This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2023.3309316

Text Entry Performance and Situation Awareness of a Joint Optical See-Through Head-Mounted Display and Smartphone System

Jens Grubert, *Member, IEEE,* Lukas Witzani, Alexander Otte, Travis Gesslein, Matthias Kranz, and Per Ola Kristensson

Abstract—Optical see-through head-mounted displays (OST HMDs) are a popular output medium for mobile Augmented Reality (AR) applications. To date, they lack efficient text entry techniques. Smartphones are a major text entry medium in mobile contexts but attentional demands can contribute to accidents while typing on the go. Mobile multi-display ecologies, such as combined OST HMD-smartphone systems, promise performance, and situation awareness benefits over single-device use. We study the joint performance of text entry on mobile phones with text output on optical see-through head-mounted displays. A series of five experiments with a total of 86 participants indicate that, as of today, the challenges in such a joint interactive system outweigh the potential benefits.

Index Terms—text entry, augmented reality, multi-display, optical see-through, head-mounted display, mobile, cross-device

1 Introduction

Optical see-through head-mounted displays (OST HMDs) open up a rich design space for human-computer interaction. OST HMDs can redesign workflows and streamline user experiences in such disparate areas as construction engineering, smart factories and triage systems. In addition, OST HMDs can potentially become the smartphone platform of the future, allowing mobile user experiences to be situated fluidly within physical reality.

However, OST HMDs also pose design challenges. One such challenge is efficient text entry. Regular text entry using only an OST HMD on a virtual keyboard is difficult as the user's typing activities have to be inferred from typically front-mounted sensors. The task of inferring users' typing in such a setup is challenging as recognition accuracy is limited by sensor range and sampling rate, and the accuracy of the inferred hand skeleton. It is also difficult for users to type in thin air as the typing poses are fatiguing, lack natural physical support, and do not provide any haptic feedback. Even smartphone typing, a widely practiced activity, provides limited haptic feedback, only informing the user of the fact that the user's finger made contact with the capacitive screen. Such feedback is not present in pure OST HMD typing. In addition, in safety-critical environments, such as construction engineering, smart factories, and triage systems, it is vital that the text entry system is certain and not subject to the noise and recognition errors that are inherent in mid-air typing solutions.

As a result of these concerns, one commercial OST HMD manufacturer, Microsoft, opted to redesign the typing process for their HoloLens product such that it involved a specific two-step target acquisition process: 1) acquire a key target on the keyboard by moving a head-tracked cursor to the target; and 2) confirm using a specific click gesture performed with the thumb and index finger. This design provides accuracy for a skilled user at the expense of speed and the need for user learning.

As OST HMDs are now beginning to slowly emerge as mainstream user interface technologies, an interesting orthogonal solution space is a joint OST HMD-smartphone system for typing (e.g., [1]) that allows users to type on a smartphone while observing the text output on the OST HMD. Such a joint system can potentially provide additional advantages for the user. First, it can provide users with additional privacy as text output is completely hidden from onlookers. Second, it can provide users with better situation awareness[1] and reduce smartphone-induced pedestrian accidents [3], [4], [5]. Third, typing on a smartphone is a relatively fast and well-practiced skill among the existing user population with low fatigue. Further, in contrast to alternative mobile text entry devices, such as chording keyboards [6], phones with physical keys, or even body-mounted desktop keyboards [7], the smartphone is a ubiquitous interaction device readily available to hundreds of millions of users

In general, a joint OST HMD-smartphone system opens up possibilities for redistributing input and output spaces in such a way that the advantages of each subsystem are maximized. However, such joint OST HMD-smartphone system performance has not been robustly investigated. Grubert et al. [1] proposed combining HMDs with smartphones but did not report on performance. To the best of our knowledge, we are the first to quantify the joint performance of text input on mobile phones with text output on OST HMDs. Specifically, we study the effects of non-spatially registered text output (i.e., a heads-up display (HUD) view) and spatially registered text output (AR view) compared to the standard text entry user interface on mobile phones (keyboard and text

- *Jens Grubert is and Alexander Otte and Travis Gesslein were with Coburg University of Applied Sciences and Arts, Germany. E-mail: jens.grubert@hs-coburg.de*
- *Lukas Witzani was and Matthias Kranz is with University of Passau, Germany.*
- *Per Ola Kristensson is with Cambridge University.*

Manuscript received December, 2022.

1. While there are multiple definitions of situation awareness, we follow Endsley's model [2] but focus on the initial perception of elements in a given situation.

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2023.3309316

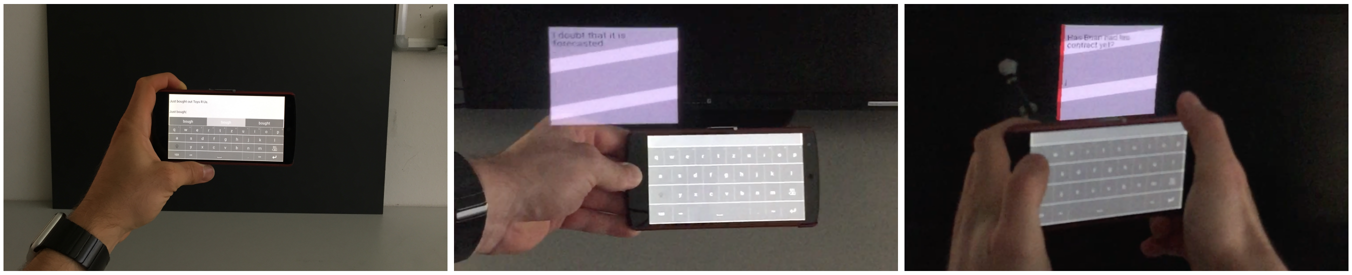

Fig. 1. First-person views of the conditions in Experiment 2 on the effects of combined typing on a joint Optical See-Through Head-Mounted Display (OST HMD) smartphone system. Left: typing on standard keyboard (condition BASELINE). Center: typing on full-screen keyboard displaying the text at a fixed position in the user's field of view (condition HUD). Right: typing on a full-screen keyboard displaying the text spatially registered above the smartphone in an Augmented Reality view (condition AR). The visual artifacts (blur, aberrations, white stripes) are due to capturing the OST HMD screen with a camera.

box on the same capacitive screen). Our central contribution is a series of five investigations that provide a nuanced understanding of the opportunities and challenges inherent with a joint OST HMD-smartphone system.

Experiment 1 ($n = 16$) served to validate the fact that larger keys on a smartphone lead to higher entry rates. However, Experiment 2 ($n = 18$) demonstrated that this potential benefit did not materialize for a joint OST HMD-smartphone system for a text entry task in a static setting with participants seated. On the contrary, text entry performance was significantly degraded.

The subsequent three experiments investigated if a positive effect of a joint OST HMD-smartphone system on situation awareness may be observed when using the system in a mobile context. Experiments 3 ($n = 24$) and 4 ($n = 14$) revealed that an OST HMD-smartphone system was not superior compared to a smartphone system in terms of text entry performance or situation awareness. Experiment 5 ($n = 14$) further showed that solely wearing an HMD, even though it is turned off, can already negatively impact users' situation awareness.

Taken together, our investigations show that while there are theoretical advantages of joint OST HMD-smartphone systems over using exclusively smartphones for text entry, those advantages do not materialize in current generation systems.

2 Related Work

Text entry is a fundamental user interface task and as a consequence, it is unsurprising that there is a large body of prior research in the area, e.g., [8], [9]. Mobile text entry has been particularly intensively studied and a standard smartphone QWERTY/QWERTZ/AZERTY keyboard might be suitable since it is portable and has a relatively high text entry rate coupled with an acceptably low error rate [10], [11].

Many improvements have also been considered for smartphone typing, such as more robust auto-correct algorithms that allow users to type entire sentences before decoding [12] or various combinations of word units [13], adding additional support for auto-correcting that incorporates information about the user's gait when walking [14], and systems that allow a user to self-regulate their certainty of their key presses by pressure regulation [15].

2.1 Text Entry in AR and VR

Prior work on providing text entry for virtual and augmented reality can be split into five categories: 1) special hardware, such as gloves, demanding considerable learning effort and limited performance e.g., [16], [17], [18]; 2) typing in thin air or on a user's palm e.g., [19], [20], [21]; 3) utilizing standard physical keyboards [22], [23], [24]. It was also investigated how to enable interaction in mobile environments using physical keyboards [25], [26], [27]; 4) using capacitive touchscreens or controllers [28], [29], [30]; and 5) head and gaze pointing-assisted text entry, such as the default text entry method in Microsoft HoloLens and the head-pointed supported gesture-keyboard [31], [32]. A recent survey on text entry in VR is presented by Dube et al. [33].

2.2 Situation Awareness and Smartphone Use

Prior research has also investigated situation awareness around smartphone use. Oulasvirta et al. [34] discovered that users tended to interact in quick bursts when walking in order to maintain awareness of their surroundings. Several situation awareness support systems have been proposed for mobile users, e.g., [35], [36], [37], [38]. Wen et al. [36] and Liu et al. [38] focus on detecting ground obstacles using ultrasound and infrared sensors, Hincapié-Ramos et al. [35] use a depth camera and Foerster et al. [37] use the smartphone back-camera. Van dam et al. [39] indicated negative effects of drivers' passive mobile phone usage on situation awareness while driving. Woodward and Ruiz [40] conducted a literature review on situation awareness in AR and presented recommendations on the visual design of AR user interfaces for increasing situation awareness. Complementary to prior work, our work focuses on investigating text entry with a smartphone while wearing an optical see-through head-mounted display.

2.3 Text Legibility and Readability on OST HMDs

Further, previous work had investigated text readability on OST HMDs. Gabbard et al. [41] investigated text readability using different text drawing styles, background textures and lighting, and later extended their investigations to industrial environments [42]. Orlosky et al. [43] proposed a dynamic text alignment system that actively maintains moving text in the user's field of view, partly inspired by early view management techniques by Blaine et al. [44]. Lucero et al. [45] investigated the presentation of icons while walking, but did not focus on text. Rzayev et al. [46] investigated different text presentation modes while reading and walking, but did not focus on text entry. Erickson et al. [47] indicated that "dark mode" user interfaces could have benefits for presenting UI elements in OST HMDs and that user preferences depend on the lighting conditions in the physical environment. Pavanoto et al. [48] identified the need to increase the font size on virtual monitors

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2023.3309316

in OST HMDs to achieve a comparable reading experience as with physical desktop monitors. Rau et al. [49] also indicated that reading on an OST HMD is slower compared to a traditional LCD display. Orlosky et al. [43] proposed a text management system that dynamically adapts the text location in order to maximize readability.

2.4 Context and Focus Switching in OST HMDs

Related work has investigated the effects of context and focus switching in OST HMDs. Huckauf et al. investigated the costs of context switching between a monocular OST HMD and a CRT placed at the same focus distance [50]. Gabbard et al. [51], [52], [53], [54] examined context switching and differing focal distances between a panel display and a monocular OST HMD (or a haploscope) using text-based visual search task. Both context switching and focal distance switching resulted in significantly reduced performance. Winterbottom et al. [55] studied multi-focal AR displays and empirically determined suitable focus distances in the context of a flight simulator. Eiberger et al. studied the impact of focus switching between an OST HMD and a handheld display [56] and found that both task completion time and error rate increased significantly when solving a visual search task jointly over the OST HMD and smartphone, compared to the OST HMD alone. Similarly, Drouot et al. [57] studied effects of context and focal distance switching using an OST HMD. Their results confirmed a negative impact for focal distance switching but not for context switching. Gabbard et al. also highlighted the need for multi-focal AR displays [58] with several research protoypes beginning to emerge [59], [60], [61]. For current single focus HMDs, Oshima et al. [62] and Cook et al. [63] proposed and evaluated a system for adaptive sharpening HMD display content. Imamov et al. mimicked context switching between a real world-task and an information display in an immersive VR display and quantified the costs of context switching for different distances [64]. In a recent survey, Koulieris et al. [65] presented an overview of see-through display technologies, such as varifocal or multiplane displays, that can aid with focus switching.

