Gaze Modulated Disambiguation Technique for Gesture Control in 3D Virtual Objects Selection

Shujie Deng, Jian Chang, Jian Jun Zhang
National Centre for Computer Animation
Bournemouth University
Poole, Dorset, United Kingdom
Email: sdeng@bournemouth.ac.uk

Shi-Min Hu
National Laboratory for
Information Science and Technology
Tsinghua University
Beijing, China

***Abstract*—Inputs with multimodal information provide more natural ways to interact with virtual 3D environment. An emerging technique that integrates gaze modulated pointing with mid-air gesture control enables fast target acquisition and rich control expressions. The performance of this technique relies on the eye tracking accuracy which is not comparable with the traditional pointing techniques (e.g., mouse) yet. This will cause troubles when fine grainy interactions are required, such as selecting in a dense virtual scene where proximity and occlusion are prone to occur. This paper proposes a coarse-to-fine solution to compensate the degradation introduced by eye tracking inaccuracy using a gaze cone to detect ambiguity and then a gaze probe for decluttering. It is tested in a comparative experiment which involves 12 participants with 3240 runs. The results show that the proposed technique enhanced the selection accuracy and user experience but it is still with a potential to be improved in efficiency. This study contributes to providing a robust multimodal interface design supported by both eye tracking and mid-air gesture control.**

***Keywords*—Multimodal interaction; eye tracking; mid-air gesture; occlusion; inaccuracy; 3D selection**

I. Introduction

Multimodal interaction is essential for immersive user experiences so it has become increasingly popular with the promotion of Virtual Reality (VR) and Augmented Reality (AR) applications [1]. This is usually achieved by enabling multiple natural modes of communication including voice, gesture, eye tracking, body movement etc. Gaze+gesture [2] is an emerging multimodal interaction technique which integrates gaze modulated pointing with mid-air gesture control. This technique allows faster target acquisition and richer control expressions comparing to conventional interaction techniques such as mouse and keyboard. Particularly, interactions with a mouse either in 2D or 3D are both based on controlling a 2D pointer while gesture control is intuitively capable of 3D manipulation with a 3D virtual representation of the hand.

Although gaze modulated pointing is capable of faster target acquisition, fundamentally it is similar to mouse pointing as the gaze interacts with the virtual world also through a 2D point on the screen. To enable 3D interaction, these techniques need extra assistance to obtain the depth information. Ray-casting is widely used for this purpose which shoots a ray from the 2D point into the projection space. The first object who is intersected with the ray is selected as the target. Therefore, a 2D point penetrating the exact target object is essential for the selection accuracy.

In addition, a virtual world with abundant information must contain plenty of virtual objects. Dense presentation of the objects often leads to close proximity between objects or even occlusion as shown in Fig. 1.

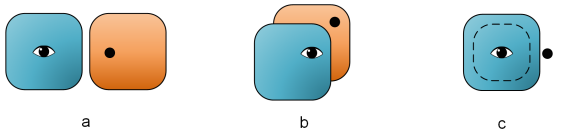

Fig. 1. Scenarios that gaze selection may be prone to errors in 3D interaction. The eye indicates where a user is actually looking at while the dot represents where the eye tracker thinks the user is looking at. Note there is an offset between the eye and the dot. (a)Proximity. (b)Partial occlusion. (c)Full occlusion. The dashed outline represents an object behind the blue cube.

Eye trackers can hardly achieve pixel-level accuracy due to two reasons. First of all, the algorithm that maps the captured eyes image to a point on the screen typically delivers some error ranging from 0.5 to 1.7cm wide [3]. Moreover, the physiological nature of our eyes constantly introduces tremor and natural random offsets [4]. It seems not a big problem when selecting in a virtual environment that is sparsely distributed with objects of decent sizes, but it can degrade selection efficiency in dense and occluded virtual scenes. As in Fig. 1 (a) and (b), the two objects are extremely proximate to each other, even partially occluded. If the purpose is to select the blue cube on the left and the user is already looking at it, the corresponding gaze point can still be mapped with a small offset to its neighbour on the right, and thus a selection error may occur.

The third scenario that not only requires accuracy but also accessibility to the depth is to select a fully occluded object. As shown in Fig. 1 (c), the target is perfectly occluded by the front blue cube. Ray-casting normally returns the front

The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union’s Seventh Framework Programme FP7/2007-2013/ under REA grant agreement No. 612627-“AniNex”. The research has been supported by Bournemouth University PhD Studentship.

object instead of the object behind, or a set of candidates intersecting with the ray for further disambiguation. It still requires accurate pointing on the exact position where the ray can go through the target, and accessing the depth at the same time. Thus, a solution is in demand to tackle both the planar offset and depth accessibility, which can take advantage of the features of gaze modulated techniques and gesture control without deteriorating the usability.

The proximity and occlusion problems were investigated previously using unimodal inputs such as game controllers [5]. As the Gaze+gesture technique has just recently emerged, no particular discussion is reported regarding whether the existing solutions still suit the new background. Thus, in this paper, we present how we adapt and develop the solutions in the Gaze+gesture background, and conduct a comparative study to understand the usability of our proposed technique. The result of the user study confirms that our technique can improve the accuracy of the Gaze+gesture technique in occluded environments although with an acceptable sacrifice of efficiency. This study aims to consummate the Gaze+gesture interaction design by supplementing the disambiguation technique and also to shed some light on future multimodal interaction design that involves gaze modulated pointing and gesture control, not only on desktop but also for VR and AR platforms.

