Comparing Elicited Gestures to Designer-Created Gestures for Selection above a Multitouch Surface

Dmitry Pyryeskin, Mark Hancock, Jesse Hoey
University of Waterloo, Ontario, Canada
{dpyryesk, mark.hancock, jhoey}@uwaterloo.ca

ABSTRACT
Many new technologies are emerging that make it possible to extend interaction into the three-dimensional space directly above or in front of a multitouch surface. Such techniques allow people to control these devices by performing hand gestures in the air. In this paper, we present a method of extending interactions into the space above a multitouch surface using only a standard diffused surface illumination (DSI) device, without any additional sensors. Then we focus on interaction techniques for activating graphical widgets located in this above-surface space. We have conducted a study to elicit gestures for above-table widget activation. A follow-up study was conducted to evaluate and compare these gestures based on their performance. Our results showed that there was no clear agreement on what gestures should be used to select objects in mid-air, and that performance was better when using gestures that were chosen less frequently, but predicted to be better by the designers, as opposed to those most frequently suggested by participants.

Author Keywords
Multimodal interaction; natural human computer interaction; surface computing; multi-touch; gestures; hoverspace.

ACM Classification Keywords
H.5.2 [Information interfaces and presentation]: User Interfaces. - Graphical user interfaces.

General Terms
Human Factors; Design; Performance; Measurement; Experimentation.

INTRODUCTION
Multi-touch technology was conceived at least as early as 1965 [14], and since then has slowly become more reliable, accurate and commonplace. Nowadays, devices equipped with multi-touch screens are becoming ubiquitous on phones and tablets, and are being researched heavily on larger surfaces, such as tables and walls. This shift provides the potential for direct interaction with on-screen objects in a fashion familiar from the physical world [1,10,26].

Recent technology, such as the Microsoft Kinect, has made cheaper the possibility of extending this physical interaction into *hover space*—the space above or in front of a multitouch display. The addition of hover space input to touch input can provide another mode of interaction, while allowing smooth transitions from one mode to another [17]. This added dimension in the interaction space can be used for a variety of purposes, for instance to manipulate 3D artifacts [13], to provide shortcuts to applications via Hover Widgets [8], or to create occlusion-aware interfaces [24].

While this design space is promising, one of the most compelling aspects of direct touch interaction is the clear and understandable way in which on-screen targets can be selected—by touching them with your hands or fingers. This physicality, however, is lost in hover space, and it becomes no longer clear how digital artifacts can and should be selected. Will people expect to be able to grab objects in mid-air, point at objects from a distance, or will they understand the need to dwell over a 3D target to select it (for example)? Currently, little work has explored what gestures people expect to be able to use to select targets above a table. In this work we study target selection in this space, with respect to both people's expectations and performance.

In this paper, we explore interaction in hover space by focusing specifically on item selection in the space above a multi-touch surface. We first present the design of a system that can approximate the height of hands above a diffused surface illumination (DSI) device. We then present the results of a pair of studies: in the first, we elicit what gestures people expect to be able to use to select on-screen targets in hover space, and in the second we explore the performance of the gestures chosen from the first study compared to several of our own designs for selection. Some of the gestures identified in our first study were beyond the capability of our hardware system, though might be possible with additional hardware (e.g., a separate motion tracking system). Thus, the focus of our second study was on evaluating the performance of gestures that were practical to implement with minimal hardware. Our results show that not only do people disagree about how to select objects in this space, but also that the less-frequently chosen designs that we (the designers) predicted to perform better, in most cases did, when compared to the most frequently chosen gestures from the first study.

RELATED WORK
In this section we focus on four related areas: detecting movement above a surface, interaction in front of a surface, how others study gestures, and in-air target selection.

ITS'12, November 11–14, 2012, Cambridge, Massachusetts, USA.

Detection of Movement above a Surface

The subject of hover space interaction is already well researched and a number of hover detection techniques have been proposed. Some of these systems rely on multiple stereo cameras to calculate the 3D location of a person's hands and fingers [16,27]. These systems require much more computational power than other vision-based systems and also require very precise camera calibration. Other systems use depth cameras [3,25], such as that found in the Microsoft Kinect. Z-Touch uses multiple layers of lasers to estimate the position of a person's hands and fingers [22], and Shadow Tracking positions infrared emitters above the table to create and track shadows of people's hands together with touch regions [5].

Most of these methods require custom-built hardware or modifications of existing multi-touch tables. We add to this work by demonstrating how to use an existing DSI setup to track hands above a surface. Our approach uses computer vision algorithms to estimate hand distance and can be considered a special case of the more general shape from shading (SFS) problem. SFS techniques compute the shape of a surface using a single greyscale image as input [19]. The computed surface may be used to estimate the height of a hand above the surface and for gesture recognition.

Interaction in Front of a Surface

Many techniques already exist for interacting in front of a large display wall or interactive surface [3,13,25,27]. Some research investigates interactive widgets in the space above a table, such as Hover Widgets [8], which include the possibility of using the space above a surface to create a new command layer that is clearly distinct from the input layer on the surface. Hover space can also be used to create techniques for more natural manipulation of 3D artifacts on a multitouch surface [13], or for faster zooming interaction [9]. Our work expands this space and explores more specifically the target selection aspect of above-surface interaction techniques, once a widget has been acquired.

Studying Gestures

Gestures can be a rich, versatile, and natural way for people to interact with each other as well as with technology. With the wide variety of techniques that allow in-air hand tracking, gesture-based interaction has been studied in depth [7,17,21]. A common approach to evaluating these gestures is to first design them, based on experience or related work, and then evaluate their performance when compared to other alternatives. An alternative approach is to first elicit what gestures people expect to correspond to a given action [18,28], rather than have the system designer determine what gesture is best-suited to an action. Other researchers have used the same methodology to develop gesture sets for other domains [6,12]. Our work uses an approach similar to Wobbrock et al. [28] to elicit gestures from participants.

