DEEP: 3D Gaze Pointing in Virtual Reality Leveraging Eyelid Movement

Xin Yi
Tsinghua University
Beijing, China
yixin@tsinghua.edu.cn

Leping Qiu
Tsinghua University
Beijing, China
qlp19@mails.tsinghua.edu.cn

Wenjing Tang
Southeast University
Nanjing, China
wenjing_tang@seu.edu.cn

Yehan Fan
Beijing University of Posts and Telecommunications
Beijing, China
fanyh@bupt.edu.cn

Hewu Li
Tsinghua University
Beijing, China
lihewu@cernet.edu.cn

Yuanchun Shi
Tsinghua University
Beijing, China
shiyc@tsinghua.edu.cn

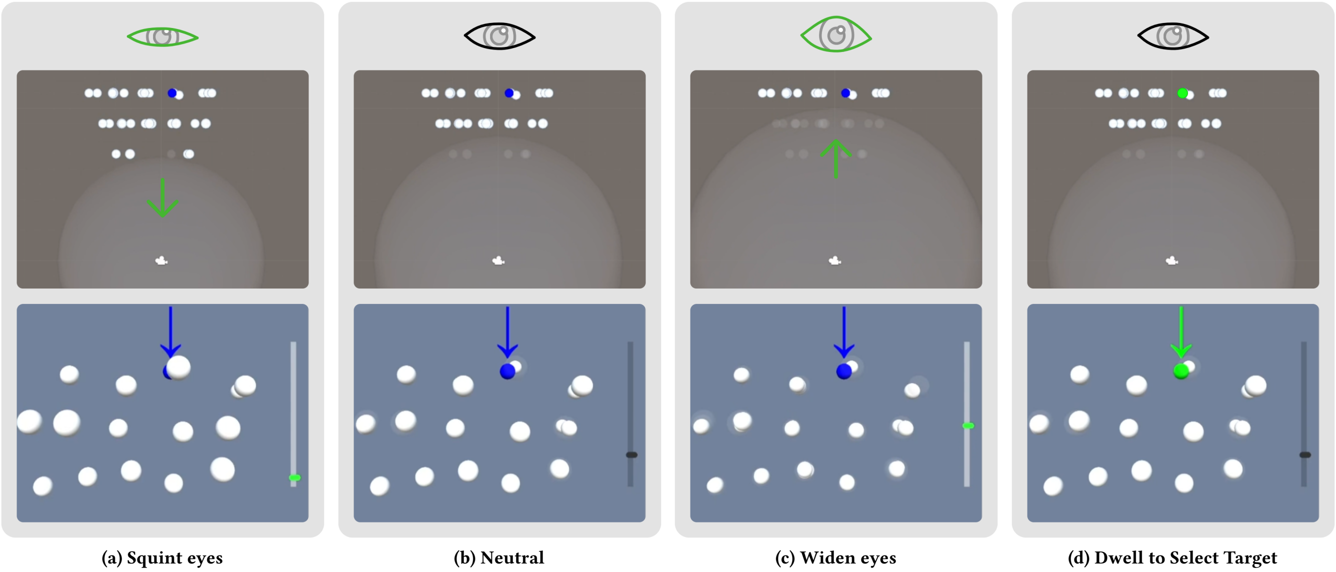


Figure 1: *DEEP* allows users to widen and squint their eyes to adjust the visual depth in the scene and reveal or hide occluded objects facilitating easier gaze pointing. This storyboard illustrates a user selecting the target marked with blue arrow using *DEEP*. The top row shows the user's eyelid movement, the middle row shows a top-down view of the layout with a gray circle representing the visual depth, the bottom row shows a front view that the user sees. *DEEP* shows a slide bar to indicated the current visual depth. Initially, the target is occluded by a sphere. The user widens his/her eyes to increase the visual depth. The slide bar turns white and the green cursor moves upward as visual feedback. During adjustment, objects that are closer than the visual depth will turn semi-transparent. Once the intended target (blue) is no longer occluded, the user stops widening the eyes. The user then dwells on the target to select it. The target turns green as visual feedback.

ABSTRACT

Gaze-based target suffers from low input precision and target occlusion. In this paper, we explored to leverage the continuous eyelid movement to support high-efficient and occlusion-robust dwell-based gaze pointing in virtual reality. We first conducted two user studies to examine the users' eyelid movement pattern both in unintentional and intentional conditions. The results proved the feasibility of leveraging intentional eyelid movement that was distinguishable with natural movements for input. We also tested the

UIST '22, October 29-November 2, 2022, Bend, OR, USA
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9320-1/22/10…$15.00
https://doi.org/10.1145/3526113.3545673

participants' dwelling pattern for targets with different sizes and locations. Based on these results, we propose *DEEP*, a novel technique that enables the users to see through occlusions by controlling the aperture angle of their eyelids and dwell to select the targets with the help of a probabilistic input prediction model. Evaluation results showed that *DEEP* with dynamic depth and location selection incorporation significantly outperformed its static variants, as well as a naive dwelling baseline technique. Even for 100% occluded targets, it could achieve an average selection speed of 2.5s with an error rate of 2.3%.

CCS CONCEPTS

• **Human-centered computing** → **Pointing**; **Virtual reality**; **User models**.

KEYWORDS

virtual reality, gaze interaction, pointing

ACM Reference Format:
Xin Yi, Leping Qiu, Wenjing Tang, Yehan Fan, Hewu Li, and Yuanchun Shi. 2022. DEEP: 3D Gaze Pointing in Virtual Reality Leveraging Eyelid Movement. In *The 35th Annual ACM Symposium on User Interface Software and Technology (UIST '22), October 29-November 2, 2022, Bend, OR, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3526113.3545673

1 INTRODUCTION

With the rapid development of virtual reality (VR) and augmented reality (AR), the pointing performance for virtual targets becomes a crucial feature for modern HMDs. Among various solutions, gazing attracts increasing attention from both the academia (e.g., [2, 18, 39, 40, 44, 46]) and the industry (e.g., HTC Vive Pro Eye, HoloLens 2, and Sony PSVR 2), due to its high moving speed and the capability of hands-free interaction. However, practical gaze pointing faces three major challenges: 1) unintentional triggering due to the continuous gaze movement (the "Midas touch problem" [55]), especially in VR where gaze tracking is always on; 2) low input precision due to jitter [50], which is more severe for small targets; 3) target occlusion in 3D scenes.

So far, researchers have proposed various techniques to facilitate gaze pointing, including dwelling [6, 18, 26, 34, 40], gaze gestures [8, 10, 27, 44, 46], eye vergence [2, 21, 25], and incorporating auxiliary modalities [26, 39, 47, 51]. These techniques have been proven effective in conventional pointing tasks. However, the problem of selecting occluded and dense targets [13] is not well-solved yet. These tasks are important in VR, which could appear in various scenarios (e.g., 3D CADs, smart room interaction and gaming).

In this paper, we propose *DEEP*, a high-efficient and occlusion-robust gaze pointing technique in VR. *DEEP* is inspired by the observation that people usually widen and squint their eyes when trying to focus on distant and close targets, respectively. The interaction of *DEEP* features two designs (see Figure 1): 1) users can continuously control the *Aperture Angle of Eyelids (AAE)* to adjust the *visual depth*, which is defined as the distance from the eyes up to which objects become semi-transparent, enabling the selection of occluded targets; 2) a probabilistic decoder can compensate the imprecision of gaze dwelling, enabling the users to accurately select small targets.

We conducted three user studies to facilitate the design of *DEEP*. Study 1 explored whether intentional eyelid movement was distinguishable from natural movements, and how comfortable it was to perform such movements. Results showed that natural eyelid movement range was small, and users were comfortable to intentionally perform eyelid movements, which provided AAE thresholds for depth adjustment detection in *DEEP*. Study 2 further examined users' ability to control eyelid movements for input. Results suggested that controlling AAE in a region was easier than holding it at precise values, which informed the depth adjustment design of *DEEP*. Study 3 examined users' gaze dwelling patterns. Results provided parameters for target disambiguation of *DEEP*.

We evaluated the interaction performance of *DEEP* in scenes with different levels of target occlusion and densities. We also tested a *Naive Dwell* technique, and two variants of *DEEP*: *L-DEEP* (L for location) and *D-DEEP* (D for depth), which statically emphasized the function of location selection and depth selection, respectively. Results showed that *L-DEEP* yielded the highest selection speed for less occluded targets, while *DEEP* or *H-DEEP* (H for hybrid) with the dynamic incorporation of both location selection and depth selection achieved the highest overall performance and user satisfaction. All three techniques performed significantly better than *Naive Dwell*.

The contributions of this paper are three-folded: 1) we systematically modeled the users' eyelid movement pattern in natural conditions and in different pointing tasks, providing empirical data on the users' ability of controlling their continuous eyelid movement; 2) we propose *DEEP*, the first technique to leverage continuous eyelid movement for visual depth adjustment, and dynamically incorporated probabilistic input prediction for dwell-based gaze pointing in VR; 3) we evaluated the interaction performance of *DEEP* with different design alternatives and in different tasks. The results prove that *DEEP* is high-efficient and occlusion-robust for 3D gaze pointing in VR. Also, the analysis on usage log sheds light on the users' pointing strategy for different targets.

2 RELATED WORKS

2.1 Gaze-based Pointing Techniques

Researchers have proposed a number of gaze-based techniques for target selection. To help resolve the false triggering problem of gaze input, or to increase the input speed, many techniques use gaze to point at the target or select a range of targets, and use auxiliary modalities (e.g., head [47], keyboard [26], touchscreen [51] and hand gesture [39]) to confirm the selection. Although effective, the involvement of additional input modalities increases the complexity of the techniques, and limits the application scenarios.

Gaze-only inputs usually employ dwell-based, gesture-based or vergence-based input techniques. Dwelling is the most common gaze pointing technique [6, 18, 26, 34, 40]. Users fixate their gaze on the target for a period of time to select it, which is intuitive and could help resolve false triggering [34]. To reduce fatigue due to long-time dwelling, researchers use Fitts' Law [18] and probabilistic model [40] to dynamically adjust the dwell time. However, applying dwell-based techniques to small-sized targets is hard due to the low input precision human eyes' natural jitters [38].

