UNIVERSITY OF BATH

*Citation for published version:*
Velloso, E, Turner, J, Alexander, J, Bulling, A & Gellersen, H 2015, An empirical investigation of gaze selection in mid-air gestural 3D manipulation. in J Abascal, S Barbosa, M Fetter, T Gross, P Palanque & M Winckler (eds), *Human-Computer Interaction – INTERACT 2015.* Lecture Notes in Computer Science, Springer International Publishing, pp. 315-330. https://doi.org/10.1007/978-3-319-22668-2_25

*DOI:*
10.1007/978-3-319-22668-2_25

*Publication date:*
2015

*Document Version*
Peer reviewed version

Link to publication

# An Empirical Investigation of Gaze Selection in Mid-Air Gestural 3D Manipulation

Eduardo Velloso[1(✉)], Jayson Turner[1], Jason Alexander[1],
Andreas Bulling[2], and Hans Gellersen[1]

[1] School of Computing and Communications, Infolab21, Lancaster University,
Lancaster LA1 4WA, UK
{e.velloso,j.turner,j.alexander}@lancaster.ac.uk,
hwg@comp.lancs.ac.uk

[2] Max Planck Institute for Informatics, Perceptual User Interfaces Group,
Campus E1 4, 66123 Saabrücken, Germany
andreas.bulling@acm.org

**Abstract.** In this work, we investigate gaze selection in the context of mid-air hand gestural manipulation of 3D rigid bodies on monoscopic displays. We present the results of a user study with 12 participants in which we compared the performance of Gaze, a Raycasting technique (2D Cursor) and a Virtual Hand technique (3D Cursor) to select objects in two 3D mid-air interaction tasks. Also, we compared selection confirmation times for Gaze selection when selection is followed by manipulation to when it is not. Our results show that gaze selection is faster and more preferred than 2D and 3D mid-air-controlled cursors, and is particularly well suited for tasks in which users constantly switch between several objects during the manipulation. Further, selection confirmation times are longer when selection is followed by manipulation than when it is not.

**Keywords:** 3D user interfaces · Eye tracking · Mid-air gestures

## 1 Introduction

Interaction fidelity—the degree with which the actions used for a task in the UI correspond to the actions used for that task in the real world [1]—is an active topic of research in 3D user interfaces (3DUI). Interfaces based upon free-space spatial input (e.g. mid-air gestures, tilt and turn gestures, magnetic trackers, etc.) offer this fidelity for 3DUI due to their multiple degrees of freedom and high integration of dimensions of control (i.e. many degrees of freedom can be controlled simultaneously with a single movement) [2, 3]. In particular, recent advances in unobtrusive motion capture (e.g. Kinect, Leap Motion) created a renewed interest in mid-air gestures for 3DUI.

In immersive virtual reality environments and on stereoscopic displays, such interactions allow users to manipulate virtual objects using interaction metaphors that relate more closely to real world interactions, for example, by using an isometric mapping between the virtual and physical spaces, users can reach virtual objects directly where they see them. However, a large number of 3D activities, such as gaming, graphic design and 3D modelling are still mostly conducted on conventional monoscopic desktop displays. This setup creates a discontinuity between the physical and the virtual environments, and therefore does not allow users to directly grasp objects in three dimensions. In this desktop context, common mid-air interaction techniques for 3D selection are *Raycasting* (in which the user's hand controls a 2D point that determines the direction of pointing) and the *Virtual Hand* (in which the user controls a 3D representation of his hand and makes selections by intersecting it with virtual objects) [2]. See Argelaguet et al. for a survey of selection techniques for 3D interaction [4].

As the eyes provide a natural indication of the focus of the user's interest, eye trackers have been used for pointing in a wide variety of contexts without necessarily requiring a representation on the screen, showing higher speeds than conventional techniques [5]. Even though gaze pointing for computing input has been investigated since the 80's [6], studies on gaze pointing for 3DUI started with work by Koons et al., who built a multimodal interface integrating speech, gaze and hand gestures [7]. Early work was also conducted by Tanriverdi and Jacob, who found it to be faster than an arm-extension technique with a 6DOF magnetic tracker in a VR environment [8]. Cournia et al. found conflicting results that suggest gaze is slower than a hand-based Raycasting technique with a wand [9]. These works only investigated selection tasks, but in practice, common 3D interaction tasks involve further manipulation steps after selection, such as translation and rotation. Given that gaze alone is impractical for all steps, several works combined gaze with additional modalities, but few explored the context of 3D user interfaces. In particular, when using gaze for selection and mid-air gestures for 3D manipulation, is there a cost in performance in switching modalities?

Even though gaze has been explored in a variety of multimodal configurations [10], few works explored the combination of gaze and mid-air gestures. Kosunen et al. reported preliminary results of a comparison between eye and mid-air hand pointing on large screen in a 2D task that indicate that pointing with the eyes is 29 % faster and 2.5 times more accurate than mid-air pointing [11]. Hales et al. describe a system in which discrete hand gestures issued commands to objects in the environment selected by gaze [12]. Pouke et al. investigated the combination of gaze and mid-air gestures, but in the form of a 6DOF sensor device attached to the hand [13]. They compared their technique with touch, and found that the touch-based interaction was faster and more accurate.

The conflicting results in the literature highlight the importance of further work investigating gaze selection for 3DUI, particularly considering that technical advances made eye tracking technology significantly more accurate, precise and robust than the devices and techniques used in previous works. In this work, we present an investigation of gaze selection for mid-air hand gestural manipulation of 3D rigid bodies in monoscopic displays. We conducted a study with three tasks. In the first task, we compared three 3D interaction techniques for selection and translation: a 2D cursor controlled by the hand based on Raycasting, a 3D cursor controlled by the hand analogous to a Virtual Hand and Gaze combined with mid-air gestures. In the second task, we also compared the same three techniques but in a selection and translation task involving multiple objects. In our pilot studies we found that when participants used the Gaze + Mid-Air Gestures technique, they reached out for objects even though they did not have to. We hypothesised that this action was due to the clutching required for manipulation.

