Exploring Cross-Modal Training via Touch to Learn a Mid-Air Marking Menu Gesture Set

Jay Henderson[1], **Sachi Mizobuchi**[2], **Wei Li**[2], **Edward Lank**[1]
[1]Cheriton School of Computer Science, Waterloo, Canada
[2]Huawei Noah's Ark Lab, Markham, Canada
{jehender,lank}@uwaterloo.ca,{wei.li.crc,sachi.mizobuchi}@huawei.com

ABSTRACT

While mid-air gestures are an attractive modality with an extensive research history, one challenge with their usage is that the gestures are not self-revealing. Scaffolding techniques to teach these gestures are difficult to implement since the input device, e.g. a hand, wand or arm, cannot present the gestures to the user. In contrast, for touch gestures, feed-forward mechanisms (such as Marking Menus or OctoPocus) have been shown to effectively support user awareness and learning. In this paper, we explore whether touch gesture input can be leveraged to teach users to perform mid-air gestures. We show that marking menu touch gestures transfer directly to knowledge of mid-air gestures, allowing performance of these gestures without intervention. We argue that cross-modal learning can be an effective mechanism for introducing users to mid-air gestural input.

CCS CONCEPTS

• **Human-centered computing → Gestural input**;

KEYWORDS

Motion input; Mobile interface; Motor learning; Training

ACM Reference format:
Jay Henderson[1], Sachi Mizobuchi[2], Wei Li[2], Edward Lank[1]. 2019. Exploring Cross-Modal Training via Touch to Learn a Mid-Air Marking Menu Gesture Set. In *Proceedings of 21st International Conference on Human-Computer Interaction with Mobile Devices and Services, Taipei, Taiwan, October 1–4, 2019 (MobileHCI '19),* 9 pages. https://doi.org/10.1145/3338286.3340119

MobileHCI '19, October 1–4, 2019, Taipei, Taiwan

ACM ISBN 978-1-4503-6825-4/19/10…$15.00
https://doi.org/10.1145/3338286.3340119

1 INTRODUCTION

Within the field of human-computer interaction, gestural input has been an active area of investigation for many years. Gestures can be either performed on a surface, captured via touch sensing or some form of tracking, or they can be performed mid-air and captured via worn or environmental sensors and trackers. Mid-air gestures are an attractive mode of interaction in the context of internet of things (IoT) and ubiquitous computing environments (ubicomp). Free space gestures have the ability to provide eyes-free input, a benefit for many IoT devices that do not have a display to guide users, e.g. a smart light-bulb or smart thermostat. In ubiquitous interaction scenarios, mid-air gestures can free the external display to hold information pertaining to it's particular context, rather than an arbitrary gesture scaffolding.

Mid-air gestures are still rarely deployed in practice due to three primary challenges: reliability in tracking and recognition [21, 22], user fatigue [18, 35], user discomfort [2], and user awareness [20] (i.e. gestures are not self-revealing [6]). In this paper, our primary interest is in examining ways we can address the challenge of user awareness of mid-air gestures; specifically, how can we help users learn an extensive mid-air gesture set.

One primary advantage of surface gestures with respect to this challenge is that surface gestures – because they are frequently performed on a display surface via direct manipulation – have a natural visual feedback mechanism via a screen that allows feed-forward techniques [5] to guide the user to a particular gesture that they wish to complete. These feed-forward, rehearsal-based techniques generally display the structure or path of a gesture, as in Bau et al.'s OctoPocus [5] and Kurtenbach et al.'s Marking Menus [23]. Mid-air motion gestures lack a natural visual display to scaffold gesture learning unless additional hardware is deployed, which, for the majority of work, is accomplished by including displays such as screens or projections [1, 3, 4, 10, 13, 16, 20, 30, 31, 34, 40, 41]. Requiring a display constrains the interaction environment – negating IoT contexts (as mentioned prior) and/or consumes all or part of that display with gesture representations – either permanently, reducing usable display space for other uses, or temporarily, requiring some means to invoke the gesture help display. Even in the presence of these training

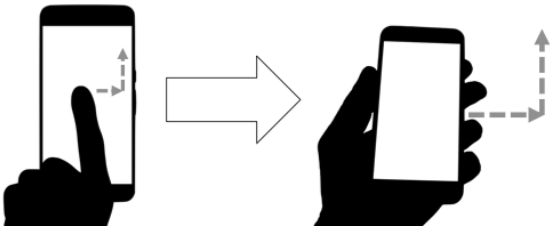


Figure 1: Visualization of transferring gestures from touch to mid-air.

mechanisms, it is often the case that training occurs as a separate task for the user [8, 20]; in contrast, feed-forward techniques allow a user to learn the gestures while performing them in context [5, 23].

