ThumbAir: In-Air Typing for Head Mounted Displays

HYUNJAE GIL, UNIST, Republic of Korea
IAN OAKLEY, UNIST, Republic of Korea

Typing while wearing a standalone Head Mounted Display (HMD)—systems without external input devices or sensors to support text entry—is hard. To address this issue, prior work has used external trackers to monitor finger movements to support in-air typing on virtual keyboards. While performance has been promising, current systems are practically infeasible: finger movements may be visually occluded from inside-out HMD based tracking systems or, otherwise, awkward and uncomfortable to perform. To address these issues, this paper explores an alternative approach. Taking inspiration from the prevalence of thumb-typing on mobile phones, we describe four studies exploring, defining and validating the performance of ThumbAir, an in-air thumb-typing system implemented on a commercial HMD. The first study explores viable target locations, ultimately recommending eight targets sites. The second study collects performance data for taps on pairs of these targets to both inform the design of a target selection procedure and also support a computational design process to select a keyboard layout. The final two studies validate the selected keyboard layout in word repetition and phrase entry tasks, ultimately achieving final WPMs of 27.1 and 13.73. Qualitative data captured in the final study indicate that the discreet movements required to operate ThumbAir, in comparison to the larger scale finger and hand motions used in a baseline design from prior work, lead to reduced levels of perceived exertion and physical demand and are rated as acceptable for use in a wider range of social situations.

164

CCS Concepts: • **Human-centered computing → Text input**; *Virtual reality*; *User studies.*

Additional Key Words and Phrases: Head Mounted Display, Text entry, Virtual Reality

ACM Reference Format:
Hyunjae Gil and Ian Oakley. 2022. ThumbAir: In-Air Typing for Head Mounted Displays. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 4, Article 164 (December 2022), 30 pages. https://doi.org/10.1145/3569474

1 INTRODUCTION

Augmented Reality (AR) and Virtual Reality (VR) Head Mounted Displays (HMDs) are rapidly developing. Recent models feature expansive high resolution, high refresh rate screens capable of providing rich experiences, while also remaining sufficiently comfortable and lightweight for prolonged use. Both end-user application areas, such as casual gaming and entertainment [15], and professional applications, such as training [48] or work and task support [52], are blossoming [57]. For AR, many of most compelling use cases involve relatively uncontrolled settings such as to enhance experiences as visitors wander around a museum [68], to support equipment maintenance tasks in the field [12], to enhance educational activities in a classroom [71] or in various accessibility scenarios [31]. In tandem with this growth, HMD sensing capabilities are also increasing, with current standalone devices supporting not only accurate and responsive controller tracking but also state-of-the-art bare hand tracking [70]. However, while these systems are impressive, they still struggle to support key high bandwidth digital input tasks, such as text entry—a study of Microsoft's HoloLens default keyboard interface,

Authors' addresses: Hyunjae Gil, UNIST, Ulsan, Republic of Korea, hyunjaegil@gmail.com; Ian Oakley, UNIST, Ulsan, Republic of Korea, ian.r.oakley@gmail.com.

2474-9567/2022/12-ART164 $15.00
https://doi.org/10.1145/3569474

for example, reported a mean WPM of just 5.86 [18]. This is problematic as text entry is an important input modality required for a very wide range of common, generic and everyday tasks such as entering identification information and authentication credentials, composing written chat/messages, accessing or searching online resources and dealing with file and system management tasks.

Reflecting the importance of text entry tasks, a wide variety of work has explored how rapid and reliable text entry can be achieved on both VR and AR HMDs. The dominant approach mimics the traditional computer typing experience—a virtual keyboard is presented to users, who must then select the desired keys in sequence. However, specific implementations of this basic design are diverse. Many rely on tracked hand held controllers and use the series of intersections between a cursor [82, 87] or ray [72, 75] and keyboard buttons generated by either a set of discrete taps [16, 73, 91] or a continuous stroke [39, 88] to specify characters. While systems of this sort benefit from the high accuracy of controller based tracking, text entry performance (in the range of 13.6 to 24.73 WPM for the representative examples cited above) may be artificially limited by their reliance of cursor-based input. In addition, using controllers makes these designs unsuitable for many AR (or casual VR) settings in which holding and managing input peripherals is either impossible (e.g., during hands-busy maintenance) or impractical (e.g., during a prolonged museum visit). Projects that eschew controllers and rely on bare hand tracking have the potential to address these limitations—to attain speeds approaching those of keyboard typing while also enabling users' hands to remain unencumbered. The existing performance data is promising. Using a high end marker based optical tracking system, Dudley et al. [17], for example, achieved mean WPMs of 42.1 for two-finger in-air typing and 34.5 for 10-finger in-air typing on a virtual keyboard. ATK [86] explored an alternative approach and combined a relatively low fidelity consumer level finger tracker positioned directly under the fingers with the simpler task of making individual finger strokes according to a touch typing scheme. In ATK, no virtual keyboard is present; rather intended key entries are derived solely from finger motions. This approach places fewer demands on tracking fidelity and the authors report relatively rapid text entry speeds of 29 WPM.

While these systems highlight the potential of achieving rapid unencumbered text entry on HMDs, they are not currently practical. The systems reported by Dudley et al. [17] and Yi et al. [86] rely on external and/or high performance trackers able to capture the hands and fingers from angles, and with a fidelity, that is not currently achievable on HMDs. Indeed, due to occlusion issues, it is unlikely that the currently dominant inside-out approach to hand tracking on HMDs will lead to reliable monitoring of the kind of downward finger strokes used in these systems: the movements are simply too small, rapid and liable to be obscured by the hand. We argue for exploring alternative approaches to bare hand text entry on HMDs and take inspiration for this from a prevalent tactic on mobile devices: two-thumb text entry. This widely studied technique [14, 27, 54, 65] involves a palms-up two-handed grip on a mobile device and the interleaved use of both thumbs to strike keys. It is prevalent—used by up to 82% of people [66]—and data from both lab [4] and field [66] studies indicate it results in high text entry speeds of, respectively, 50.1 WPM and 38 WPM, the peak performance rates reported in both these studies. In addition, the hands-up pose provides a clear view of the thumbs from cameras mounted on a HMD. Despite this advantage, the popularity of the technique and the high performance its users achieve, we are not aware of prior work implementing thumb typing for HMDs.

This paper seeks to rectify this omission and design, develop and evaluate a thumb typing interface for the bare hand tracking system integrated into a commercially available HMD (an Oculus Quest [70]). To achieve this we present four studies. In the first, we capture performance with a large set of targets displayed in a volume above the palm. We use the results to characterize in-air thumb targeting motions and, in particular, to identify an appropriate surface on which to position keyboard keys. Inspired by common thumb-based input devices, such as game controllers, we select a design based on four keys per thumb for further study. A second study captures performance on all possible pairs of these keys—a sequential input task mimicking the interleaved input that occurs during typing. We use this data to design a accurate key selection procedure and an ambiguous keyboard layout that balances maintaining QWERTY-similarity [7] with achieving reasonable word disambiguation performance [29]. A third study evaluates this layout in both a word repetition text entry task, intended to simulate expert performance, and a more realistic phrase entry task. We report final WPMs of 27.1 and 13.73 and final error rates of 3.31% (Uncorrected Error Rate) and 10.1% (Corrected Error Rate) for word repetition and phrase entry, respectively. We complement these results with a final study that compares novice performance with ThumbAir against a baseline freehand text entry technique that relies on index finger taps to an in-air QWERTY keyboard [75]. The results suggest that while initial trials with ThumbAir, when users are familiarizing themselves with the layout, may be slower than this baseline, this effect rapidly diminishes: after entering just 15 phrases mean WPMs (at 11.24 and 12.09) did not significantly differ between the conditions. In addition, ThumbAir led to reduced levels of perceived exertion ("gorilla arm" [34]) and was viewed as more discreet and social acceptable than the baseline.

The contributions of this paper are two-fold. Firstly, the design of a bi-manual, unencumbered thumb-based text entry system for a commercially available HMD. This data driven process includes a description of the viable volume for locating thumb targets during bare hand HMD input tasks, a characterization of performance during a sequential input task with a selected subset of targets located in this viable volume, the design of a system to accurately detect target selections and the final computationally designed keyboard layout. The second contribution is two evaluations of the keyboard design and layout in repetition (once) and phrase based (twice) text entry tasks. These studies demonstrate that our keyboard design enables users to achieve rapid, accurate, comfortable and socially acceptable bare hand text entry using the tracking systems available in a current consumer HMD.

2 RELATED WORK

2.1 Text-entry on HMDs

The challenge of typing on AR/VR HMDs has provoked the design of numerous text entry systems. As many headsets ship with a hand-held controller, early approaches sought to re-purpose the input capabilities of this device. However, as such devices are predominantly oriented towards pointing input, performance of the resultant systems was somewhat limited. For example, Jones et al. [41] proposed accelerometer-based gesture typing and achieved 5.4 WPM while Shoemaker et al. [72] used a ray-casting technique and achieved performance levels of 10.1-14.5 WPM. Error rates in both these projects remained high. More recent projects have sought to integrate traditional text entry devices and touch typing input styles into HMD scenarios, most commonly by combining a real physical keyboard with a virtual counterpart [30, 59, 60, 67, 79]. While this taps into users' physical typing skills and supports rapid text entry (WPMs of between 26.3 and 43.7), it essentially tethers the user to the location of the physical keyboard, typically a desk. One way of avoiding this limitation is to sense natural typing behaviors against arbitrary surfaces. Zhang et al. [90], for example, used a pair of commercial wearable devices composed of five inter-linked motion sensing finger-rings to track finger impacts on a range of surfaces. Participants operating this system achieved very rapid text entry speeds (of 70.6 WPM) by leveraging their existing typing skills. While this is impressive, it requires users to don two additional wearable devices. Recognizing that managing this additional equipment may be impractical for many HMD users (and use scenarios), other authors have designed systems that rely on sensors or surfaces built-in to the HMD itself. Numerous modalities have been proposed, such as head rotations (13.24 WPM for expert users at the end of a longitudinal study [83]), head gestures (24.73 WPM [87]) or eye-gaze assisted touch input (11.05 WPM [3]). While performance in these systems can be rapid, they suffer from the disadvantage that they dominate visual attention—in contrast to regular typing activity on a keyboard or touch screen, users must visually attend to the input surface, a typically undesirable property that precludes focusing on the written text, or anything else, while typing.

2.2 In-air Typing on HMDs

One approach that has the potential to achieve a more familiar experience is in-air typing. A very wide variety of designs have been proposed in this space. For example, Gupta et al. [32] captured the orientation of a finger ring to make input on a novel rotary keyboard layout, ultimately achieving expert performance levels of 14 WPM. A more typical design involves a QWERTY-like virtual keyboard—a grid of targets that a user strikes with one [18, 76], two [17, 75], or more fingers [17, 85, 86] to enter text. While such systems can be effective, yielding between 9.8 [75] and 42.1 WPM [17], they typically do so by either relying on high performance external optical trackers [17, 85] or cumbersome and potentially uncomfortable hand poses that clearly expose the tapping fingers to an HMD based sensor. For example, Sun et al. [76] describe a system based on index finger strokes that take place directly in front on an HMD's cameras—short stabbing motions towards the face at eye height. A more common design has been to present a virtual keyboard in front of a user, within ready reach, and require they strike the keys with one [18] or both [17, 75] of their index fingers. While this pose offers advantages, such as its resemblance to typing on a vertically mounted touchscreen, it requires maintenance of a hands-raised posture that is widely acknowledged to cause cumulative arm fatigue (or gorilla arm [34, 38]), a problem that that may ultimately restrict its practicality. In addition, while limited work has examined the social acceptability of HMD input methods [47], numerous authors have remarked that a pose entailing the arms raised and directly in front of the body may be considered awkward and undesirable [2]. As such, users may be reluctant to adopt it in a wide range of public and semi-public settings. These concerns may additionally serve to constrain the practicality of such approaches and highlight a need to develop in-air text entry methods that rely on input actions that are less strenuous and more discreet.

One approach to meet these requirements has been to capture performance with in-air typing tasks in the absence of explicit keyboards. The aim in such systems is to capture natural typing finger motions and infer intended targets. While the results suggest this can be highly effective—from 29.2 WPM in a working system [86] up to a possible maximum of 49.1 WPM if user performance is fully unconstrained [25]—the high fidelity of finger tracking required again necessitates the use of external trackers. As such, while in-air typing is a promising approach to HMD based text entry we note that current solutions are not yet practical. They are either dependant on external trackers, meaning they cannot be used standalone, or require uncomfortable, and potentially strenuous, artificial poses to ensure clear input. As such, we highlight a need for more research to identify viable in-air typing schemes that can work with current inside-out HMD based tracking systems and enable users to adopt physically comfortable and socially acceptable input poses.

2.3 Thumb Typing on HMDs

One potential solution is to use thumb input. While we are not aware of any prior implementations of in-air thumb typing for HMDs, numerous authors have explored the idea using physical controllers. For example, PizzaText [88] used the thumbsticks of a game controller to achieve a mean WPM of 13.77 while systems based on thumb-trackpads [39], trackpads-plus-hover [73], or a held smartphone [58] show peak performance levels of between 13.57 and 29.91 (after extended training). Fully wearable systems are also common. Designs include drawing gestures over the skin of the fingers [36] and thumb-to-finger taps [81] on the finger phalanxes [80] or nails [45]. While high levels of performance are frequently reported (e.g., up to 16 WPM in phrase typing task [80] and 31.3 WPM in word repetition task [45]), such systems invariably rely on elaborate wearable hardware such as touch sensing gloves [40, 49] or finger nails [45]. As such, while they highlight the strong potential of thumb based text entry for HMDs, these systems are lab prototypes that cannot be implemented on current commercial platforms.

Inspired by prior work on both in-air input and thumb-typing, the remainder of this paper seeks to implement a system that combines them to achieve a comfortable, discreet, and efficient HMD based text entry system. A key additional goal is to design a practical system that can be implemented for the inside-out tracker on a commercial HMD. The remainder of this paper describes our steps towards these goals.