2.5 Joint OST HMD-Smartphone Interaction

To the best of our knowledge, no prior research has investigated the text entry performance of joint OST HMD-smartphone system in depth. However, joint systems have been studied before (e.g., [1], [66], [67], [68], [69], [70], [71]), not focusing on text entry. The closest prior works are probably by Wolf et al. [72], which studied the performance of pointing, crossing and steering tasks in a joint OST HMD-smartwatch system but importantly did not investigate text entry and Grubert et al. [1] who proposed a joint OST HMD-smartphone text entry system but did not report on performance. Recently, Darbar et al. compared four eyes-free text selection techniques in a joint OST HMD-smartphone system [73]. Their result suggest using the smartphone as basic touchpad outperformed alternative techniques (discrete touch, in-air pointing, raycasting). Besides the use of OST HMDs, research has also explored the joint interaction between smartphones and immersive VR HMDS [74], [75].

3 Study Overview

In this paper, we investigate the performance and situation awareness aspects of a joint OST HMD-smartphone text entry system.

To this end, Experiment 1 ($n = 16$) quantifies the effects of keyboard size for text entry on a smartphone. The study demonstrates that a full-screen keyboard in landscape mode provide significantly higher text entry rates—around 22 words per minute, an increase of around 15% compared to a keyboard with standard size.

Experiment 2 ($n = 18$) then investigate if this performance gain can be transferred to a joint OST HMD-smartphone system. The main findings of Experiment 2 are that when using an HMD for text output in text entry, performance is significantly reduced compared to a standard smartphone baseline with approximately 10% lower text entry rates for the HUD condition and approximately 21% lower text entry rates for the AR condition, in addition to higher error rates. HMD-based text entry was also not preferred by participants.

Taken together experiments 1 and 2 indicate that the potential for faster text entry due to a larger keyboard are diminished by the costs of joint visual information processing across two displays.

Experiments 3 ($n = 24$), 4 ($n = 14$), and 5 ($n = 14$) investigate text entry performance in a mobile context. Experiment 3 utilizes a physical obstacle course and Experiment 4 a virtual one in order to focus on potential obstacle collisions. Experiment 3 shows that participants attend the smartphone significantly less when wearing an HMD (41% of the time) compared to a smartphone-only condition (75%) or a dual condition (71%), in which participants could see the text both on the smartphone and the HMD. However, this focus on the HMD induced significantly higher costs in terms of situation awareness (30% higher cognitive demand) and lower text entry rate (13%) compared to smartphone-only use.

Experiment 4 indicates an equivalence in object collisions in a simulated walking scenario between a smartphone-only and a joint OST HMD-smartphone not indicating an advantage of wearing an HMD for detecting obstacles while entering text. In addition, wearing an HMD also resulted in higher (but still mild) simulator sickness scores. Experiment 5 replicates the smartphone-only condition and studies the effect of wearing an HMD—without presenting any information in it. The experiment indicates equivalence for text entry performance, overall demand, and simulator sickness. However, it also indicates significantly lower attentional supply, in terms of arousal, spare mental capacity, concentration, and division of attention, when wearing an HMD. Together, experiments 3, 4, and 5 indicate that, in a mobile condition, combined text entry across a smartphone and an OST HMD does not offer substantial benefits over text entry on a smartphone alone. Unless otherwise indicated, statistical significance tests were carried out using general linear model repeated measures analysis of variance with Holm-Bonferroni adjustments for multiple comparisons at an initial significance level $\alpha = 0.05$. We indicate effect sizes whenever feasible (η_p^2). We verified that the assumptions underpinning the tests, such as normality and sphericity, were met. To aid replication and further analysis we make the anonymized experimental data available under https://gitlab.com/mixedrealitylab/ar-text-entry.

4 Experiment 1: Keyboard Size

One potential benefit of a joint OST HMD-smartphone system is the extended screen space available through the HMD. Prior work [1] suggested separating text input (on the smartphone) from output (on the HMD) but did not validate this design. Hence, as a first step, we investigated the effects of different keyboard sizes on a smartphone. While prior work studied the effect of keyboard

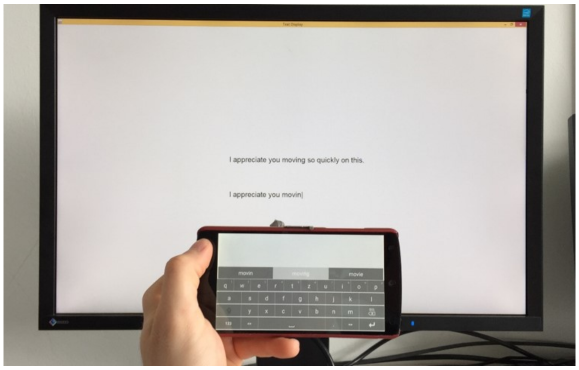

Fig. 2. The setup used for Experiment 1 on the effect of keyboard size (condition StandardKeyboard is depicted).

TABLE 1
Descriptive statistics and hypothesis test statistics for error rate and workload for Experiment 1. Grey rows indicate significant differences. CER: Character Error Rate, SK: StandardKeyboard, FK: FullscreenKeyboard, MD: Mental Demand, PD: Physical Demand, TD: Temporal Demand, P: Performance, E: Effort, F: Frustration, O: Overall Demand.

	mean (sd) SK	mean (sd) FK	Z	p	r
CER	0.008 (0.01)	0.007 (0.1)	−0.26	0.80	−0.05
MD	52.81 (25.43)	47.19 (28.75)	1.90	0.057	0.34
PD	48.75 (30.19)	36.56 (25.8)	2.30	0.03	0.41
TD	47.50 (15.71)	49.38 (20.89)	−0.08	0.75	−0.01
P	57.81 (26.20)	44.69 (29.07)	1.953	0.04	0.35
E	60.94 (26.28)	51.56 (24.95)	2.50	0.013	0.44
F	59.06 (22.53)	34.06 (22.53)	3.40	> 0.001	0.60
O	54.48 (21.32)	43.01 (21.49)	2.95	0.003	0.52

(key) size on typing performance, e.g., [76], [77], we saw it as crucial to quantify the effects at the particular operating point of a smartphone that would be used in a joint HMD-smartphone system. Specifically, based on prior work [1], we studied two-handed text entry while the smartphone was held in landscape mode. To this end, we compared a standard-sized onscreen keyboard where a proportion of the space of the display is reserved for a text box, vs. a full-screen onscreen keyboard occupying the entire display. In this experiment, no HMD was used to mitigate the potential effects of focus distance switching, which would occur in an HMD with a fixed focal distance that strongly diverges from the focus plane of a handheld display [56].

4.1 Method

The experiment was a within-subjects design with one independent variable: KeyboardSize. The independent variable KeyboardSize had two levels: typing on a normal-sized keyboard (StandardKeyboard), see Figure 2 and Figure 1, left, on a smartphone and typing on a larger full-screen keyboard (FullscreenKeyboard) on a smartphone, see Figure 1, center and right.

4.2 Participants

For this study, we recruited 17 participants from a university campus. One participant had to be excluded due to not being able to write on a smartphone. All other participants were familiar with smartphone keyboard typing. None had participated in the previous experiments. From the 16 remaining participants (mean age 26.4 years, sd = 7.5, mean height 173.3 cm, sd = 10.4, 6 male, 10 female), Five indicated to write between 0.1-1 h daily using their smartphone, 7 to write between 1-2 h and 2 to write 2-4 h per day. Ten participants indicated having visual restrictions which were corrected by contact lenses or glasses during the experiment.

4.3 Apparatus and Materials

Participants were shown stimulus phrases randomly drawn from a mobile email phrase set [78] on an external monitor (see Figure 2). To ensure comparability between normal and full-screen keyboard, the stimulus text as well as the entered text by the user was shown on an external PC monitor and not on the smartphone directly. A LG Nexus 5 was used as the smartphone.

The software keyboard was implemented in two different sizes. The dimensions of the half-screen-sized keyboard were ($w \times h$) 109 $\times$ 31 mm and the dimensions of the keys were 11.0 $\times$ 7.6 mm, see Figure 1, left. The dimensions of the full-screen keyboard were ($w \times h$) 109 $\times$ 53 mm and the dimensions of the keys were 11 $\times$ 13 mm, see Figure 1, center.

The layout of both keyboards were identical. Arrow buttons were added to be able to move the cursor to a different text position during writing. A view showing three suggestions for the currently composed word was shown above the keyboard. The leftmost suggestion showed the currently typed word by the user while the middle and rightmost suggestions showed the most likely and second most likely word respectively. All suggestions could be selected by the user. When a suggestion was selected the keyboard switched the currently composed word into the selected suggestion and moved the cursor to the end of the word. Both keyboards used auto-correction. In the case the user entered a word separator (a space or punctuation character), the keyboard automatically replaced the currently composed word with the most likely word shown as the top middle suggestion. If the user explicitly selected the leftmost suggestion, the auto-correction would not replace the word.

4.4 Procedure

Each participant filled out a demographic questionnaire, was then shown a five-minute video containing information about the study, the participant's task as well as explanations of how the keyboard works. The order of the conditions was counterbalanced across all participants. In either condition, participants were shown a series of stimulus sentences. For an individual stimulus sentence, participants were asked to enter it as quickly and as accurately as possible. Participants were allowed to use the backspace key to correct errors. Participants typed stimulus sentences for a total of 15 minutes in each condition with a short 30 second break after every 5 minutes of typing. The conditions were separated by a 5-minute break in which participants filled out the NASA TLX questionnaire [79]. A final questionnaire about the participant's self-assessment was filled out after the second condition. The experiment was carried out in a single 60-minute session structured as a 15-minute introduction and briefing phase, a 40-minute testing phase (15 minutes per condition + five-minute breaks including questionnaires), and five minutes for final questionnaires, interview, and debriefing.

4.5 Results

Text Entry Rate and Error Rate: Entry rate was measured in words-per-minute (wpm), with a word defined as five consecutive characters, including spaces. The time frame used for calculating the entry rate for each sentence started at appearing of the stimulus text and ended when hitting the enter button.

For StandardKeyboard, the mean entry rate was 19.51 wpm (sd = 4.4). For FullscreenKeyboard, the mean entry rate was 22.37 wpm (sd = 4.45). A paired two-tailed t-test showed that the entry rate difference between StandardKeyboard and FullscreenKeyboard was statistically significant ($t(15) = 3.5862$, $p = 0.0027$) with an effect size of $Cohen's\ d_z = 0.93$ (calculated after [80]).

Error rate was measured as character error rate (CER). CER is the minimum number of character-level insertion, deletion and substitution operations required to transform the response text into the stimulus text, divided by the number of characters in the stimulus text. Results for character error rate can be seen in Table 1. No significant statistical difference was detected. In other words, FullscreenKeyboard resulted in significant higher text entry rates compared to StandardKeyboard.

Workload: Scores for the workload of both conditions measured by the unweighted NASA TLX questionnaire are depicted in Table 1. Wilcoxon signed rank tests did reveal significant differences for physical demand, performance, effort, frustration, and overall demand, but not for mental demand and temporal demand. In other words, FullscreenKeyboard resulted in significant lower workload compared to StandardKeyboard.

Preferences and Open Comments: Participants were asked to rank the conditions from least preferred to most preferred. Ten out of the 16 participants preferred the bigger keyboard. A binomial test did not indicate a significant difference ($p = 0.454$). Ten participants felt increased typing speed with the full-screen keyboard. 15 participants estimated to mistype more often when using the small keyboard. Qualitative feedback also revealed that participants found the big keyboard to be easier to use and its buttons easier to hit. They felt to have more precision and less times correcting mistakes. Reasons for preferring the small keyboard were either the habit to use this size for a keyboard or the smaller distance the finger had to travel while typing.