In this paper we explore whether we can leverage surface-based gesture representations to teach users mid-air gestures. Leveraging marking menus [23, 41], we train users in one of two-ways: 1) We show marking menus on an external display to reveal gestures to users [20]; and 2) We show marking menus performed and displayed on a touch-screen and explore whether users can map 2D actions learned on the touchscreen onto mid-air actions, i.e. *cross-modal* training. We evaluate both the error rate and speed of interaction and find no statistically significant differences between touch and mid-air training on mid-air performance of gestures. We find, somewhat counter-intuitively, that error rate is unaffected for participants trained on a mid-air gesture set using touch-based training. The only cost we observe in touch-based training of mid-air gestures is in the first block of the experimental phase of our two-phase (training and experimental phases) study, where participants' speeds differed significantly for only that first block as they move from one modality (touch) in the training phase to a new modality (mid-air) in the experimental phase [11]. To the best of our knowledge, this work represents the first instance of evaluating how well touch can be leveraged to train users to perform mid-air gestures.

2 BACKGROUND AND RELATED WORK

Our work sits at the intersection of device-motion gestural interaction and the learning of gestural interfaces. In this section, we describe related work in mid-air gesture input, learning, and marking menus.

Mid-Air Motion Interaction

Early mid-air gesture systems follow a common pattern: pointing in a direction followed by gesturing to indicate an action [7, 9, 42]. Wilson and Shafer's XWand [42] was designed for an intelligent environment, where users would point at a location in a room (recognized using stereo-vision) and then complete a gesture, for example pointing at a television set followed by rotating the wand to indicate volume up or down. The VisionWand [9] used classical computer vision colour detection and stereo-vision to obtain orientation information with an extensive gesture set allowing rotations, pull, push, tap, tilt and selection via a pie menu. All of these follow the classic multimodal paradigm introduced by Bolt [7] of pointing followed by action invocation.

Pointing interactions have flourished in gaming systems with devices like Nintendo's Wiimote [43], a ray-casting device for interaction mid-air. Input is captured by in-device sensors including an accelerometer, gyroscope, and IR emitter. Despite the Wiimote being a dedicated ray-casting device, Pietroszek et al. showed a smartphone had similar performance [29]. Vatavu et al. compared user-defined gestures through free-form (i.e. hand movement) and via a handheld device. Users deemed handheld gestures less difficult to execute, which the researchers hypothesize is due to their familiarity with such devices [39]. Jakobsen et al. compared touch to mid-air techniques for large display interaction. While their analysis found touch superior, their results suggest situations where mid-air techniques would be optimal (e.g. walking to type on a keyboard) [19].

Learning

Cockburn et al. review four domains of research that help users transition from novice to expert modes of interaction [11]. Two of these are applicable to the current work: intramodal improvement, which is concerned with the performance improvement of a single interactive method, and intermodal improvement, which is concerned with ways to assist users in switching to a faster method of accessing a particular action - in this case, switching between surface and mid-air interaction of the same interactive method. While intermodal involves switching to a faster method, the initial switch in input method is often accompanied by a dip in performance [11, 33]. The intramodal improvement domain includes rehearsal based interfaces, which have long been used to guide users on how to perform gestural interaction. The principle of rehearsal states that "making physical actions in novice actions as similar as possible to the form of the expert's actions will facilitate skill development" [11].

In their seminal work on the Charade system, Baudel and Beaudoin-Lafon [6] note that one primary problem with mid-air gestures is that gestures are not self-revealing - the user must know the set of gestures that the system can recognize and their associated functionality. Most often, systems designed to teach mid-air interaction techniques require the use of additional hardware to *reveal* gestures to the user, usually in a visual [1, 3, 4, 10, 12, 13, 15, 16, 20, 30, 34, 36, 40],

auditory [25, 30], or haptic form [30, 34]. Mirrored representations of the user are one common form of user training [1, 3, 4, 12, 13, 20, 31]. In terms of ubiquitous public display interaction, Vatavu suggested not requiring users to learn at all, but, rather, to use a *preferred, familiar* gesture set that is individual to each user. For example, a user who often uses a Kinect gaming system may use mid-air gestures that they have used previously, where as a user who only uses touch screen manipulations such as *pan*, *tap*, or *pinch* could use those gestures [37].

Marking Menus

Intramodal improvement is often embodied as a rehearsal based interface, leveraging feed-forward techniques to guide users on how to perform interaction. One of the most well researched examples of a feed-forward, rehearsal-based input is Kurtenbach et al.'s Marking Menu [23, 24], which uses directional gestures (or "strokes" on a 2D surface) to indicate a selection in a hierarchical menu. Users begin to learn through a graphical representation of the menu, displayed as they perform the directional strokes. The marks used to activate commands are not self-revealing [24]; therefore learning a mark involves memorizing it's mapping to a command, like an accelerator key but lacking a mnemonic device. Once users memorize the spatial content of the menu, they no longer need the visible menu, and interact based on associating the directional strokes to the desired selection. Thus, the user performs an identical interaction, whether the GUI is visible or not.

While marking menus are a relatively straight-forward set of two-dimensional gestures, significant work exists in exploring their use in mid-air interaction [26, 41]. As one example, Oakley et al. conceptualized a mid-air motion marking menu, using a smart device, with a 90 degree range of motion [26, 27]. In a between subjects study, they used a feed-forward UI either on a distant screen (a laptop in front of the user) or on the smart device [27]. They note the local screen was more difficult to observe, due to the nature of the interaction, and that participants found it more time consuming to read experimental instructions. However, performance was not greatly effected by screen placement [27].