Gesture-based techniques seek to overcome the "Midas touch problem" [8, 10, 27, 44, 46]. Users move their gaze along a predefined path or the target to trigger selection, which helps disambiguate users' input with natural gaze movements. However, this is less intuitive than dwell-based pointing, and frequently performing rapid gaze movement may cause discomfort [46].

Vergence-based techniques allow users to gaze in the scene, and then cross their eyes to trigger input [2, 21, 25]. In theory this achieves high input performance as the interaction behavior is subtle. However, divergence-based pointing is not easy to perform, which leads to considerable learning effort. In comparison, *DEEP* also uses subtle movement around the users' eyes to facilitate gaze pointing, but the eyelid movement design is more close to the experience in daily lives.

2.2 Pointing for Small and Occluded Targets in Virtual Reality

Small and occluded targets post significant challenge in VR pointing [3, 13]. Existing techniques focus on enhancing the ray-casting selection mechanism of the controller (e.g., [3, 13, 33, 41, 49, 58]). These techniques demonstrate good performance in pointing partially occluded objects [54]. In particular, Depth Ray [13] and Raycursor [3] add a cursor on the ray to enable 3D selection, but users cannot visually acquire fully occluded objects for selection. Alpha Cursor [58] turns the objects semi-transparent when moving the cursor, which inspires *DEEP*'s visual depth adjustment design. Noticeably, all of these techniques require auxiliary input (e.g., touch pad and joy stick) that cannot be used in gaze-only scenarios.

Some techniques explore to rearrange the objects or change the user's perspective (e.g., [5, 22, 28]) to avoid occlusion. However, these techniques usually suffer from low pointing speed (e.g., $> 5s$ for high density and occluded targets [5, 22]), and can create unnecessary visual clutter, making them unsuitable for gaze interaction. Some techniques use statistical model [37] and scoring function [7] to distinguish target objects without rearranging them, which inspires our location selection mechanism. However, these techniques are not applied to dwell-based gaze input, and do not support selecting fully occluded targets.

For gaze input, Outline Pursuits [46] allows the users to select partially occluded targets by following a moving point along the target's outline with their gaze. However, it has difficulty distinguishing multiple objects with similar outline shapes. In comparison, *DEEP* can select all viewable targets without additional visual clutter. VOR Depth Estimation [35, 36] also allows users to select partially occluded objects by gazing towards the target and shake their heads. However, the selection can be unintentionally triggered during natural head movements. In addition, selecting fully-occluded targets is still not possible. In comparison, with visual depth adjustment, *DEEP* enables robust selection for all targets while minimizing the possibility of false triggering.

2.3 Input Techniques Leveraging Facial Expressions

Facial expression is widely used for interaction due to its rich expressiveness and naturalness [9]. Applications include emotion recognition [1, 4, 15, 30, 31], user intent recognition [11, 57] and gesture input [17, 19, 53]. As *DEEP* leverages the eyelid movement, we focus on the facial expressions around the users' eyes.

Many works explore using facial gesture for intentional input, including using the eye's closure and blink gestures. Jota et al. [19] explores the design space of eyelid gestures and classifies eye gestures into three discrete levels: open, closed and half-closed. Ku et al. [24] adds more levels to eye gestures: gaze, enlarge, frown, wink, raise eyebrow and squint, and tests them on an AR device. Li et al. [29] proposes gesture grammar and eyelid gesture detection mechanism.

Some researchers combine eye gesture with gaze. Heikkilä et al. [14] uses the closure of both eyes as confirmation. Gomez et al. [12] explores the closure of one eye and the gaze movement of the other to perform drag and drop. Similar to other gesture input techniques, these techniques face the challenge of gesture memorability, and are not suitable for target selection tasks.

The users' eyelid control ability is mainly investigated clinically [23, 45]. However, to the best of our knowledge, no one has formally investigate the users' ability of controlling the continuous movement of the eyelid. Accordingly, *DEEP* is the first to leverage the continuous movement of eyelid for interaction.

3 STUDY 1: EXAMINING THE NATURALNESS OF EYELID MOVEMENT

In order to design a robust and natural eyelid interaction mechanism for *DEEP*, we first conducted a user study to explore whether or not intentional eyelid movement is distinguishable from natural movements and easy to perform in order to provide the AAE thresholds for depth adjustment detection. We are interested in the range of natural eyelid movement during dwelling tasks rather than glancing tasks in existing work (e.g., [48]) as *DEEP* utilizes dwell detection.

3.1 Participants and Apparatus

We recruited 14 participants (7 male, 7 female) from the campus, with an average age of 21.0 (SD = 1.3). 9 of them reported occasional or no VR experience, while 5 of them had daily to monthly experience. Each participant was compensated $10.

We used an HTC Vive Pro Eye headset as the apparatus, which has a 90Hz display with a resolution of 1440 × 1600 and a field of view (FOV) of 110° per eye. The gaze data was obtained through the SRanipal SDK at a frame rate of 120Hz and $0.5 - 1.1^{\circ}$ gaze tracking precision using built-in eye tracker cameras [16].

We defined the *Aperture Angle of Eyelids (AAE)* to quantify the users' eyelid movement. It is calculated based on the "eye wide" and "eye blink" values from the SDK (both ranged from 0 to 1). Both data are 0 in neutral condition. "Eye wide" reaches 1 when the eyes is fully opened, and "eye blink" reaches 1 when the eyes are fully closed. As the eyelid movement range of different users varied, some participants were not able to reach an "eye wide" of 1. Therefore, we designed an AAE calibration process that measures the range of the values for each participant, and normalizes the data to calculate AAE (ranges from -1 to +1). In result, an AAE of -1, +1 and 0 indicates fully closed, widen to the extreme, and neutral condition, respectively.

3.2 Experiment Design

We arranged the targets on 5 concentric rings 5° apart to cover ±25° of the participants' FOV. The number of targets on each ring was 8, 8, 12, 12 and 16 respectively, resulting in 56 targets (plus 1 target at the center). All targets had a visual angular radius of 1° (see Figure 2a).

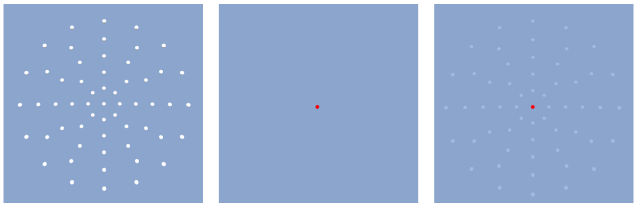

(a) (b) (c)

Figure 2: Experiment platform: (a) Target layout. (b) In the first task, the current target was highlighted, and other targets were hidden. (c) In the second task, the current target was highlighted, while other targets were semi-transparent for reference.

We designed two sub tasks in this experiment. The first task aims to examine the natural movement pattern of the participants' eyelid in gazing tasks. Therefore, we did not restrict the participants' head movement (i.e. rotation) during the tasks. However, to avoid parallax, they were not allowed to walk in the room or move their upper body. The second task aims to examine whether intentionally controlling the eyelid movement was comfortable for the users. We want to cover the full range of gaze angles, therefore head movement was restricted. During the experiment, no interaction feedback was provided to the users, which ensured the most natural behavior without potential bias towards any specific interaction design.

3.3 Procedure

Participants were seated during the experiment. They first performed eye tracker calibration of the headset, and our AAE calibration process (less than 10 seconds). In the first task they were asked to gaze at each of the target twice in random order. The current target was highlighted in red, while other targets were hidden to avoid visual distractions (see Figure 2b). They were asked to gaze at the target naturally for 2 seconds, and not blink during gazing. After that, they pressed the space key to continue to the next target. They were allowed to rest if they felt tired. The experiment took about 20 minutes. In the second task, the current target was also highlighted in red, but other targets were translucent to facilitate participants compare and rate (see Figure 2c). During gazing, they were asked to try to widen and squint their eyelids as much as possible, and rate the comfort level when doing this from 1 (impossible) to 5 (very easy).

3.4 Results

3.4.1 Eyelid Movement Range. We analyzed the AAE distribution for all targets based on the results in the first task, as shown in Figure 3. AAE roughly followed a Gaussian distribution, with the mean being 0.04. This confirmed that in most cases, the participants kept a neutral AAE that was close to 0. Meanwhile, the standard deviation of the distribution was 0.31, suggesting that even during natural movements, the participants' AAE could vary significantly when gazing at targets at different angles. Some AAE data were beyond the the maximum eyelid movement range ([-1, +1]) as users' eyelids might instantaneously exceed this region during natural gaze movement and lead to outliers. However, they only accounted for 0.7% of all data points, and had minimal impact on the results.

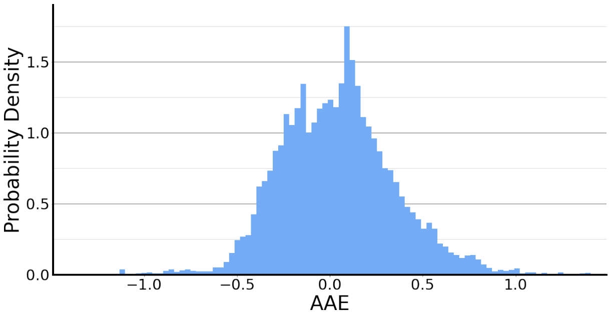


Figure 3: AAE distribution of all participants.

The 90% confidence interffval of the AAE distribution was [-0.47, 0.55], suggesting that the distribution was skewed towards widening the eyes. However, this was still significantly smaller than the maximum eyelid movement range, highlighting the possibility of leveraging intentional eyelid movement out of this region for input.

