Sridhar, Srinath; Feit, Anna; Theobalt, Christian; Oulasvirta, Antti

Investigating the Dexterity of Multi-Finger Input for Mid-Air Text Entry

Published in:
33rd Annual ACM Conference on Human Factors in Computing Systems, CHI'15, Seoul, Korea, April 18 - 23, 2015

DOI:
10.1145/2702123.2702136

Published: 01/01/2015

Document Version
Peer reviewed version

Please cite the original version:
Sridhar, S., Feit, A., Theobalt, C., & Oulasvirta, A. (2015). Investigating the Dexterity of Multi-Finger Input for Mid-Air Text Entry. In *33rd Annual ACM Conference on Human Factors in Computing Systems, CHI'15, Seoul, Korea, April 18 - 23, 2015* ACM. https://doi.org/10.1145/2702123.2702136

Investigating Multi-Finger Gestures for Mid-Air Text Entry

Anna Maria Feit[1] Srinath Sridhar[2] Christian Theobalt[2] Antti Oulasvirta[1]
[1]Aalto University
{anna.feit,antti.oulasvirta}@aalto.fi
[2]Max Planck Institute for Informatics
{ssridhar,theobalt}@mpi-inf.mpg.de

ABSTRACT
Mid-air input, enabled by recent progress in computer vision based marker-less hand tracking, is an exciting candidate for text entry where direct touch input is not practicable or not available. In this paper we investigate the use of chord-like multi-finger gestures for entering text in mid-air. In contrast to previous methods, they require no extrinsic targets and can be performed eyes-free. We systematically explore the design space of hand gestures by computationally optimizing the letter-to-gesture mapping with respect to multiple objectives: performance, anatomical comfort, learnability and mnemonics. First investigations of one optimization case show entry rates of 22 words per minute. While this is promising, our study reveals several limitations of both, the mapping design as well as the available tracking methods. We discuss open challenges in mid-air input and conclude with recommendations for future work.

Categories and Subject Descriptors
H.5.m. [**Information Interfaces and Presentation (e.g. HCI)**]: Miscellaneous

Keywords
Mid-air interaction; multi-finger gestures; text entry;

1. INTRODUCTION
The hand is the most dexterous of the human extremities. The many degrees of freedom (DOFs) that it exhibits allow us to perform fast and precise movements, such as communication via sign language at rates of up to 225 words per minute (WPM) [2]. Recent advances in computer vision based, markerless hand tracking [15, 20] have increased expectations to finally exploit the full potential of the hand for computer input. In particular, mid-air text entry is an exciting and promising input modality for wearable and mobile devices, or large interactive displays. As shown in Figure 1, it can allow for freehand input for example onto a small smartwatch or a distant TV screen. In both cases direct touch is not practicable or not possible.

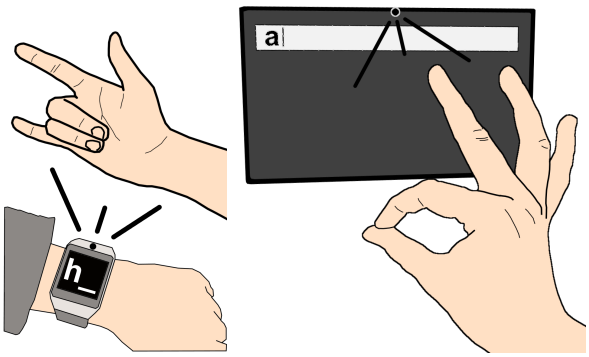


Figure 1: Mid-air input allows text entry on any device where direct touch is not practicable, such as smartwatches (left) or TV screens (right). Here, 'h' and 'a' are entered by a *multi-finger gesture*.

Previous attempts have used a single end effector for selecting 3D key targets or performing continuous gestures [10, 13, 18]. This does not exploit the full capacity of the hand, is slow and tiring. Instead, this paper studies the use of chord-like multi-finger gestures as shown in Figure 1. These rely only on proprioceptive feedback of joint angles and make use of many degrees of freedom. The method allows input without looking and is literally right at your fingertips.

We use a computational optimization method to identify gesture sets that allow for high-throughput. Building on prior work in keyboard optimization [4, 23], our previous studies [19], and existing literature, we construct a novel objective function called PALM that optimizes a letter-to-gesture mapping with respect to performance (P), anatomical comfort (A), learnability (L), and mnemonics (M). We empirically study one outcome of this approach, called FastType. We built a prototype system for tracking the hand with the Leap Motion sensor[1] and recognizing hand gestures as defined by FastType. First investigations with 10 participants showed promising input rates [19] with an average of 22 WPM and peak performances of up to 38 WPM. However, they cannot keep up yet to those of physical keyboards or sign language. We identify several challenges with respect to individual differences, learning time and technological limitations and end with a list of opportunities for future research to overcome those limitations.