User interface guidelines such as those for Microsoft's Windows and Google's Android stress the importance of consistent touch gestures across applications so that users can transfer knowledge between different contexts [14]. A related question is whether, if the gesture is consistent across modalities, touch and device motion, i.e. in the same 2D shape, users can transfer this knowledge from one modality to another. In other words, it is an open question as to whether rehearsing a touch-based gesture can serve as an effective guide to learning mid-air motion paths.

Synthesis and Problem Space

The above observations give rise to a rationale for exploring cross-modal training. First, we believe, alongside many past researchers, that mid-air gestural interaction is effective in interactive environments because it supports target selection and action in a single, unified input modality. Second, marking menus are an effective and elegant way to scalably learn rich command sets due to their ability to map commands into submenus that can be represented as compound actions to be sensed and captured. Finally, while personal devices – via motion gestures [26, 32] – are an effective means of capturing and mapping gestural commands, as Oakley et al. note [27], there are challenges with teaching users mid-air inputs using a touchscreen representation because, during performance, the screen may not be visible. For us, the attendant question is then: Can we leverage touch-based marking menu training to teach users mid-air marking menu input?

3 ASSESSING TOUCH-BASED TEACHING OF MID-AIR GESTURES

In past work, the primary mechanism for providing feedback and teaching users to perform mid-air gestures is through a visual representation of the current and required movement path, typically via an external display [1, 3, 4, 10, 12, 13, 16, 20, 30, 40]. While this makes sense in environments where external displays exist, in other environments (e.g. environments populated with IoT artifacts), it may be the case that the environment does not have physical displays and provides output more subtly. The other tension within this space is, if we wish to train with surface gestures, then why move to motion gestures? We believe that, alongside motion gestures representing an effective means for combining target and command, motion gestures can serve as an eyes-free shortcut to command with the attendant benefit that the touchscreen is not impacted. We view motion gesture input as a form of short-cut, analogous to a system-wide hotkey, to support dedicated gestural input.

While our earlier discussion treated mid-air gestural input as a generalized input modality, in our evaluation we focus specifically on motion gesture input, i.e. mid-air gestural input where the user moves a mobile device in order to issue commands. Our rationale for this is similar to that of Vatavu et al. [38]: users are familiar with their personal devices, almost always have them with them, and we can leverage the display as an opportunistic training platform and as a convenient, on-hand, input sensor to capture interactions. That said, we believe our results should generalize to bare-hand mid-air input, provided the environment supports hand-tracking to capture gesture input.

In this section, we describe participants, apparatus, and an experiment that explores the comparison of touch and mid-air training of mid-air gestures in detail.

Participants

Fourteen participants between the ages of 21 and 28 volunteered for the study. Average age was 24.47 (SD = 2.36). Six participants identified as female and the remaining eight identified as male. Participants were remunerated $15 for their participation. 13 participants were post-secondary students and one was a teacher. Five participants had experience with marking menus (e.g. Maya, Pinterest for iPad, or other application contexts) and the remaining nine did not.

Apparatus

The visual interface was displayed either on an ASUS PB287Q monitor with 1080p resolution using an Nvidia Shield TV running Android 8.0 or on a smartphone. Participants interacted using a Huawei Nexus 6P running Android 8.0, with dimensions: 159.3 × 77.8 × 7.3 mm, a weight of 178g and a 5.7" screen. Cross-device communication was facilitated through a dedicated WiFi network.

Mid-Air Pointing

Mid-air interaction was captured through sensor fusion on the mobile device, i.e. obtaining rotation of the Android smartphone as described in [28]. Our rotation technique maps changes in device orientation on the Yaw and Pitch axis, which can then be converted to a 2D position relative to the center of the display. Participants were asked to keep the device's roll with the screen facing up within 45 degrees in each direction, as this facilitated stability in orientation detection.

Task and Stimulus

The experiment was a 2-factor between-subjects experiment. Participants completed a demographic questionnaire followed by a two phase study, a training phase that leverages touch of mid-air followed by an experimental phase involving mid-air input.

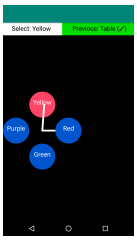
In the *training phase* participants had a visual representation of a 4 × 4, 16 item marking menu. For each trial in the training phase, a prompt was displayed on the top-left corner of the screen indicating what selection to make. Participants selected a command based on the prompt, which then displayed on the top-right hand side of the screen, in either green with a check-mark or red with an ×, to indicate a correct or incorrect selection, respectively. Each new prompt was one of 8 items from the marking menu and only displayed upon correct selection. Prompts were presented in random order for each block. Short breaks were given every 32 correct selections (4 blocks). The *training phase* consisted of 20 blocks of 8 selections for 160 selections in total. Participants were assigned to one of two conditions in the *training phase*, either Touch or Mid-Air, as follows.