In-air Target Selection

Device-free interaction makes signalling selection non-trivial. One of the most basic solutions which has been adopted by many eye-tracking systems [11] is to use dwell time threshold to indicate a click event. Some of the other proposed gestures are: pinching in the air [13], *SideTrigger* [2], *AirTap* [23] and *ThumbTrigger* [23]. As Wobbrock et al. [28] put forward, while gestures defined by the system designers certainly produce good results, they may not reflect people's expectations, and their development may be influenced by concerns for reliable recognition [18].

Our work builds upon existing research in each of these areas by examining the specific domain of above-surface target selection. We present a technique to detect the height of a person's hand above the table, which we use to study above-surface selection. We draw from existing methodology by first eliciting preferred gestures, and then comparing their performance to designer-chosen gestures.

SELECTION ABOVE A MULTI-TOUCH SURFACE

The introduction of inexpensive hardware to detect the 3D shape of hands and fingers in mid-air has led to a surge in research that explores interaction in the space above a table. While this direction seems promising, the interaction space has some significant qualitative differences from on-surface interaction. One of the most salient of these differences is the inability to select on-screen targets by touching or clicking on them. A variety of gestures have been suggested in the research to perform such selections, but it remains an open question how people will expect selections to occur.

In this section, we first describe the design challenges associated with a person's expectations specifically about target selection in the space above a table, and then describe our system for detecting the height of a person's hand above the table, which can be used to make these selections.

Design Challenges for Above-Surface Selection

A number of factors make the design of above-surface selection particularly challenging. Many of these challenges have been raised in prior work on above-surface interaction, but we describe the most pertinent challenges specific to selection: non-physicality, the use of layers, the transition between above-surface and touch interaction, and fatigue.

Non-Physicality

One of the most important challenges when selecting objects in mid-air is the lack of a reference point. On a 2D surface, people can select artifacts such as buttons, menus, and images by touching the surface directly. In contrast, when a target is in hover space, it cannot be represented in the air (unless some sort of holographic or virtual reality technology is used); therefore the direct interaction paradigm breaks down. It is not clear how this lack of physical connection will manifest in a person's expectation about how to select on-screen targets.

Layers/Precision

Defining a number of layers in hover space can be useful for increasing the dimensionality of control space [22], and can provide an indication of the precision with which the system can detect hand height. Objects can be placed in

separate layers above one another and selected by moving a hand up and down. However, it is not clear whether people will understand or expect these layers to exist, and whether such constraints will improve selection performance.

Transition between Hover and Touch

As discussed by Spindler et al [21], hover and touch data should be considered as a unified space. That is, there should be no separation between interactions in hover space and on the surface, and no need to switch the modality of interaction. Another challenge in designing a selection technique is to ensure that this seamless transition is supported. The most common selection technique on existing surfaces typically manifests at this point of transition (e.g., on touch down), and so the designer must still determine a suitable way of handling the response just before this event. It is again unclear what people's expectations are about how this transition should happen, and what gestures should cause on-screen changes.

Fatigue

Fatigue is a well-known issue with mid-air interactions. However, one common technique for mid-air selections is the use of a dwell time (e.g., Microsoft's Kinect interface on the Xbox). When multiple selections are required, this added wait time has the potential to exacerbate fatigue effects when a person is required to hold their hand steady. Ergonomics of the surface and the size of hoverspace are also important factors that affect users' fatigue. For example, Spindler et al. tried to reduce the effects of fatigue in their studies by limiting the height of hoverspace to shoulder height for standing users [21].

In our pair of studies, we attempted to determine what gestures people expect to be able to use to select targets above a table, and whether non-physicality, layers, transitions, and fatigue play a role in those expectations. To better elicit these expectations, we first developed a system to detect the height of a person's hand above a multitouch surface.

System Description

We implemented two techniques for estimating the height of hands and objects above a DSI multitouch table, both based on a vision-based tracker. While we found the second approach to be more simple and useful in our study of target selection, we include the description of the first as well, as it provides more information than just hand height, and could be useful in other applications.

Vision-Based Tracker

Our system is based on a common pipeline used in other vision-based multitouch trackers, for example, Community Core Vision (ccv.nuigroup.com) and reacTIVision (reactivision.sourceforge.net). We add one or more additional pipelines (in addition to the touch pipeline) to detect hands above the surface that each use a lower threshold (i.e., detects dimmer blobs) and a mean-shift filter [4] to reduce noise. Standard blob finding and tracking algorithms are applied to the result of all pipelines and the data is sent to client applications using the TUIO [15] protocol.

Hover Height Estimation

Currently, there are many methods that can be used to estimate the height of a palm above a surface: stereo cameras [16,27], depth cameras (like Kinect) [3,25], multiple layers of lasers [22], infrared emitters above a table [5] and so on. What distinguishes our tracker from other methods is the ability to estimate the height of a hovering hand above the surface of the table, without additional sensing technology. Some computer vision approaches, like shape-from-shading (SFS) techniques [19], may compute the shape of a palm using an image from a regular camera. In the process of software development, we explored two methods of palm height estimation based on the more general SFS problem.

Slices Method. This method uses nine additional hover pipelines instead of just one, gradually decreasing in cut-off value, and each representing a different slice of height above the table (the algorithm was adapted from [20]). This method presents interesting opportunities for gesture recognition, because it produces an approximate 3D surface of a hovering hand. A potential weakness of this method is the low resolution and high computational demand of the blob-finding algorithm run on each slice of the image.

Center-weighted Average Method. Another method that proved simpler and required much less processing power is a "center-weighted average" approach. This method computes an average value of the 16×16 pixel square located in the geometric center of each blob in the hover layer. The resulting value is then mapped to the height of a hovering hand using a formula derived from linear regression.

Our proposed height estimation methods inherit limitations and difficulties of other SFS techniques, such as the fact that the source image is heavily influenced by the properties of the surface (such as reflectivity) and lighting (direction and magnitude). To address these issues we have chosen to solve the problem empirically, not analytically. We have recorded 5-second videos of 16 participants holding their dominant hand at various heights above the surface (from 2cm to 20cm with a step of 2cm). An exponential function was then fitted to the relationship between the measured value and the actual height and was used in the tracker to estimate the height of a hand.