The rest of the paper is outlined as follows. We first briefly review existing disambiguation techniques in section II. Then we introduce the details of our solution in section III. Further, we describe the aim and design of the comparative experiment in section IV, followed by its results in section V. We discuss the relationship between accuracy and efficiency, as well as limitations in section VI and then conclude the paper in section VII.

II. Related Work

Because the target cube in Fig. 1 (a) and (b) are still visible from the scene, the gaze selection ambiguity in these two scenarios can be alleviated using the same concepts. Thus, we define them as the proximity problem. We define the ambiguity in Fig. 1 (c) as the occlusion problem because the target cube is out of sight, and consequently, typical solutions for the proximity problem are not feasible here. In this section, we briefly review how previous studies tackle these two types of problems.

A. Proximity Problem

Stellmach and Dachselt [6] summarised the following solutions for the inaccuracy issue of gaze interaction. These methods are specifically developed for selecting small targets from crowded clutters, which is essentially the same issue with the proximity problem, so they can share the solutions.

1) Magnified Target: Because of the limited accuracy of the eye trackers, an intuitive solution is to enlarge or magnify the virtual target either visibly or invisibly. The most common technique is the two-step magnification ([6]–[8]) who follows a coarse-to-fine pattern. It divides the point-and-select task into separated pointing and selecting operations. The first step locates the surrounding area of the gaze and pops up a magnified view of this area. In the second step, if the target is inside this magnified area, the selection can be activated by a precise pointing on the target.

2) Manual Fine Selection: Gaze provides coarse estimation of the target location while some hand-based manipulation can achieve precisely fine selection, such as a mouse, so combining gaze and hand-based manipulation can compensate the gaze inaccuracy. It still follows the idea of the two-step coarse-to-fine selection. In the first step, the gaze conducts a coarse selection, then in the second step, the manual input enforces the fine selection. For example, the MAGIC pointing [9] firstly warps the mouse cursor to the vicinity of the gaze position, then finish the fine selection using the mouse.

3) Target Estimation: Prediction approach is used to model users' visual attention so that the object of interest can be estimated and corrected from the noisy gaze data. For example, Salvucci and Anderson [10] described a probabilistic algorithm in order to interpret gaze focus in a WIMP (Window, Icon, Menu, Pointer) example. This method estimates the object of interest based on the semantic meaning of the visual contents and the impact of the context.

B. Occlusion Problem

The previously reviewed techniques can help select an object that is still visible. To select a fully occluded object, we need to firstly see it. Based on Elmqvist and Tsigas's taxonomy of 3D occlusion management [11], there are five types of solutions to improve visibility of the occluded objects: multiple views, transparency, distortion, volumetric probes, and tour planner. Tour planner presents all targets in a scene by precomputing a path through all of them, and then guides users to interactively explore the scene following the path. It is not very common in object selection tasks, so we only discuss the first four techniques in particular.

1) Multiple Views: An intuitive reaction when a full occlusion occurs is to change the viewing perspective. In 3D modelling software such as CAD and Maya, multiple views in three orthogonal perspectives are provided, as well as an interactive way to manually rotate the model or the camera. Guidelines of multiple views system design were presented by Wang Baldonado et al. [12].

2) Transparency: The idea takes advantage of transparency to reveal the occluded object. The basic concept is to directly remove part of the occluding layer to show the details inside, such as complex anatomy and engineering graphs [13]. An interactive way is to allow the users to cut holes into the occluding object by themselves [14]. In order to retain the geometry information of the cut-away layer as a reference, semi-transparency [15] or phantom outlines of the transparent objects [16] can be applied.

3) Distortion: Usually, a linear projection is used to display a 3D virtual scene, typically perspective or parallel projection. An occluded object in one projection may be seen in another, so distortion uses this projective difference to reduce occlusion. The simplest way is to change the projection method of the whole scene. For example, Elmqvist and Tsigas [17] applied an animation to switch from perspective projection to parallel projection when occlusion occurred.

4) Volumetric Probes: This method utilises a volume instead of a point to coarsely select a set of candidates among which the final target is included for later fine selection. It conforms with the two-step coarse-to-fine pattern. There are several alternatives of the volume, for example, spotlight [18], sphere-casting [19], and cone [20]. It always involves a rearrangement of the coarsely selected candidates, typically reposition them to avoid occlusions and proximity.

The geometry of the volumetric probe can be used to define how the cluttered objects should be mapped to their new positions to maintain visual consistency, such as spherical scatter using the spherical BalloonProbe and wedge-shaped scatter using the wedge-shaped Balloonprobe [21].

Volumetric probes and distortion can also solve the proximity problem. Considering the consistency of interactions, it is desired to use the same interactive pattern under all circumstances, which can preferably reduce user's learning time and confusion during interactions [22]. Among the two, distortion is preferred by global tasks because it provides more context information while volumetric probe deals with local scope. Considering gaze is a natural local filter, we design a volumetric probe using a gaze cone and a gaze probe to solve both proximity and occlusion problems in this study.

