Gaze-Hand Alignment: Combining Eye Gaze and Mid-Air Pointing for Interacting with Menus in Augmented Reality

MATHIAS N. LYSTBÆK, Aarhus University, Denmark
PETER ROSENBERG, Aarhus University, Denmark
KEN PFEUFFER, Aarhus University, Denmark and Bundeswehr University Munich, Germany
JENS EMIL GRØNBÆK, Aarhus University, Denmark
HANS GELLERSEN, Lancaster University, UK and Aarhus University, Denmark

Gaze and freehand gestures suit Augmented Reality as users can interact with objects at a distance without need for a separate input device. We propose Gaze-Hand Alignment as a novel multimodal selection principle, defined by concurrent use of both gaze and hand for pointing and alignment of their input on an object as selection trigger. Gaze naturally precedes manual action and is leveraged for pre-selection, and manual crossing of a pre-selected target completes the selection. We demonstrate the principle in two novel techniques, Gaze&Finger for input by direct alignment of hand and finger raised into the line of sight, and Gaze&Hand for input by indirect alignment of a cursor with relative hand movement. In a menu selection experiment, we evaluate the techniques in comparison with Gaze&Pinch and a hands-only baseline. The study showed the gaze-assisted techniques to outperform hands-only input, and gives insight into trade-offs in combining gaze with direct or indirect, and spatial or semantic freehand gestures.

CCS Concepts: • **Human-centered computing → Mixed / augmented reality**; **Pointing**; **Interaction design theory, concepts and paradigms**.

Additional Key Words and Phrases: eye-tracking, gaze interaction, pointing, mid-air gestures, augmented reality, menu selection

ACM Reference Format:
Mathias N. Lystbæk, Peter Rosenberg, Ken Pfeuffer, Jens Emil Grønbæk, and Hans Gellersen. 2022. Gaze-Hand Alignment: Combining Eye Gaze and Mid-Air Pointing for Interacting with Menus in Augmented Reality. *Proc. ACM Hum.-Comput. Interact.* 6, ETRA, Article 145 (May 2022), 18 pages. https://doi.org/10.1145/3530886

1 INTRODUCTION

Gaze and freehand gestures are attractive input modalities as they enable users to interact directly with their environment through movements of their eyes and hands. Users are not reliant on any input device, can avoid physical contact where this is inconvenient or undesirable, and are empowered to interact beyond reach. With these properties, gaze and gestures are well suited to complement Augmented Reality (AR) and interactions that are situated in the world [9]. Current trends in AR technology reflect this with the integration of both hand- and eye-tracking in head-worn display (HMD) devices. However, the design of input techniques that rely on gaze and freehand gestures is challenging, as it requires robust segmentation of input from the continuous movement of the user's eyes and hands.

In this work, we consider the combination of gaze and freehand input for pointing and menu selection at a distance. Either modality can be used by itself for pointing at objects, but a separate confirmatory action is required to complete a selection and avoid "Midas Touch" input [11]. In menus specifically, users need to be able to inspect and traverse objects with their pointer without

Authors' addresses: Mathias N. Lystbæk, mathiasl@cs.au.dk, Aarhus University, Denmark; Peter Rosenberg, peterrosenberg71@gmail.com, Aarhus University, Denmark; Ken Pfeuffer, ken@cs.au.dk, Aarhus University, Denmark and Bundeswehr University Munich, Germany; Jens Emil Grønbæk, jensemil@cs.au.dk, Aarhus University, Denmark; Hans Gellersen, h.gellersen@lancaster.ac.uk, Lancaster University, UK and Aarhus University, Denmark.

2022. 2573-0142/2022/5-ART145 $15.00
https://doi.org/10.1145/3530886

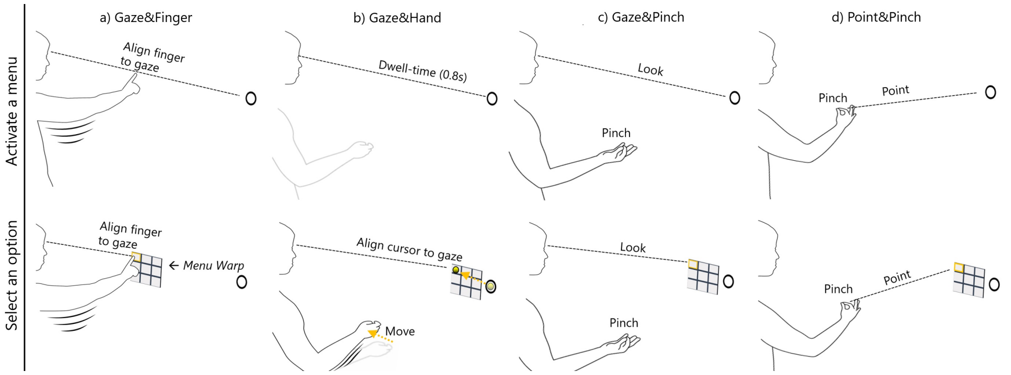


Fig. 1. Gaze&Finger and Gaze&Hand are novel techniques for menu selection based on Gaze-Hand Alignment. With Gaze&Finger, both menu activation and item selection is by alignment of a finger in the line of sight to the target (a). With Gaze&Hand, the activation step is by dwell time to instantiate a hand-controlled cursor, followed by alignment of gaze and cursor for item selection (b). In an evaluation, we compare both techniques against baselines of Gaze&Pinch (c), and Point&Pinch (d).

triggering input, prior to finalising their selection. For gaze-only input, it is common to use dwell time to confirm a selection, requiring the user to maintain a gaze fixation on the target for a longer duration to signal their intent [23]. In gestural interfaces, manual pointing is completed by a distinct gesture such as a tap on a surface, or a pinch in mid-air [5]. Where both modalities are available, gaze lends itself better for the initial pointing step, as our eyes naturally focus on objects that we aim to manipulate, whereas our hands are more effective for deliberate input to complete a selection [30]. In past work, this has been demonstrated by combining gaze pointing for instance with mouse click [55], key press [11] or touch [46], and in recent work by combining gaze with pinch as the delimiting mid-air gesture [6, 33]. Gaze&Pinch is also supported by emerging AR headsets (such as the HoloLens 2) as state of the art gaze-based selection technique.

We propose *Gaze-Hand Alignment* as a principle for gaze and freehand input. In contrast to Gaze&Pinch and comparable techniques, we are using *both* modalities for pointing, and alignment of their input as selection trigger. The key idea is to leverage that the eyes naturally look ahead to a pointing target, followed by the hands [55]. This enables us to use gaze for pre-selection, and manual crossing of a pre-selected target to trigger a selection as soon as the hand catches up with the eyes and aligns with gaze. We introduce *Gaze&Finger* and *Gaze&Hand* as novel techniques for AR context menus based on the concept. Gaze&Finger (shown in Fig. 1a) combines gaze with perspective-based manual pointing where a ray is cast from the eye position over the user's index finger. A user looks at targets of interest and completes selection by lifting their finger into the line of sight. We use the same principle to invoke a menu and to select from it. The menu is warped to the user's hand to avoid parallax issues and scaled to target distance to support an illusion of direct touch. Gaze&Hand (Fig. 1b) combines gaze with indirect manual input to reduce effort and arm fatigue. The technique uses dwell time for menu activation and instantiation of a cursor, and alignment of cursor and gaze for selection from the menu. Both techniques have in common with Gaze&Pinch (Fig. 1c) that initial selection is by gaze and confirmation by manual gesture, but alignment is spatial and implicitly guided by gaze, whereas a pinch is semantic and requires separate attention to gaze.

A number of studies have found the combination of gaze and hand-based input to outperform selection by the hands alone, but these were based on manual input mediated by physical controls [15, 56]. We contribute a study in which we compare Gaze&Finger, Gaze&Hand, and Gaze&Pinch with Point&Pinch (Fig. 1d), a hands-only baseline where a pointing ray cast from the hand is combined with pinch as selection trigger. The four techniques were evaluated on a task designed to represent contextual AR menus. Figure 1 illustrates the techniques for the two steps of activating a menu on a target object and selecting an item from the menu. The task was chosen as Gaze&Hand requires an initial step to instantiate a cursor whereas the other techniques are symmetric in using the same selection principle for both steps. We found that all three gaze-based techniques outperformed the manual condition. Gaze&Pinch was fastest for menu activation and Gaze&Finger most efficient for item selection. Gaze&Hand was perceived to be least physically demanding but had the highest error rate. While providing evidence of the efficacy of gaze-assistance for mid-air interaction, the study also contributes insight into trade-offs in the design of selection techniques, specifically spatial versus semantic and direct versus indirect use of gestures.