[1]https://www.leapmotion.com/

ACM womENcourage '15, 24 – 25 September, Uppsala, Sweden.

2. RELATED WORK

In the following we discuss recent developments in mid-air input from two viewpoints. First, we give a quick overview of technological advances in markerless hand tracking during the last 4-5 years. Then we review previous text entry methods for mid-air input.

2.1 Markerless Hand Tracking

We limit our discussion to methods that work in real time without the use of markers. Oikonomidis *et al.* [14] proposed a method for tracking the hand using a single depth camera, which achieved frame rates of 15 frames per second (fps). Wang and Popović [22] proposed a two-camera setup and a method to track hands without gloves at real time speeds. Melax *et al.* [11] proposed a method for tracking hands directly in depth by efficient parallel physics simulations. They achieve 60 fps, suitable for interactive applications. Kim *et al.* [5] presented a method for tracking finger articulations using a wrist-worn device that can be used for simple interactions on-the-go. Recently, Qian *et al.* [15] proposed a robust method for hand tracking based on hybrid optimization that runs at 25 fps. Commercial solutions such as Three Gear[2] and Leap Motion are also capable of tracking full hand motion as a skeleton at high frame rates.

In our experiments we used the Leap Motion as the primary tracker and a modified version of Sridhar *et al.* [20]. However, our approach is independent of the tracking device, but only requires high framerate input of tracked hand motion as a *kinematic skeleton*.

2.2 Mid-Air Text Entry

Although mid-air text entry seems a naturally fast input method, previous work has shown limited performances in comparison to physical or virtual keyboards. Generally, the methods can be classified into two categories:

Selection-based techniques require pointing on external key targets either on a 2D plane or a 3D grid. Previous studies have evaluated different keyboard layouts and the QWERTY keyboard was found to be the fastest with entry rates between 13 and 19 WPM [9, 18]. Markussen *et al.* [10] extended a previously introduced shape writing keyboard [7] to mid-air. Letters were selected by continuous movements through the respective keys. They achieved a rate of 28.1 WPM on a small phrase set, using a marker-based finger tracking system. Recently there have been efforts to implement existing keyboard layouts with the Leap Motion, such as the Minuum keyboard[3] and Dasher[4].

Gesture-based techniques map free-hand gestures (or static postures) to characters or words. Different approaches have been proposed to extend handwriting and unistroke methods to 3D [6, 13, 16], reaching entry rates of 11 WPM [13]. Sign language is the most prominent example of gesture-based, word-level "text entry" with rates in the range of 175 to 225 WPM [2]. However, current technology is not yet capable of fully tracking complex sign language that employs discrete postures as well as continuous movements.

[2] `http://www.threegear.com/`

[3] `http://minuum.com/future-of-wearable-typing/`

[4] `http://www.inference.phy.cam.ac.uk/dasher`

3. MULTI-FINGER GESTURES

The above efforts show large interest in mid-air text entry. However, they cannot compete with traditional physical or even virtual keyboards. We identify several reasons for this: (1) The capacity of the hand is limited by the use of only one end-effector. (2) Extrinsic targets require high perceptual attention for accurate pointing. (3) In some methods, the 3D interaction space of the hand is reduced to a 2D plane in front of the screen. (4) Pointing and gesturing involving the whole upper extremity is slow and tiring (Gorilla arm).

To overcome these limitations, our work focuses on chord-like motions in mid-air performed by multiple fingers (Figure 1). In this input paradigm, the involved fingers are extended and flexed at a single joint to a discriminable end posture. This requires no external target, allows eyes-free input controlled by proprioception and is independent of the global position of the hand. In the following, we use the terms 'gesture' and 'posture' interchangeably, denoting a combination of the static joint angles of each finger.

3.1 Mapping Gestures to Letters

The space of possible gesture-to-letter mappings is (exponentially) large. Nevertheless, we can systematically explore it by employing a computational optimization method, as done in previous work on keyboard optimization (*e.g.* [4, 23]). Our design task is to maximize the *usability* U of a letter assignment. We assume that there are n characters and m discrete postures. As $n \ll m$, this makes our optimization task an instance of the Generalized Assignment Problem (GAP) [3]. To characterize U, we formulate a multi-objective function called PALM which addresses four factors: Performance, Anatomical Comfort, Learnability, and Mnemonics. U is then defined as a weighted sum of these factors. In the following we go through the four objectives and give an intuition on how we derive their respective scores. Further details can be found in [19].