We used our system to perform two experiments to evaluate people's expectations of how to select targets in hover space, as well as the performance of techniques derived from these expectations and our own designs.

EXPERIMENT 1: EXPECTED SELECTION GESTURES

The purpose of the first study was to elicit expectations of how objects should be selected above a multitouch surface. We used the methodology of Wobbrock et al. [28] to elicit these gestures. Participants were asked to acquire targets in the space above a table, and then asked to perform the gesture that they expected would make a selection.

Apparatus

We used an unmodified DSI multi-touch table setup for the experiment. This setup uses EndLighten acrylic that scatters infrared (IR) light evenly throughout the table's surface. An 81cm × 61cm sheet of acrylic forms the surface of the table. It is edge-illuminated by a strip of 850nm IR diodes. The display is rear-projected with resolution of 1024×768 pixels. An Unibrain Fire-i™ camera equipped with an 850nm band-pass IR filter is mounted behind the screen. The table is powered by a Windows XP computer.

Participants

Sixteen paid participants (6 female) took part in the study. All were right-handed and the average age was 24.8 years ($SD = 5.08$). 13 participants had previously used smartphones equipped with a multitouch screen, and 7 had previously used tablet PCs and public multitouch devices such as bank machines or airport check-in kiosks.

Design

We used a within-participants factorial design with the following three factors:

- Widget (bar, circle, button, menu)
- Number of hands (one, two)
- Anchoring (screen, cursor)

Task

Participants were shown one of the four widgets (described below) and asked to interact by moving their hand in the 3D space above the multitouch surface. When the system detected a hand hovering above the surface, a cursor was displayed. The cursor is represented as a semi-transparent circle with its center directly below the geometric center of a hovering palm and its size proportionately related to the height of the palm above the surface. Participants were then asked to move this cursor to the target (which varied by widget) and then demonstrate what gesture they would use to select that target. The software did not attempt to recognize or act on the gestures performed by the participants; it only recorded activity above the surface. We also video recorded participants' hands using a camera positioned above the screen.

Participants were asked to perform gestures using four different visual widgets. These four widgets were designed to be both abstract representations of targets (bar and circle) as well as closer approximations of widgets that could be used in an application (button and menu). Each visual widget is controlled in the same way: by moving one's hand to a predetermined target in the 3D space above the table.

The *bar* widget (Figure 1a) is an abstract rectangular control with a small slider moving up and down along the midline. The distance of a person's hand above the table determines the position of this slider along the vertical axis. The target is represented as a differently-coloured portion of the rectangle, which turns red when the slider is placed inside.

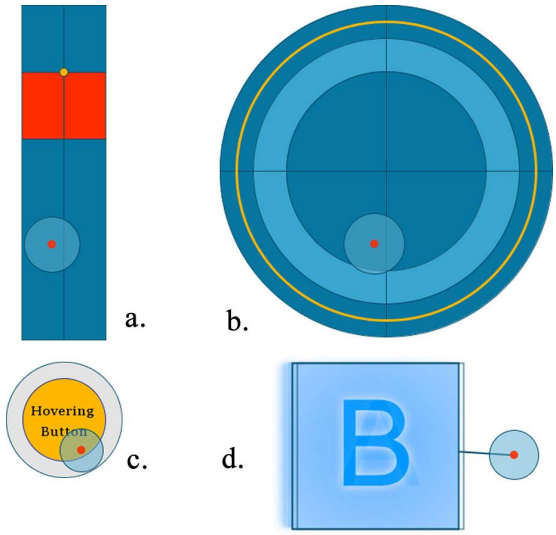

Figure 1: Above-surface widgets:
a. bar, b. circle, c. button and d. menu

The *circle* widget (Figure 1b) is similar to the bar widget, but shaped differently. The slider is represented as a ring and the target is represented as a differently coloured band, which turns red when the cursor is placed inside.

The *button* widget (Figure 1c) is a version of a standard circular GUI button, but adapted to be acquired by moving one's hand to a predetermined 3D position above the table. The circle changes colour when this correct location is acquired to indicate this hover state. Its height above the surface is represented by a grey band around the inner circle. The width of this band is the same as the radius of the cursor, when the cursor's height matches the button's height.

The *menu* widget (Figure 1d) is again adapted to hover space and is similar to the menu described by Benko et al. [25]. The current menu item is made visible by moving a hand up or down. Items that are located above or below the current item are blurred to simulate the way human eye focuses on objects. Items in our study are abstracted to the letters A through E.

Number of Hands and Anchoring

Participants were asked to demonstrate a gesture using both one hand and two hands separately. The widgets were also shown to be anchored either to the centre of the screen (i.e., remained stationary in x and y), or to the cursor (i.e., the x and y position of the widget moved with the hand).

Procedure

The order of events for each participant can be described algorithmically as follows:

1. The idea of hover space selection was introduced
2. For each widget (× 4):
   a. For each combination of hands and anchoring (× 4):
      i. Practice using the widget for as long as desired
      ii. When ready, demonstrate a gesture to select the corresponding target

Gesture	Description	Freq. 1 hand	Freq. 2 hands
Off-hand tap	Tapping the screen with a single or several fingers of the off hand	N/A	37.5%
Grab	Grabbing or pinching gesture with one or both hands	35.2%	23.4%
Push with a finger	Downwards motion of a single or several fingers	26.6%	10.2%
Snapping/Clapping	Clapping hands together or snapping fingers (sound-based interaction)	9.4%	7.0%
Spread/Expand	Both hands moving horizontally from the target to the edges of the surface	N/A	6.3%
Push	Downwards motion of a full hand, by bending wrist, elbow or shoulder joint. A version of the gesture was performed by bending all fingers downwards.	9.4%	3.9%
Tap	Tapping the screen with a single or several fingers	7.8%	0.0%
Dwell	A hand is held steady in the same place for a set period of time	6.3%	0.0%
Shake hand	A hand is held in the same place and shaken	3.9%	2.3%
Swipe	Horizontal motion above the surface	0.8%	6.3%
Other	Other gestures, such as rotating a palm or bumping palms together	0.8%	3.1%

Table 1. The frequencies with which each gesture was demonstrated in the first experiment.