To test this hypothesis, users performed a third task, in which we compared the selection time in the case where users were only required to select an object to the case where they also had to translate the object after selecting it.

Our results show that gaze selection is faster and more preferred than conventional mid-air selection techniques, particularly when users have to switch their focus between different objects. We also discovered a significant difference in the time to pinch after the object was gazed at between selection only tasks and selection followed by translation, indicating that the context of the selection impact the selection confirmation time.

## 2 Related Work

### 2.1 Human Prehension

Prehension is formally defined as “the application of functionally effective forces by the hand to an object for a task, given numerous constraints” [14], or more informally as the act of grasping or seizing. Different authors proposed ways of modelling this process. In Arbib’s model, the eyes (perceptual units), arms and hands (motor units) work together, but under distributed control to reach and grasp objects [14, 15]. The perceptual schema uses the visual input provided by the eyes to locate the object and recognise its size and orientation. The motor schema can be divided into two stages: reaching (comprised of a quick ballistic movement followed by an adjustment phase to match the object’s location) and grasping (including adjusting the finger and rotating the hand to match the object’s size and orientation, followed by the actual grasping action). Paillard’s model begins with the foveal grasping, in which the head and the eyes position themselves towards to object. Then, according to shape and positional cues, the arms and hands locate and identify the object, in open and closed loops, until finally grasping it, performing mechanical actions and sensory exploration [14, 16].

In the context of mid-air gestures for 3D user interfaces, reaching is analogous to selection and grasping to the confirmation of the selection. In this work, we investigate how human prehension can be supported in a desktop monoscopic 3D environment. In all conditions we studied, grasping (confirmation) was performed by a pinch gesture, similar to how we would grasp physical objects, but the selection step varied across conditions. The 3D cursor includes a reaching step similar to normal prehension, only offset due to the discontinuity between the virtual and physical worlds. The 2D cursor also contains a reaching step, but only in two dimensions. The Gaze condition only requires foveal grasping, as when the user looks at the object, she only needs to pinch to confirm the selection. However, as we show in the results of task 3, when the user grasps the object for further manipulation, she still reaches out for it.

### 2.2 Mid-Air Interaction for 3D Manipulation

Due to our familiarity in manipulating physical objects with our hands, a considerable effort of the HCI community has been put into developing input devices and interaction techniques that leverage our natural manual dexterity to interact with digital content.

An important interaction paradigm in 3D interaction is isomorphism: a strict, geometrical, one-to-one correspondence between hand motions in the physical and virtual worlds [2]. Even though isomorphic techniques are shown to be more natural, they suffer from the constraints of the input device (e.g. the tracking range of the device) and of human abilities (e.g. the reach of the arm). When targets are outside the user's arm reach, techniques such as Go-Go [17] and HOMER [18] can be used to extend the length of the virtual arm [4]. Other works have explored different modalities for 3D interaction, including feet movements [19], tangible interfaces [20], and computer peripherals (e.g. 3D Connexion SpaceNavigator).

### 2.3 Gaze in Multimodal Interactions

Gaze-based interaction is known to suffer from a few challenges [21]: inaccuracy (due to the jittery nature of eye movements and technological limitations), double-role of visual observation and control, and the Midas Touch problem (the unintentional activation of functionality due to eye tracking being always-on [22]). To address these problems gaze is usually combined with other input modalities and devices.

Stellmach et al. investigated combinations of gaze with a wide variety of modalities [10], including a keyboard [23], tilt gestures [23, 24], a mouse wheel [24], touch gestures [24–26] and foot pedals [27]. A common interaction paradigm in gaze-based interaction is that of *gaze-supported interaction*—gaze suggests and the other modality confirms [25]. An example of a gaze-supported interaction technique is MAGIC pointing, which warps the mouse cursor to the area around the gaze pointing [28]. Fine positioning and selection confirmation are performed normally with the mouse.

These works have shown that multimodal gaze-based techniques are intuitive and versatile enough to work in a wide variety of contexts, ranging from small mobile devices to large public displays [29].

In this work, we have a similar goal to Stellmach and Dachselt, that of *seamless* selection and positioning [26]. Whereas in their work, they achieved this with different touch-based techniques for mobile devices, the context of 3D user interfaces requires extra degrees of freedom that are better suited for mid-air gestures.

### 2.4 Gaze and Mid-Air Gestures

Kosunen et al. reported preliminary results of a comparison between eye and mid-air hand pointing on large screen in a 2D task that indicates that pointing with the eyes is 29 % faster and 2.5 times more accurate than mid-air pointing [11]. The techniques investigated in their paper are analogous to our 2D Cursor and our Gaze technique, as they also used a pinch gesture for selection confirmation. In this paper, we extend their work to 3D manipulation, also comparing them to a 3D cursor. Also, their task involved 2D translation of objects, whereas ours involves 3D translation.

Pouke et al. investigated the combination of gaze and mid-air gestures, but in the form of a 6DOF sensor device attached to the hand [13]. Their system supported tilt, grab/switch, shake and throw gestures. They compared their technique with a touch-based one, and found that the touch-based interaction was faster and more accurate, mainly due to accuracy issues with their custom-built eye tracker. We aimed to minimise tracker accuracy problems, by using a commercial eye tracker with a gaze estimation error of 0.4 degrees of visual angle. Our study also differs from theirs in that the mid-air gestures investigated by them were based on a tangible device, rather than hands-only gestures.

Yoo et al.'s system tracked the user's head orientation (as an approximation for the gaze point) and the 3D position of the hands for interacting with large displays [30]. Bowman et al. investigated pointing in the direction of gaze, but also approximating it to the head orientation [2]. Such approximations only work when the user is looking straight ahead. Hence they are only suitable for large scale interactions, such as with large displays and fully immersive virtual environments. In a desktop setting, the head orientation is not a good approximation for the point of regard, as the user is constantly facing the same direction.

