Enns, "Depth Perception," in *Encyclopedia of Human Behavior*, 2nd ed. V.S. Ramachandran, Ed. Cambridge: Academic Press, pp. 690–696, 2012.

[5] I.P. Howard, *Perceiving in Depth*. New York: Oxford University Press, 2012.

[6] 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, Aug. 2017.

[7] K. Essig, M. Pomplun, and H. Ritter, "A neural network for 3D gaze recording with binocular eye trackers," *The International Journal of Parallel, Emergent and Distributed Systems*, vol. 21, pp. 79–95, 2006.

[8] R.I. Wang, B. Pelfrey, A.T. Duchowski, and D.H. House, "Online 3d gaze localization on stereoscopic displays," *ACM Trans. Appl. Percept.*, vol. 11, no. 1, pp. 1–21, Apr. 2014.

[9] T. Pfeiffer, "Measuring and visualizing attention in space with 3D attention volumes," *Proceedings of the Symposium on Eye Tracking Research and Applications*, pp. 29–36, 2012.

[10] A.T. Duchowski, D.H. House, J. Gestring, R. Congdon, L. Swirski, N.A. Dodgson, K. Krejtz, and I. Krejtz, "Comparing estimated gaze depth in virtual and physical environments," *Proceedings of the Symposium on Eye Tracking Research and Applications*, pp. 103–110, 2014.

[11] E.G. Mlot, H. Bahmani, S. Wahl, and E. Kasneci, "3D Gaze Estimation using Eye Vergence," *Proceedings of the 9th International Joint Conference on Biomedical Engineering Systems and Technologies*, pp. 125–131, 2016.

[12] D. H. Brainard, "The psychophysics toolbox," *Spatial Vis.*, vol. 4, no. 4, pp. 433–436, 1997.