2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct) | 978-1-6654-5365-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISMAR-ADJUNCT57072.2022.00132

An Evaluation of Caret Navigation Methods for Text Editing in Augmented Reality

Jinghui Hu[*] John J. Dudley[†] Per Ola Kristensson[‡]

University of Cambridge

Abstract

A common task in text editing is navigating the text caret as part of the text editing process. We explore three key dimensions in the design space of caret navigation techniques: cursor placement method, visual expansion in the form of a magnifying lens, and display distance. In addition to two commonly used cursor placement methods, direct touch and raycast, we also present a novel multimodal text cursor placement method combining eye gaze and touch gestures for optical see-through augmented reality (AR). This method allows the user to refine the caret position with an indirect mid-air virtual touch block after the caret has snapped to a location provided by an initial eye fixation. We derive eight combinations from three design dimensions and study their performance in a user study with 24 participants. Our results reveal that: 1) raycast delivered the fastest completion times among all combinations evaluated; 2) in near-field conditions the multimodal method can achieve similar performance as direct touch input with less physical effort; and 3) magnifying lenses offer no significant performance advantages for caret navigation in AR.

Index Terms: Human-centered computing—Human computer interaction (HCI)—Interaction techniques—Text input; Computing methodologies—Computer graphics—Graphics systems and interfaces—Mixed / augmented reality

1 Introduction

For augmented reality (AR) headsets to gradually become mainstream, it will be necessary to increasingly provide good support for common tasks, such as text entry and text editing [11]. Adding annotations, sending short emails or messages, reading and taking notes all have the potential to become common practices on AR headsets in the near future. However, while text entry in AR has been widely studied (e.g. [1, 2, 7, 8]), our understanding of text editing interactions in AR is still in its infancy. Text editing essentially tackles a complex set of operations involving caret navigation, text selection and error correction. Navigating the text caret is the first step and integral to that process. Caret navigation essentially requires the user to perform rapid and accurate selections of small targets roughly equivalent to the width of individual characters. Supporting accurate navigation between characters is often difficult in AR headsets with the absence of a keyboard, mouse or touchscreen. Although alternative input modalities available on modern AR headsets, such as hand tracking, eye and head tracking, provide other design opportunities, they tend to be too noisy and error-prone to allow precise text caret navigation.

Several studies have explored multimodal target selection techniques incorporating eye gaze and manual input [5, 22]. Eye gaze movement is relatively fast but inaccurate while hand movement can offer a wide variety of expressive gestures and manipulations in a continuous manner. Therefore, the two modalities are by nature complementary. A rapid and precise input can be offered by a combination of eye and manual manipulations which could be a suitable solution for navigating a text caret in AR. Further, since different conditions of the visual display influence the performance of target selection [23] and text editing [9] in AR, the visual features presented to the user are also likely to influence the user experience of caret navigation.

In this paper we explore the design space for caret navigation methods for AR and evaluate the influence of the display distance, any visual expansion (magnifying lenses), and the specific cursor placement method. We find that: 1) raycast delivers the fastest completion times; 2) in near-field conditions a combination of eye gaze and touch can achieve similar performance as direct touch input but with less physical effort; and 3) magnifying lenses offer no significant performance advantages.

2 Related Work

Text editing is a complex task that requires constant switching between several subtasks. There is very limited prior research on text editing in AR. To the best of our knowledge, all of the previously proposed techniques for text caret navigation require some form of peripheral input device and none of the prior studies solely use an HMD. Ghosh et al. [9] combined speech input with a hand-held controller to enable two modes of heads-up text editing in an on-the-go situation. Darbar et al. [6] explored eyes-free text selection techniques with a smartphone.

The idea of combining gaze with manual input was first introduced by Zhai et al. [24] in a desktop environment to reduce physical effort and fatigue. They used gaze to support faster cursor movement which was subsequently refined with a mouse to improve accuracy.

This work was later followed by other explorations combining eye movement with distant displays [22,23], desktop interactions [5], and touchscreen interactions with finger [15, 17] or pen input [16]. Kyto et al. [12] adapted several multimodal selection techniques for AR headsets pairing eye gaze with head movement, hand gestures and a handheld device. The refinement modalities showed drastic improvements in the speed and accuracy of selection tasks.

Text editing using gaze-supported input on touchscreens has also been studied. Rivu et al. [19, 20] extended touchscreen text editing on tablets by incorporating gaze interaction. In their design, a gaze button activates various gaze interactions when tapped or held, thus boosting the expressiveness of the interface. Sindhwani et al. [21] used gaze as an additional modality for positioning small text edits, such as error corrections. Users retype the correct words and gaze at the error location to modify the text automatically without using a mouse. Zhao et al. [25] used eye gaze to locate users' intended words in a hands-free error correction experience on mobile devices.