Cha and Maier proposed a combination of gaze and mid-air gestures for a multi-display use case [31]. These authors presented architectural implementation details of their system, but did not present any evaluation results or interaction design decisions.

## 2.5 Gaze and 3D User Interfaces

Stellmach and Dachselt proposed two ways in which 3D user interfaces can benefit from eye tracking: first, understanding how users visually perceive 3D scenes can assist the design of new 3DUIs and second, eye trackers can be used for direct control of 3DUIs [32].

Examples of the first group of applications include studying players' gaze patterns in 3D video games to improve level design and graphics [33], to improve gaze behaviour and body animations of virtual agents [34] and to enhance the rendering of depth-of-field blur effects [35].

In the second group are applications controlled directly by gaze. In the past few years, companies such as Tobii and SMI started marketing eye trackers for the wider consumer market aimed primarily at gaming, which stimulated developers to create the first commercial gaze-enabled games [36]. The research community has also demonstrated several examples of interaction techniques and game prototypes in this context (see Sundstedt for an overview [37]).

The popularity of head-mounted displays such as the Oculus Rift created renewed interest in exploring eye tracking within Virtual Reality. Tanriverdi and Jacob compared selection time and the users' ability to recall spatial information in a VR between two techniques: the gaze position and pointing with a magnetic tracker [8]. They found that the gaze technique was significantly faster, but led to more difficulties in recalling the locations of items they had interacted with. On the other hand, Cournia et al. compared gaze and hand pointing in VR and found hand pointing to perform better [9]. Following Cournia et al.'s recommendation, we implemented our 2D cursor using raycasting, as it seemed to outperform arm extension techniques (such as the one used by Tanriverdi and Jacob [8]). These works, however, investigate 3D interaction in an immersive VR environment, whereas we use a monoscopic display. Duchowski et al. also investigated eye tracking in virtual reality in applications ranging from monitoring users' gaze for aircraft inspection training [38] to providing a visual deictic reference in collaborative environments [39].

## 3 Experimental Setup

We recruited 12 right-handed participants (6 M/6F), aged between 20 and 43 years (median = 28). Three wore glasses and one wore contact lenses in the study. Figure 1 shows our experimental setup. Participants sat in front of an 18'' laptop running a custom application built in the *Unity* game engine. Gaze was tracked at 30fps using a Tobii EyeX tracker mounted under the display, with an average gaze estimation accuracy of 0.4 degrees of visual angle. Hands were tracked using an Asus Xtion PRO LIVE sensor, with resolution of 640 × 480 (30 Hz), mounted facing down on a 0.82 m × 1.0 m rig. Pose estimation and gesture recognition were performed using 3Gear Systems' *Nimble SDK*.

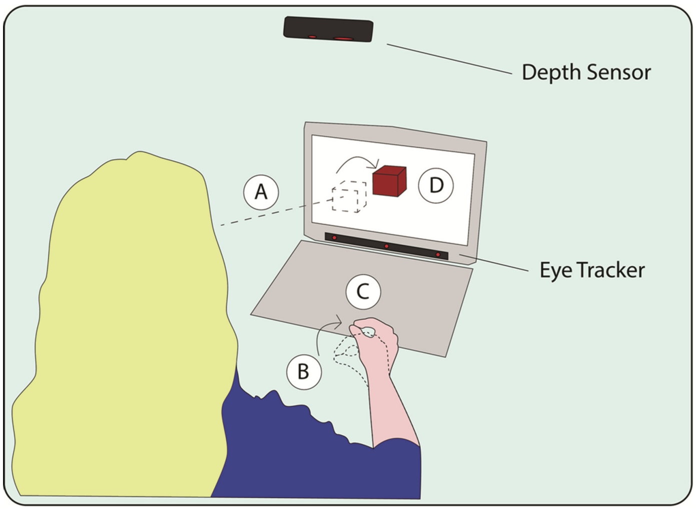

**Fig. 1.** Gaze selection for 3DUI: The user selects the object by looking at it (A), pinches (B), and moves her hand in free-space (C) to manipulate it (D).

We implemented three interaction techniques for selecting and translating objects in our 3D scene:

- **Gaze (Gaze-Supported Mid-Air Gestures):** the user looks at the object he wishes to select, pinch, move his hand to translate the object, and releases the pinch to disengage from the interaction.
- **2D Cursor (Raycasting):** the user moves his hand on the plane parallel to the screen (up/down and left/right), which moved a cursor on the camera plane of the scene (moving the hand towards and away from the screen had no effect on the cursor). Targets were selected by hovering over them, (similar to a mouse cursor) and pinching. Then, the user moved his hand to translate the object and released the pinch to disengage from the interaction. Note that in this interaction technique, whereas the selection step uses only the XY coordinates of the hand, the translation step uses all three (XYZ).
- **3D Cursor (Virtual Hand):** the user moves his hand around the space above the desk, which moved a sphere cursor in the virtual environment in three dimensions. Because we used an isomorphic mapping between the physical space and the 3D

scene, any movement of the hand was directly translated in an equivalent movement of the cursor. To select an object, the user intersects the sphere cursor with the desired object and pinches. The user then moves his hand to translate the object and releases the pinch to disengage from the interaction.

Upon arrival, participants completed a consent form and a demographics questionnaire. We calibrated the eye and hand trackers with the manufacturers' default procedures. Participants then performed three 3D interaction tasks, described in the following sections. After all tasks were completed, we conducted an open-ended interview about their experience in using the interaction techniques.

## 4 Task 1: Translating a Single-Object

In Task 1, we compared completion times for two hand-based and one gaze-based selection techniques in a translation task. Participants were presented with a 3D environment containing one blue and one red cube (see Fig. 2). The task was to pick up the blue cube with a pinch gesture using each technique to select it, match its position to that of the red cube by moving their right hand whilst pinching, and drop it at the position of the red cube by releasing the pinch. When the blue cube intersected with the red cube, the red cube would turn green, indicating that the object could be released.