3 PLATFORM AND ENVIRONMENT

All work in this paper was conducted using an Oculus Quest VR HMD. We note VR HMDs are often used to design and prototype interactions intended for both AR and VR scenarios [19, 37, 87] due to the greater maturity of the VR product category. This device features an advanced ”inside-out” bare hand tracking system capable of capturing finger and thumb movements without additional hardware [70]. More specifically, it features a camera-based multi-stage hand tracking pipeline that implements processes of hand detection, hand keypoint identification, and model-based tracking [1]. It has an average finger joint angle error of 9.6° and an average temporal delay of 38ms [1]. In all studies, we configured both the frame rate of HMD display and the hand tracking system to be 60Hz. All studies and applications were developed using Unity3D and the Oculus Quest SDK. The size of virtual hand used in all work reported in this paper was the default size specified in the Oculus Quest SDK. Virtual hands did not vary in size by participant. To best support this device, we minimized variability in tracking performance due to environmental issues by conducting all work in the same laboratory room and ensuring it was always lit solely by constant fluorescent lighting. Windows were blocked by blackout curtains. Finally, we note that all studies reported in this paper were approved by the local IRB and conducted in full compliance with all national and institutional rules and recommendations relating to social distancing.

4 RANGE OF MOTION STUDY

Key location is a critical aspect of a keyboard design—a comfortable, effective keyboard is composed of keys in easy and ready reach. While numerous authors have conducted studies to establish appropriate key locations in scenarios such as touch typing on a tablet [65], thumb typing on an index finger [84] or even during in-air touch typing [86], no prior work has explored this issue for in-air thumb typing. There is reason to suspect performance from other settings will generalize to an in-air experience—the specific open-hand, palm-up we use, the complex articulation of thumb movement and the lack haptic feedback inherent during in-air input will likely create a unique performance profile [44]. To address this omission, we therefore conducted a study to establish thumb input performance over a large set of in-air targets. The goals are to improve our understanding of in-air thumb targeting motions and support the selection of a subset of target locations that support fast, accurate and reliable user performance.

4.1 Methods

4.1.1 Participants. Sixteen participants (mean age 23.8 (SD=3.7), eight female, eight male) completed this study. They were screened for right-handedness. They rated their familiarity with computers (5.0/5.0) and smartphones (5.0/5.0) as high, and with HMDs (1.94/5.0, SD=0.93) as relatively low. The study took approximately 50 minutes to complete and participants were compensated with the equivalent of 15 USD in local currency.

4.1.2 Study Design. The study was descriptive in design: all participants completed a single condition—tapping in-air targets with their thumbs—in which we sought to characterize performance. Participants were required to hold their palm up, with thumbs clearly visible to the HMD tracking system. We arranged a grid of targets above each palm with the goal of densely populating the full range of locations comfortably in reach of the thumbs. Through a process of subjective experimentation, we first defined a viable volume for these targets with respect to both center of the palm and the fixed virtual hand size, as reported by the Oculus SDK. This volume ultimately covered a range across the palm that spanned the region underneath the ring, middle and index fingers, and a range along the palm from the base of the thumb to the base of the fingers. Due to discomfort experienced when trying to actually touch the palm, we ensured the lowest targets were situated 2cm above the palm and defined

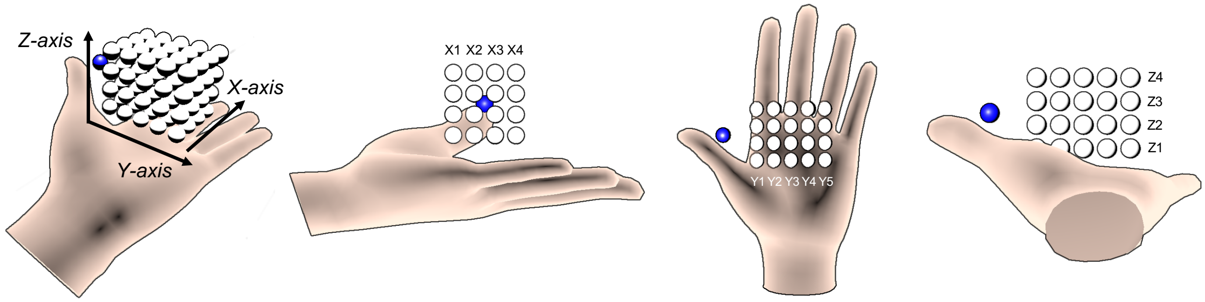

Fig. 1. Target locations with x, y, and z axes marked in the range of motion study. The blue sphere is thumb starting point.

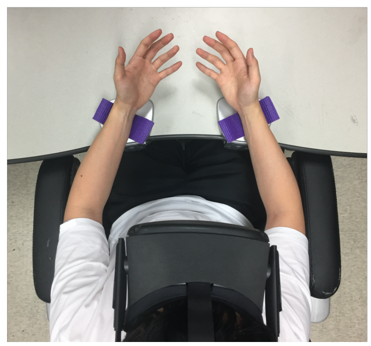

Fig. 2. Participant in the range of motion study. Participants selected a comfortable palm-up hand pose and used wrist supports (in blue) to avoid fatigue.

the limit for highest targets based on an assessment of the vertical reach of the thumb above the palm. Based on these constraints we ultimately selected a volume that was 6cm by 4.75cm by 4.75cm, centered 3.875 cm above the palm, precisely centered across the palm and offset 1.25cm forward, towards the the finger tips. We then selected a typical target shape (sphere), size (1cm diameter) and spacing (0.25cm) and uniformly populated this volume. This resulted in a five by four by four grid of targets positioned relative to the center of the each palm: 80 targets per hand, 160 in total. In addition to this layout, we specified a start point for each thumb motion. This was 2.5cm beyond the outer edge of the grid of targets, aligned with the grid's center. The targets and thumb starting position are illustrated in Figure 1.

Each trial in the study started with a blue highlight over the left or right thumb start point. After moving their corresponding thumb to highlighted location, participants were required to dwell for 0.5 seconds, during which time the hand model was greyed out. The hand model was then returned to skin color, one of the grid targets was displayed in red and the participant's task was to move their thumb quickly and accurately to intersect it. No distractor targets were shown during this task. Trials were arranged in blocks of 160: one unique occurrence of each target. Trials in each block were delivered in a random order and the study consisted of six blocks, the first of which was discarded as practice. In this way, each we retained 12,800 trials (16 participants by 5 blocks by 160 trials) over the study. Each trial timed-out after a maximum of three seconds and we logged the success rate (whether or not the participant touched the target), the touch time (measured from initial display of the target until first contact) and thumb velocity during this period. Additionally, we recorded global changes in wrist angle

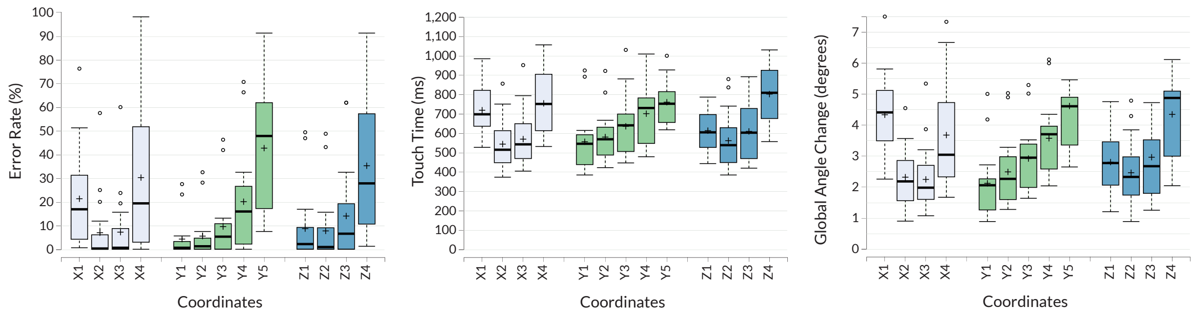

Fig. 3. Error rate, touch time, and global changes in wrist angle in wrist for range of thumb motion study. Data is combined from both left and right hands.

in order to estimate difficulty—the intuition being that increased wrist motions correspond to more challenging targeting movements [9].

4.1.3 Procedure. The study took place in an empty office lit solely by uniform fluorescent lighting. Participants were seated in front of a desk. They first read study instructions, and completed consent and basic demographics forms. They then viewed a video clip depicting the study procedures and had the opportunity to ask any questions to an experimenter. After indicating they clearly understood the study task, participants adjusted the position of two wrist supporters mounted on the desk to achieve what they considered to be a comfortable palm-up hand pose (see Figure 2). We used these wrist supporters to minimize the impact of “gorilla arm” [34], or upper limb fatigue, during the course of this quite intensive, repetitive and prolonged study. Next, participants donned and adjusted the HMD, then returned their wrists to the desk-mounted supports and began the study trials. There was an enforced rest of three minutes between blocks. At the end of the study, an experimenter measured their thumb size.

4.2 Results and Discussion

We first examined mean data for time, errors and wrist angle change. We conducted paired t-tests for each measure to examine performance variations due to the handedness. No differences between dominant and non-dominant hands were observed (the lowest p value was 0.225), so we combined data between hands for all subsequent analysis. The resultant mean data for these measures are shown in Figure 3, organized to depict variations along targets in the x, y, and z axes. The overall mean success rate per target was 83.6% (SD 23.3%), indicating that many targets were hard to reach. This was expected: the target set was designed to clearly document the viable range of thumb motion by extending beyond it into more challenging territory. For those target selections that were successful, the mean touch time was 646ms (SD 179ms). This indicates that the in-air tapping task could be executed relatively rapidly. For example, prior thumb targeting data reports targeting times of between 560ms [61] on a mobile phone and 1500ms [46] on the fingertips. This high performance was enabled by a rapid (and not plotted, as it was highly uniform) mean thumb velocity of 0.138m/s (SD=0.028m/s) during targeting motions. This suggests that one factor supporting the rapid performance we observe was that participants took advantage of the purely virtual nature of the targets to approach at high speed, pass through the targets and decelerate in the space beyond them. We note that while this strategy is effective in the single target task studied here, it may be less useful in a more complex task, such as typing, that involves a sequence of selections. In such a task, an extended targeting motion for one target may make a subsequent targeting motion slower. In addition, if targets are densely arranged, such a strategy may result in intersections with multiple targets. This result suggests that

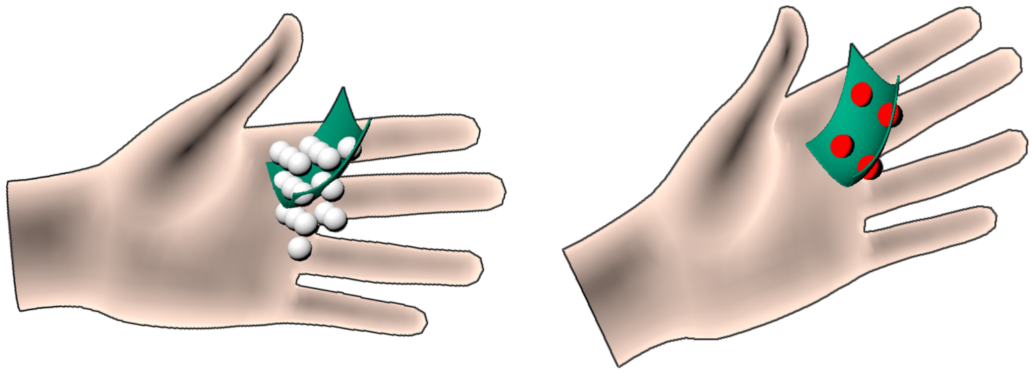

Fig. 4. Surface (shown in green) derived from the 18 selected targets (shown in white, left image) in the range of motion study. The key locations (shown in red, right image) used in bigram study.

an in-air thumb keyboard will need to be carefully design to prevent such inadvertent activations. We opted not to analyze this data statistically, as formally establishing the presence of performance variations would not serve our objectives. Finally, we note that the measure of wrist angle change (Mean = 3.14°, SD=1.66°), which was intended to capture subtler variations in the difficulty of different targeting motions, closely followed the trends in the time and error data (with Pearson correlations of between 0.831 and 0.995). This suggests these traditional metrics were able to effectively capture key trends in user performance.

We also examined the relationship between the thumb length (the distance from the end of thumb to the tip of thumb) and time and error data. The mean thumb length among our participants was 5.59cm (SD=0.39cm, Min=4.9cm, Max=6.4cm). However, Pearson correlations indicated that variations in thumb length were not linked to either success rate ($r = -0.223$, $p = 0.406$) or touch time ($r = 0.325$, $p = 0.219$). This suggests that participants with different thumb lengths were able to use the system with equal effectiveness.

4.3 Selecting Target Locations

To determine appropriate target locations, we first created a single *performance score* for each target. This was based on normalized scores for success rate, touch time, and wrist angle; thumb velocity data was excluded as it showed highly limited variability. Data for each of these three metrics was first inverted to ensure larger scores corresponded to more desirable performance (i.e., higher success rate, lower times, and reduced hand rotation). The *performance score* for each target was then calculated as the magnitude of the vector composed of these three features. We then selected the top 20 scoring targets for each hand (25%) and then, as we intend to support bi-manual input, filtered these by those that appear in both hand's data sets. Ultimately, we retained 36 targets, or 18 on each hand. These are shown in Figure 4 and their coordinates, and the normalized scores recorded for all metrics, are reported in Appendix A. They were arranged in a tapered wedge aligned with the movement of the thumb tip as it rotates around the metacarpal joint. In this set, targets above the center or ulnar (little finger) side of the palm were physically occluded by those on the radial (thumb) side—to reach targets over the ulnar region, the thumb would need to pass through targets above the radial region. As such, many of these target locations were incompatible with one another.