In the context of virtual and augmented reality, Pfeuffer et al. [18] proposed Gaze+Pinch to enable distant object manipulation in a manner similar to direct manipulation. Two recent works have also integrated gaze into the typing experience. He et al. [10] used eye gaze to select between word predictions in VR. Lystbæk et al. [14] explored freehand text entry in AR with the assistance of gaze.

[*] e-mail: jh2265@cam.ac.uk

[†] e-mail: jjd50@cam.ac.uk

[‡] e-mail: pok21@cam.ac.uk

2771-1110/22/$31.00 ©2022 IEEE
DOI 10.1109/ISMAR-Adjunct57072.2022.00132

Authorized licensed use limited to: CAMBRIDGE UNIV. Downloaded on January 05,2023 at 17:00:09 UTC from IEEE Xplore. Restrictions apply.

		Display Distance	
		Near-field	Far-field
Visual Expansion	No Lens	Near: DirectTouch	Far: DirectTouch
		Near: Raycast	Far: Raycast
		Near: EyeGazeTouch	Far: EyeGazeTouch
	With Lens	Near: DirectTouch+Lens	Far: DirectTouch+Lens
		Near: Raycast+Lens	Far: Raycast+Lens
		Near: EyeGazeTouch+Lens	Far: EyeGazeTouch+Lens

Figure 1: Summary of the design space exploration covering alternative configurations for CURSOR PLACEMENT METHOD, DISPLAY DISTANCE and VISUAL EXPANSION. In total, eight TECHNIQUES were tested in the experiment. As DirectTouch is only compatible with near-field situations and Raycast is designated for far-field situations, they are mutually exclusive. The four incompatible and infeasible combinations are shaded in the table.

Chatterjee et al. [5] proposed a 'gaze plus free-space gesture' to accommodate fast and accurate interactions. An example application mentioned in their work for their Gaze+Gesture technique was to perform text selection in a word processor scenario. This work by Chatterjee et al. [5] is the closest conceptually to the multimodal cursor navigation technique we propose in this paper. However, we utilize a mid-air touchpad instead of using a pinch gesture to allow for more expressiveness in the later design of selection and other text editing operations.

Caret navigation and text selection are analogous to target acquisition but caret navigation is unique in the sense that characters are small in size but large in number. Thus, we propose to take advantage of eye gaze for fast cursor movements over large distances and a secondary hand input for precise refinement in order to support accurate caret placement.

3 DESIGN EXPLORATION

Caret navigation in AR has not been systematically investigated or evaluated. The study in this paper investigates three key dimensions that are particularly relevant to the design of caret navigation methods: cursor placement method, display distance and visual expansion. Each of these design dimensions is described in detail below. As shown in Figure 1, there are eight feasible techniques representing different combinations of these key dimensions.

3.1 Cursor Placement Method

DirectTouch The text field displayed in AR is treated as an imaginary mid-air 2D touchscreen (shown in Figure 2 (a)). Users first touch the virtual text plane to move the caret. The finger position indicates the absolute position of the caret. The navigation is confirmed when the finger touch leaves the virtual plane. There is a small depth threshold (0.5 cm) that delineates the touch-on and touch-off states. The depth position of the fingertip also imitates the duration of maintaining a press action on a touchscreen. DirectTouch is an absolute and direct method working on a 1:1 control-display ratio and for near-field interactions only. In the scene in the study, the touch-on state results in visual feedback in the form of the caret turning from red to green.

Raycast A raycast method is commonly used with 6DoF controllers on immersive headsets. HoloLens 2 uses raycast as the default far-field interaction technique. This default raycast method (see Fig, 2 (b)) uses a ray projected from the hand and an 'air-tap' or pinch gesture. Our implementation of the Raycast technique displays a caret icon when the ray is intersecting with the text plane. This is similar to other conditions and users' previous experience of text editing on touchscreens and laptops/workstations.

EyeGazeTouch This multimodal technique combines eye fixation with hand input and offers potential advantages for fast and accurate targeting of small objects. Similar concepts have been tested in target selection tasks [5, 12]. We redesigned and modified this approach of using tracked eye gaze with gestures by adding a trackpad-like touch block for the hand refinement phase. As shown in Figure 1 (c)–(d), a vertical input block is placed within the user's reach and at the *wrist* height, as suggested by prior work [4]. Importantly, this input block works in a relative mode. We set it to a low control-display ratio to ensure accurate and precise navigation between characters. The dimensions of the touch block in our study were set at $30 \times 20 \times 10$ cm.

This multimodal technique leverages the strengths of each input modality: eye gaze ensures the swift movement of the caret across large distances while a touch gesture allows for subtle but precise fine-tuning.