5 Experiment 2: Effects of Combined Typing on a joint OST HMD-Smartphone System

Experiment 2 investigated the effects of combined typing on a smartphone and an OST HMD. We compared QWERTZ keyboard text entry on a mobile phone (Figure 1, left) with text entry on a joint OST HMD-smartphone system in two configurations. In the first OST HMD-smartphone configuration, as proposed by Grubert et al. [1], we used a full-screen keyboard on the smartphone for text entry combined with a spatially registered AR view on an OST HMD (Figure 1, right). In the second OST HMD-smartphone configuration, we investigated a non-spatially registered view (HUD), see Figure 1, center. Hence, text input was always conducted on the smartphone, but text output appeared either on the smartphone or on the OST-HMD (spatially registered or fixed). In addition, to Experiment 1, we also added NASA TLX [79], in order to receive workload indications, the Simulator Sickness Questionnaire (SSQ) [81], in order to investigate if the spatially registered text might lead to higher simulator sickness, and the Flow-Short-Scale questionnaire (FSS) [82], in order to investigate if the HMD condition would lead to a higher flow.

5.1 Method

The experiment was a within-subjects design with one independent variable Interface. The independent variable Interface had three levels: Baseline, HUD and AR. In the Baseline condition, text entry was done using the halfscreen-sized mobile keyboard with the text shown on the smartphone itself. In the HUD condition, text entry was done using the full-screen keyboard and text output was on the OST HMD with a HUD-like display, i.e. with screen-aligned text that was not spatially registered. Third, in the AR condition, text entry was also done using the full-screen keyboard, but text output was spatially registered to the smartphone as proposed by Grubert et al. [1]. In this mode, the user sees the display through the glasses as a virtual extension to the smartphone screen, i.e. only if the head points toward the smartphone. The order of the starting condition was balanced across participants. The task was to write as quickly and as accurately as possible.

5.2 Participants

For this study, we recruited 18 participants (mean age 22.9 years, sd = 1.54, mean height 172.7 cm, sd = 7.79, 5 male, 13 female) from a university campus. Eleven participants indicated having visual restrictions: ten used contact lenses to correct these and one participant used her glasses during the experiment. Interpupillary distance was measured with a ruler (mean = 63.06 cm, sd = 3.24) to adjust the virtual camera positions for the AR condition. All participants were familiar with smartphone keyboard typing. Fourteen participants never used an HMD before, two once, and two rarely.

5.3 Apparatus and Materials

Stimulus sentences were drawn from a mobile email phrase set [78]. Participants were shown stimulus phrases randomly drawn from the set. An LG Nexus 5 was used as the smartphone. We used the same software keyboard as in Experiment 1.

An Epson Moverio BT-300 (23°diagonal field of view, 1280 × 720 pixel resolution) was used as the OST HMD. Participants were sitting on a chair and could rest their arms on a table if they wanted to. The font size was set empirically to approximately 0.9°on the HMD, such that the text was still legible (and matched on the smartphone accordingly). An Optitrack Flex 13 outside-in tracking system was used for spatial tracking of smartphone and HMD for spatial tracking in the AR condition. For this purpose, the smartphone and the OST HMD were equipped with retroreflective markers. We used a fixed optical see-through calibration for all users (Single Point Active Alignment, SPAAM [83]) but ensured that all participants had a correct spatially registered view in the AR condition.

5.4 Procedure

Each participant filled out a demographic questionnaire. Then, each participant was shown a five minute video containing information about the study, the study task, as well as explanations on how the keyboard worked. The interpupilary distance was measured to ensure correct stereo rendering for the AR condition. Participants typed stimulus sentences for a total of 15 minutes in each condition with a short 30 second break after every 5 minutes of typing. In

TABLE 2
Descriptive statistics and hypothesis test statistics for text entry rate and error rate for Experiment 2. Grey rows indicate significant differences. B: BASELINE, WPM: Words Per Minute. CER: Character Error Rate.

	mean (sd) B	mean (sd) HUD	mean (sd) AR	$F_{2,16}$	p	η_p^2
WPM	22.73 (4.02)	20.57 (4.65)	17.95 (4.68)	12.68	0.001	0.61
CER	0.007 (0.006)	0.021 (0.016)	0.097 (0.07)	9.30	0.002	0.54

TABLE 3
Descriptive statistics and hypothesis test statistics for workload, simulator sickness, and flow for Experiment 2. Grey rows indicate significant differences. B: BASELINE, MD: Mental Demand, PD: Physical Demand, TD: Temporal Demand, P: Performance, E: Effort, F: Frustration, O: Overall Demand. NA: Nausea, OM: Oculo-Motor, DI: Disorientation, TS: Total Severity, FF: Flow, FA: Anxiety subscale of FSS, FC: Challenge subscale of FSS.

	mean (sd) B	mean (sd) HUD	mean (sd) AR	$\chi^2(2)$	p
MD	35.28 (20.25)	42.5 (20.88)	58.61 (21.95)	24.80	< 0.001
PD	27.50 (19.95)	33.61 (20.42)	49.72 (27.63)	24.49	< 0.001
TD	48.33 (23.01)	43.89 (17.87)	48.06 (23.34)	0.49	0.79
P	35.56 (13.60)	41.11 (15.10)	51.11 (22.12)	9.94	0.007
E	44.17 (21.71)	50.83 (17.34)	63.61 (19.46)	8.85	0.012
F	36.94 (16.37)	41.39 (19.61)	60.83 (20.45)	28.90	< 0.001
O	37.96 (13.66)	42.22 (13.66)	55.32 (16.20)	23.01	< 0.001
NA	44.3 (44.4)	47.4 (37.9)	71.5 (50.8)	12.03	0.002
OM	21.5 (31.8)	21.9 (28.5)	28.6 (33.4)	8.33	0.016
DI	43.3 (28.3)	46.4 (43.0)	82.7 (73.6)	10.34	0.006
TS	44.3 (44.4)	47.4 (37.9)	71.5 (50.8)	9.77	0.008
FF	4.56 (1.14)	4.54 (1.05)	3.99 (1.13)	5.51	0.063
FA	3.13 (1.65)	2.87 (1.53)	2.94 (1.60)	2.51	0.285
FC	3.50 (0.86)	3.67 (0.77)	4.61 (1.15)	16.69	< 0.001

the 5-minute break between conditions, participants filled out the unweighted NASA TLX [79], the Simulator Sickness Questionnaire (SSQ) [81] and the Flow-Short-Scale questionnaire (FSS) [82]. Participants also filled out a final self-assessment questionnaire after the last condition, followed by a semi-structured interview. The conditions were balanced across participants. In either condition, participants were shown a series of stimulus sentences and asked to type them as quickly and as accurately as possible. Participants were allowed to use the backspace key to correct errors. The experiment was carried out in a single 85-minute session structured as a 20-minute introduction and briefing phase, a 60-minute testing phase (15 minutes per condition + five-minute breaks including questionnaires), and 5 minutes for filling out the final questionnaire, the interview, and debriefing.

5.5 Results

Text Entry Rate and Error Rate: Descriptive statistics and results of RM-ANOVA omnibus tests for entry rate and character error rate are depicted in Table 3. For text entry rate, Holm-Bonferroni adjusted post-hoc testing revealed that there were significant differences between BASELINE and HUD ($p = 0.006$) and AR ($p < 0.001$), as well as between HUD and AR ($p = 0.003$). For character error rate, Holm-Bonferroni adjusted post-hoc testing revealed that there were significant differences between BASELINE and HUD ($p = 0.002$) and AR ($p = 0.002$) as well as between HUD and AR ($p = 0.004$). In summary, HUD resulted in a 10% lower text entry rate and three times higher character error rate (albeit on a low absolute level) compared to BASELINE. AR resulted in a 21% lower text entry rate and a 14 times higher character error rate compared to BASELINE.

NASA TLX: The scores as well as the statistics of Friedman omnibus tests for workload as measured by Nasa TLX are depicted in Table 3. Post-hoc analysis with Wilcoxon signed-rank tests and Holm-Bonferroni correction revealed that there were significant differences for mental demand between AR and BASELINE ($p < 0.001$), AR and HUD ($p < 0.001$), for physical demand between all conditions ($p < 0.016$), for performance between AR and HUD ($p = 0.005$), for effort, frustration and overall demand between AR and BASELINE ($p < 0.001$). No other significant differences were indicated. In other words, AR led to significantly higher mental and physical demand compared to both BASELINE and HUD, lower performance compared to HUD, as well as higher effort, frustration, and overall demand compared to BASELINE.

TABLE 4
Reference SSQ score ranges for none to severe symptoms.

Level	Nausea	Oculo-motor	Disorientation	Total
none	0	0	0	0
slight	66.8	53.1	97.4	78.5
moderate	133.6	106.1	194.9	157.1
severe	200.3	159.2	292.3	235.6

Simulator Sickness:

The scores for the SSQ scales and for a Friedman omnibus test are depicted in Table 3. For reference purposes, the possible score ranges are depicted in Table 4. Post-hoc analysis with Wilcoxon signed-rank tests and Holm-Bonferroni correction revealed that there were significant differences for oculo-motor between AR and BASELINE ($p = 0.009$), AR and HUD ($p = 0.009$), for nausea, between AR and BASELINE ($p = 0.007$) and AR and HUD ($p = 0.006$), for disorientation, between AR and BASELINE ($p = 0.006$) and AR and HUD ($p = 0.003$), and, for total severity, again between AR and BASELINE ($p = 0.004$) and AR and HUD ($p = 0.006$). No other significant differences were indicated. In other words, AR led to significantly higher SSQ scores in all dimensions compared to both BASELINE and HUD. Symptoms can be regarded as slight (total severity, oculo-motor, disorientation for all conditions) to moderate (nausea for AR).

Flow: The scores for the flow scales and for a Friedman omnibus test are depicted in Table 3. Post-hoc analysis with Wilcoxon signed-rank tests and Holm-Bonferroni correction revealed that there were significant differences for challenge between AR and BASELINE ($p = 0.001$) as well as between AR and HUD ($p = 0.004$). No other significant differences were indicated. In other words, AR led to a significantly lower flow score compared to both HUD and BASELINE.

Preferences and Open Comments: When asked to rank all three conditions, 13 out of 18 participants preferred BASELINE, four HUD and one AR. Five participants preferred the BASELINE condition due to habits.

Friedman's tests revealed statistically significant differences on ranks ($\chi^2(2) = 17.4$, $p < 0.001$). Post-hoc analysis revealed that there were significant differences between all conditions ($p < 0.02$).

Also, four didn't enjoy wearing the OST HMD in general. Five participants explicitly preferred HUD over AR because of the display being always visible. They liked the freedom of head movements and disliked the restrictions that AR would force on them. Two felt the display to be sharper in HUD than in AR mode. Positive comments on AR included comments about the overlay being in the same plane as the smartphone's screen. Two participants liked this mode more because it felt easier for them. AR was indicated to be the most exhausting of all three types of display. People disliked the fact that they would need to have the

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2023.3309316

Fig. 3. Apparatus for Experiment 3. Left: schematic top-down view on 7 × 7 m obstacle course (black: tables, red numbers: entries into course, red line: instruction monitor blue: boundary tables). Center: Obstacle course. Right: User with HMD and smartphone.

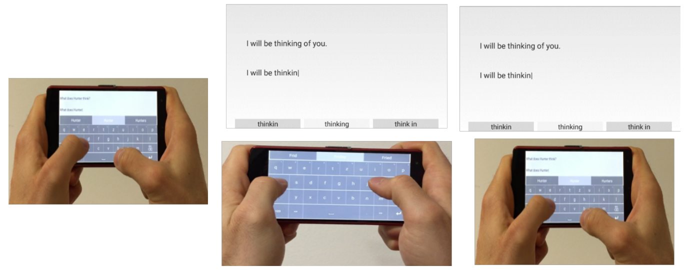

Fig. 4. Conditions for Experiment 3, obstacle course. Left: View on smartphone in condition BASELINE. Center: view on smartphone (bottom) and HMD view (top, as screenshot) for condition HUD, view on smartphone (bottom). Right: HMD view (top, again as screenshot) in condition DUAL. Please note, that the HMD views are depicted as screenshots but were visible to users on the see-through HMD (i.e. not opaque).

smartphone at one exact position in relation to the head to be able to read the text.