In sum, the novel contributions of this work are:

- the concept of Gaze-Hand Alignment for pointing and selection at a distance in freehand interfaces
- the Gaze&Finger and Gaze&Hand techniques, demonstrating the concept for menu selection in AR
- evidence of the efficacy of gaze-assisted techniques in conjunction with mid-air input
- insight into design trade-offs in the combination of gaze with freehand gestures for pointing and selection

2 RELATED WORK

Gaze and mid-air gestures have long been studied for interaction at a distance in real, virtual, and mixed realities. We are building on insights from work that has considered the modalities separately and in combination, and from work on eye-hand coordination and alignment of input with different modalities.

2.1 Pointing in Gaze and Mid-Air Interfaces

Gaze corresponds to the user's focus of attention which makes it natural and fast for pointing at objects of interest [4]. A wide range of work has harnessed gaze for implicit interaction, for example, to render interfaces attentive to the user [49] and adapt information displays [29]. Gaze has also been adopted for explicit input, based on interaction techniques that extend gaze pointing with a selection method equivalent to a mouse "click". Gaze-only selection methods require the user to deviate from natural gaze behaviour to signal their intent and avoid Midas Touch, for example by using dwell time on a target [11, 52], or by saccading from the target to a confirmation button [22, 27]. Although successful for accessibility, such techniques are commonly experienced as uncomfortable and error-prone, in comparison with manual input [34, 56]. The idea of Gaze-Hand Alignment is to take advantage of both modalities in tandem, and to make more natural use of gaze as a modality that implicitly guides manual pointing.

Freehand gestures, also referred to as mid-air gestures, afford more deliberate control than gaze. The hands are also more expressive, for pointing as well as semantic input [6, 33]. For remote control, gesture interfaces tend to either employ pointing with cursor feedback [50], or a library of predefined gestures for direct selection of commands [48]. In 3D user interfaces, freehand input can be based on a *Virtual Hand* metaphor for direct interaction, or a *Virtual Pointer* for interaction at a distance with a ray cast from the user's hand [38]. With the hands in mid-air, pointing is intuitive but there is no obvious "click method" for selection [50]. One approach is to virtually extend the reach of the hand for direct selection with non-linear input mappings [8, 37]. More commonly, pointing gestures are combined with a delimiting gesture, for example, a "pinch" as employed with current AR devices (Microsoft HoloLens *Point-and-Commit*[1], Oculus Quest *Point and Pinch* [2]). In this work, we compare multimodal eye- and hand-based techniques against Point&Pinch as a hands-only baseline.

Freehand pointing is commonly based on raycasting from the hand. As an alternative, perspective-based pointing is based on projecting a ray from the position of the user's eye over a point in a space that the user controls, such as the tip of a finger [2, 18]. The technique has been widely explored in virtual reality (VR) and is similar to direct input, as input and feedback coincide in the user's visual space [13]. Perspective-based pointing is therefore also associated with occlusion of targets as the selection mechanism [1, 17]. However, even though the pointing ray is projected from the eye, it is solely based on manual input and prone to Midas Touch. Prior work has proposed to address this with direct manipulation in the image plane, for example by pinching the image of an object as it appears between finger and thumb in order to select it [36]. In Gaze&Finger, we instead combine perspective-based pointing with actual tracking of gaze, casting two independent rays that coincide when the user aligns their finger in the line of sight to an object.

2.2 Eye-Hand Coordination and Combination of Gaze and Gestures

Gaze is involved in the planning of hand movements by fixating relevant landmarks of a target before its manual acquisition [12, 16, 43, 53]. That gaze precedes action has been employed in a range of multimodal techniques. In the MAGIC technique, a mouse pointer is warped toward the gaze position to move it closer to an anticipated target [55]. In Gaze-touch and Gaze-shifting, touch input is dynamically shifted to where the user looks, to effectively extend the reach of manual input [30–32]. In Gaze-Hand Alignment, we exploit the natural coordination of eye and hand in a novel way. Even though gaze is used for pre-selection in our techniques, users need not consider this as an explicit step to complete prior to alignment of manual input. Rather, users are free to explore potential targets with their gaze, and will implicitly gaze at the intended target once they decide on a selection, in order to guide the manual input to visually align with the target. In the process, the manual pointer can traverse other objects without Midas Touch effect and a selection can be triggered instantly when the manual pointer crosses onto a target that is pre-selected by gaze.

Existing techniques that combine input from eyes and hand are characterised by a division of labour and separate metaphors for the actions performed with either modality, using the eyes to *point*, and the hand to *click*, *tap* or *pinch* [28, 30, 33, 40, 56]. Some techniques have in common with ours that they use both gaze and hand for pointing input, however with the latter for a separate concurrent task such as rotate and scale [47] or a subsequent task of refining the gaze position which still requires a further action to trigger selection [6, 15]. Closer to our work is the idea of using gaze to delimit raycast input such that it only triggers input when it is within a predefined gaze range [39]. In Gaze-Hand Alignment, gaze and hand act as mutual delimiters so that neither triggers input unless they become spatially aligned in pointing at the same object. Other work has explored the idea of mutually delimited input by eye and hand based on temporal alignment, specifically of blink and touch events [51].

Most existing techniques rely on input devices for manual input, but gaze has also been combined with mid-air gestures and specifically pinch [6, 33]. A recent study compared mid-air pinch, dwell time and button click for confirming gaze input, and found the combination of gaze and pinch to be

[1] https://docs.microsoft.com/en-us/windows/mixed-reality/design/point-and-commit - accessed 2nd of September 2021.
[2] https://support.oculus.com/articles/headsets-and-accessories/controllers-and-hand-tracking/hand-tracking-gestures/?locale=en_US - accessed 2nd of September 2021.

faster than gaze-only input [28]. Gaze&Pinch thus constitutes a suitable baseline for comparison with our alignment-based techniques.

2.3 Alignment of Input from Separate Pointing Modalities

A number of other techniques are based on input from two separate pointing modalities and concepts of alignment. Toolglass and magic lenses are based on the alignment of see-through tools and objects in the user interface, where tool alignment can be performed by one hand while the other hand is used for pointing and selection through the tool [3]. EyeSeeThrough is an adaptation of the Toolglass concept for 3D user interfaces and involves alignment of a menu in the line of sight to the target object, either by hand or head movement [24]. In these techniques, pointing modalities are mapped to separate tasks in the interface, whereas our techniques employ gaze and hand in natural coordination on the shared task of selecting an object.

Gaze naturally involves head movement in conjunction with eye movement, in particular for interaction over wider fields of view such as in AR and VR [41], but recent work has also considered gaze and head pose as separate inputs that can be combined for pointing and selection [15, 42, 44]. Two techniques in particular have inspired our work. In Eye&Head Convergence, a gaze selection becomes confirmed by moving the head pointer into a focal zone around the gaze point, based on the observation that gaze and head rays do not normally align during visual exploration of potential targets. We found the idea transferable to eye and hand interfaces, as we are not normally looking at our hands during visual exploration, but naturally do so when we manipulate a target [43]. In Look&Cross, eye and head are aligned for selection in a radial menu [45]. As in Look&Cross we are adopting gaze for pre-selection in menus, but we use manual target crossing for confirmation instead of crossing by head pointer. The key difference in using the hand instead of the head is that eye and hand naturally align for positioning a finger or a cursor on a target, whereas head-based alignment requires the user to override the natural behaviour of orienting the head no further toward a target than necessary for maintaining a comfortable eye-in-head position [41].

3 DESIGN OF GAZE-HAND ALIGNMENT INTERACTION TECHNIQUES

We propose Gaze-Hand Alignment as a general principle for taking advantage of the natural coordination of eye and hand in the selection of input. The idea lends itself to the design of techniques that follow the same principle while addressing different design goals. In this work, we develop the idea specifically for menu-based input in AR, for which we have designed two techniques that differ in direct versus indirect use of hand input. Gaze&Finger is designed for direct touch in the image plane, appealing to familiarity with mobile touch input. Gaze&Hand is designed for indirect input with less physical effort, building on relative hand input as familiar from use of a computer mouse.