Performance: Finger movement performance can be quantified by the time it takes for one end-effector to reach a target from a given position. Fitts' law predicts the movement time (MT) to increase logarithmically as a function of movement distance and target width. It has been highly successful for predicting MT with traditional input devices [8], such as mouse or virtual keyboards. In [19] we derive Fitts' law models for each finger and explain their use in predicting MT for *multi-finger* gestures.

Anatomical Comfort: Due to anatomical constraints, the single fingers cannot be controlled fully independently of each other. The result is *unintended co-activation* of non-instructed fingers during movement. For example, it is not possible to move the ring finger without unvoluntary movement of the little finger. Hand gestures should minimize the extent to which non-instructed fingers move, as it can cause recognition errors. Schieber [17] proposed an *index of individuation* that quantifies how independently an instructed finger can be moved from all others. In [19] we conduct a study to derive the individuation indices for each finger and describe how these can be used to measure the anatomical comfort of a multi-finger gesture.

Learnability: Learnability is an important factor to consider for any activity involving rapid and careful articula-

Finger Flexion	Character	Finger Flexion	Character
0,1,0,0,0	_	1,1,0,0,0	n
1,0,0,0,0	a	1,0,0,1,0	o
0,0,1,0,1	b	0,0,0,1,1	p
1,1,0,1,0	c	0,1,1,1,1	q
0,1,1,1,0	d	0,1,0,1,0	r
0,0,0,1,0	e	0,1,1,0,0	s
1,1,1,1,0	f	0,0,1,0,0	t
0,1,0,0,1	g	0,0,0,0,1	u
0,0,1,1,0	h	1,0,0,1,1	v
1,0,1,0,0	i	1,0,0,0,1	w
0,1,1,0,1	j	0,0,1,1,1	x
1,1,0,0,1	k	0,1,0,1,1	y
1,1,1,0,0	l	1,0,1,0,1	z
1,0,1,1,0	m		

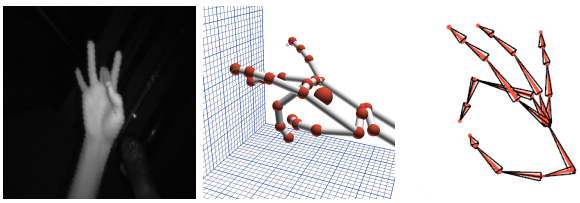
Table 1: FastType is optimized for performance, respecting anatomical constraints and learnability. The 0-1 strings describe each gesture: ordered from thumb to little finger, a 1 denotes that the finger is flexed, while 0 means extended.

tion of multiple joints. To develop a score for learnability of a gesture, we build on some prevalent theories of motor learning that view learning as a *hierarchical combination of primitives* [12]. According to this view, the brain simplifies multi-dimensional motor control by collapsing it into a few dimensions. Practicing a complex gesture gradually increases hierarchical organization and decreases reliance on feedback. The consequence is for example, that the fewer fingers a gesture involves, the easier it will be to learn. A mathematical formulation of this effect is described in [19].

Mnemonics: While our learnability score looks at motor learning "from scratch", the mnemonics score focuses on easily memorable gestures. Studies of human memory suggest that categorization, chunking, and mnenomics help forming more durable long-term memory traces among otherwise unrelated materials [21]. To identify finger mnemonics, we build on a recent study of multi-finger chord gestures that showed a positive effect on learning when assigning gestures with respect to mnemonics families [21]. In particular, we include the following mnemonics families: neighboring fingers (*e.g.* thumb and little finger together), base (*e.g.* thumb with other fingers), and one single finger. The combination of these rules into a mnemonics score M is described in [19].

3.2 Outcome: FastType

The optimization method can be used to generate different mappings, depending on the weighting of the objectives. The exploration of the design space by varying these weights is described in [19]. There we also compare the predicted performance of various outcomes with that of existing methods, such as fingerspelling. Here we concentrate on only one of the outcomes: FastType maps the letters *a-z* (including space), to multi-finger gestures including all 5 fingers. FastType is optimized with a focus on performance, lowering the weights of the other objectives. Expert motor performance is predicted in [19] to be 54.7 WPM. The mapping is shown in Table 1. Each letter is assigned a 0-1 string which describes the gesture. Ordered from thumb to little finger, a 1 denotes flexion of the finger, while 0 means comfortably extended. Optimized for performance, the mapping ensures that frequent letter pairs, such as 'th', are mapped to gestures that require minimal movement for transition.