III. Design of Gaze Modulated Disambiguation

Here we describe the details of how we solve the proximity and occlusion problems using the multimodal features of gaze and gesture. This design follows the coarse-to-fine selection patterns with the two steps: ambiguity detection and decluttering.

A. Conical Ambiguity Detection

We use a right circular gaze cone [23] that is invisible to the users to realise volume selection, so small targets and missed targets caused by the gaze mapping offset are captured. Fig. 2 shows an example of a gaze cone. The height of the cone should be long enough to reach the far clipping plane of the camera. Typically, the cone is always centred with the user's gaze ray, so from the user's perspective, the cone always looks like a circle, as shown in Fig. 2a. We can define the size of the cone using the diameter of the circle. The distance between the screen and the user usually retains in a limited range, if we treat it as a fixed value, we can also set the size of the cone to a defined value. If the size of the cone is too large, too many objects will be included so the filtering effect is not significant. If the size is too small, there will be little difference with the ray-casting selection and our problem remains unsolved. Here we set the size of the cone about 5° of visual angle. The visual acuity area extends about 10° around the centre of the retina combining the 2° highest visual acuity area in the centre and 4° parafovea around it [24], so 5° is the median size of the 10° visual acuity angle. It can tolerate upto 2.5° eye tracking errors.

We consider an object is inside the cone if its centroid is inside. If more than one object are inside the cone, ambiguity exists. The objects inside the cone are defined as the ambiguous candidates.

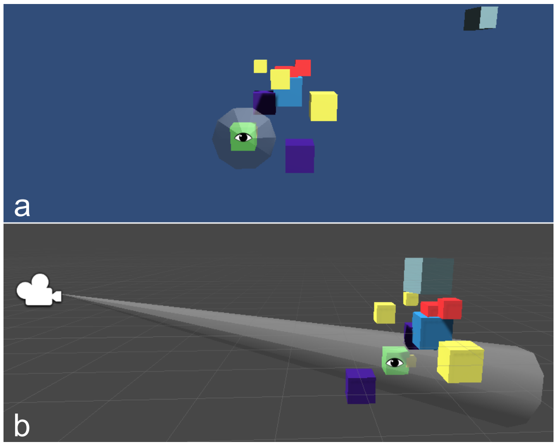

Fig. 2. Illustration of a gaze cone example. (a)The scene that a user sees. The user is looking at the green cube. The top of the cone is determined by transforming the 2D gaze position on the screen to the camera's near clip plane. (b)The right view of the scene. The gaze cone shoots from the camera's near clip plane in the direction to the gaze. In this frame, there are two selection candidates, the green cube and the small yellow cube behind it.

B. Gaze Probe Decluttering

Once a set of ambiguous candidates are determined, we want to declutter them. Firstly, we find the average centroid of all the candidates. Using this average position as the centre, we reposition all the candidates around it like a circle with equal intervals. As illustrated in Fig. 3b, the gaze probe is a circle centred with the average position but not the centre of the gaze cone projection. It is because gaze lacks the depth dimension which is required to set the new positions of the candidates. Also, gaze keeps jittering, but the candidates that the gaze cone covers do not change too much with the jittering. This can filter the gaze input and stabilise the new positions of the ambiguous candidates.

Because the gaze cone projection is a circle, we design that the candidates declutter in a circular pattern for visual consistency. In BalloonProbe [21], the objects are projected onto a sphere surface as their new positions, so the candidates are with different depths. We instead separate the objects into their new positions with the same depth which is determined by the average centre so that they are scattered on the same vertical plane. This is for avoiding new occlusions after the decluttering.

We use a mask to blur out the other objects in order to stand out the ambiguous candidates and the users can only select from the outstanding objects (Fig. 3a). The background is gradually blurred out and in the meantime, the objects are animated from their original positions to the new positions. To avoide the Midas Touch problem, the mask is triggered by a combination status of the gaze and gesture. When it detects

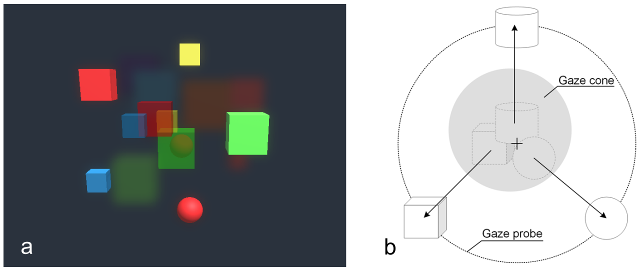

Fig. 3. Gaze probe. (a) An exmaple of a scene when a gaze probe is applied. The semi-transparent cubes in the middle are the original locations of the ambiguous candidates. Note they are not displayed in real applications because of distraction. (b) Overview of the gaze probe. The grey disk is the projection of the gaze cone and the central crosshair indicates the average centroid of the three candidates that are covered by the grey disk. The cluttered objects are separated equally around the circular gaze probe which is centred with the crosshair. The centre of the gaze cone projection is not necessarily aligned with the crosshair.

an eye fixation over $200ms$ and the hand is swiping towards the gaze position, the ambiguous candidates are determined and they will start to separate in the circular pattern. Once the selection is aborted or the target is locked and confirmed, the mask disappears and the distracting candidates recover their original positions. Whether the selected target recovers its position depends on the purpose of the tasks. For visualisation purposes only, it may recover its position with a highlighted visual effect. For manipulation purposes, it may stay at the new position and moves with the hand.