In order to establish a more limited set of possibilities for final target locations that would avoid issues of physical occlusion, we fit a paraboloid to the positions of the full set of targets (R^2=0.154), aligned and clipped so that it faced the thumb of each hand, maintained rectilinear edges and did not extend beyond the range of the 18 original targets. This surface specifies an appropriate set of locations on which to situate targets that will be

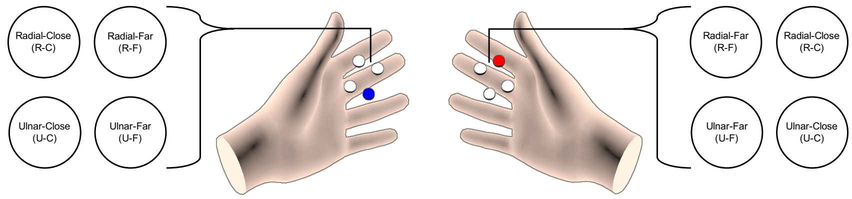

Fig. 5. Study task in the Bigram study showing both hands and all eight keys. Participants were required to touch red target first and blue target second. Annotations indicate naming conventions for the targets and were not shown during the study.

rapidly and accurately in reach. We located candidate targets on this surface by considering our design goals of supporting quick, error (and inadvertent activation) free targeting performance on an ambiguous keyboard. As performance with larger numbers of targets will inevitably reduce performance we opted to locate four targets on each hand or a total of eight over both hands—eight targets are sufficient to support a wide range of ambiguous keyboard designs [45, 81]. We located these four targets by the simple expedient of dividing the paraboloid surface into four quarters, and placing a target at the center of each quarter. We note this four target design bares similarity to common input devices such as directional thumb joy-pads. We believe the familiarity of this arrangement may help to reduce novelty effects with our system. In addition, we validated these target locations with respect to the reported sensing resolution of the Quest HMD used in this work. Measured from the trapeziometacarpal joint of the thumb, the angular gaps between the four targets range from 17.0° to 35.2°, comfortably exceeding the Quest's 9.6° joint angle accuracy [1]. This suggests the performance of the hand tracking system used in this work should not overly impact or constrain the performance participants can achieve when selecting these targets.

5 BIGRAM STUDY

We conducted a second study to complement the data captured in our first study. It had two objectives. First, it captured data in a sequential input task. This is important because ambiguous keyboards [28, 43, 45, 49, 50] feature relatively few keys (in the range of between 5 and 10), each of which is mapped to multiple characters. They take advantage of the fact that the vast majority of possible character sequences do not represent valid words to create accurate and effective text entry systems—although individual key entries may be highly ambiguous, the vast majority of words can be precisely specified by appropriate sequences of ambiguous selections. However, in order to map characters to keys, a process known as the letter assignment problem [20], in arrangements that support good user performance, data about single input events, such as that gathered in first study reported in this paper, is insufficient. Rather, typing is a continuous task and data about performance of sequential input actions—how users select targets one after the other—is required. This study aimed to capture such data. In addition, we also sought to characterise the distribution of thumb motions made during targeting. Data of this sort has previously been captured for touch screen keyboards and used for a range of purposes, such as understanding typing behaviors [22] and optimizing or customizing [21, 26] key locations and sizes. We captured such data in our study to support the similar objectives: to use it to adjust our target locations and/or target selection process in order to improve user performance.

5.1 Methods

5.1.1 Participants. Sixteen new participants (right-handed, mean age 25.8 (SD=4.1), seven female, nine male) completed this study. They again rated themselves as highly familiar with computers (5.0/5.0) and smartphones

(5.0/5.0) but with relatively low familiarity with HMDs (1.94/5.0, SD=0.93). The study took approximately 50 minutes to complete and participants were compensated with the equivalent of 15 USD in local currency.

5.1.2 Study Design. The study again featured a single condition. It had the goal of characterizing user performance in the task of sequential in-air thumb taps. The candidate layout selected from the results of the first study features eight 1cm diameter targets, so this study used the full set of 64 possible sequential pair-touches (see Figure 5). The study was again designed with six blocks, each of which featured a single randomly ordered occurrence of each possible pair-touch, with the first two blocks discarded as practice.

To characterize thumb motions, we opted for the simple expedient of logging the area of all thumb contact with the paraboloid surface during each targeting task. This approach is directly analogous to closely related prior work on touch screen typing [21, 22] which logs all finger contact points during key presses. To achieve this we created a grid of 0.15cm diameter transparent spherical colliders (15x16) distributed evenly over the paraboloid surface and spaced at 0.15cm intervals. Effectively, this formed a solid mesh of separate colliders. As the visually displayed targets were spherical, we offset this collider mesh towards the thumb and along the paraboloid's normal vector by 0.5cm—the radius of the visually presented targets. This ensured that we captured data from even shallow contact with the displayed targets. In addition, to log full data from each touch we considered each touch to start and end when the thumb made and released contact with the collider grid—a style of input analogous to interacting with a touch screen. To ensure that we captured the most diverse set of valid thumb motions toward each target we treated any touch that contacted the correct target to be correct, irrespective of whether or not it came into contact with other targets beforehand or afterwards. As such errors were recorded only if participants failed to make contact with one or both of the correct targets in a trial in the correct order. In addition, we recorded a timeout if no contact with any target occurred after five seconds. To ensure we logged a complete set of data for each target-pair, error and timeout trials were repeated in a new random order. In this way, the study retained data from 4096 successful trials (4 blocks by 64 pairs by 16 participants), with each trial composed of a pair of target selections.

Trials in the study were broadly similar to the those in the first study. Each trial started by tapping a start button, following by a fixation period during which time the hand model and all eight targets were greyed out. After two seconds, the hand model was re-colored and two targets were highlighted - red for the first touch and blue for the second touch. If the first and second touch were due on the same key, participants were required to tap this key twice. When touching one of the targets, feedback was provided by applying a green highlight.

5.1.3 Procedure. The procedure broadly followed the first study. It was completed in the same environment, and participants completed similar consent and demographic forms. Instructions were again read and exemplified via a video. To increase ecological validity, and due to the reduced fatigue expected during the somewhat shorter data collection period in this study, participants did not use the wrist supporters in this study. Instead, they adopted a comfortable hands-free, palm-up, in-air posture of their own choosing. Prior to the main study, participant's completed an informal practice session (max five minutes) in which the eight keys were shown and they were able to practice touching them with their thumbs. Participants then began the main study. A break of three minutes was enforced between each block and participants were additionally able to rest between each trial if needed. After completing the study, an experimenter measured their thumb size.

5.2 Results and Discussion

We recorded a total of 4223 trials, including 4096 successful trials and 127 failure trials in which the participant did not touch the specified targets in the specified order. From this set of trial failures, we excluded the 26 trials involving selection of a first target on the wrong hand—these failures likely represented confusion with regards to the trial instructions rather than performance of the actual study task. Similarly, we excluded the 13 timeout trials,

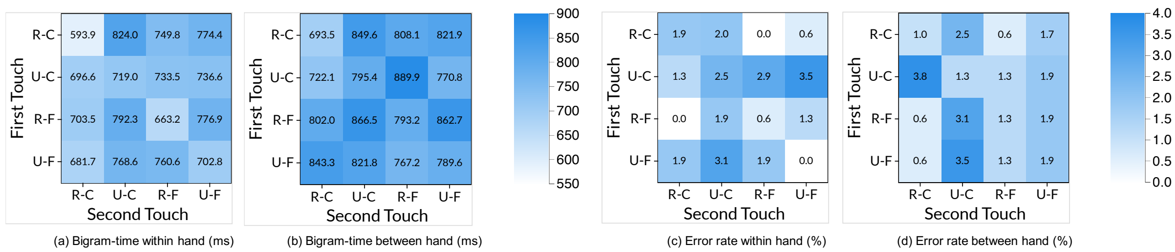

Fig. 6. Bigram-time and error rate divided according to trials in which both targets were on the same hand (within hand) and those in which they were on opposite hands (between hands). Targets are described by their position: R(adial) or U(lnar) and F(ar) or C(lose). See Figure 5 for an illustration of this mapping

thus retaining a total 88 errors for analysis. Based on this data set, we first examined the impact of participant thumb length (Mean 6.01cm, SD=0.47cm, Min=5.5cm, Max=7.0cm) on performance using Pearson correlations, finding no relationship with either bigram-time (p=0.465 with r=0.197) or error rate (p=0.365 with r=0.243). As in the first study, participants' performance was not linked to the size of their physical thumbs.

We then plotted the data—Figure 6 shows confusion matrices showing the mean bi-gram time and error rate for all possible pairs of touches. We first analysed this data by conducting two-way RM ANOVAs on the measures of bigram-time and error rate using the independent variables of first-touch (left or right hand) and second-touch (same or different hand). These tests sought to establish whether there were performance variations due to handedness or the possibility of interleaving the target selections—of overlapping input in trials in which it was bi-manually split between the hands. As with the range of motion study, we did not detect any differences in performance between dominant and non-dominant hands in either time or errors, suggesting that handedness did not exert an effect. This result is positive—users should be able to use a bi-manual system effectively from the outset. Based on this result we combined data from both the hands during all subsequent analysis. However, while there were no differences in terms of errors, we did observe a single significant difference for the second-touch variable: a main effect of bigram-time (F(1, 60)=4.1, p=0.047). This indicates that two touches on the same hand showed were performed faster (730ms, SD=134ms) than a pair of touches composed of a single touch to each hand (806ms, SD=162ms). This result suggests that participants were unable to effectively interleave input between their hands. A potential explanation for this is that they may have needed to direct visual attention to each individual target selection task and faced challenges with shifting visual attention to a new target on different hand in a short period of time—due to a lack of familiarity with the in-air input task, participants may have visually monitored each thumb movement, in effect placing a central processing limit on performance.

The overall error rate in this study was a mean of 1.67% (SD=1.47%) over all target pairs. This high level of performance indicates all targets were in easy reach of the thumb. While this figure is inarguably low, we note it also applies an unrealistic decision criteria for determining target selection: any contact with the specified target, irrespective of whether it proceeds or succeeds contact with another target, is considered to be a successful trial. Use of this criteria enabled collection of the most diverse set of valid targeting motions possible. However, it cannot be implemented in a realistic system as intended targets are not known in advance. Accordingly, we opted to explore error rates with more realistic and unambiguous decision criteria. Specifically, we examined error rates based on treating the initially touched target as the intended selection and those on treating the last touched target as the intended selection—conceptually equivalent to touch-down and touch-up events. The error rates calculated according to these criteria were substantially elevated at means of 9.01% (SD=6.17%) and 20.89% (SD=9.54%), respectively. These increases are likely due to a number of factors. Firstly, physical occlusion of the

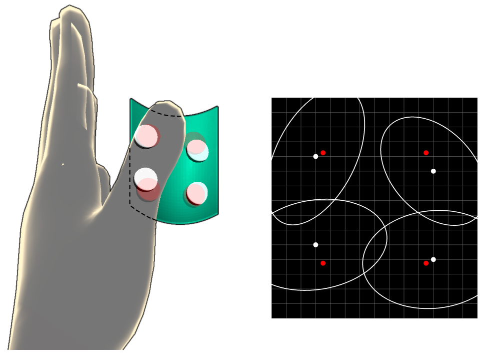

Fig. 7. Four white keys for text entry study are the refined target locations (left hand) with the mean target centroid positions for each key. Four red keys are the previous target locations used in bigram study.

intended target by other targets. For example, of ulnar targets by radial ones. Such occlusion would result in the thumb path intersecting an unintended target on its way to or from the intended target. Secondly, overshooting an intended target and subsequently colliding with an unintended target. A final factor is doubtless the definition of errors used in the study. As we did not penalize contact with non-targets, participants did not explicitly seek to minimize these collisions. While a study applying a stricter error criteria would likely reduce the prevalence of unintended collisions we note that it would also represent a fundamentally more difficult task, in which the burden of avoiding unintentional collisions is placed entirely on the user. This would likely lead to elevated errors and time data. Based on these observations, we determined that developing a technique to accurately and unambiguously detect intended target selections from user's thumb motions is a key next step for this work.

To address this problem, we turned to the touch profile contact data recorded from the collider mesh. As we previously noted no performance differences dues to handedness, we first collapsed data between the hands. From each trial we then aggregated data from all frames in which the thumb made contact with the mesh, ultimately creating a 15x16 pixel binary image (a *collision map*) representing the full set of these intersections. We then explored the design of machine learning classifiers capable of determining the intended target from the collision maps. To build these models we constructed neural networks using PyTorch. We selected neural network models as they can process both raw image data and other forms of numerical data (e.g., summary statistics derived from the images). In addition PyTorch neural networks can run naively in Unity and on the Quest headset when converted to an Open Neural Network Exchange (ONNX) format. To construct classifiers we used Leave-One-Out cross validation (LOOCV), an approach in which each participant's data serves as the test set for a model built on all other participant's data. This reduces the risk of over-fitting. Other aspects of the models and training procedures were customized to match the features we explored.

Specifically, we examined performance with two feature sets. First, we calculated image moments for each collision map. We used these to extract the centroids [22, 24] of the touched regions. We designed a simple model for this data comprised of an input layer, a hidden layer, and a softmax output layer, all fully-connected using linear transformation. To support multi-label classification, we used Adaptive Moment Estimation (ADAM) as an optimizer and Cross-Entropy-Loss as the loss function. When training the model, we avoided under or over-fitting by monitoring changing trends in training and validation loss per epoch. This model ultimately attained a LOOCV classification accuracy of 97.6% (SD=1.8%). Second, we used the raw collision maps. We first down-sampled these using linear interpolation to create a 3x4 grey-scale image. We then constructed a simple model composed of one layer for linear transformation. We used Stochastic Gradient Descent (SGD) as an optimizer and Cross-Entropy-Loss as a loss function. This model achieved an LOOCV accuracy of 98.2% (SD=1.5%).