(a) (b) (c)

Figure 2: Tracking challenge: Error made by the Leap Motion tracking a pinching posture. (a) The input posture, (b) the posture as recognized by the Leap Motion tracker (c) the true posture.

4. INVESTIGATING PERFORMANCE

In order to investigate the performance achievable with mid-air text entry in practice, we conducted a preliminary evaluation of FastType with 10 users. We followed a word-level paradigm previously used by Bi *et al.* [1]. Here, a randomly sampled word is practiced until performance peaks. The benefit of this is that the upper boundary of entry performance can be estimated even without having to learn the full gesture set. We built a prototype that allowed users to enter text and recorded performance of typed words. Our gesture recognizer used joint angle data from the Leap Motion and used a combination of dwell times and signal peak detection to detect when users made a particular posture. Further details can be found in [19].

Results: Overall, the 10 users entered 53 words at an average peak performance of 22.25 WPM (SD 8.9). For analyzing the peak performance of each word, we extracted the top three repetitions with an error rate less than 15% (measured by Damerau-Levenshtein distance). Three words had to be excluded due to this restriction. The remaining words were typed with an average error rate of 2.3% (SD 0.04). Average peak performances of users were significantly different (one-way ANOVA on WPMs: $F(9, 49) = 7.68$, $p < 0.001$), ranging from 13 WPM to 38.1 WPM.

5. DISCUSSION

This first exploration of mid-air text entry with multi-finger gestures showed promising entry rates, motivating the need for further research in this area. However, several issues persist, such as low entry rates and learnability of gestures. In this section we go beyond our previous work in [19]. Based on our observations we identify several reasons for these issues and discuss the open challenges of mid-air text entry.

Large individual differences: We found significant individual differences among the users which can be attributed to: (1) anatomical differences in finger individuation and articulation, and (2) cognitive differences in motor control and learning. The design of hand gestures, tracking, and recognition algorithms need to account for these differences.

Learning: The learnability of gestures is a pragmatic obstacle for multi-finger input. If a gesture set for text entry is prohibitively time consuming to learn it will affect large-scale adoption. With PALM, we proposed a method to optimize for learnability. However, further research is needed to evaluate the involved models and their influence on learning time. In practice, users required clear instructions on

how to perform efficiently and often made errors in their movements. However, after many repetitions, some users reported quick and fast recall for some pre-learned gestures.

Technological difficulties: Markerless tracking of hands is a challenging, high-dimensional optimization problem. In order for the tracked motion to be usable for input it is essential to track all 26 DOFs of the hand. Due to anatomical dependencies, not all DOFs are independent. Therefore, many tracking methods simplify the DOFs using constraints during tracking. However, the dependencies are not uniform over the population. This leads to inaccuracies in tracking when using such constraints. We have found evidence of this with the Leap Motion, particularly with pinching gestures (see Figure 2). In our study we observed poor tracking and gesture recognition to limit user performance. When told to enter a word without looking at visual feedback, participants could perform faster, but tracking was often not reliable.

Feedback: Frequent and easy gestures were observed to be reliably performed eyes-free. Only during early stages of training participants frequently looked at their hands. This may help to avoid involuntary movements of non-instructed fingers. Auditory feedback at the input of each letter was helpful for users. Touch feedback could potentially be leveraged for pinching or fisting gestures, but tracking with the Leap Motion was poor in these cases (Figure 2).

Ergonomics: The effect of fatigue in multi-finger input is not fully understood yet. In contrast to prior work, users were not required to control the upper arm and shoulder but only wrist and finger movements. Thus, "Gorilla arm" problems could be avoided. However, users in our study reported discomfort in their lower arm and wrist.

6. CONCLUSION AND FUTURE WORK

Mid-air input by chord-like hand gestures, as deployed here, can work as an eyes-free, always-on input method. This seems an ideal candidate for mobile and wearable devices or large displays, as well as under special conditions that require touch-free input. The entry rates achieved so far are promising. However, we discussed several challenges that need to be addressed before mid-air input can be effectively used as an alternative to the standard keyboard. Therefore, future work should study:

1. The use of more DOFs or finer granularity of each DOF (*e.g.* different degrees of flexion–extension to map several letters to one finger).
2. Integration of different feedback modalities such as haptic, visual or auditory feedback (*e.g.* optional combination with surfaces or on-body interaction).
3. The influence of gesture design on learning time (motoric as well as cognitive learning).
4. Recognition algorithms and gesture sets designed for individual differences in anatomy and finger dexterity.
5. More sophisticated tracking solutions and integration of technological limitations into the design process.