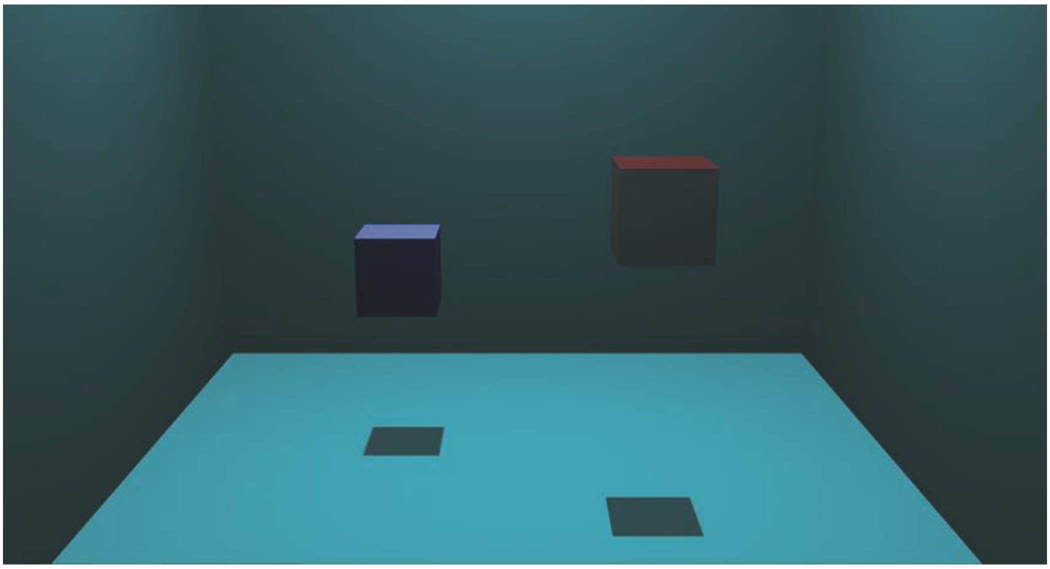

**Fig. 2.** Task 1: Users picked up the blue cube using each of the three techniques to select it and pinching to confirm the selection. They then moved this cube until it touched the red cube, which, in turn, would change its colour to green. The trial was concluded by releasing the pinch (Color figure online).

Participants performed the tasks in three blocks, each with 18 trials for each technique, in a counter-balanced order, for a total of 3 blocks × 3 techniques × 18 trials = 162 interactions. In each trial, the starting position of the cubes changed, but the distance between them remained constant. In the final block, after completing all trials for each technique, participants completed a questionnaire in which they rated each technique on a 7-point scale with respect to speed, accuracy, ease of learning and use, eye, hand and arm fatigue, intuitiveness, mental and physical effort, comfort, suitability for the task and personal preference. After completing all blocks they ranked the techniques in terms of speed, accuracy, comfort and personal preference. We discarded the first block from further analyses as a practice round.

## 4.1 Results

We compared the mean completion times between each technique across all trials, as well as the times of each step of the task, namely the time to acquire the blue cube (Acquisition), the time to pinch to confirm the selection (Confirmation) and the time to move it to the red cube (Translation). We tested the effects of the technique on the dependent variables using a one-way repeated-measures ANOVA (Greenhouse-Geisser corrected in case Mauchly's test revealed a violation of sphericity) and post hoc pairwise t-tests (Bonferroni corrected).

The mean trial completion time using Gaze (3.76 s) was 21.3 % shorter than using the 2D Cursor (4.78 s) and 37.0 % shorter than using the 3D Cursor (5.97 s) (see Fig. 3). The effect of technique on mean completion time was significant ($F_{2,22} = 24.5$, $p < .01$) with significant differences between all combinations of techniques ($p < .05$).

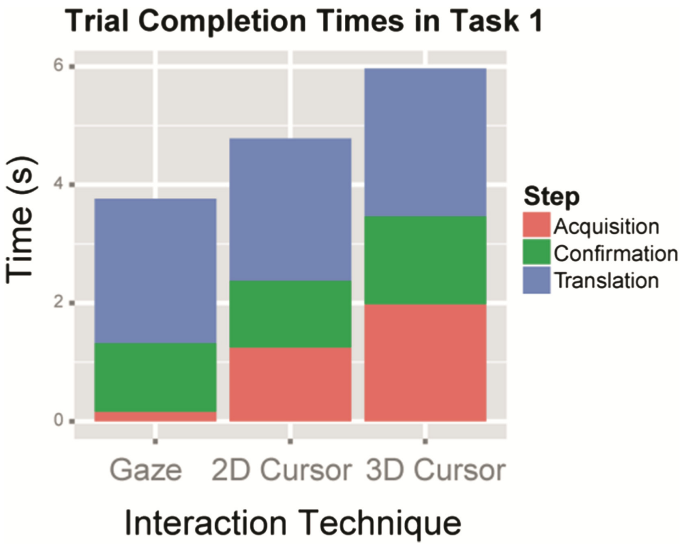

**Fig. 3.** Mean task 1 completion time split by step

The Acquisition Time using Gaze (161 ms) was 87.2 % shorter than using the 2D Cursor (1.25 s), and 91.9 % shorter than using the 3D Cursor (1.98 s), with a significant effect of the technique ($F_{2,22} = 194.5$, $p < .01$). Post hoc tests showed significant differences between all combinations of techniques at $p < .05$. We did not find a significant effect of the technique neither on the confirmation time ($F_{1.2,13.2} = 3.1$, $p = .07$) nor on the translation time ($F_{2,22} = .12$, $p = 0.88$).

