How to Click in Mid-Air

Florian van de Camp[1*], Alexander Schick[1*], and Rainer Stiefelhagen[2]

[1] Fraunhofer Institute of Optronics, System Technologies and Image Exploitation, Karlsruhe, Germany
{florian.vandecamp, alexander.schick}@iosb.fraunhofer.de
[2] Karlsruhe Institute of Technology (KIT)
rainer.stiefelhagen@kit.edu
[*]These authors contributed equally to this work

Abstract. In this paper, we investigate interactions with distant interfaces. In particular, we focus on how to issue mouse click like commands in mid-air and we propose a taxonomy for distant one-arm clicking gestures. The gestures are divided into three main groups based on the part of the arm that is responsible for the gesture: the fingers, the hand, or the arm. We evaluated nine specific gestures in a Wizard of Oz study and asked participants to rate each gesture using a TLX questionnaire as well as to give an overall ranking. Based on the evaluation, we identified groups of gestures of varying acceptability that can serve as a reference for interface designers to select the most suitable gesture.

1 Introduction and Related Work

In this paper, we investigate interaction with distant interfaces. In particular, we focus on how to interact with distant objects that are, for example, displayed on a videowall in a control room or on a shopping window. We are interested in distant interaction because growing screen sizes require, and improved computer vision technologies allow for, interaction without actually touching the displays. We further selected the clicking gesture because it is one of the most basic forms of interaction which allows for very powerful and universal interaction designs as the computer mouse shows.

There exists a large body of literature about gestures for human-computer interaction, e.g. [6, 9]. Here, we focus on clicking gestures that can be performed with one arm only. Alternatives are, e.g., two-arm clicking gestures or the use of additional devices. However, we feel that device-free one-arm clicking gestures place the least restrictions on users and allow a more general use. We also believe that they are more natural and intuitive and can more easily be related to touch.

We also focus on clicking gestures only. We do not investigate how to translate a pointing gesture to display coordinates (see [12, 13] for recent examples). We also do not consider gestures that can be used as shortcuts to perform more complicated actions, e.g. [1], or on gestures for other actions, e.g. pan-and-zoom [8]. We focus on clicking because it is one of the most fundamental human-computer interactions, but at the same time one of the most useful and general ones.

The clicking gestures in this paper are all based on a pointing gesture to specify the area of interaction on the display. Pointing gestures have a long history in human-computer interaction, e.g. Bolt's "Put that there!" [2], Vogel and Balakrishnan [13], Schick et al. [12], to name just a few. To use pointing gestures for interaction, two main problems have to be solved: first, how is the pointing gesture translated to display coordinates, e.g. ray-casting [12, 13] or relative pointing [13], and second, how is the interaction triggered, e.g. with speech [2], hand gestures [1, 13], or dwell-based [12]. Here, we solely focus on how to trigger the interaction. Even though different clicking gestures have been proposed and compared to each other, there are two drawbacks in previous work. First, the number of clicking gestures is relatively small and the selection arbitrary, thus not exploring the full design space of clicking gestures [13]. Second, they are usually implemented with a given system, e.g. based on Vicon markers [13] or video cameras [1, 12]. Such systems are never absolutely perfect, e.g. due to measurement noise or physical limitations like resolution. In addition, some gestures are more difficult to recognize than others. Unfortunately, imperfections in the recognition system affect the users' perception of the gesture and, consequently, their evaluation. In this paper, we present a systematic evaluation of clicking gestures and evaluate them in a Wizard of Oz study [4], thereby eliminating the bias of an imperfect system.

There exists a large body of research on how to classify gestures in human-computer interaction and several taxonomies have been proposed [3, 7, 10, 11, 14]. A recent overview about gesture taxonomies can also be found in [14]. The proposed taxonomies usually have a relatively wide scope to capture a large range of gestures that can occur in different contexts. Even though it is possible to describe clicking gestures based on these taxonomies, almost all clicking gestures would fall into the same category. Therefore, we will introduce a new taxonomy that focuses only on one-arm clicking gestures. This allows for a more diverse categorization and a better exploration of the design space of clicking gestures.

In the remainder of this paper, we will first introduce a new taxonomy before choosing nine specific gestures that were evaluated in a Wizard of Oz study. After presenting the results of this study, we conclude with a discussion.

2 Taxonomy

In this paper, we focus on clicking gestures to interact with objects that are displayed on a distant vertical display (Figure 1). We assume that the basis of each clicking gesture is a pointing gesture to specify the object to interact with. This is very natural for humans. Pointing gestures can be split into three phases: preparation, stroke, and recovery [5]. The clicking gesture always occurs in the stroke phase.

