Loriette, A., Murray-Smith, R., Stein, S. and Williamson, J. (2017) Gesture Typing on Virtual Tabletop: Effect of Input Dimensions on Performance. In: ACM International Conference on Interactive Surfaces and Spaces (ISS '17), Brighton, UK, 17-20 Oct 2017, pp. 330-335. ISBN 9781450346917 (doi:10.1145/3132272.3135074)

This is the author's final accepted version.

There may be differences between this version and the published version. You are advised to consult the publisher's version if you wish to cite from it.

http://eprints.gla.ac.uk/146963/

Deposited on: 19 December 2017

Enlighten – Research publications by members of the University of Glasgow
http://eprints.gla.ac.uk

Gesture Typing on Virtual Tabletop: Effect of Input Dimensions on Performance

Antoine Loriette
Dept. Computing Science
University of Glasgow
Scotland, UK
antoine.loriette@glasgow.ac.uk

Sebastian Stein
Dept. Computing Science
University of Glasgow
Scotland, UK
sebastian.stein@glasgow.ac.uk

Roderick Murray-Smith
Dept. Computing Science
University of Glasgow
Scotland, UK
rod@dcs.gla.ac.uk

John Williamson
Dept. Computing Science
University of Glasgow
Scotland, UK
jhw@dcs.gla.ac.uk

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author.
ISS '17, October 17-20, 2017, Brighton, United Kingdom.
© 2017 Copyright is held by the owner/author(s).
ACM ISBN 978-1-4503-4691-7/17/10.
https://doi.org/10.1145/3132272.3135074

Abstract
The association of tabletop interaction with gesture typing presents interaction potential for situationally or physically impaired users. In this work, we use depth cameras to create touch surfaces on regular tabletops. We describe our prototype system and report on a supervised learning approach to fingertips touch classification. We follow with a gesture typing study that compares our system with a control tablet scenario and explore the influence of input size and aspect ratio of the virtual surface on the text input performance. We show that novice users perform with the same error rate at half the input rate with our system as compared to the control condition, that an input size between A5 and A4 present the best tradeoff between performance and user preference and that users' indirect tracking ability seems to be the overall performance limiting factor.

Author Keywords
Indirect Interaction; Gesture Input; Gesture Keyboard; Mobile; Continuous Interaction; Tabletop; Text Input

ACM Classification Keywords
H.5.2 [Information interfaces and presentation (e.g. HCI)]: User Interfaces.

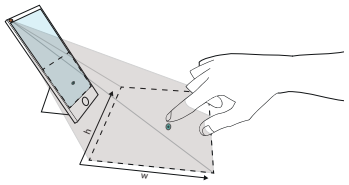

Figure 1. Potential interaction setup. A mobile device whose optical sensor creates an on-demand touch surface offers width and height as free parameters.

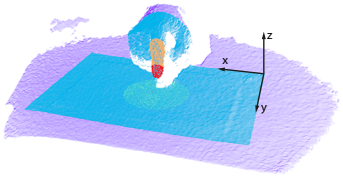

Figure 2. Segmented sensor image with points within the interactive surface (blue), the ROI plane (cyan), the detected fingers (yellow) and the user pointer (red).

Introduction

The combination of indirect optically tracked input and potentially projected display ("virtual surfaces") has several interesting properties. Virtual surfaces, as opposed to touch screens, are well suited to tackle the palm rejection problem and the occlusion problem, offer a choice of size, aspect ratio and texture of the input space, and allow interactions with dirty or wet hands.

Depth cameras [4, 12, 14] have usually been employed to create such surfaces, while other work have combined different sensor sources [13]. However, the touch classification, critical to the quality of the interaction remains challenging [14] and is traditionally addressed by hand-tuning parameters and thresholding.

In this context, the task of gesture typing [6] is an interesting research focus. Text-input is a major activity [9] on mobile device taking up to 40% of the user interaction time. The technique lends itself well to optical systems by limiting the requirement for repetitive target acquisition or touch classification. Yet, little is known about the impact of input size (one of the free parameters of virtual surfaces) on writing performance as other works focused on pointing [2, 3] or bimanual task [11].

This work specifically focuses on two goals: to report on a supervised learning approach to the fingertip touch classification task and to study the influence of input space (size and aspect ratio) on gesture typing performance using virtual tabletops.

System Overview

The envisioned interaction scenario is shown in Figure 1. A camera overlooks the interaction surface and a computing device performs the hand and fingers tracking as well as the touch detection. The system map in an indirect manner the user's pointer from the virtual surface onto the device's screen. A 3-state button model is used with visual feedback indicating hover positions and continuous touches. Audio cues indicate whenever a new touch down event is detected.