The order in which the widgets appeared was counterbalanced using a random Latin square. The number of hands and target anchoring parameters were combined into a single 4-value parameter and counterbalanced using random Latin squares (one for every widget). Since the software had no means to recognize a gesture, participants were asked to indicate verbally when their gestures were complete. The experimenter then pressed a button to indicate to the software when to stop recording. With 16 participants, 4 visualizations, 1 or 2 hands and 2 anchoring methods, a total of $16 \times 4 \times 2 \times 2 = 256$ gestures were performed.

Gesture Classification

Once the gestures were collected we analyzed the videos recorded by the above-table camera. We manually classified the gestures performed along 3 dimensions: *palm shape*, *magnitude* and *motion*. Palm shape was specific to one hand; therefore the shape of the second palm was also analyzed in two-handed gestures. Examples of the *palm shape* category are: open palm, closed fist, closed fist with an extended index finger. The gesture *magnitude* describes how much of the palm was involved in the gesture; it ranged from full palm gestures to single finger gestures. The gesture *motion* category describes the path that the hand or a finger followed during the gesture. Depending on the magnitude of the gesture, the motion can describe the path of the full palm or a single finger.

Similar to the findings of [28], we noticed that participants did not attach significance to which fingers were used in a gesture and how many fingers were involved. Some participants performed gestures using their index finger interchangeably with their middle finger or thumb.

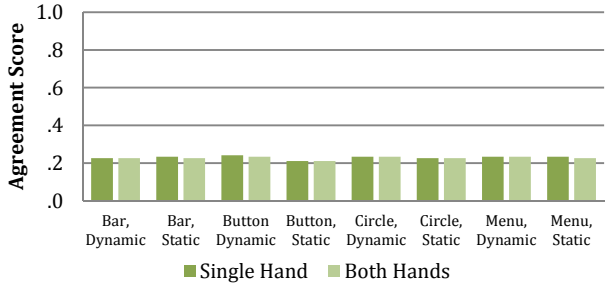

Figure 2. Agreement for each condition

Agreement Scores

We have grouped gestures with similar *palm shape, magnitude,* and *motion* into 11 groups (see Table 1). Group size was then used to compute an agreement score A that reflects, in a single number, the degree of consensus among participants (this process was adopted from [28]).

$$A = \sum_{P_i \subseteq P} \left(\frac{|P_i|}{|P|}\right)^2$$

P is the set of all proposed gestures for a certain condition (defined by widget, number of hands and anchoring) and P_i is a set of identical gestures from P.

Results

The average agreement score for each condition was .23 ($SD = .008$). This small variability indicates that the study factors had very little effect on agreement (see Figure 2). The overall agreement for one-handed and two-handed gestures was .22. In contrast, the agreement scores of most gestures selected for the user-defined set in [28], were between .30 and 1.0 for a single hand and between .30 and .60 for both hands. Our result indicates that there is no clear agreement between participants about how selection should be performed above a surface.

While the agreement scores were low, there were several gestures that were performed more often than others. With one hand, the *grab* gesture was performed 45 times (35.2%) and *push with a finger* was performed 34 times (26.6%). With two hands, *off-hand tap* (37.5%) and *grab* (23.4%) were performed more often than the others. Overall participants preferred one-handed gestures, as indicated either verbally or in the post-study survey. This finding agrees with the results of a gesture-elicitation study in [28].

We noticed that most participants (12 of 16) consistently used the same one-handed gesture and the same two-handed gesture for each experimental condition. Based on this observation, and in order to reduce the complexity of the factorial design, we opted to use only one visualisation in the second experiment, instead of testing all four.

EXPERIMENT 2: GESTURE PERFORMANCE

From the first experiment, we were able to identify *off-hand tap, grab* and *push with a finger* as possible candidates for a selection technique for targets above a multi-touch surface.

The second experiment was designed to measure the performance of these candidate methods of above-surface selection, as well as some techniques we suspected might be effective. We were interested in comparing three properties of each gesture: how fast it can be performed, how accurately it can be performed, and how difficult it would be for the computer to recognize and disambiguate the gesture. To do so, we designed an experiment where the human subject had to first *acquire* a target in hover space (above the screen), and then *perform* one of the gestures to select the target (like a button click). Unfortunately, the system we used was not accurate enough to recognize some of the gestures due to the low resolution and ambiguities inherent to shape-from-shading techniques. Too many factors can change the way a palm looks in greyscale apart from its height; for example the reflectivity of the skin on the top and bottom of the human palm is different, so the system would not be able to differentiate between a hand held palm-down higher above the table and a hand held palm-up closer to the surface. The *grab* gesture looks to our software exactly like lifting a palm higher above the surface. *Push with a finger* could not be recognized because the change in the image of the hand was too small to be differentiated from noise. We were similarly unable to include such designer-defined gestures as *SideTrigger* [2] or *ThumbTrigger* [23]. However, we hope these limitations will be addressed in future work using this research as a foundation.

As a result of these limitations, we decided to focus our second study on evaluating the performance of gestures that were practical to implement with minimal hardware. We chose *push* (as a close approximation to *push with a finger*) and selected the most common two-handed gesture suggested by the participants: *off-hand tap*. In contrast to elicited gestures, *dwell* and *tap* were also included in the experiment as those we expected to perform well based on our design experience (i.e., designer-defined), even though they were infrequently chosen by participants in our first study. We also used the same apparatus as in the first.

Participants

Sixteen paid participants (5 female) took part in the study. One participant was left-handed and the average age was 24.1 years ($SD = 3.17$). Some participants also took part in the first study. All participants were local university students and most majored in computer science or engineering.