In the questionnaires, Gaze received higher scores than the other two techniques along all dimensions, except for eye fatigue, for which it scored the lowest of all three (but the difference was not statistically significant). Eleven participants ranked gaze as their preferred technique overall, with only one user preferring the 2D cursor. Nine users indicated the 3D cursor as the worst technique and three indicated the 2D cursor. A similar pattern was found for Accuracy, Speed and Comfort rankings.

### 4.2 Discussion

The results from Task 1 are in line with Tanriverdi and Jacob [8]. Even though their setup was VR-based, it seems that Gaze also outperform other 3D selection techniques in monoscopic displays. Unlike Cournia et al., Gaze also outperformed Raycasting for selection, but as suggested by these authors, Raycasting performed better than Virtual Hand [9].

Both Tanriverdi and Jacob and Cournia et al. investigated 3D selection, but not in the context of further manipulation. We also included a translation task to analyse whether the selection technique influenced the completion time of subsequent manipulation tasks (for example, by requiring clutching or adjustment of the hand position after selection). Because we found no significant difference in the confirmation and translation tasks, we cannot affirm that these interaction techniques have any effects on the manipulation task time, even though we observed certain hand clutching in the Gaze and 2D Cursor conditions. As shown in Fig. 3, the only significant cause for the difference in the task completion time was in the object acquisition.

## 5 Task 2: Sorting Multiple Objects

In Task 1, we showed that the acquisition time using gaze is significantly shorter than using the other techniques. However, Gaze is known to suffer from inaccuracies, due to the jittery nature of eye movement, calibration issues and gaze estimation error. The goals of the second task were twofold: to investigate whether eye tracking inaccuracies would impair object selection in cluttered environments and to investigate how the faster selection times enabled by gaze can speed up tasks in which the user is required to rapidly manipulate different objects in sequence. We hypothesised that, because users do not have to necessarily move their hands to pick up new objects with Gaze, the fact that they could start the manipulation from wherever their hands were would speed up switching between objects.

Participants were presented with the same environment, now containing six black and six white chess pieces (see Fig. 4). The right and left walls were coloured in white and black, respectively. Participants were asked to pick up each chess piece and move it to the appropriate wall. When the object collided with the corresponding wall, it disappeared. If the object collided with the wrong wall, it did not disappear, but we incremented an error counter. Each participant performed three trials with 12 pieces, totalling 3 trials × 3 techniques × 12 pieces = 108 interactions. In the last trial, after each technique, they answered the same questionnaire as before. After all trials were completed, they completed the preference ranking questionnaire again. We discarded the first trial of each technique as a practice round.

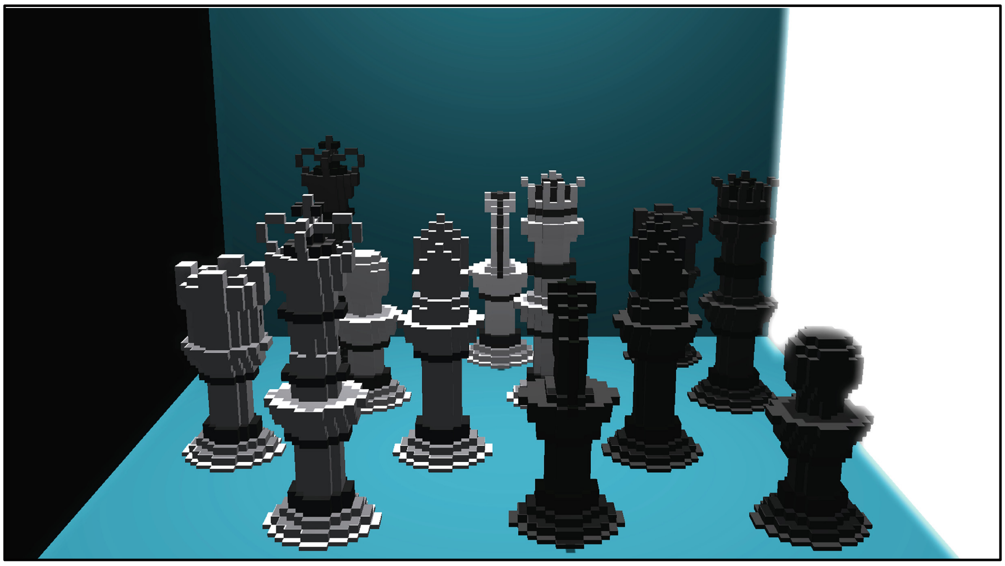

**Fig. 4.** Task 2: Users picked up each chess piece and moved it to the appropriate side of the virtual environment.

### 5.1 Results

The mean time to put away each piece with Gaze (2.27 s) was 38.4 % shorter than with the 2D cursor (3.67 s) and 49.4 % shorter than the 3D cursor (4.47 s) (see Fig. 5). We found a significant effect of the technique on Completion Time ($F_{2,22} = 37.7$, $p < .01$) and significant differences between all combinations at $p < .05$.

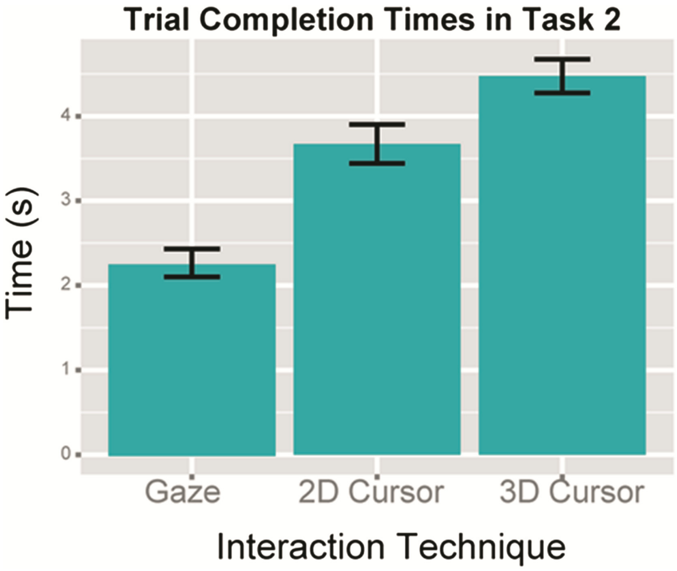

**Fig. 5.** Mean trial completion times in task 2. Gaze was significantly faster than the other two techniques.

The mean rate of incorrectly placed pieces with the 3D Cursor (1.92 %) was 71.3 % smaller than with the 2D cursor (6.70 %) and 82.7 % smaller than the Gaze (11.1 %). We found a significant effect on Error Rate ($F_{2,22} = 8.19$, $p < .01$). The post hoc tests showed significant differences only between Gaze and the 3D Cursor ($p < .05$). No considerable differences were found in the questionnaire responses between the first and second task.