Touch Classification

This task is well suited to a supervised learning approach provided a training dataset is available. One of the authors video-taped 6 minutes of interactions creating 3 distincts datasets for 3 different fingers (index, thumb and pinky) totalling 10800 frames. The datasets are balanced with regards to touching frames and hovering frames and include a minor fraction of out-of-range frames.

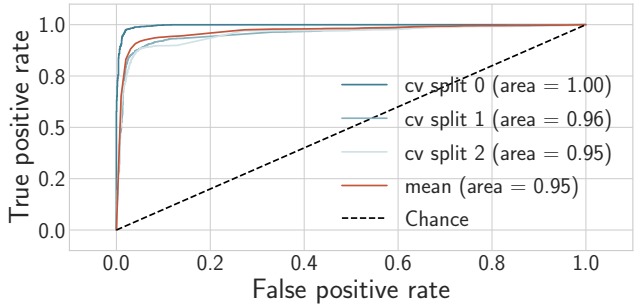

Figure 3. Performance of the fingertip touch classification with 3-fold cross validation. Each split trains on two finger types (among three) and validates against the remaining one.

The classification features are chosen as 20-bins histograms of z-values of the detected fingertip pointcloud, in red on Figure 2, which provides some orientation and shape invariance. This is fed to a binary neural network classifier: 20 inputs dimensions, 2 layers of

Figure 4. Picture of the user study setup. A tablet device provides the audiovisual feedback while a depth camera mounted on a tripod creates a virtual touch input surface.

64 hidden units with rectified linear (ReLU) activation and 50% dropout, and a fully connected output layer with sigmoid activation. We use the binary cross-entropy as loss function and rmsprop for the optimiser. Finally, we train the model over 50 epochs and perform a 3 folds cross-validation across the 3 datasets, different fingers are thus validated against each others. This provides an estimate for the generalisation power of the model when classifying unseen users as shown by the averaged 0.96 AUC for the ROC, see Figure 3.

The model used for the user study was trained over the full dataset over 75 epochs. As an improvment, it is worth noting the possibility to delegate the feature extraction step to a model capable of inferring features directly, such as a convolutional neural network.

User Study

We ran an experiment with two research goals in mind. First we wanted to evaluate the usability of our system, for this purpose we included a condition with an interaction directly on the tablet. Second, given the nature of the task and its inherent difficulty for novice users, we investigated the influence of the physical space on performance. Accot et al. [1] demonstrated a U-shaped curve in performance against input scale. Our independent variable were thus the control surface dimensions, which effectively change the control-display gain (CD_{gain}) defined as in [2] by the the ratio of the pointer velocity to device velocity.

We recruited 12 participants, all right handed, without requirement on gesture typing experience. A repeated measures within-subjects design was used. There were 3 conditions (DEVICE, SIZE and ORIENTATION) and 5 level combinations were evaluated, see Table 1 below for details. We adopted a 3-symbols naming convention: the first letter represents the DEVICE, the second marks the ORIENTATION and the number represents the scaling factor as SIZE.

LEVEL	*DEVICE*	*width*	*height*	*area*	*ratio*	CD_{gain}
OP1	OPTICAL	9.4	4.7	44.2	2	1
OP2	OPTICAL	18.8	9.4	176.7	2	$1/2$
OL2	OPTICAL	25.6	6.9	176.6	3.7	$1/1.7$
OP4	OPTICAL	37.7	18.9	712.5	2	$1/4$
TP1	TABLET	9.4	4.7	44.2	2	1

Table 1. Design of the experiment. Level name, device type, dimensions (in cm), area (in cm^2), ratio and control/display gain for all 5 combinations used in the experiment.

The apparatus, see Figure 4, implements the system description above. An Intel Realsense camera is used, and the processing is performed on a desktop computer. An Android tablet running custom software produces the audiovisual feedback. The pointer tracked by the system is defined as the point cloud closest to the camera in the depth direction (y on Figure 2); there is no explicit finger or hand modeling performed. Instead, the assumption is that the interacting finger is the furthest protruding object from the user.

The task was to write 20 words per level (words taken at random from the most common english words with length between 2 and 5 letters) with maximum 7 attempt to complete 5 correct input for each word. The design of the task is similar to [10] and allows novice users to focus on the physical execution instead of the shape recollection, effectively emulating a proficient bahavior even for novice users. After each level, participants were offered to take a

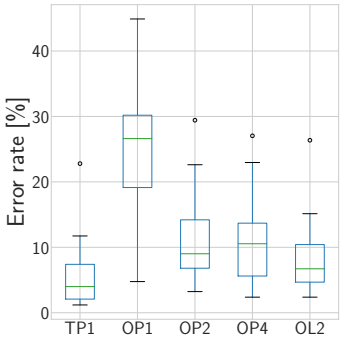

Figure 5. Effect of levels on error rate.