Given established taxonomies, one-arm clicking gestures can, for example, be categorized as deictic or manipulative [10, 11]. However, such a categorization is very coarse and does not capture the possible variations, as we will show now.

Fig. 1. The interface for the user study.

A one-arm clicking gesture requires some form of movement over time. Given a pointing gesture, this can either be a movement of the arm, of the hand, or the fingers. This leads to the following categorization (see also Figure 2).

If the clicking gesture is based on arm movement, it must not affect the pointing location. This reduces possible movements to orthogonal ones (towards and away from the display) and rotation. Not moving the arm is the third option. If based on hand movement, the clicking gesture can be expressed by a bending movement, e.g. by vertically bending the hand. The most variations are possible when the clicking gesture is based on finger movements. We characterize these by the number of fingers that are part of the gesture, starting from one and up to five. We found that having three or four fingers perform a gesture are the least natural options.

In summary, the proposed taxonomy classifies one-arm clicking gestures for distant interaction based on two characteristics: first, the body part that is mainly involved, and second, the type or direction of the movement. These characteristics also influence the difficulty of implementing a system to recognize the gesture (the larger the body part, the easier it is to recognize), and how stressful it is for the body to execute the movements depending on size and how natural the movement is (e.g. push versus pull arm movements).

A clicking gesture is categorized by one single leaf node; however, even though it would increase physical and cognitive stress, it is also possible to combine leaf nodes, e.g. arm and finger movements.

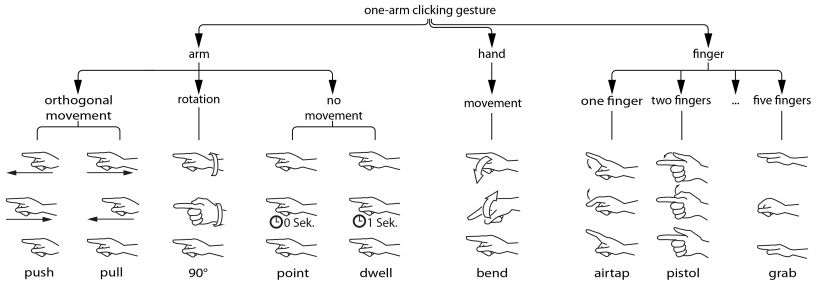

Fig. 2. Taxonomy for distant one-arm clicking and examples for each leaf node of the taxonomy. The nine representatives for clicking gestures at the bottom are the ones evaluated in the user study. Each gesture is split into three consecutive phases that are shown from top to bottom. The clicking event is triggered in the last phase.

2.1 Clicking Gestures

For our experiments, we chose for most leaf nodes of the taxonomy at least one representative. They are shown at the bottom in Figure 2 and will now be explained.

For arm movements, we chose five representatives: push and pull, 90° rotation, point, and dwell. We set the dwell time for the dwell gesture to one second. We chose this duration after experiments in our laboratory and looking at existing interfaces, e.g. Microsoft Kinect applications. The point gesture is different from dwell-based interaction, in that the clicking event is triggered as soon as the arm movement stops.

For hand movements, we chose a downward vertical bending of the hand. We found other hand movements similar, but much more stressful on the wrist.

For finger movements, we chose one, two, and five finger movements: airtap [13], pistol [13], and grab. We categorized pistol as a two-finger movement because it requires two fingers to perform the gesture. Not moving a finger is similar to the point and dwell gestures. We found three and four finger movements either similar to other gestures or as not as natural.

When comparing the gestures to mouse clicks, it is interesting to note that all gestures have phases that can be compared to hold and release (except point and dwell). Even though not required for our study, this is important for applications that require something similar to a mouse-down event. Also note that all gestures can be implemented in a real-world system; in fact, most of them are already available at our lab.