In the first phase, a subtle eye cursor follows the user's fixation on the text display. When the user performs a touch-on action on the blue block (see Fig. 2 (c)–(d)), the caret will snap to the eye fixation position on the text display. The user then moves their fingertip inside the block to refine the caret position. This addresses the fact that coarse eye tracking is unreliable for selection of character-level targets. Finally, retracting the fingertip from the block confirms the navigation result. In our implementation, a 4 cm movement of the fingertip moves the caret by one character (informed by a pilot study). This multimodal technique can be applied in both near-field and far-field settings. We chose to use the mid-air touch block for the second refinement phase instead of a pinch gesture in order to leave more design space available for other text editing operations, such as text selection, which will inevitably be required for a complete AR text editing system. An alternative approach would be to drag while holding a pinch gesture in order to refine the target position and then releasing the pinch to confirm. However, utilizing the pinch gesture in this way would then limit its use in subsequent editing tasks, such as to indicate the start/stop of a drag selection.

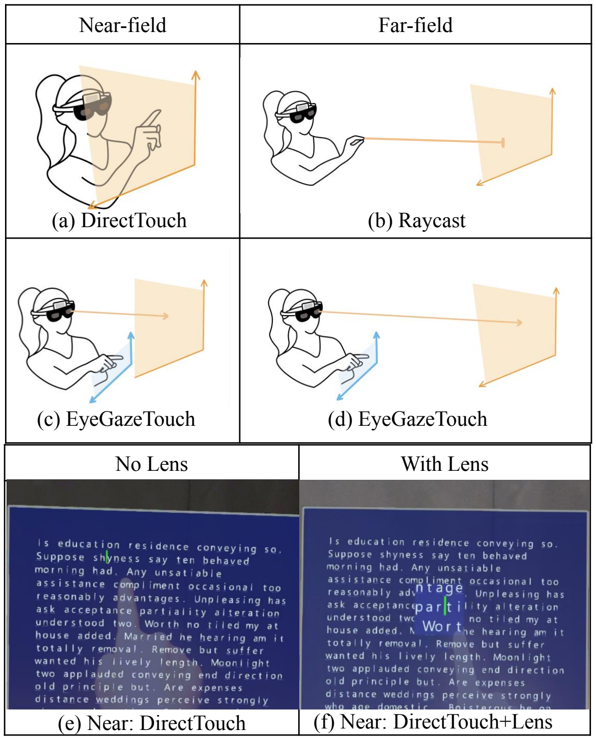
3.2 Visual Expansion

A magnifying lens is commonly used in caret navigation on touchscreens to both accommodate the narrow width of characters and eliminate occlusion problems. Previous studies have explored the use of magnifying lenses in AR [23]. To evaluate the potential benefit of this form of visual expansion for text editing in AR, we implemented a magnifying lens interaction as shown in Figure 2 (f). For DirectTouch mode, the magnifying lens appears at the user's fingertip location when it is in the touch-on state on the text plane. For Raycast mode, the lens shows at the ray interaction point. A caret cursor is also shown in the middle of the lens to ensure these conditions are consistent with those without a magnifying lens. For EyeGazeTouch mode, we keep the lens invisible in the eye gaze phase. The lens will only appear during the refinement phase when users touch on the touchpad block. For all the conditions with a magnifying lens, the size (5×5 cm) and magnification times (1.5) are kept identical.

3.3 Display Distance

DISPLAY DISTANCE describes the two distinct categories of content presentation and interaction (near- and far-field) that can coexist within an AR application. Direct manipulation and Raycast are the two default input methods used in HoloLens 2 for near-field and far-field interactions respectively. One crucial advantage of AR is the spatial freedom of content. Text and other objects can be placed near the user or away from the user. We evaluate caret navigation by exploiting the existing interaction system to ensure the function is viable for most situations. We use two display distances formed by

Authorized licensed use limited to: CAMBRIDGE UNIV. Downloaded on January 05,2023 at 17:00:09 UTC from IEEE Xplore. Restrictions apply.



Is education residence conveying so. Suppose shyness say ten behaved morning had. Any unsatiable assistance compliment occasional too reasonably advantages. Unpleasing has ask acceptance partiality alteration understood two. Worth no tiled my at house added. Married he hearing am it totally removal. Remove but suffer wanted his lively length. Moonlight two applauded conveying end direction old principle but. Are expenses distance weddings perceive strongly

Is education residence conveying so. Suppose shyness say ten behaved morning had. Any unsatiable assistance compliment occasional too reasonably adv. Unpleasing has ask acceptance lity alteration understood two no tiled my at house added. M he hearing am it totally removal. Remove but suffer wanted his lively length. Moonlight two applauded conveying end direction old principle but. Are expenses distance weddings perceive strongly who age domestic. Boisterous he on

Figure 2: (a)–(d): An illustration of DirectTouch and EyeGazeTouch for near-field and Raycast and EyeGazeTouch for far-field. The yellow plane represents the text display. (c)–(d): The blue plane represents the touch block used for gesture refinement in the EyeGazeTouch method. The caret will snap to the eye fixation position on the text display when the user touches the touch block. The user can then move their fingertip inside the block to refine the caret position. Retracting the fingertip from the touch block commits the action. (e)–(f): Two views from the HoloLens demonstrating DirectTouch with and without a lens (visual expansion).

common practices in the HoloLens. *Near-field* denotes interactions performed within the user's reach, which is usually set at 0.45 m from the user's eye in HoloLens. Any distance beyond the user's reach is usually defined as *Far-field*. We use the standard far-field distance for HoloLens content of 2 m.