The fixation threshold is set to $200ms$ because a delay over $250ms$ can be clearly perceived and the users may start to feel the system is lagging [25]. Less than $200ms$ may make the system too sensitive to prevent unexpected triggering.

The radius of the gaze probe cannot be too small as it is difficult to accommodate the candidates with clear gaps. A clear gap should satisfy the condition that no ambiguity will be detected by the gaze cone in the new positional layout. The radius cannot be too large either because it will cost users more time to relocate the final target for the fine selection, either manually or by gaze. However, we prefer to use gaze selection here because gaze is much faster in target acquisition when there is no proximity or occlusion problems.

C. Gesture Control

We designed three intuitive gestures in a selection task, swipe, palm close and palm open. To select an object, a user only needs to fixate their gaze on the target and then swipes their hand towards it. As we mentioned earlier the fixation should be over $200ms$. If there is no ambiguity, the object will be highlighted, then a palm close gesture can select it. If there is ambiguity, the ambiguous candidates will be decluttered, and the user can further select the target using their gaze. In the gaze selection, the object with gaze upon will be highlighted. It is still a palm close gesture that can confirm the gaze selection. To cancel a highlighted candidate or a decluttering, the user only needs to swipe their hand away. A highlighted candidate indicates a selection that is not yet confirmed by a palm close. To cancel a selection, i.e., the selection is already confirmed by a palm close, the user can simply open their palm.

IV. Experiment

The aim of the experiment is to evaluate the efficiency and accuracy of the gaze modulated disambiguation technique in selection tasks using eye tracking and gesture control, especially when the proximity or occlusion problem occurs. Therefore, we designed a task to select the only sphere from many distracting cubes. We have the following hypotheses based on the aim: comparing with the default Gaze+gesture technique, 1) the proposed technique has equivalent accuracy when no proximity or occlusion occurs; 2) the proposed technique can improve accuracy when proximity and occlusion occurs; 3) introducing disambiguation may degrade interactive efficiency.

To better understand the proposed technique, not only compared it with the default Gaze+gesture technique, we also compared it with the conventional mouse interaction. Thus, there are three techniques we compared under three conditions, no occlusion, partial occlusion (proximity), and full occlusion (see Fig. 4 for examples).

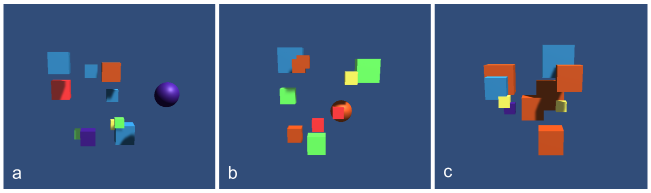

Fig. 4. Examples of the trial scene under three conditions. (a) No occlusion. (b) Partial occlusion (proximity). (c) Full occlusion.

The three techniques are:

a) Mouse (M): The cubes can be dragged by the mouse. The first click on the sphere indicates a successful selection. When occlusion occurs, users can drag the distracting cubes away to reveal the sphere.

b) Gaze+Gesture (GG): This is the default Gaze+gesture technique. To select an object, the user needs to look at it and make a gesture, for example, a grab. If the target is occluded, the user needs to move the occluding cubes away to reveal the sphere. Once the sphere is selected, the selection is marked as successful.

c) Gaze+Gesture+Disambiguation (GGD): This technique was elaborated in section III. One thing to add is that if the selected object is a cube, the participant can freely move it or open their palm to release/deselect it; if the object is a sphere, a successful selection will be admitted.

We measured the task completion time, errors and user preference for the usability comparison study.

A. Apparatus

As shown in Fig. 5, participants sat $55cm$ away from a desktop screen running the experiment built with the Unity3D game engine. The display was a Lenovo LS2323 23” wide LCD monitor with a frame rate set to $60Hz$. The resolution was 1920×1080. We used a Tobii EyeX tracker mounted on the bottom edge of the display with estimated 0.4 degrees of visual angle accuracy and the sampling rate used was the same as the frame rate. The viewing was binocular and the calibration was conducted with both eyes. The hand was tracked by a Leap Motion sensor placed facing up about $50cm$ away from the display and $17cm$ lower than the eye tracker. We used the SDK provided by Leap Motion for gesture recognition. The mouse was a Logitech M280 with the sensitivity set at 1000 DPI (dots per inch).

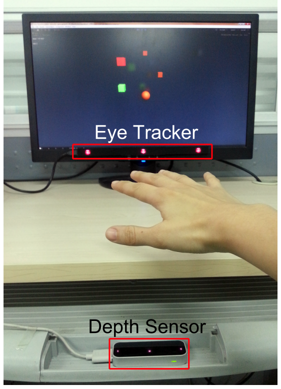

Fig. 5. Experiment setup.