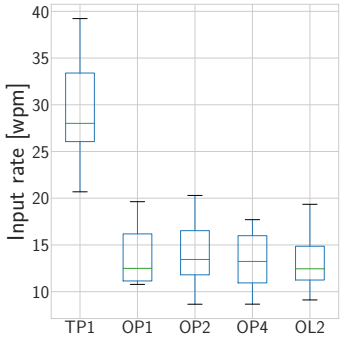

Figure 6. Effect of levels on input rate.

break before moving to the next one. Finally, participants were asked for their feedback using the NASA Task Load Index [5] (NASA-TLX) to assess the perceived workload of completed level.

The experimental design was thus: 12 participants $\times$ 5 LEVEL $\times$ 20 WORD $=$ 1200 trials. For each trial, we had 5 to 7 attemps depending on the error rate, which means a total of 6,000 to 8,400 total samples. After the experiment, we actually recorded 6963 samples.

Results

The dependent variables we analyse are the success rate, the time taken per trial and the trace data when available. From these, we computed the dependent measure error rate defined as the percentage of unsucessful attempts as well as the text entry rate measured in Words Per Minute (WPM), as in [8], according to the following formula: $WPM = |T|/s \times 60/5$ where $|T|$ is the length of the transcribed string and s is time in seconds.

We found one input word (LAY) presented some very unusual behavior. Its error rate was at 89.5% while the WORD mean was 20.1% and no other word had an error rate higher than 30%. The reason was that the recogniser promoted words of higher prior probability in the language model, *"Larry"*, *"last"* or *"Katy"*. For the all subsequent analysis of error rate, LAY was removed from the dataset.

Effect on error rate

A statistical analysis showed a significant main effect of DEVICE ($F_{1,11} = 37.77$, $p < 0.0001$) on error rate with mean value for TP1 and OP1 equal to 6.2% and 26.1%, respectively, see Figure 5. A statistical analysis also showed a significant main effect of SIZE ($F_{2,22} = 10.99$, $p < 0.001$) on the error rate. Finally, an ANOVA on TP1, OP2, OP4 and OL2 did not show a significant main effect ($p = 0.11$) even though OPTICAL is on average higher than TABLET. In other words, among all levels of the experiment, only OP1 shows a significant higher error rate.

Effect on input rate

A statistical analysis showed a significant main effect of DEVICE ($F_{1,11} = 90.15$, $p < 0.0001$) on input rate, see Figure 6, with mean values for TP1 and OP1 equal to 29.2 WPM and 13.7 WPM, respectively. We also looked at pairwise comparison for OPTICAL and could not find a statistical difference in the mean input rate. The input rate achieved by the participants in TP1 is in-line with what can be expected from novice users after the time of the experiments [6]. The averaged 54% lower input rate in OPTICAL should be compared with other similar results, as [8] with 57% after 10 sessions, that compare direct and indirect input modality for mid-air gesture typing.

Effect of orientation

The design of the experiment also includes ORIENTATION. This condition has so far been excluded from the SIZE analysis and can be difficult to apprehend since not only the ORIENTATION of the visual feedback changes, but its size also. The main result is the average pointer speed in display space at $303.75 pixel/s$, higher than the PORTRAIT orientation at an average $228 pixel/s$. It is important to keep in mind that the keyboard area in LANDSCAPE ($61 cm^2$), bigger than PORTRAIT ($44 cm^2$), which can explain the faster pointer speed but is not responsible for a higher input rate due to longer traces to produces.

Qualitative data

At the end of the experiment, participants were invited to provide some informal feedback, rank the different level in order of preference and fill in the NASA-TLX form, the tablet interaction was ranked best by all participants

except one. OP4 was consistently ranked last, while OP2 and OL2 had equal ranking in second position. The NASA-TLX data shows that participants describe OP4 as the most physically demanding interaction. Participants graded OP1 and OP4 20% higher than OP2 and OL2 on the scale of effort and frustration. Finally, OP1 was graded as having lowest level of performance and highest mental and temporal demand.

Discussion

We conducted a user study investigating the usability of the system and showed that for control dimension at least twice the display dimension novice users were capable the same error rate at half the input rate. We also showed that at the same CD_{gain}, the error rates were dramatically higher. Also, the experiment showed a constant input rate across display sizes and aspect ratios. Accot et al. [1] showed that task with high index of difficulty would exhibits a u-shaped curved with size. Since we did not observe such a behavior, this puts in question whether gesture typing for novice users is "hard enough".