### 5.2 Discussion

The task completion times in Task 2 were significantly shorter than in Task 1. The reason for this is that whereas the selection step required precision, the translation step did not—as soon as the object hit the correct wall, the task was complete. For the hand-based tasks this represented a similar gain in speed (30 % for the 2D Cursor and 34 % for the 3D Cursor), but a much higher gain in speed for the gaze technique (66 %). This shows that, even though it comes at a price of accuracy, Gaze is particularly well suited for tasks in which there is constant switching between different objects being manipulated. Examples of such tasks include organising furniture in architectural applications, playing speed-based 3D puzzle games and switching between different tools in a 3D modelling application.

## 6 Task 3: Selection Only vs. Selection and Manipulation

In our pilot studies, we noticed an interesting phenomenon when observing participants using gaze-assisted mid-air gestures. In the Gaze condition, once participants looked at the object, they could pinch from wherever their hands were and start manipulating the object from there. However, users still slightly reached out to the general position of the object, either to open up space for subsequent manipulations or due to a natural tendency to reach out as when handling real objects.

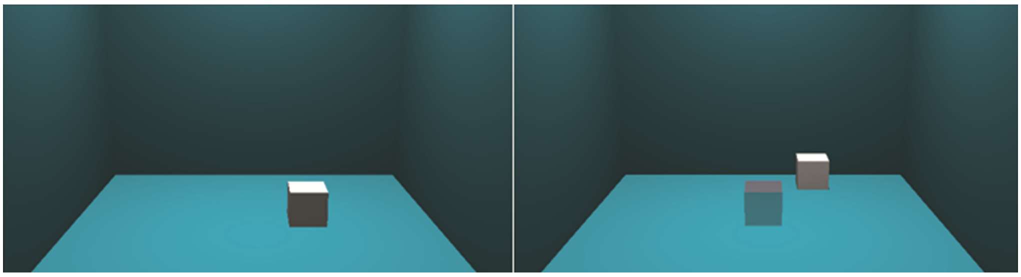

**Fig. 6.** Task 3: We compared the selection time between only selecting the object to the selection time with subsequent manipulation.

We hypothesized that this “clutching” before the translation would delay the selection confirmation when compared to selecting the object without any subsequent manipulation. To test this, we conducted a third task with the same 3D environment (see Fig. 6-A). In one condition, a white cube appeared at random positions, but always at the same Y coordinate and at one of two Z coordinates (one in the foreground and one in the background). To reset the gaze point, between each trial the cube would show up at the centre of the environment. Participants were asked to look at the cube and make a pinch gesture, after which the cube would disappear. To avoid participants predicting the timing of the pinch gestures, we added a random delay uniformly distributed between 500 ms and 1.0 s before each trial. The second condition was similar to the first, but after pinching, the user was asked to drag the white cube to a red cube at centre of the environment (see Fig. 6-B). Participants performed three blocks, each containing 20 trials of each task (not counting the gaze-resetting steps), for a total of 3 blocks × 2 tasks × 20 trials = 120 interactions.

## 6.1 Results

We compared the time to perform the pinch gesture after having acquired the object with their gaze. The time in the Selection Only condition (557 ms) was 26.5 % shorter than in the Selection + Translation (758 ms) (see Fig. 7). A Welch's t-test revealed that this difference was significant ($t_{11} = -2.69$, $p < .05$).

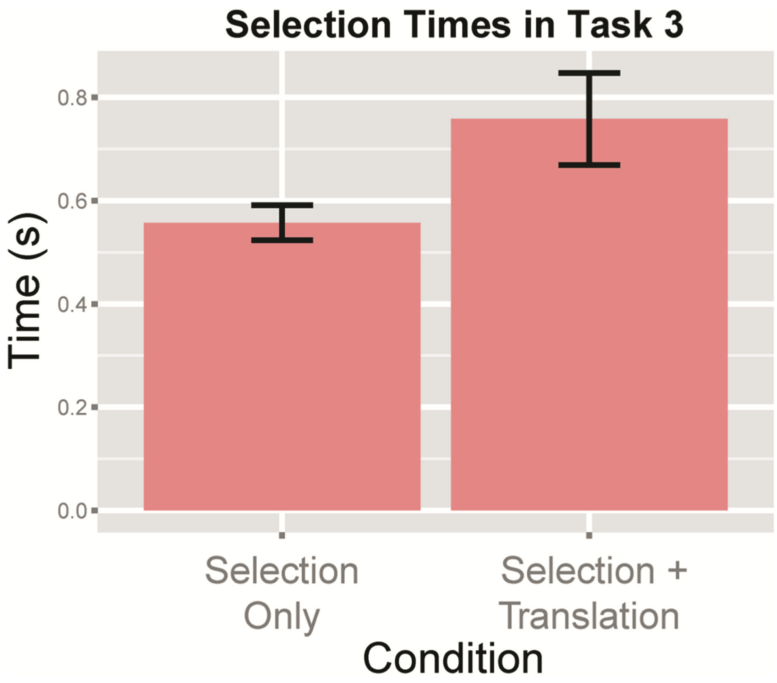

**Fig. 7.** Selection times in task 3. Users took longer to select the object when they were going to manipulate it afterwards.