Design

We used a within-participants factorial design with the following two factors:

- Gesture (dwell × 3, push, tap, off-hand tap)
- Location (dominant, middle, non-dominant)

Task

To begin each trial, the participant was asked to touch a specific area of the screen labelled "parking area" with their finger. When the participant touched this parking area, a target was displayed. The target is identical to the circle widget used in the first study, except for an adjustment in color scheme (see Figure 3). The participant was asked to acquire this target, using the same cursor as in the first experiment, by moving their hand so that the centre of their palm was directly above the centre of the ring (i.e., at the crosshair), and the height of their hand made the cursor radius match the target radius. Since the target was virtually located above the surface, the participants had to not only match its x, y position on the surface, but also its z position, or height above the surface. Once acquired, the participant then performed one of six selection gestures: *short dwell*, *medium dwell*, *long dwell*, *push*, *tap*, or *off-hand tap*.

The three dwell gestures required participants to hold their hand above the target for 500ms, 1000ms, and 2000ms, respectively. The push gesture required participants to move their hand rapidly in the downward direction, and was detected when the speed of the hand was above ~10m/s. To perform the tap gesture, a participant had to move their hand down rapidly (with the same speed as the push gesture) and then touch the screen. The last z position of the cursor prior to the start of this rapid motion was saved, and when a touch event was detected, it was used to determine if the target was activated successfully. Off-hand tap gestures were completed when the participant touched the table anywhere with a finger on the non-dominant hand.

Targets appeared in one of the three x, y locations: on the dominant side (right for right-handed participants, and left for left-handed), in the middle, or on the non-dominant side of the screen. The height and distance of the target from the "parking area" was kept constant, so that participants' hands had to be at 14cm above the table and 25.4cm from the start position along the table's surface. The participant was asked to acquire the target and perform a gesture as quickly as possible.

To indicate the state of the target, the following color scheme was used: initially the target was red, when the centre of the cursor was within 1cm of the crosshair, the target changed to yellow, and when a gesture was completed, the

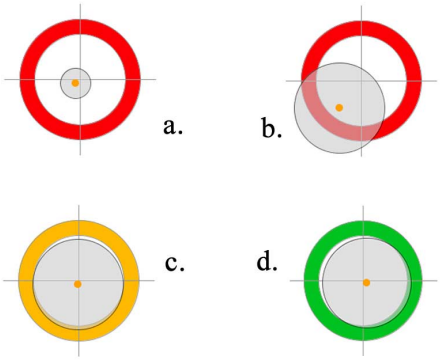

Figure 3. The targets (circular rings) and cursor used in experiment 2. Targets first appears red (a. & b.), turns yellow when acquired (c.) and green upon selection (d.). The radius of the cursor is determined by the height of the participant's hand above the table.

target changed to green (see Figure 3). For both the *push* and *tap* gestures, movement beyond the target boundaries was required to perform the gesture, and so once the target was first acquired (yellow), the target would not return to its non-acquired state (red).

Procedure

The order of events for each participant can be described algorithmically as follows:

1. The idea of hover space selection was introduced
2. For each gesture (× 6):
   a. The experimenter demonstrated the gesture
   b. The participant performed a practice trial
   c. For each trial (3 locations × 3 repetitions = 9), the participant was asked to:
      i. Move their hand to the "parking area"
      ii. Acquire and selected the target

As the dwell gestures all required only one explanation, the three dwell gestures were presented together (one after the other) in random order. The order of dwell, push, tap, and off-hand tap was then counterbalanced using a random Latin square. The order of the 3 locations was randomized.

The application recorded the path of the hand and the timing of events, as well as target loss and successful/failed gestures. The trial was considered successful when a gesture was recognized while the target was acquired. If a gesture was performed outside of the target or a gesture was never performed, the trial was marked as failed.

After each block of 9 trials, a participant answered three 7-point Likert scale questions about the ease of navigation to the target, performing the gesture and the overall experience. With 16 participants, 6 gestures and 9 repetitions, a total of $16 \times 6 \times 9 = 864$ gestures were performed.

Results & Discussion

Due to the shape and ergonomics of the experimental multi-touch table, most participants had difficulty acquiring targets on the side of the table opposite their dominant hand. The table was shaped as a coffee table and its small height (51 cm) forced participants to bend over the table or kneel next to it; which meant that to reach the left side of the screen, right-handed people had to rotate their torso and/or shoulders, making the target acquisition awkward and uncomfortable. We performed a one-way analysis of variance (ANOVA) on the location factor, which showed a significant effect ($F(2,30) = 32.19, p < .001$). Post-hoc pairwise comparison of the location factor showed a significant difference between the non-dominant location and both other locations ($p < .001$), while the middle and dominant locations were not significantly different ($p = .113$). We thus removed data from the non-dominant level of the location factor from the remainder of our analysis. Therefore, a total of $16 \times 6 \times 6 = 576$ gestures were considered in the analysis. We analyzed three main dependent measures: gesture speed, gesture accuracy, and participant preference.

Gesture Speed

The time to perform a gesture can be broken down into three parts: acquisition time, jitter time, and selection time.

Acquisition time was measured as the time it takes to move a hand from the starting area to the target. Using a 6 (gesture) × 2 (location) repeated-measures ANOVA, we found a significant main effect of gesture on acquisition time ($F(5,75) = 5.528, p < .001$). In particular, the *short dwell* gesture had significantly smaller acquisition times than all other gestures (*medium dwell*: $p < .01$, *long dwell:* $p = .024$, *push:* $p < 0.001$, *tap:* $p < 0.01$; *off-hand tap:* $p < 0.001$). Acquisition in the medium dwell was also significantly faster than *push* ($p = .035$), *tap* ($p = .028$), and *off-hand tap* ($p < .01$). The *long dwell* was also significantly faster than *off-hand tap* ($p = .039$). Acquisition times for *push*, *tap,* and *off-hand tap* were not significantly different ($p > .05$). There was also no main effect of location ($F(1,15) = 2.630, p > .05$), nor interaction between gesture and location ($F(5,75) = 1.544, p > .05$).