3.1 Gaze&Finger

The Gaze&Finger technique uses eye-tracking to determine a gaze vector into the Field Of View (FOV). When the gaze vector intersects with interactive objects, the object is pre-selected and visual feedback is provided to indicate its “hover” state. Simultaneously, the hand is tracked to determine a perspective-based ray cast from the position of the eyes over the index finger into the scene. When the ray intersects with a pre-selected object, selection is triggered. Other than in response to the selection, no feedback is provided on the manual input, as users rely on proprioception to guide their hand into their line of sight, and as their view over the raised index finger itself guides accurate positioning on a target.

Figure 2 a-c illustrate the technique for selection from a context menu in AR. To open a menu on an interactive object, in this case a smart lamp, the user needs to raise their hand into the visual

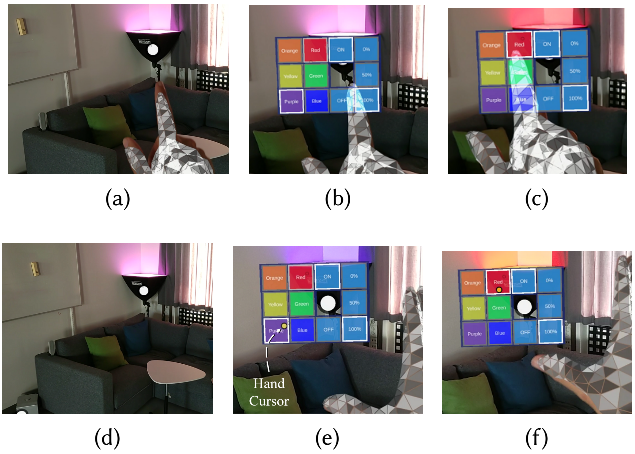


Fig. 2. Illustration of Gaze&Finger (a-c) and Gaze&Hand (d-f) for selection from AR context menus in a smart home context. With Gaze&Finger, the user lifts their finger into the line of sight of a smart lamp (a); this triggers a menu displayed at the user's fingertip (b); a selection is made by shifting gaze and finger onto one of the items(c). With Gaze&Hand, the user gazes at the lamp using dwell time to invoke the menu (d); this opens the menu and instantiates a cursor controlled by relative hand movement (e); a selection is completed by aligning gaze and cursor on one of the items (f).

field, to align their finger over the lamp while looking at it. When the menu has been opened, any item can be selected by a small shift of the finger in the image plane. This is implicitly guided and preceded by gaze so that users can initiate the manual selection without first having to complete an explicit gaze task. As soon as the finger aligns with gaze on the target, the selection is triggered, reinforcing a sense of direct touch where the response is instantaneous.

As Gaze&Finger is based on input in the image plane, parallax presents a challenge. We project the manual ray from a central position between the eyes over the finger, whereas users normally align their fingertip along the line of sight from one of their eyes, depending on eye dominance and whether the target appears towards the left or the right in the visual field [14]. This causes an apparent displacement of the target that increases with distance from the finger and reduces pointing accuracy. For the activation step (i.e., triggering the context menu), we address this by increasing the target selection area around the visual representation of objects, but accuracy could also be improved by compensating for systematic displacement [25]. However, for the selection step, we address parallax by rendering the context menu at the depth of the user's finger while also scaling the menu so that its apparent size remains consistent with target distance. This strategy is inspired by *automatic scaling* for direct manipulation in VR [26]. It requires accommodation of gaze depth after menu activation but eliminates any parallax issue for accurate selection within the menu.

3.2 Gaze&Hand

The Gaze&Hand technique employs gaze in the same manner for object pre-selection as Gaze&Finger but tracks the hand for indirect control of a cursor. Instead of employing a cursor that is continuously present, we use gaze dwell time for dynamic instantiation of a cursor at the current gaze position. In this way, a cursor is generated in the area of interest and need not be moved across a larger space or different levels of depths. Once the cursor has been instantiated, it is controlled by relative movement of the hand for translation in 2D while visual depth remains fixed.

Figure 2 d-f illustrate the technique for the smart home context menu. The user can open a menu without any hand movement but they need to maintain their gaze on a target object for a dwell time required to signal their intent. Once the dwell time has been reached, the menu is opened adjacent to the object, and a cursor is rendered at the gaze position, i.e. on the object. Any item in the menu is then selected by looking at it and moving the cursor onto it. As cursor control is indirect, there is no need to lift arm and hand into the visual field, and alignment with gaze can be achieved with small hand movement. However, we also constrain cursor movement to a circular bounding area around the menu, to avoid that any unintended larger hand movement results in losing the cursor from view. Gaze&Hand minimises physical effort, avoid arm fatigue and prohibits occlusion of the target area by the hand. There are no parallax issues, as object, menu and cursor feedback are all presented at the same depth of view, and there is also no occlusion of the target area by the hand.

Gaze&Hand reduces physical effort but is more complex in design, with a number of parameters to consider, in particular dwell time for activation, and control-display (CD) gain for cursor movement. We implemented and evaluated the technique with a dwell time of 800 ms, the default setting for dwell time in the device we used for our study (HoloLens 2). A shorter dwell time would be possible to speed up access to the menu but that would increase the likelihood of unintended activation. The CD gain determines the amount of hand movement needed for translation of the cursor in the 3D user interface. We optimised the technique for a menu distance of 1.8 meters, where we found a CD gain of 8.5 effective for minimising required hand movement while ensuring robust control.

3.3 Implementation and Application

We implemented our techniques for the Microsoft HoloLens 2, using the Mixed Reality Toolkit (MRTK) in Unity 2020. For demonstration, we developed a smart home application scenario, where the techniques are used for selecting commands on Philips Hue smart lamps in the environment, as illustrated in Figure 2. The menu provides buttons for turning the lamp on or off, picking a colour, and changing brightness. When the menu loads, it polls the colour, state, and brightness of the specific lamp using the Philips Hue REST API, and the corresponding options are marked as “selected” in the menu. When the user selects an item in the menu by Gaze-Hand Alignment, the previous selection is cleared, and a query is sent through the API to affect the selected changes in the lighting.

4 STUDY DESIGN FOR COMPARATIVE EVALUATION

We had three general objectives for the evaluation of our techniques. First, to gain insight into how our two techniques compare as alternative designs that are both based on the concept of Gaze-Hand Alignment. Secondly, to compare the techniques against a state of the art gaze-assisted freehand selection technique. Thirdly, to compare all three gaze-based techniques against as baseline of freehand selection without any gaze assistance. We designed our study accordingly as a factorial experiment with interaction technique as the main factor and Gaze&Finger, Gaze&Hand, Gaze&Pinch and Point&Pinch as the four conditions.

4.1 Task Design

We designed a compound object and menu selection task for the study, where the user activates a menu in a first subtask and then selects an item in the menu in a second subtask. The compound task is grounded in contextual user interfaces in AR where interfaces appear spatially anchored to objects in the physical world [9], which others have also explored as a use case for gaze-based

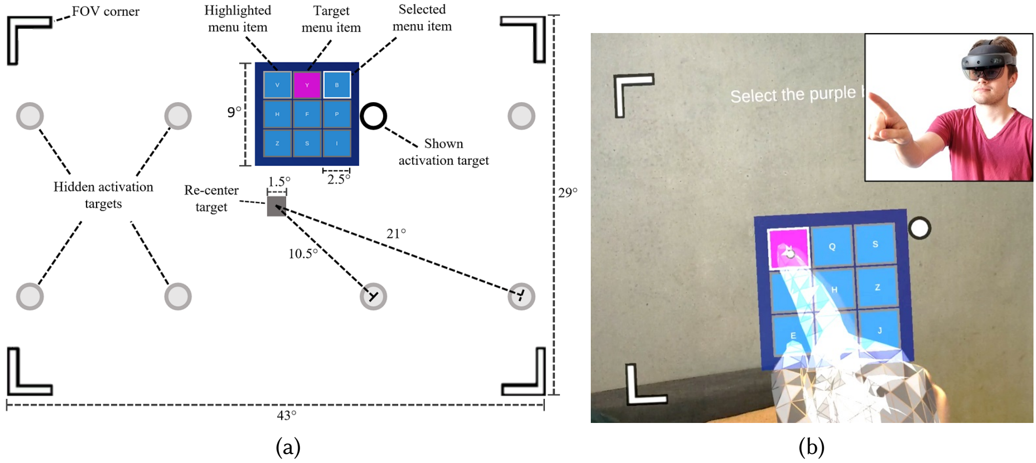


Fig. 3. User study task. (a) Layout of activation targets in the FOV, and of a menu for item selection. The task involved first selection of activation target appearing in one of eight possible positions, and then selection of an item in a menu opened next the activation target. (b) View of the AR interface for the task, and of a participant interacting with the interface.

interaction [19–21, 29, 35]. The task structure allows us to investigate selection subtasks separately, where the first selection is an open environment while the second requires selection from multiple targets presented in close proximity. For a UI design representative of head-worn displays, we use an abstract version of the HoloLens 2 "Start Menu". We chose a 3 × 3 cell-matrix with a cell size of about 2.5° visual angle, to have realistic target sizes and visual parameters in relation to the capabilities of state of the art head-mounted AR systems.