- **Touch**: Participants learned the marking menu via a touch interface on mobile device. They were instructed to press their finger down on the screen, navigate through the menu (drawing a gesture on the screen), until they have reached the item they wish to select. Once they have completed the gesture, they release their finger. The experimenter gave a brief demonstration with no interface on the mobile device's screen of how to complete a gesture. The interface was displayed on the mobile device screen.
- **Mid-Air**: Participants learned the marking menu via an external display positioned in front of them, motivated by prior work [1, 3, 4, 10, 12, 13, 16, 20, 30, 40]. They were instructed to navigate through a menu, using a mobile device as a "wand" to point where to select on the screen. Their movement path was displayed on the screen as they completed a motion. To begin a gesture, they press down on the volume down button on the device. When they have completed the gesture, they release the button. The experimenter gave a brief demonstration with no interface on the monitor of how to complete a gesture. Menu and motion path are displayed on the external display (to avoid obfuscation as in the Oakley et al. training [27]).

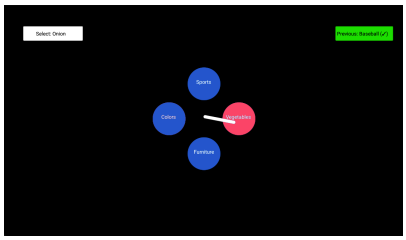
In both conditions, participants were asked to learn the menu as best as possible, because in the *experiment phase* they would have to complete the gestures in Mid-Air with no visual interface guiding them. In the *training phase* participants were required to select the correct target before moving to the next selection.

The *experiment phase* followed the exact procedure of the *training phase*, with the exception that no visual marking menu interface was displayed (Figure 2c) and prompts changed upon every selection made (regardless of correctness, to keep a consistent experiment time and reduce frustration in the case participants did not learn the menu). The experiment phase was always conducted with Mid-Air gestures. Prompts and correct/incorrect indicators were displayed on the external monitor. The experimenter gave a brief demonstration with no interface on the monitor of how to complete a Mid-Air gesture, but participants were not permitted to confirm whether or not their interpretation was correct.

In each phase, participants were permitted a break after 32 (4 BLOCKS × 8) selections. After each phase, both *training* and *experiment*, participants completed the NASA-TLX [17].



(a) Training phase interface for the Touch condition, showing the dragged motion path in white.



(b) Training phase for the Mid-Air condition on a monitor (external display), showing the movement motion path (i.e. pointer path) from the mobile device.

(c) Experiment phase interface, always in Mid-Air on a monitor (external display). No visual guidance is provided.

Figure 2: Interface displays used in the study.

Design and Analysis

Our 2-factor between-subjects experiment included the following parameters: Training Condition (Touch vs. Mid-Air, between-subjects), × Blocks (20) × commands (8) × participants (14) for a total of 2240 command selections in the experiment phase.

Our analysis focuses on whether cross-modal training (i.e. learning mid-air gestures via touch input) is as effective as learning mid-air gestures via mid-air training. As a result, dependent measures are error rate and time to perform mid-air gestures in the experimental phase of the experiment. We also assess, quantitatively, whether mid-air training provides a statistical advantage versus touch training for learning mid-air gestures, and qualitatively, on how large the advantage of consistency in training (mid-air to mid-air) is compared to cross-modal (touch to mid-air) training.

4 RESULTS

Our quantitative analysis consists of the error rate and timing of mid-air gestures. Error rate represents the fraction of mid-air gestures ([0, 1]) performed correctly in the experimental block given two training conditions – touch or mid-air training. Likewise, time represents the time from prompt or from gesture initiation for mid-air gesture input in the experiment phase given each training condition in the training phase.

Before beginning our analysis, we performed a power and sample size determination. We set a threshold of 100ms for a temporal cost that would represent a significant difference in temporal performance. The standard deviation of our data set for time from prompt to selection was 1876ms and time from beginning a gesture to selection was 804ms. Because our data was not normally distributed, we applied a log-normal transform of our data, yielding a normal distribution. We found that, to support a 95% confidence interval for a 2-tailed analysis of effects, we required a sample size of at least 5 participants per condition, or 10 participants. Our data set of 14 participants exceeds this threshold.

Error Rate

Figure 3 shows the fraction of gestures performed correctly in each condition. When participants train to perform mid-air gestures via mid-air training, 89% of gestures are performed correctly; interestingly, with cross-modal or touch training for mid-air gestures, the rate of correct gestures is slightly *higher*, 94%. We performed a χ^2 analysis of error rate and found that the difference was not significant ($p = 0.45$), indicating no significant difference in error rate for mid-air versus touch training of mid-air gestures.

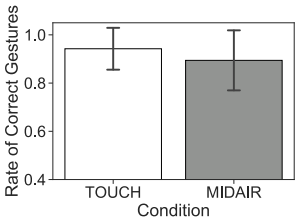


Figure 3: Rate of correct selections across conditions (error bars indicate SD).