7. REFERENCES

[1] X. Bi, B. A. Smith, and S. Zhai. Multilingual touchscreen keyboard design and optimization. *Human–Computer Interaction*, 27(4):352–382, 2012.

[2] H. Brashear, T. Starner, P. Lukowicz, and H. Junker. Using multiple sensors for mobile sign language recognition. In *Proc. ISWC*, Oct. 2003.

[3] D. G. Cattrysse and L. N. Van Wassenhove. A survey of algorithms for the generalized assignment problem. *Europ. Journ. of Operational Research*, 60(3):260–272, 1992.

[4] M. Dunlop and J. Levine. Multidimensional Pareto optimization of touchscreen keyboards for speed, familiarity and improved spell checking. In *Proc. CHI*, pages 2669–2678, 2012.

[5] D. Kim, O. Hilliges, S. Izadi, A. D. Butler, J. Chen, I. Oikonomidis, and P. Olivier. Digits: freehand 3d interactions anywhere using a wrist-worn gloveless sensor. In *Proc. CHI*, pages 167–176, 2012.

[6] P. O. Kristensson, T. Nicholson, and A. Quigley. Continuous recognition of one-handed and two-handed gestures using 3d full-body motion tracking sensors. In *Proc. IUI*, pages 89–92, 2012.

[7] P.-O. Kristensson and S. Zhai. SHARK2: A large vocabulary shorthand writing system for pen-based computers. In *Proc. UIST*, pages 43–52, 2004.

[8] I. S. MacKenzie. Fitts' law as a research and design tool in human-computer interaction. *Human-Computer Interaction*, 7(1):91–139, 1992.

[9] A. Markussen, M. R. Jakobsen, and K. Hornbaek. Selection-based mid-air text entry on large displays. In *Proc. INTERACT*, number 8117, pages 401–418. Springer Berlin Heidelberg, jan 2013.

[10] A. Markussen, M. R. Jakobsen, and K. Hornbaek. Vulture: A mid-air word-gesture keyboard. In *Proc. CHI*, pages 1073–1082, 2014.

[11] S. Melax, L. Keselman, and S. Orsten. Dynamics based 3D skeletal hand tracking. In *Proc. i3D*, pages 184–184, 2013.

[12] S. Mitra, P. G. Amazeen, and M. T. Turvey. Intermediate motor learning as decreasing active (dynamical) degrees of freedom. *Human Movement Science*, 17(1):17–65, 1998.

[13] T. Ni, D. Bowman, and C. North. AirStroke: Bringing unistroke text entry to freehand gesture interfaces. In *Proc. CHI*, pages 2473–2476, 2011.

[14] I. Oikonomidis, N. Kyriazis, and A. Argyros. Efficient model-based 3d tracking of hand articulations using kinect. In *Proc. BMVC*, pages 101.1–101.11, 2011.

[15] C. Qian, X. Sun, Y. Wei, X. Tang, and J. Sun. Realtime and robust hand tracking from depth. In *Proc. CVPR*, 2014.

[16] A. Schick, D. Morlock, C. Amma, T. Schultz, and R. Stiefelhagen. Vision-based handwriting recognition for unrestricted text input in mid-air. In *Proc. ICMI*, pages 217–220, 2012.

[17] M. H. Schieber. Individuated finger movements of rhesus monkeys: a means of quantifying the independence of the digits. *J Neurophysiology*, 65(6):1381–91, 1991.

[18] G. Shoemaker, L. Findlater, J. Q. Dawson, and K. S. Booth. Mid-air text input techniques for very large wall displays. In *Proc. GI*, pages 231–238. Canadian Information Processing Society, 2009.

[19] S. Sridhar, A. M. Feit, C. Theobalt, and A. Oulasvirta. Investigating the dexterity of multi-finger input for mid-air text entry. In *Proc. of CHI*. ACM, 2015.

[20] S. Sridhar, A. Oulasvirta, and C. Theobalt. Interactive markerless articulated hand motion tracking using RGB and depth data. In *Proc. ICCV*, pages 2456–2463, Dec. 2013.

[21] J. Wagner, E. Lecolinet, and T. Selker. Multi-finger chords for hand-held tablets: recognizable and memorable. In *Proc. CHI*, pages 2883–2892, 2014.

[22] R. Wang, S. Paris, and J. Popovi{\'c}. 6d hands: markerless hand-tracking for computer aided design. In *Proc. UIST*, pages 549–558, 2011.

[23] S. Zhai, M. Hunter, and B. A. Smith. The Metropolis Keyboard - an exploration of quantitative techniques for virtual keyboard design. In *Proc. UIST*, pages 119–128, 2000.