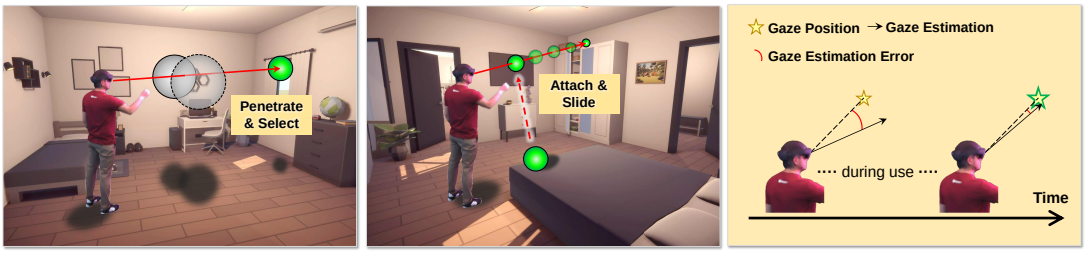
Exploring 3D Interaction with Gaze Guidance in Augmented Reality

Yiwei Bao* Jiaxi Wang† Zhimin Wang‡ Feng Lu§

State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, China



(a) Penetrate obstacles & accurate select **in 3D** (b) Fast Pickup & move to target position **in 3D** (c) Calibration on the fly: more accurate **in use**

Figure 1: We explore the potential of gaze as a guide for 3D interaction in AR. a) Accurate object selection in the case of heavy occlusions in 3D. b) Easy object translation in 3D by attaching to and sliding along the gaze beam. c) Our implicit online calibration avoids cumbersome explicit gaze calibration and allows gaze improvement during use.

Abstract

Recent research based on hand-eye coordination has shown that gaze could improve object selection and translation experience under certain scenarios in AR. However, several limitations still exist. Specifically, we investigate whether gaze could help object selection with heavy 3D occlusions and help 3D object translation in the depth dimension. In addition, we also investigate the possibility of reducing the gaze calibration burden before use. Therefore, we develop new methods with proper gaze guidance for 3D interaction in AR, and also an implicit online calibration method. We conduct two user studies to evaluate different interaction methods and the results show that our methods not only improve the effectiveness of occluded objects selection but also alleviate the arm fatigue problem significantly in the depth translation task. We also evaluate the proposed implicit online calibration method and find its accuracy comparable to standard 9 points explicit calibration, which makes a step towards practical use in the real world.

Index Terms: Human-centered computing—Human computer interaction (HCI)—HCI design and evaluation methods—User studies; Human-centered computing—Human computer interaction (HCI)—Interaction paradigms—Mixed/augmented reality

1 Introduction

Virtual Reality (VR) and Augmented Reality (AR) technologies break through the boundary between the digital world and the physical world. They have enabled a number of applications, such as online education and remote healthcare [10, 11, 49]. Compared with the conventional interaction medium, i.e., a 2D computer screen, VR/AR brings users an increased sense of immersion. While the mouse and keyboard are the two most commonly used tools for interaction with conventional devices, they are not designed for interaction in VR/AR, which is more challenging. Thus, developing efficient and intuitive interaction tools is necessary.

*joint first author, e-mail: baoyiwei@buaa.edu.cn
†joint first author, e-mail: wangjiaxi@buaa.edu.cn
‡e-mail: zm.wang@buaa.edu.cn
§corresponding author, e-mail: lufeng@buaa.edu.cn

In existing VR and AR devices, users usually interact with virtual objects only via hand rather than multimodal cues. For example, commercial VR devices such as HTC Vive and Oculus Quest use a hand-held controller as their primary interaction approach. These controllers are usually integrated with multiple hardware like sensors, buttons, and joysticks to provide various functions. For AR devices like Hololens, they are more a fan of the hand gesture as these systems usually equip one or more scene cameras that capture hands simultaneously.

Besides the hand, another important channel for our daily communication is the eye. Eye gaze indicates the places of interest, which offers ample spatial information. Gaze-based interactions have also been explored recently. Compared with hand-based interaction, gaze exhibited advantages and limitations. The advantages of gaze include: 1) Fast. The rotation speed of eyeball exceeds $300^\circ/s$ in certain motions [2]; 2) Intuitive. Gaze is a kind of natural behavior of we human beings, which points to the desired target. It happens all the time, so it does not require any extra effort. The shortcomings of gaze are also significant: 1) Calibration. Current gaze-based interaction systems usually require personal calibration before usage, and the accuracy of gaze is still unsatisfactory for precise interaction. 2) Midas Touch problem. Users may accidentally select the wrong object by just observing it. As a result, whether gaze is an efficient input modality remains to be explored.

A natural idea then is to take advantage of both hand and eye, by applying hand-eye coordination in VR/AR interactions. Hand-eye coordination techniques leverage both eye gaze and hand gesture to provide a better interactive experience. Since gaze is fast and intuitive while hand gesture is accurate, many studies follow the principle of "gaze selects, hand manipulates" [9, 29, 37] for a natural and fast selection experience. These methods have been proven to be efficient in scenarios with a limited number of large-size objects like furniture [47]. Some other studies combine gaze and hand to perform precise selection like menu selection in the same 2D plane [21]. Gaze has also been employed in object translation. Teleporting objects to the gazed position is easier than long-distance translation with a hand [21]. In a nutshell, gaze has been proven to be a beneficial modality when: 1) objects are not occluded with each other in 3D space, 2) translation mainly happens in lateral and longitudinal direction instead of depth direction, and 3) note that all of the above techniques rely on accurate gaze calibration in advance, which takes extra time and effort.

Considering the above limitations, we argue that the advantages

of gaze-based interaction have not yet been fully exploited. We aim to investigate whether gaze could help to deal with more complicated settings in AR with hand-eye coordination. Specifically, we investigate the *corresponding three challenges*: 1) Object selection in 3D, 2) Object translation in 3D, and 3) Gaze calibration on the fly. For the first challenge, the difficulty is about the occlusion between objects, since the target object may be largely invisible and untouchable to the user. For the second challenge, the difficulty lies in the translation in depth. Conventional hand translation methods require users to straighten their arms in midair to move objects away, which quickly leads to noticeable arm fatigue. Thus, we aim to use both gaze and hand to achieve more affordable object translation in depth. Finally, we claim that calibrating user gaze during interaction is possible. This requires new techniques to simplify or even avoid explicit gaze calibration before interaction, and still provide good gaze-tracking accuracy.

Consequently, this paper delivers new hand-eye coordinate methods for 3D interaction in AR and an implicit online gaze calibration method. We conduct two user studies to evaluate various interaction methods. In the first study, we evaluate different interaction methods in two different tasks related to selection in 3D and translation in 3D, respectively. The results show that with proper cooperation with the hand, gaze not only improves 3D selection efficiency but also alleviates arm fatigue problem significantly for 3D translation, especially in the depth dimension. In the second study, using our implicit online gaze calibration method, users can avoid cumbersome explicit calibration before the interaction task. The user's gaze is calibrated and improved quietly during the task and achieves even better accuracy than the standard 9 points calibration.

In summary, the contributions of this paper mainly include:

- We propose a hand-eye coordination method for object selection in 3D. Experiments show that the proposed method is efficient and accurate to select target objects within multiple occlusion objects.
- We explore the potential of gaze in 3D translation. With the design of sliding objects along the gaze beam, the arm fatigue problem in 3D translation is alleviated, especially when moving objects away along the depth axis.
- We propose an implicit online gaze calibration method that improves the usability of gaze by simplifying or even avoiding explicit gaze calibration. The proposed method improves both gaze estimation accuracy and actual interaction experience without users' active cooperation.

2 Related Work

2.1 Interaction Techniques in AR

As the next-generation device platform [19], AR systems have utilized different modalities like handheld controllers, freehand gestures, and eye gaze for interaction.

Handheld input. Handheld controllers are widely employed in early commercial AR devices [46]. These controllers consist of different hardware, *e.g.*, wearable sensors [5], buttons [40], joysticks [33], and smartphones [27, 51]. The handheld devices usually provide a ray-casted cursor for primary pointing and are used for confirming the selection. Bowman *et al.* first proposed to move the selected object closer or away by two buttons like a fishing-reel [7]. It has been reported that handheld input requires less physical demand than freehand gestures [46]. However, this input has lower user preference due to its traditional interaction way and external devices requirement [45].

Freehand gesture. Freehand gesture is commonly used in AR devices, due to its convenience and stable hand tracking [8, 31]. With spatially tracked hand position, users can select objects with the pointing cone attached to the hand [28] or even perform 6 degree-of-freedom (DOF) object manipulation in AR. Gesture-based interaction usually has higher accuracy than other modalities and thus offers fine-grained manipulation control [41]. Nevertheless, researchers have found that this modality falls short in supporting long-time usage due to noticeable arm muscle fatigue [15].