When asked if the participants would use OST HMDs for privacy reasons during text entry on smartphones, results of answering a 10-item Likert scale (1 not at all - 10 at all times) resulted in a mean of 3.76 (sd = 1.97). Reasons for this varied. Seven people indicated not to see a need for privacy, and, if so, there would exist other ways to achieve privacy e.g., covering the smartphone with one's body. Six people described the OST HMD for this use as too inconvenient. Two people wouldn't feel comfortable in public space using the HMD and three described the use as exhausting. However, five people could see use cases where OST HMDs could be useful. Mainly, they saw the use in occupational environments where privacy and secrecy have more important roles than during leisure activities.

5.6 Discussion

Experiment 2 indicated, that wearing an HMD for text output resulted in significantly lower text entry performance compared to a standard smartphone. Further, the spatially registered AR view performed worse than the HUD condition. The lower text entry performance is accompanied by significantly higher workload scores for AR compared both to BASELINE and HUD, significantly higher nausea and oculo-motor problems of AR compared to BASELINE and a higher challenge level for AR compared both to BASELINE and HUD.

A potential factor for this outcome is the low vertical field-of-view (ca. 11°), which forced participants to tilt their head in order to see the virtual text above the smartphone in the AR condition. Further, the vergence-accomodation conflict potentially influenced participants' performance and well-being. In summary, the joint OST HMD-smartphone system did not provide performance benefits compared to standard smartphone typing. We conjecture this result fundamentally depends on two different factors: 1) the interaction context; and 2) the parameters of the subsystems in the joint OST HMD-smartphone system.

6 Experiments 3, 4 and 5: Text Entry while Walking

We hypothesized that the static context in Experiment 2 did not provide an opportunity for the joint HMD-smartphone system to provide any detectable advantages for users. Specifically, while users could potentially recognize their physical surrounding due to looking heads-up at the OST HMD instead of looking heads-down on a smartphone, this potential benefit could not be validated in the static context of Experiment 2.

To better understand this factor, we conducted three experiments focusing on user behaviour while walking. In Experiment 3, we exposed participants to a physical obstacle course. In Experiment 4, we replaced the physical obstacle course with a virtual one to be able to investigate the effect of obstacle collision risks without endangering participants health. Finally, in Experiment 5, we studied if wearing an HMD alone, without any information displayed, would have potential negative impacts on user behaviour. For these experiments we included additional measures— number of collisions, gaze, and the Situational Awareness Rating Technique (SART) questionnaire [84], in order to investigate situation awareness in a mobile context.

6.1 Experiment 3: Obstacle Course

Experiment 3 focused on text entry while walking. To this end, we employed a physical obstacle course, in which participants were asked to walk and type simultaneously.

6.1.1 Method

The experiment was a within-subjects design with one independent variable INTERFACE. The independent variable INTERFACE had three levels: BASELINE, HUD and DUAL. In the BASELINE condition, text entry was done using the standard-sized mobile keyboard that was used in the previous experiments. In the HUD condition, as previously investigated in Experiment 2, the full-screen keyboard was used for text input and the HMD for output. As a progression from Experiment 2, we now used an Android view in the HMD to improve text clarity, see Figure 4. In the DUAL condition, text entry was carried out with a standard keyboard on the smartphone and text output was shown both on the HMD (as in condition HUD) and on the smartphone (as in condition BASELINE). This was done to investigate user behaviour when users are not forced to look at the HMD. The order of the starting condition was balanced across participants. The task was to write as quickly and as accurately as possible while walking through a physical obstacle course at a self-paced speed.

6.1.2 Participants

In this study, 24 volunteers participated (mean age 24.75 years, sd = 3.14, mean height 175.58 cm, sd = 9.89, 13 male, 11 female). None had participated in the previous experiment. Eight participants indicated to have visual restrictions. Six used contact lenses to correct these. No participant used additional glasses during the experiment. All participants were familiar with typing on a

TABLE 5
Descriptive statistics and hypothesis test statistics for text entry rate and error rate for Experiment 3. Grey rows indicate significant differences. B: BASELINE, H: HUD, D: DUAL, WPM: Words Per Minute. CER: Character Error Rate.

	mean (sd) B	mean (sd) H	mean (sd) D	$F_{2,21}$	p	η_p^2
WPM	18.70 (3.64)	16.24 (3.62)	18.05 (4.31)	11.94	< .001	0.53
CER	0.087 (0.082)	0.018 (0.015)	0.011 (0.010)	5.89	.009	0.53

smartphone. Sixteen participants never used an HMD before, 6 once, and two rarely. Data for gaze and text entry metrics from one participant needed to be excluded due to logging errors.

6.1.3 Apparatus and Materials

The apparatus is shown in Figure 3. A physical maze with dimension 7 × 7 m was used (see Figure 3, left and middle). On each side, a monitor displayed which entry participants should take into the maze. The participants were then instructed to go either to the monitor on the opposite side of the maze or to a monitor left or right to the current one outside of the maze. This was done to generate a variety of possible paths through the maze. We used an LG Nexus 5 as the smartphone and an Epson Moverio BT-300 as the OST HMD. The OST HMD was equipped with retro-reflective markers on a rod to be able to record participants' movements through the maze. We ensured that the rod had no substantial effect on the comfort of wearing the headset. Users also did not report on any such issues. Also, the HMD was equipped with a mobile Pupil Labs eye-tracker (connected to another smartphone for recording) with an accuracy of 0.6°. The text size for the displayed text output was the same across conditions.

6.1.4 Procedure

Before each condition, we calibrated the eye tracker. Participants typed stimulus sentences from the same sentence set as in Experiment 3 for a total of 10 minutes in each condition. While typing, they walked through the maze and followed instructions on which way to take through the maze. While the instructions, which way to follow, were randomized, it was possible that participants took a specific route multiple times. There were 7-minute breaks between conditions. During the breaks participants filled out the same questionnaires as in Experiment 2 (TLX, SSQ, FSF) and, in addition, the SART questionnaire. Towards the end, participants filled out a final self-assessment questionnaire, followed by a semi-structured interview. In either condition, participants were shown a series of stimulus sentences (drawn from the same phrase set as in the other experiments). The order of conditions was counterbalanced. The experiment was carried out in a single 90-minute session structured as a 20-minute introduction and briefing phase, a 65-minute testing phase (five minutes for calibration + 10 minutes per condition + seven-minute breaks including questionnaires), and 5 minutes for the final questionnaire, interview, and debriefing.

6.1.5 Results

Text Entry Rate and Error Rate:

Descriptive statistics and results of RM-ANOVA omnibus tests for entry rate and character error rate are depicted in Table 5. For, text entry rate, Holm-Bonferroni adjusted post-hoc testing revealed that there were significant differences between BASELINE and HUD ($p < 0.001$) as well as between DUAL and HUD ($p = 0.004$), but not BASELINE and DUAL ($p = 0.812$). For character error rate, Holm-Bonferroni adjusted post-hoc testing revealed that there was a significant difference between BASELINE and HUD ($p = 0.006$). No other significant differences were detected.In other words, HUD led to a significantly lower text entry rate compared to both BASELINE and DUAL, and to a significantly higher error rate compared to BASELINE.

TABLE 6
Relative gaze duration on smartphone for Experiment 3, obstacle course.

	DUAL B	HUD	baseline
mean	0.67 B	0.41	0.75
sd	0.24 B	0.17	0.22

Eye Gaze: We investigated the duration participants spent looking at the smartphone screen. To this end, for each frame, we checked if the reported normalized 2D gaze pointer would fall into the bounding box of the smartphone screen. Data from one participant had to be excluded due to non-working eye-tracking. The mean relative duration (i,e. duration spent looking at the smartphone relative to the duration of the condition) for BASELINE was 74.53% (sd = 22.87), for HUD 41.09% (sd = 17.48) and for DUAL 70.32% (sd = 19.93), see also Figure 6. An omnibus test revealed that the difference in gaze duration was statistically significant ($F_{2,20} = 46.244$, $\eta_p^2 = 0.822$, $p < 0.001$). Holm-Bonferroni adjusted post-hoc testing revealed that there were significant differences between HUD and BASELINE ($p < 0.001$), HUD and DUAL ($p < 0.001$), but not between BASELINE and DUAL ($p = 0.70$). In other words, participants looked at smartphone content in the HUD condition significantly less compared to both BASELINE and DUAL.

Workload: The mean mental demand score for workload as measured by Nasa TLX was 54.58 (sd = 19.83) for BASELINE, 57.29 (sd = 20.95) for HUD and 66.04 (sd = 23.077) for DUAL. The mean overall score for workload as measured by Nasa TLX was 47.22 (sd = 15.17) for BASELINE, 48.85 (sd = 16.19) for HUD and 53.06 (sd = 16.33) for DUAL. Friedman's tests revealed statistically significant differences for mental demand ($\chi^2(2) = 12.602$, $p = 0.002$) and overall demand ($\chi^2(2) = 6.796$, $p = 0.033$), but not for temporal demand. Post-hoc analysis with Wilcoxon signed-rank tests and Holm-Bonferroni correction revealed that there were significant differences for mental demand between DUAL and BASELINE ($p = 0.002$) as well as for DUAL and HUD ($p = 0.009$). No other significant differences were indicated. Hence, further descriptive statistics are omitted for brevity. In other words, DUAL led to a significantly higher mental demand compared to both BASELINE and HUD.

Situation Awareness:

SART scores are depicted in Table 7. Friedman's tests revealed a statistically significant difference for demand subscale ($\chi^2(2) = 10.05$, $p = 0.007$), but not for the overall SART score or the understanding and supply subscales. For demand, post-hoc analysis with Wilcoxon signed-rank tests and Holm-Bonferroni correction revealed that their were significant differences between HUD and BASELINE ($p = 0.002$), as well as HUD and DUAL ($p = 0.008$). In other words, HUD led to a significantly higher demand compared to both BASELINE and DUAL.

Simulator Sickness: SSQ scores are depicted in Table 8. Friedman's tests revealed statistically significant differences for nausea ($\chi^2(2) = 8.03$, $p = 0.018$), disorientation ($\chi^2(2) = 7.05$, $p = 0.029$) and total severity ($\chi^2(2) = 10.05$, $p = 0.007$).

TABLE 7
Average SART results for Experiment 3 with standard deviation in parenthesis. SA: Spatial Awareness. For SA, supply and understanding, higher scores are better, for demand, lower scores are better. Grey rows and bold numbers indicate scales with significant differences.

Scale	BASELINE	HUD	DUAL
SA	14.6 (4.0)	12.2 (4.4)	14.0 (4.6)
Demand	**9.6 (3.3)**	**12.8 (3.3)**	**10.66 (3.1)**
Supply	16.5 (4.0)	18.5 (3.7)	17.5 (4.0)
Understanding	7.7 (2.1)	6.5 (2.7)	7.0 (1.6)

TABLE 8
Average SSQ results for Experiment 3 with standard deviation in parenthesis. Grey rows and bold numbers indicate scales with significant differences.

Scale	BASELINE	HUD	DUAL
Total	**36.5 (36.0)**	45.3 (38.9)	**47.8 (41.1)**
Nausea	**46.1 (38.8)**	58.4 (44.8)	**58.8 (43.4)**
Oculo-Motor	17.1 (21.4)	19.3 (25.2)	21.2 (22.9)
Disorientation	**37.1 (47.5)**	48.1 (47.3)	**53.4 (58.4)**

Post-hoc analysis with Wilcoxon signed-rank tests and Holm-Bonferroni correction revealed that there were significant differences between DUAL and BASELINE for nausea ($p = 0.007$), for disorientation ($p = 0.003$) and for total severity ($p = 0.004$). No other significant differences were indicated. In other words, DUAL led to significantly higher SSQ scores (except oculo-motor) compared to BASELINE. The symptoms can be regarded as slight.

Flow: The mean overall score for flow as measured by FSS was 4.75 (sd = 0.63) for BASELINE, 4.58 (sd = 0.67) for HUD and 4.21 (sd = 0.79) for DUAL. For anxiety, it was 2.49 (sd = 1.116) for BASELINE, 3.03 (sd = 1.55) for HUD and 2.71 (sd = 1.27) for DUAL. For challenge, it was 3.96 (sd = 0.55) for BASELINE, 4.17 (sd = 0.48) for HUD and 4.46 (sd = 0.83) for DUAL.