B. Participants

Twelve volunteers (two females) participated in the study, aged 22 to 30 (Mean ± SD = 24.9 ± 2.3). None of the participants had any eye movement, hand movement or neurological abnormalities. The participants either had adequate natural visual acuity or corrected vision with glasses. All participants reported being right-handed. Written consent has obtained from each participant after explanation of the experiment. Before starting the tasks, participants were asked to answer some background questions by rating a 5-point Likert scale from $1-$ Strongly disagree to $5-$ Strongly agree. All the participants stated that they mainly use mouse and keyboard for computer interaction. Most participants never used mid-air gesture control except for four participants (Mean ± SD = 1.5 ± .8). As for using eye tracker as an interaction interface with computers, only one reported he had some experiences (Mean ± SD = 1.2 ± .6).

C. Procedure

The user study started with a brief introduction and a demographic questionnaire as described in the previous section. The participants were instructed to sit fairly still without restricting their head movements. Before recording the experiment data, a 7-point calibration was performed (three points separated equally on the top edge and the bottom edge of the monitor respectively, and one point in the middle). Then one technique at a time was described to the participants and they were asked to practice the technique until they felt confident. The practice usually took less than 5 minutes for each technique. After that we started to record the data of this technique. The order of the three techniques was randomised. Each technique had 90 trials tested, in which each occlusion condition was tested in 30 trials. The order of the 90 trials was randomised to guarantee that the occurrence of the three different occlusion conditions were randomised as well. Thus, 12 participants × 30 trials × 3 conditions × 3 techniques = 3240 runs were tested in total.

The scene was the same for each trial which contained one sphere and ten cubes spreading within 10° of visual angle in the middle of the full screen. The task was to select the only sphere amongst the ten cubes. There was another cube at the top right corner of the scene marking the destination of where to drag the sphere to. The destination was about 10° of visual angle away from the centre of the clutter containing the sphere and the cubes. The size, colour and position of the sphere and the cubes were randomly generated at the beginning of each trial. Once a successful selection was admitted, i.e., the sphere was selected, it would be automatically moved to the destination and disappear when it collided with the object. The automatic movement after the selection was only for indicating a successful trial, it was not included in the data recording because our aim was to evaluate the selection performance, not including the manipulation following it.

As selection is more about position acquisition, the orientation of the objects was eliminated from the task implementation in order to remove redundant noises. Thus, the gesture could only employ 3 degrees of freedom (DOF) on the virtual objects.

After all trials were completed, the participant was given a SUS (System Usability Scale) [26] questionnaire to fill for evaluation of each technique.

The whole process typically took 40 minutes to complete for each participant. The experimenter would remind the participants to have a break every 15 trials but they could skip the break and continue with the trials. They could still ask for a break at any time during the tests when they felt necessary.

D. Measures

The quantitative evaluation included three parts: the task completion time, the error count, and the SUS score.

a) Task completion time: The timer started when the scene was displayed and stopped when the sphere was grabbed and marked as “selected” when it was about to automatically move to the destination.

b) Error count: We measured two types of errors in the experiment, the *cube error* and the *decluttering error*. The cube error count (N_{ce}) increases by one when a cube is selected instead of the target sphere. The decluttering error count (N_{de}) increases by one when no successful selection is registered in a decluttered scene, i.e., a scene that the ambiguous candidates are presented in a circularly scattered way (Fig. 3a). In fact, based on the technique design, the cube error will hardly occur in Gaze+gesture+disambiguation, while the decluttering error will hardly occur in mouse and Gaze+gesture. The total error count of one trials is $N_{ce} + N_{de}$.

c) SUS score: The SUS questionnaire was presented with 10 statements with a 5-point Likert scale from 1− Strongly disagree to 5− Strongly agree. The range of a SUS score is between 0 and 100 from low to high satisfactory.

V. Results

To understand the usability of the disambiguation technique, we evaluated its efficiency by measuring the task completion time, and evaluated the accuracy by measuring the error count. The SUS score revealed user preference among the different techniques.

A. Usability

We evaluated the usability of the three techniques under three occlusion conditions, thus nine combinations of $technique \times occlusion$ were tested, for each combination we collected data of 30 trials from each participant. We obtained the average of the 30 trials from each participant and applied a repeated measures two-way ANOVA to estimate the impact of the task completion time and error count introduced by the different techniques and occlusion conditions. *Post hoc* Tukey test was applied to identify specific techniques and occlusion conditions who caused the significant differences. All statistical significance were determined at the level of 5%.

1) Efficiency: The variance analysis showed that the different techniques were associated with different completion times, $F(2, 22) = 13.18$, $p < .001$. The different occlusion conditions affected the completion time as well ($F(2, 22) = 103.2$, $p < .0001$). The result also yielded a statistical significance of the interaction between the two factors ($F(4, 44) = 13.22$, $p < .0001$). Thus, the task completion time depends on the technique and also the occlusion level.

We were more interested in the impacts of the techniques, so we further analysed which technique was significantly different with the others under each condition. Fig. 6 illustrates the task completion time for each technique grouped by the occlusion conditions.