Eye gaze. More recent efforts have been made to explore gaze-based interaction [16, 42]. Gaze positions intuitively indicate where a user is interested and are 'always on'. Therefore, compared with hand-based interaction, this modality takes less physical effort [21]. However, limited by the development of gaze estimation technology, time-consuming calibration is necessary for gaze-based AR systems [44]. The Midas Touch problem also degrades the user experience, where the user triggers a wrong target with an unintentional glance [26].

In short, single-modal interactions have their own pros and cons [3]. To tackle these challenges, multimodal interaction has become a research hot spot [41]. Since eye movement is fast and effortless while hand-based interaction is accurate, many researchers try to find an optimal hand-eye coordination strategy to benefit from their complementary natures.

2.2 Hand-Eye Coordination Techniques

The combination of the eyes and hands lies in the core of our daily activities and interactions with surrounding objects [12, 50]. This form allows our eyes to guide the hand towards the target and can complete more fine-grained operations. Many researchers have investigated how to fuse these two modalities simultaneously [30,47]. These works mainly focus on two aspects: object selection and manipulation.

Object selection. Target selection has higher accuracy with gaze-gesture modality than using gaze-only input in AR [20, 41]. In these methods, the line of sight intersects with the first object in the scene or indicates the closest object. The selection is then confirmed once a pinch gesture is detected [47]. This modality has some limitations for crowded objects because the gaze drift might yield false positive selection errors [30]. Kytö *et al.* proposed a solution that the user could fine-tune the gaze position for menu selection with a pinch gesture, which has been proven to be effective [21]. However, the selection becomes challenging when the target is occluded by multiple objects [25]. Although many prior works have tried to solve selection ambiguity with different techniques [1, 4, 23, 34, 48], relatively few studies have focused on the gaze-guided selection of occluded objects. Sidnmark *et al.* proposed the "outline pursuits" technique that leverages the gaze to follow the moving stimulus around the occluded object to select it [35]. However, this method is suitable for the selection of partial occlusion. Lee *et al.* used a virtual mirror to reflect the occluded object from the side and selected it via gaze-based pointing [22]. This mirrored selection changes position relations between objects, which is less intuitive and natural. Therefore, we investigate how to naturally and effectively select fully occluded objects.

Object manipulation. Object manipulations include rotation, scaling, and translation [30]. In this work, we focus on the most common operations in the real world, *i.e.*, object translation. In this process, The prior works implement two different ways of hand-eye coordination [47]. One is following the principle of "gaze selects, hand manipulates" [9, 29]. After target selection, the user employs direct or indirect freehand gestures to translate the object without gaze control until gesture release. This method is widely used due to its intuitiveness and accuracy [37, 39]. However, as mentioned before, arm muscle fatigue is still the key problem after prolonged usage. Another way fuses gaze movement into the coarse translation process [36, 38]. The target attaches to the gaze ray and translates with eye movement. The pinch gesture releases to drop the object. This technique benefits from fast and effortless gaze movement. However,

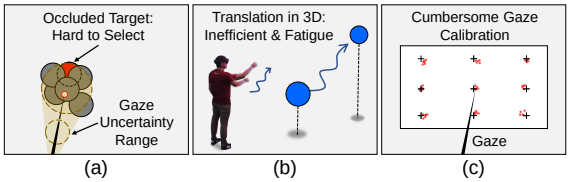


Figure 2: Challenges of hand-eye coordination techniques in AR.

this movement is 2 DOF and thus translates objects in the lateral and longitudinal directions. Therefore, we explore whether gaze could help to translate objects in a depth direction.

2.3 Near-eye Gaze Estimation and Calibration

Limited by the development of gaze estimation technology, calibration is necessary for a common head-mounted eye tracker. We discuss common gaze estimation methods and the role of calibration.

Pupil Center Corneal Reflection (PCCR) based Methods. PCCR based methods establish a 3D model of eye structures with corneal reflections [13, 24]. These methods calculate gaze direction based on the anatomical structure of the human eye. Usually, 9 points calibration is required to derive user-specific eyeball parameters like corneal curvature and eyeball radius. Although these methods are theoretically more robust to device slippage, they require extra devices like multiple infrared light sources.

Glint-Free Methods. Glint-free methods usually estimate the point of gaze (PoG) directly from pupil center coordinates [18]. Blignaut *et al.* proposed a polynomial mapping model to estimate the point of gaze [6]. The estimation error of these methods increases when device slippage occurs, as the mapping function from pupil center coordinates to PoG is changed. Santini *et al.* proposed a slippage-robust gaze estimation method [32]. For these methods, calibration is required to determine the personal parameters before use.

Recently, Kytö *et al.* proposed to further improve gaze accuracy during AR interaction on the basis of traditional gaze calibration [21]. A correction vector is determined by the user interaction procedure to rectify gaze estimation results. As we will discuss in Sec. 3.3, deviation of gaze could be more complicated than a simple bias. Thus we propose to calibrate the personal coefficients of users during the interaction.

3 Exploring Gaze for Interaction in 3D Space

Interacting with an object in 3D space mainly includes two steps: selection and manipulation. Here we choose the most common manipulation task, *i.e.*, translation, as an example. In this section, we first discuss the difficulties of selection and translation in 3D space. Then, we discuss the role of gaze during the interaction. Finally, we discuss the possibility of implicit gaze calibration brought by hand-eye coordination.

3.1 Object Selection in 3D

Object Selection Techniques in 2D. In the case without occlusion, selecting an object in 3D space is equivalent to selection in the 2D plane, since objects are not overlapped. There are various interaction modalities. Hand-based selection techniques are the most commonly used ones, where users select the target object through hand gestures directly or remotely. To achieve a more intuitive experience, gaze has also been employed. The objects are selected when the user's line of sight intersects with the objects, and such selection is usually confirmed by hand gesture [9]. Due to the limited accuracy, the estimated gaze points in a random direction within its uncertainty range, which may lead to the wrong object selection. To compensate for the low accuracy of gaze, eye-hand coordination techniques have been proposed, in which the gaze selection is fine-adjusted by hand [25]. These techniques have been proven to be effective for object selection in the 2D imaging plane when objects are close to each other like menus.

Difficulties of Object Selection in 3D. When an occlusion occurs, interaction becomes much more complicated. Since the occluded objects overlap in the 2D imaging plane, object selection has to be considered in 3D space, as shown in Fig. 2 (a). To address this, several 3D object selection techniques have been proposed, where they reposition potential objects to a non-occluded pattern such as a grid [48]. These techniques are less intuitive as they disturb the relative position of objects, especially when the appearance of objects is similar or even identical.

The Role of Gaze in 3D Object Selection. The role of gaze during 3D object selection has not yet been fully exploited. Sidenmark *et al.* proposed a gaze-assisted occluded object selection approach called Online Pursuits [35]. For each potential object, a stimulus moves around its outline and users select the target by following the stimulus with their gaze. The pursuit eye movement is usually slower than simple fixation. In fact, it takes around 4 seconds to confirm selection after the stimulus is displayed as reported in the paper. Still, Online Pursuits proves that, as an indicator of user attention, gaze helps to narrow down the range of potential objects. Thus, in this paper, we investigate what and how gaze could help in 3D object selection with occlusions.

3.2 Object Translation in 3D

Hand-based Object Translation. Object translation in AR is another typical 3D interaction task. Here, we refer to object translation in '3D' as the movement along the depth axis. Translating objects by hand has been widely adopted in AR systems. Users can either translate objects by hand directly similar to the interaction in the real world or select objects remotely. Hand-based translation techniques are popular due to their intuitiveness and accuracy. However, they are not perfect. For long-distance or long-time translation, hand translation techniques require users to move their arm in midair constantly, which causes arm fatigue and inefficient translation routes as shown in Fig. 2 (b).

Troublesome Depth Translation. To alleviate arm fatigue problem, gaze has been concluded to cooperate with hand in many recent studies. Objects are usually translated by gaze first to accomplish long-distance movements and then adjusted by hand later for precision [47]. Objects either follow the movement of the gaze or attach to the gaze location when triggered. Because the rotation of eyeball is fast and effortless, gaze is a good replacement to perform long-distance translation. However, for traditional gaze translation approaches, the problem is that gaze is only helpful for translation along a lateral and longitudinal direction. The reason is that gaze is usually defined as a direction vector without length. Although a human could focus on objects at different depths, accurate and smooth gaze depth estimation still remains unsolved [43]. In existing gaze-related interaction techniques, moving objects along a depth direction is still the duty of hand-only. Unfortunately, moving objects away in the depth direction by hand is extra tiresome because it requires users to straighten their arms in midair, which increases the workload of arms. Thus, in this paper, we investigate whether gaze could help to translate objects in 3D, particularly for translation in the depth direction.