## 6.2 Discussion

Our results show that the time taken to select an object with Gaze is significantly longer when the user plans on manipulating it afterwards. We offer three possible explanations for this phenomenon. First, when the user must translate the object after picking it, there is an additional planning step in the prehension process, adding some extra time for cognitive processing. Second, it is our natural behaviour to reach out in the general direction of where objects are. Third, with Gaze, even though the object can be selected from wherever the hand is, this initial position must allow enough room for the subsequent manipulation. Therefore, if the user's hand is not in an appropriate position, she must clutch it before picking the object up. From our observations, we believe the third explanation to be the most likely one.

## 7 Discussion

The results of Task 1 show that the acquisition time varied significantly between techniques, with Gaze being the fastest, followed by the 2D Cursor. We did not find a significant modality switch latency, as once the object was acquired, participants took approximately the same time to pick it up with a pinch gesture and move it to the target. As the results from Task 2 show, the advantage of Gaze is even stronger in tasks in which multiple objects are manipulated in sequence. This gain in speed comes at the cost of accuracy, particularly in densely populated environments. Participants' opinions on the techniques also confirmed that Gaze was the most popular technique.

In task 3, we discovered a significant difference in the time to pinch after the object was gazed at between selection only tasks and selection followed by translation. Although this difference is negligible for practical purposes, it reveals an interesting aspect of human behaviour when interacting using gaze. Even though the system was calibrated so that no clutching was necessary, participants still reached out in the general direction of where the object was positioned before pinching, similarly to how they would do with physical objects. Gaze selection elegantly supports this natural behaviour. This result suggests that gaze selection should be analysed in the context of the subsequent tasks, and not as an independent phenomenon.

We conducted our experiment in a desktop environment. The presented techniques could, however be extended to standing interaction with large displays and immersive environments. Moreover, in stereoscopic displays, as the hands do not need to intersect the objects, Gaze-selection can be used without breaking the 3D illusion. Another limitation was that we only looked at translation tasks, but the same could be investigated for rotation and scaling.

Gaze-assisted mid-air manipulation allows users to select objects far away and manipulate them comfortably as if they were within reach. This allows users to rest their wrists on the desk, minimising the *Gorilla Arm* problem. This technique is are also particularly useful for monoscopic displays, where the inherent discontinuity between the virtual and the physical spaces do not allow for direct manipulation and often require an extra step for positioning the cursor on the target. In fact, participants reported not having to think about this step at all and that all they had to do was to think about the object and pinch, allowing for an arguably more immersive experience and an interaction with more fidelity.

## 8 Conclusion

In this work we evaluated gaze as a modality for object selection in combination with mid-air hand gestures for manipulation in 3DUI. Whereas previous work has found conflicting results on the performance of gaze for 3D interaction, we found that gaze outperforms other mid-air selection techniques and supports users' natural behaviours when reaching out for objects. Our findings suggest that gaze is a promising modality for 3D interaction and that it deserves further exploration in a wider variety of contexts. In particular, in future work we would like to explore how gaze can modulate the mapping between the physical and virtual environments, making it easier to reach distant objects, for example. Another avenue for investigation is how gaze can be incorporated into existing 3D applications.

## References

1. Bowman, D.A., McMahan, R.P., Ragan, E.D.: Questioning naturalism in 3D user interfaces. Commun. ACM **55**, 78–88 (2012)
2. Bowman, D.A., Kruijff, E., LaViola Jr, J.J., Poupyrev, I.: 3D User Interfaces: Theory and Practice. Addison-Wesley, Boston (2004)
3. Hinckley, K., Pausch, R., Goble, J.C., Kassell, N.F.: A survey of design issues in spatial input. In: Proceedings of the 7th Annual ACM Symposium on User Interface Software and Technology, pp. 213–222. ACM (1994)
4. Argelaguet, F., Andujar, C.: A survey of 3D object selection techniques for virtual environments. Comput. Graph. **37**, 121–136 (2013)
5. Sibert, L.E., Jacob, R.J.: Evaluation of eye gaze interaction. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 281–288. ACM (2000)
6. Ware, C., Mikaelian, H.H.: An evaluation of an eye tracker as a device for computer input. In: Proceedings of the SIGCHI/GI Conference on Human Factors in Computing Systems and Graphics Interface, pp. 183–188. ACM, Toronto (1987)
7. Koons, D.B., Sparrell, C.J., Thorisson, K.R.: Integrating simultaneous input from speech, gaze, and hand gestures. In: Maybury, M.T. (ed.) Intelligent Multimedia Interfaces, pp. 257–276. American Association for Artificial Intelligence, Menlo Park (1993)
8. Tanriverdi, V., Jacob, R.J.: Interacting with eye movements in virtual environments. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 265–272. ACM (2000)
9. Cournia, N., Smith, J.D., Duchowski, A.T.: Gaze-vs. hand-based pointing in virtual environments. In: CHI 2003 Extended Abstracts on Human Factors in Computing Systems, pp. 772–773. ACM (2003)
10. Stellmach, S.: Gaze-supported Multimodal Interaction (2013). http://www.dr.hut-verlag.de/978-3-8439-1235-8.html
11. Kosunen, I., Jylha, A., Ahmed, I., An, C., Chech, L., Gamberini, L., Cavazza, M., Jacucci, G.: Comparing eye and gesture pointing to drag items on large screens. In: ITS, pp. 425–428. ACM (2013)
12. Hales, J., Rozado, D., Mardanbegi, D.: Interacting with objects in the environment by gaze and hand gestures. In: ECEM (2011)
13. Pouke, M., Karhu, A., Hickey, S., Arhippainen, L.: Gaze tracking and non-touch gesture based interaction method for mobile 3D virtual spaces. In: OzCHI, pp. 505–512. ACM (2012)
14. MacKenzie, C.L., Iberall, T.: The Grasping Hand. Elsevier, Amsterdam (1994)
15. Arbib, M.A.: Perceptual structures and distributed motor control. Comprehensive Physiology (1981)
16. Paillard, J.: Le corps situé et le corps identifié. Rev. Méd. Suisse Romande. 100 (1980)
17. Poupyrev, I., Billinghurst, M., Weghorst, S., Ichikawa, T.: The go-go interaction technique: non-linear mapping for direct manipulation in VR. In: Proceedings of the 9th Annual ACM Symposium on User Interface Software and Technology, pp. 79–80. ACM, Seattle (1996)