Figure 3 illustrate the task. Each task trial starts when the user centers their FOV by facing a rectangular target of 1.5° size located 1.8 meters in front of the user at eye level. To guide the participant in this step, a cursor is shown in the direction of the user's head pose, which they need to keep on the target for 800 ms to ensure that each trial starts from a central position. Once a trial has been triggered, a circular target of 2° width appears for menu activation. The activation target appears in one of eight positions in the FOV, rendered at 1.8m as a representative depth for remote selection where targets are out-of-reach but still clearly visible [15]. Once the participant has selected the activation target, a menu is displayed next to it, with one of 9 items highlighted as the selection target. The menu always appears to the left of the activation target and is always in full view as users naturally re-align the FOV with a head shift when an activation target appears closer to the display edge. Feedback on Pre-selection of an item is presented as a grey border around the item. Once the item has been selected, the compound task is completed.

4.2 Experimental Design and Procedure

We used the Microsoft HoloLens 2 as apparatus for our study, with a head-mounted display of 43° × 29° FOV (60Hz) and built-in 6DOF tracking, hand tracking, and eye-tracking[3]. All study software was implemented on the HoloLens 2.

We used a within-subject design, where all participants completed the task with all four interaction techniques. Gaze&Finger and Gaze&Hand were implemented as described above. Gaze&Pinch

[3] https://docs.microsoft.com/en-us/HoloLens/HoloLens2-hardware - accessed 1st of September 2021.

was implemented based on the MRTK which provides input events for the gaze position and pinch gestures. A selection is triggered at the time of receiving a pinch-in event. Point&Pinch was implemented based on the MRTK with a ray projected from the user's hand and is, in essence, controlled by the direction from the user to their hand.. Visual feedback is provided by a dashed line from the user's hand to a cursor shown where the ray intersects the depth plane of the target. A selection is triggered when a pinch-in event is received from the MRTK.

In addition to interaction technique, we manipulated activation target distance as factor with two levels (10.5°, 21°), see Fig. 3a. We counterbalanced the two factors across participants using a Latin Square. We further controlled task variation with presentation of the activation target in 4 different positions for each distance, and with highlighting different items in the menu for selection. We used five of the nine menu items as pseudo-random targets (corner and centre item) for variation while limiting the total number of trials per technique to $4 \times 2 \times 4 \times 5 = 160$ trials per participant.

At the start, participants filled out consent and demographic forms, and were then introduced to the study. For each technique, participants first calibrated the eye tracker and performed 10 training trials before conducting the 160 test trials. Participants received verbal assistance from the experimenter in training. Participants were instructed to be as fast and error-free as possible. If an error occurred (i.e. a wrong item was selected in the menu), the task was considered completed incorrectly and participants continued with the next trial. After each technique, participants had a break and filled out a usability questionnaire. After completion of all conditions, participants filled out a ranking and conducted a short interview. The study took about 40 minutes per participant.

4.3 Evaluation Measures and Data Analysis

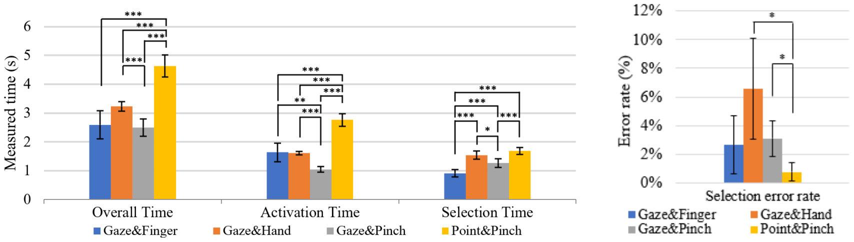
We used the following measures for evaluation:

(1) **Completion Time**: For performance measurement, task time started when an activation target appeared. We measured *overall time* for completion of the compound menu task, and *activation time* for completion of the subtask of selecting an activation target. We calculated *selection time* by subtracting activation time from overall time, to capture performance of the item selection subtask.
(2) **Error rate**: Error rates were measured by counting trials in which the wrong item was selected as *selection error* and trials that timed out after a limit of 10 seconds as *timeout error*.
(3) **User-reported metrics ratings**: We used rating scales (0-21) for six questions from the NASA TLX [7] and an additional question on Eye Fatigue, and had participants rank techniques in order of preference.
(4) **User Feedback**: Comments on user experience were collected in the post-task questionnaires and a post-session interview.

Task completion times were analysed with two-way repeated-measures ANOVA. Greenhouse-Geiser correction was applied when Mauchly's sphericity test indicated a violation and Bonferroni adjustments were applied for post-hoc comparisons. For completion time analysis, trials that had timed out were excluded. Data is reported to pass normality as skewness and kurtosis were within ±1.5. For the remaining dependent variables, we use the Friedman test with post-hoc Bonferroni corrected Wilcoxon tests. Effect sizes are reported as partial eta squared (η_p^2) and error bars in diagrams show 95% CI.

5 STUDY RESULTS

We recruited 16 volunteers for the user study (2 female). Participants were between 17 and 59 years old with a mean age of 35.88 (SD = 15.2). Out of the 16 participants, one wore contact lenses, three wore glasses, and all were right-handed. 14 participants reported no or little prior with AR/VR



(a) Overall, activation, and selection completion times. (b) The selection error rate.

Fig. 4. Results on task completion times (a) and error rates (b). Statistical significance shown as * for $p < .05$, ** for $p < .01$, *** for $p < .001$.

while 2 reported some experience. All 16 reported no or little prior experience with eye-tracking and 3D freehand gestures. The study took place in participants' homes under the guidance of the experimenter, who ensured comparable lighting conditions for participants to easily see their hands through the HMD. Participants performed the experimental task in standing position, facing a wall or surface chosen to minimise visual distraction.

In total we collected data for 10,240 trials. We removed outliers from the analysis where the task timed out (approx. 3.36% of trials) or where task completion time exceeded ($M \pm 4 \cdot SD$) (approx. 3.95%), to limit the influence of eye and hand tracking issues on analysis of technique performance. Timeout errors were observed across all conditions, however with a higher rate for Point&Pinch (($M = 6.25\%$). The performance results are summarised in Figure 4 and the user-reported metrics in Figure 5. We discuss the results for the compound task performance, performance of the menu activation and item selection subtasks, and user-reported data.

5.1 Compound Menu Task Performance

The key performance measures for the menu task on the whole are overall time for completion (Fig. 4a, left) and error rate (Fig. 4b). We found a significant main effect of technique on overall time ($F3, 45 = 48.544, p < .001, \eta_p^2 = .764$) and post-hoc analysis showed that Gaze&Finger, Gaze&Hand, and Gaze&Pinch were significantly faster than Point&Pinch ($p < .001$). Gaze&Pinch was significantly faster than Gaze&Hand ($p < .001$). As expected, activation target distance also had a significant effect on overall time, but we did not find any significant two-way interaction between the factors. We also found a significant effect of technique on error rate using the Friedman test ($\chi^2 = 11.504, p = .009$), with post-hoc analysis showing that Point&Pinch is less error-prone than Gaze&Hand ($p = .018$) and Gaze&Pinch ($p = .048$). The mean error rates were lowest for Point&Pinch ($M = .78\%, SD = 1.2\%$) followed by Gaze&Finger ($M = 2.66\%, SD = 3.82\%$), Gaze&Pinch ($M = 3.13\%, SD_x = 2.33\%$), and Gaze&Hand ($M = 6.56\%, SD = 6.57\%$) but there was high variation in individual performance specifically with the alignment-based techniques.