Results from both classifiers are strong, reporting accuracies close to the original 98.33% recorded in the study. This suggests that contact with non-intended targets is, in general, brief and/or slight and does not preclude reliable inference of intended targets from thumb motions. Based on these results, we made two revisions to the target layout and selection process used in further studies in this paper. First, we opted to use the collision maps and a version of the image-based classifier built using the full set of user data to determine selected targets. This is due to modestly higher accuracy and lower variability in the LOOCV results compared to the centroid based classifier. One implication of this choice is that target selections are calculated on release of the collider grid (an event equivalent to finger-up on a touch screen) rather than in relation to contact with visually displayed targets. Second, we refined the location of the visual targets. We achieved this by calculating all ellipses from the collision maps for each target and using these to calculate the mean target centroid position. These positions represent, on aggregate, the locations participant's actually moved towards during trials. To reflect these motions, we placed the final target locations at each of these mean positions. This process follows prior work identifying optimal key locations for touchscreen typing [21]. The revised target locations are illustrated in Figure 7 and reported in Appendix A.

In sum, this study recorded mean bigram-times and error rates that that were were stable across both hands and showed relatively minor advantages for uni-manual input over bi-manual input. While error rates are somewhat high using simple (touch-down/touch-up) selection criteria, we show how simple classifiers can be used to dramatically improve these. It is worth contextualizing this data. In a broadly similar study involving pairs of thumb (and finger) taps to the nails of the same hand, Lee et al. [45] report mean bigram-times of approximately 550ms and error rates of approximately 4.5%. While our times are somewhat elevated compared to this prior study (at 768ms, SD=74ms), error rates from our classifiers are notably reduced. We argue the results are strong enough to validate our refined target locations and target selection procedures and support further investigation into their ability to support effective in-air thumb typing.

6 KEYBOARD LAYOUT SELECTION

We used the data from the Bigram study to inform a keyboard layout design process. The goal was to map characters to the eight keys in our system in such a way that they can support rapid, accurate, unambiguous and familiar input. We achieved this through three mechanisms. First we specified a limited set of layouts that retain a strong similarity to QWERTY. Second, we defined four metrics and calculated these for each candidate layout. Next we reviewed the layouts and filtered them based the minimum and maximum number of characters assigned to each key. We then reviewed their performance on all metrics and selected a balanced candidate with a strong performance profile for further study. These processes are described and defined in the following sections.

6.1 Layout Constraints

We restricted the character layouts to closely resemble QWERTY. We achieved this by considering our bi-manual eight key layout as two rows of four keys. The top row of QWERTY characters was mapped to the top four keys, the bottom row of QWERTY was mapped to the bottom four keys and the middle row of QWERTY was assigned to either top or bottom keys. We also restricted column assignments by stipulating each row of QWERTY characters was divided as equally as possible over the four key columns—there could be either two or three characters per key for the top and middle QWERTY rows and one or two keys for the bottom QWERTY row. We then laid out all possible character to key assignments, following the QWERTY layout and respecting the possible sequences of character to key counts. For example, 'q' and 'w' could only be assigned to the top left key while 'e' could be assigned to either the top-left key, or the next key over. We determined all possible key assignments

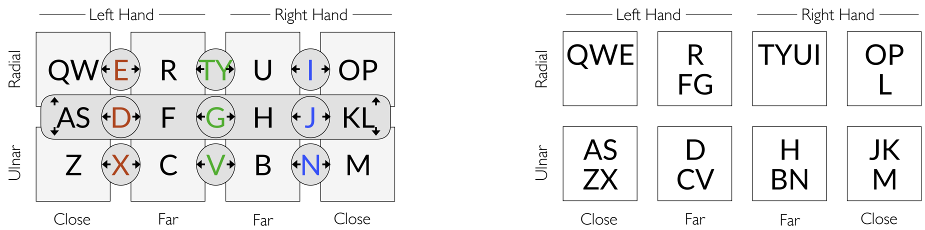

Fig. 8. Keyboard layouts. Left shows the character to key mappings considered in the layout selection process. The eight keys are shown as grey squares, overlaid with the characters that could be assigned to them. Characters in black have a fixed key column. Characters in red could appear on the first or second column of keys (left hand); in green on the second or third columns (between the hands) and in blue between the third and fourth columns (right hand). In addition, while top row and bottom row QWERTY characters always appeared on, respectively, the top and bottom row of keys, all characters on the middle QWERTY row could appear on either the top or bottom row of keys. This limited set of variations was designed to ensure all layouts considered retain a close similarity to QWERTY. Right shows the final key layout selected for further study.

according to these constraints (illustrated shown in Figure 8, left) and then calculated all possible layouts: 524288 in total. Following prior work dealing with such restricted sets, we opted to calculate performance metrics for all layouts [84]. The number of candidate layouts is sufficiently small that a computational optimization process [45, 49] is not necessary.

6.2 Metrics

6.2.1 Speed. We combined our empirically captured bigram entry times with word frequencies from Norvig [62] to calculate the text entry speed for a given layout using the standard quadratic formulation used for the letter assignment problem [20]. This calculates, for a given layout, a single measure that expresses how quickly text can be entered.

6.2.2 Accuracy. Accuracy was defined similarly to speed. We used the success rate of each bigram entry, together with Norvig [62]'s word frequencies, to calculate a measure of the accuracy of text entry on each layout.

6.2.3 Qwerty Similarity. We defined a metric to model the extent to which layouts resembled the rows and columns in standard QWERTY. Row similarity was defined as the Levenshtein distance between the characters in each row of the QWERTY layout and the order in which those characters were assigned to the keys in the layout. Column difference examined the characters in each QWERTY column (e.g., "QAZ") and calculated how many different key columns these were assigned to in each layout. We subtracted one from the value for each column to attain zero for a column arrangement than matched QWERTY. We summed these two measures to create a final metric that expressed how distant character assignment in a layout was from the rows and columns in QWERTY.

6.2.4 Confusability. A key measure for ambiguous keyboard layouts is how unambiguous they are—how uniquely sequences of key presses will accurately specify intended words. To estimate this, we calculated Lesher et al. [50]'s confusability matrices. We closely followed a recent implementation for confusability matrices [45], which we briefly review here. Confusability matrices assess ambiguity by calculating, for a given text prediction algorithm, the frequency with which all pairs of letters are mistakenly selected for each other in a given text corpus. The relative confusability of a particular key layout can then be estimated simply by summing, for each key, the matrix cells for the characters assigned to it. We created confusability matrices using Lesher et al. [50]'s k-gram algorithm (implemented via a dictionary of the thirty thousand most common words [11]) and a 322210 word

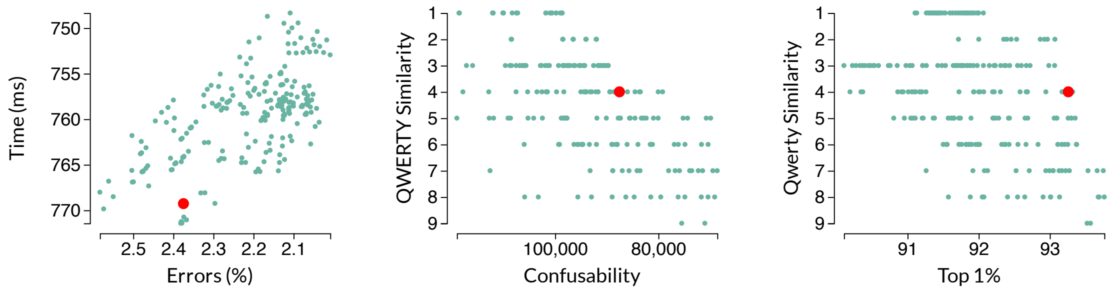

Fig. 9. Layout scores for metrics of time/error (left), QWERTY similarity/confusability (center) and QWERTY similarity/Top 1% (right). Scales adjusted to show more desirable scores to the top-right. To facilitate illustration we show only the non-dominated solutions (i.e., the Pareto sets). Scores for the selected layout are manually highlighted in red. Note that due to the highly limited range of values calculated for time and error metrics (left), we did not consider them during layout selection.

corpus formed by combination of three mobile text entry data sets [63, 77, 78]. The confusability scores generated from these matrices are quick to calculate and effectively capture the relatively ambiguity of each layout.

6.3 Filtering and Selection

In order to facilitate layout selection, we explored methods to filter them. We first removed layouts with less than three or greater than four characters assigned to any key. The intuition here was that a relatively even distribution of characters would likely yield strong layout candidates (i.e., by avoid overloaded keys with many characters and high ambiguity) while also helping ensure the geometric arrangements of QWERTY were maintained. This led to a set of 10782 layouts. We then reviewed the raw data for metrics. Speed and accuracy showed highly limited variability (748-780ms and 2.0%-2.8%, respectively) over the full set of layouts; Pareto sets for these data are shown in Figure 9. The small ranges for these two variables likely reflects the fact that we examined a relatively small subset of possible layouts that all closely resembled QWERTY and, thus, all showed similar empirical performance profiles. As such we judged these two variables to have limited impact and did not consider them during layout selection. Due to challenges in contextualizing the confusability scores (68597-156939), we additionally calculated Gong et al. [28]'s disambiguation scores for all layouts. These scores express, for a given text corpus, set of word frequencies, number of character inputs, number of predicted words and key layout, the mean frequency with which all words in the corpus will be correctly predicted. This is a very practical measurement of how a given layout will behave—it estimates, as a probability, how often a user's inputs will result in prediction of the word they intend. We calculated disambiguation scores, assuming input of three characters, for most frequently predicted word (Top 1%). Figure 9 shows the distribution of scores in the non-dominated set of layouts for the metrics of QWERTY similarity, confusability and Top 1% scores. We note Top 1% scores were strongly correlated with the original confusability scores ($r(10782) = -0.748$). This indicates the measures assessed similar properties.

Based on these data, we then selected a balanced candidate attaining a good performance profile across QWERTY similarity, confusability and Top 1% score. Our goal was to support a familiar text entry experience without sacrificing the ability of the layout to unambiguously specify words. The character to key mappings in this layout are shown in Figure 8 (right) while its scores for the metrics of QWERTY similarity, confusability and Top 1% are illustrated with red highlights in Figure 9. It is worth contextualizing the Top 1% score. Notably, the whole range for this metric is elevated compared to typical examples in the literature. In Gong et al. [28]'s WrisText, for example, the Top 1% score for the layout they develop is 85.9%; for the layout selected here it is

93.3%, an increase which we believe will lead to markedly better word suggestion performance. In addition, we also calculated the Top 2% score for our proposed layout—the proportion of intended words that would be returned within the first two suggestions—and found this to be 98.1%. This suggests that a word selection interface featuring just two options would likely be highly effective when combined with our proposed layout. Based on these relatively strong results, we moved forward to further studies evaluating our key layout during actual text entry tasks.

7 TYPING STUDY: EVALUATING THE PERFORMANCE OF THUMBAIR

We conducted a study to evaluate the performance of our in-air thumb typing system. During this study we used the target locations, target selection procedure and collider mesh based target classifier discussed and defined in the second study. To provide a rounded assessment we targeted capture of both an approximation of expert performance (via a word repetition task) and novice performance (via a traditional phrase entry task).

7.1 Methods

7.1.1 Participants. Fourteen new participants (right-handed, mean age 24.5 (SD=3.3), six female, eight male) completed this study. All were non-native English speakers enrolled in a full time English language degree program. They self-rated their familiarity with computers (5.0/5.0) and smartphones (5.0/5.0) as high, but they had passing experience with HMDs (1.86/5.0, SD=0.53). They took approximately 70 minutes to complete this study and each received the equivalent of 20 USD in local currency.

7.1.2 Study Design. This study was designed to assess typing performance with the keyboard and key layout designed based on data from prior studies reported in this paper. To do this, we included two stages. Participants first completed a word repetition task [8, 45] designed to mimic expert performance. We followed closely related prior work [8, 45, 89] by using a 20 item word set ("the, and, you, that, is, in, of, know, not, they, get, have, were, are, bit, quick, fox, jumps, lazy, on") which includes all English letters and approximates monogram and bigram frequencies. Participants were presented with each word in this set in a random order and were required to enter it seven times. There was no support for correcting entered text. We recorded Words Per Minute (WPM) [56] from all entered words and calculated Minimum String Distance (MSD) error rate [74] at the character level. In this stage of the study, we collected 1960 typed words (composed of 6468 characters): seven repetitions by 20 words by 14 participants.

The second stage of the study was a phrase typing task. This more realistic typing task better reflects novice performance with the system. In this task, participants were required to correctly type a total of 40 phrases (1061 characters plus 240 spaces). We used this instruction to simulate a careful typing experience in which accuracy is emphasized over speed. Half of phrases consisted of selected ones from MacKenzie and Soukoreff [55]'s widely used set. In order to cover all English letters, the rest of phrases were composed of two pangrams repeated ten times ("the quick brown fox jumped over the lazy dog" and "pack my box with five dozen liquor jugs"), an approached borrowed from closely related prior work optimizing touch screen typing [22]. Participants were able to use basic error correction in the form of character deletion. We implemented space and delete actions via the simple gesture of touching the collider surface with the index fingers: left for delete and right for space. These actions involved simple finger flexion and did not require gross movements of the hand or wrist. In this task we logged WPM, CER (Corrected Error Rate at the character level) [74], and MSD error rate [53]. We collected 560 typed phrases (40 phrases by 14 participants) in this phase of the study.