3 Experiments

Every gesture recognition is biased by the accuracy of the underlying recognition system. To overcome this problem, we used a Wizard of Oz

setup [4] where participants are interacting with a pretense system that is controlled by a hidden human experimenter. This allows users to experience each gesture as if a perfect recognition system would be present.
We evaluated all techniques in our laboratory (Figure 1). The interface was displayed on the right half of a 4m by 1.5m videowall with the highest point being at 2.37m and a display resolution of 4096 by 1536 pixels. The effective interaction space was 2m by 1.5m with a resolution of 2048 by 1536 pixels. In our experiment participated 5 females and 13 males ranging from the age of 20 to 64. Two participants were left-handed. The participants were students from university and employees from our research institution.
We presented each participant with all nine clicking gestures in randomized order. The task was to click a button that was displayed on the wall (Figure 1). The size of the button was 13.6cm. Due to the Wizard of Oz setup, the size did not affect the recognition accuracy of the clicking gesture. In each run, the participants were allowed to try each gesture. Then, a succession of 25 buttons appeared that they had to click. When a button was clicked, it disappeared and the next button was displayed. The buttons were equally distributed across the screen and their order of appearance randomized.
There were two experimenters. One experimenter was present in the room and guided the participants through the study. The second, hidden experimenter was seated in a separate room and observed the participants with multiple cameras that are part of our regular system setup. By pushing a button, the hidden experimenter could trigger a clicking event. By carefully observing the scene, it was possible to only trigger clicking events when the participants were actually pointing at the button (which some did not to test the system limits). The perceived system reaction time was minimal and only affected by the latency of the cameras and the reaction time of the hidden experimenter.
After each clicking gesture, the participants were asked to rate the gesture based on a NASA TLX questionnaire. The NASA TLX questionnaire contains questions about mental, physical, and temporal demand, overall performance, frustration level, and effort. We asked participants to give their ratings for each of these categories on a 7-point Likert scale. Then, the next gesture was presented in the same fashion. After the experiment, the participants were presented with an additional questionnaire where they were asked for further comments and to select which gestures they considered generally useful. Most importantly, we asked them to rank the gestures based on how they liked them, starting from 1 for their most and ending at 9 for their least favored method.

4 Results

For presenting the results, we mainly focus on the overall ranking of the gestures because it summarizes the overall perception of the participants and we show the physical and temporal demands results from the TLX questionnaire. Table 1 shows the mean values as well as standard deviation for each gesture and the pair-wise significance comparisons. Tables 2

	airtap	point	pistol	90°	bend	grab	dwell	push	pull
μ	2.78	3.00	4.61	4.89	5.11	5.28	5.39	6.22	7.72
σ	1.93	2.13	2.16	2.26	2.38	2.28	2.31	2.15	1.19
airtap	-	0.95	**< 0.01**	**< 0.01**	**≪ 0.01**	**≪ 0.01**	**≪ 0.01**	**≪ 0.01**	**≪ 0.01**
point	0.95	-	**0.02**	**0.02**	**0.01**	**< 0.01**	**≪ 0.01**	**≪ 0.01**	**≪ 0.01**
pistol	**< 0.01**	**0.02**	-	0.62	0.54	0.41	0.30	**0.03**	**≪ 0.01**
90°	**< 0.01**	**0.02**	0.62	-	0.78	0.75	0.54	0.10	**≪ 0.01**
bend	**≪ 0.01**	**0.01**	0.54	0.78	-	0.94	0.72	0.18	**≪ 0.01**
grab	**≪ 0.01**	**< 0.01**	0.41	0.75	0.94	-	0.87	0.28	**< 0.01**
dwell	**≪ 0.01**	**≪ 0.01**	0.30	0.54	0.72	0.87	-	0.28	**≪ 0.01**
push	**≪ 0.01**	**≪ 0.01**	**0.03**	0.10	0.18	0.28	0.28	-	**0.04**
pull	**≪ 0.01**	**≪ 0.01**	**≪ 0.01**	**≪ 0.01**	**≪ 0.01**	**< 0.01**	**≪ 0.01**	**0.04**	-

Table 1. The gesture ranking evaluation. Each participant ranked the gestures from 1 (liked best) to 9 (liked worst). The table shows the resulting ranking from left (best) to right (worst) with mean and standard deviation in the top rows and the significance analysis results (p-values) for the pair-wise comparisons below (based on Wilcoxon rank sum test).

and 3 show results for physical and temporal demands. Because the data did not always follow a normal distribution, we used the Wilcoxon rank sum test. However, a t-test showed comparable results. The significance level was 0.05. While the other categories of the TLX questionnaire are in line with the ranking, they are not as significant. This is most likely due to the fact that the TLX questions are tailored towards more complex tasks. While physical and temporal demands can be directly related and are meaningful for the simple task, the other categories like mental demand do not fit very well and might lead to inconclusive ratings.

5 Discussion

The ranking of the gestures follows a smooth descent from airtap, as the best rated gesture, to pull, as the worst. The pair-wise significance analysis (Table 1) shows that there are three groups of gestures that have similar ratings within the group but significantly differ from other groups. The results from the physical and temporal demands ratings are in line with the ranking of the gestures and support it.