3.4.2 Correlation Between Gaze and Eyelid Movement. To further investigate the effect of target location on eyelid movement, we analyzed the correlation between gaze and AAE based on the results in the first task. Figure 4 shows the AAE value at different gaze locations. Compared with the distribution of the targets, the range of gaze point movement was small and centralized, indicating that the participants tended to rotate their heads to avoid gazing at targets with great angles. The center of the gaze point distribution was lower than the center of FOV. The distribution was bilaterally symmetrical, and was wider vertically than horizontally.

A significant correlation between gaze point location and AAE was observed. Generally, AAE increased as the users gazed upwards and decreased as they gazed downwards. This result corroborated with existing finding [56] that eyelid movements had strong correlation with gaze. This can be explained as the muscle of human eyes is connected with the eyelid, causing correlated movements [20, 43].

3.4.3 Ease of Intentional Eyelid Movements. The results for the first task suggested that leveraging the eyelid movement out of the neutral region for input was possible. While in the second task, we analysed participants' ease of intentional eyelid movements. Figure 5 shows the subjective ratings for the comfort of squinting and widening the eyes. As expected, performing eyelid movement around the center of FOV received the highest rating. Generally, when gazing at targets with greater angles, both the ratings for

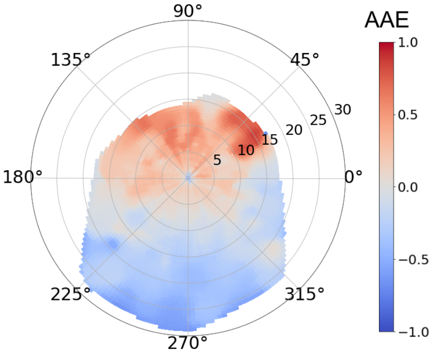
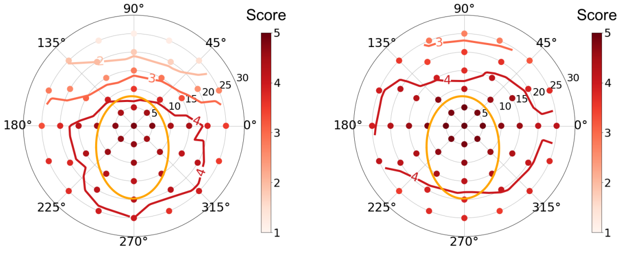


Figure 4: AAE value at different gaze locations.



(a) Squinting **(b) Widening**

Figure 5: Average subjective ratings on the comfort of (a) squinting and (b) widening the eyelid for targets at different locations (1: impossible, 5: very easy). The color bar indicates the rating scores. We plotted the isograms of scores from 2 to 4 after linear interpolation. The orange ellipse is the 90% confidence ellipse of gaze point distribution.

squinting and widening would drop monotonically. This implied that moving the eyelid when gazing at extreme angels was difficult. This trend was especially observable for squinting. Friedman test found that across all targets, the ratings for squinting was significantly lower than that for widening ($\chi^2(1) = 632, p < .001$), with the average rating being 3.59 (SD = 1.02) and 4.15 (SD = 0.56) respectively. Noticeably, within the 90% confidence ellipse of natural gaze point distribution, all targets (15/15) received a rating higher than 4. This proved that during natural gaze movement, intentionally controlling the eyelid movement for interaction was comfortable for the participants.

4 STUDY 2: EXAMINING EYELID MOVEMENT CONTROL ABILITY

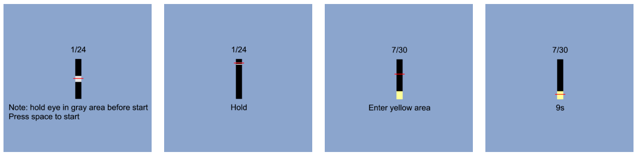
In Study 1, we verified that intentional eyelid movement is distinguishable from natural eyelid movement and comfortable to perform. In this study, we aim to further examine users' precise eyelid control abilities in order to inform the eyelid movement interaction design of *DEEP*.

4.1 Participants and Apparatus

We recruited 14 participants (7 male, 7 female) from the campus, with an average age of 21.5 (SD = 1.2). 7 of them reported occasional or no VR experience, while 7 of them had daily to monthly experience. 3 of them have participated in study 1, but did not exhibit learning effect because study 1 did not require holding eyelid movement. Each participant was compensated $10. We used the same apparatus as in previous study.

4.2 Experiment Design

To test the participants' ability of controlling their AAE, we designed two kinds of targets: line target and segment target. A line target indicated a specific value that the participants should keep their AAE at. And a segment target indicated a range that the participants should keep their AAE within. To evenly test the entire range of AAE ([-1, 1]), we arranged 8 line targets with an AAE of ±0.2, ±0.4, ±0.6 and ±0.8, respectively. Meanwhile, as we aimed to leverage the AAE out of the central region for interaction, we designed the segment targets to to be [-1, -X] and [X, 1], with X being 0.2, 0.4, 0.6, 0.8 and 0.9, respectively, resulting in 10 segment targets in total.



Note: hold eye in gray area before start
Press space to start

Hold

Enter yellow area

9s

(a) **(b)** **(c)** **(d)**

Figure 6: Experiment platform. (a) At the start of each trial, participants' AAE should be within the gray neutral region. (b) For line targets, participants adjusted their AAE to the target value, kept dwelling and pressed a button to start a 10-second recording. (c) For segment targets, participants adjusted their AAE into the yellow segment, (d) and automatically triggered the recording. Participants were asked to keep within the segment for as long as possible.

Figure 6 shows the visual feedback. A red line on a black bar indicates the current AAE value within the range of [-1, 1]. Each line target is shown as a white line (see Figure 6b), and each segment target is shown as a yellow segment (see Figure 6c).

4.3 Procedure

Participants were seated during the experiment. They first performed eye tracker calibration of the headset, and our AAE calibration process. They then completed the tasks for 8 line targets and 10 segment targets in random order. In each trial, they first kept their AAE within the neutral gray region ([-0.15, +0.15]). They then pressed the space key to show the target, and adjusted the AAE to hit the target or enter the segment as quickly as possible. For line targets, the participants pressed the space key again when they felt they had reached the target, and then they kept their AAE at the value for 10 seconds for recording. For segment targets, the recording would automatically start when AAE entered the segment, and the participants were required to keep within the segment for as long as possible. We restricted the maximal dwelling time to 15 seconds to avoid fatigue. During the experiment, the participants were allowed to rest when they felt tired. The experiment took about 30 minutes.

4.4 Results

4.4.1 Pointing Time. Pointing time is measured as the time elapse between the moment the target appeared and AAE reached the target (pressed the key for line targets, or entered the segment for segment targets). Figure 7 shows the pointing time for different targets. Each segment target is represented by the AAE value of its inner boundary. As expected, the pointing time for line targets was consistently longer than that for segment targets with the same AAE. This result confirmed that adjusting the AAE to a specific value was more difficult than keeping it within a region.

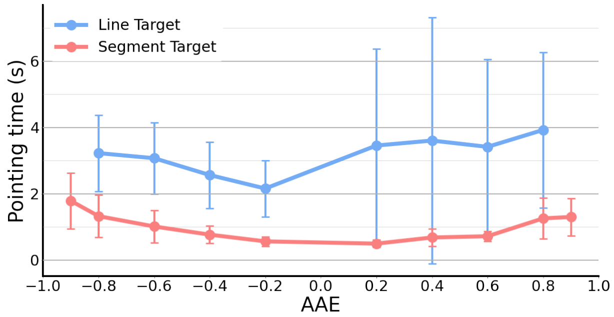


Figure 7: Pointing time for different line and segment targets. Error bar indicates one standard deviation.

For line targets with $AAE < 0$, the pointing time increased monotonically as the targets became further from the neutral position. However, this trend was not observed for line targets with $AAE > 0$. According to RM-ANOVA, significant effect of AAE was found on pointing time in the former condition ($F_{3,39} = 5.98, p < .01$), but not in the latter condition ($F_{3,39} = 0.64, p = .59$). Interestingly, compared with squinting ($AAE < 0$), the pointing time during widening ($AAE > 0$) appeared to be longer with greater variance. This suggested that performing fine control during widening may be more difficult than during squinting.

For segment targets, the pointing time increased monotonically as the targets became further from the neutral position. According to RM-ANOVA, significant effect of AAE was found on both sides ($AAE < 0$: $F_{4,52} = 20.3, p < .001$; $AAE > 0$: $F_{4,52} = 16.5, p < .001$). Opposite from the results of line targets, the pointing time during widening ($AAE > 0$) was shorter ($F_{1,13} = 1.41, p < .05$) with much smaller variance. This suggested that for all the participants, widening the eyes to enter a segment was easier than squinting.

4.4.2 Holding Time within Segment Targets. The above results suggested that using segment targets for eyelid interaction was more efficient and robust than line targets. To further investigate the users' AAE holding abilities within segment targets, we analyzed the holding time measured as the time elpase between the moment when AAE entered the segment and AAE exiting the segment or the maximal dwelling time was reached.

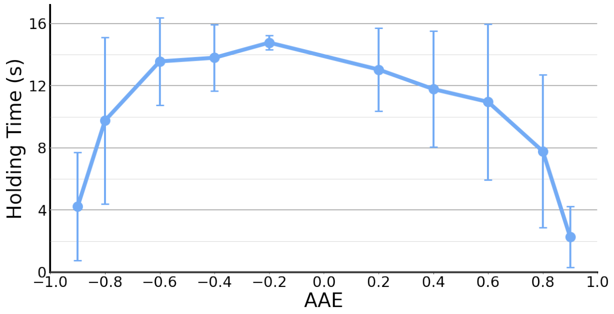


Figure 8: Holding time for different segment targets. Error bar indicates one standard deviation.

Figure 8 shows the holding time for different segment targets. In both widening and squinting conditions, the holding time decreased monotonically as the segment was narrower (also further from the neutral position). According to RM-ANOVA, significant effect of AAE was found on the holding time ($AAE < 0$: $F_{4,52} = 32.1, p < .001$; $AAE > 0$: $F_{4,52} = 43.0, p < .001$). *Post hoc* analysis found that targets with $|AAE| \geq 0.8$ yielded significant shorter holding time than $|AAE| = 0.2$. This suggested that the participants' AAE holding ability was relatively stable, which only dropped at extreme AAE values.