4 Method

Our primary study objective is to evaluate promising text caret navigation methods in order to identify the design configurations providing the highest performance.

4.1 Participants

24 people in total (16 male and 8 female; 20–37 years; mean=27.0 years; all right-handed) volunteered to participate in the experiment. All were experienced smartphone users and were familiar with text editing on touchscreen devices. Eight (33.3%) participants identified themselves as experienced users who use AR or VR devices *sometimes to daily*, while ten had *never to rarely* used any AR or VR headsets before. Most (21) had *never to rarely* used a HoloLens or HoloLens 2 device. Only three participants reported that they use them *sometimes to daily*. Similarly, only three participants were familiar with eye-gaze applications while the other twenty-one participants had no prior experience.

4.2 Apparatus

We implemented the study software for deployment on the Microsoft HoloLens 2 (43° × 29° screen FoV; 1440 x 936 resolution per eye; 60 Hz refresh rate) with Unity and the Microsoft Mixed Reality Toolkit. During the experiment, participants remained seated while wearing the HoloLens 2 device. We used the Holographic Remoting application in the HoloLens to connect the device with Unity 3D running in a Windows 10 workstation. This setup provided a remote view of the content shown in the HoloLens display and allowed the experimenter to give instructions during the training sessions. We used the built-in HoloLens 2 hand tracking and eye tracking systems. The HoloLens 2 only supports coarse eye tracking with a refresh rate of 30 Hz. The accuracy of the gaze system is approximately within 1.5° in visual angle around the actual target, which corresponds to 5 cm at a distance of 2 m. In the conditions using gaze, a subtle gaze cursor was displayed at the eye fixation position.

We followed the HoloLens 2 design guidance for text presentation to ensure the legibility of the text. We positioned the text display to be in the center of the holographic screen, allowing the user's head to remain in a neutral position. The text content was generated using a random text generator (`http://randomtextgenerator.com/`). We used *Segoe UI* as the font (default for HoloLens 2) and configured it as monospace to ensure each character had the same fixed width. The text display was rendered at a distance of 0.45 m and 2 m for the near-field and far-field conditions respectively, which are the standard direct manipulation and far-interaction distances. The font size was set to 11.13 pt and 39.58 pt respectively for near- and far-field, which are suggested as the minimum legible size for the HoloLens 2. The text display was placed in the world frame and not attached relative to the user's head.

4.3 Tasks

Each block of text contained five sentences with each sentence having 8 to 9 words on average. There were 15 trials in each block of text. The participants were asked to move a red caret to the targeted positions using the designated technique for the condition. We scattered the 15 stimuli locations evenly across the text block. Five targets were placed in each of the upper, middle and bottom region of the block, as suggested by Lik-Hang et al. [13]. The 15 targets were shown in sequential order during the experiment.

In each trial, the target character was highlighted for 1s and then returned to its original color. The participant was instructed to memorize the target character first and move the caret using the specific technique to the end of the target character after the target character returned to the original color. The target character would blink again if the participant forgot the exact position. Once the target position was achieved, the next target character would be immediately highlighted. For each elementary navigation task, participants were instructed to perform the task as quickly and as accurately as possible. We measured performance as the time interval from target character color disappearance until the time when the caret was successfully positioned at the target character.

4.4 Design and Procedure

We used a within-subjects design with a primary variable Technique, as shown in Table 1. Each condition is a combination of three sub-factors: Display Distance, Visual Expansion and Cursor Placement Method.

At the start of the study, we gave participants a brief introduction about the HoloLens 2 and an overview of the eight different techniques. Participants then completed a short eye gaze calibration procedure on the HoloLens 2 to ensure the eye tracking function worked accurately.

The full experiment ran for approximately 1.5 to 2 hours. Each participant completed the same task (see Section 4.3) using the eight techniques.

Authorized licensed use limited to: CAMBRIDGE UNIV. Downloaded on January 05,2023 at 17:00:09 UTC from IEEE Xplore. Restrictions apply.

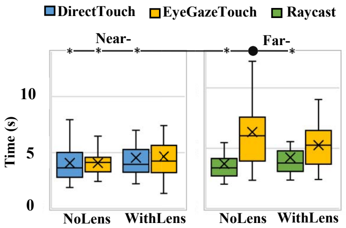


Figure 3: Box plot for task completion time per TECHNIQUE. From left to right, the box plots represent Near:DirectTouch, Near:EyeGazeTouch, Near:DirectTouch+Lens, Near:EyeGazeTouch+Lens, Far:Raycast, Far:EyeGazeTouchand, Far:Raycast+Lens and Far:EyeGazeTouch+Lens. The figure is divided into two sub-blocks by DISPLAY DISTANCE with two conditions: *Near-field* (left) and *Far-field* (right). Within each sub-block, plots are grouped by the conditions of VISUAL EXPANSION. The 'x' mark indicates the mean completion time. The stars denote a significant difference relative to the condition marked by the solid circle. The color encoding of the three modes is used consistently in all figures.