As main finding, techniques that support freehand input with gaze enable significantly faster selection from a context menu, compared to unimodal freehand input. Users are more error-prone with gaze-assisted input, explained by the need to coordinate input across modalities in time, but observed error rates can be considered low for novel and unfamiliar techniques.

5.2 Menu Activation Subtask

The menu activation subtask was a basic point-and-select task, for which we captured activation time as performance measure (Fig. 4a, middle). Note that for this subtask, Gaze&Hand was a standard dwell time technique. A main effect on activation time was observed for technique (($F_{1.754,26.304} = 61.813, p < .001, \eta_p^2 = .805$)) as well as distance (($F_{1,15} = 34.734, p < .001, \eta_p^2 = .698$)). Post-hoc analysis showed that Gaze&Finger, Gaze&Hand, and Gaze&Pinch were significantly faster than Point&Pinch ($p < .001$). Gaze&Pinch was faster than Gaze&Finger ($p = .009$) and Gaze&Hand ($p < .001$). The mean activation time was fastest for Gaze&Pinch ($M = 1.05s, SD = .19s$), followed by Gaze&Hand ($M = 1.61s, SD = .12s$), Gaze&Finger ($M = 1.63s, SD = .61s$) and Point&Pinch ($M = 2.76s, SD = .41s$).

There are several findings to note here, in addition to the speed advantage of the gaze-based techniques over freehand selection. Comparison of Gaze&Pinch and Gaze&Hand replicates results from a recent study of pinch versus dwell for confirming gaze input [28]. That study had also found dwell time slower than pinch even though they had used a much shorter dwell time of 300ms compared to 800ms in our study. Gaze&Finger is slower than Gaze&Pinch, explained by the need to lift arm and finger into the visual field, but significantly faster than Point&Pinch as a selection is completed as soon as the manual pointer crosses onto the target, without further confirmation.

5.3 Item Selection Subtask

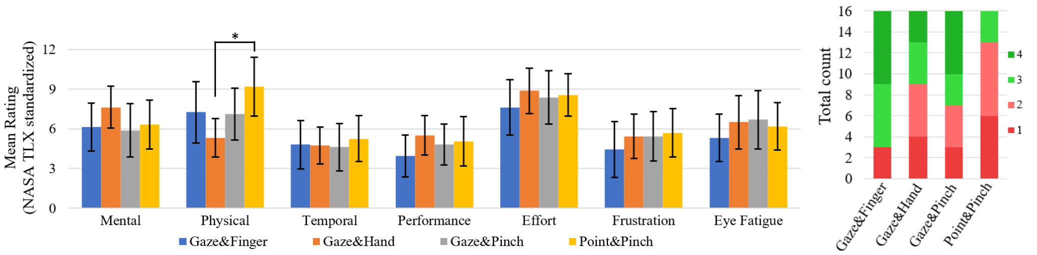
Item selection presents a distinct selection task as it starts from a position close to available targets, and as it requires selection of a specific target in a dense presentation of options. The key performance measures are selection time (Fig. 4a, right) and selection error (Fig. 4b) which we reported above. A significant main effect on selection time was observed for both technique (($F_{3,45} = 50.649, p < .001, \eta_p^2 = .772$)) and activation distance ($F_{1,15} = 16.249, p = .001, \eta_p^2 = .520$). Post-hoc analysis showed that Gaze&Finger was significantly faster than all other techniques ($p < .001$). Gaze&Pinch was significantly faster than Gaze&Hand ($p = .018$) and than Point&Pinch ($p < .001$). The mean selection time was fastest with Gaze&Finger ($M = 0.91s, SD = .25s$) followed by Gaze&Pinch ($M = 1.26s, SD = .28s$), Gaze&Hand ($M = 1.54s, SD = .28s$) and Point&Pinch ($M = 1.68s, SD = .23s$).

There are several findings to highlight. Gaze&Finger performs best for this subtask, indicating that is efficient for selections once arm and finger have already been lifted into visual field for a prior task. Gaze&Pinch is significantly slower which is interesting as a pinch gesture does not require more movement than an alignment gesture. This indicates that Gaze&Finger benefits from the natural coordination of eye and hand on a spatial task, compared to coordination of gaze with a separate semantic gesture. Gaze&Hand was slower and more error-prone, pointing to problems in replacing direct finger input with an indirectly controlled cursor. For alignment of a finger, users benefit from proprioception and focus gaze on the target, whereas with a cursor there is also tendency to gaze at the cursor. In our implementation, selection was triggered whenever gaze and cursor aligned, leading to a higher error rate with Gaze&Hand.

5.4 User-Reported Data and Feedback

Statistical analysis of rating scales did not reveal significant differences between techniques, except for *Physical Demand* ($\chi^2 = 8.562, p = 0.036$) (Fig. 5a). Participants perceived Point&Pinch as more demanding than Gaze&Hand ($p = .012$). Gaze&Hand had the lowest rating for physical demand ($M = 5.3, SD = 3.16$), followed by Gaze&Pinch ($M = 7.1, SD = 4.27$), Gaze&Finger ($M = 7.3, SD = 5.07$) and Point&Pinch ($M = 9.2, SD = 4.86$). In ranking of preferences, all participants preferred one or more of the gaze-based techniques over Point&Pinch.

We collected informal feedback from users on their experience with the four techniques. The feedback was not rigorously analysed but provides complementary insight into how techniques were perceived. Several participants commented on ease of use of Gaze&Finger, for example "*It was very nice that you can point and then click afterwards in one motion.*". One participant remarked "*It*



(a) NASA TLX post-task questionnaire results

(b) Preference Ranking

Fig. 5. (a) NASA TLX results (lower is better). Statistical significance shown as * for $p < .05$. (b) Ranking from 1 (least) to 4 (most preferred).

was so intuitive that I didn't even remember that you had to use the eyes.", indicating that the technique design is effective in letting the user focus on the manual alignment while gaze pre-selection is achieved implicitly.

Several participants commented that they found it harder to select activation targets that were further from the centre of the display. Other work on selection in HMDs found that some users use more eccentric eye movement to gaze at targets closer to the display edge, whereas others support their gaze with head movement which in an HMD moves targets closer to the display centre where eye-tracking typically performs better [41]. For Gaze&Hand, participants commented on unintended input that was triggered by looking at the cursor before reaching the target, explaining the higher error rate we observed. Another issue was associated with hand tracking (e.g. "*It was annoying that it couldn't track my hand when it was closer to the body*"). The technique afforded more flexibility as users did not have to lift their arm our setup was limited by the HoloLens' tracking range and had not been optimised for more casual input.

Gaze&Pinch received positive comments on speed, precision and ease of use but users also reported on issues with gesture detection (e.g. "*It was annoying that it didn't already register my click*") and late-trigger errors "*I thought it was going to be the right choice, but it then it changed at the last moment*"). This is a known problem in confirmation of gaze input by a separate "click" modality, as gaze has a tendency to move ahead of manual input, leaving a target before the click has been completed [42, 55]. With Point&Pinch, some participants specifically liked ray feedback, but a larger number of participants reported difficulty in pointing accurately while also performing a pinch for selection. Use of a pinch made the technique more comparable with Gaze&Pinch, but other trigger gestures (e.g. ThumbTrigger [50]) might align better with raycasting.

6 DISCUSSION

The idea that motivated this work is to use alignment of gaze and freehand pointing as a selection mechanism. We proposed Gaze-Hand Alignment as a general selection principle that can be implemented in different ways, as demonstrated with the design of Gaze&Finger and Gaze&Hand. The defining characteristics of Gaze-Hand Alignment are that both gaze and manual input are used simultaneously as pointing devices, and that a selection results from coincidence of both pointers on the same object in the interface. The advantage we expected Gaze-Hand Alignment to have over other existing strategies for combining gaze and hand input is that manual pointing implicitly involves gaze pointing to guide the hand. We presumed that gaze could be used for pre-selection of targets without users having to focus on this as an explicit step. In our study, we found this to work well with Gaze&Finger, where the user's focus is on alignment of their finger in the line of sight, as one integrated action of eye-hand coordinated pointing and selection. This is compelling as people are not normally inclined to pay attention to the agency of their eyes [10].