3.3 Calibration on the Fly

Cumbersome but Necessary Gaze Calibration. Existing gaze-tracking techniques usually require personal calibration before usage. The most commonly-used calibration approach is explicit calibration, where the user is asked to stare at 5 or 9 target positions for a few seconds [6] as illustrated in Fig. 2 (c). The target positions and corresponding eye images are used to calibrate user-specific eye parameters, *i.e.*, eyeball radius and kappa angle. The explicit calibration process is tedious and significantly diminishes the practicality of

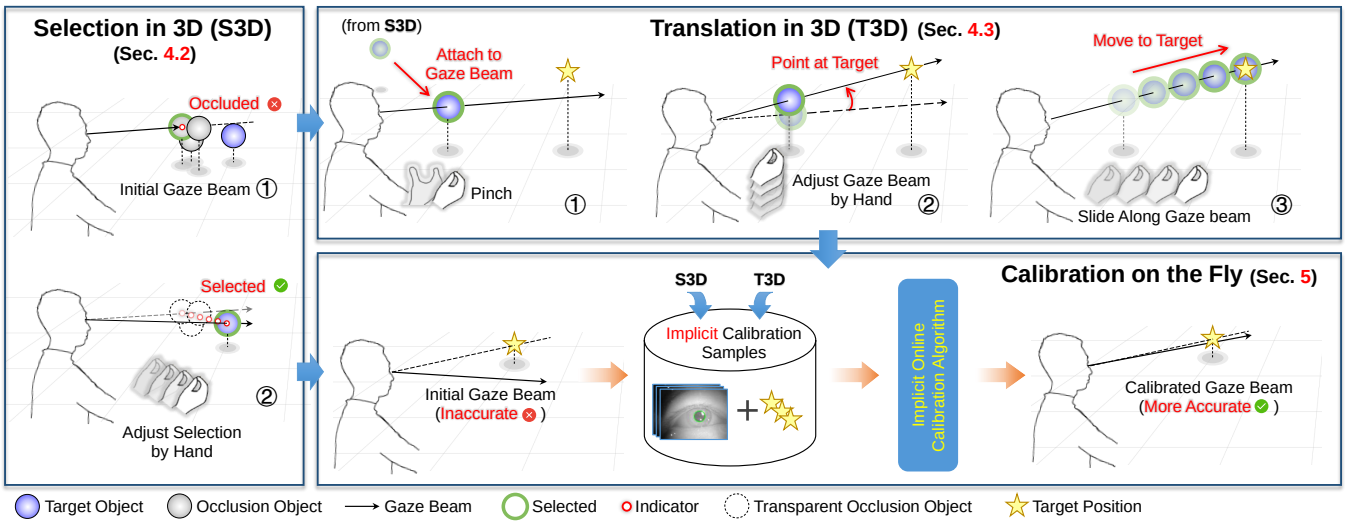


Figure 3: Overview of the proposed *Gaze Beam Guided Interaction* for 3D interaction in AR. By cooperating eye with hand, this method allows quick selection of occluded objects in 3D space. For translation in 3D, *Gaze Beam Guided Interaction* uses the gaze beam as the direction of object translation, and then allows a quick slide along the direction without lifting the user's arm to higher positions. In addition, we propose an implicit online calibration method to avoid explicit calibration and improve gaze estimation accuracy on the fly.

gaze. Even after gaze calibration, due to the device slippage caused by head movement, *etc.*, the deviation of gaze would increase significantly. Re-calibration could fix such deviation, however, frequent calibration hurts the user experience.

Benefit of Gaze Interaction: Calibration on the Fly. Recent studies have focused on gaze calibration on the fly, which is also known as implicit gaze calibration. Similar to explicit gaze calibration, samples of target gaze location and corresponding eye images are employed to correct the gaze deviation in an online manner. This brings us to another potential benefit of gaze-related interactions: the interaction content may contain clues about where the user is looking. If these clues are decrypted, it is possible to calibrate the user's gaze continuously and implicitly during the interaction, which saves much time and effort. Following this idea, previous work proposed calibrating gaze with a single correction vector derived from the interaction process [21]. We argue that the deviation of gaze in practice is polygenetic. The deviation could be caused by different factors, *e.g.*, bad calibration, device slippage. Implicit calibration should be considered in the estimation process of gaze instead of a simple offset on the result. Thus, in this paper, we aim to find a more reasonable way to calibrate user gaze implicitly during the normal interaction process.

4 Interaction Technique Design

Based on the above identified research topics, we developed three gaze-related interaction techniques and an implicit online calibration algorithm. In this chapter, we introduce these three interaction techniques to study object selection and translation in 3D. The implicit online calibration algorithm we proposed will be introduced in Sec. 5. Here, we implemented the developed techniques in Microsoft HoloLens2. These techniques could be easily applied to other HDM devices with pupil detection and hand gesture detection.

4.1 Remote Hand (RH)

With *Remote Hand*, we follow the common principle of "gaze select, hand manipulate". Users select the object with their eye gaze and use a pinch gesture to confirm the selection. If the user's gaze intersects with an object, the object is selected directly. Else, the closest object within 30cm range around the gaze ray is selected. The selection range is chosen according to gaze estimation deviation in the pilot test. If an object is selected, translation is performed by the user's hand remotely while keeping the pinch gesture. The translation is stopped once the pinch gesture is released. While moving objects, the moving distance of the hand is amplified to expand the range of translation. This is the baseline method following [47].

4.2 Gaze Position Guided Interaction (GP)

Gaze Position Guided Interaction cooperates gaze with hand in both the selection and translation stages. The user first points at the target with gaze and then adjusts the selection target by hand remotely when pinching. Once a pinch is detected during the selection stage, a small indicator is displayed at the gaze position at a fixed distance (2 meters). The indicator is adjusted by hand remotely in 3D space so that users can skip occlusions and select the occluded target directly. The closest object to the globe is selected once the pinch gesture is released. Objects that are closer than the indicator become transparent so that occluded objects can be observed. After an object is selected, the object is attached to the gaze position by pinch gesture. Users could move the object in 3D space by remote hand until the pinch is released. In *Gaze Position Guided Interaction*, we cooperate gaze with hand for occluded object selection. Although gaze is involved in the translation mechanism, the translation in depth dimension is still performed by hand alone in *GP*.

4.3 Gaze Beam Guided Interaction (GB)

Gaze Beam Guided Interaction applied the same object selection technique as *Gaze Position Guided Interaction*. The translation of *Gaze Beam Guided Interaction* follows the key idea of "gaze beam guided translation", which consists of two steps. In the first step, like *Gaze Position Guided Interaction*, the selected object attaches to the gaze position on a pinch. At this point, the hand could only adjust the direction of the gaze ray. The first step finishes when the pinch is released and the direction of the gaze ray is fixed. The user enters the second step by pinching again. In the second step, the user moves the object away or closer along the fixed gaze ray by moving the hand away or closer to him/herself. The release of pinch indicates the end of translation. In *Gaze Position Guided Interaction*, the moving direction of the object is identical to the pinching hand.

Table 1: Definition of three kinds of implicit gaze calibration samples.

No.	Time	Assumed PoG
1	0.1s before user release pinch and confirm selection in selection phase.	center of the indicator.
2	0.1s before user finish adjusting gaze beam.	center of the selected object.
3	0.1s before user finish adjusting object alone gaze beam .	center of the translation object.

In *Gaze Beam Guided Interaction*, since the moving direction of the object is determined as the direction of gaze, the depth of the object is controlled by the distance between the pinching hand and users, neglecting the specific moving direction of the hand.

5 Calibration on the Fly

In this section, we introduce the Implicit Online Gaze Calibration (IOGC) method, which simplifies the traditional explicit calibration procedure. The proposed method aims to calibrate user gaze implicitly in the process of experiencing *GB* in Microsoft HoloLens 2 so that the traditional explicit 9 points calibration could be simplified for a better user experience. First, we introduce the gaze estimation method and the simplified explicit calibration procedure we implemented. Then, we define Implicit Calibration Samples that are generated during the interaction and calibrate user gaze without any interference to users.

5.1 Gaze Estimation Method

Previous studies have proved that fewer calibration points mean larger gaze deviation [6]. As discussed in Sec. 3.3, the deviation cannot be compensated by a simple offset on the estimation result. This challenge needs to be considered from a more fundamental aspect, *e.g.*, optimizing the estimation process of gaze directly. Unfortunately, Microsoft HoloLens 2 only provides access to the results of gaze estimation to developers. To study the calibration on the fly, we develop our own glint-free gaze estimation method following [6].