The typing system in both study phases was identical and is shown in Figure 10. As previously, the basic scene featured the hands and the eight keyboard keys. In addition, we added a transparent object indicating the location of the collider mesh—the touch surface on which the keys were situated. This was important as the classifier based target selection system used in the study (see Section 5.2) was based on touching and releasing this mesh,

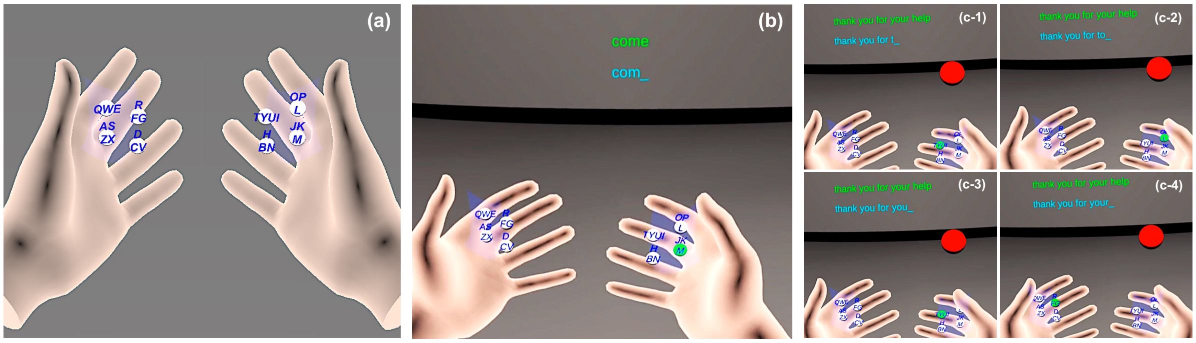

Fig. 10. ThumbAir keyboard layout (a), screenshot showing the word repetition task (b), and screenshot showing the phrase typing task (c). In (b) and (c), the trial word or phrase is displayed in green with typed text in emerald green directly underneath. In addition, (c) shows feedback from the word prediction system—although the target word is "your", a higher probability word stem ("to") is displayed after selection of the first two correct keys (top row, right). Characters assigned to each key were always orientated towards the HMD to ensure they could be clearly seen by participants. Key selection was triggered by releasing thumb from surface. The selected target was highlighted in green.

rather than the actual keys. We used this surface to convey this functionality to participants. The characters assigned to each key were shown directly over it. Each word repetition or phrase task began with display of a large (10 cm) target directly in front of the participant. During this period, the keyboard keys and hands were shown in grey. After tapping this initial target, the trial word or phrase was shown and the hand and keys were re-colored. Text entry could then begin. Key selection was triggered by releasing the thumb from the collider surface and the selected target was then shown by briefly highlighting it in green. In each trial, the word or phrase to be typed was shown in green directly in front of participants at a height of 70cm; entered text was shown in emerald green underneath this. Feedback for entered text was predictive: for each typed word, we showed the most common word or word prefix based on the currently selected sequence of keys. However, we did not implement feedback or commands to view or navigate predicted word sets. Rather we manually adjusted word frequency data to ensure all words in our test sets were displayed in response to entry of the full set of characters they were composed of. This simplified the experimental system and task and matched our intention to focus on text entry performance in the form of entering a sequence of predefined characters (i.e., a phrase from MacKenzie and Soukoreff [55]'s set) rather than the design and use of a word prediction system and interface. We note that while this manipulation ensured that correct task performance could be achieved by selecting the full and correct key sequence required to enter each word, it also provided typical dynamic word prediction feedback (including display of non-target word stems that have high probability) during typing (see Figure 10-c).

7.1.3 Procedure. The study followed similar procedures to the prior studies: it took place in same room, and followed similar instruction delivery and enrolment processes, including the use of a video to exemplify instructions and inclusion of a practice session. This session initially mimicked practice in the bigram study and involved free tapping of the thumb targets for a maximum of five minutes. Participants were then exposed to the full text entry system and provided with an explanation of the qwerty-like keyboard mapping and operation of the predictive text feedback. They then practiced the full word repetition task with five words: "my, see, run, come, there". After completing this practice session, the word repetition task began. After completing all 20 words, there was a five minute break. Participants were then introduced to the space and delete gestures and completed a practice session involving typing three phrases (not overlapped with the set in the study) from MacKenzie

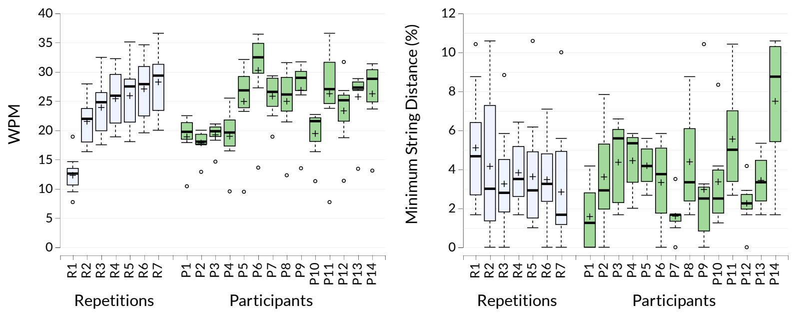

Fig. 11. WPM and MSD error rate in the word repetition task. Mean WPM was 27.1 WPM over the last three repetitions while MSD error rate was 3.31% and gradually decreased over the full set of seven repetitions. Per participant data is included to highlight variability in the word repetition task across different users.

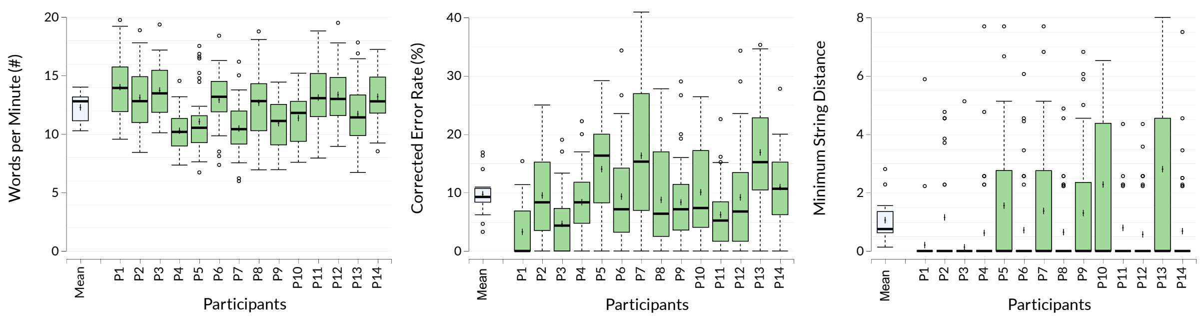

Fig. 12. WPM, CER, and MSD error rate in the phrase typing task. Mean WPMs, CER and MSD figures over the whole data were, respectively, 12.27, 9.7% and 1.06%. In addition means from the final five phrases were, respectively, 13.73, 10.1% and 1.15%. Per participant data is included to highlight variability in the phrase typing task across different users.

and Soukoreff [55]'s set. They then completed the phrase entry section of the study. As with the bigram study, participants held their arms in a comfortable, unsupported pose throughout the study.

7.2 Results and Discussion

In the word repetition stage of the study, participants completed a total of 1960 trials. Figure 11 shows the overall mean WPMs for the entered words plotted by repetition and participant. It shows a very substantial improvement between first (12.30 WPM) and second (21.5 WPM) repetition and a more gradual upward trend thereafter. This suggests the task was relatively easy to pick up—after just a single experience. In line with prior work, we estimate expert performance as the mean of the final three repetitions: 27.1 WPM. This is on par with closely related work employing word repetition tasks during thumb to finger typing [45]. The fact performance during in-air thumb typing can match that attained in systems based on directly instrumenting the hands to precisely sense physical contact is a positive result. It suggests that current HMD trackers, when combined with careful keyboard design, are of sufficient quality to support expert text entry at rates previously achievable only by instrumenting the hands and fingers with dedicated worn sensors. Figure 11 shows the MSD error rate gradually decreased over the full set of seven repetitions: from 5.1% in the first repetition to 2.8% in the final repetition. The mean of the final three repetitions is 3.31%. This suggests that, as participants grew more familiar with the task, they were able to achieve speed improvements while maintaining or improving accuracy. We note such downward trends in error rate are common in word repetition tasks [45].

In the phrase typing task, we collected data from a total of 560 trials. We excluded two trials which featured fewer than 75% of the typed phrase from further analysis, leaving data from 558 phrase entries for analysis. Figure 12 shows WPM, CER and MSD error rate data from these trials. Performance differed substantially from that in the word repetition stage. Perhaps most clearly, it was slower. The mean WPM was 12.27WPM (the mean of the final five trials: 13.73WPM), less than half that achieved in the word repetition task (27.1WPM). This suggests that the phrase task achieved our objective of capturing novice performance levels—participants were not able to fully automate their input in the phrase task and resorted to the more traditional hunt-and-peck style typing behaviors common to inexperienced users. However, there is also evidence that they benefited in this task from our work to maintain the familiar QWERTY layout. WPMs over sequentially ordered trials do not show signs of an elbow point—a linear regression on trial sequence and WPM reveals a steady upward trend (slope=0.09, R^2=75.6%). This suggests participants were successfully and incrementally adapting their knowledge of QWERTY to the new format of our layout. Based on the performance levels achieved in the word repetition task and the steady slope we observe here, we suggest further practice would result in further WPM improvements. The WPM data also stacks up well to that reported in prior systems. For example, Fashimpaur et al. [19]'s PinchType achieves 12.54 in thumb to finger typing, but requires a high performance external optical tracking system. The benefits conveyed by physical feedback are also clear however. In a thumb typing task on a VR controller surface, Son et al. [73] report WPMs of up to 20.56. We suggest that the inherent lack of haptic feedback may increase the challenges of in-air typing for novice users [42].

While WPMs showed signs of steady improvement, error data remained relatively flat over all trials, showing no substantial relationship with increased experience in the task (R^2=0.077). In terms of corrected error rate, mean performance was 9.70% (SD=3.92%) and the mean of the final five trials was 10.1%. This is a typical score for CER in novice typing performance. Prior work reports 9.75% CER [73] in text entry on handheld controllers and 13.3% CER [19] with a system based on in-air hand tracking. Furthermore, the uncorrected error rate, which we calculated as the MSD error rate, was low over all the trials (1.06%, SD=0.76%), with many participants recording a median of zero. In addition the mean over the final five trials was 1.15%. This indicates that participants performed the phrase typing task carefully throughout the study, correcting the vast majority of errors they committed.

Beyond these analysis of the two typing tasks, we also explore two further fundamental issues. First, as in prior studies, we examined the relationship between the thumb length and performance in both tasks using Pearson correlation. Thumb lengths in this study were an average of 5.81cm (SD=0.34cm, min=5.0cm, max=6.3cm). Variations in thumb length were not significantly related to any of the metrics from either study task. We once again that conclude thumb length did not impact performance: users can operate our system effectively, regardless of their hand size. Second, we explored the key level accuracy, defined as the proportion of times participants selected the correct key during the phrase typing task, to gain a more detailed understanding of how our target selection classifier performed in practice. This measure combines incorrect target selections due to cognitive errors (i.e., choosing to select the wrong key) with performative (or classifier) errors in which a user attempts to select the correct key, but fails and selects another key instead. Mean key level accuracy was 92.6% (SD=2.4%), or 5.6% lower than the accuracy reported during the LOOCV procedures used to assess the classifiers developed in the second study (98.2%). While some of these additional errors are doubtless due to failures in the classifier, we suggest the majority are likely cognitive in nature and due to, for example, problems interpreting feedback from the word prediction system (which may display non-target word prefixes during partial word entry). Evidence to support this assertion comes from closely related prior work. Examinations of mean key level accuracy for, respectively, two-thumb text entry systems on soft tablet keyboards [65], miniature physical keyboards [14], and soft smartphone keyboards [4] are 94.8%, 93.9% and 89.2%. Our data fall squarely in the middle of this range, suggesting they may represent typical rates for thumb typing in general. This analysis, combined with the generally good performance recorded throughout the study, suggest our classifiers performed more or less as expected during this study. Indeed, the results of this study arguably serve as an effective validation of the performance of our key classifier as it includes both a new participant group and a more complex and naturalistic task.

8 COMPARISON STUDY: COMPARING THUMBAIR AGAINST A BASELINE

Building on the results of the typing study, we conducted a follow-up study that sought to contextualize and ground the performance data we report for ThumbAir. We achieved this by exploring typing performance in a more realistic pose (standing), by capturing salient aspects of participants' experiences and opinions via standard exertion, workload and social acceptability questionnaires as well as through short interviews and, finally, by contrasting ThumbAir's performance against that achieved in a baseline in-air typing system implemented following a design that has been frequently proposed in prior work [18, 75]. This design requires a user to use their the index fingers to strike keys on a QWERTY layout keyboard positioned directly in front of their body. We opted to compare ThumbAir against this design due to both its prevalence in the literature, and the equivalence of its design goals to our own: this design also enables unencumbered, internally tracked, in-air typing input.

8.1 Methods

8.1.1 Participants. We recruited twelve new participants (right-handed, mean age 20.33 (SD=2.64), five female, seven male). They self-rated their familiarity with computers (4.75/5.0), smartphones (5.0/5.0), and thumb typing on smartphones (5.0/5.0) as high, but they had relatively low familiarity with HMDs (1.92/5.0, SD=0.51) and very limited experience with typing on HMDs (1.0/5.0). They were all non-native English speakers enrolled in a full time English language degree program. They took approximately 60 minutes to complete this study and each received the equivalent of 15 USD in local currency.

8.1.2 Materials and Study Design. This study contrasted two conditions: ThumbAir and a Baseline. ThumbAir was configured as in the typing study (see Section 7.1.2), while the Baseline was implemented following a frequently proposed and studied in-air keyboard design [18, 75]. It is illustrated in Figure 13 (left) and involves a QWERTY keyboard located in the space immediately in front of a user; text is entered by striking keys with either index finger. Following recommendations in the literature, we situated this keyboard 35cm below and 50cm forward of the HMD [18, 75]. Keys were 22mm square [18]. We included space and delete keys on the right side of the keyboard, provided a green highlight when contact with a key was made and implemented a simple debounce routine that blocked multiple key selections within 200ms of each other. For both ThumbAir and the Baseline system, we displayed the instructions (i.e., the phrases to type) and entered text just above the input surface—see Figure 13.

All participants completed both conditions in a fully balanced repeated measures arrangement. Each condition was composed of two sequential sessions, a manipulation we included to enable us to examine short term learning rates. Each session featured 15 randomly selected phrases from Mackenzie and Soukoreff's phrase set [55]. We filtered the phrases to ensure they only contain words that appear in ThumbAir's dictionary. In this way, in total, we collected 720 phrases (2 conditions by 2 sessions by 15 phrases by 12 participants). Metrics were both objective (WPM, CER) and also subjective: for each condition we captured standard measures of perceived exertion (BORG CR10 [10]) and workload (NASA TLX [33]). We also used a questionnaire for social acceptability [2] that asks participants to report on the places (e.g., home, street) in which they would be willing to perform a input task and individuals/groups (e.g., alone, partner, colleagues) they would be willing to perform that task in front of. Finally, we asked participants for their comments and opinions on each of the text entry systems.