We designed Gaze&Hand as an alternative to Gaze&Finger to implement the same concept with indirect manual input, to reduce physical effort. As with Gaze&Finger, the assumption was that gaze would consistently lead the hand. However, replacing direct touch in the image plane with a cursor affected gaze behaviour. Users do not look back and forth between target and finger when they guide the finger to a target as they can rely on proprioception, but with a cursor, we saw a stronger tendency of using gaze to check the cursor position while it was being moved toward the target. In our study, this led to Midas Touch input which could be avoided by an improved implementation of Gaze&Hand that only triggers a selection when the cursor aligns with a gaze target, and not when gaze aligns with a cursor target. Gaze&Hand would likely compare more favourably as a result. However, we nonetheless see a clear trade-off between maximising benefit of eye-hand coordination through direct input in the line of sight versus reducing physical effort through use of indirect input.

The comparison of the Gaze-Hand Alignment techniques with Gaze&Pinch is interesting as the principal difference is in the type of manual gesture used in conjunction with gaze. In Gaze&Pinch the gesture is of a certain shape with predefined meaning and can be performed independently of position, relying on temporal coordination to associate the selection trigger with a gaze position. In Gaze&Finger, the gesture is spatial and has to be performed in the users FOV, using direct eye-hand coordination for selection. In Gaze&Hand, the manual gesture is spatial and relative to the display, relying on spatial alignment of a cursor with a gaze position. This affords flexibility for the gesture to be initiated from any starting position but requires an additional step for cursor instantiation. Gaze&Pinch and Gaze&Finger are more directly comparable for immediate selection, where Gaze&Pinch was more efficient for selection of targets presented at a distance from the hand (as in the menu activation subtask), whereas Gaze&Finger proved more efficient for selection of targets presented in proximity of the hand (as in the item selection subtask).

Semantic confirmation of gaze input may be advantageous for one-off selections anywhere on the interface, whereas Gaze-Hand Alignment may be more efficient for consecutive selections that are spatially connected, for example in nested menus, form filling, marking, input on keypads, or drawing applications. In our study, we have compared techniques for discrete selection. Gaze-Hand Alignment could also be extended to continuous input, where a widget selected by alignment could then be directly manipulated with further hand movement. However, while a control selected by pinch-in would be naturally released by pinch-out, a different mechanism will be needed when the selection is by alignment. Gaze would naturally focus on visual feedback provided for continuous input, and gaze shift away from the task could serve to disengage control.

While the study exposed trade-offs in comparison of the three gaze-based techniques, we found all of them strikingly faster in comparison with the manual baseline. This provides clear evidence of the utility of gaze in supporting freehand input. In other recent studies of gaze-with-hand versus hand-only input in VR the benefit of gaze support had not been clear [34, 54]. Our choice of task may have favoured gaze input but the result also suggests that the specific way in which we integrate input from eye and hand is effective. However, gaze has lower accuracy for pointing and the alignment-based techniques may perform less well on tasks that require more fine-grained selection. Alignment-based techniques could also be extended for fine-grained selection, for example with gaze pre-selecting multiple candidate targets and manual alignment employed to resolve ambiguity.

Both alignment techniques we introduced have specific limitations. Gaze&Hand requires a cursor for indirect manual input, and an additional mechanism to deploy the cursor “where the action is”. A cursor could also be provided continuously but that would introduce other problems such as movement over larger display space and the need for a clutching mechanism. Gaze&Finger is limited by parallax issues caused by depth disparity input and output in the FOV. For menu selection, we addressed the issue of warping and scaling the menu to the depth level of the user's finger but it remains a limitation for initial selection of targets at a distance, exacerbating accuracy limitations of gaze. As proposed in related work, this could be addressed by adapting the display to eye dominance [24, 25].

The study we reported has several limitations. We conducted the study under restrictions imposed by the COVID-19 pandemic. Participants were recruited through the personal network of the researchers, resulting in an almost exclusively male sample. This is not reflecting intended target usage across genders. However, we would expect our study results to generalise across genders, as past comparable experiments in VR have not indicated gender differences in task performance and preference. The study was conducted in the participants' homes instead of controlled lab space due to COVID-19 restrictions. However, it was guided by an experimenter to ensure comparable conditions for task performance and data collection. All participants were right-handed and our menu layout favoured right-handed usage by spawning to the left of an object selected, however an adaptation to left-handed usage would be straightforward. Data collection was limited by the apparatus used. A main technical limitation was hand tracking range, with tracking errors occurring when gestures were performed with the hand too close to the body, or when users performed a shift of their head and the head-mounted tracking system to the left, away from their input with the right hand. We used outlier removal to limit the influence of these tracking issues on our results.

7 CONCLUSION

The relationship of eye and hand is intricate. Spatial tasks such as pointing inherently require gaze to guide the hand into position. Our work on Gaze-Hand Alignment contributes to understanding how the natural coordination between eye and hand can be leveraged for multimodal input. There are several conclusions we can draw from the work.

- If gaze is tracked in conjunction with manual pointing, then it lends itself for pre-selection of targets as we naturally look ahead to pointing targets. This defines an additional input state akin to a hover and in mouse interaction. In Gaze-Hand Alignment, we have applied this to delimit manual input to pre-selected targets, addressing Midas Touch issues of mid-air pointing, and to trigger selection immediately when the manual pointer reaches the target, without need for a "click" method. What makes gaze compelling in this context is that its utility can remain transparent to users, to focus on manual pointing while using gaze naturally.
- If we consider both gaze and hand as pointing devices, then we can treat their alignment as meaningful input. In our implementation and study of Gaze-Hand Alignment, we demonstrate the efficacy of alignment of the two modalities as selection trigger. Gaze-Hand Alignment is conceptually interesting as it captures a state associated with more intent than "just looking". Even though eye and hand coordinate closely in pointing, their movement is not synchronous but effortlessly aligned at points of interest.
- The concept of Gaze-Hand Alignment opens a new design space in particular for input beyond manual reach, in 3D and AR. Our design of the Gaze&Finger and Gaze&Hand techniques demonstrates how the principle can be applied in fundamentally different ways, for user experiences of direct versus indirect input. Their comparison against established baselines contributes insight into design challenges and trade-offs, including parallax, feedback for relative input, and benefits of spatial versus semantic gestures for selection. These insights

have practical relevance for the design of emerging interfaces, enabled by HMDs that support eye- and hand-tracking technologies.

Overall, as a fundamental take-away, the work demonstrates value in grounding the design of multimodal gaze and mid-air techniques in natural eye-hand coordination. While this might seem obvious, it requires a shift in perspective, to view gaze not as entirely separate from other modes of input, but as working in synergy with our body and hands.

ACKNOWLEDGMENTS

This work has received funding by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grants no. 101021229 GEMINI, and no. 740548 CIO) and by the Innovation Fund Denmark, as part of the Manufacturing Academy of Denmark (MADE) FAST project.

REFERENCES

[1] Ferran Argelaguet and Carlos Andujar. 2013. A survey of 3D object selection techniques for virtual environments. *Computers & Graphics* 37, 3 (2013), 121–136.

[2] Amartya Banerjee, Jesse Burstyn, Audrey Girouard, and Roel Vertegaal. 2012. MultiPoint: Comparing laser and manual pointing as remote input in large display interactions. *International Journal of Human-Computer Studies* 70, 10 (2012), 690–702.

[3] Eric A. Bier, Maureen C. Stone, Ken Pier, Ken Fishkin, Thomas Baudel, Matt Conway, William Buxton, and Tony DeRose. 1994. Toolglass and Magic Lenses: The See-through Interface. In *Conference Companion on Human Factors in Computing Systems* (Boston, Massachusetts, USA) *(CHI '94)*. Association for Computing Machinery, New York, NY, USA, 445–446. https://doi.org/10.1145/259963.260447

[4] Richard A. Bolt. 1981. Gaze-Orchestrated Dynamic Windows. *SIGGRAPH Comput. Graph.* 15, 3 (Aug. 1981), 109–119. https://doi.org/10.1145/965161.806796

[5] Doug Bowman, Chadwick Wingrave, Joshua Campbell, and Vinh Ly. 2001. Using pinch gloves for both natural and abstract interaction techniques in virtual environments. *Proceedings of HCI International* (01 2001).