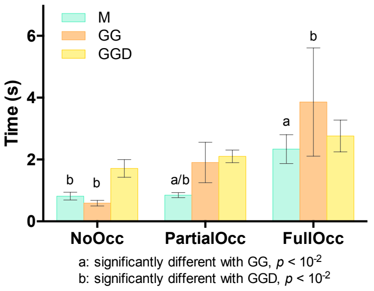

Fig. 6. Completion time for each technique under the three occlusion conditions. Error bar indicates the standard deviation.

The *Post hoc* results can be simply interpreted this way. When there was no occlusion, Gaze+gesture with disambiguation took significantly longer time than the other two; when there was partial occlusion, mouse was more efficient than the other two using Gaze+gesture; when there was full occlusion, Gaze+gesture was not as efficient as the other two while Gaze+gesture with disambiguation achieved a comparable efficiency with mouse. This result showed no advantage in efficiency of the disambiguation technique when occlusion was not severe. However, it did improve the selection efficiency to a close level of the mouse when full occlusion happened.

It was not our main concern about the impacts of the occlusion condition because we assumed that increasing the level of occlusion would lead to longer completion time in all techniques. We still briefly ran the Tukey test to testify this assumption. The result showed that only full occlusion impacted the efficiency of mouse. The two techniques using Gaze+gesture followed our assumption, only that the increasing rate of completion time was much larger when there was no disambiguation. Comparing to the dramatic increase of Gaze+gesture, adding the disambiguation yielded a fairly flat increase. It indicates that Gaze+gesture was extremely sensitive with occlusions and the disambiguation technique could largely alleviate this sensitivity.

2) Accuracy: The variance analysis result shows that the number of errors occurred was associated with the level of occlusion ($F(2, 22) = 220.5, p < .0001$) and interaction technique ($F(2, 22) = 247.8, p < .0001$). The interaction of this two factors also yielded significance, so the relationship between the number of errors and the techniques is affected by the level of occlusion ($F(4, 44) = 132.8, p < .0001$).

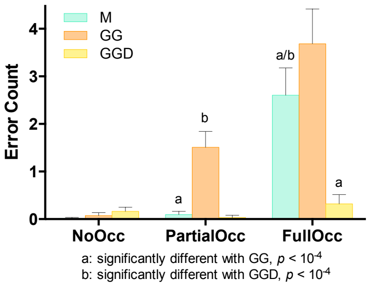

Fig. 7. Error count of each technique under each occlusion condition. Error bar indicates the standard deviation.

Fig. 7 shows the average number of errors occurred in a trial of each technique under each occlusion condition. Note that this result combined the cube errors and the decluttering errors. For a breakdown of the two errors in each bar please refer to Table I. The *Post hoc* test reveals that when there was no occlusion, the occurrence of both errors were close to zero in all three techniques; when there was partial occlusion, only Gaze+gesture had 1.51 cube errors while other techniques barely had any errors; when there was full occlusion, only Gaze+gesture featured with disambiguation had nearly zero occurrences of errors and the sorted accuracy among the three techniques is $GGD > M > GG$.

A comparison across the three different occlusion levels grouped by techniques shows that Gaze+gesture was prone to errors as long as there existed occlusion. The mouse pointing was capable of fine selection, so it could still per-

TABLE I

BREAKDOWN OF THE AVERAGE ERROR COUNT PER TRIAL. $\bar{N}_{ce}$ AVERAGE CUBE ERROR COUNT, $\bar{N}_{de}$ AVERAGE DECLUTTERING ERROR COUNT.

	NoOcc		PartialOcc		FullOcc	
Technique	$\bar{N}_{ce}$	$\bar{N}_{de}$	$\bar{N}_{ce}$	$\bar{N}_{de}$	$\bar{N}_{ce}$	$\bar{N}_{de}$
M	.014	0	.10	0	2.61	0
GG	.08	0	1.51	0	3.68	0
GGD	.003	.16	.01	.03	.003	.32

form well in the partial occlusion condition. However, in the full occlusion condition, the cube error count of the mouse technique would greatly surge. The fluctuation of the accuracy maintained fairly flat across all occlusion levels for Gaze+gesture+disambiguation. It suggests that this technique is robust to prevent both types of errors.

B. Preference

To evaluate the user preference among the three techniques, a SUS score between 0 to 100 of each technique was obtained. This score was an overall evaluation without considering the occlusion conditions separately. The SUS score (Fig. 8) from high to low is mouse (84.38), Gaze+gesture+disambiguation (80.83), Gaze+gesture (61.04).

Mouse is a mature and the most familiar interaction technique to the users, so it was supposed to score high as a reference for us to evaluate the other techniques. It could score higher if it was not tested in the full occlusion condition. It shows that our disambiguation technique helped Gaze+gesture achieve the second highest score even though the efficiency was degraded. Compared to the low score of Gaze+gesture, it reveals that users prefer steady accurate selection instead of inaccurate fast selection.