Time

We ran an independent samples t-test on the mean time from prompt to selection and mean time from beginning a gesture to selection for participants in our study. Figures 4 and 5 show the length of time from prompt to command selection and from beginning of gesture (as measured by displacement of the device) and command selection. Differences were not significant ($T_{12} = -0.076, p = 0.94$ and $T_{12} = 0.085, p = 0.93$ respectively). Upon observation, time from prompt to gesture was slightly shorter and time to gesture slightly longer for the touch training condition, but neither measure exhibits statistical significance.

We also analyzed gesture time per block. One factor that does seem to differentiate touch-based training of mid-air gestures is that the first mid-air gesture performed in the experimental block is significantly slower (longer time) than

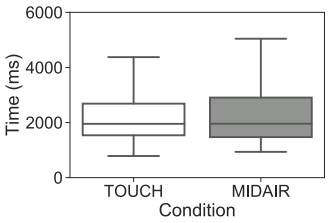


Figure 4: Time from prompt appearing to selection across conditions.

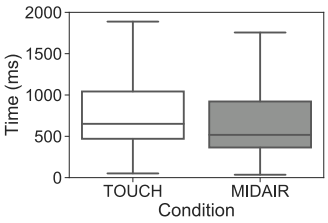


Figure 5: Time from beginning a gesture to selection across conditions.

subsequent blocks. Figure 6 demonstrates this effect, a result of the initial cost of switching modalities [11]. However, participants quickly converged on equivalent performance, and, by the second and subsequent blocks, performance was indistinguishable.

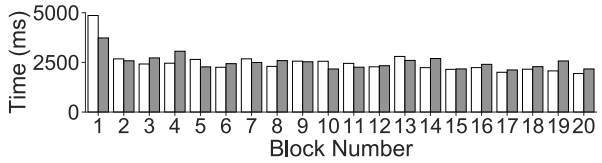
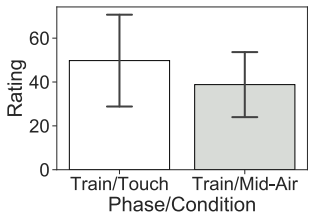
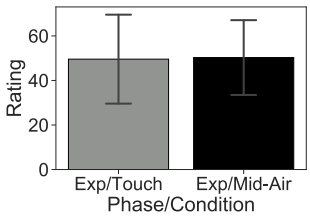


Figure 6: Time from prompt appearing to selection across blocks by condition.



(a) Training phase.



(b) Experimental phase.

Figure 7: Overall NASA TLX scores across phases and conditions (error bars indicate SD). Training phase conditions are performing via touch or mid-air. Experimental phase is always performed in mid-air, conditions are mode which training phase was completed.

NASA TLX

Recall after each phase in the user study, training and experiment, we asked participants to complete the six category NASA Task Load Index [17]. We performed a multivariate analysis (MANOVA) on reported TLX scores and found no significance across each study phase (*training* or *experiment*) and each condition (Touch or Mid-Air). These results are depicted in Figure 7. Between subjects tests indicated significance for reported physical demand scores ($F_{3,24} = 3.869$, $p < 0.05$). Post-hoc analysis using Tukey's HSD indicated a difference in physical demand between the *training* and *experimental* phase for participants who trained using touch and between *training* via touch and the *experiment* in mid-air (after mid-air training).

5 DISCUSSION

With quantitative analyses as presented in results, it is often desirable to examine hypotheses, but, in this work, our goal was slightly different. We expected touch-based training to be worse than mid-air training for mid-air gestural input, and our goal was to quantitatively assess a qualitative question: *How much worse is touch-based training for learning mid-air input?*.

Surprisingly, our results argue that touch training is as effective as mid-air training to learn mid-air motion gesture input. There is no statistically significant difference between error rate and time, and, qualitatively, overall values of performance appear similar in the experimental phase: touch training results in slightly higher accuracy scores and both training mechanisms exhibit similar interaction time. Stated more succinctly, we found that participants were able to effectively learn marking menu gestures through an alternative mechanism (touch).

Our control condition, of teaching mid-air gestures from scaffolding on a distant display, echoes findings of previous literature – users are able to learn bodily motions, gestures, in particular – through a representation of the required movement path on an external display [1, 3, 4, 20, 30, 31]. Our result – of cross-modal learning – echoes findings of Kamal et al., who showed that an alternative mechanism of revealing a gesture (on-device video plus recognizer feedback) was as effective as instruction via an external display representation of the gesture [20]. We note the difficulty in comparing our findings with the literature as we are unaware of prior studies that observe learning free-space (mid-air) gestures via touch.

In our view, mid-air gestural input as we have formulated the problem has much in common with keyboard accelerators. Users typically use one, sub-optimal modality (e.g. a menu or toolbar) to perform commands. However, use of these sub-optimal commands provides them with an awareness of an expert mode - a keyboard accelerator - that can sufficiently enhance performance. Cockburn et al. [11], in studying this learning of expert performance, note that moving from a novice to expert mode may have a performance cost, and we do see a brief performance cost in our experimental phase during the first block of eight gestures as participants habituated themselves to the new input modality.