The first group consists of the two highest rated gestures: airtap and point. Airtap requires minimal effort and, as several users pointed out, has a high resemblance to the use of a computer mouse thus making it very intuitive. The high rating of the point gesture is not surprising as it requires no additional effort besides pointing itself and, therefore, was perceived as very convenient and fast. It has to be pointed out, however, that a large number of users noted that they would expect a lot of errors when operating a real interface as they did not have the same level of

	point	airtap	90°	pistol	bend	push	dwell	grab	pull
μ	-2.83	-2.28	-1.39	-1.11	-1.11	-0.89	-0.89	-0.67	0.00
σ	0.37	0.93	1.70	1.63	1.49	1.45	1.94	1.73	1.49
point	-	**0.04**	**$\ll$ 0.01**	**$\ll$ 0.01**	**$\ll$ 0.01**	**$\ll$ 0.01**	**$\ll$ 0.01**	**$\ll$ 0.01**	**$\ll$ 0.01**
airtap	**0.04**	-	0.15	**0.03**	**0.02**	**$<$ 0.01**	**0.04**	**$<$ 0.01**	**$\ll$ 0.01**
90°	**$\ll$ 0.01**	0.15	-	0.55	0.47	0.31	0.48	0.22	**0.02**
pistol	**$\ll$ 0.01**	**0.03**	0.55	-	0.95	0.59	0.85	0.47	**0.05**
bend	**$\ll$ 0.01**	**0.02**	0.47	0.95	-	0.68	0.85	0.47	**0.05**
push	**$\ll$ 0.01**	**$<$ 0.01**	0.31	0.59	0.68	-	0.82	0.68	0.09
dwell	**$\ll$ 0.01**	**0.04**	0.48	0.85	0.85	0.82	-	0.68	0.13
grab	**$\ll$ 0.01**	**$<$ 0.01**	0.22	0.47	0.47	0.68	0.68	-	0.28
pull	**$\ll$ 0.01**	**$\ll$ 0.01**	**0.02**	**0.05**	**0.05**	0.09	0.13	0.28	-

Table 2. Physical exhaustion. How physically exhaustive was interaction with the given technique with ratings from -3 (not exhaustive at all) to +3 (very exhaustive).

	point	airtap	bend	pistol	90°	grab	push	pull	dwell
μ	-2.94	-2.06	-2.00	-1.83	-1.78	-1.28	-1.22	-0.28	0.06
σ	0.23	1.22	1.00	1.30	1.23	1.59	1.69	1.56	2.12
point	-	**$<$ 0.01**	**$\ll$ 0.01**	**$<$ 0.01**	**$\ll$ 0.01**	**$\ll$ 0.01**	**$\ll$ 0.01**	**$\ll$ 0.01**	**$\ll$ 0.01**
airtap	**$<$ 0.01**	-	0.67	0.61	0.47	0.14	0.20	**$\ll$ 0.01**	**$\ll$ 0.01**
bend	**$\ll$ 0.01**	0.67	-	0.89	0.69	0.23	0.25	**$\ll$ 0.01**	**$\ll$ 0.01**
pistol	**$<$ 0.01**	0.61	0.89	-	0.82	0.33	0.33	**$<$ 0.01**	**$<$ 0.01**
90°	**$\ll$ 0.01**	0.47	0.69	0.82	-	0.41	0.40	**$<$ 0.01**	**$<$ 0.01**
grab	**$\ll$ 0.01**	0.14	0.23	0.33	0.41	-	0.97	0.07	0.06
push	**$\ll$ 0.01**	0.20	0.25	0.33	0.40	0.97	-	0.10	0.06
pull	**$\ll$ 0.01**	**$\ll$ 0.01**	**$\ll$ 0.01**	**$<$ 0.01**	**$<$ 0.01**	0.07	0.10	-	0.68
dwell	**$\ll$ 0.01**	**$\ll$ 0.01**	**$\ll$ 0.01**	**$<$ 0.01**	**$<$ 0.01**	0.06	0.06	0.68	-

Table 3. Temporal demand. How long did it take to execute the gesture from -3 (very short) to +3 (very long).