The differences in acquisition times were surprising, and cannot be easily explained, since the acquisition task did not differ for any of the gestures; all trials required acquiring a target at the same distance from the starting location. This indicates that people adjusted their behaviour depending on the gesture they were performing. Specifically, people moved more quickly toward targets that required only a dwell. We hypothesize that this was due to the inaccuracy of, in particular, the short dwell for selection (described below). This inaccuracy perhaps led to a speed-accuracy trade-off. That is, participants may have noticed an inability to accurately select targets, and so increased their speed. However, we note that this same trade-off did not occur for *off-hand tap*. Further studies are required to better isolate this effect.

Jitter time was measured as the time between the initial target acquisition and the final one. In other words, jitter represents a phase when the participant lost and reacquired the target (perhaps several times) before performing the gesture.

We performed a 6 (gesture) × 2 (location) repeated measures ANOVA on the jitter times. There was a significant main effect of gesture ($F(5,75) = 10.275, p < .001$). Post-hoc analysis revealed that jitter times for *long dwell* were not significantly different than *off-hand tap* ($p = .212$), and jitter times of both were significantly longer than jitter times of all other gestures ($p < .012$). *Short dwell* had significantly less jitter than both other dwells ($p < .002$), but similar to *push* and *tap* gestures ($p > .05$). Jitter times for *medium dwell, push* and *tap* were not significantly different. There was also no main effect of location ($F(1,15) = 0.065, p > .05$), nor interaction between gesture and location ($F(5,75) = 0.260, p > .05$).

The increase in jitter is expected as the length of dwell increases, since it is difficult for a person to hold their hand

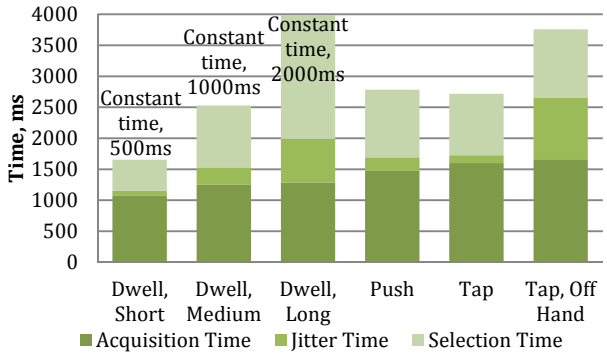

Figure 4. Acquisition, jitter, and selection times.

steadily in the same place. *Off-hand tap* had more jitter than all other gestures except for *long dwell*, perhaps due to the ergonomics of our table or to the fact that participants had trouble holding their main hand steady while touching the screen with the other hand. The same effect appeared in the number of times a target was lost (see below). The other four gestures had comparable amount of jitter.

Selection time was measured as the time it takes to perform the gesture after the last acquisition (i.e., after acquisition + jitter). For the *dwell* gestures, this selection time is constant, and so was not included in the analysis. We performed a 3 (gesture) × 2 (location) repeated measures ANOVA on the remaining three gestures, but found no significant main effects or interactions (gesture: $F(2,30) = 0.372, p > .05$; location: $F(1,15) = 0.252, p > .05$; gesture × location: $F(2,30) = 0.44, p > .05$).

Overall, although we broke down our analysis by acquisition, jitter, and selection time, Figure 4 shows how these times would accumulate in practice. *Short dwell* was the fastest to perform, while *long dwell* was the slowest. *Push*, *tap*, and *medium dwell* were comparable in speed, while *off-hand tap* performed almost as badly as *long dwell* (likely due to the high jitter times).

Gesture Accuracy

To evaluate the precision of each gesture we measured the number of times a target was lost when performing a gesture and the overall number of failed trials.

Target lost count was a measure of the difficulty in keeping one's hand on the target while selecting a gesture. This can be thought of as a count of the number of jitters per trial, rather than the time taken for jitter. The 6 (gesture) × 2 (location) ANOVA was performed on target lost count. There was a main effect of gesture ($F(5,75) = 12.947, p < 0.001$). Post-hoc analysis revealed that the target lost count of *long dwell* and *off-hand tap* gestures were significantly different from all others ($p < .034$). The target lost counts of *short* and *medium dwells* were also different from all others ($p < .030$) except for the *push* and *tap* gestures. There was also no main effect of location ($F(1,15) = 0.181, p > .05$), nor interaction between gesture and location ($F(5,75) = 0.368, p > .05$).

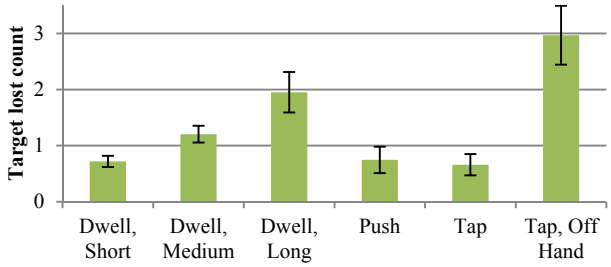

Figure 5. The target lost counts for each gesture.

As expected, it becomes harder to stay on-target as the length of the *dwell* gesture increases. *Off-hand tap* again performed the worst for this measure. *Push* and *tap* performed as well as *short* and *medium dwells*.

Trial failure frequency was measured as the proportion of unsuccessful trials, which were recorded if (a) the target was never acquired, (b) no selection gesture was recorded, or (c) the target was not in its acquired state when the selection gesture was performed. We performed a Cochran's Q test to analyze this binary data (each trial was either successful or not) for the gesture factor, and included each location as a repetition. We found a significant difference between the failure frequencies of the gestures ($Q_{df=5,N=96} = 50.282, p < .001$). A Post-hoc McNemar's test revealed that the *short dwell* resulted in significantly more failures than the rest of the gestures ($\chi^2_{N=96} > 6.568, p < .01$); *long dwell* resulted in fewer failures than *push* ($\chi^2_{N=96} = 8.828, p < .01$); *tap* also had fewer unsuccessful trials than *push* ($\chi^2_{N=96} = 6.323, p = .012$).