[6] Ishan Chatterjee, Robert Xiao, and Chris Harrison. 2015. Gaze+Gesture: Expressive, Precise and Targeted Free-Space Interactions. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction* (Seattle, Washington, USA) *(ICMI '15)*. Association for Computing Machinery, New York, NY, USA, 131–138. https://doi.org/10.1145/2818346.2820752

[7] Lacey Colligan, Henry W.W. Potts, Chelsea T. Finn, and Robert A. Sinkin. 2015. Cognitive workload changes for nurses transitioning from a legacy system with paper documentation to a commercial electronic health record. *International Journal of Medical Informatics* 84, 7 (2015), 469–476. https://doi.org/10.1016/j.ijmedinf.2015.03.003

[8] Tiare Feuchtner and Jörg Müller. 2017. *Extending the Body for Interaction with Reality.* Association for Computing Machinery, New York, NY, USA, 5145–5157. https://doi.org/10.1145/3025453.3025689

[9] Jens Grubert, Tobias Langlotz, Stefanie Zollmann, and Holger Regenbrecht. 2016. Towards pervasive augmented reality: Context-awareness in augmented reality. *IEEE transactions on visualization and computer graphics* 23, 6 (2016), 1706–1724.

[10] Ouriel Grynszpan, Jérôme Simonin, Jean-Claude Martin, and Jacqueline Nadel. 2012. Investigating social gaze as an action-perception online performance. *Frontiers in human neuroscience* 6 (2012), 94.

[11] Robert J. K. Jacob. 1990. 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* (Seattle, Washington, USA) *(CHI '90)*. Association for Computing Machinery, New York, NY, USA, 11–18. https://doi.org/10.1145/97243.97246

[12] Roland S Johansson, Göran Westling, Anders Bäckström, and J Randall Flanagan. 2001. Eye–hand coordination in object manipulation. *Journal of neuroscience* 21, 17 (2001), 6917–6932.

[13] Ricardo Jota, Miguel A. Nacenta, Joaquim A. Jorge, Sheelagh Carpendale, and Saul Greenberg. 2010. A Comparison of Ray Pointing Techniques for Very Large Displays. In *Proceedings of Graphics Interface 2010*. Canadian Information Processing Society, CAN, 269–276.

[14] Aarlenne Z Khan and J Douglas Crawford. 2003. Coordinating one hand with two eyes: optimizing for field of view in a pointing task. *Vision research* 43, 4 (2003), 409–417.

[15] Mikko Kytö, Barrett Ens, Thammathip Piumsomboon, Gun A. Lee, and Mark Billinghurst. 2018. Pinpointing: Precise Head- and Eye-Based Target Selection for Augmented Reality. In *Proceedings of the 2018 CHI Conference on Human

Factors in Computing Systems (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173655

[16] Michael Land, Neil Mennie, and Jennifer Rusted. 1999. The roles of vision and eye movements in the control of activities of daily living. *Perception* 28, 11 (1999), 1311–1328.

[17] Gun A. Lee, Mark Billinghurst, and Gerard Jounghyun Kim. 2004. Occlusion Based Interaction Methods for Tangible Augmented Reality Environments. In *Proceedings of the 2004 ACM SIGGRAPH International Conference on Virtual Reality Continuum and Its Applications in Industry* (Singapore) *(VRCAI '04)*. Association for Computing Machinery, New York, NY, USA, 419–426. https://doi.org/10.1145/1044588.1044680

[18] Sangyoon Lee, Jinseok Seo, Gerard Jounghyun Kim, and Chan-Mo Park. 2003. Evaluation of pointing techniques for ray casting selection in virtual environments, In Third international conference on virtual reality and its application in industry. *Proceedings of SPIE - The International Society for Optical Engineering* 4756, 38–44. https://doi.org/10.1117/12.497665

[19] Feiyu Lu and Doug A. Bowman. 2021. Evaluating the Potential of Glanceable AR Interfaces for Authentic Everyday Uses. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*. IEEE Computer Society, Los Alamitos, CA, USA, 768–777. https://doi.org/10.1109/VR50410.2021.00104

[20] Feiyu Lu, Shakiba Davari, Lee Lisle, Yuan Li, and Doug A. Bowman. 2020. Glanceable AR: Evaluating Information Access Methods for Head-Worn Augmented Reality. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE Computer Society, Atlanta, Georgia, USA, 930–939. https://doi.org/10.1109/VR46266.2020.00113

[21] Yiqin Lu, Chun Yu, and Yuanchun Shi. 2020. Investigating Bubble Mechanism for Ray-Casting to Improve 3D Target Acquisition in Virtual Reality. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, Atlanta, GA, USA, 35–43. https://doi.org/10.1109/VR46266.2020.00021

[22] Christof Lutteroth, Moiz Penkar, and Gerald Weber. 2015. Gaze vs. Mouse: A Fast and Accurate Gaze-Only Click Alternative. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology* (Charlotte, NC, USA) *(UIST '15)*. Association for Computing Machinery, New York, NY, USA, 385–394. https://doi.org/10.1145/2807442.2807461

[23] Päivi Majaranta, Ulla-Kaija Ahola, and Oleg Špakov. 2009. Fast Gaze Typing with an Adjustable Dwell Time. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 357–360. https://doi.org/10.1145/1518701.1518758

[24] Diako Mardanbegi, Benedikt Mayer, Ken Pfeuffer, Shahram Jalaliniya, Hans Gellersen, and Alexander Perzl. 2019. EyeSeeThrough: Unifying Tool Selection and Application in Virtual Environments. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, Osaka, Japan, 474–483. https://doi.org/10.1109/VR.2019.8797988

[25] Sven Mayer, Katrin Wolf, Stefan Schneegass, and Niels Henze. 2015. Modeling Distant Pointing for Compensating Systematic Displacements. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 4165–4168. https://doi.org/10.1145/2702123.2702332

[26] Mark R. Mine, Frederick P. Brooks, and Carlo H. Sequin. 1997. Moving Objects in Space: Exploiting Proprioception in Virtual-Environment Interaction. In *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '97)*. ACM Press/Addison-Wesley Publishing Co., USA, 19–26. https://doi.org/10.1145/258734.258747

[27] Emilie Møllenbach, John Paulin Hansen, and Martin Lillholm. 2013. Eye movements in gaze interaction. *Journal of Eye Movement Research* 6, 2 (05 2013), 1–15. https://doi.org/10.16910/jemr.6.2.1

[28] Aunnoy K Mutasim, Anil Ufuk Batmaz, and Wolfgang Stuerzlinger. 2021. *Pinch, Click, or Dwell: Comparing Different Selection Techniques for Eye-Gaze-Based Pointing in Virtual Reality.* Association for Computing Machinery, New York, NY, USA, Chapter 15, 7. https://doi.org/10.1145/3448018.3457998

[29] Ken Pfeuffer, Yasmeen Abdrabou, Augusto Esteves, Radiah Rivu, Yomna Abdelrahman, Stefanie Meitner, Amr Saadi, and Florian Alt. 2021. ARtention: A design space for gaze-adaptive user interfaces in augmented reality. *Computers & Graphics* 95 (2021), 1–12. https://doi.org/10.1016/j.cag.2021.01.001

[30] Ken Pfeuffer, Jason Alexander, Ming Ki Chong, and Hans Gellersen. 2014. Gaze-Touch: Combining Gaze with Multi-Touch for Interaction on the Same Surface. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology* (Honolulu, Hawaii, USA) *(UIST '14)*. Association for Computing Machinery, New York, NY, USA, 509–518. https://doi.org/10.1145/2642918.2647397

[31] Ken Pfeuffer, Jason Alexander, Ming Ki Chong, Yanxia Zhang, and Hans Gellersen. 2015. Gaze-Shifting: Direct-Indirect Input with Pen and Touch Modulated by Gaze. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology* (Charlotte, NC, USA) *(UIST '15)*. Association for Computing Machinery, New York, NY, USA, 373–383. https://doi.org/10.1145/2807442.2807460

[32] Ken Pfeuffer and Hans Gellersen. 2016. Gaze and Touch Interaction on Tablets. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (Tokyo, Japan) *(UIST '16)*. Association for Computing Machinery,