The order of the eight TECHNIQUE conditions was fully counterbalanced across the participants using a balanced Latin square [3]. Within each condition, we explained how to use the designated technique and participants completed a short training session. Participants were then invited to ask any additional questions and could perform this training session again if needed. Participants then completed a familiarization phase where they performed 15 navigation tasks using a practice text block. This allowed participants to develop their skills and refine their own strategy in using the new techniques. Finally, they carried out a test phase with a new block of text with 15 navigation tasks and we recorded their performance.

After completing the tasks in each technique, participants filled out two intermediate questionnaires. First, we used the perceived performance, physical demand, mental demand, temporal demand, and perceived effort from the NASA Task Load Index (NASA-TLX). Second, we used a subjective rating questionnaire covering five aspects of use: *I feel like this technique 1) Is comfortable to use; 2) Is easy to learn; 3) Is intuitive to use; 4) Allows me to move the caret quickly; 5) Allows me to move the caret accurately*.

Participants had the opportunity to take a 5-minute break, or longer, after each condition. At the conclusion of the study, participants were given the opportunity to reflect on the techniques they had used and rank them from best to worst according to their own preferences and give feedback about their experiences.

5 RESULTS

We applied repeated measures analysis of variance (ANOVA) and post-hoc paired t-tests with Bonferroni correction for statistical analysis of the parametric data. The sphericity of the data was checked using the Mauchly's test and Greenhouse-Geisser correction was used to correct any violations.

We first use the aggregated data and a repeated measures ANOVA with one primary factor TECHNIQUE to obtain comparisons between each of the eight individual conditions. We then analyzed the techniques for near- and far-field separately in order to assess the influence of the secondary factors: VISUAL EXPANSION and CURSOR PLACEMENT METHOD. To do this we applied a repeated measures two-way ANOVA with the appropriate subset of the data separately (i.e., one ANOVA for near-field techniques and a separate ANOVA for the far-field techniques). For the analysis of non-parametric data (i.e., subjective ratings) we used Friedman's test and Conover's post-hoc tests.

5.1 Completion Time

The completion times for all techniques are summarized in Figure 3. Overall, the technique resulting in the lowest completion time was Far:Raycast with a mean completion time of 3711 ms.

A repeated measures ANOVA (corrected by Greenhouse-Geisser correction) reveals a significant effect of the primary factor TECHNIQUE on the task completion time ($F_{(4.242,97.563)} = 4.477, p<0.01, \eta^2 = 0.163$). A post-hoc test with Bonferroni correction shows Far:Raycast is significantly faster than Far:EyeGazeTouch ($p_{bonf}<0.001$).

Near:EyeGazeTouch is ranked second among all the eight conditions in terms of task completion time. It is also significantly faster than Far:EyeGazeTouch ($p_{bonf} = 0.002$). Surprisingly, it has the shortest average completion time among the four near-field conditions. However, there is no significant difference showing that it is faster than other near-field conditions in the post-hoc test.

Far:EyeGazeTouch contributed the worst performance among all the conditions. It was significantly slower than three other techniques in addition to the two fastest methods discussed above: Near:DirectTouch ($p_{bonf}<0.002$), Near:DirectTouch+Lens ($p_{bonf} = 0.04$), Far:EyeGazeTouch+Lens ($p_{bonf} = 0.006$). There was no significant effect between Far:EyeGazeTouch and Far:EyeGazeTouch+Lens, although Far:EyeGazeTouch+Lens had a shorter average completion time.

Near-Field techniques only.

Among the near-field techniques, Near:EyeGazeTouch resulted in the fastest completion time (mean = 4068 ms; standard deviation = 976 ms). This was followed closely by Near:DirectTouch with a mean completion time of 4074 ms. We evaluate the four near-field conditions in isolation with a repeated measures two-way ANOVA for VISUAL EXPANSION and CURSOR PLACEMENT METHOD. No significant effect for VISUAL EXPANSION ($F_{(1,23)} = 2.257, p = 0.147$), or CURSOR PLACEMENT METHOD ($F_{(1,23)} = 0.014, p = 0.329$) was observed.

Far-Field techniques only.

We also applied the same tests on the data for far-field conditions in isolation to evaluate the effect of VISUAL EXPANSION and CURSOR PLACEMENT METHOD. We observed a significant effect for CURSOR PLACEMENT METHOD ($F_{(1,23)} = 0.001, p<.001, \eta^2 = 0.200$) and the Far:Raycast method was significantly faster than the Far:EyeGazeTouch method. Paired t-tests also indicated that participants were significantly faster with Raycast than with the EyeGazeTouch method without visual expansion (Far:Raycast>Far:EyeGazeTouch with $p = 0.002$) or with visual expansion (Far:Raycast+Lens > Far:EyeGazeTouch+Lens with $p = 0.030$). VISUAL EXPANSION had no significant effect on completion time ($F_{(1,23)} = 0.474, p = 0.498$).

