Camera-Based Selection with Cardboard Head-Mounted Displays

Siqi Luo[1], Robert J. Teather[2(✉)], and Victoria McArthur[3]

[1] School of Computer Science, Carleton University, Ottawa, ON, Canada
[2] School of Information Technology, Carleton University, Ottawa, ON, Canada
rob.teather@carleton.ca
[3] School of Journalism and Communication, Carleton University, Ottawa, ON, Canada

Abstract. We present two experiments comparing selection techniques for low-cost mobile VR devices, such as Google Cardboard. Our objective was to assess the feasibility of computer vision tracking on mobile devices as an alternative to common head-ray selection methods. In the first experiment, we compared three selection techniques: air touch, head ray, and finger ray. Overall, hand-based selection (air touch) performed much worse than ray-based selection. In the second experiment, we compared different combinations of selection techniques and selection indication methods. The built-in Cardboard button worked well with the head ray technique. Using a hand gesture (air tap) with ray-based techniques resulted in slower selection times, but comparable accuracy. Our results suggest that camera-based mobile tracking is best used with ray-based techniques, but selection indication mechanisms remain problematic.

Keywords: Mobile VR · Selection · Google Cardboard

1 Introduction

Combining cheap and lightweight cardboard-style HMDs with mobile devices makes VR more accessible to people than ever before. Devices such as Google Daydream (which includes a plastic HMD shell for a mobile phone and a touchpad controller) or Google Cardboard allow users to employ their mobile phone as a VR head-mounted display (HMD). Considerably more affordable than a dedicated VR device, Daydream is priced at around $140, and Cardboard is around $20. Both devices make it possible for more people to experience VR using ordinary mobile phones. However, one drawback is that Cardboard and similar devices do not offer complex interaction techniques. Google Cardboard is, simply put, a cardboard box to contain the mobile, with two focal length lenses. See Fig. 1. The mobile's built-in inertial measurement unit (IMU) tracks head orientation. The button located on the side of Google Cardboard can be pressed to provide input. Their low cost and simplicity may attract a larger user base than conventional HMDs.

However a major limitation of these devices is that mobile sensors offer low-fidelity tracking, and cannot provide absolute 6DOF position and orientation tracking. As a

C. Stephanidis et al. (Eds.): HCII 2020, LNCS 12428, pp. 383–402, 2020.
https://doi.org/10.1007/978-3-030-59990-4_29

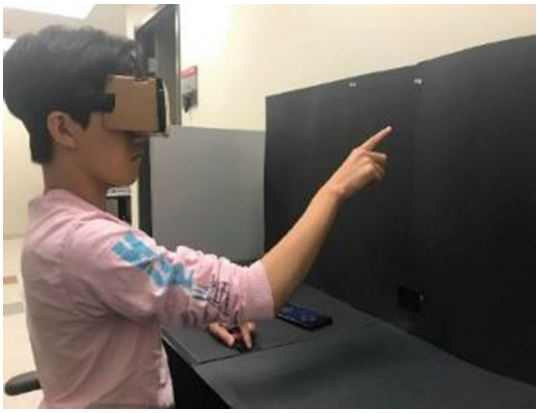

Fig. 1. Participant performing a selection task wearing a Google Cardboard HMD. The button can be seen at the top-right of the device. No other external controller is provided.

result, interaction on mobile VR is more limited than with trackers offered by high-end HMDs. Many past VR interaction techniques require absolute position tracking; without it, only a few of these techniques are compatible with mobile VR [17]. Browsing the Google Play store, one can see that the variety of applications is limited. Most are "look-and-see" type applications, which involve a fairly passive user experience of watching videos in 3D, sometimes using IMU-based head tracking to allow the user to look around the scene [17]. Additionally, there is little research on whether using the type of button provided on cardboard devices is the best design alternative for selecting targets in mobile VR applications. Yet, most mobile VR applications rely on the Cardboard button, to confirm selections despite measurably worse user feedback than alternative approaches [23].

Our research evaluates selection methods using "Cardboard-like" HMDs. Specifically, we investigate the performance of the built-in cameras available on virtually all modern cellphones. With appropriate software, such cameras can track the hands to provide absolute pose information, and also support gesture-recognition [5, 8]. Hence, they may provide a good alternative to head-based selection using a button. Although the tracking quality is low, such research can guide future mobile device development, for example, if higher resolution or depth cameras are required to provide better mobile VR interaction.