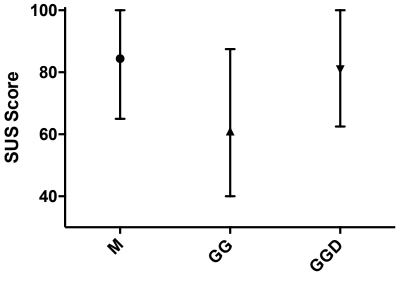

Fig. 8. SUS score mean of each technique. The top and bottom whiskers indicate the max and minimum score.

VI. DISCUSSION

Our three hypotheses of the experiment received positive answers. We validated that Gaze+gesture was highly accurate either with or without the disambiguation technique when there was no occlusion. While in occluded conditions, the disambiguation reduced the surging error count of the original Gaze+gesture down to nearly zero. However, the disambiguation was not advanced regarding efficiency, especially in the unoccluded condition. Mouse measured in the experiment as a benchmark also supported that the proposed technique did well in preventing errors but still not as fast as the mouse.

As Cashion et al. [5] commented, there was no best technique for all situations. Each technique is dependent on specific conditions to fit, or to be tailored, as the best solution.

A. Relationship between Accuracy and Efficiency

Overall, we observed that the efficiency was positively correlated with the accuracy of every technique. The task completion time increased with the growth of error count from no occlusion to full occlusion. Moreover, the coefficient of the correlation between completion time and error count was different with each technique. Some had fewer errors but longer completion time, and vice versa, such as mouse and Gaze+gesture+disambiguation under full occlusion.

However, as the unoccluded condition involves no errors, the completion time under this condition was irrelevant to the number of errors. It defined the nature of each technique. The Gaze+gesture+disambiguation technique is a good example. Its performance without occlusions was already slower than the others, so its poorer efficiency was not caused by the errors but the nature of its design. This could explain why the proposed technique had better accuracy but poorer efficiency.

The technique used a coarse-to-fine two-step design in selection. Compared to the direct selection in mouse and Gaze+gesture, it would double the completion time because it basically consists of two direct selections. As shown in the fully occluded condition, the completion time of the mouse increased because of the errors while the Gaze+gesture+disambiguation barely had any error. However, their completion times were close to each other. It indicates that a balance between the accuracy and efficiency has been reached using the Gaze+gesture+disambiguation technique in the fully occluded condition. Similarly, Gaze+gesture also reached a balance with our proposed technique in the partially occluded condition but only earlier because of its sensitivity to occlusions.

This is inline with the Fitts' Law [27] which points out that reaching a small target will cost more time. It suggests that accuracy and efficiency cannot be fulfilled at the same time in certain circumstances. Although not quite the same with reaching a small target, we would consider reaching an occluded object shares this common feature and requires a trade-off between accuracy and efficiency in interaction design. Elmqvist and Tudoreanu reported the same argument in an occlusion management comparison study [21].

B. Limitation

The current design of this disambiguation technique has limitations to be used when the ambiguous candidates are of various sizes and shapes. For example, a case that a small object fully occluded by a very large object that is much bigger than the gaze cone projection can be better dealt with using transparency or changing the viewport.

Selecting an occluded object does not always involve positional manipulation, such as for display-only purposes. Besides, repositioning loses the context of the original scene, so possible confusion can be introduced when the selected ambiguous candidates share the same features especially in shape and colour. In these cases, some further constraints and visualisation are necessary. For example, the selected object can return to its original position with other ambiguous candidates but visually highlighted for users to view.

Moreover, there is a limitation of the maximum number of ambiguous candidates the gaze probe can accommodate as the decluttering circle has limited space. Possible solutions are to add more layers of circles or to adjust the circle size but the usability remains unknown.

VII. Conclusion

This paper presented a two-step disambiguation technique to facilitate gestural selection in occluded 3D virtual environments, which combined a gaze cone for ambiguity detection, i.e., coarse selection, and a gaze probe to declutter the ambiguous objects for fine selection. The technique was evaluated in a user study to provide useful insight for further improvements. Using the presented technique we could prevent selection errors caused by inaccuracy of eye tracking. The enhanced user preference admitted the positive effect of the technique. By comparing with a well-established pointing technique, mouse, we understood that the current usability was acceptable but it could still be improved especially in efficiency to be pervasively used in daily life. Moreover, as long as the gaze modulated pointing exists, the gesture control can be replaced by other manual input methods, but the naturalness and feasibility may not be as good depending on the platform it is applied.

For future work, we firstly plan to improve the efficiency when no occlusion is involved. Furthermore, because it is awkward to declutter objects with significant size and shape difference, we need to adapt better solutions to facilitate a more robust and generic interaction technique. A possible approach could be context awareness which applies different techniques depending on the virtual contents covered by the gaze cone.

Acknowledgment

The authors would like to thank Dr Song-Hai Zhang and the anonymous reviewers for their valuable discussions and comments.

References

[1] G. Burdea, P. Richard, and P. Coiffet, "Multimodal virtual reality: Inputoutput devices, system integration, and human factors," *International Journal of Human-Computer Interaction*, vol. 8, no. 1, pp. 5–24, 1996.

[2] I. Chatterjee, R. Xiao, and C. Harrison, "Gaze+gesture: Expressive, precise and targeted free-space interactions," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, Seattle, Washington, 2015, pp. 131–138.