Specifically, we install two infrared cameras on HoloLens 2 to capture user eye images. The 2D point of gaze (PoG) $G = \{X, Y\}$ in the virtual screen is calculated by a polynomial model based on detected pupil center coordinates $\{x_p, y_p\}$ for each eye:

$$\begin{aligned} X &= a_0 + a_1 x_p + a_2 x_p^2 + a_3 y_p + a_4 y_p^2 + a_5 x_p y_p + a_6 x_p^2 y_p^2, \\ Y &= b_0 + b_1 x_p + b_2 x_p^2 + b_3 y_p + b_4 y_p^2 + b_5 x_p y_p + b_6 x_p^2 y_p^2, \end{aligned} \tag{1}$$

where a_i, b_i are personal coefficients that are normally determined through an explicit calibration like 9 points calibration. In our implementation, we define pupil center coordinates $\{x_p, y_p\}$ as the relative pixel coordinates to the pupil center coordinate while users are looking at the center of the screen (we named the pixel coordinates of the pupil center at this point as the reference coordinate). The final estimation is the average of two eyes' PoGs. To simplify the calibration process, we established a set of average coefficients by collecting the average of 15 users' personal coefficients offline. For a new user, the reference point is acquired during the interaction implicitly and user gaze is estimated by average coefficients. We refer to this method as average coefficients only (Avg.) in Sec. 7. Our final target is to alleviate gaze deviation by acquiring personal coefficients during user interaction.

5.2 Implicit Online Calibration Strategy

To optimize personal coefficients, we first find Implicit Calibration Samples which consist of $\{x_p^l, y_p^l, x_p^r, y_p^r, X, Y\}$ from the user's interaction behavior without the user's notice. $\{x_p^l, y_p^l\}$ and $\{x_p^r, y_p^r\}$ are pixel coordinates of the user's left and right pupil center and $\{X, Y\}$ are the PoG truth. Once enough Implicit Calibration Samples are collected, personal coefficients are calculated by minimizing the L2 error of PoG estimation according to Eq. (1). In other words, the question is to find where the user is looking at a certain moment.

Luckily, in *Gaze Beam Guided Interaction*, the cooperation between hand and eye gives us enough clues. Specifically, we collect Implicit Calibration Samples from the following assumptions: 1) In the object selection phase, when users finish adjusting the indicator, they are assumed to be looking at the indicator so that they can confirm that the indicator is at the right place. 2) In the first step of the object translation phase, users are assumed to be looking at the center of the object so that they know the gaze beam is aligned with the target position. 3) Similar to assumption 2, in step two of the object translation phase, when users finish object translation by releasing a pinch gesture, they are assumed to be looking at the object so that they can confirm that the object is moved to the right place. A detailed definition is shown in Tab. 1. Each time the user selects and translates an object, three samples are collected. We only use the latest 60 samples to calculate the personal coefficient so that the personal coefficient will be updated quickly when device slippage occurs.

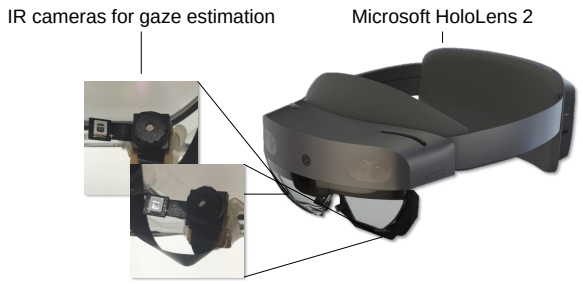


Figure 4: We conducted all the experiments on our modified Microsoft HoloLens 2. Two infrared cameras were installed on the bottom of the eye screens for the gaze estimation method we implemented.

6 Study 1

In study 1, we evaluate and compare three hand-eye coordination techniques we developed to investigate whether gaze could help object selection and translation in 3D. We design two tasks for all three interaction techniques to study object selection and translation, respectively. In the Heavy Occlusion task, participants need to select target objects within multiple occlusions and move them to four target positions. In the Varies Depth task, participants are required to move non-occluded objects to different target positions with a maximum depth distance of 6.5 meters.

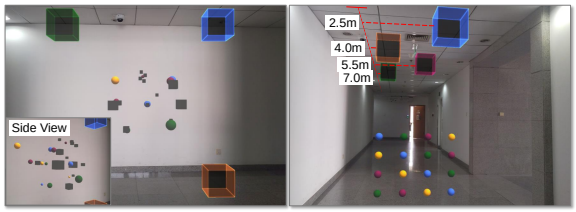
6.1 Participants and Devices

For study 1, we recruit 15 university students (14 males and 1 female) ranging in age from 22 to 29 years old (mean=24.2). All participants are familiar with computers and digital games.

We conduct our experiments using Microsoft HoloLens 2 on which we installed two infrared cameras with 30 FPS and 320 X 240 resolution for the gaze estimation method we implemented, as shown in Fig. 4. The software was implemented in C# with Unity3D.

6.2 Task Design

We design two search and move tasks in study 1, namely Heavy Occlusion (HO) task and Varies Depth (VD) task, respectively. In both tasks, the goal is to find spheres with four different colors and move them to target areas with the corresponding color as shown in Fig. 5. The radius of the spheres varies from 0.06 to 0.2 meters.



(a) Heavy Occlusion (HO) Task (b) Varies Depth (VD) Task

Figure 5: Illustration of two tasks in study 1. In the Heavy Occlusion (HO) Task, about half of the target objects (spheres) are completely occluded. There are 12 objects in the HO task which are shown in the side view.

Target areas are cube-shaped spaces 0.5 meters in length. Translation succeeds if the sphere is completely inside the cube. Head movements are allowed as long as the user remains in the original position. The parameters of the task are determined according to the common indoor interaction situations and fine-tuned during our pilot test.

Heavy Occlusion task: In the Heavy Occlusion task, there are 12 target spheres in total, *i.e.*, 3 spheres for each color. 12 gray cubes are distributed in the space as interference. Only 2 target spheres are completely visible to participants, while 6 are partially occluded and 4 are completely occluded. All objects are placed in a 1.5m X 1.2m space from 2m to 5m away from participants. Four target positions are on the corners of a 3m X 3m rectangle which is 3m away.

Varies Depth task: For the Varies Depth task, participants are asked to move objects away in the depth axis for different distances. There are 16 target spheres and no interference cubes. Target spheres are placed in a 1.2m X 0.9m grid that is 2.7m away. Target positions are 1.5m high and 2.5m, 4m, 5.5m, 7m away in the depth axis.

6.3 Evaluation Metrics

6.3.1 Objective Measures

Measurements for Selection in 3D: we define three objective metrics to evaluate the performance of participants in selection.

- **Total Selection Time:** total completion time minus total translation time. This is the time participants spend observing and selecting.
- **Total Selection Count:** the total number of selections participants have made during the HO task. More selection numbers means more redundant work.
- **Invalid Selection Ratio:** proportion of interference cube selection count to Total Selection Count. Lower ratios means more effective selection in 3D spaces with occlusion.

Measurements for Translation in 3D: we define four objective metrics to evaluate the performance of participants in translation in 3D. These metrics are evaluated on both tasks.

- **Hand Translation Distance:** the total distance of hand translations. Note that it is the accumulation of translation distance in each frame. Longer distance means more noneffective translation. This represents the accuracy of translation techniques.
- **Average Translation Count:** selection count of target objects divided by the number of target objects. This represents how many translations the participant takes to move an object to the target position successfully.

We also record the Completion Time of each task as an overall evaluation metric.

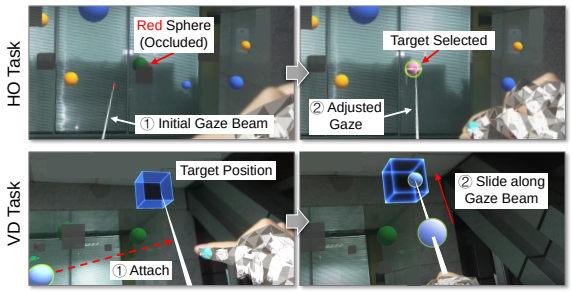


Figure 6: Illustration of the proposed *Gaze Beam Guided* technique's interaction procedure in the HO task and the VD task.

6.3.2 Subjective Measures

We also evaluate the techniques based on subjective measures of workload, arm fatigue, frustration and user preference. Subjective measures are collected after participants complete both tasks with each technique.