2 Related Work

2.1 Mid-air Interaction

Interaction in 3D is more complex than 2D depending on the task [12]. Interaction in 2D only requires up to 3 degrees of freedom (DOF) including translation in the x and y-axes and rotation around the y-axis. In contrast, full 3D interaction requires three additional DOFs: z-axis translation and two more rotational DOFs around the z- and x-axes. Consequently, 2D interaction initially appears to be a poor match for 3D scenarios

[1]. However, while additional DOFs can give more freedom and support a wider range of interaction techniques, they can also be a source of frustration [14]. Previous work suggests minimizing the required DOF for manipulating virtual objects. Generally, the more simultaneous DOFs required, the greater the difficulty to control the interaction technique [2, 7, 20]. Additionally, the absence of tactile feedback and latency and noise common to motion tracking systems also impacts user performance [5, 15]. Current alternatives involve using DOF-limiting techniques in 3D environments [18, 20], for example, modelled after mouse control.

2.2 Selection in 3D

Selection is a fundamental task in VR [1, 12], typically preceding object manipulation. Improving selection time can improve overall system performance [1]. Many factors impact selection accuracy, including the target's size and distance [6], display and input device properties [19], object density [22], etc. For example, input device degrees of freedom (DOF) influence selection, with lower DOFs generally yielding better performance [4]. Display size and resolution also affect performance [15].

Two major classes of VR selection techniques include ray-casting and virtual hands [12]. Past research has shown that virtual hands tend to perform better in high accuracy (and nearby) tasks [13]. This is likely due to a combination of proprioception [14] and good visual feedback. However, in mobile VR scenarios, good 3D pose data is unavailable; such devices only provide head orientation, but not position. As a result, users lose the advantage of head motion parallax depth cues which have long been known to be beneficial in 3D selection [2] and other 3D tasks [11]. Since they rely on good depth perception, virtual hand techniques may offer demonstrably worse performance than ray-casting when implemented on mobile VR platforms.

2.3 ISO 9241 and Fitts' Law

Our experiments employ the ISO 9241-9 standard methodology for pointing device evaluation [9]. The standard is based on Fitts' law [6] and recommends the use of throughput as a primary performance metric.

Fitts' law models the relationship between movement time (MT) and selection difficulty, given as index of difficulty (ID). ID, in turn, is based on target size (W) and distance to the target (D):

$$ID = \log_2 \frac{D}{w} + 1 \tag{1}$$

Throughput is calculated as:

$$TP = \frac{\log_2\left(\frac{De}{We} + 1\right)}{MT} \tag{2}$$

where

$$W_e = 4.133 \times SD_x \tag{3}$$

W_e is *effective* width and D_e is the *effective* distance of movements. D_e, is calculated as the average movement distance from the previous selection coordinate to the current one. W_e is calculated as the standard deviation (SD_x) of the distance between the selection coordinate and the target, multiplied by 4.133 [9]. This adjusts target size post-experimentally to correct the experiment error rate to 4% (i.e., 4.133, or ±2.066 standard deviations from the mean, corresponding to 96% of selections hitting the target). We adopted a previously validated methodology for extending the standard into 3D scenarios [19, 20]. This approach projects cursor/ray selection coordinates into the target plane and uses the projected coordinates in calculating both D_e and W_e, to address the angular nature of ray-based selections. An alternative approach proposed by Kopper et al. [10] instead employs the angular size of motions and targets to address this problem. We employ the projection method, since we also use a direct touch (i.e., virtual hand) technique, where angular measures do not apply.

3 Common Methodology

We first present the elements common to our two experiments, with experiment-specific details appearing in subsequent sections. The first experiment compared three different selection techniques using a mobile VR HMD, while the second focused on evaluating different selection indication mechanisms.

3.1 Participants

The same 12 participants took part in both experiments. There were 2 female and 10 male participants, aged between 18 and 30 years old (mean ≈ 22.67 years old). Two were left-handed. All had normal or corrected-to-normal stereo vision.

3.2 Apparatus

We used a Samsung Galaxy S8 smartphone as the display device. The device has a 5.8 in. screen at 1440 x 2960 pixel resolution and 12-million-pixel main camera. We used a Google Cardboard v2 as the HMD (Fig. 1). The Cardboard has a button on the right side; pressing the button taps the touchscreen inside the HMD.

The virtual environment (Fig. 2, right) was developed using Unity3D 5.5 and C# and presented a simple selection task based on ISO 9241-9 [9]. The overall target sequence is seen in Fig. 2 (left). A single target sphere appeared per selection trial at a specified position. Target distance was always fixed at 0.8 m.