RM-ANOVA found the holding time during widening was slightly shorter than during squinting ($F_{1,13} = 148.4, p < .05$). However, for all targets with $|AAE| \leq 0.8$, the average holding time was over 7 seconds, which was sufficient for interaction in *DEEP*. In comparison, the average time elapsed between eye blinks was only 2.8 seconds [52].

5 STUDY 3: MODELING THE 2D GAZE DWELLING BEHAVIOR

Previous studies informed *DEEP*'s interaction design employing eyelid movement. However dwell-based selection is imprecise on small or dense targets due to gaze jitter [6]. In 3D pointing, this becomes a major challenge as partially occluded targets can be dense and small from the user's point of view, and using *DEEP* to adjust the visual depth and revealing these targets is not enough for high precision selection. Therefore, in this study, we examine users' dwelling patterns to provide parameters for *DEEP*'s probabilistic decoder.

5.1 Participants and Apparatus

We recruited 12 participants (5 male, 7 female) from the campus, with an average age of 20.4 (SD = 1.1). 6 of them reported occasional or no VR experience, while 5 of them had daily to monthly experience. 9 of them participated in previous studies. However as the interaction tasks of the three studies were different (natural gaze movement vs. intentional eyelid movement vs. gaze dwelling), participants did not exhibit learning effects. Each participant was compensated $10. We used the same apparatus as in previous studies.

5.2 Experiment Design

We used a single-factor within-subjects design, with *Target Size* as the only factor. We tested five levels of target sizes (measured in visual radius): 1°, 1.5°, 2°, 3° and 4°. As shown in Figure 9, all targets were arranged within an ellipses with a major axis of 12.5° and a minor axis of 10°. The center of the ellipse was set at (0°, -2.5°). These values were determined according to the 90% confident ellipse of the gaze point distribution in natural conditions in Study 1 (see Figure 5). The number of targets in each condition was 25, 17, 17, 9 and 5, respectively.

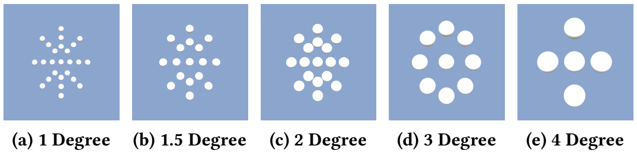

(a) 1 Degree (b) 1.5 Degree (c) 2 Degree (d) 3 Degree (e) 4 Degree

Figure 9: Target layout of the five target sizes in Study 3.

The goal of this study was to model the effect of target size and location on the users' 2D gaze dwelling behavior. Therefore, the position of all targets were fixed to the participants' FOV, regardless of head rotation. Meanwhile, we provided no visual feedback, in order to avoid any potential effect on their behavior. We were interested in the upper bound of the dwelling precision that the participants could achieve given sufficient time. Therefore, we did not restrict their pointing speed. The recording began after they stabilized their gaze and pressed a button.

5.3 Procedure

Participants were seated during this experiment. They first performed eye tracker calibration of the headset, and then performed gazing tasks for the five levels of target sizes in random order. In each condition, they pointed at each target twice, in random order. In each task, the target would be highlighted in red, the participants moved their gaze to the target and dwell on it. When they felt they had stabilized the gaze, they press a button to start a 2-second recording. After that, they pressed the button again to continue to the next target. The participants were allowed to rest if they felt tired. The experiment took about 30 minutes.

5.4 Results

5.4.1 Collective Gaze Point Distribution. We analyzed all the recorded gaze points. Figure 10 shows the collective gaze point distribution and the 90% confidence ellipses. The percentage of gaze points that fell within the target boundary for increasing sizes of targets was 63.7%, 82.6%, 92.6%, 97.2% and 99.2%, respectively. This suggested that even for 4° targets, the gaze point could still fall out of the target boundary during dwelling. For smaller targets, this percentage dropped monotonically, and even fell below 64% for 1° targets, making it nearly impossible to select the corresponding targets using dwelling. Therefore, a robust dwelling selection algorithm was necessary.

5.4.2 Dwelling Precision. We measured the participants' dwelling precision by calculating the length of the semi-major and semi-minor axes of the 90% confidence ellipses, as shown in Figure 11. Across all target sizes, the length of the semi-major axes varied between 0.93° and 1.01°, and the length of the semi-minor axes varied between 0.39° and 0.42°, which was comparable with the target sizes. According to RM-ANOVA, target size yielded no significant effect on the length of either axis (semi-major axis: $F_{4,44} = 3.4, p < .05$, semi-minor axis: $F_{4,44} = 2.8, p < .05$). This suggested that the participants' dwelling precision was relatively consistent, regardless of the target size. Even for larger sizes of targets, they still tended to gaze at a focused point.

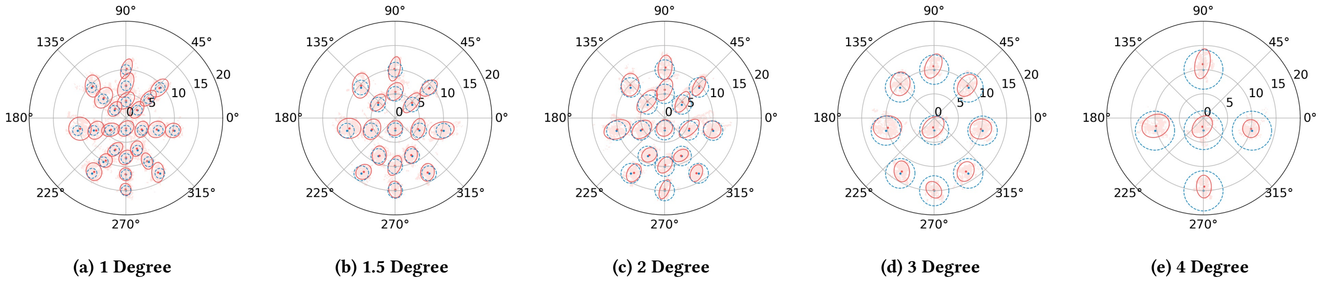
5.4.3 Systematic Offset. We calculated the angular offsets between the center of the 90% confidence ellipses and the target center, and reported horizontal offsets, vertical offsets and offset distances. Horizontal and vertical offsets are the angular distances between the center of the confidence ellipse and that of the target in the horizontal and vertical directions. A positive horizontal and vertical offset for example represents the center of the confidence ellipse is to the right and upper side of the target respectively. As shown in Figure 12, the average horizontal offset was very small (< 0.1°), while the average vertical offset was relatively greater (0.19° to 0.83°). The offset distance increased monotonically from 0.53° to 0.98°, but not as fast as target size. RM-ANOVA found that target size yielded a significant effect on vertical offset ($F_{4,44} = 8.30, p < .001$) and offset distance ($F_{4,44} = 7.54, p < .001$), but not on horizontal offset ($F_{4,44} = 0.59, p = .67$). In general, the participants' gaze points tended to land slightly above the target center, which corroborated with existing work [42]. Combining the results above, we could found that the users' dwelling precision was relatively high, but the systematic offset mainly led to a low dwelling accuracy for small targets.

6 DEEP: DESIGN AND IMPLEMENTATION

6.1 Interaction Design

Traditional dwell-based gaze pointing techniques are error-prone when pointing occluded or dense targets [34]. To tackle these challenges, *DEEP* has two design goals: 1) the users can adjust the visual depth to select occluded targets (depth selection); 2) the users can dwell to discern targets close to each other (location selection). *DEEP* is designed as the first technique to leverage continuous eyelid movement for visual depth adjustment, and dynamically incorporate probabilistic input prediction for dwell-based gaze pointing.

Inspired by the observation that people usually widen their eyes when focusing on distant objects, and squint for close objects, *DEEP* leverages eyelid movement for visual depth control. When using *DEEP*, the user can intentionally widen or squint his/her eyes to continuously control the visual depth (see Figure 1). Objects whose depth is smaller than the visual depth (closer) are rendered semi-transparent, and are unselectable. Meanwhile, objects whose depth are greater than the visual depth (further) are unaffected. Compared to techniques that re-arranged the positions of 3D targets (e.g. [5]), *DEEP* provides a smoother interaction experience for the user.



(a) 1 Degree (b) 1.5 Degree (c) 2 Degree (d) 3 Degree (e) 4 Degree

Figure 10: Collective gaze point distribution across all participants in different conditions. Red dots and solid lines indicates the center and boundary of the 90% confidence ellipses, respectively. Blue dots and dashed lines show the target center and boundary for reference.

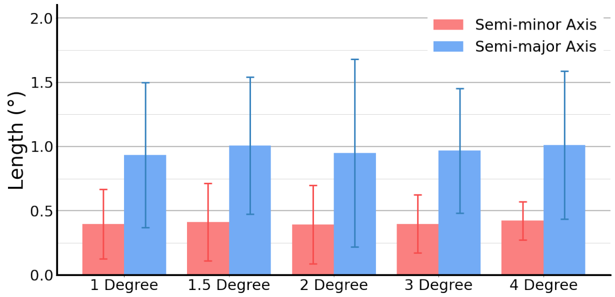


Figure 11: Dwelling precision (the length of the semi-major and semi-minor axes of the 90% confidence ellipses) for different sizes of targets. Error bar indicates one standard deviation.

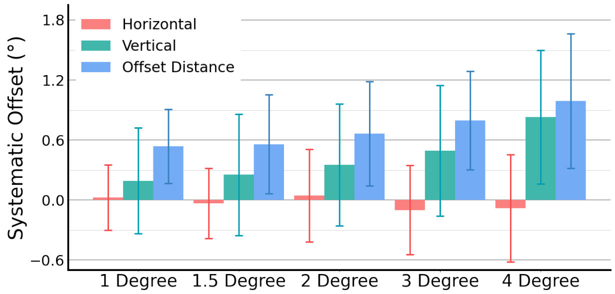


Figure 12: Systematic offset for different sizes of targets. Error bar indicates one standard deviation.