Friedman's tests revealed statistically significant differences for the flow ($\chi^2(2) = 7.298$, $p = 0.026$) and challenge ($\chi^2(2) = 6.045$, $p = 0.049$) subscales, but not for anxiety. For flow, post-hoc analysis with Wilcoxon signed-rank tests and Holm-Bonferroni correction revealed that there was a significant difference between HUD and BASELINE ($p = 0.013$). No other significant differences were indicated. In other words, HUD led to a significantly lower flow score (12%) compared to BASELINE.

Preferences and Open Comments:

Participants were asked to rank the conditions from least preferred to most preferred. Nineteen out of 24 participants preferred BASELINE, two HUD and three DUAL.

Friedman's tests revealed statistically significant differences on ranks ($\chi^2(2) = 22.6$, $p < 0.001$). Post-hoc analysis revealed that there were significant differences between BASELINE and HUD ($p < 0.001$) and BASELINE and DUAL ($p < 0.001$), but not between HUD and DUAL.

Five participants explicitly highlighted that they could concentrated best in the smartphone only condition, with one mentioning "The less devices I had, the better I could concentrate" and another "I was least overwhelmed by stimuli". One participant explicitly mentioned, that she "needed to switch attention less often, as keyboard and text output where spatially very close" and another one "using only the smartphone, I can still see my feet." However, one participant also mentioned "I tend to forget to concentrate on the environment [using solely the smartphone]" and another related to nausea when using the smartphone while walking: "I got sick when using the smartphone, using the HMD along with the smartphone reduced this sickness."

Participants also mentioned benefits of using the HMD, e.g. "reading on OST HMD was more comfortable" and "reading on the HMD made me more aware of the environment". One participant explicitly stated that the HUD condition was "best for writing and [simultaneous] walking". On the other hand, comments addressing problematic effects included "switching between depth layers was annoying", three participants mentioned readability issues on the HMD. Also, participants mentioned the need for context switching resulting in higher workload with comments such as "I needed to reorient myself constantly", "repeated switching between smartphone and HMD lead to concentration errors", or "the attention I gain for the environment is mitigated by the attention I loose for text input." One participants would have liked haptic feedback on the smartphone keyboard to help with writing.

Additionally, for condition DUAL participants mentioned, that they used the HMD less often with comments such as "I barely looked at the HMD" or "This condition leads to irritation where to actually look". However, one participant also stated "It is relieving to be able to read the text both on the smartphone as well as in front of you."

6.2 Experiment 4: Obstacle Collisions

Experiment 3 used a physical obstacle course and revealed that participants looked significantly less at the smartphone in the HUD condition at the expense of text entry performance and attentional demand. Subjective feedback from participants indicated a potentially higher awareness of the environment (which was not reflected in SART scores). As eye gaze alone is not a suitable indicator of true attention to the environment (participants could have solely focused their attention on the virtual text in the HMD), we set out to further investigate this potential factor in another experiment. As participants tended to not make use of the HMD in the DUAL condition in Experiment 3, we removed this condition for Experiment 4. Instead of a physical course, a virtual obstacle course was set up. This was done for two reasons. First, the physical obstacle course led to a frequent change in walking speed (starting to walk and stopping again after a few meters). In addition, we wanted to investigate behavior on an extended straight boardwalk. Second, we wanted to mitigate the risks of injuries when running into physical obstacles.

6.2.1 Method

The experiment was a within-subjects design with one independent variable INTERFACE. The independent variable INTERFACE had two levels: BASELINE and HUD, which were the same conditions used in Experiment 2. The order of the starting condition was balanced across participants. The task was to write as quickly and as accurately as possible while walking on a physical treadmill and avoiding virtual obstacles. A screen in front of the participant showed a virtual path moving at the same speed (1.5 km/h) as the physical treadmill. On that path, virtual obstacles appeared

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2023.3309316

in the middle, to the left, and to the right of the participant. The participant was asked to indicate if they noticed the virtual obstacles in the middle of the path (as they could bump into them) with a button attached at the back of the smartphone. The button press was registered to the nearest obstacle. On average, the participants had between 5 and 19 seconds to react to the closest visible obstacle.

6.2.2 Participants

We recruited 17 volunteers. None had participated in the previous experiments. Three volunteers had to abort the study due to incompatibility with the HMD. Those participants experienced visual problems (diplopia, focussing issues) with the HMD during the experiment. For the remaining 14 volunteers (mean age 26.71 years, sd = 3.36, mean height 174.36 cm, sd = 8.41, 8 male, 6 female), six indicated to have visual restrictions. Three used contact lenses to correct these. No participant used additional glasses during the experiment. All participants were familiar with typing on a smartphone. Eight participants never used an HMD before, one once, four rarely and one often.

6.2.3 Apparatus and Materials

The study setup is shown in Figure 5. For this study, an Epson Moverio BT-300 was used along with a Moto Z2 Play smartphone for text entry. The HMD was equipped with a Pupil Labs mobile eye-tracker (attached to a stationary PC for recording) and with retro-reflective markers. An Optitrack V120:Trio system was used for tracking of the head tilt. Additionally, cameras were mounted to the side and in front of the participants. The treadmill was a DeskFit200 with conveyor belt dimensions of 40x90.5 cm. The treadmill was secured with styrofoam on the sides and one experimenter was positioned behind the participants to address participants potentially slipping off the treadmill. Four Samsung UE55MU6179 4K monitors were used together as a single large display with the dimensions 248 $\times$ 143 cm and were placed 100 cm in front of the participant. This resulted in an approximate horizontal field of view of the virtual path of 102° and vertical 71°. The obstacles were placed on an endless virtual path which had a width of 5 meters. Obstacles with 60 cm width appeared in three distinct lanes (left - 1.5 m from the center, middle, right - 1.5 m from the center). Obstacle frequency and placement was randomized for each lane, with all lanes using the same parameters, which were as follows: a new obstacle was placed after an average of 5 m, with the distances offset using a uniformly distributed random value in a range of 2 m to 8 m. Center obstacles were also randomly offset to the left or right by 40 cm. Outer obstacles were offset randomly in a range of 20 cm to the center, also uniformly distributed.

6.2.4 Procedure

Participants were asked to fill out a demographic questionnaire and then tested to ensure they were able to read text on the HMD using a test text. Participants' heights were used to adjust the virtual camera height for the obstacle course. Participants were then asked to get used to the treadmill in a 2 minute test run without wearing the HMD or smartphone in order to familiarize themselves with the equipment. We calibrated the eye-tracker and conducted a 5 minute testing phase before each condition. When a condition had been completed participants filled out the same questionnaires as in Experiment 2 (TLX, SSQ, SART, FSS) and were given an additional 5 minutes for resting. After the final condition, participants were asked to fill out a self-assessment questionnaire,

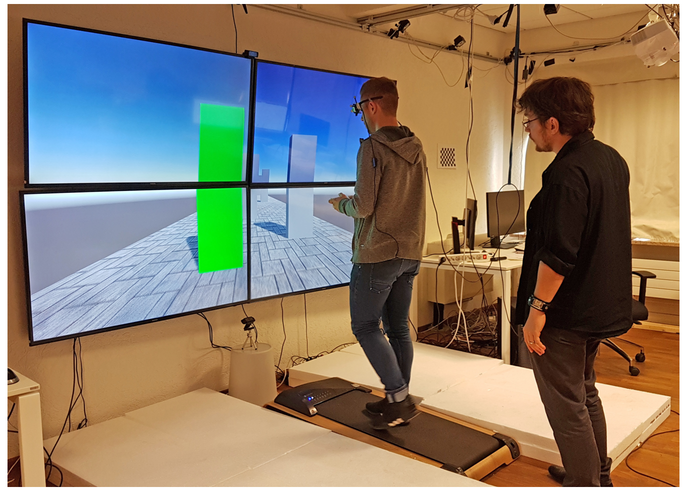

Fig. 5. Setup for Experiment 4, obstacle collisions: A participant (middle person) being secured by an experimenter (right person) and use of styrofoam on the side to prevent potentially slipping off the treadmill. On the screen in front of the participant an obstacle has been triggered as detected (highlighted in green) by the participant pushing a button on the back of the smartphone.

which was followed by a semi-structured interview. The order of the starting condition was balanced across participants. In either condition, participants were shown a series of stimulus sentences. The experiment was carried out in a 75-minute session. Starting with a 20-minute introduction, briefing phase and calibration followed by a 45-minute testing phase (5 minutes calibration per condition + 15 minutes per condition + 10 minutes questionnaires + 5 minute break between conditions) and 10 minutes for final questionnaire, interview and debriefing.

6.2.5 Results

In addition to significance tests, we also conducted two-one sided paired t-test (TOST) for equivalence testing with equivalence bounds of $d_z = \pm 0.56$ (based on a critical t-value of 2.16 at an alpha value of 0.05 and a sample size of 14, c.f. [80], [85].

Obstacle collisions: We measured the collision ratio (between 0 and 100%) as the number of undetected collisions relative to the overall number of possible collisions. For BASELINE, the mean collision ratio was 19.84% (sd = 11.43) and for HUD, it was 19.35% (sd = 14.67). We also measured a false positive ratio, i.e. the number of falsely identified obstacles (e.g., obstacles left or right of the user). For BASELINE, the false positive ratio was 23.64% (sd = 9.55) and for HUD, it was 24.10% (sd = 10.53). Paired two-tailed t-tests did not reveal a significant difference ($t(13) = 0.257, p = 0.8$). TOST indicated equivalence between conditions for both collision ratio (results for the larger of the two p-values: ($t(13) = 1.84, p = 0.045$) and false positive ratio ($t(13) = -1.83, p = 0.045$). In other words, both conditions were equivalent in terms of obstacle collisions.

Eye Gaze and Head Tilt: We investigated the duration participants spent looking at the smartphone. Data from three participants had to be excluded due to non-working eye-tracking. The mean relative duration for BASELINE was 38.70% (sd = 21.05) and for HUD 24.67% (sd = 12.00). Paired two-tailed t-tests did reveal a significant difference ($t(10) = 3.172, p = 0.10$).

We also measured the head tilt of participants, with 0° looking straight ahead and -90° looking straight down. For BASELINE,

TABLE 9
Average SART results for Experiment 4 with standard deviation in parenthesis. SA: Spatial Awareness. For SA, supply and understanding, higher scores are better, for demand, lower scores are better.

Scale	BASELINE	HUD
SA	13.9 (5.1)	11.9 (5.6)
Demand	12.1 (3.4)	13.2 (3.9)
Supply	19.4 (3.3)	19.5 (3.5)
Understanding	6.6 (2.3)	5.6 (2.1)

TABLE 10
Average SSQ results for Experiment 4 with standard deviation in parenthesis. Grey rows and bold numbers indicate scales with significant differences.

Scale	BASELINE	HUD
Total	**17.9 (18.6)**	**26.2 (19.9)**
Nausea	22.5 (21.0)	33.4 (23.9)
Oculo-Motor	11.4 (10.6)	12.5 (10.6)
Disorientation	**12.9 (22.2)**	**25.9 (28.3)**

the mean head tilt was -39.12° (sd = 46.64) and for HUD, it was -28.58° (sd = 42.13). Paired two-tailed t-tests did not reveal a significant difference. In other words, participants looked significantly less on the smartphone in the HUD condition compared to BASELINE.

Text Entry Rate and Error Rate: For BASELINE, the mean entry rate was 13.84 wpm (sd = 4.09) and for HUD 12.84 wpm (sd = 5.46). The mean character error rate was 0.018 (sd = 0.023) for BASELINE and for HUD, it was 0.045 (sd = 0.070). Paired two-tailed t-tests did not reveal any significant differences. Also, no equivalence between conditions could be detected. In other words, no significant differences for text entry or error rate could be detected, but the conditions cannot be seen as equivalent.

Workload: The mean overall score for workload as measured by Nasa TLX was 58.45 (sd = 9.37) for BASELINE and 61.67 (sd = 11.05) for HUD. Wilcoxon signed rank tests did not reveal statistically significant differences for overall demand score ($p > 0.05$) or any of the subscales, which are omitted for brevity. Also, no equivalence between conditions could be detected. In other words, no significant differences for workload could be detected but conditions cannot be considered equivalent.