[3] A. Monden, K.-i. Matsumoto, and M. Yamato, "Evaluation of gaze-added target selection methods suitable for general guis," *International journal of computer applications in technology*, vol. 24, no. 1, pp. 17–24, 2005.

[4] O. Špakov, "Comparison of gaze-to-objects mapping algorithms," in *Proceedings of the 1st Conference on Novel Gaze-Controlled Applications*, Karlskrona, Sweden, 2011, pp. 1–8.

[5] J. Cashion, C. Wingrave, and J. J. L. Jr., "Dense and dynamic 3d selection for game-based virtual environments," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 4, pp. 634–642, 2012.

[6] S. Stellmach and R. Dachselt, "Look & touch: gaze-supported target acquisition," in *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*. ACM, 2012, pp. 2981–2990.

[7] C. Lankford, "Effective eye-gaze input into windows," in *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications*, ser. ETRA '00, Palm Beach Gardens, Florida, USA, 2000, pp. 23–27.

[8] M. Kumar, A. Paepcke, and T. Winograd, "Eyepoint: Practical pointing and selection using gaze and keyboard," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '07, San Jose, California, USA, 2007, pp. 421–430.

[9] S. Zhai, C. Morimoto, and S. Ihde, "Manual and gaze input cascaded (MAGIC) pointing," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Pittsburgh, Pennsylvania, 1999, pp. 246–253.

[10] D. D. Salvucci and J. R. Anderson, "Intelligent gaze-added interfaces," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '00, The Hague, The Netherlands, 2000, pp. 273–280.

[11] N. Elmqvist and P. Tsigas, "A taxonomy of 3d occlusion management for visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 5, pp. 1095–1109, 2008.

[12] M. Q. Wang Baldonado, A. Woodruff, and A. Kuchinsky, "Guidelines for using multiple views in information visualization," in *Proceedings of the Working Conference on Advanced Visual Interfaces*, ser. AVI '00, Palermo, Italy, 2000, pp. 110–119.

[13] W. Li, L. Ritter, M. Agrawala, B. Curless, and D. Salesin, "Interactive cutaway illustrations of complex 3d models," in *ACM SIGGRAPH 2007 Papers*, ser. SIGGRAPH '07, San Diego, California, 2007.

[14] C. Coffin and T. Hollerer, "Interactive perspective cut-away views for general 3d scenes," in *3D User Interfaces (3DUI'06)*, 2006, pp. 25–28.

[15] L. Chittaro and I. Scagnetto, "Is semitransparency useful for navigating virtual environments?" in *Proceedings of the ACM Symposium on Virtual Reality Software and Technology*, ser. VRST '01, Baniff, Alberta, Canada, 2001, pp. 159–166.

[16] J. Diepstraten, D. Weiskopf, and T. Ertl, "Transparency in interactive technical illustrations," *Computer Graphics Forum*, vol. 21, no. 3, pp. 317–325, 2002.

[17] N. Elmqvist and P. Tsigas, "View-projection animation for 3d occlusion management," *Computers & Graphics*, vol. 31, no. 6, pp. 864–876, 2007.

[18] J. Liang and M. Green, "Jdcad: A highly interactive 3d modeling system," *Computers & Graphics*, vol. 18, no. 4, pp. 499 – 506, 1994.

[19] R. Kopper, F. Bacim, and D. A. Bowman, "Rapid and accurate 3d selection by progressive refinement," in *2011 IEEE Symposium on 3D User Interfaces (3DUI)*, 2011, pp. 67–74.

[20] A. Steed, "Towards a general model for selection in virtual environments," in *3D User Interfaces (3DUI'06)*, 2006, pp. 103–110.

[21] N. Elmqvist and M. E. Tudoreanu, "Occlusion management in immersive and desktop 3d virtual environments: Theory and evaluation," *IJVR*, vol. 6, no. 2, pp. 21–32, 2007.

[22] T. Mandel, *The Elements of User Interface Design*. New York, NY, USA: John Wiley & Sons, Inc., 1997.

[23] A. Forsberg, K. Herndon, and R. Zeleznik, "Aperture based selection for immersive virtual environments," in *Proceedings of the 9th Annual ACM Symposium on User Interface Software and Technology*, ser. UIST '96, Seattle, Washington, USA, 1996, pp. 95–96.

[24] A. Duchowski, *Eye tracking methodology: Theory and practice*. London: Springer, 2007, vol. 373.

[25] J. Kangas, J. Rantala, D. Akkil, P. Isokoski, P. Majaranta, and R. Raisamo, "Delayed haptic feedback to gaze gestures," in *Proceedings of Haptics: Neuroscience, Devices, Modeling, and Applications: 9th International Conference, EuroHaptics 2014*, Versailles, France, 2014, pp. 25–31.

[26] J. Brooke, "Sus-a quick and dirty usability scale," *Usability evaluation in industry*, vol. 189, no. 194, pp. 4–7, 1996.

[27] P. M. Fitts, "The information capacity of the human motor system in controlling the amplitude of movement," *Journal of experimental psychology*, vol. 47, no. 6, pp. 381–391, 1954.