New York, NY, USA, 301–311. https://doi.org/10.1145/2984511.2984514

[33] Ken Pfeuffer, Benedikt Mayer, Diako Mardanbegi, and Hans Gellersen. 2017. Gaze + Pinch Interaction in Virtual Reality. In *Proceedings of the 5th Symposium on Spatial User Interaction* (Brighton, United Kingdom) *(SUI '17)*. Association for Computing Machinery, New York, NY, USA, 99–108. https://doi.org/10.1145/3131277.3132180

[34] Ken Pfeuffer, Lukas Mecke, Sarah Delgado Rodriguez, Mariam Hassib, Hannah Maier, and Florian Alt. 2020. Empirical Evaluation of Gaze-Enhanced Menus in Virtual Reality. In *26th ACM Symposium on Virtual Reality Software and Technology* (Virtual Event, Canada) *(VRST '20)*. Association for Computing Machinery, New York, NY, USA, Article 20, 11 pages. https://doi.org/10.1145/3385956.3418962

[35] Robin Piening, Ken Pfeuffer, ANDTim Mittermeier Augusto Esteves, Sarah Prange, Philippe Schroeder, and Florian Alt. 2021. Looking for Info: Evaluation of Gaze Based Information Retrieval in Augmented Reality. In *Proceedings of the 18th IFIP TC 13 International Conference on Human-Computer Interaction* (Bari, Italy) *(INTERACT '21)*. Springer, Berlin-Heidelberg, Germany, 544–565.

[36] Jeffrey S. Pierce, Andrew S. Forsberg, Matthew J. Conway, Seung Hong, Robert C. Zeleznik, and Mark R. Mine. 1997. Image Plane Interaction Techniques in 3D Immersive Environments. In *Proceedings of the 1997 Symposium on Interactive 3D Graphics* (Providence, Rhode Island, USA) *(I3D '97)*. Association for Computing Machinery, New York, NY, USA, 39–ff. https://doi.org/10.1145/253284.253303

[37] Ivan Poupyrev, Mark Billinghurst, Suzanne Weghorst, and Tadao Ichikawa. 1996. The Go-Go Interaction Technique: Non-Linear Mapping for Direct Manipulation in VR. In *Proceedings of the 9th Annual ACM Symposium on User Interface Software and Technology* (Seattle, Washington, USA) *(UIST '96)*. Association for Computing Machinery, New York, NY, USA, 79–80. https://doi.org/10.1145/237091.237102

[38] I. Poupyrev, S. Weghorst, M. Billinghurst, and T. Ichikawa. 1998. Egocentric Object Manipulation in Virtual Environments: Empirical Evaluation of Interaction Techniques. *Computer Graphics Forum* 17 (1998), 12 pages.

[39] Robin Schweigert, Valentin Schwind, and Sven Mayer. 2019. EyePointing: A Gaze-Based Selection Technique. In *Proceedings of Mensch Und Computer 2019* (Hamburg, Germany) *(MuC'19)*. Association for Computing Machinery, New York, NY, USA, 719–723. https://doi.org/10.1145/3340764.3344897

[40] Linda E. Sibert and Robert J. K. Jacob. 2000. Evaluation of Eye Gaze Interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (The Hague, The Netherlands) *(CHI '00)*. Association for Computing Machinery, New York, NY, USA, 281–288. https://doi.org/10.1145/332040.332445

[41] Ludwig Sidenmark and Hans Gellersen. 2019. Eye, Head and Torso Coordination During Gaze Shifts in Virtual Reality. *ACM Trans. Comput.-Hum. Interact.* 27, 1, Article 4 (dec 2019), 40 pages. https://doi.org/10.1145/3361218

[42] Ludwig Sidenmark and Hans Gellersen. 2019. Eye&Head: Synergetic Eye and Head Movement for Gaze Pointing and Selection. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (New Orleans, LA, USA) *(UIST '19)*. Association for Computing Machinery, New York, NY, USA, 1161–1174. https://doi.org/10.1145/3332165.3347921

[43] Ludwig Sidenmark and Anders Lundström. 2019. Gaze Behaviour on Interacted Objects during Hand Interaction in Virtual Reality for Eye Tracking Calibration. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications* (Denver, Colorado) *(ETRA '19)*. Association for Computing Machinery, New York, NY, USA, Article 6, 9 pages. https://doi.org/10.1145/3314111.3319815

[44] Ludwig Sidenmark, Diako Mardanbegi, Argenis Ramirez Gomez, Christopher Clarke, and Hans Gellersen. 2020. BimodalGaze: Seamlessly Refined Pointing with Gaze and Filtered Gestural Head Movement. In *ACM Symposium on Eye Tracking Research and Applications* (Stuttgart, Germany) *(ETRA '20 Full Papers)*. Association for Computing Machinery, New York, NY, USA, Article 8, 9 pages. https://doi.org/10.1145/3379155.3391312

[45] Ludwig Sidenmark, Dominic Potts, Bill Bapisch, and Hans Gellersen. 2021. Radi-Eye: Hands-Free Radial Interfaces for 3D Interaction Using Gaze-Activated Head-Crossing. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 740, 11 pages. https://doi.org/10.1145/3411764.3445697

[46] Sophie Stellmach and Raimund Dachselt. 2012. Look & Touch: Gaze-Supported Target Acquisition. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 2981–2990. https://doi.org/10.1145/2207676.2208709

[47] Jayson Turner, Jason Alexander, Andreas Bulling, and Hans Gellersen. 2015. Gaze+RST: Integrating Gaze and Multitouch for Remote Rotate-Scale-Translate Tasks. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 4179–4188. https://doi.org/10.1145/2702123.2702355

[48] Radu-Daniel Vatavu. 2012. User-Defined Gestures for Free-Hand TV Control. In *Proceedings of the 10th European Conference on Interactive TV and Video* (Berlin, Germany) *(EuroITV '12)*. Association for Computing Machinery, New York, NY, USA, 45–48. https://doi.org/10.1145/2325616.2325626

[49] Roel Vertegaal et al. 2003. Attentive user interfaces. *Commun. ACM* 46, 3 (2003), 30–33.

[50] Daniel Vogel and Ravin Balakrishnan. 2005. Distant Freehand Pointing and Clicking on Very Large, High Resolution Displays. In *Proceedings of the 18th Annual ACM Symposium on User Interface Software and Technology* (Seattle, WA, USA) *(UIST '05)*. Association for Computing Machinery, New York, NY, USA, 33–42. https://doi.org/10.1145/1095034.1095041

[51] Bryan Wang and Tovi Grossman. 2020. *BlyncSync: Enabling Multimodal Smartwatch Gestures with Synchronous Touch and Blink*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376132

[52] Colin Ware and Harutune H. Mikaelian. 1986. An Evaluation of an Eye Tracker as a Device for Computer Input2. In *Proceedings of the SIGCHI/GI Conference on Human Factors in Computing Systems and Graphics Interface* (Toronto, Ontario, Canada) *(CHI '87)*. Association for Computing Machinery, New York, NY, USA, 183–188. https://doi.org/10.1145/29933.275627

[53] Pierre Weill-Tessier and Hans Gellersen. 2017. Touch input and gaze correlation on tablets. In *International Conference on Intelligent Decision Technologies*. Springer, Springer International Publishing, Cham, 287–296.

[54] Difeng Yu, Xueshi Lu, Rongkai Shi, Hai-Ning Liang, Tilman Dingler, Eduardo Velloso, and Jorge Goncalves. 2021. *Gaze-Supported 3D Object Manipulation in Virtual Reality*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3411764.3445343

[55] Shumin Zhai, Carlos Morimoto, and Steven Ihde. 1999. Manual and Gaze Input Cascaded (MAGIC) Pointing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Pittsburgh, Pennsylvania, USA) *(CHI '99)*. Association for Computing Machinery, New York, NY, USA, 246–253. https://doi.org/10.1145/302979.303053

[56] Xuan Zhang and I Scott MacKenzie. 2007. Evaluating eye tracking with ISO 9241-Part 9. In *International Conference on Human-Computer Interaction*. Springer, Springer Berlin Heidelberg, Berlin, Heidelberg, 779–788.

Received November 2021; revised January 2022; accepted April 2022