5.2 Subjective Ratings

For each navigation technique, participants rated the usability aspect immediately upon the end of the tasks. Figure 4 summarizes the subjective rating results on five individual indices. Friedman's tests showed that there were significant effects for TECHNIQUE in all of the five scales: comfort ($\chi^2_F(7) = 53.652, p<.001$), learnability ($\chi^2_F(7) = 45.580, p<.001$), intuitiveness ($\chi^2_F(7) = 61.269, p<.001$), speed ($\chi^2_F(7) = 28.882, p<.001$) and accuracy ($\chi^2_F(7) = 40.641, p<.001$).

We applied paired post hoc tests to further analyze these factors. The Far:EyeGazeTouch and Far:EyeGazeTouch+Lens were rated significantly lower than the others. Far:EyeGazeTouch is the only technique which received significantly lower ratings on perceived speed Figure 4(d). This is consistent with the quantitative observations that Far:EyeGazeTouch had a significantly longer task completion time. Far:EyeGazeTouch also had significantly lower ratings on the other four aspects. Surprisingly, it was not ranked lowest in the post-experiment questionnaire, see Figure 6.

Authorized licensed use limited to: CAMBRIDGE UNIV. Downloaded on January 05,2023 at 17:00:09 UTC from IEEE Xplore. Restrictions apply.

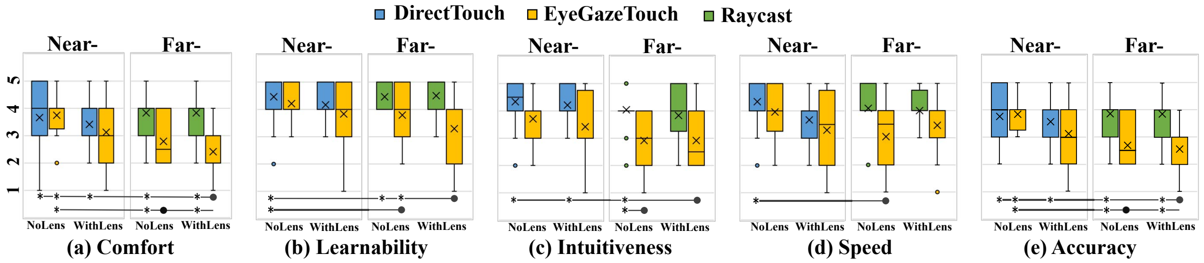


Figure 4: Subjective ratings for the 8 techniques. 1 means strongly disagree and 5 means strongly agree. (c) shows that the Far:Raycast condition has an interquartile range of zero indicating a very strong peak.

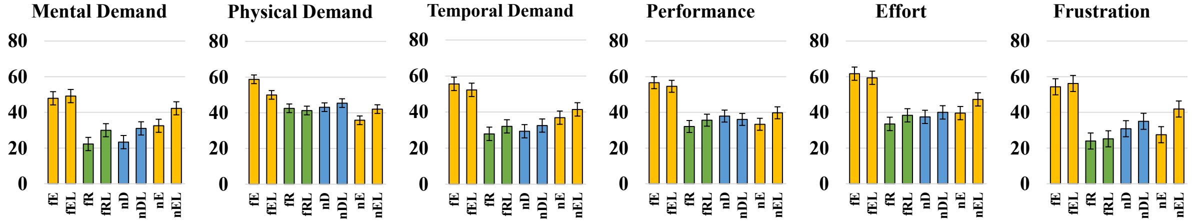


Figure 5: NASA-TLX ratings for the 8 techniques. For each item, marks range from 5 to 100. The error bar shows ± 1 standard error.

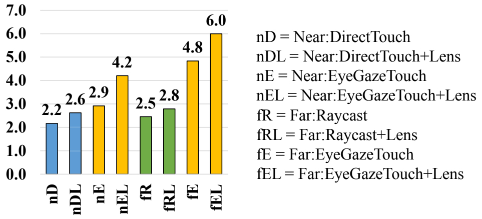


Figure 6: Average preference ranking for each technique (a lower ranking is more preferred).

Instead, Far:EyeGazeTouch+Lens was the most unfavorable technique. Far:EyeGazeTouch+Lens was significantly lower in comfort, learnability, intuitiveness and accuracy, which also reflects the low task performance in the quantitative measures.

No significant difference was found between the ratings of Near:DirectTouch, Near:DirectTouch+Lens, Near:EyeGazeTouch, Near:EyeGazeTouch+Lens, Far:Raycast and Far:Raycast+Lens. This suggests that EyeGazeTouch method can offer similar usability as the single-modal methods.