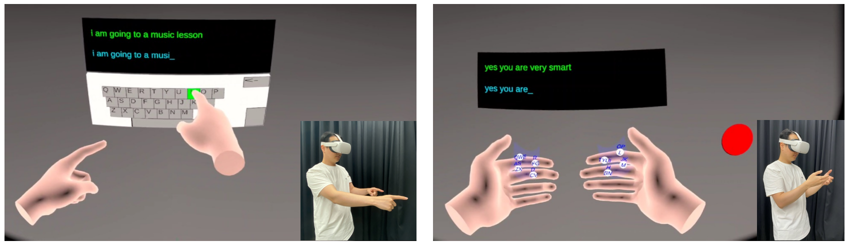

Fig. 13. Baseline (left) and ThumbAir (right) interfaces in the comparison study. Participants stood throughout the study.

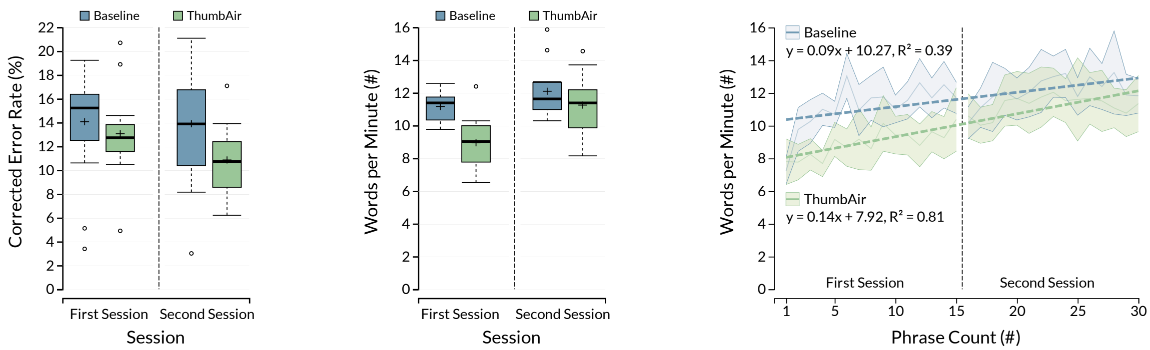

Fig. 14. Box plots showing CER (left) and WPM (center) recorded in the comparison study. Right figure shows how WPM varies as more phrases are typed (shaded area represents 95% CI) and includes linear regression lines (which are dotted), equations and fit. Note that high variability for each per-phrase mean is expected and that interpretation of this plot should focus on the regression lines and fit—these reflect the observed trends over the whole duration of the study.

8.1.3 Procedure. The study used the same setting and equipment, and followed broadly similar procedures, to prior studies reported in this paper. Participants first read study instructions and watched a video demonstrating both typing systems. They did not complete any practice sessions; we sought to assess genuine novice performance. In addition, they conducted all tasks standing, a realistic HMD use posture for a very broad range of scenarios, and one that has previously been used to study performance in the design that inspired our Baseline system [75]. They next completed two sequential sessions, separated by a break of three minutes, for one of the typing systems. They then filled in study questionnaires and gave qualitative comments about their experience. After another five minute break, they completed the same process with the other typing system.

8.2 Results and Discussion

We captured 720 typed phrases in this study; five phrases produced by two participants in the Baseline condition were truncated, most likely by inadvertent and premature contact with the enter key. We excluded these five from our analysis and then calculated WPM and CER from the remaining set of 715 phrases. In addition to dividing the data by typing condition, we also divided it by session to enable us to examine short term learning rates. This data is shown in Figure 14 (center, left) and we analyzed it with a pair of two-way repeated-measures ANOVAs. For WPM, we recorded a full suite of significant differences: the interaction (F(1, 11)=10.545, p=0.008, η^2_G=0.050) and main effects of condition (F(1, 11)=9.377, p=0.011, η^2_G=0.206) and session (F(1, 11)=70.960, p<0.001, η^2_G=0.224). The interaction effect indicates that performance changes between the two sequential sessions differed between the

Table 1. Mean subjective data recorded in the baseline study for NASA TLX and Borg CR10. Figures in brackets show SD.

	Nasa TLX						Borg CR10				
	Mental	**Physical**	**Temporal**	**Perf.**	**Effort**	**Frust.**	**Shoulder**	**Arm**	**Forearm**	**Wrist**	**Hand**
Baseline	2.83 (2.1)	6.17 (3.9)	4.92 (4.6)	6.5 (4.4)	7.25 (5.3)	4.58 (5.8)	2.79 (2.2)	1.71 (1.9)	1.5 (2)	1.46 (1.9)	1.5 (2.2)
ThumbAir	3.33 (2.3)	2.58 (2.6)	4.17 (4.2)	6.42 (5.2)	7.42 (5.5)	4.42 (4.5)	1.08 (0.8)	0.83 (0.7)	0.75 (0.7)	0.83 (0.9)	2.04 (1.6)

two typing conditions. To explore these trends, we ran paired t-tests on the data from each session individually. The results show that while Baseline leads to a significantly higher WPM in the first session (p=0.001), there is no significant difference in second session data (p=0.167). This suggests that the main effect of condition is primarily due to the improved performance recorded for Baseline in the first session and that the main effect of session mainly reflects the performance improvement observed for ThumbAir in the second session. To explore this change in more detail, we plotted WPM data by phrase entered (see Figure 14 right) and fit linear regression lines: ThumbAir shows a steeper slope (0.1402 vs 0.0877 for Baseline) and good fit (r^2 = 0.81). While per-phrase means show, as expected, high variability, the trends over the full set remain relatively clear and consistent. We conclude that participants initially struggled due to unfamiliarity with ThumbAir's layout. However, they were able to rapidly pick it up: after typing 15 phrases, performance was not significantly different to that attained using a more standard QWERTY design. This strong performance likely reflects our emphasis on maintaining close QWERTY similarity during the design of ThumbAir's keyboard layout. We also note that basic performance levels recorded are representative of those in the literature: our mean Baseline WPM of 12.09 WPM is 20% higher than that recorded for the keyboard design and implementation it is based on [75]. This boost that may be due to a wide variety of factors, most likely to improvements in finger tracking technology. In terms of CER, no significant differences were recorded in either interaction or main effect. In addition, we note these CER figures are typical of those reported in closely related studies of HMD based bare hand text entry (7.6% to 13.3%) [19, 75], but somewhat elevated when compared more mature platforms such as smartphones (approximately 5%) [5]. Finally, MSD error rates were low in both ThumbAir (Mean=0.34%, SD=0.30%) and Baseline (Mean=0.15%, SD=0.27%). This indicates that participants corrected the vast majority of typing errors that occurred in the study.

Subjective feedback data for Borg CR10 and NASA TLX are shown in Table 1. We analyzed this data with Bonferroni corrected paired t-tests, ultimately uncovering significant differences in two measures. Participants reported higher levels of perceived exertion at the shoulder (*p-value* = 0.049) in the Baseline condition. Specifically, they reported that the Baseline condition led to moderate levels of perceived shoulder exertion versus very weak levels during use of ThumbAir. In addition, they reported increased physical demand in the Baseline condition (*p-value* = 0.045). These data suggest that the Baseline condition suffered from "gorilla arm" [34, 38] but that the more relaxed posture of ThumbAir, with the hands lowered and the arms close to the body, largely avoided this. This result is in line with prior work in which such poses have been suggested as targets for reducing exertion levels during input tasks [34]. In addition, we analyzed the social acceptance questionnaire by the simple expedient of summing the number of locations and groups participants selected. This creates counts for each condition of the number of locations in which participants would perform input, and the number of groups they would perform it in front of. We then tested these counts using paired t-tests, recording significant differences for both locations (p=0.002) and groups (p=0.001). Specifically, while participants indicated they would use the Baseline condition in a mean of 2.33 locations (SD 1.23)), they indicated they would use ThumbAir in a mean of 4.5 locations (SD 1.78). For groups this data was 4.08 (SD 1.08) for Baseline and 5.42 (SD 0.79) for ThumbAir. These results indicate that participants felt it would be socially acceptable to use ThumbAir in a greater number of real world contexts. Qualitative recorded at the end of the study shed light on this issue. With ThumbAir six participants noted the input actions resembled current practices, such as use of a smartphone. This familiarity increased their sense the input would be acceptable in social situations. Additionally five described the small inputs motions as discreet or hard to observe, factors that were also viewed to make them more acceptable. These findings are in line with prior work [47]. In contrast for Baseline, five participants noted the input would be "strange", "weird" or "odd", and four further suggested this problem is due to the fact the required finger motion is "uncommon" or otherwise unintelligible to bystanders. More positively, two noted that they felt that it would be more socially acceptable if they provided an explanation for their behavior, or readily envisaged a future when this type of input becomes socially acceptable due to its prevalence.

9 DISCUSSION AND CONCLUSION

This paper describes the process of designing an in-air thumb typing keyboard for the inside-out hand tracker available on a standalone off-the-shelf HMD. It targets a casual use-scenario in which an AR or VR HMD user lacks external input devices but wishes to type on their headset to meet various needs, such as sporadic messaging [35], note taking [13], authentication [23] or other personal information management tasks [51]. It takes account of factors such as comfort [10], workload [33] and, reflecting the fact that input may take place in a range of settings, the social acceptability [2] of the input by focusing on small and discreet input actions [47]. Our design process was data-driven and empirical. First, we sampled performance over a broad range of target locations to understand viable regions in which to locate keys and how users would move towards them. Based on the results, we designed an arrangement of four targets for each hand in ready reach of thumb when it is rapidly articulated with gross rotations of the metacarpal joint. These movements are comfortable, quick and relatively large scale. As such, they are easy to track (comfortably within the HMD's reported finger tracking accuracy [1]) and also easy to execute in the absence of haptic feedback (from the user's point of view). We then used these targets in a follow up study examining bigram performance—tapping targets in sequential pairs. We used the data from this study to refine target locations and design a novel target selection system (a classifier based on finger intersections with a virtual surface) that achieves a high level of accuracy (98.2%). In addition we used this data in a computational process to inform the design of a novel keyboard layout. The layout we ultimately selected emphasized a high level of QWERTY similarity and low levels of ambiguity with respect to entered words. In a third study, we evaluated the target locations, selection system and keyboard layout in word repetition and phrase typing tasks, ultimately achieving final WPMs of 27.1 and 13.73, respectively. We believe this level of performance, achieved on an on-the-shelf HMD hand tracker, represents a meaningful achievement. Both word repetition and phrase entry speeds approached or exceeded those previously attained in closely related prior work that relies on dedicated worn hardware [36, 45, 80] or external optical tracking systems [17, 19, 73]. We ground this data with a final study that compares performance with ThumbAir against a baseline in-air typing design in a phrase-entry task. In addition to objective measures we collect qualitative measures of perceived exertion, workload and social acceptability. The results indicate that while objective performance between the two designs is broadly similar, ThumbAir offers advantages in terms of the qualitative metrics: it induces significantly lower levels of perceived exertion and physical demand and is rated as acceptable for use in a wider range of situations that the baseline. These are key qualities that match our goal of creating a typing system to support a range of casual HMD use scenarios. In sum, our work demonstrates that rapid, accurate, comfortable and socially acceptable in-air thumb typing performance can be achieved using currently available consumer hardware.

Despite these positive conclusions, there are a number of limitations to this work and, similarly, a wide range of directions for future study. Perhaps most foundationally, while the metacarpal flexes we study represent the dominant thumb motion performed by our users, it is simple and arguably fails to take advantage of the high dexterity, and degrees-of-freedom, available to the thumb [44]. Future work might further investigate the potential for detailed in-air thumb motions involving lateral movements, flexes of inter-phalangeal joint, or more complex compound actions such as ”pokes”. It would also be important to extend the studies reported in this paper with work that examines a range of more diverse settings. We used a seated pose to avoid fatigue during our first three prolonged studies and a standing pose in our shorter fourth study. A valuable complement to this lab-based work would be an exploration of in-air typing performance while walking [45]). These more mobile scenarios are commonplace in HMD applications and particularly relevant for AR: our system needs be assessed in them. The performance we observe while seated and standing may not fully generalize to more dynamic poses and settings. It would also be interesting to explore the impact of customization on performance. While we note our studies showed no variations that could be explained by hand size, and we used a fixed virtual hand size throughout our work, mismatches between a users' body and its representation in AR/VR are known to exert a wide range of effects [69]. It would be interesting to evaluate whether using scaling or other types of customization to minimize mismatches between a user's real and virtual hands can boost performance in the types of text entry tasks studied here. In addition, while the studies in the paper target various aspects of typing performance, we did not conduct a prolonged, multi-session study [65, 73]. Data about in-air thumb-typing from such a study would closely complement the results we report here.

Finally, we also note that while this paper documents the design and evaluation of a keyboard layout, a full typing system is a more sophisticated artifact. Immediate future work will need to redesign or integrate a wide range of features into our system to truly enable users to type. For example, as our keyboard is ambiguous, a interface for word selection needs to be defined. As the proposed layout shows a high Top 2% [29] score of 98.1%, a simple system featuring two word choices is likely viable. These limited selections could therefore be activated by simple hand gestures, such as flexing all fingers (or making a fist) with the left or right hands. However, it is also worth noting that word selection and prediction systems do not universally result in improved text entry speeds [6]. Although a two choice design may be practical for ThumbAir, it would also need be highly accurate in order to allow users to achieve greater WPMs. The cost of errors in word prediction systems are high. Alternatively, an effective design might integrate advanced feedback, such as display of confidence in its predictions [6] which may allow users to better recover from errors. Finally, we note a realistic keyboard will also require mode switching to support entry of numbers or symbols. This might be achieved by a palm open gesture with one hand which could be overlaid with a numeric keypad or other symbol entry system [64]. By continuing to develop our system to achieve full text entry functionality we hope to enable users of current AR and VR HMDs to use the hand tracking built into their existing headsets to type comfortably, discreetly, rapidly and accurately.