However, after that one habituation block, participant performance converged to being qualitatively indistinguishable for the remaining blocks in the experimental phase.

There are some potential slight differences in NASA TLX ratings that might merit further investigation. In Figure 7, it is observed that training in mid-air resulted in the lowest overall average TLX score. Participants who trained in the touch condition provided some insight of why this may be, for instance, "the occlusion of the menu by my finger" [P16, trained via touch]. Since they were aware that they would have to later perform gestures in mid-air, one participant noted "learning the gestures on the phone in part 1 was somewhat stressful" [P10, trained via touch].

Future Work

While our results demonstrate that we can leverage touch-based training to teach in-air marking menus, the marks represented in marking menus and evaluated in our study are comprised only of two straight-line segments where the two segments vary only in direction. One area of future work is to assess cross-modal gesture learning on more complex gesture sets, including ideographic, alphanumeric and multi-stroke gesture sets [35] to see if our findings hold. Furthermore, alongside mid-air gestures that can be rendered on a 2D plane (i.e. planar gestures, as in the gestures by Siddhpuria et al. [35]), one could imagine teaching 3-dimensional rotation gestures with a hand-held device, where the user could rotate their finger or add an additional finger to indicate a rotation and we could explore mappings to mid-air such that non-planar mid-air gestures could be trained via touch. Another area extending from the current work is the incorporation of haptic feedback with our cross-modal learning technique [30, 34]. Exploration of this space could include haptic feedback while learning on a touch surface in an effort to further ingrain gestures into a users memory, providing haptic effects while interacting in mid-air, or using both of these in conjunction to create a more congruent mapping between modalities.

6 LIMITATIONS

In our study, the participant sample size is limited. To counter this, we performed a sample size power estimate to ensure that our sample generates sufficient statistical power to identify discrepancies at the level we wish. Alongside this, we also note that, qualitatively, sample size is a highly questionable critique given that the error rate for touch training is actually lower than the error rate for mid-air training and given that temporal profiles are so similar – again touch training resulting in lower prompt to selection timing. Thus, in conjunction with our sample size power estimate, it would be highly unlikely to identify significant differences with a higher sample size, meaning that our likelihood of type 2 error is quite low.

7 CONCLUSION

One challenge with mid-air gesture input is that gestures are not self-revealing, so to teach users gestures, it is common to use an external display and/or tracking to provide guidance and feedback to the user. In this work, we examine whether or not we can leverage another modality, touch, to teach users a mid-air gesture set. Leveraging the paradigm of free-space marking menus, and a smartphone, as a motion-gesture-style input device, we explore how well learning in touch transfers to mastery in mid-air input when compared to learning and performing in mid-air. Our results argue that transferring spatial knowledge of marking menus between touch and mid-air exhibits similar performance as obtaining such knowledge in mid-air, with only minimal, short-term performance cost.

REFERENCES

[1] Christopher Ackad, Andrew Clayphan, Martin Tomitsch, and Judy Kay. 2015. An In-the-wild Study of Learning Mid-air Gestures to Browse Hierarchical Information at a Large Interactive Public Display. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. ACM, New York, NY, USA, 1227–1238. https://doi.org/10.1145/2750858.2807532

[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 (MobileHCI '14)*. ACM, New York, NY, USA, 193–202. https://doi.org/10.1145/2628363.2628381

[3] Florian Alt, Sabrina Geiger, and Wolfgang Höhl. 2018. ShapelineGuide: Teaching Mid-Air Gestures for Large Interactive Displays. In *Proceedings of the 7th ACM International Symposium on Pervasive Displays (PerDis '18)*. ACM, New York, NY, USA, Article 3, 8 pages. https://doi.org/10.1145/3205873.3205887

[4] Fraser Anderson, Tovi Grossman, Justin Matejka, and George Fitzmaurice. 2013. YouMove: Enhancing Movement Training with an Augmented Reality Mirror. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology (UIST '13)*. ACM, New York, NY, USA, 311–320. https://doi.org/10.1145/2501988.2502045

[5] Olivier Bau and Wendy E. Mackay. 2008. OctoPocus: A Dynamic Guide for Learning Gesture-based Command Sets. In *Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology (UIST '08)*. ACM, New York, NY, USA, 37–46. https://doi.org/10.1145/1449715.1449724

[6] Thomas Baudel and Michel Beaudouin-Lafon. 1993. Charade: Remote Control of Objects Using Free-hand Gestures. *Commun. ACM* 36, 7 (July 1993), 28–35. https://doi.org/10.1145/159544.159562

[7] Richard A Bolt. 1980. *"Put-that-there": Voice and gesture at the graphics interface*. Vol. 14. ACM.