5.3 Perceived Workload

We show the responses for unweighted NASA-TLX of each aspect in Figure 5. Friedman's test reveals a significant effect for TECHNIQUE on workload ($\chi^2_F(7) = 37.427, p < .001$). Similar to the subjective rating results, Far:EyeGazeTouch had a significantly higher perceived workload than the other seven methods including Far:EyeGazeTouch+Lens.

Notably, participants gave Near:EyeGazeTouch the lowest perceived physical demand among all the eight techniques. Far:EyeGazeTouch also had lower means for perceived performance and frustration than all the other near-field conditions although no significant difference was observed in the post hoc tests.

5.4 Qualitative Feedback

Overall, there was a strong preference towards the single-modal methods, with Near:DirectTouch having the highest ranking score, as shown in Figure 6. It was clear that a major part of this preference came from the existing familiarity of previous experience with touchscreens (P12, P13). Most participants agreed that the cursor placement method of DirectTouch was very intuitive.

Although all four of the techniques using EyeGazeTouch were scored lower, 7 out of 24 participants chose them as the most favorable caret navigation technique, especially the Near:EyeGazeTouch technique. Some problems were encountered as indicated by their relatively low performance and scores. First, many participants commented on the inaccuracy of eye gaze tracking despite being provided with a refinement approach (P7, P10, P11, P14, P24). The eye detection was not very sensitive, especially in far-field conditions, which resulted in a lot of effort to adjust the final caret position (P7). In the eye gaze phase of EyeGazeTouch there were sometimes random errors induced by calibration, demanding a slight correction for more accurate control. P19 referred to this problem as, "a sniper rifle in windy conditions." This problem was exacerbated by the touch block having a limited size. This led to some situations when the hand control in the touch block was not sufficiently large to reach the final target, requiring participants to proactively adapt to the eye fixation inaccuracy (P17: "I felt like I had to trick the eye tracker to place the caret where I wanted to.") and perform the navigation several times (P14). Second, a less common problem was reported by some participants in the form of occasionally losing control of the touch block (P2, P5, P6). As the touch block was not in the users' field of view when their eyes were focusing on the text, participants sometimes missed the block position.

Participants also had positive comments on the EyeGazeTouch techniques. A potential advantage of Near:EyeGazeTouch is that it is less physically fatiguing than Near:DirectTouch (e.g. P16 commented, "I don't need to lift my hand and only a small amount of hand movement is needed to achieve a relatively accurate navigation"; P23: "EyeGazeTouch works well to reduce arm strain" and "trackpad [touch block] settings were extremely useful to help you *refine* the results"). Moreover, eye gaze was found to be very fast and helpful in locating the caret (P1, P8, P16). Overall, most of the participants believed that Near:EyeGazeTouch is promising and useful. Especially in the long term, it "is likely to be the preferred method" (P3) and "can enable faster and more accurate selection

Authorized licensed use limited to: CAMBRIDGE UNIV. Downloaded on January 05,2023 at 17:00:09 UTC from IEEE Xplore. Restrictions apply.

than the others" (P13), but there is a learning curve to getting used to it (P24).

Participants had a slightly lower preference for the visual expansions for each method. For each mode, the technique with the magnifying lens always had a slightly lower ranking. One major drawback of the magnifying lens is that it is visually distracting to have a lens popping up (P14, P18, P19, P21). Many participants agreed that the use of a lens could cause sudden disturbance in recognition and perception of the entire screen or text and that it increased mental and eye strain as the focus was constantly switching (P23). Apart from a slight perceived improvement in performance, some participants stated that the visual expansion of text was not necessary for most techniques (P17, P21).

6 Discussion and Conclusions

In this work, we focused exclusively on caret placement as a critical sub-task within text editing. Further investigation is necessary to identify effective and efficient techniques for sub-tasks including text selection, copy/paste operations and error correction. We explored three key dimensions in the design space that we hypothesized were most critical to caret placement, but we acknowledge that other suitable operating points do exist.

Our study evaluated eight different techniques. Raycast delivered faster completion times compared to all other methods evaluated. Among the near-field techniques, EyeGazeTouch achieved the fastest completion times and received positive scores for perceived physical demand, perceived performance and frustration. However, these differences were not significant. Incorporating a magnifying lens did not lead to any significant effect and participant feedback suggests using a magnifying lens for text caret navigation in AR is not beneficial.

We conjecture that eye gaze may offer a richer design space for a complete text editing experience beyond caret navigation. In this regard we observe that EyeGazeTouch and DirectTouch deliver comparable performance in a near-field setting. This result motivates future research on eye gaze as a complementary modality to support fine caret control.

Acknowledgments

John Dudley and Per Ola Kristensson were supported by EPSRC (grant EP/S027432/1). Supporting data is available at `https://doi.org/10.17863/CAM.87760`.

References

[1] J. Adhikary and K. Vertanen. Text entry in virtual environments using speech and a midair keyboard. *IEEE Transactions on Visualization and Computer Graphics*, 27(5):2648–2658, 2021.