ACKNOWLEDGMENTS

Thanks to all our study participants. This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (2020R1F1A1070699).

REFERENCES

[1] Diar Abdlkarim, Massimiliano Di Luca, Poppy Aves, Sang-Hoon Yeo, R. Chris Miall, Peter Holland, and Joseph M. Galea. 2022. A Methodological Framework to Assess the Accuracy of Virtual Reality Hand-Tracking Systems: A case study with the Oculus Quest 2. *bioRxiv* (2022). https://doi.org/10.1101/2022.02.18.481001 arXiv:https://www.biorxiv.org/content/early/2022/02/20/2022.02.18.481001.full.pdf

[2] David Ahlström, Khalad Hasan, and Pourang Irani. 2014. Are You Comfortable Doing That? Acceptance Studies of around-Device Gestures in and for Public Settings. In *Proceedings of the 16th International Conference on Human-Computer Interaction with Mobile Devices & Services* (Toronto, ON, Canada) *(MobileHCI '14)*. Association for Computing Machinery, New York, NY, USA, 193–202. https://doi.org/10.1145/2628363.2628381

[3] Sunggeun Ahn and Geehyuk Lee. 2019. Gaze-Assisted Typing for Smart Glasses. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (New Orleans, LA, USA) *(UIST '19)*. Association for Computing Machinery, New York, NY, USA, 857–869. https://doi.org/10.1145/3332165.3347883

[4] Shiri Azenkot and Shumin Zhai. 2012. Touch Behavior with Different Postures on Soft Smartphone Keyboards. In *Proceedings of the 14th International Conference on Human-Computer Interaction with Mobile Devices and Services* (San Francisco, California, USA) *(MobileHCI '12)*. Association for Computing Machinery, New York, NY, USA, 251–260. https://doi.org/10.1145/2371574.2371612

[5] Nikola Banovic, Varun Rao, Abinaya Saravanan, Anind K. Dey, and Jennifer Mankoff. 2017. Quantifying Aversion to Costly Typing Errors in Expert Mobile Text Entry. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 4229–4241. https://doi.org/10.1145/3025453.3025695

[6] Nikola Banovic, Ticha Sethapakdi, Yasasvi Hari, Anind K. Dey, and Jennifer Mankoff. 2019. The Limits of Expert Text Entry Speed on Mobile Keyboards with Autocorrect. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services* (Taipei, Taiwan) *(MobileHCI '19)*. Association for Computing Machinery, New York, NY, USA, Article 15, 12 pages. https://doi.org/10.1145/3338286.3340126

[7] Xiaojun Bi, Barton A. Smith, and Shumin Zhai. 2010. Quasi-Qwerty Soft Keyboard Optimization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 283–286. https://doi.org/10.1145/1753326.1753367

[8] Xiaojun Bi and Shumin Zhai. 2016. IJQwerty: What Difference Does One Key Change Make? Gesture Typing Keyboard Optimization Bounded by One Key Position Change from Qwerty. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 49–58. https://doi.org/10.1145/2858036.2858421

[9] Benjamin I Binder-Markey and Wendy M Murray. 2017. Incorporating the length-dependent passive-force generating muscle properties of the extrinsic finger muscles into a wrist and finger biomechanical musculoskeletal model. *Journal of biomechanics* 61 (2017), 250–257.

[10] Gunnar Borg, Gunilla Ljunggren, and Ruggero Ceci. 1985. The increase of perceived exertion, aches and pain in the legs, heart rate and blood lactate during exercise on a bicycle ergometer. *European journal of applied physiology and occupational physiology* 54, 4 (1985), 343–349.

[11] Thorsten Brants. 2006. Web 1T 5-gram Version 1. *http://www. ldc. upenn. edu/Catalog/CatalogEntry. jsp? catalogId= LDC2006T13* (2006).

[12] Alisa Burova, John Mäkelä, Jaakko Hakulinen, Tuuli Keskinen, Hanna Heinonen, Sanni Siltanen, and Markku Turunen. 2020. Utilizing VR and Gaze Tracking to Develop AR Solutions for Industrial Maintenance. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376405

[13] Yi-Ting Chen, Chi-Hsuan Hsu, Chih-Han Chung, Yu-Shuen Wang, and Sabarish V Babu. 2019. Ivrnote: design, creation and evaluation of an interactive note-taking interface for study and reflection in vr learning environments. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, IEEE Computer Society, USA, 172–180.

[14] Edward Clarkson, James Clawson, Kent Lyons, and Thad Starner. 2005. An Empirical Study of Typing Rates on Mini-QWERTY Keyboards. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems* (Portland, OR, USA) *(CHI EA '05)*. Association for Computing Machinery, New York, NY, USA, 1288–1291. https://doi.org/10.1145/1056808.1056898

[15] Carolina Cruz-Neira, Marcos Fernández, and Cristina Portalés. 2018. Virtual reality and games.

[16] Rajkumar Darbar, Joan Odicio-Vilchez, Thibault Lainé, Arnaud Prouzeau, and Martin Hachet. 2021. Text selection in ar-hmd using a smartphone as an input device. In *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, IEEE Computer Society, USA, 526–527.

[17] John Dudley, Hrvoje Benko, Daniel Wigdor, and Per Ola Kristensson. 2019. Performance Envelopes of Virtual Keyboard Text Input Strategies in Virtual Reality. In *2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE Computer Society, Washington, DC, USA, 289–300. https://doi.org/10.1109/ISMAR.2019.00027

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

[19] Jacqui Fashimpaur, Kenrick Kin, and Matt Longest. 2020. PinchType: Text Entry for Virtual and Augmented Reality Using Comfortable Thumb to Fingertip Pinches. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3334480.3382888

[20] Anna Maria Feit. 2018. *Assignment Problems for Optimizing Text Input*. Doctoral thesis. School of Electrical Engineering. http://urn.fi/URN:ISBN:978-952-60-8016-1

[21] Leah Findlater and Jacob Wobbrock. 2012. Personalized Input: Improving Ten-Finger Touchscreen Typing through Automatic Adaptation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 815–824. https://doi.org/10.1145/2207676.2208520

[22] Leah Findlater, Jacob O. Wobbrock, and Daniel Wigdor. 2011. Typing on Flat Glass: Examining Ten-Finger Expert Typing Patterns on Touch Surfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 2453–2462. https://doi.org/10.1145/1978942.1979301

[23] Ceenu George, Mohamed Khamis, Emanuel von Zezschwitz, Marinus Burger, Henri Schmidt, Florian Alt, and Heinrich Hussmann. 2017. Seamless and secure vr: Adapting and evaluating established authentication systems for virtual reality. NDSS.

[24] Hyunjae Gil, DoYoung Lee, Seunggyu Im, and Ian Oakley. 2017. TriTap: Identifying Finger Touches on Smartwatches. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 3879–3890. https://doi.org/10.1145/3025453.3025561

[25] Hyunjae Gil, Yonghwan Shin, Hyungki Son, Inwook Hwang, Ian Oakley, and Jin Ryong Kim. 2020. Characterizing In-Air Eyes-Free Typing Movements in VR. In *26th ACM Symposium on Virtual Reality Software and Technology* (Virtual Event, Canada) *(VRST '20)*. Association for Computing Machinery, New York, NY, USA, Article 7, 10 pages. https://doi.org/10.1145/3385956.3418963

[26] Kentaro Go and Yuki Endo. 2007. CATKey: Customizable and Adaptable Touchscreen Keyboard with Bubble Cursor-Like Visual Feedback. In *Human-Computer Interaction – INTERACT 2007*, Cécilia Baranauskas, Philippe Palanque, Julio Abascal, and Simone Diniz Junqueira Barbosa (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 493–496.

[27] Mayank Goel, Leah Findlater, and Jacob Wobbrock. 2012. 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* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 2687–2696. https://doi.org/10.1145/2207676.2208662

[28] Jun Gong, Zheer Xu, Qifan Guo, Teddy Seyed, Xiang 'Anthony' Chen, Xiaojun Bi, and Xing-Dong Yang. 2018. WrisText: One-Handed Text Entry on Smartwatch Using Wrist Gestures. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173755

[29] Jun Gong, Xing-Dong Yang, and Pourang Irani. 2016. WristWhirl: One-Handed Continuous Smartwatch Input Using Wrist Gestures. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (Tokyo, Japan) *(UIST '16)*. Association for Computing Machinery, New York, NY, USA, 861–872. https://doi.org/10.1145/2984511.2984563

[30] Jens Grubert, Lukas Witzani, Eyal Ofek, Michel Pahud, Matthias Kranz, and Per Ola Kristensson. 2018. 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 Computer Society, USA, 159–166. https://doi.org/10.1109/VR.2018.8446059

[31] Ru Guo, Yiru Yang, Johnson Kuang, Xue Bin, Dhruv Jain, Steven Goodman, Leah Findlater, and Jon Froehlich. 2020. HoloSound: Combining Speech and Sound Identification for Deaf or Hard of Hearing Users on a Head-Mounted Display. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, Greece) *(ASSETS '20)*. Association for Computing Machinery, New York, NY, USA, Article 71, 4 pages. https://doi.org/10.1145/3373625.3418031

[32] Aakar Gupta, Cheng Ji, Hui-Shyong Yeo, Aaron Quigley, and Daniel Vogel. 2019. RotoSwype: Word-Gesture Typing Using a Ring. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300244

[33] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

[34] Juan David Hincapié-Ramos, Xiang Guo, Paymahn Moghadasian, and Pourang Irani. 2014. Consumed Endurance: A Metric to Quantify Arm Fatigue of Mid-Air Interactions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 1063–1072. https://doi.org/10.1145/2556288.2557130

[35] Ching-Yu Hsieh, Yi-Shyuan Chiang, Hung-Yu Chiu, and Yung-Ju Chang. 2020. Bridging the Virtual and Real Worlds: A Preliminary Study of Messaging Notifications in Virtual Reality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376228

[36] Da-Yuan Huang, Liwei Chan, Shuo Yang, Fan Wang, Rong-Hao Liang, De-Nian Yang, Yi-Ping Hung, and Bing-Yu Chen. 2016. DigitSpace: Designing Thumb-to-Fingers Touch Interfaces for One-Handed and Eyes-Free Interactions. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 1526–1537. https://doi.org/10.1145/2858036.2858483

[37] Wahyu Hutama, Hikari Harashima, Hironori Ishikawa, and Hiroyuki Manabe. 2021. HMK: Head-Mounted-Keyboard for Text Input in Virtual or Augmented Reality. In *The Adjunct Publication of the 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21)*. Association for Computing Machinery, New York, NY, USA, 115–117. https://doi.org/10.1145/3474349.3480195

[38] Sujin Jang, Wolfgang Stuerzlinger, Satyajit Ambike, and Karthik Ramani. 2017. Modeling Cumulative Arm Fatigue in Mid-Air Interaction Based on Perceived Exertion and Kinetics of Arm Motion. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 3328–3339. https://doi.org/10.1145/3025453.3025523

[39] Haiyan Jiang and Dongdong Weng. 2020. HiPad: Text entry for Head-Mounted Displays Using Circular Touchpad. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, IEEE Computer Society, USA, 692–703. https://doi.org/10.1109/VR46266.2020.00092

[40] Haiyan Jiang, Dongdong Weng, Zhenliang Zhang, and Feng Chen. 2019. Hifinger: One-handed text entry technique for virtual environments based on touches between fingers. *Sensors* 19, 14 (2019), 3063.

[41] Eleanor Jones, Jason Alexander, Andreas Andreou, Pourang Irani, and Sriram Subramanian. 2010. GesText: Accelerometer-Based Gestural Text-Entry Systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 2173–2182. https://doi.org/10.1145/1753326.1753655

[42] Jin Ryong Kim and Hong Z. Tan. 2014. A study of touch typing performance with keyclick feedback. In *2014 IEEE Haptics Symposium (HAPTICS)*. IEEE Computer Society, USA, 227–233. https://doi.org/10.1109/HAPTICS.2014.6775459

[43] Andreas Komninos and Mark Dunlop. 2014. Using a smart-watch as a input device for text. *IEEE Pervasive Computing* 13, 4 (2014), 50–58.

[44] Catherine E Lang and Marc H Schieber. 2004. Human finger independence: limitations due to passive mechanical coupling versus active neuromuscular control. *Journal of neurophysiology* 92, 5 (2004), 2802–2810.

[45] DoYoung Lee, Jiwan Kim, and Ian Oakley. 2021. FingerText: Exploring and Optimizing Performance for Wearable, Mobile and One-Handed Typing. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 283, 15 pages. https://doi.org/10.1145/3411764.3445106

[46] DoYoung Lee, SooHwan Lee, and Ian Oakley. 2020. Nailz: Sensing Hand Input with Touch Sensitive Nails. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376778

[47] DoYoung Lee, Youryang Lee, Yonghwan Shin, and Ian Oakley. 2018. Designing Socially Acceptable Hand-to-Face Input. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology* (Berlin, Germany) *(UIST '18)*. Association for Computing Machinery, New York, NY, USA, 711–723. https://doi.org/10.1145/3242587.3242642

[48] Gyusung I Lee and Mija R Lee. 2018. Can a virtual reality surgical simulation training provide a self-driven and mentor-free skills learning? Investigation of the practical influence of the performance metrics from the virtual reality robotic surgery simulator on the skill learning and associated cognitive workloads. *Surgical endoscopy* 32, 1 (2018), 62–72.

[49] Lik Hang Lee, Kit Yung Lam, Tong Li, Tristan Braud, Xiang Su, and Pan Hui. 2019. Quadmetric optimized thumb-to-finger interaction for force assisted one-handed text entry on mobile headsets. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–27.