When the intended target is no longer occluded, the user can dwell on the target to select it. It is worth mention that when using *DEEP*, the user can freely choose to perform selection after all occlusions are gone, or when the target is partially occluded. However, as we showed in Study 3, for partially occluded targets, using its boundary for dwelling detection could lead to input error due to its limited size. Therefore, instead of manipulating the dwelling gaze points to always fall within the target's boundary we adopt a probabilistic modeling to discriminate ambiguous gaze targets. This design improves dwelling selection performance, and allows users to point with less stress.

During the pilot study, we discovered that when adjusting the visual depth, the users' gaze location usually stayed still, which may cause unintentional triggering. Therefore, we design all targets to be unselectable during visual depth adjustment. In addition, we found that displaying the users' gaze ray was distracting, and would harm the user's confidence due to its jitter. Therefore, we hide the gaze ray from users.

6.2 Visual Depth Adjustment

Figure 13 shows the algorithm pipeline of *DEEP*. Visual depth adjustment is achieved by controlling the AAE of the user. In Study 1, we found that the AAE value in natural conditions mainly fell within a limited range around the neutral point. Therefore, we design a region with two thresholds $[AAE_{low}, AAE_{high}]$. When the AAE falls within this range, the visual depth adjustment is unable to be triggered. In Study 2, we found that adjusting the AAE to a specific value was difficult. Therefore, we mimic the segment target condition, and design the visual depth to increase when AAE falls within $[AAE_{high}, 1]$, and to decrease when AAE falls within $[-1, AAE_{low}]$. According to results from Study 1 and 2, we determine AAE_{low} and AAE_{high} to be -0.5 and +0.6, respectively, which balances the interaction performance and false triggering.

In practice, blinking also occasionally causes false triggering. Therefore, we design a time threshold of 0.5s before the visual depth adjustment is triggered (see Figure 13) based on the average human blink time of 0.1 to 0.4 seconds [52]. A visual depth speed too slow introduces extra fatigue and harm the performance, while a speed too fast would make it harder to precisely control the visual depth. Therefore, we pre-tested different speeds, and dynamically adjust the speed so that the adjustment time from the nearest point to the furthermost point in the scene is 3 seconds.

In order to facilitate AAE control, we design visual feedback for *DEEP*. As shown in Figure 1, we display a slide bar on the right side of the FOV that mapped to [-1, 1], and a cursor that indicated the current AAE. By default, the slide bar is shown in gray to minimize distraction. And when the visual depth adjustment is triggered, the bar is highlighted in white, and the cursor changes to green. The

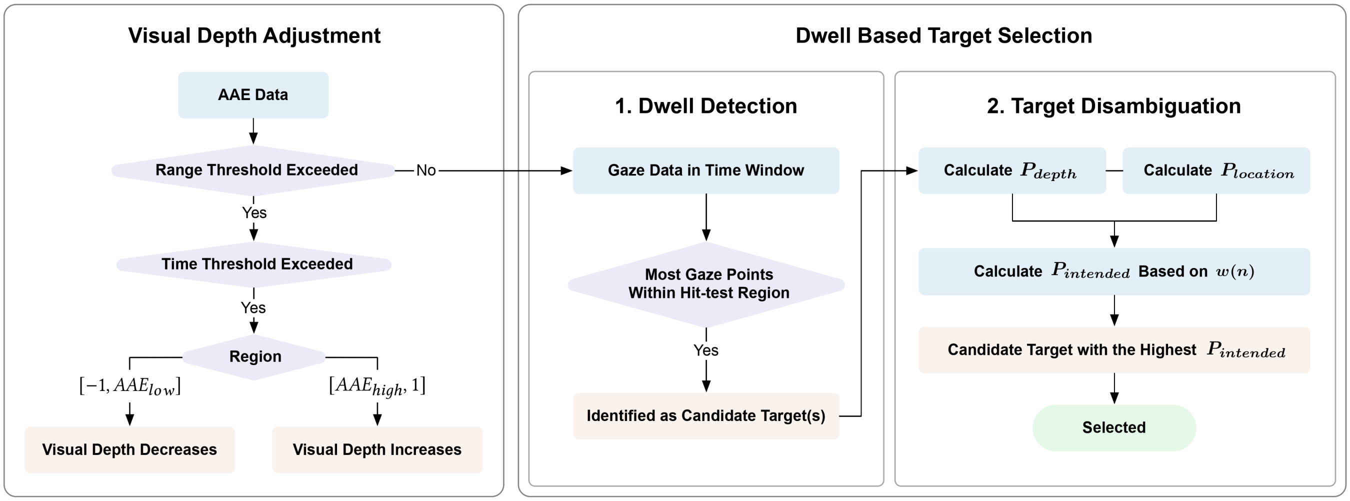


Figure 13: Algorithm Pipeline of *DEEP*.

adjustment stops once the AAE returns to the central region, and the slide bar turns to gray.

6.3 Dwell-Based Target Selection

As mentioned above, to help select dense and small targets, we incorporate a probabilistic model into the classical dwell-based target selection technique, which consists two steps:

Step 1: Dwell detection. Based on the typical dwell time for gaze selection of 0.6 to 1.0 seconds [34], we use a 0.8 seconds time window to detect dwelling. According to Study 3, we enlarge the radius of the "hit-test region" for each target to be at least 2° (larger targets are not affected) (see Figure 14a). Once 90% of the gaze point within the time window falls within the hit-test region, *DEEP* identifies the corresponding target as a candidate target (see Figure 14b). Compared with other target-expansion techniques (e.g. Expand [5]), this design is more conservative, and will not significantly increase the false triggering rate.

Step 2: Target Disambiguation. It is possible that in step 1, multiple targets are identified as the candidate (e.g. when they are very close or overlapped in depth). Therefore, we design an algorithm to disambiguate these candidates. Specifically, for each candidate target k ($1 \leq k \leq n$), *DEEP* calculates the probability of it being the intended target $P_{intented}(k)$, and selects the one with the highest probability. The calculation is:

$$P_{intended}(k) = w(n) \times P_{depth}(k) + [1 - w(n)] \times P_{location}(k). \quad (1)$$

In Equation 1, P_{depth} quantifies the selection probability in the depth dimension. Note that all targets whose depth is smaller than the visual depth are unselectable. Therefore we design P_{depth} to be a linear model that yields smaller probability for further targets:

$$P_{depth}(k) = (d_{max} - d_k)/\sum_{i=1}^{n}(d_{max} - d_i) \quad (2)$$

where d_i indicates the depth of the ith candidate. d_{max} is the depth of the furthermost candidate (see Figure 14c).

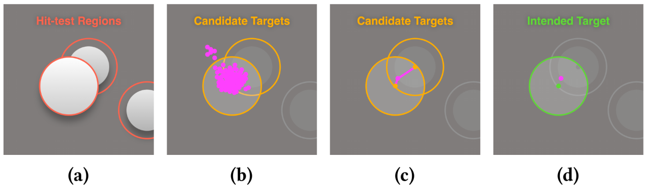


Figure 14: Dwell Based Target Selection: (a) The hit-test regions for targets smaller than 2° are enlarged to 2°. (b) Targets with 90% or more gaze points falling in their hit-test region within the 0.8s time window are marked as candidate targets. (c) P_{depth}: Target in the front has higher probability of selection. $P_{location}$: Target closer to the center of gaze points has higher probability of selection. (d) $P_{intended}$: Weighs the two probabilities and select the intended target.

Meanwhile, $P_{location}$ quantifies the dwelling selection probability on the x-y plane. According to the results in Study 3, we use a univariate Gaussian distribution to model the offset distance between the centroid of the dwelling gaze points and the target center. We then perform linear interpolation to calculate the model parameters for targets with arbitrary sizes. After that, we calculate $P_{location}(k)$ based on the offset distance between the gaze point and the center of the kth candidate target using the distribution Figure 14d.

Finally, $w(n)$ serves as a weighting factor between the above two probabilities, where n is the number of candidate targets. A great n value usually appears when multiple targets are very close in location, but overlapped in depth. While a small n value usually means that users can easily distinguish between them using location selection (e.g., by shifting the gaze location away from the distraction). Therefore, we design $w(n)$ to be a piecewise function that increased with n. To determine the value of $w(n)$, we conducted a pilot study, in which 4 participants performed 200 pointing tasks using different $w(n)$ values ranging from 0.1 to 0.9, and different cut points from 2 to 5. Finally, we set $w(n) = 0.3$ for $n \leq 2$, and $w(n) = 0.9$ for $n \geq 3$.

7 STUDY 4: INTERACTION PERFORMANCE EVALUATION

In this section, we introduce a user study to evaluate the interaction performance of *DEEP* in scenes with different target layouts. We are also interested in testing the performance of the dynamic adjustment strategy (w(n)) compared with static strategies.

7.1 Participants and Apparatus

We recruited 13 participants (6 male, 7 female) from the campus, with an average age of 20.2 (SD = 0.2). 6 of them reported occasional or no VR experience, while 7 others had daily to monthly experience. 6/13 users in Study 4 have participated in the previous studies. However, they were all new to DEEP's algorithms and tasks in Study 4, therefore did not exhibit learning effect. Each participant was compensated $15. Tasks were developed using the same platform and hardware as in previous studies.

7.2 Experiment Design

We designed five scenes to mimic different real-life pointing scenarios (see Figure 15). The *50% Occluded* scene, *75% Occluded* scene and *100% Occluded* scene shared the same 2-layer sparse layout with different occlusion levels, which was similar with existing work [46]. The *Complex Depth* scene used a 4-layer layout with up to 100% occlusion. The *Complex Density* scene used a 2-layer dense layout with an average of 50% occlusion. We used targets with radius from 1° to 3°. The number of targets in the scenes were 20, 20, 20, 50 and 50 respectively.