Situation Awareness: Descriptive statistics for SART scores are depicted in Table 9. Wilcoxon signed rank tests did not reveal statistically significant differences for the overall SART score ($p > 0.05$) or any of the subscales, which are omitted for brevity. Also, no equivalence between conditions could be detected. In other words, no significant differences for situation awareness could be detected, but conditions cannot be considered equivalent.

Simulator Sickness: The scores for the SSQ scales are depicted in Table 10. Wilcoxon signed-rank tests revealed significant differences between BASELINE and HUD for disorientation ($Z = 2.36, p = 0.018$) and total severity ($Z = 2.11, p = 0.035$) but not for nausea or oculo-motor. In other words, HUD led to significantly higher disorientation and total severity scores than the baseline condition, but those scores can still be considered to indicate mild symptoms.

Flow: The mean overall score for flow as measured by FSS was 3.58 (sd = 2.02) for BASELINE and 3.35 (sd = 2.09) for HUD. For anxiety, it was 2.62 (sd = 2.34) for BASELINE and 2.69 (sd = 2.34) for HUD. For challenge, it was 3.36 (sd = 1.865) for BASELINE and 3.43 (sd = 1.95) for HUD. Wilcoxon signed rank tests did not reveal statistically significant differences and TOST did not indicate equivalence. In other words, no significant differences between the conditions could be detected for flow.

Preferences and Open Comments:

Participants were asked to rank the conditions from least preferred to most preferred. Nine out of 14 participants preferred BASELINE, five the HUD condition. A binomial test did not indicate a significant difference ($p = 0.424$). Two participants mentioned that using only the smartphone lead to more collisions. “I ran into more obstacles, because I was fixated on the smartphone.” For the benefits of using only the smartphone, three participants said, that they are used to writing and reading on it. “I am using it daily. I can write without looking on it.” Two participants mentioned, that they had difficulty switching between the different depth layers. “It’s hard for me to adjust my focus on three different things. It’s inconvenient.” Four participants mentioned that they had a hard time to get a clear view on the text in the HMD: “The text was blurry until I concentrated hard on reading it,” “My eyes needed time to adjust to HMD.” This was reconfirmed by two participants mentioning that their eyes were strained after the HMD condition. One participant stated that the HMD display obstructed the view of the obstacles: “I couldn’t see the obstacles because the projection was in the way.” As for the benefits of the HMD, five participant mentioned that the HMD allowed them to see more of their surroundings, with one participant saying “It feels that I can see more of my surroundings and the obstacles with the HMD.” Two participants also mentioned that they feel safer using it: “I was maybe slower with the HMD, but I feel safer.” Two participants imagined that further practice with the HMD could lead to better results: “Given more time and getting used to the HMD, I could have avoided more obstacles.”

6.3 Experiment 5: Effects of Wearing an HMD

As prior research has indicated that wearing an HMD could lead to perceptual effects, for example, on distance perception [86] due to a restricted field of vision [87], we investigated if there is an effect of wearing an HMD (that is turned off) on obstacle avoidance. To this end, we reran Experiment 4, but with the following changes:

The independent variable INTERFACE had two levels: BASELINE and HMD-OFF. HMD-OFF used the same input and output channels (the smartphone) as BASELINE, but, in addition, required users to wear the same HMD as in condition HUD, which was turned off. In the procedure, we omitted eye tracking as the input and output happened on the same device.

Fourteen volunteers participated in the study (mean age 27.5 years sd = 8.5, 10 male, 4 female). All participants were familiar with typing on a smartphone. Six participants never wore an HMD before, 4 once, 2 rarely, 1 occasionally and 1 frequently.

6.3.1 Results

In addition, to significance tests, we also conducted two-one sided paired t-test (TOST) for equivalence testing with equivalence bounds of $d_z = \pm 0.58$ (based on a critical t-value of 2.16, an alpha

TABLE 11
Average SART results for Experiment 5 with standard deviation in parenthesis. SA: Spatial Awareness. For SA, supply and understanding, higher scores are better, for demand, lower scores are better. Grey rows and bold numbers indicate significant differences.

Scale	BASELINE	HMD-OFF
SA	14.3 (5.9)	12.4 (5.9)
Demand	11.9 (3.7)	12.4 (4.0)
Supply	**19.1 (4.1)**	**17.4 (4.9)**
Understanding	7.1 (2.6)	7.5 (2.3)

TABLE 12
Average SSQ results for Experiment 5 with standard deviation in parenthesis.

Scale	BASELINE	HMD-OFF
Total	35.0 (41.3)	35.8 (35.3)
Nausea	42.2 (49.3)	42.2 (40.9)
Oculo-Motor	20.0 (19.6)	21.7 (19.9)
Disorientation	31.8 (50.2)	31.8 (46.1)

value of 0.05 and a sample size of 14. Data from two participants for the challenge subscale of the flow questionnaire were lost due to logging errors.

Obstacle collisions: For BASELINE, the mean collision ratio was 15.6% (sd = 6.5) and for HMD-OFF it was 20.0% (sd = 11.0). For BASELINE, the false positive ratio was 19.9% (sd = 8.9) and for HMD-OFF, it was 21.7% (sd = 12.2). Paired two-tailed t-tests did not reveal a significant difference and TOST did not indicate equivalence.

Text Entry Rate and Error Rate: For BASELINE, the mean entry rate was 14.8 wpm (sd = 4.4) and for HMD-OFF 14.7 wpm (sd = 3.8). The mean character error rate was 0.016 (sd = 0.020) for BASELINE and for HMD-OFF, it was 0.014 (sd = 0.013). TOST did indicate equivalence for text entry rate (results for the larger of the two p-values: ($t(13) = 1.771, p = 0.005$)) but not character error rate. To summarize, the text entry but not error rate was equivalent between both conditions.

Workload: The mean overall score for workload as measured by Nasa TLX was 62.02 (sd = 13.93) for BASELINE and 61.37 (sd = 12.53) for HMD-OFF. Wilcoxon signed rank tests did not reveal statistically significant differences for overall demand score ($p > 0.05$) or any of the subscales, which are omitted for brevity. TOST indicated equivalence for physical demand (results for the larger of the two p-values: $t(13) = 1.86, p = 0.043$), frustration ($t(13) = -1.99, p = 0.034$) and overall demand ($t(13) = 1.86, p = 0.043$), but not for mental demand, temporal demand, effort or performance. In other words, the conditions can be seen as equivalent in terms of overall demand (but not for each subscale).

Situation Awareness: SART results are depicted in Table 11 Wilcoxon signed rank tests indicated a significant difference for the supply subscale (i.e. how users can supply their attentional resources using items on arousal, spare mental capacity, concentration and division of attention) between BASELINE and HMD-OFF (Z = 2.17, p = 0.03, Cohen's d = 0.9), but not for demand, understanding or overall. Also, no equivalence between conditions for any of those measures was indicated. In other words, wearing an HMD, even though it was turned off led to a significantly lower supply rating compared to BASELINE.

Simulator Sickness: The scores for the SSQ scales are depicted in Table 12. TOST indicated equivalence for nausea (results for the larger of the two p-values: $t(13) = 2.10, p = 0.028$), disorientation $t(13) = 2.10, p = 0.028$), and total severity ($t(13) = -197, p = 0.035$), but not for oculo-motor ($t(13) = -165, p = 0.061$). In other words, both conditions can be considered equivalent in terms of induced simulator sickness for all but the oculo-motor component.

Flow: The mean overall score for flow as measured by FSS was 4.57 (sd = 1.11) for BASELINE and 4.61 (sd = 1.02) for HMD-OFF. For anxiety, it was 3.74 (sd = 1.50) for BASELINE and 2.95 (sd = 1.51) for HMD-OFF. For challenge, it was 4.67 (sd = 0.99) for BASELINE and 4.92 (sd = 0.90) for HMD-OFF. Wilcoxon signed rank tests did not reveal statistically significant differences. TOST indicated equivalence for flow (results for the larger of the two p-values: $t(13) = -3.44, p = 0.002$), but not for anxiety and challenge. In other words, no significant differences between conditions could be detected for flow.

Open Comments:

In a semi-structured interview, participants were asked if they felt that wearing the HMD (even though it was turned off) was impacting the task. Seven participants did not notice a difference between BASELINE and HMD-OFF. Five participants, mentioned that HMD-OFF was negatevily impacting the task. All of those five participants usually do not wear corrective glasses. Two participants, mentioned that the HMD was limiting the available field-of-view. Another one mentioned that wearing the HMD felt unusual. Yet, another two participant explicitly mentioned, that they felt that they were recognizing the targets less often when wearing the HMD, with one mentioning looking through the HMD "felt like looking through a veil". Two participants, preferred wearing the HMD, with one mentioning a learning effect (the participant conducted the HMD-OFF condition after BASELINE) and another one mentioning fatiguing effects (HMD-OFF was conducted before BASELINE) as reasons for this preference.

7 DISCUSSION

In this paper we have studied possible merits and challenges of a joint OST HMD-smartphone system for mobile text entry.

In Experiment 1, we quantified the performance gains a full-screen smartphone keyboard could have over a standard-size smartphone keyboard. We verified that the full-screen keyboard lead to a significantly higher text entry rate (around 15%) with no significant difference in error rate. In addition, the larger keyboard also led to a lower workload, which was also supported by participants' remarks.

In Experiment 2, we found that wearing an HMD for text output significantly reduced text entry performance compared to a standard smartphone baseline. In addition, we found that a spatially registered AR text output view performed worse than a HUD text output view. This result is supported by the literature on context and focus distance switching (e.g., [54], [56]), which indicate substantial costs when visual information is processed across an OST HMD and another display. Specifically, Eiberger et al. [56]

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2023.3309316

studied a setup very similar to ours, with a comparable HMD model and a secondary display at a typical reading distance. While we could not isolate the effects of the joint visual information processing in experiment 2, the overall results, together with the results from experiment 1 (higher text entry performance with full-screen keyboard), and results reported in prior work, it seems clear that there are major costs introduced due to the need for switching between the HMD and the smartphone. In addition, while we ensured that the AR condition was spatially registered for each participant, we did not use a user-specific calibration. In future work, one should further investigate that factor by either utilizing user-specific calibrations or replicating the experiment with a high-resolution video see-through display. Also, while previous work [1] considered joint OST HMD-smartphone systems having potential benefits for supporting privacy, the subjective feedback in Experiment 2 did not confirm this.

Due to observations that participants looked down at the smartphone considerably less in Experiment 2, we hypothesized that there could be potential advantages in mobile scenarios in which users need to divide attention between the interactive system and the surrounding environment. Hence, we ran another set of experiments focusing on the need for splitting the attention between text entry and the physical environment. To this end, Experiment 3 used a physical obstacle course and Experiments 4 and 5 used a treadmill and virtual obstacles to compare typing using the joint OST HMD-smartphone system with typing on a smartphone only.

Experiment 3 (physical obstacle course) showed that typing with a joint OST HMD-smartphone system led to significantly slower and more error prone text entry compared to smartphone-only typing. This experiment also indicated that users tended to use the smartphone if the text was simply mirrored on the HMD. Subjective feedback indicated that participants experienced significantly more oculo-motor symptoms when wearing the HMD and most participants (19 out of 24) preferred smartphone-only typing. However, this was not reflected in the overall workload or flow ratings. Also, we found that users tended to look on the smartphone about 70% of the time when using the smartphone for text input and output compared to approximately 40% of the time when using the HMD for text output. This was accompanied by user comments on increased awareness of the environment (even though this was not reflected in SART scores).

However, the design of Experiment 3 meant that we were unable to detect whether participants' awareness had actually improved as additional attention might have been simply directed towards the HMD. In addition, the physical obstacle course did not allow us to study extended continuous walking as participants needed to slow down and change direction after a few meters. Therefore Experiment 4 used virtual obstacles on a treadmill, which allowed us to investigate whether participants were able to avoid obstacles less with an HMD. Experiment 4 revealed that the obstacle collision rate was equivalent (around 20 %) for both conditions and that the HUD condition resulted in significantly higher (but still mild) simulator sickness in terms of disorientation and total severity.