[8] Andrew Bragdon, Robert Zeleznik, Brian Williamson, Timothy Miller, and Joseph J. LaViola, Jr. 2009. GestureBar: Improving the Approachability of Gesture-based Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. ACM, New York, NY, USA, 2269–2278. https://doi.org/10.1145/1518701.1519050

[9] Xiang Cao and Ravin Balakrishnan. 2004. VisionWand: Interaction Techniques for Large Displays Using a Passive Wand Tracked in 3D. *ACM Trans. Graph.* 23, 3 (Aug. 2004), 729–729. https://doi.org/10.1145/1015706.1015788

[10] Christopher Clarke and Hans Gellersen. 2017. MatchPoint: Spontaneous Spatial Coupling of Body Movement for Touchless Pointing. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology (UIST '17)*. ACM, New York, NY, USA, 179–192. https://doi.org/10.1145/3126594.3126626

[11] Andy Cockburn, Carl Gutwin, Joey Scarr, and Sylvain Malacria. 2014. Supporting Novice to Expert Transitions in User Interfaces. *ACM Comput. Surv.* 47, 2, Article 31 (Nov. 2014), 36 pages. https://doi.org/10.1145/2659796

[12] William Delamare, Céline Coutrix, and Laurence Nigay. 2015. Designing Guiding Systems for Gesture-based Interaction. In *Proceedings of the 7th ACM SIGCHI Symposium on Engineering Interactive Computing Systems (EICS '15)*. ACM, New York, NY, USA, 44–53. https://doi.org/10.1145/2774225.2774847

[13] William Delamare, Thomas Janssoone, Céline Coutrix, and Laurence Nigay. 2016. Designing 3D Gesture Guidance: Visual Feedback and Feedforward Design Options. In *Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI '16)*. ACM, New York, NY, USA, 152–159. https://doi.org/10.1145/2909132.2909260

[14] Tilman Dingler, Rufat Rzayev, Alireza Sahami Shirazi, and Niels Henze. 2018. Designing Consistent Gestures Across Device Types: Eliciting RSVP Controls for Phone, Watch, and Glasses. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 419, 12 pages. https://doi.org/10.1145/3173574.3173993

[15] Dustin Freeman, Hrvoje Benko, Meredith Ringel Morris, and Daniel Wigdor. 2009. ShadowGuides: Visualizations for In-situ Learning of Multi-touch and Whole-hand Gestures. In *Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces (ITS '09)*. ACM, New York, NY, USA, 165–172. https://doi.org/10.1145/1731903.1731935

[16] Euan Freeman, Stephen Brewster, and Vuokko Lantz. 2016. Do That, There: An Interaction Technique for Addressing In-Air Gesture Systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 2319–2331. https://doi.org/10.1145/2858036.2858308

[17] Sandra G. Hart and L E. Stavenland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. (12 1988), 139- pages. https://doi.org/10.1016/S0166-4115(08)62386-9

[18] 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 (CHI '14)*. ACM, New York, NY, USA, 1063–1072. https://doi.org/10.1145/2556288.2557130

[19] Mikkel R. Jakobsen, Yvonne Jansen, Sebastian Boring, and Kasper Hornbæk. 2015. Should I Stay or Should I Go? Selecting Between Touch and Mid-Air Gestures for Large-Display Interaction. In *Human-Computer Interaction – INTERACT 2015*, Julio Abascal, Simone Barbosa, Mirko Fetter, Tom Gross, Philippe Palanque, and Marco Winckler (Eds.). Springer International Publishing, Cham, 455–473.

[20] Ankit Kamal, Yang Li, and Edward Lank. 2014. Teaching Motion Gestures via Recognizer Feedback. In *Proceedings of the 19th International Conference on Intelligent User Interfaces (IUI '14)*. ACM, New York, NY, USA, 73–82. https://doi.org/10.1145/2557500.2557521

[21] Keiko Katsuragawa, Ankit Kamal, Qi Feng Liu, Matei Negulescu, and Edward Lank. 2019. Bi-Level Thresholding: Analyzing the Effect of Repeated Errors in Gesture Input. *ACM Trans. Interact. Intell. Syst.* 9, 2-3, Article 15 (April 2019), 30 pages. https://doi.org/10.1145/3181672

[22] Keiko Katsuragawa, Krzysztof Pietroszek, James R. Wallace, and Edward Lank. 2016. Watchpoint: Freehand Pointing with a Smartwatch in a Ubiquitous Display Environment. In *Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI '16)*. ACM, New York, NY, USA, 128–135. https://doi.org/10.1145/2909132.2909263

[23] Gordon Kurtenbach and William Buxton. 1994. User Learning and Performance with Marking Menus. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '94)*. ACM, New York, NY, USA, 258–264. https://doi.org/10.1145/191666.191759

[24] Gordon Paul Kurtenbach. 1993. *The Design and Evaluation of Marking Menus*. Ph.D. Dissertation. Toronto, Ont., Canada, Canada. UMI Order No. GAXNN-82896.