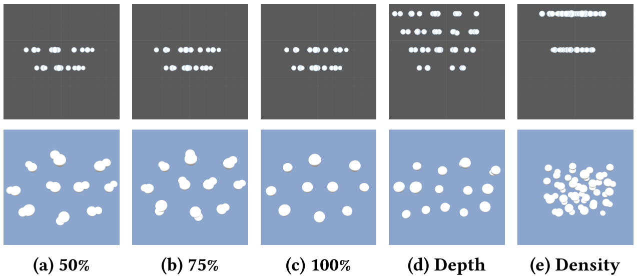


Figure 15: The five target layouts in Study 4, shown in top and front views. (a) 50% Occluded layout (b) 75% Occluded layout (c) 100% Occluded layout (d) Complex Depth layout (e) Complex Density layout.

We tested four techniques, which shared the same layouts:

- *Naive Dwell [34].* Visual depth adjustment and probabilistic dwelling selection was not supported. A target would be selected when 100% gaze points within a 0.8 time window fell within the target boundary.In *Naive Dwell*, target boundary is the visual outline of a target.
- *L-DEEP (L for location).* A static variant of *DEEP*. We set $w(n) \equiv 0.3$, and disabled the visual depth adjustment function to emphasize the location selection function.
- *D-DEEP (D for depth).* A static variant of *DEEP*. We set $w(n) \equiv 0.9$ to emphasize the depth selection function.
- *H-DEEP (H for hybrid).* The complete *DEEP* with dynamic weighting factors. We renamed it for presentation clarity.

L-DEEP and *D-DEEP* are DEEP variants, while *Naive Dwell* is the baseline. We did not test other techniques mentioned in Section 2 as their interaction requirements are different from DEEP (e.g., not hands-free [3, 13, 58], requires different target shapes[46], modifies target positions [5]). We also did not test VOR [32] as some participants can not perform it successfully.

The goal of this study is to test the performance of the techniques in real use. Therefore the participants were free to rotate their heads, but they were not allowed to walk or move their upper body. To improve the internal validity of the results, 20 pre-determined targets in different layers in each scene were used as the tasks. During the experiment, the target was marked blue with an arrow pointing at it, so that the participants could find the target even if it was occluded (see Figure 1). Since the participants were unable to adjust visual depth when using *Naive Dwell* and *L-DEEP*, it would be impossible for them to select fully or heavily occluded targets. Therefore we allowed the participants to give up on a task by pressing a button after trying.

7.3 Procedure

Participants were seated during the experiment. They first performed eye tracker calibration of the headset, and our AAE calibration process. They then familiarized themselves with the techniques for about 2 minutes, and completed four sessions of tasks in random order, each corresponding to one technique. Each session was consisted of five blocks, corresponding to the five scenes, in random order. In each block, they performed selection for the 20 targets in random order. For each target, they first pressed a button to show the target, and then select the target "as fast as possible". Each task was completed when the correct target was selected, or when the participant gave up on that task. A 5-minute break was enforced between different sessions. The participants were also allowed to rest if they felt tired. Finally, we gathered their subjective feedback towards the techniques through questionnaires and interviews.

7.4 Results

7.4.1 Selection Time.

Selection time is measured as the elapse between when the target displays and is selected, excluding aborted trials. Figure 16 shows the selection time for different techniques and scenes. Scenes with heavier occlusion led to longer selection time. *Complex Depth* yielded the longest selection time for all the techniques, while *Complex Density* only introduced severe challenge for *Naive Dwell*.

RM-ANOVA found a significant difference between the selection time of different techniques ($F_{3,36} = 103, p < .0001$). *Post hoc* analysis found that the three variants of *DEEP* all performed significantly better than *Naive Dwell* in all scenes. Comparing *L-DEEP* and *D-DEEP*, *L-DEEP* was faster in less occluded scenes (*50% Occluded*, *75% Occluded* and *Complex Density*), whereas *D-DEEP* was faster in more occluded scenes (*100% Occluded* and *Complex Depth*). Meanwhile, *H-DEEP* achieved competitive performance with the best performing technique in all the scenes. This proved that the

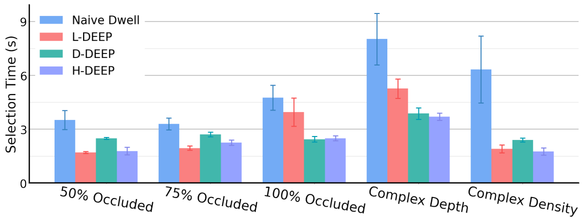


Figure 16: Selection time for different techniques and scenes. Error bar indicates one standard deviation.

dynamic algorithm of *H-DEEP* could leverage the advantages of both location and depth selection, making it adaptive in both dense and occluded scenes.

7.4.2 Error and Abort Rate. We calculated the *error rate* and *abort rate* as the percentage of incorrect and aborted selections out of all selections. Additionally, we calculated *failure rate* as the sum of error and abort rate, as shown in Figure 17. RM-ANOVA found significant difference between the techniques (Error rate: $F_{3,36} = 31.9, p < .0001$, Abort rate: $F_{3,36} = 24.8, p < .0001$, failure rate: $F_{3,36} = 66.2, p < .0001$). The abort rate of *Naive Dwell* in all the five scenes was all significantly higher than the other three techniques, suggesting that participants found it to be unresponsive.

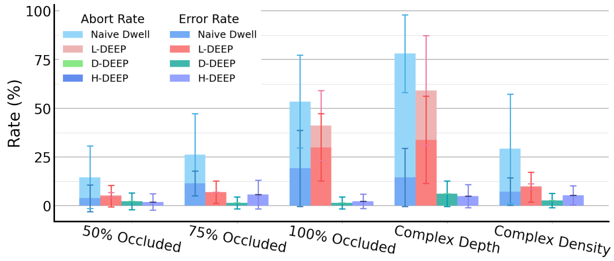


Figure 17: Error rate and abort rate for different techniques and scenes. The stacked bars indicated failure rate. Error bar indicates one standard deviation.

L-DEEP performed well in *50% Occluded*, *75% Occluded* and *Complex Density*, but its error rate dramatically increased to over 25% in the other two scenes. Surprisingly, no participant aborted selection when using *D-DEEP* and *H-DEEP*. And the error rate of these two techniques were very low (< 6.2%) in all the scenes. This suggested that compared with location selection, depth selection may benefit the selection success rate more effectively. Again, *D-DEEP* performed well in all the five scenes, proving its robustness.

7.4.3 Strategy of Visual Depth Adjustment. When selecting partially occluded targets using *D-DEEP* and *H-DEEP*, it was possible that the user could successfully perform the selection without visual depth adjustment. Therefore, we performed analysis on the participants' strategy of using visual depth depth adjustment, which could provide more information on the intrinsic advantage of different techniques.

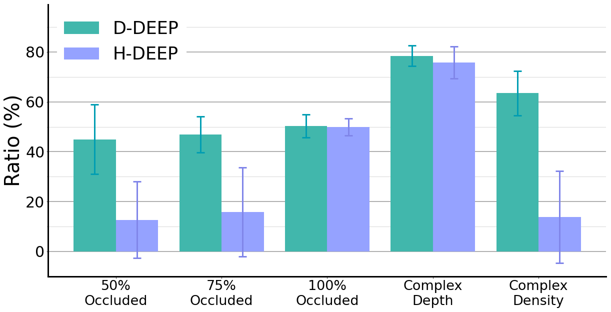


Figure 18: Ratio of selections that used visual depth adjustment. Error bar indicates one standard deviation.

Figure 18 shows the ratio of selections in each scene that used visual depth adjustment. Across different scenes, the ratio of *H-DEEP* varied from 13% to 78%, while that of *D-DEEP* were all above 40%. We speculated that the location selection feature of *H-DEEP* could help the users when distinguishing adjacent targets, therefore avoiding the need for depth selection for partially occluded targets. In comparison, as *D-DEEP* emphasized more on target depth, even the target was only partially occluded, the participants still needed to adjust the depth to hide all the closer occlusions. Even in *75% Occluded*, the ratio of *H-DEEP* was still below 20%, suggesting that the users generally preferred location selection, and only used depth selection when necessary.

7.4.4 Subjective Results. We asked the participants to rate all the techniques in terms of different dimensions from 1 to 5. Higher scores indicate higher preference. Dimensions included learnability (1: Difficult to learn, 5: Easy to learn), ease of use (1: Difficult to use, 5: Easy to use), performance (1: Unable to complete selection, 5: Able to complete speedy selection), fatigue (1: Tiring, 5: Easy), frustration (1: Very frustrating, 5: Not frustrating) and overall satisfaction (1: Not satisfied, 5: High satisfaction). Cronbach's α of the questionnaire was 0.85, confirming the internal consistency of the survey. Figure 19 shows the scores.

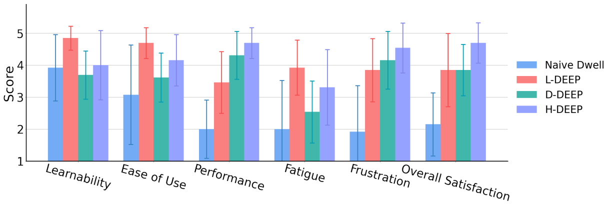


Figure 19: Subjective ratings for different techniques on a 5-point survey. Error bar indicates one standard deviation.

The scores of *L-DEEP* and *H-DEEP* were above 3 on all dimensions, indicating that the participants were generally positive towards the two techniques. Friedman test found that the scores of the techniques were significantly different on all dimensions: learnability ($\chi^2(3) = 13.6, p < .01$), ease of use ($\chi^2(3) = 14.5, p < .01$), performance ($\chi^2(3) = 31.5, p < .001$), fatigue ($\chi^2(3) = 18.8, p < .001$), frustration ($\chi^2(3) = 27.1, p < .001$) and overall satisfaction

($\chi^2(3) = 28.4, p < .001$). *L-DEEP* appeared to be the most learnable, easy to use technique, and caused the least faitigue, which was consistent with the above results. While *H-DEEP* achieved the best performance, caused the least frustration and had the highest overall satisfaction. This confirmed the effectiveness of both depth selection and location selection in 3D pointing. We also asked the participants to rank these techniques, and 12/13 participants chose *H-DEEP* to be the best.