The number of unsuccessful trials for the *short dwell* gesture was very high ($M = .47, SD = .502$). This high error rate was not unexpected as it has been noted before and dubbed the "Midas Touch" effect [11], which negates any speed benefit noted before by resulting in many unintentional selections. Moreover, the *tap* gesture resulted in fewer errors than the *push* gesture, and with a similar overall speed for these two gestures, this indicates that *tap*'s overall performance was better. While the *long dwell* had similarly few errors, the added time for dwell means that *tap* also outperforms *long dwell*.

Participant Preference

We also analyzed participants' preferences based on three

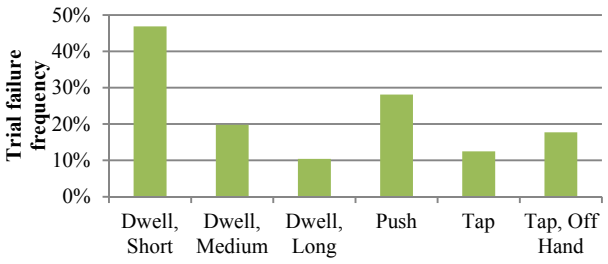

Figure 6. The frequency of unsuccessful trials.

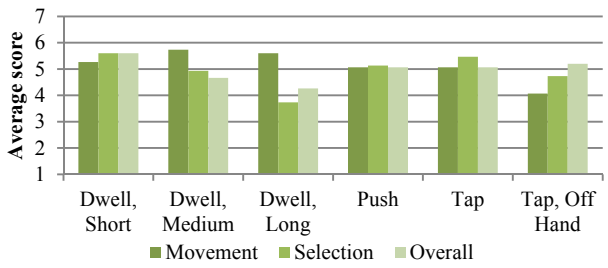

Figure 7. Participant preferences.

7-point Likert scale statements. The first was: "It was easy to ***move*** to the target", the second was: "It was difficult it to ***select*** the target" and the last was: "The selection technique is easy ***overall***". The words "easy" and "difficult" were used alternatingly to avoid influencing the responses; for clarity, we present the results of the responses to the scales with the word "difficult" backwards (i.e., answers 1 and 7, 2 and 6, 3 and 5 were interchanged). One of the participants did not rate one of the gestures.

A Friedman's test was performed on each of the three scales. Significant effects were found in the ***move*** and ***overall*** categories ($\chi^2_{N=15} = 17.509, p < .01$ and $\chi^2_{N=15} = 12.164, p = .033$ respectively). The ***select*** category's differences were only marginally significant ($\chi^2_{N=15} = 10.719, p = .057$). Post-hoc pairwise Wilcoxon tests were performed which revealed the following significant results:

- For ***move***, *off-hand tap* was rated significantly lower than all other gestures and *push* was rated lower than *medium dwell*.
- For ***select***, *long dwell* was rated lower than all other gestures except for the *off-hand tap*.
- For ***overall***, *short dwell* was rated higher than *medium* and *long dwells*.

The fact that most participants found moving a hand to a target in the *off-hand tap* gesture more difficult is consistent with our findings in performance. Selection using the *long dwell* gesture was rated low as expected, since two seconds is a long time to hold a hand steady in the same place. *Short dwell* was preferred to other dwells overall, but due to the unacceptably high false activation rate, we would not recommend that it be used in a real-world application. *Push* and *tap* were rated consistently high on all 3 scales.

Unlike the participants of the first experiment, the second experiment's volunteers had to perform a gesture multiple times while worrying about the speed and precision of their gestures. Therefore their preferences may be a better indication of which gestures are more suited for an application.

DISCUSSION

The results of this pair of studies provide some insight into the design of above-surface selection techniques. When selecting with one hand, people most frequently expect to be able to *grab* on-screen objects from a location in mid-air above that object, and with two hands expect to be able to *tap* with their other hand. However, this expectation was not agreed upon by all participants (35.2% and 37.5%, respectively; only between 1/3 and 2/5 of participants). Although, due to system limitations we could not easily investigate the preferred one-handed grab gesture, our investigation of a close approximation of their second choice in our *push* gesture and the *off-hand tap* two-handed gesture revealed that they underperformed when compared to the *tap* gesture, as expected by designers. More specifically, while we found no difference in selection time between *push* or *off-hand tap* and the one-handed *tap*, participants frequently drifted off of the target when using *off-hand tap*, and frequently missed the target with *push*.

In addition, the common *dwell* technique did not result in a suitable alternative. In particular, a trade-off between dwell time and accuracy was revealed; when the dwell time is low enough to improve speed beyond the best-performing *tap*, the number of errors increased dramatically.

Based on these results, we recommend the use of a single-handed tap gesture for selection of targets in hover space. However, we suggest some caution to designers in this interpretation, as our system was not capable of detecting *grab*, the most preferred selection gesture from the first experiment. Nonetheless, we note that accurate detection of a grab gesture is not simple in any of the existing hardware systems that we are aware of, whereas tap can easily be detected by tracking sudden acceleration and using the existing touch capabilities of an interactive surface. We demonstrate the use of tap to select colours from a colour palette in our example application.

EXAMPLE APPLICATION

We designed Hover Paint (Figure 8) using the results of our two experiments, which allows people to paint on a multi-touch surface using their fingers. Hover space interaction is used to control the colour and size of the brush. To activate the colour wheel, a person can lift their hand above the table. The *x* and *y* position of the hand can then be used to control the colour, and the height of the hand can control the brush size; the current selection is displayed as a circular cursor located under the palm. The selection can then be made by moving the hand down quickly and tapping the screen (i.e., the *tap* gesture).

CONCLUSION AND FUTURE WORK

We presented a system that can estimate the height of a person's hand above a DSI multitouch table. With this system, we performed two experiments, the results of which showed that people do not have a clear idea how to select

Figure 8. Our demo paint application allows people to select the colour and size of a brush using the space above the table.

objects in this hover space. In addition, gestures that were frequently chosen tended to underperform those both less-frequently chosen and predicted by designers to perform better. Specifically, based on our pair of studies, we suggest using a tap gesture. That is, people can perform selections by transitioning from the movement occurring above the table to tapping or touching the surface. In addition, our study shows that this technique will perform better than a tap with the other hand, a push gesture, and dwell. In the future we hope to address the limitations of our system that made us unable to analyze some of the gestures frequently selected in the first study or defined by other designers.