- **NASA-TLX [14]:** 21-point Likert scale to measure the mental demand, physical demand, temporal demand, effort, performance and frustration level of participants.
- **Borg CR10 [17]:** 10-point scale to measure the level of arm fatigue. It uses verbal anchors and numbers to map the magnitude of exertion to a scalar invariance scale.
- **Subjective Ranking:** a measure of participants' preference across all techniques.
- **Open Questions:** open questions about general evaluation, intuitiveness, frustration level, suggestion for improvement, and comparison to former techniques.

6.4 Experiment Procedure

Participants first fill in a pre-study questionnaire to collect how familiar they are with gaze-related interaction techniques, hand-related interaction techniques, and AR systems. Then, they are given a brief introduction to all three interaction techniques. We design a warm-up trial with random objects for participants to get familiar with all techniques. Participants could interact with objects freely in this trial until they fully master these techniques. Then, participants are introduced to the formal experiments. In the formal experiments, participants need to complete the HO task and the VD task in order by three techniques, respectively. The order of techniques is counter balanced following the Latin Square approach. As we study calibration on the fly in study 2, in study 1, participants complete a standard 9 points calibration at the beginning of each technique for accurate gaze estimation. After two tasks are completed by each technique, participants fill in a questionnaire to collect subjective measures and take a break for 5 minutes to relax their arms before the next technique. The whole experiment lasts about 68 minutes.

6.5 Results

6.5.1 Results of Objective Measures

The results of objective measures are shown in Fig. 7. We conducted repeated-measures ANOVAs ($\alpha = 0.05$) and post hoc pairwise t-tests to judge whether a certain metric is significantly different across techniques. Repeated-measures ANOVAs show that techniques have significant main effects on all subjective measures ($p = 0.0011$ for Completion Time in the HO task and $p < 0.001$ for others). We first analyze objective measures for selection. Results on the Total Selection Time demonstrated that *Remote Hand* spends significantly

Figure 7: Bar charts of the techniques' performance under different measurements. Error bars indicate the standard error. The statistical significances are labeled with ** ($p < 0.05$).

more time than others (*RH-GP*, $p = 0.023$; *RH-GB*, $p = 0.008$) in the HO task. *Remote Hand* spends about double the time on object selection than others, proving that the selection method we proposed in *GP* and *GB* achieves effective selection across multiple occlusions. Interestingly, *Remote Hand* also took significantly longer time to select in the VD task (*RH-GP*, $p = 0.002$; *RH-GB*, $p = 0.002$). We assume the reason is that objects in the VD task are close to each other. Although there is no occlusion in the VD task, it is still hard to select objects that are close by gaze alone. Results of the Total Select Count show that *Remote Hand* took significantly more selections to complete the HO task (*RH-GP*, $p < 0.001$; *RH-GB*, $p < 0.001$). There are two reasons for this. First, for *Remote Hand*, users must select and move occlusion objects away before selecting the target, while *GP* and *GB* could skip the occlusion objects. This assumption is proved by Invalid Selection Ratio where the ratio of invalid sections for *Remote Hand* is significantly higher (*RH-GP*, $p < 0.001$; *RH-GB*, $p < 0.001$). Second, it may take more than one selection and translation to move an object to its target location in *Remote Hand*. This is proved by the results of Average Translation Count which we will discuss next. Although Total Selection Time and Total Selection Count of *GB* in the HO task are significantly lower than *GP* ($p = 0.042, p = 0.01$, respectively), the values of these metrics for *GP* and *GB* are quite close (Total Completion Time:[133.66s, 83.27s, 70.25s], Total Selection Count: [40.8, 16.73, 14.73]) compared to *Remote Hand*.

Then, we analyze objective measures for translation. Results of Average Translation Count showed that *Remote Hand* made more attempts for moving an object to the target position for both the HO task (*RH-GP*, $p = 0.001$; *RH-GB*, $p = 0.001$) and the VD task (*RH-GP*, $p < 0.001$; *RH-GB*, $p < 0.001$). As the length of the arm is limited, the attach mechanism in *GP* and *GB* helps to reduce the translation distance of hand, saving the trouble of multiple moves. The results of Hand Translation Distance support our assumption. Hand Translation Distance of *Remote Hand* is significantly longer than others in both the HO task (*RH-GP*, $p < 0.001$; *RH-GB*, $p < 0.001$) and the VD task (*RH-GP*, $p < 0.001$; *RH-GB*, $p < 0.001$). Post hoc pairwise t-tests also show that the Average Translation Count of *GP* in the HO task is significantly larger than *GB* ($p = 0.042$), suggesting that "the gaze beam guided depth translation" in *GB* might be more accurate than remote hand in certain conditions.

Overall, the measurement of Completion Time shows that *Remote Hand* took significantly more time to complete the HO task(*RH-GP*, $p = 0.01$; *RH-GB*, $p = 0.021$), suggesting that *GP* and *Gaze Beam Guided Interaction* work better in occlusion settings. For task2, *GP* spends significantly less time (*RH-GP*, $p = 0.001$; *GP-GB*, $p = 0.016$). Compared to *GP*, the time disadvantage of *GB* may come from the multi-step translation design in translation demanding settings.

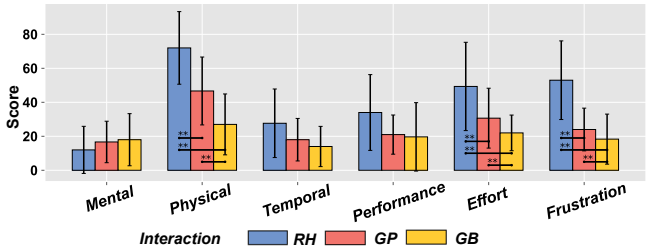


Figure 8: Bar charts of scores on the NASA-TLX questionnaire for the three interaction techniques. Error bars indicate the standard error. The statistical significances are labeled with ** ($p < 0.05$).

6.5.2 Results of Subjective Measures

Repeated-measures ANOVA on the NASA TLX questionnaire demonstrated that three techniques had a significant main effect on physical demand ($F_{2,28} = 22.656, p < 0.001, \eta^2 = 0.618$), effort ($F_{2,28} = 12.09, p = 0.002, \eta^2 = 0.463$) and frustration ($F_{2,28} = 14.928, p < 0.001, \eta^2 = 0.516$), as shown in Fig. 8. The physical demand for *Gaze Beam Guided Interaction* is significantly lower than other techniques (all $p < 0.001$). The physical demand for *GP* is also significantly lower than that for *Remote Hand* ($p = 0.0058$). Similarly, the effort of users had a trend of decreasing for three techniques (*RH-GP*, $p = 0.009$; *RH-GB*, $p = 0.002$; *GP-GB*, $p = 0.005$). The same trend continued for frustration (*RH-GP*, $p = 0.004$; *RH-GB*, $p = 0.001$; *GP-GB*, $p = 0.01$).

Repeated-measures ANOVA on other subjective measures demonstrated that three techniques had a significant difference in arm fatigue ($F_{2,28} = 44.352, p < 0.001, \eta^2 = 0.76$) and user preference ($F_{2,28} = 23.784, p < 0.001, \eta^2 = 0.629$), as shown in Fig. 7. Perception of arm fatigue decreases significantly for three techniques (for all pairwise comparison, $p < 0.001$), with *GB* causing the least arm fatigue. Interestingly, the Hand Translation Distant of *GP* and *GB* does not show a significant difference with respect to the arm fatigue level. This indicates that the moving distance of the hand does not reflect the level of arm fatigue, which we will discuss in Sec. 6.6.2. For preference, participants prefer *GB* to the other two