18. Bowman, D.A., Hodges, L.F.: An evaluation of techniques for grabbing and manipulating remote objects in immersive virtual environments. In: Proceedings of the 1997 Symposium on Interactive 3D Graphics, pp. 35–38. ACM (1997)
19. Simeone, A., Velloso, E., Alexander, J., Gellersen, H.: Feet movement in desktop 3D interaction. In: Proceedings of the 2014 IEEE Symposium on 3D User Interfaces. IEEE (2014)
20. Kitamura, Y., Itoh, Y., Kishino, F.: Real-time 3D interaction with ActiveCube. In: CHI 2001 Extended Abstracts on Human Factors in Computing Systems, pp. 355–356. ACM, Seattle (2001)
21. Stellmach, S., Dachselt, R.: Still looking: investigating seamless gaze-supported selection, positioning, and manipulation of distant targets. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 285–294. ACM (2013)
22. Jacob, R.J.: What you look at is what you get: eye movement-based interaction techniques. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 11–18. ACM (1990)
23. Stellmach, S., Stober, S., Nürnberger, A., Dachselt, R.: Designing gaze-supported multimodal interactions for the exploration of large image collections. In: Proceedings of the 1st Conference on Novel Gaze-Controlled Applications, p. 1. ACM (2011)
24. Stellmach, S., Dachselt, R.: Investigating gaze-supported multimodal pan and zoom. In: Proceedings of the Symposium on Eye Tracking Research and Applications, pp. 357–360. ACM (2012)
25. Stellmach, S., Dachselt, R.: Look & touch: gaze-supported target acquisition. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 2981–2990. ACM (2012)
26. Stellmach, S., Dachselt, R.: Still looking: Investigating seamless gaze-supported selection, positioning, and manipulation of distant targets. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 285–294. ACM (2013)
27. Göbel, F., Klamka, K., Siegel, A., Vogt, S., Stellmach, S., Dachselt, R.: Gaze-supported foot interaction in zoomable information spaces. In: CHI 2013 Extended Abstracts on Human Factors in Computing Systems, pp. 3059–3062. ACM (2013)
28. Zhai, S., Morimoto, C., Ihde, S.: Manual and gaze input cascaded (MAGIC) pointing. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 246–253. ACM (1999)
29. Turner, J., Alexander, J., Bulling, A., Schmidt, D., Gellersen, H.: Eye pull, eye push: moving objects between large screens and personal devices with gaze and touch. In: Kotzé, P., Marsden, G., Lindgaard, G., Wesson, J., Winckler, M. (eds.) INTERACT 2013, Part II. LNCS, vol. 8118, pp. 170–186. Springer, Heidelberg (2013)
30. Yoo, B., Han, J.-J., Choi, C., Yi, K., Suh, S., Park, D., Kim, C.: 3D user interface combining gaze and hand gestures for large-scale display. In: CHI 2010 Extended Abstracts on Human Factors in Computing Systems, pp. 3709–3714. ACM (2010)
31. Cha, T., Maier, S.: Eye gaze assisted human-computer interaction in a hand gesture controlled multi-display environment. In: Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction, p. 13. ACM (2012)
32. Stellmach, S., Dachselt, R.: Looking at 3D user interfaces. In: CHI 2012 Workshop on The 3rd Dimension of CHI (3DCHI): Touching and Designing 3D User Interfaces, pp. 95–98 (2012)
33. El-Nasr, M.S., Yan, S.: Visual attention in 3D video games. In: Proceedings of the 2006 ACM SIGCHI International Conference on Advances in Computer Entertainment Technology, p. 22. ACM (2006)

34. Vinayagamoorthy, V., Garau, M., Steed, A., Slater, M.: An eye gaze model for dyadic interaction in an immersive virtual environment: Practice and experience. Comput. Graph. Forum **23**, 1–11 (2004). Wiley Online Library
35. Hillaire, S., Lécuyer, A., Cozot, R., Casiez, G.: Using an eye-tracking system to improve camera motions and depth-of-field blur effects in virtual environments. In: Virtual Reality Conference, VR 2008. IEEE, pp. 47–50. IEEE (2008)
36. Turner, J., Velloso, E., Gellersen, H., Sundstedt, V.: EyePlay: applications for gaze in games. In: Proceedings of the First ACM SIGCHI Annual Symposium on Computer-Human Interaction in Play, pp. 465–468. ACM (2014)
37. Sundstedt, V.: Gazing at games: using eye tracking to control virtual characters. In: ACM SIGGRAPH 2010 Courses, p. 5. ACM (2010)
38. Duchowski, A.T., Shivashankaraiah, V., Rawls, T., Gramopadhye, A.K., Melloy, B.J., Kanki, B.: Binocular eye tracking in virtual reality for inspection training. In: Proceedings of the 2000 Symposium on Eye Tracking Research & Applications, pp. 89–96. ACM (2000)
39. Duchowski, A.T., Cournia, N., Cumming, B., McCallum, D., Gramopadhye, A., Greenstein, J., Sadasivan, S., Tyrrell, R.A.: Visual deictic reference in a collaborative virtual environment. In: Proceedings of the 2004 Symposium on Eye Tracking Research and Applications, pp. 35–40. ACM (2004)