Finally, based on prior indications [87], Experiment 5 investigated whether simply wearing an OST HMD that is turned off would have an impact on user behavior in the same task as in Experiment 4. The results indicated equivalence for text entry speed, overall demand and simulator sickness (except the oculo-motor subscale). At the same time, wearing an HMD led to significantly lower attentional supply.

In summary, the five experiments have contributed to a nuanced understanding of the performance of a joint OST HMD-smartphone system. In a static setting, the joint OST HMD-smartphone system has significantly reduced performance compared to a smartphone-only baseline. In a dynamic setting, where users must maintain awareness of their surroundings, the joint OST HMD-smartphone system does reduce the number of times users look at the smartphone. However, users are still not further aware of their surroundings as evidenced by the obstacle collision rate being similar for both systems. Hence, further research is needed to investigate if it is possible, and if so how, to translate the theoretical performance benefits of a joint OST HMD-smartphone system into mobile settings.

For example, the setup used a current-generation OST HMD with a single focal plane. Prior work has demonstrated costs of focal plane switching (e.g. [53], [54], [56]), which might also have impacted text entry performance in our study. Possibly, varifocal or multiple foci OST HMDs [59], [60], [61], [88] would reduce this negative impact by aligning both the focal plane of the smartphone and the HMD, or of the HMD and the physical environment (if users are looking at the physical surroundings). Still, even if this could be realized, there would still remain costs of context switching [50], [53], [54]. Hence, it remains to be seen if future OST HMD systems could substantially improve the situation awareness, or text entry performance. Until then, we would caution against proposing OST HMDs as a likely solution for reducing distractions for mobile text entry or other smartphone-related activities demanding visual attention.

Our experiments could be modified by changing the parameters, such as changing the frequency of obstacles. It is also possible to change the nature of the obstacles, such as having obstacles on the ground. While it is possible to conceive many alternative experimental designs, on balance, we believe our results are robust to minor adjustments in the design and procedure. Further experiments could consider alternative ways of presenting text on an HMD [46].

While it was not viable to increase exposure time in our experimental setups, we did get indications from some participants that they could possibly increase their performance with practice. In general, we suggest an interesting avenue for future work would be to study OST HMD-smartphone systems, and other complex designs, longitudinal in real interaction contexts, possibly using the Experience Sampling Methodology [89].

Finally, the studies were relatively short interventions. Hence, we could not investigate long-term learning effects. Quantifying potentially different learning rates and performance bounds is one avenue of future work.

8 Conclusions

The central contribution of this paper is a series of investigations that provide a nuanced understanding of the opportunities and challenges inherent with a joint OST HMD-smartphone system. To date, OST HMDs lack efficient text entry methods. Since smartphones are a major text entry medium in mobile contexts but their attentional demands can contribute to accidents while typing on the go, we explored the possible performance benefits of a joint OST HMD-smartphone system. In a series of five experiments with a total of 86 participants we found that, as of today, the challenges inherent in a joint OST HMD-smartphone system outweigh the potential benefits compared to a smartphone-only baseline. The experiments confirmed previous findings about the performance issues in multi-depth information environments that come along with today's single focus HMD design. We hope that joint HMD-smartphone ecologies get a boost from upcoming multi-focus and variofocal HMD designs.

REFERENCES

[1] J. Grubert, M. Heinisch, A. Quigley, and D. Schmalstieg, "Multifi: Multi fidelity interaction with displays on and around the body," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2015, pp. 3933–3942.

[2] M. R. Endsley, "Toward a theory of situation awareness in dynamic systems," in *Situational awareness*. Routledge, 2017, pp. 9–42.

[3] J. L. Nasar and D. Troyer, "Pedestrian injuries due to mobile phone use in public places," *Accident Analysis & Prevention*, vol. 57, pp. 91–95, 2013.

[4] D. C. Schwebel, D. Stavrinos, K. W. Byington, T. Davis, E. E. ONeal, and D. De Jong, "Distraction and pedestrian safety: how talking on the phone, texting, and listening to music impact crossing the street," *Accident Analysis & Prevention*, vol. 45, pp. 266–271, 2012.

[5] A. K. Yadav and N. Velaga, "A systematic review of observational studies investigating the influence of mobile phone distraction on road crossing behaviour of pedestrians," *Available at SSRN 4047909*.

[6] K. Lyons, T. Starner, D. Plaisted, J. Fusia, A. Lyons, A. Drew, and E. Looney, "Twiddler typing: One-handed chording text entry for mobile phones," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2004, pp. 671–678.

[7] D.-M. Pham and W. Stuerzlinger, "Hawkey: Efficient and versatile text entry for virtual reality," in *25th ACM Symposium on Virtual Reality Software and Technology*, 2019, pp. 1–11.

[8] P. O. Kristensson, "Next-generation text entry." *IEEE Computer*, vol. 48, no. 7, pp. 84–87, 2015.

[9] ——, "Five challenges for intelligent text entry methods," *AI Magazine*, vol. 30, no. 4, p. 85, 2009.

[10] G. González, J. P. Molina, A. S. García, D. Martínez, and P. González, "Evaluation of text input techniques in immersive virtual environments," in *New Trends on Human–Computer Interaction*. Springer, 2009, pp. 109–118.

[11] S. Reyal, S. Zhai, and P. O. Kristensson, "Performance and user experience of touchscreen and gesture keyboards in a lab setting and in the wild," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, ser. CHI '15. New York, NY, USA: ACM, 2015, pp. 679–688. [Online]. Available: http://doi.acm.org/10.1145/2702123.2702597

[12] K. Vertanen, H. Memmi, J. Emge, S. Reyal, and P. O. Kristensson, "Velocitap: Investigating fast mobile text entry using sentence-based decoding of touchscreen keyboard input," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2015, pp. 659–668.

[13] K. Vertanen, C. Fletcher, D. Gaines, J. Gould, and P. O. Kristensson, "The impact of word, multiple word, and sentence input on virtual keyboard decoding performance," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 2018, p. 626.

[14] M. Goel, L. Findlater, and J. Wobbrock, "Walktype: using accelerometer data to accomodate situational impairments in mobile touch screen text entry," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2012, pp. 2687–2696.

[15] D. Weir, H. Pohl, S. Rogers, K. Vertanen, and P. O. Kristensson, "Uncertain text entry on mobile devices," in *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2014, pp. 2307–2316.

[16] D. A. Bowman, C. J. Rhoton, and M. S. Pinho, "Text input techniques for immersive virtual environments: An empirical comparison," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 46, no. 26. SAGE Publications, 2002, pp. 2154–2158.

[17] H. Jiang, D. Weng, Z. Zhang, and F. Chen, "Hifinger: One-handed text entry technique for virtual environments based on touches between fingers," *Sensors*, vol. 19, no. 14, p. 3063, 2019.

[18] D. Brun, S. George, and C. Gouin-Vallerand, "Keycube: Text entry evaluation with a cubic device," in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–9.

[19] 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)*, vol. 25, no. 6, pp. 1–40, 2018.

[20] 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)*. IEEE, 2019, pp. 289–300.

[21] Y. Wang, Y. Wang, J. Chen, Y. Wang, J. Yang, T. Jiang, and J. He, "Investigating the performance of gesture-based input for mid-air text entry in a virtual environment: a comparison of hand-up versus hand-down postures," *Sensors*, vol. 21, no. 5, p. 1582, 2021.

[22] S. Kim and G. J. Kim, "Using keyboards with head mounted displays," in *Proceedings of the 2004 ACM SIGGRAPH international conference on Virtual Reality continuum and its applications in industry*, 2004, pp. 336–343.

[23] J. Grubert, L. Witzani, E. Ofek, M. Pahud, M. Kranz, and P. O. Kristensson, "Text entry in immersive head-mounted display-based virtual reality using standard keyboards," in *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2018, pp. 159–166.

[24] ——, "Effects of hand representations for typing in virtual reality," in *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2018, pp. 151–158.

[25] A. Otte, T. Menzner, T. Gesslein, P. Gagel, D. Schneider, and J. Grubert, "Towards utilizing touch-sensitive physical keyboards for text entry in virtual reality," in *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2019, pp. 1729–1732.

[26] A. Otte, D. Schneider, T. Menzner, T. Gesslein, P. Gagel, and J. Grubert, "Evaluating text entry in virtual reality using a touch-sensitive physical keyboard," in *2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE, 2019, pp. 387–392.

[27] T. Menzner, A. Otte, T. Gesslein, J. Grubert, P. Gagel, and D. Schneider, "A capacitive-sensing physical keyboard for vr text entry," in *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2019, pp. 1080–1081.

[28] J. Gugenheimer, D. Dobbelstein, C. Winkler, G. Haas, and E. Rukzio, "Facetouch: Enabling touch interaction in display fixed uis for mobile virtual reality," in *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. ACM, 2016, pp. 49–60.

[29] M. Speicher, A. M. Feit, P. Ziegler, and A. Krüger, "Selection-based text entry in virtual reality," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–13.

[30] W. J. Norton, P. Pitura, and D. Gerhard, "Design and implementation of a chorded-keyboard mapping for existing vr hand controllers," in *Proceedings of the Future Technologies Conference*. Springer, 2021, pp. 443–454.

[31] W. Xu, H.-N. Liang, Y. Zhao, T. Zhang, D. Yu, and D. Monteiro, "Ringtext: Dwell-free and hands-free text entry for mobile head-mounted displays using head motions," *IEEE transactions on visualization and computer graphics*, vol. 25, no. 5, pp. 1991–2001, 2019.

[32] W. Xu, H.-N. Liang, A. He, and Z. Wang, "Pointing and selection methods for text entry in augmented reality head mounted displays," in *2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 2019, pp. 279–288.

[33] T. J. Dube and A. S. Arif, "Text entry in virtual reality: A comprehensive review of the literature," in *International Conference on Human-Computer Interaction*. Springer, 2019, pp. 419–437.

[34] A. Oulasvirta, S. Tamminen, V. Roto, and J. Kuorelahti, "Interaction in 4-second bursts: the fragmented nature of attentional resources in mobile hci," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2005, pp. 919–928.

[35] J. D. Hincapié-Ramos and P. Irani, "Crashalert: enhancing peripheral alertness for eyes-busy mobile interaction while walking," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2013, pp. 3385–3388.

[36] J. Wen, J. Cao, and X. Liu, "We help you watch your steps: Unobtrusive alertness system for pedestrian mobile phone users," in *Pervasive Computing and Communications (PerCom), 2015 IEEE International Conference on*. IEEE, 2015, pp. 105–113.

[37] K.-T. Foerster, A. Gross, N. Hail, J. Uitto, and R. Wattenhofer, "Spareeye: enhancing the safety of inattentionally blind smartphone users," in *Proceedings of the 13th international conference on mobile and ubiquitous multimedia*. ACM, 2014, pp. 68–72.

[38] X. Liu, J. Cao, J. Wen, and S. Tang, "Infrasee: An unobtrusive alertness system for pedestrian mobile phone users," *IEEE Transactions on Mobile Computing*, vol. 16, no. 2, pp. 394–407, 2017.

[39] J. Van Dam, S. J. Kass, and L. VanWormer, "The effects of passive mobile phone interaction on situation awareness and driving performance," *Journal of Transportation Safety & Security*, vol. 12, no. 8, pp. 1007–1024, 2020.

[40] J. Woodward and J. Ruiz, "Analytic review of using augmented reality for situational awareness," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 4, pp. 2166–2183, 2022.

[41] J. L. Gabbard, J. E. Swan, and D. Hix, "The effects of text drawing styles, background textures, and natural lighting on text legibility in outdoor augmented reality," *Presence: Teleoperators & Virtual Environments*, vol. 15, no. 1, pp. 16–32, 2006.

[42] M. Gattullo, A. E. Uva, M. Fiorentino, and J. L. Gabbard, "Legibility in industrial ar: text style, color coding, and illuminance," *IEEE computer graphics and applications*, vol. 35, no. 2, pp. 52–61, 2015.

[43] J. Orlosky, K. Kiyokawa, and H. Takemura, "Dynamic text management for see-through wearable and heads-up display systems," in *Proceedings of the 2013 international conference on Intelligent user interfaces*. ACM, 2013, pp. 363–370.