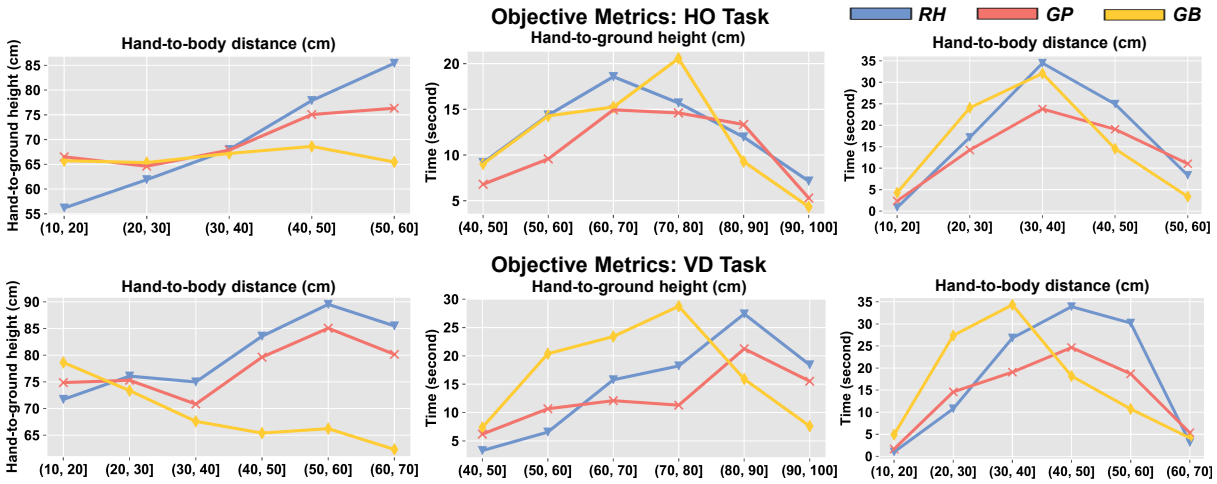


Figure 9: Objective metrics we defined to evaluate the level of arm fatigue. Results in (a) have shown that with our proposed *Gaze Beam Guided Interaction* techniques, users keep their hands at a lower position when their hands are distant from their bodies. Results in (b) and (c) have shown that users spend more time when their hands are lower and closer to their bodies with *Gaze Beam Guided Interaction*. The conclusion of these objective metrics coincides with the results of subjective measures, proving that *Gaze Beam Guided Interaction* alleviates arm fatigue problem significantly. In the HO task, hand-to-body distances are less than 60cm as the target positions are closer to participants.

techniques (*RH*-*GB*, $p < 0.001$; *GP*-*GB*, $p = 0.005$). They also prefer *GP* to *Remote Hand* ($p < 0.001$).

6.6 Discussion

6.6.1 Comments from Participants

The comments on open questions for participants ($N = 15$) also show some interesting patterns. Most participants commented that *Remote Hand* was "inaccurate" ($N = 9$) and "very tiresome for the arm" ($N = 10$), which could be the reason for obvious frustration in NASA TLX. Among those who commented that it was inaccurate, over half of them clearly mentioned that the inaccurate came from "hard to select the right object with gaze" ($N = 5$). We also found that some participants ($N = 5$) considered it "counter-intuitive and strange" to correct gaze deviation through eye rotation, because they had to "look away from the target". Overall, the performances of participants were significantly worse when the gaze estimation was less accurate in *Remote Hand*. In *GP* and *GB*, the impact of gaze estimation accuracy was minor because participants could easily adjust selection and manipulation outcomes by hand. This explains the bigger variation of *Remote Hand* in objective metrics.

Most participants ($N = 9$ for both techniques) commented that both *GP* and *GB* were "accurate". For *GP*, most participants ($N = 9$) still felt "tiresome for the arm", "especially for translations that are far from them" ($N = 5$). Only few participants ($N = 2$) felt "tiresome for the arm" for *GB* and some of them ($N = 3$) said it was "less tired". Few participants ($N = 2$) mentioned that the multi-step design was "less convenient" for *GP* ($N = 2$) and *GB* ($N = 4$). Some of them ($N = 4$) also mentioned that "it requires some learning to master *GB*". User comments coincide with the minor increasing mental demand in NASA TLX which is not significant.

6.6.2 Objective Measurements for Arm Fatigue

In subjective metrics, the results of Borg CR10 show that *Gaze Beam Guided Interaction*causes less arm fatigue than other techniques. To verify this conclusion, we seek some objective metrics to measure the level of arm fatigue. A simple way is to measure the total distance of hand movement for each technique. But it is obvious that moving the same distance at a farther position from the user's body should cause more fatigue as the arm of force is longer. Some previous works have modeled accurate levels of fatigue following the anatomy and physics of human arm [15, 17]. Although these methods usually require more detailed data like the condition of muscles, the basic idea that the level of fatigue depends on how much work the arm has done is still inspiring for us. Following the formula $W = fs$, where W, f, s represents work, force, and time, we decouple the level of arm fatigue to two factors: 1) the height of the hand across different distances to the user's body. 2) the duration of the hand in different heights and distances. The first factor represents f, *i.e.*, it takes more effort to raise your arm in a distant position because of the longer arm of force. The second factor represents s, *i.e.*, the duration of different forces.

The results of factor f are shown in Fig. 9 (a) and the results of factor s are shown in Fig. 9 (b) and (c). In Fig. 9 (a), it is obvious that with *GB*, users kept their hands in a lower position when their hands were distant from their bodies. This proves the advantage of the *Gaze Beam Guided Interaction*. When moving objects away in the depth axis, they only need to move their hands away from their bodies instead of moving their hands in the same direction as the object. In this way, users lowered their hands naturally when their hands were distant from their bodies, alleviating the arm fatigue problem. This pattern becomes more obvious for the VD task as it requires farther depth translation. The results of Fig. 9 (b) and (c) show that for *GB*, users spent more time while their hands are lower and closer to their bodies. As shown in the second row of Fig. 9 (b), the duration that users raise their hands to 90-100cm of *GB* is significantly lower because of the long-distance translation demand in the VD task. In Fig. 9 (c), users spent more time when their hands were close to their bodies with *GB*. This coincides with our design purpose for *GB* because users first adjust the gaze direction to align the object with the target position. As depth translation is not involved, this could be done with their hands close to their bodies. After which, users adjust the depth alone, making the adjustment in the depth axis simpler and thus faster. The duration is similar when users' hands reach the farthest distance because users need to straighten their arms to the limits for all techniques in order to reach the farthest target position. The overall results of these metrics verify the conclusion that *GB* causes significantly less arm fatigue.

6.6.3 Summary of Key Findings in Study 1

Based on the above results and analysis of Study 1, we summarize the following key findings:

- The hand-eye coordination selection technique we designed in *Gaze Position Guided Interaction* and *Gaze Beam Guided Interaction* achieves efficient and accurate selection within multiple occlusions. Objective measures show that it enables the ability to skip occlusions and takes less time to complete the selection with and without occlusion.

- Although gaze could not translate objects in the depth dimension directly, gaze provides guidance for the translation direction in the depth dimension. The proposed *Gaze Beam Guided Interaction* alleviates arm fatigue significantly for translation in 3D, especially in the depth dimension.

7 Study 2

In study 2, we evaluate the proposed Implicit Online Gaze Calibration method in actual user interactions. Participants are required to complete a search and move task similar to the Heavy Occlusion task in study 1 with *Gaze Beam Guided Interaction*. During this period, all Implicit Calibration Samples are recorded. After participants finish the task, we compare the accuracy of the online calibrated gaze estimation with other baselines.

7.1 Participants and Devices

We recruited 12 university students (9 males and 3 females) from 22 to 29 years old (mean=23.42) for study 2. All of them are familiar with traditional interaction techniques like mouse and keyboard. We used the same device as in study 1.

7.2 Task Design and Experiment Procedure

Similar to study 1, participants need to find spheres with different colors and move them to corresponding target positions. Implicit Calibration Samples are generated during the interaction process.

The specific experiment procedure is as follows. First, before the formal task, participants are introduced to a warm-up trial where they can interact with objects freely with *Gaze Beam Guided Interaction*. After they are familiar with *Gaze Beam Guided Interaction*, the formal task starts. At the beginning of the formal trial is a standard 9 points calibration. Note that the data of 9 points calibration will not be used during the whole task. This calibration is only used to establish baseline gaze estimation error for comparison. Then, participants press a button at the center of the screen to start the search and move task. We acquire the reference coordinate at this point quietly so that the user gaze is estimated by the Avg. method described in Sec. 5.1 during the task. 16 target globes are randomly placed in the frontal space of participants. The task is completed once they move all globes to corresponding target positions that split at the corners. Personal coefficients are derived from the Implicit Calibration Samples generated during the task. At last, participants are required to stare at 9 points for 2s each as the test set. The parameters of the calibration are determined according to the calibration procedure in the existing work [6] and our pilot test. We compare the angular error between the estimated gaze ray and the ground truth gaze ray of three methods: 1) Implicit Online Gaze Calibration strategy (IOGC); 2) standard 9 points calibration (9 Pts); 3) average coefficients only (Avg.). The formal trial lasts about 11 minutes. Device slippage is highly likely to happen as users are allowed to rotate their heads freely, causing the head acceleration.

7.3 Results and Discussion

The average gaze estimation error of all participants are shown in Fig. 10. We delete the results from P12 because the gaze estimation error in P12's trail is around 16°, which might be caused by absent-minded behaviour during testing.