The constant input rate however shows that participants are capable of adapting their motor speed across all investigated scales. Participants also varied the display pointer speed, especially in OL2 when presented with a bigger display surface. Because participants are capable of adapting their motor behavior and their control bahavior, another explanation for the upper bound in text input is that the indirect nature of the interaction is the limiting factor.

Participants feedback showed that OP1 was a demanding interaction, and that increasing the surface size to the extent of OP4 was also physically demanding. We would therefore recommend sizes OP2 or OL2 for interaction, noting that the performance is also dependent on the CD_{gain}). Some participants reported the texture having an impact of the interaction. Levesque et al. [7] have used friction in a dynamic manner to improve target acquisition. Given the relatively low performance of the participants, it could be interesting to explore if tactile feedback could help controlling an indirect pointer.

Conclusion

We developed a system that affords gesture typing on arbitrary flat surfaces using depth camera tracking, and demonstrated a machine learning approach to the issue of detecting touch for fingertips which is effective for this mode of control. We designed an experiment with two research goals in mind. First, we compared the typing performance of our optical indirect system against a control condition which was a direct interaction on a tablet. Second, we studied the influence of size and aspect ratio of the input surface on gesture typing performance. We showed that the participants could enter text at half the input rate and the same error rate on surface at least twice the size of the visual feedback. We also showed that the input rate is largely independent of surface size across the range of sizes we were able to examine. Gesture typing has promise for interaction outside of mobile devices, e.g. for motor impaired users who struggle with capacitive touch technologies. This paper indicates that while camera-tracked gesture typing performance is usable, input rates are lower than touchscreen performance, and this is not influenced by scaling of the input.

Acknowledgements

This research is funded by the E.U. Horizon 2020 research and innovation programme under project Moregrasp, award number 643955.

References

[1] Accot, J., and Zhai, S. Scale effects in steering law tasks. *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '01* (2001), 1–8.

[2] Casiez, G., Vogel, D., Balakrishnan, R., and Cockburn, A. The Impact of Control-Display Gain on User Performance in Pointing Tasks. *Human-Computer Interaction 23*, 3 (jul 2008), 215–250.

[3] Gilliot, J., Casiez, G., and Roussel, N. Impact of form factors and input conditions on absolute indirect-touch pointing tasks. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14*, ACM Press (New York, New York, USA, 2014), 723–732.

[4] Harrison, C., Benko, H., and Wilson, A. D. OmniTouch. In *Proceedings of the 24th annual ACM symposium on User interface software and technology - UIST '11*, ACM Press (New York, New York, USA, oct 2011), 441.

[5] Hart, S. G., and Staveland, L. E. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. 1988, 139–183.

[6] Kristensson, P.-O., and Zhai, S. SHARK^2. In *Proceedings of the 17th annual ACM symposium on User interface software and technology - UIST '04*, ACM Press (New York, New York, USA, oct 2004), 43.

[7] Levesque, V., Oram, L., MacLean, K., Cockburn, A., Marchuk, N. D., Johnson, D., Colgate, J. E., and Peshkin, M. A. Enhancing physicality in touch interaction with programmable friction. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*, ACM Press (New York, New York, USA, may 2011), 2481.

[8] Markussen, A., Jakobsen, M. R., and Hornbæk, K. Vulture: a mid-air word-gesture keyboard. *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14* (2014), 1073–1082.

[9] McGregor, M., Brown, B., and McMillan, D. 100 days of iPhone use. In *Proceedings of the extended abstracts of the 32nd annual ACM conference on Human factors in computing systems - CHI EA '14*, ACM Press (New York, New York, USA, sep 2014), 2335–2340.

[10] Quinn, P., and Zhai, S. Modeling Gesture-Typing Movements. *Human Computer Interaction* (aug 2016), 1–47.

[11] Schmidt, D., Block, F., and Gellersen, H. A Comparison of Direct and Indirect Multi-touch Input for Large Surfaces. In *Proceedings of the 12th IFIP TC 13 International Conference on Human-Computer Interaction: Part I*. Springer-Verlag, 2009, 582–594.

[12] Sridhar, S., Markussen, A., Oulasvirta, A., Theobalt, C., and Boring, S. WatchSense. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*, ACM Press (New York, New York, USA, 2017), 3891–3902.

[13] Wen, E., Seah, W., Ng, B., Liu, X., and Cao, J. UbiTouch. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '16*, ACM Press (New York, New York, USA, 2016), 286–297.

[14] Xiao, R., Hudson, S., and Harrison, C. DIRECT. In *Proceedings of the 2016 ACM on Interactive Surfaces and Spaces - ISS '16*, ACM Press (New York, New York, USA, 2016), 85–94.