[50] Gregory W Lesher, Bryan J Moulton, and D Jeffery Higginbotham. 1998. Optimal character arrangements for ambiguous keyboards. *IEEE Transactions on Rehabilitation Engineering* 6, 4 (1998), 415–423.

[51] Jonathan Liebers, Patrick Horn, Christian Burschik, Uwe Gruenefeld, and Stefan Schneegass. 2021. Using Gaze Behavior and Head Orientation for Implicit Identification in Virtual Reality. In *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology* (Osaka, Japan) *(VRST '21)*. Association for Computing Machinery, New York, NY, USA, Article 22, 9 pages. https://doi.org/10.1145/3489849.3489880

[52] Mark A. Livingston, Lawrence J. Rosenblum, Dennis G. Brown, Gregory S. Schmidt, Simon J. Julier, Yohan Baillot, J. Edward Swan, Zhuming Ai, and Paul Maassel. 2011. Military Applications of Augmented Reality. (2011), 671–706. https://doi.org/10.1007/978-1-4614-0064-6_31

[53] I. Scott MacKenzie and R. William Soukoreff. 2002. A Character-Level Error Analysis Technique for Evaluating Text Entry Methods. In *Proceedings of the Second Nordic Conference on Human-Computer Interaction* (Aarhus, Denmark) *(NordiCHI '02)*. Association for Computing Machinery, New York, NY, USA, 243–246. https://doi.org/10.1145/572020.572056

[54] I Scott MacKenzie and R William Soukoreff. 2002. Text entry for mobile computing: Models and methods, theory and practice. *Human–Computer Interaction* 17, 2-3 (2002), 147–198.

[55] I. Scott MacKenzie and R. William Soukoreff. 2003. Phrase Sets for Evaluating Text Entry Techniques. In *CHI '03 Extended Abstracts on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI EA '03)*. Association for Computing Machinery, New York, NY, USA, 754–755. https://doi.org/10.1145/765891.765971

[56] Scott MacKenzie. 2021. A note on calculating text entry speed. https://www.yorku.ca/mack/RN-TextEntrySpeed.html

[57] Federico Manuri and Andrea Sanna. 2016. A survey on applications of augmented reality. *ACSIJ Advances in Computer Science: an International Journal* 5, 1 (2016), 18–27.

[58] Fabrice Matulic, Aditya Ganeshan, Hiroshi Fujiwara, and Daniel Vogel. 2021. Phonetroller: Visual Representations of Fingers for Precise Touch Input with Mobile Phones in VR. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 129, 13 pages. https://doi.org/10.1145/3411764.3445583

[59] Mark McGill, Daniel Boland, Roderick Murray-Smith, and Stephen Brewster. 2015. A Dose of Reality: Overcoming Usability Challenges in VR Head-Mounted Displays. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 2143–2152. https://doi.org/10.1145/2702123.2702382

[60] Mark McGill, Stephen Brewster, Daniel Pires De Sa Medeiros, Sidney Bovet, Mario Gutierrez, and Aidan Kehoe. 2022. Creating and Augmenting Keyboards for Extended Reality with the K eyboard A ugmentation T oolkit. *ACM Transactions on Computer-Human Interaction* 29, 2 (2022), 1–39.

[61] Alexander Ng, Stephen A. Brewster, and John H. Williamson. 2014. Investigating the Effects of Encumbrance on One- and Two- Handed Interactions with Mobile Devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 1981–1990. https://doi.org/10.1145/2556288.2557312

[62] Peter Norvig. 2009. Natural language corpus data. *Beautiful data* (2009), 219–242.

[63] Daniel R. O'Day and Ricardo A. Calix. 2013. Text message corpus: Applying natural language processing to mobile device forensics. In *2013 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*. IEEE, IEEE Computer Society, USA, 1–6. https:

//doi.org/10.1109/ICMEW.2013.6618380
[64] Taihei Ogitani, Yoshitaka Arahori, Yusuke Shinyama, and Katsuhiko Gondow. 2018. Space Saving Text Input Method for Head Mounted Display with Virtual 12-key Keyboard. In *2018 IEEE 32nd International Conference on Advanced Information Networking and Applications (AINA)*. IEEE, IEEE Computer Society, USA, 342–349. https://doi.org/10.1109/AINA.2018.00059
[65] Antti Oulasvirta, Anna Reichel, Wenbin Li, Yan Zhang, Myroslav Bachynskyi, Keith Vertanen, and Per Ola Kristensson. 2013. Improving Two-Thumb Text Entry on Touchscreen Devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 2765–2774. https://doi.org/10.1145/2470654.2481383
[66] Kseniia Palin, Anna Maria Feit, Sunjun Kim, Per Ola Kristensson, and Antti Oulasvirta. 2019. How Do People Type on Mobile Devices? Observations from a Study with 37,000 Volunteers. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services* (Taipei, Taiwan) *(MobileHCI '19)*. Association for Computing Machinery, New York, NY, USA, Article 9, 12 pages. https://doi.org/10.1145/3338286.3340120
[67] Duc-Minh Pham and Wolfgang Stuerzlinger. 2019. HawKEY: Efficient and Versatile Text Entry for Virtual Reality. In *25th ACM Symposium on Virtual Reality Software and Technology* (Parramatta, NSW, Australia) *(VRST '19)*. Association for Computing Machinery, New York, NY, USA, Article 21, 11 pages. https://doi.org/10.1145/3359996.3364265
[68] Christina Pollalis, Amanda Gilvin, Lauren Westendorf, Lauren Futami, Bella Virgilio, Dana Hsiao, and Orit Shaer. 2018. ARtLens: Enhancing Museum Visitors' Engagement with African Art. In *Proceedings of the 2018 ACM Conference Companion Publication on Designing Interactive Systems* (Hong Kong, China) *(DIS '18 Companion)*. Association for Computing Machinery, New York, NY, USA, 195–200. https://doi.org/10.1145/3197391.3205435
[69] Maria Pyasik, Gaetano Tieri, and Lorenzo Pia. 2020. Visual appearance of the virtual hand affects embodiment in the virtual hand illusion. *Scientific reports* 10, 1 (2020), 1–11.
[70] Oculus Quest. 2021. VR Headsets, Games, and Equipment. https://www.oculus.com/
[71] Shwetha Rajaram and Michael Nebeling. 2022. Paper Trail: An Immersive Authoring System for Augmented Reality Instructional Experiences. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 382, 16 pages. https://doi.org/10.1145/3491102.3517486
[72] Garth Shoemaker, Leah Findlater, Jessica Q. Dawson, and Kellogg S. Booth. 2009. Mid-Air Text Input Techniques for Very Large Wall Displays. In *Proceedings of Graphics Interface 2009* (Kelowna, British Columbia, Canada) *(GI '09)*. Canadian Information Processing Society, CAN, 231–238.
[73] Jeongmin Son, Sunggeun Ahn, Sunbum Kim, and Geehyuk Lee. 2019. Improving Two-Thumb Touchpad Typing in Virtual Reality. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3312926
[74] R. William Soukoreff and I. Scott MacKenzie. 2003. Metrics for Text Entry Research: An Evaluation of MSD and KSPC, and a New Unified Error Metric. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI '03)*. Association for Computing Machinery, New York, NY, USA, 113–120. https://doi.org/10.1145/642611.642632
[75] Marco Speicher, Anna Maria Feit, Pascal Ziegler, and Antonio Krüger. 2018. Selection-Based Text Entry in Virtual Reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3174221
[76] Ke Sun, Wei Wang, Alex X. Liu, and Haipeng Dai. 2018. Depth Aware Finger Tapping on Virtual Displays. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services* (Munich, Germany) *(MobiSys '18)*. Association for Computing Machinery, New York, NY, USA, 283–295. https://doi.org/10.1145/3210240.3210315
[77] Keith Vertanen and Per Ola Kristensson. 2011. 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* (Stockholm, Sweden) *(MobileHCI '11)*. Association for Computing Machinery, New York, NY, USA, 295–298. https://doi.org/10.1145/2037373.2037418
[78] Keith Vertanen and Per Ola Kristensson. 2021. Mining, analyzing, and modeling text written on mobile devices. *Natural Language Engineering* 27, 1 (2021), 1–33.
[79] James Walker, Bochao Li, Keith Vertanen, and Scott Kuhl. 2017. Efficient Typing on a Visually Occluded Physical Keyboard. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 5457–5461. https://doi.org/10.1145/3025453.3025783
[80] Eric Whitmire, Mohit Jain, Divye Jain, Greg Nelson, Ravi Karkar, Shwetak Patel, and Mayank Goel. 2017. Digitouch: Reconfigurable thumb-to-finger input and text entry on head-mounted displays. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–21.
[81] Pui Chung Wong, Kening Zhu, and Hongbo Fu. 2018. FingerT9: Leveraging Thumb-to-Finger Interaction for Same-Side-Hand Text Entry on Smartwatches. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3173574.3173752
[82] Wenge Xu, Hai-Ning Liang, Anqi He, and Zifan Wang. 2019. 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, IEEE Computer Society, USA,

279–288. https://doi.org/10.1109/ISMAR.2019.00026

[83] Wenge Xu, Hai-Ning Liang, Yuxuan Zhao, Tianyu Zhang, Difeng Yu, and Diego Monteiro. 2019. Ringtext: Dwell-free and hands-free text entry for mobile head-mounted displays using head motions. *IEEE transactions on visualization and computer graphics* 25, 5 (2019), 1991–2001.

[84] Zheer Xu, Weihao Chen, Dongyang Zhao, Jiehui Luo, Te-Yen Wu, Jun Gong, Sicheng Yin, Jialun Zhai, and Xing-Dong Yang. 2020. BiTipText: Bimanual Eyes-Free Text Entry on a Fingertip Keyboard. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376306

[85] Hui-Shyong Yeo, Xiao-Shen Phang, Taejin Ha, Woontack Woo, and Aaron Quigley. 2017. TiTAN: Exploring Midair Text Entry Using Freehand Input. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI EA '17)*. Association for Computing Machinery, New York, NY, USA, 3041–3049. https://doi.org/10.1145/3027063.3053228

[86] Xin Yi, Chun Yu, Mingrui Zhang, Sida Gao, Ke Sun, and Yuanchun Shi. 2015. ATK: Enabling Ten-Finger Freehand Typing in Air Based on 3D Hand Tracking Data. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology* (Charlotte, NC, USA) *(UIST '15)*. Association for Computing Machinery, New York, NY, USA, 539–548. https://doi.org/10.1145/2807442.2807504

[87] Chun Yu, Yizheng Gu, Zhican Yang, Xin Yi, Hengliang Luo, and Yuanchun Shi. 2017. Tap, Dwell or Gesture? Exploring Head-Based Text Entry Techniques for HMDs. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 4479–4488. https://doi.org/10.1145/3025453.3025964

[88] Difeng Yu, Kaixuan Fan, Heng Zhang, Diego Monteiro, Wenge Xu, and Hai-Ning Liang. 2018. PizzaText: Text entry for virtual reality systems using dual thumbsticks. *IEEE transactions on visualization and computer graphics* 24, 11 (2018), 2927–2935.

[89] Shumin Zhai and Per Ola Kristensson. 2008. Interlaced QWERTY: Accommodating Ease of Visual Search and Input Flexibility in Shape Writing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 593–596. https://doi.org/10.1145/1357054.1357149

[90] Mingrui Ray Zhang, Shumin Zhai, and Jacob O. Wobbrock. 2022. TypeAnywhere: A QWERTY-Based Text Entry Solution for Ubiquitous Computing. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 339, 16 pages. https://doi.org/10.1145/3491102.3517686

[91] Fengyuan Zhu and Tovi Grossman. 2020. BISHARE: Exploring Bidirectional Interactions Between Smartphones and Head-Mounted Augmented Reality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376233

A TARGETS AND KEY LOCATIONS

In the range of motion study, we chose 18 target locations above each hand based on users' ability to select them rapidly and accurately with their thumbs. We used these locations to derive a paraboloid surface and an arrangement of four key locations, which we use in bigram study. We present the locations of these targets here. Table 2 shows coordinates for the 18 selected targets. Table 3 shows the coordinates of the refined key locations, which we use in text entry study. The refined key locations were derived with the mean target centroid position by calculating all ellipses from the collision maps for each target.

Table 2. Coordinates of the 18 targets on the left hand used to derive the paraboloid surface and four key locations used in bigram study. Units are in mm and all coordinates are relative to the center of the palm (as returned by the Oculus SDK). Locations were mirrored for the right hand. The table also depicts the normalized scores of the three metrics (success rate, touch time, and wrist angle) used to calculate the target performance scores (see Section 4.3).

Target ID	1	2	41	42	43	81	82	9	49	50	89	90	17	57	58	97	98	65
X Coordinate	6.25	18.75	6.25	18.75	31.25	6.25	18.75	6.25	6.25	18.75	6.25	18.75	6.25	6.25	18.75	6.25	18.75	6.25
Y Coordinate	-20	-20	-32.5	-32.5	-32.5	-45	-45	-20	-32.5	-32.5	-45	-45	-20	-32.5	-32.5	-45	-45	-32.5
Z Coordinate	-25	-25	-25	-25	-25	-25	-25	-12.5	-12.5	-12.5	-12.5	-12.5	0	0	0	0	0	12.5
Success rate	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Touch time	0.91	0.80	1.00	0.95	0.74	0.92	0.94	0.81	0.95	0.88	0.88	0.91	0.83	0.90	0.87	0.88	0.87	0.87
Wrist angle	0.93	0.86	1.00	1.00	0.87	0.91	0.94	0.86	0.95	0.93	0.85	0.93	0.78	0.84	0.84	0.80	0.87	0.80

Table 3. Coordinates of the refined key locations used in text entry study for the left hand. Units are in mm and all coordinates are relative to the center of the palm (as returned by the Oculus SDK). Locations were mirrored for the right hand.

Key Name	Radial Close	Radial Far	Ulnar Close	Ulnar Far
X Coordinate	28.89	26.96	15.65	12.97
Y Coordinate	-20.73	-42.29	-20.73	-42.29
Z Coordinate	-28.84	-21.17	-16.78	-10.05