[44] B. Bell, S. Feiner, and T. Höllerer, "View management for virtual and augmented reality," in *Proceedings of the 14th annual ACM symposium on User interface software and technology*. ACM, 2001, pp. 101–110.

[45] A. Lucero and A. Vetek, "Notifeye: using interactive glasses to deal with notifications while walking in public," in *Proceedings of the 11th Conference on Advances in Computer Entertainment Technology*. ACM, 2014, p. 17.

[46] R. Rzayev, P. W. Wozniak, T. Dingler, and N. Henze, "Reading on smart glasses: The effect of text position, presentation type and walking," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 2018, p. 45.

[47] A. Erickson, K. Kim, A. Lambert, G. Bruder, M. P. Browne, and G. F. Welch, "An extended analysis on the benefits of dark mode user interfaces in optical see-through head-mounted displays," *ACM Transactions on Applied Perception (TAP)*, vol. 18, no. 3, pp. 1–22, 2021.

[48] L. Pavanatto, C. North, D. A. Bowman, C. Badea, and R. Stoakley, "Do we still need physical monitors? an evaluation of the usability of ar virtual monitors for productivity work," in *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2021, pp. 759–767.

[49] P.-L. P. Rau, J. Zheng, and Z. Guo, "Immersive reading in virtual and augmented reality environment," *Information and Learning Sciences*, 2021.

[50] A. Huckauf, M. H. Urbina, J. Grubert, I. Böckelmann, F. Doil, L. Schega, J. Tümler, and R. Mecke, "Perceptual issues in optical-see-through displays," in *Proceedings of the 7th Symposium on Applied Perception in Graphics and Visualization*, 2010, pp. 41–48.

[51] D. Gupta, "An empirical study of the effects of context-switch, object distance, and focus depth on human performance in augmented reality," Ph.D. dissertation, Virginia Tech, 2004.

[52] J. Gabbard, D. G. Mehra, and J. E. Swan II, "Effects of ar display context switching and focal distance switching on human performance," *IEEE Transactions on Visualization and Computer Graphics*, 2018.

[53] M. S. Arefin, N. Phillips, A. Plopski, J. L. Gabbard, and J. E. Swan II, "Impact of ar display context switching and focal distance switching on human performance: Replication on an ar haploscope," in *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 2020, pp. 571–572.

[54] M. S. Arefin, N. Phillips, A. Plopski, J. L. Gabbard, and J. Swan, "The effect of context switching, focal switching distance, binocular and monocular viewing, and transient focal blur on human performance in optical see-through augmented reality," 2022.

[55] M. D. Winterbottom, R. Patterson, B. J. Pierce, C. M. Covas, and J. Winner, "Depth of focus and visual recognition of imagery presented on simultaneously viewed displays: implications for head-mounted displays," *Human factors*, vol. 49, no. 5, pp. 907–919, 2007.

[56] A. Eiberger, P. O. Kristensson, S. Mayr, M. Kranz, and J. Grubert, "Effects of depth layer switching between an optical see-through head-mounted display and a body-proximate display," in *Symposium on Spatial User Interaction*, 2019, pp. 1–9.

[57] M. Drouot, N. Le Bigot, J.-L. de Bougrenet, and V. Nourrit, "Effect of context and distance switching on visual performances in augmented reality," in *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 2021, pp. 476–477.

[58] J. L. Gabbard, G. M. Fitch, and H. Kim, "Behind the glass: Driver challenges and opportunities for ar automotive applications," *Proceedings of the IEEE*, vol. 102, no. 2, pp. 124–136, 2014.

[59] D. Dunn, C. Tippets, K. Torell, P. Kellnhofer, K. Akşit, P. Didyk, K. Myszkowski, D. Luebke, and H. Fuchs, "Wide field of view varifocal near-eye display using see-through deformable membrane mirrors," *IEEE transactions on visualization and computer graphics*, vol. 23, no. 4, pp. 1322–1331, 2017.

[60] K. Akşit, W. Lopes, J. Kim, P. Shirley, and D. Luebke, "Near-eye varifocal augmented reality display using see-through screens," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 6, p. 189, 2017.

[61] A. Wilson and H. Hua, "High-resolution optical see-through vari-focal-plane head-mounted display using freeform alvarez lenses," in *Digital Optics for Immersive Displays*, vol. 10676. International Society for Optics and Photonics, 2018, p. 106761J.

[62] K. Oshima, K. R. Moser, D. C. Rompapas, J. E. Swan, S. Ikeda, G. Yamamoto, T. Taketomi, C. Sandor, and H. Kato, "Sharpview: Improved clarity of defocussed content on optical see-through head-mounted displays," in *Virtual Reality (VR), 2016 IEEE*. IEEE, 2016, pp. 253–254.

[63] T. Cook, N. Phillips, K. Massey, A. Plopski, and C. Sandor, "User preference for sharpview-enhanced virtual text during non-fixated viewing," in *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2018, pp. 1–400.

[64] S. Imamov, D. Monzel, and W. S. Lages, "Where to display? how interface position affects comfort and task switching time on glanceable interfaces," in *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2020, pp. 851–858.

[65] G. A. Koulieris, K. Akşit, M. Stengel, R. K. Mantiuk, K. Mania, and C. Richardt, "Near-eye display and tracking technologies for virtual and augmented reality," in *Computer Graphics Forum*, vol. 38, no. 2. Wiley Online Library, 2019, pp. 493–519.

[66] D. Wenig, J. Schöning, A. Olwal, M. Oben, and R. Malaka, "Watchthru: Expanding smartwatch displays with mid-air visuals and wrist-worn augmented reality," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 2017, pp. 716–721.

[67] P. Mohr, M. Tatzgern, T. Langlotz, A. Lang, D. Schmalstieg, and D. Kalkofen, "Trackcap: Enabling smartphones for 3d interaction on mobile head-mounted displays," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–11.

[68] F. Zhu and T. Grossman, "Bishare: Exploring bidirectional interactions between smartphones and head-mounted augmented reality," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–14.

[69] S. Hubenschmid, J. Zagermann, S. Butscher, and H. Reiterer, "Stream: Exploring the combination of spatially-aware tablets with augmented reality head-mounted displays for immersive analytics," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–14.

[70] R. Langner, M. Satkowski, W. Büschel, and R. Dachselt, "Marvis: Combining mobile devices and augmented reality for visual data analysis," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–17.

[71] P. Knierim, D. Hein, A. Schmidt, and T. Kosch, "The smartphone controller," *i-com*, vol. 20, no. 1, pp. 49–61, 2021.

[72] D. Wolf, J. J. Dudley, and P. O. Kristensson, "Performance envelopes of in-air direct and smartwatch indirect control for head-mounted augmented reality," in *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2018, pp. 347–354.

[73] 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.

[74] E. Normand and M. J. McGuffin, "Enlarging a smartphone with ar to create a handheld vesad (virtually extended screen-aligned display)," in *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 2018, pp. 123–133.

[75] H. Bai, L. Zhang, J. Yang, and M. Billinghurst, "Bringing full-featured mobile phone interaction into virtual reality," *Computers & Graphics*, vol. 97, pp. 42–53, 2021.

[76] A. Sears, D. Revis, J. Swatski, R. Crittenden, and B. Shneiderman, "Investigating touchscreen typing: the effect of keyboard size on typing speed," *Behaviour & Information Technology*, vol. 12, no. 1, pp. 17–22, 1993.

[77] J. H. Kim, L. S. Aulck, O. Thamsuwan, M. C. Bartha, C. A. Harper, and P. W. Johnson, "The effects of touch screen virtual keyboard key sizes on typing performance, typing biomechanics and muscle activity," in *International Conference on Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management*. Springer, 2013, pp. 239–244.

[78] K. Vertanen and P. O. Kristensson, "A versatile dataset for text entry evaluations based on genuine mobile emails," in *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, ser. MobileHCI '11. New York, NY, USA: ACM, 2011, pp. 295–298. [Online]. Available: http://doi.acm.org/10.1145/2037373.2037418

[79] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," *Advances in psychology*, vol. 52, pp. 139–183, 1988. [Online]. Available: https://dx.doi.org/10.1016/S0166-4115(08)62386-9

[80] D. Lakens, "Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and anovas," *Frontiers in psychology*, vol. 4, p. 863, 2013.

[81] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal, "Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness," *The international journal of aviation psychology*, vol. 3, no. 3, pp. 203–220, 1993.

[82] F. Rheinberg, S. Engeser, and R. Vollmeyer, "Measuring components of flow: the flow-short-scale," *Proceedings of the 1st International Positive Psychology Summit*, 2002.

[83] M. Tuceryan, Y. Genc, and N. Navab, "Single-point active alignment method (spaam) for optical see-through hmd calibration for augmented reality," *Presence: Teleoperators & Virtual Environments*, vol. 11, no. 3, pp. 259–276, 2002.

[84] R. Taylor, "Situational awareness rating technique (sart): The development of a tool for aircrew systems design," in *Situational Awareness*. Routledge, 2017, pp. 111–128.

[85] D. Lakens, A. M. Scheel, and P. M. Isager, "Equivalence testing for psychological research: A tutorial," *Advances in Methods and Practices in Psychological Science*, vol. 1, no. 2, pp. 259–269, 2018.

[86] R. Messing and F. H. Durgin, "Distance perception and the visual horizon in head-mounted displays," *ACM Transactions on Applied Perception (TAP)*, vol. 2, no. 3, pp. 234–250, 2005.

[87] S. Darius, J. Marquez, O. Chegrynets, R. Mecke, and I. Boeckelmann, "Investigation of the field of vision when using various head-mounted displays," *ZENTRALBLATT FUR ARBEITSMEDIZIN ARBEITSSCHUTZ UND ERGONOMIE*, vol. 65, no. 4, pp. 203–209, 2015.

[88] Y. Itoh, T. Langlotz, J. Sutton, and A. Plopski, "Towards indistinguishable augmented reality: A survey on optical see-through head-mounted displays," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–36, 2021.

[89] S. Consolvo and M. Walker, "Using the experience sampling method to evaluate ubicomp applications," *IEEE Pervasive Computing*, vol. 2, no. 2, pp. 24–31, 2003.

Jens Grubert is Professor of Human-Computer Interaction in the Internet of Things at Coburg University of Applied Sciences and Arts. He has more than 15 years of academic and industrial experience in fields such as human-computer interaction, mixed reality, computer graphics and computer vision. His research interests include supporting future knowledge work with mixed reality, multimodal and mutli-device interaction in augmented and virtual reality. He is an avid supporter of interdisciplinary work.

Lukas Witzani graduated from the University of Passau were he was conducting research on text entry in virtual reality.

Alexander Otte worked at the mixedrealitylab at the University of Applied Sciences and Arts. He researched human-computer interaction in virtual and augmented reality with a focus on text entry. Specifically, he focused on building novel hardware prototypes and conduction human-subject exeriments.

Travis Gesslein worked at the mixedrealitylab at the University of Applied Sciences and Arts. He researched interactive protoypes for augmented and virtual reality. Specificially, Travis worked on combining multiple input modalities for efficient cross-device interaction between head-mounted displays and touch devices such as smartphones and tablets.

Matthias Kranz is Professor of Embedded Systems at the Faculty for Informatics and Mathematics at Passau University. His research focuses on embedded systems and human-computer interaction.

Per Ola Kristensson is Professor of Interactive Systems Engineering in the Department of Engineering at the University of Cambridge and a Fellow of Trinity College, Cambridge. He leads the Intelligent Interactive Systems group, which belongs to the Engineering Design Centre. He is also a co-founder and co-director of the Centre for Human-Inspired Artificial Intelligence at the University of Cambridge.

APPENDIX

NOMENCLATURE

AR	augmented reality
CER	character error rate
CRT	cathode-ray tube
FSS	flow-short-scale
HMD	head-mounted display
HUD	heads-up display
OST	optical see-through
RM-ANOVA	repeated measures analysis of variance
SART	situational awareness rating technique
SD	standard deviation
SSQ	simulator sickness questionnaire
TLX	task load index
TOST	two-one sided paired t-test
VR	vitual reality