ACKNOWLEDGEMENTS

We would like to thank the Natural Sciences and Engineering Council of Canada (NSERC), NSERC Surfnet, and the Graphics Animation & New Media (GRAND) NCE for funding. We also thank the reviewers for their thoughtful and constructive suggestions.

REFERENCES

1. Agarawala, A. and Balakrishnan, R. Keepin' it real: pushing the desktop metaphor with physics, piles and the pen. *Proc. CHI*, ACM Press (2006), 1283–1292.
2. Banerjee, A., Burstyn, J., Girouard, A., and Vertegaal, R. Pointable: an in-air pointing technique to manipulate out-of-reach targets on tabletops. *Proc. ITS*, ACM Press (2011), 11–20.
3. Benko, H. and Wilson, A.D. DepthTouch : Using Depth-Sensing Camera to Enable Freehand Interactions On and Above the Interactive Surface. In *Technical Report MSR-TR-2009-23, Microsoft Research*. 2009.
4. Comaniciu, D. and Meer, P. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence 24*, 5 (2002), 603–619.
5. Echtler, F., Huber, M., and Klinker, G. Shadow tracking on multi-touch tables. *Proc. AVI*, ACM Press (2008), 388–391.
6. Frisch, M., Heydekorn, J., and Dachselt, R. Investigating multi-touch and pen gestures for diagram editing on interactive surfaces. *Proc. ITS*, ACM Press (2009), 149–156.
7. Grossman, T. and Balakrishnan, R. The design and evaluation of selection techniques for 3D volumetric displays. *Proc. UIST*, ACM Press (2006), 3–12.
8. Grossman, T., Hinckley, K., Baudisch, P., Agrawala, M., and Balakrishnan, R. Hover Widgets: Using the Tracking State to Extend the Capabilities of Pen-Operated Devices. *Proc. CHI*, ACM Press (2006), 861–870.
9. Han, S. and Park, J. A study on touch & hover based interaction for zooming. *Proc. CHI*, ACM Press (2012), 2183–2188.
10. Hancock, M., ten Cate, T., and Carpendale, S. Sticky tools: full 6DOF force-based interaction for multi-touch tables. *Proc. ITS*, ACM Press (2009), 133–140.
11. Hansen, J.P., Tørning, K., Johansen, A.S., Itoh, K., and Aoki, H. Gaze typing compared with input by head and hand. *Proc. ETRA*, ACM Press (2004), 131–138.
12. Henze, N., Löcken, A., Boll, S., Hesselmann, T., and Pielot, M. Free-hand gestures for music playback. *Proc. MUM*, ACM Press (2010), 1–10.
13. Hilliges, O., Izadi, S., Wilson, A.D., Hodges, S., Garcia-Mendoza, A., and Butz, A. Interactions in the air: Adding Further Depth to Interactive Tabletops. *Proc. UIST*, ACM Press (2009), 139–148.
14. Johnson, E.A. Touch display—a novel input/output device for computers. *Electronics Letters 1*, 8 (1965), 219–220.
15. Kaltenbrunner, M., Bovermann, T., Bencina, R., and Costanza, E. TUIO: A protocol for table-top tangible user interfaces. *Proc. ISon*, (2005).
16. Leibe, B., Starner, T., Ribarsky, W., et al. The Perceptive Workbench: toward spontaneous and natural interaction in semi-immersive virtual environments. *Proc. VR*, IEEE Comput. Soc (2000), 13–20.
17. Marquardt, N., Jota, R., Greenberg, S., and Jorge, J. The Continuous Interaction Space: Interaction Techniques Unifying Touch and Gesture On and Above a Digital Surface. *Proc. INTERACT*, Springer-Verlag (2011), 461–476.
18. Nielsen, M., Störring, M., Moeslund, T.B., and Granum, E. A procedure for developing intuitive and ergonomic gesture interfaces for HCI. *Proc. GW*, Springer (2003), 409–420.
19. Prados, E. and Faugeras, O. Shape from shading. In Springer, ed., *Handbook of Mathematical Models in Computer Vision*. 2006, 375–388.
20. Pyryeskin, D., Hancock, M., and Hoey, J. Extending interactions into hoverspace using reflected light. *Proc. ITS*, ACM Press (2011), 262–263.
21. Spindler, M., Martsch, M., and Dachselt, R. Going beyond the surface: studying multi-layer interaction above the tabletop. *Proc. CHI*, ACM Press (2012), 1277–1286.
22. Takeoka, Y., Miyaki, T., and Rekimoto, J. Z-touch: An Infrastructure for 3D gesture interaction in the proximity of tabletop surfaces. *Proc. ITS*, ACM Press (2010), 91–94.
23. Vogel, D. and Balakrishnan, R. Distant freehand pointing and clicking on very large, high resolution displays. *Proc. UIST*, ACM Press (2005), 33–42.
24. Vogel, D. and Balakrishnan, R. Occlusion-aware interfaces. *Proc. CHI*, ACM Press (2010), 263–272.
25. Wilson, A.D. and Benko, H. Combining Multiple Depth Cameras and Projectors for Interactions On, Above, and Between Surfaces. *Proc. UIST*, ACM Press (2010), 273–282.
26. Wilson, A.D., Izadi, S., Hilliges, O., Garcia-Mendoza, A., and Kirk, D. Bringing physics to the surface. *Proc. UIST*, ACM Press (2008), 67–76.
27. Wilson, A.D. TouchLight: An Imaging Touch Screen and Display for Gesture-Based Interaction. *Proc. ICMI*, ACM Press (2004), 69–76.
28. Wobbrock, J.O., Morris, M.R., and Wilson, A.D. User-defined gestures for surface computing. *Proc. CHI*, ACM Press (2009), 1083–1092.