Repeated-measures ANOVA test shows that the gaze estimation method has significant main effects on gaze estimation error ($F_{2,20} = 6.093, p = 0.014, \eta^2 = 0.379$). The error of IOGC (1.87°) is significantly smaller than 9 Pts (2.57° error, $p = 0.009$) and Avg. (2.55° error, $p = 0.032$). The results prove that the proposed IOGC method improves the performance of gaze estimation after a certain time of interaction. The accuracy of 9 Pts is similar to Avg.. We assume the reason is that the accuracy of 9 Pts is severely decreased by device slippage, while Avg. is derived from an average situation. The results of each participant are shown in Fig. 10. When 9 Pts and Avg. perform worse (most likely due to device slippage), IOGC method makes the most significant improvement as large as 2°.

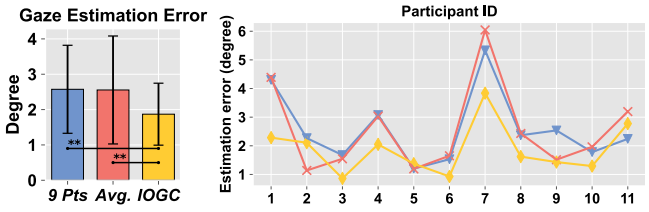


Figure 10: Comparison on average gaze estimation results of the three methods (left). Detailed results of 11 valid participants are shown on the right.

Overall, the estimation errors of the three methods have similar variation trends across different participants as they share the same gaze estimation method. The estimation error varies from around 1° to 6° due to factors like glasses, the appearance of the eye, *etc.* But the proposed IOGC method achieves the best performance in most situations (8 out of 11 participants).

8 Limitations, Future Work and Conclusion

8.1 Limitations and Future Work

The proposed *Gaze Beam Guided Interaction* causes significantly less arm fatigue. However, there are still some limitations. The two-step translation design of *Gaze Beam Guided Interaction* has made some participants confused. As two steps are triggered by the same gesture *i.e.*, pinch, a few participants ($N = 2$) mentioned that sometimes they were not sure about which step it was, especially when the user's pinch was wrongly detected. Some others ($N = 2$) also mentioned that it was inconvenient because they had to go through both steps even if they did not wish to adjust the depth.

In the future, we may cooperate current translation mechanism in *Gaze Beam Guided Interaction* with more than one gesture. For example, two steps in translation could be controlled by different gestures so that users could jump into any step as they wish.

8.2 Conclusion

In this research, we explore the potential of gaze in AR. Specifically, we investigate whether gaze could help the troublesome selection and translation in 3D. Results from the first study show that under proper cooperation with the hand, gaze could achieve efficient selection within multiple occlusions and alleviate arm fatigue problem significantly especially for translation in the depth dimension. Based on the intention of users revealed by the proposed interaction technique, we further propose an Implicit Online Gaze Calibration method which calibrates user gaze implicitly so that traditional explicit calibration is completely avoided. Experiments in study 2 have shown that the proposed method achieves even better accuracy than the standard 9 points calibration after a certain time of interaction. Overall, we show that gaze could help improve interaction experience in complicated settings, and gaze-based interaction techniques may free gaze estimation from annoying explicit calibration in the future.

REFERENCES

[1] N. A. B. Abdul Halim and A. W. B. Ismail. Raycasting method using hand gesture for target selection on the occluded object in handheld augmented reality. In *2021 6th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*, vol. 6, pp. 1–6, 2021. doi: 10.1109/ICRAIE52900.2021.9704035 2

[2] A. N. Angelopoulos, J. N. P. Martel, A. P. S. Kohli, J. Conradt, and G. Wetzstein. Event-based near-eye gaze tracking beyond 10, 000 hz. *IEEE Trans. Vis. Comput. Graph.*, 27(5):2577–2586, 2021. doi: 10.1109/TVCG.2021.3067784 1

[3] S. K. Badam, A. Srinivasan, N. Elmqvist, and J. Stasko. Affordances of input modalities for visual data exploration in immersive environments. In *Proc. Workshop on Immersive Analytics at IEEE VIS*, 2017. 2

[4] M. Baloup, T. Pietrzak, and G. Casiez. Raycursor: A 3d pointing facilitation technique based on raycasting. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, p. 1–12. Association for Computing Machinery, 2019. doi: 10.1145/3290605.3300331 2

[5] C. Bermejo and P. Hui. A survey on haptic technologies for mobile augmented reality. *ACM Comput. Surv.*, 54(9), 2021. doi: 10.1145/3465396 2

[6] P. Blignaut. A new mapping function to improve the accuracy of a video-based eye tracker. In *Proceedings of the South African Institute for Computer Scientists and Information Technologists Conference*, SAICSIT '13, p. 56–59. Association for Computing Machinery, 2013. doi: 10.1145/2513456.2513461 3, 5, 9

[7] D. A. Bowman and L. F. Hodges. An evaluation of techniques for grabbing and manipulating remote objects in immersive virtual environments. In *Proceedings of the 1997 Symposium on Interactive 3D Graphics, SI3D '97*, pp. 35–38, 182. ACM, 1997. doi: 10.1145/253284.253301 2

[8] N. Chaconas and T. Höllerer. An evaluation of bimanual gestures on the microsoft hololens. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces, VR 2018*, pp. 33–40. IEEE Computer Society, 2018. doi: 10.1109/VR.2018.8446320 2

[9] 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*, pp. 131–138. ACM, 2015. doi: 10.1145/2818346.2820752 1, 2, 3

[10] P. M. Emmelkamp, K. Meyerbröker, and N. Morina. Virtual reality therapy in social anxiety disorder. *Current psychiatry reports*, 22(7):1–9, 2020. doi: 10.1007/s11920-020-01156-1 1

[11] C. Faure, A. Limballe, B. Bideau, and R. Kulpa. Virtual reality to assess and train team ball sports performance: A scoping review. *Journal of sports Sciences*, 38(2):192–205, 2020. doi: 10.1080/02640414.2019.1689807 1

[12] S. Feng, W. He, S. Zhang, and M. Billinghurst. Seeing is believing: Ar-assisted blind area assembly to support hand–eye coordination. *The International Journal of Advanced Manufacturing Technology*, 119:1–10, 04 2022. doi: 10.1007/s00170-021-08546-6 2

[13] E. D. Guestrin and M. Eizenman. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Trans. Biomed. Eng.*, 53(6):1124–1133, 2006. doi: 10.1109/TBME.2005.863952 3

[14] S. G. Hart. Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, vol. 50, pp. 904–908, 2006. doi: 10.1177/154193120605000909 6

[15] J. D. Hincapié-Ramos, X. Guo, P. Moghadasian, and P. Irani. Consumed endurance: a metric to quantify arm fatigue of mid-air interactions. In *CHI Conference on Human Factors in Computing Systems, CHI'14*, pp. 1063–1072. ACM, 2014. doi: 10.1145/2556288.2557130 2, 8

[16] T. Hirzle, J. Gugenheimer, F. Geiselhart, A. Bulling, and E. Rukzio. A design space for gaze interaction on head-mounted displays. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019*, p. 625. ACM, 2019. doi: 10.1145/3290605.3300855 2

[17] S. Jang, W. Stuerzlinger, S. Ambike, and K. Ramani. Modeling cumulative arm fatigue in mid-air interaction based on perceived exertion and kinetics of arm motion. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, p. 3328–3339. Association for Computing Machinery, 2017. doi: 10.1145/3025453.3025523 6, 8

[18] M. Kassner, W. Patera, and A. Bulling. Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing: Adjunct publication*, pp. 1151–1160, 2014. doi: 10.1145/2638728.2641695 3

[19] J. J. Kim, Y. Wang, H. Wang, S. Lee, T. Yokota, and T. Someya. Skin electronics: Next-generation device platform for virtual and augmented reality. *Advanced Functional Materials*, 31(39):2009602. doi: 10.1002/adfm.202009602 2

[20] T. Kosch, A. Matviienko, F. Müller, J. Bersch, C. Katins, D. Schön, and M. Mühlhäuser. Notibike: Assessing target selection techniques for cyclist notifications in augmented reality. *Proc. ACM Hum.-Comput. Interact.*, 6(MHCI), 2022. doi: 10.1145/3546732 2

[21] M. Kytö, B. Ens, T. Piumsomboon, G. A. Lee, and M. Billinghurst. Pinpointing: Precise head-and eye-based target selection for augmented reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2018. doi: 10.1145/3173574.3173655 1, 2, 3, 4

[22] J.-J. Lee and J.-M. Park. 3d mirrored object selection for occluded objects in virtual environments. *IEEE Access*, 8:200259–200274, 2020. doi: 10.1109/ACCESS.2020.3035376 2

[23] N. Li, Z. Zhang, C. Liu, Z. Yang, Y. Fu, F. Tian, T. Han, and M. Fan. 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*, CHI '21. Association for Computing Machinery, 2021. doi: 10.1145/3411764.3445537 2