We interviewed participants about their experience. They reported *Naive Dwell* to be the most strenuous and least accurate technique. *"There was little room for error." (P5).* They reported *L-DEEP* to be simple and easy to use for less-occluded targets. *"I could achieve acceptable selection performance." (P3).*. They reported *D-DEEP* with its depth control ability to be attractive and functional. *"Controlling depth with my eyelid is super cool!" (P11).* However, frequently performing depth control could cause fatigue. *"Controlling depth with my eyelid is super cool!" (P11).* They reported *H-DEEP* to be well-balanced and flexible. *"The best technique among the four, because I could choose to adjust depth to select occluded objects, and simply stare to select less occluded objects." (P9)*

We also interviewed them about their strategy when using *H-DEEP* which fell into three categories: 4 of them only adjusted depth for heavily occluded targets (95%-100% occluded), because they discovered that they could select other targets by gazing at the rim with the help of the location selection algorithm. 7 of them adjusted depth for objects that were partially occluded (50%-100% occluded), which best balance input performance and fatigue. While 2 of them adjusted depth for nearly all the targets, which ensured the most accurate selection.

8 DISCUSSION

8.1 Feasibility of DEEP in 3D Gaze Pointing

In Study 4, all three *DEEP* techniques showed significant advantage over *Naive Dwell* in terms of speed, accuracy, and user preference, demonstrating *DEEP* as an effective technique for 3D gaze pointing, especially for dense and occluded targets. As with other dwell-based techniques, the speed improvement was mainly brought by the shortening of time window. Compared with existing dwell time reduction techniques [18, 40], *DEEP*'s location selection improved jitter tolerance without requiring prior knowledge of users' pointing task (e.g., language model in text entry) or changing dwell time that may make users feel out of control.

The accuracy improvement was brought by visual depth adjustment that revealed occluded target and the probabilistic selection model for target disambiguation. We drew inspirations from existing depth adjustment implementations [3, 13, 58], and optimized them for gaze and fully occluded scenarios. Compared with other techniques that also comprehensively considered factors from different dimensions [37], our algorithm was derived from the behavior model built from user studies.

Our comparison of the three *DEEP* techniques in Study 4 revealed their unique strengths. *L-DEEP* performs best in lightly occluded scenes, and is perceived best in learnablity, ease of use and fatigue. However, its performance reduces significantly in heavily occluded scenes. It is useful when simplicity is required. In comparison, *D-DEEP* performs well in heavily occluded scenes, and is the most precise technique with the lowest failure rate. This makes it suitable for VR scenarios that require high robustness. In general, *H-DEEP* is overall the fastest, most accurate and most preferred technique due to its dynamic incorporation of both depth selection and location selection. *H-DEEP* incurs more fatigue than *L-DEEP* due to its additional depth adjustment actions. However, the difference was insignificant ($p = .28$). Therefore, we recommend *H-DEEP* in most application scenarios (e.g., 3D CAD, gaming, and for users with limited motor functions).

In practice *DEEP* can also be applied AR/MR HMDs (e.g., HoloLens 2 and Magic Leap 1) as long as the eye tracker (or external camera) can capture users' eyelids. The AAE calculation algorithm can be easily developed using computer vision.

8.2 Leveraging Eyelid Movement for Input in VR

DEEP demonstrated the feasibility of leveraging continuous eyelid movement in VR interaction. Comparing with gesture-based pointing techniques [46] that only tenseutilize gaze point location, eyelid movement serves an integral role in *DEEP* by providing additional input, allowing *DEEP*'s gaze interaction to be more relaxing while achieving better performance.

The results in Study 2 showed that adjusting AAE to a specific value took significantly longer time than reaching for a segment, and the time that users could keep their AAE within the segment targets could be relatively long (> 6s). Therefore, we suggested that continuous eyelid movement was more suitable for rough controlling tasks rather than fine controlling tasks. Figure 5 shows that intentionally controlling the AAE value was natural and comfortable for the users, and the interaction range of AAE was distinguishable with that during natural movements (see Figure 3). These proved the potential of eyelid interaction in terms of robustness and naturalness. The subjective ratings of *D-DEEP* and *H-DEEP* (see Figure 19) revealed that performing eyelid movement for a long time could lead to fatigue. Therefore we recommended that eyelid movement be utilized for quick (e.g., confirmation) rather than heavy tasks (e.g., text editing).

Based on the robustness and ease of use of eyelid interaction, we expected it to be applied in various VR scenarios. In addition to depth adjustment in *DEEP*, the added input space from eyelid movements can be used as triggers for menu or address the "Midas touch problem" using eyelid movement to enable and disable dwell selection.

9 LIMITATION AND FUTURE WORKS

DEEP features both depth selection and location selection. In this paper, we used $w(n)$ to dynamically combine these two components. Although effective, the current implementation was relatively simple and heuristic, and could be further improved. For example, leveraging more contextual information (e.g., task sequence) and a more sophisticated weighting model (e.g., machine learning) could potentially further increase the performance of *DEEP*. We planed this in future work.

Personalization served as a crucial part for any gaze based interaction techniques. In this work, we designed a short AAE calibration process to resolve the difference in eyelid moving range between different participants. Our user study results found that there were still a number of factors that could be considered in personalization (e.g., gaze-AAE correlation and 2D gaze point distribution during dwelling). Therefore, designing online personalization algorithms that learn from user's gaze data can also improve the gaze selection performance.

The current prototype of *DEEP* was tested under an abstract VR environment. We used sphere targets with no background distractions to ensure the internal validity of our results. It is worthwhile to test the interaction performance of *DEEP* and the user behavior in real-world scenarios with complex distractions and targets with different form factors. We also deffer this to future work.

10 CONCLUSION

In this paper, we present *DEEP*, a novel gaze pointing technique that leverages eyelid movement for pointing dense and occluded targets in VR. Our design followed the results of three user studies. Study 1 examined the naturalness of leveraging eyelid movement for interaction. Study 2 explored the users' eyelid movement control ability for different line and segment targets. Study 3 investigated the users' gaze point distribution in dwell selection tasks in terms of systematic offset and dwelling precision. In the evaluation study, we compared the interaction performance of *DEEP* and three baseline techniques. Results demonstrated that *H-DEEP* with a dynamic incorporation of depth selection and location selection significantly outperformed the other techniques in terms of selection time and accuracy. It was also the most preferred technique by the participants in general. We conclude that *DEEP* is an attractive solution that achieves efficient and occlusion-robust gaze pointing in VR scenarios.

ACKNOWLEDGMENTS

This work was supported by the Natural Science Foundation of China (NSFC) under Grant No. 61902208, and the grant from the Institute for Guo Qiang, Tsinghua University No. 2019GOG0003.

REFERENCES

[1] Faiza Abdat, Choubeila Maaoui, and Alain Pruski. 2011. Human-Computer Interaction Using Emotion Recognition from Facial Expression. *2011 UKSim 5th European Symposium on Computer Modeling and Simulation* 5 (2011), 196–201.

[2] Sunggeun Ahn, Jeongmin Son, Sangyoon Lee, and Geehyuk Lee. 2020. Verge-It: Gaze interaction for a binocular head-worn display using modulated disparity vergence eye movement. In *Extended abstracts of the 2020 CHI conference on human factors in computing systems*. 1–7.

[3] Marc Baloup, Thomas Pietrzak, and Géry Casiez. 2019. Raycursor: A 3d pointing facilitation technique based on raycasting. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[4] Marian Stewart Bartlett, Gwen Littlewort, Ian Fasel, and Javier R Movellan. 2003. Real time face detection and facial expression recognition: development and applications to human computer interaction.. In *2003 Conference on computer vision and pattern recognition workshop*, Vol. 5. IEEE, 53–53.

[5] Jeffrey Cashion, Chadwick Wingrave, and Joseph J LaViola Jr. 2012. Dense and dynamic 3d selection for game-based virtual environments. *IEEE transactions on visualization and computer graphics* 18, 4 (2012), 634–642.

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

[7] Gerwin De Haan, Michal Koutek, and Frits H Post. 2005. IntenSelect: Using Dynamic Object Rating for Assisting 3D Object Selection.. In *Ipt/egve*. Citeseer, 201–209.

[8] Morten Lund Dybdal, Javier San Agustin, and John Paulin Hansen. 2012. Gaze input for mobile devices by dwell and gestures. In *Proceedings of the Symposium on Eye Tracking Research and Applications*. 225–228.

[9] Paul Ekman. 1993. Facial expression and emotion. *American psychologist* 48, 4 (1993), 384.

[10] Augusto Esteves, Eduardo Velloso, Andreas Bulling, and Hans Gellersen. 2015. Orbits: Gaze interaction for smart watches using smooth pursuit eye movements. In *Proceedings of the 28th annual ACM symposium on user interface software and technology*. 457–466.

[11] Miguel García-García, Alice Caplier, and Michèle Rombaut. 2018. Sleep deprivation detection for real-time driver monitoring using deep learning. In *International conference image analysis and recognition*. Springer, 435–442.

[12] Argenis Ramirez Gomez, Christopher Clarke, Ludwig Sidenmark, and Hans Gellersen. 2021. Gaze+ Hold: Eyes-only Direct Manipulation with Continuous Gaze Modulated by Closure of One Eye. In *ACM Symposium on Eye Tracking Research and Applications: Bridging Communities*. Association for Computing Machinery.

[13] Tovi Grossman and Ravin Balakrishnan. 2006. The design and evaluation of selection techniques for 3D volumetric displays. In *Proceedings of the 19th annual ACM symposium on User interface software and technology*. 3–12.

[14] Henna Heikkilä and Kari-Jouko Räihä. 2012. Simple gaze gestures and the closure of the eyes as an interaction technique. In *Proceedings of the symposium on eye tracking research and applications*. 147–154.