control compared to the other gestures. Given a real interface, the pointing gesture would most likely only be useful if there is no potential for false positives. Therefore, we recommend airtap as the primary gesture for click-like commands.
The pistol, bend, 90°, grab, dwell, and push gestures make up the second group. They were rated worse than the first group but were still named when participants were asked which gestures they considered to be useful in everyday life. These gestures could be used for secondary tasks like opening a context menu.
The remaining gesture, pull, forms the third group. Almost no participant could imagine using the pull gesture for an actual application and in all ratings, pull was among the worst rated gestures. This seems to result from the fact that pulling the arm away from the display to click something is very counterintuitive. We advise not to use this gesture at all for clicking gestures.
As a whole, the ranking follows the general observation that a gesture with less required effort resulted in a better overall rating. While the gestures in the first group add little to no additional strain to the always required pointing gesture, the pull gesture of the third group requires movement of the complete arm which, depending on the execution of the gesture, can even include the upper body. In between are the gestures of the second group that mostly require movement of multiple fingers or the whole hand. In case of the dwell gesture, no movement is necessary but the long delay during which the arm has to remain extended is tiring.
While the point gesture and the dwell gesture are very similar, the dwell gesture performed significantly worse in the ranking. Of course, the dwell gesture has other advantages such as being easy to detect and is, therefore, robust but the delay was not well perceived by the participants. We see the dwell gesture as a good choice if robust detection of other gestures cannot be guaranteed. However, given similar robustness for any of the better rated gestures, it could be a valid design choice to consider them over the dwell gesture.
As pointed out, the ranking of the gestures shows a smooth descent which indicates that several gestures can be considered useful. This leads to the conclusion that gestures of the first group could be used for common operations, like clicking, because they were generally perceived as being faster and more efficient. Gestures of the second group would then be a good choice for less frequent but still common operations such as drag-and-drop or opening a context menu.

6 Conclusion

We presented a taxonomy specifically aimed at one-arm clicking gestures for distant interaction and evaluated nine specific gestures in a Wizard of Oz study. The rankings indicate which gestures were considered more useful over the others. Design choices based on the presented results can help to provide an improved user experience for distant interaction. As a conclusion, we recommended airtap as the primary clicking gesture and gave recommendations for secondary gestures that can be used for shortcut functions, e.g. opening a context menu.

References

1. Bader, T., Räpple, R., Beyerer, J.: *Fast Invariant Contour-Based Classification of Hand Symbols for HCI*, Proc. Computer Analysis of Images and Patterns, 2009, pp. 689–696
2. Bolt, R.A.: *"Put-that-there": Voice and Gesture at the Graphics Interface.* SIGGRAPH Computer Graphics, 1980, 14(3), pp. 262–270
3. Grossman, T., Wigdor, Daniel: *Going Deeper: a Taxonomy of 3D on the Tabletop* Proc. Horizontal Interactive Human-Computer Systems, 2007, pp 137–144
4. Kelley, J.F.: *An Empirical Methodology for Writing User-friendly Natural Language Computer Applications* ACM SIGCHI Conference on Human Factors in Computing Systems, 1983, pp. 193–196
5. Kendon, A.: *Gesticulation and speech: two aspects of the process of utterance.* The Relationship of Verbal and Nonverbal Communication. Key, M.R., ed., 1980, pp. 207–227
6. Kendon, A.: *Gesture: Visible Action as Utterance* Cambridge University Press, 2004
7. McNeill, D.: *Hand and Mind: What Gestures Reveal about Thought.* University of Chicago Press, 1992
8. Nancel, M., Wagner, J., Pietriga, E., Chapuis, O., Mackay, W.: *Mid-air pan-and-zoom on wall-sized displays* Proc. SIGCHI Conference on Human Factors in Computing Systems, 2011, pp 177–186
9. Pavlovic, V.I., Sharma, R., Huang, T.S.: *Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review.* Pattern Analysis and Machine Intelligence, 19(7), 1997, 677–695
10. Quek, F.K.H.: *Toward a Vision-Based Hand Gesture Interface.* Proc. Virtual Reality Software and Technology, 1994, pp. 17–29
11. Quek, F.K.H.: *Eyes in the Interface.* Image and Vison Computing 13(6), 1995, pp. 511–525
12. Schick, A., van de Camp, F., Ijsselmuiden, J., Stiefelhagen, R.: *Extending Touch: Towards Interaction with Large-Scale Surfaces* Proc. Interactive Tabletops and Surfaces, 2009, pp. 127–134
13. Vogel, D., Balakrishnan, R.: *Distant Freehand Pointing and Clicking on Very Large, High Resolution Displays* ACM symposium on User Interface Software and Technology, 2005, pp. 33–42
14. Woobrock, J.O., Morris, M.R., Wilson, A.D.: *User-Defined Gestures for Surface Computing* Proc. Human Factors in Computing Systems, 2009, pp 1083–1092