[24] C. Lu, P. Chakravarthula, Y. Tao, S. Chen, and H. Fuchs. Improved vergence and accommodation via purkinje image tracking with multiple cameras for ar glasses. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 320–331. IEEE, 2020. doi: 10.1109/ISMAR50242.2020.00058 3

[25] M. N. Lystbæk, P. Rosenberg, K. Pfeuffer, J. E. Grønbæk, and H. Gellersen. Gaze-hand alignment: Combining eye gaze and mid-air pointing for interacting with menus in augmented reality. *Proc. ACM Hum. Comput. Interact.*, 6(ETRA):145:1–145:18, 2022. doi: 10.1145/3530886 2, 3

[26] P. Mohan, W. B. Goh, C. Fu, and S. Yeung. Dualgaze: Addressing the midas touch problem in gaze mediated VR interaction. In *IEEE International Symposium on Mixed and Augmented Reality, ISMAR 2018 Adjunct*, pp. 79–84. IEEE, 2018. doi: 10.1109/ISMAR-Adjunct.2018.00039 2

[27] A. Mossel, B. Venditti, and H. Kaufmann. 3DTouch and HOMER-S: intuitive manipulation techniques for one-handed handheld augmented reality. In *Virtual Reality International Conference - Laval Virtual, VRIC 2013*, pp. 12:1–12:10. ACM, 2013. doi: 10.1145/2466816.2466829 2

[28] A. Olwal, H. Benko, and S. Feiner. Senseshapes: Using statistical geometry for object selection in a multimodal augmented reality system. In *2003 IEEE / ACM International Symposium on Mixed and Augmented Reality (ISMAR 2003)*, pp. 300–301. IEEE Computer Society, 2003. doi: 10.1109/ISMAR.2003.1240730 2

[29] K. Pfeuffer, J. Alexander, M. K. Chong, and H. Gellersen. Gaze-touch: combining gaze with multi-touch for interaction on the same surface. In *The 27th Annual ACM Symposium on User Interface Software and Technology, UIST '14*, pp. 509–518. ACM, 2014. doi: 10.1145/2642918.2647397 1, 2

[30] K. Pfeuffer, B. Mayer, D. Mardanbegi, and H. Gellersen. Gaze + pinch interaction in virtual reality. In *Proceedings of the 5th Symposium on Spatial User Interaction, SUI 2017*, pp. 99–108. ACM, 2017. doi: 10.1145/3131277.3132180 2

[31] T. Piumsomboon, D. Altimira, H. Kim, A. J. Clark, G. A. Lee, and M. Billinghurst. Grasp-shell vs gesture-speech: A comparison of direct and indirect natural interaction techniques in augmented reality. In *IEEE International Symposium on Mixed and Augmented Reality, ISMAR 2014*, pp. 73–82. IEEE Computer Society, 2014. doi: 10.1109/ISMAR.2014.6948411 2

[32] T. Santini, D. C. Niehorster, and E. Kasneci. Get a grip: slippage-robust

and glint-free gaze estimation for real-time pervasive head-mounted eye tracking. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications, ETRA 2019*, pp. 17:1–17:10. ACM, 2019. doi: 10.1145/3314111.3319835 3

[33] G. Schall, E. Méndez, E. Kruijff, E. E. Veas, S. Junghanns, B. Reitinger, and D. Schmalstieg. Handheld augmented reality for underground infrastructure visualization. *Pers. Ubiquitous Comput.*, 13(4):281–291, 2009. doi: 10.1007/s00779-008-0204-5 2

[34] J. Schjerlund, K. Hornbæk, and J. Bergström. Ninja hands: Using many hands to improve target selection in vr. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21. Association for Computing Machinery, 2021. doi: 10.1145/3411764.3445759 2

[35] L. Sidenmark, C. Clarke, X. Zhang, J. Phu, and H. Gellersen. Outline pursuits: Gaze-assisted selection of occluded objects in virtual reality. In *CHI '20: CHI Conference on Human Factors in Computing Systems*, pp. 1–13. ACM, 2020. doi: 10.1145/3313831.3376438 2, 3

[36] S. Stellmach and R. Dachselt. Still looking: investigating seamless gaze-supported selection, positioning, and manipulation of distant targets. In *2013 ACM SIGCHI Conference on Human Factors in Computing Systems, CHI '13*, pp. 285–294. ACM, 2013. doi: 10.1145/2470654.2470695 2

[37] S. Stellmach, S. Stober, A. Nürnberger, and R. Dachselt. Designing gaze-supported multimodal interactions for the exploration of large image collections. In *NGCA 2011, First Conference on Novel Gaze-Controlled Applications*, p. 1. ACM, 2011. doi: 10.1145/1983302.1983303 1, 2

[38] J. Turner, J. Alexander, A. Bulling, D. Schmidt, and H. Gellersen. Eye pull, eye push: Moving objects between large screens and personal devices with gaze and touch. In *Human-Computer Interaction - INTERACT 2013 - 14th IFIP TC 13 International Conference, Proceedings, Part II*, vol. 8118 of *Lecture Notes in Computer Science*, pp. 170–186. Springer, 2013. doi: 10.1007/978-3-642-40480-1_11 2

[39] J. Turner, A. Bulling, J. Alexander, and H. Gellersen. Cross-device gaze-supported point-to-point content transfer. In *Eye Tracking Research and Applications, ETRA '14*, pp. 19–26. ACM, 2014. doi: 10.1145/2578153.2578155 2

[40] E. E. Veas and E. Kruijff. Handheld devices for mobile augmented reality. MUM '10. Association for Computing Machinery, 2010. doi: 10.1145/1899475.1899478 2

[41] Z. Wang, H. Wang, H. Yu, and F. Lu. Interaction with gaze, gesture, and speech in a flexibly configurable augmented reality system. *IEEE Trans. Hum. Mach. Syst.*, 51(5):524–534, 2021. doi: 10.1109/THMS.2021.3097973 2

[42] Z. Wang, H. Yu, H. Wang, Z. Wang, and F. Lu. Comparing single-modal and multimodal interaction in an augmented reality system. In *2020 IEEE International Symposium on Mixed and Augmented Reality Adjunct, ISMAR 2020 Adjunct*, pp. 165–166. IEEE, 2020. doi: 10.1109/ISMAR-Adjunct51615.2020.00052 2

[43] Z. Wang, Y. Zhao, and F. Lu. Control with vergence eye movement in augmented reality see-through vision. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 548–549, 2022. doi: 10.1109/VRW55335.2022.00125 3

[44] Z. Wang, Y. Zhao, and F. Lu. Gaze-vergence-controlled see-through vision in augmented reality. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–11, 2022. doi: 10.1109/TVCG.2022.3203110 2

[45] M. Whitlock, E. Harnner, J. R. Brubaker, S. K. Kane, and D. A. Szafir. Interacting with distant objects in augmented reality. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces, VR 2018*, pp. 41–48. IEEE Computer Society, 2018. doi: 10.1109/VR.2018.8446381 2

[46] J. Wither and T. Hollerer. Evaluating techniques for interaction at a distance. In *Eighth International Symposium on Wearable Computers*, vol. 1, pp. 124–127, 2004. doi: 10.1109/ISWC.2004.18 2

[47] D. Yu, X. Lu, R. Shi, H. Liang, T. Dingler, E. Velloso, and J. Gonçalves. Gaze-supported 3d object manipulation in virtual reality. In *CHI '21: CHI Conference on Human Factors in Computing Systems*, pp. 734:1–734:13. ACM, 2021. doi: 10.1145/3411764.3445343 1, 2, 3, 4

[48] D. Yu, Q. Zhou, J. Newn, T. Dingler, E. Velloso, and J. Gonçalves. Fully-occluded target selection in virtual reality. *IEEE transactions on visualization and computer graphics*, 26(12):3402–3413, 2020. doi: 10.1109/TVCG.2020.3023606 2, 3

[49] S. N. V. Yuan and H. H. S. Ip. Using virtual reality to train emotional and social skills in children with autism spectrum disorder. *London journal of primary care*, 10(4):110–112, 2018. doi: 10.1080/17571472.2018.1483000 1

[50] X. Zhou, Y. Jin, L. Jia, and C. Xue. Study on hand–eye cordination area with bare-hand click interaction in virtual reality. *Applied Sciences*, 11(13), 2021. doi: 10.3390/app11136146 2

[51] F. Zhu and T. Grossman. Bishare: Exploring bidirectional interactions between smartphones and head-mounted augmented reality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, p. 1–14. Association for Computing Machinery, 2020. doi: 10.1145/3313831.3376233 2