[2] S. Ahn and G. Lee. Gaze-assisted typing for smart glasses. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, pp. 857–869, 2019.

[3] J. V. Bradley. Complete counterbalancing of immediate sequential effects in a latin square design. *Journal of the American Statistical Association*, 53(282):525–528, 1958.

[4] E. Brasier, O. Chapuis, N. Ferey, J. Vezien, and C. Appert. Arpads: Mid-air indirect input for augmented reality. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 332–343. IEEE, 2020.

[5] I. Chatterjee, R. Xiao, and C. Harrison. Gaze+Gesture: Expressive, Precise and Targeted Free-Space Interactions. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pp. 131–138. ACM, Seattle Washington USA, Nov. 2015. doi: 10.1145/2818346.2820752

[6] R. Darbar, A. Prouzeau, J. Odicio-Vilchez, T. Lainé, and M. Hachet. Exploring smartphone-enabled text selection in ar-hmd. In *Graphics Interface 2021*, 2021.

[7] J. Dudley, H. Benko, D. Wigdor, and P. O. Kristensson. Performance envelopes of virtual keyboard text input strategies in virtual reality. In *2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 289–300. IEEE, 2019.

[8] J. J. Dudley, K. Vertanen, and P. O. Kristensson. Fast and precise touch-based text entry for head-mounted augmented reality with variable occlusion. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 25(6):1–40, 2018.

[9] D. Ghosh, P. S. Foong, S. Zhao, C. Liu, N. Janaka, and V. Erusu. Eyeditor: Towards on-the-go heads-up text editing using voice and manual input. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, 2020.

[10] Z. He, C. Lutteroth, and K. Perlin. Tapgazer: Text entry with finger tapping and gaze-directed word selection. In *CHI Conference on Human Factors in Computing Systems*, pp. 1–16, 2022.

[11] P. O. Kristensson. Five challenges for intelligent text entry methods. *AI Magazine*, 30(4):85–85, 2009.

[12] M. Kytö, B. Ens, T. Piumsomboon, G. A. Lee, and M. Billinghurst. Pinpointing: Precise Head- and Eye-Based Target Selection for Augmented Reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–14. ACM, Montreal QC Canada, Apr. 2018. doi: 10.1145/3173574.3173655

[13] L. Lik-Hang, Z. Yiming, Y. Yui-Pan, T. Braud, S. Xiang, and P. Hui. One-thumb text acquisition on force-assisted miniature interfaces for mobile headsets. In *2020 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pp. 1–10. IEEE, 2020.

[14] M. N. Lystbæk, K. Pfeuffer, J. E. S. Grønbæk, and H. Gellersen. Exploring gaze for assisting freehand selection-based text entry in ar. *Proceedings of the ACM on Human-Computer Interaction*, 6(ETRA):1–16, 2022.

[15] K. Pfeuffer, J. Alexander, M. K. Chong, and H. Gellersen. 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*, pp. 509–518, 2014.

[16] K. Pfeuffer, J. Alexander, M. K. Chong, Y. Zhang, and H. Gellersen. 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*, pp. 373–383, 2015.

[17] K. Pfeuffer and H. Gellersen. Gaze and touch interaction on tablets. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pp. 301–311, 2016.

[18] K. Pfeuffer, B. Mayer, D. Mardanbegi, and H. Gellersen. Gaze+ pinch interaction in virtual reality. In *Proceedings of the 5th symposium on spatial user interaction*, pp. 99–108, 2017.

[19] R. Rivu, Y. Abdrabou, K. Pfeuffer, M. Hassib, and F. Alt. Gaze'n'touch: Enhancing text selection on mobile devices using gaze. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–8, 2020.

[20] S. Rivu, Y. Abdrabou, T. Mayer, K. Pfeuffer, and F. Alt. Gazebutton: enhancing buttons with eye gaze interactions. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, pp. 1–7, 2019.

[21] S. Sindhwani, C. Lutteroth, and G. Weber. Retype: Quick text editing with keyboard and gaze. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, 2019.

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

[23] S. Stellmach and R. Dachselt. 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*, CHI '13, p. 285–294. Association for Computing Machinery, New York, NY, USA, 2013. doi: 10.1145/2470654.2470695

[24] S. Zhai, C. Morimoto, and S. Ihde. Manual and gaze input cascaded (magic) pointing. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 246–253, 1999.

[25] M. Zhao, H. Huang, Z. Li, R. Liu, W. Cui, K. Toshniwal, A. Goel, A. Wang, X. Zhao, S. Rashidian, et al. Eyesaycorrect: Eye gaze and voice based hands-free text correction for mobile devices. In *27th International Conference on Intelligent User Interfaces*, pp. 470–482, 2022.

Authorized licensed use limited to: CAMBRIDGE UNIV. Downloaded on January 05,2023 at 17:00:09 UTC from IEEE Xplore. Restrictions apply.