[15] Steven Hickson, Nick Dufour, Avneesh Sud, Vivek Kwatra, and Irfan Essa. 2019. Eyemotion: Classifying facial expressions in VR using eye-tracking cameras. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1626–1635.

[16] HTC. 2019. Vive Pro Eye Specs. https://www.vive.com/us/product/vive-pro-eye/specs/

[17] Mirja Ilves, Yulia Gizatdinova, Veikko Surakka, and Esko Vankka. 2014. Head movement and facial expressions as game input. *Entertainment Computing* 5, 3 (2014), 147–156.

[18] Toshiya Isomoto, Toshiyuki Ando, Buntarou Shizuki, and Shin Takahashi. 2018. Dwell time reduction technique using Fitts' law for gaze-based target acquisition. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research and Applications*. 1–7.

[19] Ricardo Jota and Daniel Wigdor. 2015. Palpebrae superioris: Exploring the design space of eyelid gestures. In *Proceedings of the 41st Graphics Interface Conference*. 273–280.

[20] DW Kennard and GL Smyth. 1963. The causes of downward eyelid movement with changes of gaze, and a study of the physical factors concerned. *The Journal of physiology* 166, 1 (1963), 178.

[21] Dominik Kirst and Andreas Bulling. 2016. On the verge: Voluntary convergences for accurate and precise timing of gaze input. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human factors in Computing Systems*. 1519–1525.

[22] Regis Kopper, Felipe Bacim, and Doug A Bowman. 2011. Rapid and accurate 3D selection by progressive refinement. In *2011 IEEE symposium on 3D user interfaces (3DUI)*. IEEE, 67–74.

[23] Rakshit Kothari, Zhizhuo Yang, Christopher Kanan, Reynold Bailey, Jeff B Pelz, and Gabriel J Diaz. 2020. Gaze-in-wild: A dataset for studying eye and head coordination in everyday activities. *Scientific reports* 10, 1 (2020), 1–18.

[24] Pin-Sung Ku, Te-Yan Wu, and Mike Y Chen. 2017. EyeExpression: exploring the use of eye expressions as hands-free input for virtual and augmented reality devices. In *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*. 1–2.

[25] Shinya Kudo, Hiroyuki Okabe, Taku Hachisu, Michi Sato, Shogo Fukushima, and Hiroyuki Kajimoto. 2013. Input method using divergence eye movement. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*. 1335–1340.

[26] Manu Kumar, Andreas Paepcke, and Terry Winograd. 2007. Eyepoint: practical pointing and selection using gaze and keyboard. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 421–430.

[27] Andrew Kurauchi, Wenxin Feng, Ajjen Joshi, Carlos Morimoto, and Margrit Betke. 2016. EyeSwipe: Dwell-free text entry using gaze paths. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 1952–1956.

[28] Nianlong Li, Zhengquan Zhang, Can Liu, Zengyao Yang, Yinan Fu, Feng Tian, Teng Han, and Mingming Fan. 2021. vMirror: Enhancing the interaction with occluded or distant objects in VR with virtual mirrors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–11.

[29] Zhen Li, Mingming Fan, Ying Han, and Khai N Truong. 2020. iWink: Exploring eyelid gestures on mobile devices. In *Proceedings of the 1st International Workshop on Human-Centric Multimedia Analysis*. 83–89.

[30] Christine L Lisetti and Diane J Schiano. 2000. Automatic facial expression interpretation: Where human-computer interaction, artificial intelligence and cognitive science intersect. *Pragmatics & cognition* 8, 1 (2000), 185–235.

[31] Zhentao Liu, Min Wu, Weihua Cao, Luefeng Chen, Jianping Xu, Ri Zhang, Mengtian Zhou, and Junwei Mao. 2017. A facial expression emotion recognition based human-robot interaction system. *IEEE/CAA Journal of Automatica Sinica* 4, 4 (2017), 668–676.

[32] Mark A Livingston, J Edward Swan, Joseph L Gabbard, Tobias H Hollerer, Deborah Hix, Simon J Julier, Yohan Baillot, and Dennis Brown. 2003. Resolving multiple

occluded layers in augmented reality. In *The Second IEEE and ACM International Symposium on Mixed and Augmented Reality, 2003. Proceedings.* IEEE, 56–65.

[33] Yiqin Lu, Chun Yu, and Yuanchun Shi. 2020. Investigating bubble mechanism for ray-casting to improve 3d target acquisition in virtual reality. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 35–43.

[34] Päivi Majaranta and Kari-Jouko Räihä. 2002. Twenty years of eye typing: systems and design issues. In *Proceedings of the 2002 symposium on Eye tracking research and applications*. 15–22.

[35] Diako Mardanbegi, Christopher Clarke, and Hans Gellersen. 2019. Monocular Gaze Depth Estimation Using the Vestibulo-Ocular Reflex. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research and Applications* (Denver, Colorado) *(ETRA '19)*. Association for Computing Machinery, New York, NY, USA, Article 20, 9 pages. https://doi.org/10.1145/3314111.3319822

[36] Diako Mardanbegi, Tobias Langlotz, and Hans Gellersen. 2019. Resolving target ambiguity in 3d gaze interaction through vor depth estimation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[37] Alex Olwal, Hrvoje Benko, and Steven Feiner. 2003. Senseshapes: Using statistical geometry for object selection in a multimodal augmented reality. In *The Second IEEE and ACM International Symposium on Mixed and Augmented Reality, 2003. Proceedings.* IEEE, 300–301.

[38] Abdul Moiz Penkar, Christof Lutteroth, and Gerald Weber. 2012. Designing for the eye: design parameters for dwell in gaze interaction. In *Proceedings of the 24th Australian Computer-Human Interaction Conference*. 479–488.

[39] Ken Pfeuffer, Benedikt Mayer, Diako Mardanbegi, and Hans Gellersen. 2017. Gaze + Pinch Interaction in Virtual Reality *(SUI '17)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3131277.3132180

[40] Jimin Pi and Bertram E Shi. 2017. Probabilistic adjustment of dwell time for eye typing. In *2017 10th International Conference on Human System Interactions (HSI)*. IEEE, 251–257.

[41] Ivan Poupyrev, Mark Billinghurst, Suzanne Weghorst, and Tadao Ichikawa. 1996. The go-go interaction technique: non-linear mapping for direct manipulation in VR. In *Proceedings of the 9th annual ACM symposium on User interface software and technology*. 79–80.

[42] Jenny L Reiniger, Niklas Domdei, Frank G Holz, and Wolf M Harmening. 2021. Human gaze is systematically offset from the center of cone topography. *Current Biology* 31, 18 (2021), 4188–4193.

[43] K. Ruhland, C. E. Peters, S. Andrist, J. B. Badler, N. I. Badler, M. Gleicher, B. Mutlu, and R. Mcdonnell. 2015. A Review of Eye Gaze in Virtual Agents, Social Robotics and HCI: Behaviour Generation, User Interaction and Perception. *Computer Graphics Forum* (2015).

[44] Simon Schenk, Marc Dreiser, Gerhard Rigoll, and Michael Dorr. 2017. GazeEverywhere: enabling gaze-only user interaction on an unmodified desktop PC in everyday scenarios. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 3034–3044.

[45] KLAUS SCHMIDTKE and JEAN A BÜTTNER-ENNEVER. 1992. Nervous control of eyelid function: a review of clinical, experimental and pathological data. *Brain* 115, 1 (1992), 227–247.

[46] Ludwig Sidenmark, Christopher Clarke, Xuesong Zhang, Jenny Phu, and Hans Gellersen. 2020. Outline pursuits: Gaze-assisted selection of occluded objects in virtual reality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[47] Ludwig Sidenmark and Hans Gellersen. 2019. Eyehead: Synergetic eye and head movement for gaze pointing and selection. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. 1161–1174.

[48] Oleg Špakov and Darius Miniotas. 2007. Visualization of eye gaze data using heat maps. *Elektronika ir elektrotechnika* 74, 2 (2007), 55–58.

[49] Anthony Steed and Chris Parker. 2004. 3D selection strategies for head tracked and non-head tracked operation of spatially immersive displays. In *8th international immersive projection technology workshop*, Vol. 2.

[50] RM Steinman, WB Cushman, and AJ Martins. 1982. The precision of gaze. *Human neurobiology* 1 (1982), 97–109.

[51] Sophie Stellmach and Raimund Dachselt. 2012. Look & touch: gaze-supported target acquisition. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2981–2990.

[52] HR Taschenbuch Verlag Schiffman. 2001. Sensation and Perception: An Integrated Approach.

[53] Alan Transon, Adrien Verhulst, Jean-Marie Normand, Guillaume Moreau, and Maki Sugimoto. 2017. Evaluation of facial expressions as an interaction mechanism and their impact on affect, workload and usability in an AR game. In *2017 23rd International Conference on Virtual System & Multimedia (VSMM)*. IEEE, 1–8.

[54] Lode Vanacken, Tovi Grossman, and Karin Coninx. 2009. Multimodal selection techniques for dense and occluded 3d virtual environments. *International Journal of Human-Computer Studies* 67, 3 (2009), 237–255.

[55] Boris Velichkovsky, Andreas Sprenger, and Pieter Unema. 1997. Towards gaze-mediated interaction: Collecting solutions of the "Midas touch problem". In *Human-Computer Interaction INTERACT'97*. Springer, 509–516.

[56] Kang Wang, Rui Zhao, and Qiang Ji. 2018. A Hierarchical Generative Model for Eye Image Synthesis and Eye Gaze Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[57] Yukang Yan, Chun Yu, Wengrui Zheng, Ruining Tang, Xuhai Xu, and Yuanchun Shi. 2020. FrownOnError: Interrupting Responses from Smart Speakers by Facial Expressions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[58] Difeng Yu, Qiushi Zhou, Joshua Newn, Tilman Dingler, Eduardo Velloso, and Jorge Goncalves. 2020. Fully-occluded target selection in virtual reality. *IEEE Transactions on Visualization and Computer Graphics* 26, 12 (2020), 3402–3413.