[25] Sarah Morrison-Smith, Megan Hofmann, Yang Li, and Jaime Ruiz. 2016. Using Audio Cues to Support Motion Gesture Interaction on Mobile Devices. *ACM Trans. Appl. Percept.* 13, 3, Article 16 (May 2016), 19 pages. https://doi.org/10.1145/2897516

[26] Ian Oakley and Junseok Park. 2007. A Motion-based Marking Menu System. In *CHI '07 Extended Abstracts on Human Factors in Computing Systems (CHI EA '07)*. ACM, New York, NY, USA, 2597–2602. https://doi.org/10.1145/1240866.1241048

[27] Ian Oakley and Junseok Park. 2009. Motion Marking Menus: An Eyes-free Approach to Motion Input for Handheld Devices. *Int. J. Hum.-Comput. Stud.* 67, 6 (June 2009), 515–532. https://doi.org/10.1016/j.ijhcs.2009.02.002

[28] Alexander Pacha. 2013. Sensor fusion for robust outdoor Augmented Reality tracking on mobile devices. (2013).

[29] Krzysztof Pietroszek, Anastasia Kuzminykh, James R. Wallace, and Edward Lank. 2014. Smartcasting: A Discount 3D Interaction Technique for Public Displays. In *Proceedings of the 26th Australian Computer-Human Interaction Conference on Designing Futures: The Future of Design (OzCHI '14)*. ACM, New York, NY, USA, 119–128. https://doi.org/10.1145/2686612.2686629

[30] Otniel Portillo-Rodriguez, Oscar O. Sandoval-Gonzalez, Emanuele Ruffaldi, Rosario Leonardi, Carlo Alberto Avizzano, and Massimo Bergamasco. 2008. Real-Time Gesture Recognition, Evaluation and Feed-Forward Correction of a Multimodal Tai-Chi Platform. In *Haptic and Audio Interaction Design*, Antti Pirhonen and Stephen Brewster (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 30–39.

[31] Gustavo Rovelo, Donald Degraen, Davy Vanacken, Kris Luyten, and Karin Coninx. 2015. Gestu-Wan - An Intelligible Mid-Air Gesture Guidance System for Walk-up-and-Use Displays. In *INTERACT*.

[32] Jaime Ruiz, Yang Li, and Edward Lank. 2011. User-defined Motion Gestures for Mobile Interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 197–206. https://doi.org/10.1145/1978942.1978971

[33] Joey Scarr, Andy Cockburn, Carl Gutwin, and Philip Quinn. 2011. Dips and Ceilings: Understanding and Supporting Transitions to Expertise in User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 2741–2750. https://doi.org/10.1145/1978942.1979348

[34] Christian Schönauer, Kenichiro Fukushi, Alex Olwal, Hannes Kaufmann, and Ramesh Raskar. 2012. Multimodal Motion Guidance: Techniques for Adaptive and Dynamic Feedback. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction (ICMI '12)*. ACM, New York, NY, USA, 133–140. https://doi.org/10.1145/2388676.2388706

[35] Shaishav Siddhpuria, Keiko Katsuragawa, James R. Wallace, and Edward Lank. 2017. Exploring At-Your-Side Gestural Interaction for Ubiquitous Environments. In *Proceedings of the 2017 Conference on Designing Interactive Systems (DIS '17)*. ACM, New York, NY, USA, 1111–1122. https://doi.org/10.1145/3064663.3064695

[36] Rajinder Sodhi, Hrvoje Benko, and Andrew Wilson. 2012. LightGuide: Projected Visualizations for Hand Movement Guidance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 179–188. https://doi.org/10.1145/2207676.2207702

[37] Radu-Daniel Vatavu. 2012. Nomadic Gestures: A Technique for Reusing Gesture Commands for Frequent Ambient Interactions. *J. Ambient Intell. Smart Environ.* 4, 2 (April 2012), 79–93. http://dl.acm.org/citation.cfm?id=2350758.2350765

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

[39] Radu-Daniel Vatavu. 2013. A Comparative Study of User-defined Handheld vs. Freehand Gestures for Home Entertainment Environments. *J. Ambient Intell. Smart Environ.* 5, 2 (March 2013), 187–211. http://dl.acm.org/citation.cfm?id=2594684.2594688

[40] David Verweij, Vassilis-Javed Khan, Augusto Esteves, and Saskia Bakker. 2017. Multi-User Motion Matching Interaction for Interactive Television Using Smartwatches. In *Adjunct Publication of the 2017 ACM International Conference on Interactive Experiences for TV and Online Video (TVX '17 Adjunct)*. ACM, New York, NY, USA, 67–68. https://doi.org/10.1145/3084289.3089906

[41] Robert Walter, Gilles Bailly, and Jörg Müller. 2013. StrikeAPose: Revealing Mid-air Gestures on Public Displays. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 841–850. https://doi.org/10.1145/2470654.2470774

[42] Andrew Wilson and Steven Shafer. 2003. XWand: UI for Intelligent Spaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '03)*. ACM, New York, NY, USA, 545–552. https://doi.org/10.1145/642611.642706

[43] C. A. Wingrave, B. Williamson, P. D. Varcholik, J. Rose, A. Miller, E. Charbonneau, J. Bott, and J. J. LaViola. 2010. The Wiimote and Beyond: Spatially Convenient Devices for 3D User Interfaces. *IEEE Computer Graphics and Applications* 30, 2 (March 2010), 71–85. https://doi.org/10.1109/MCG.2009.109