Targets were initially displayed in blue and became pink when intersected by the selection ray/cursor. The first target appeared at the top of the ring cycle. Upon clicking, the target disappeared, and the next target appeared, whether the target was hit or not. An example of two subsequent targets being selected, and the pink highlight are shown in Fig. 3

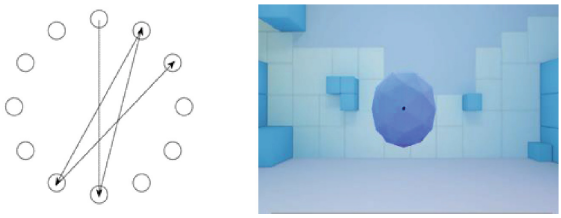

Fig. 2. (Left) ISO 9241-9 selection task target ordering pattern. (Right) The participant's view of a target in the virtual environment.

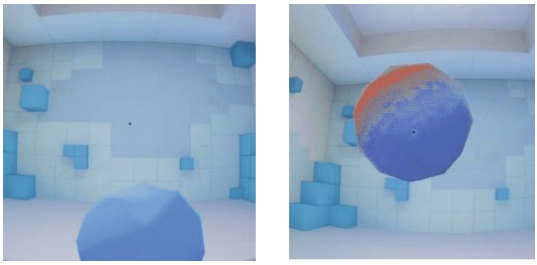

Fig. 3. (Left) Bottom target prior to selection, partially cut off by the edge of the HMD field of view. (Right) Subsequent (top-left) target depicting cursor intersection highlight. (Color figure online)

Selection Techniques. Our study included three selection techniques: head ray, finger ray, and air touch. All three techniques were evaluated in Experiment 1, while only head ray and finger ray were included in Experiment 2.

Head ray is a typical interaction technique used in mobile VR. Selection is performed using a ray originating at the user's head [24] and the ray direction is controlled by the orientation of their head via the mobile IMU (see Fig. 4). A black dot in the center of the viewport provided a cursor to use for selection. When the participant turned their head, the cursor always remained in the center of the viewport.

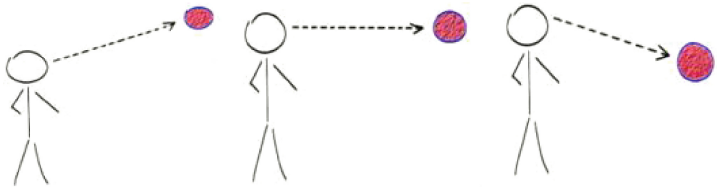

Fig. 4. Target selection with head ray.

The finger ray technique is similar to image-plane interaction [16], and also uses a ray originating at the head. The direction of the ray is controlled by tracking the user's index fingertip with the mobile camera (Fig. 5). A black dot at the end of the ray acts as a cursor.

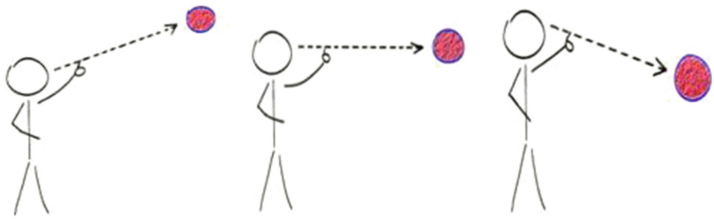

Fig. 5. Target selection with finger ray.

Finally, air touch is a representative virtual hand selection technique and mimics real-world selection. With air touch, the user must physically tap the targets in space, and thus requires depth precision (Fig. 6).

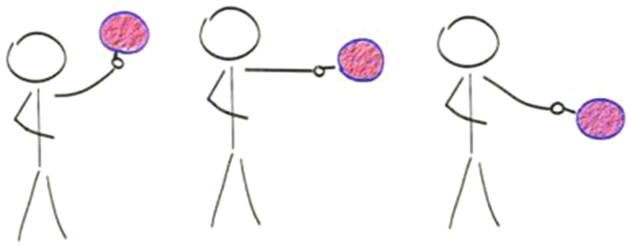

Fig. 6. Target selection with air touch.

We used the Manomotion SDK[1] to acquire the hand position for both the finger ray and air touch conditions. Manomotion uses the built-in RGB camera on the back of a smartphone to track the user's hand, providing coordinates for their fingertips and the center of their palm. Since the built-in mobile device RGB camera has no depth sensor, hand depth was determined by the SDK's proprietary algorithms. The further the hand moves from the camera the larger the reported z value. To ensure the farthest targets were still reachable with the air touch condition (which required directly touching targets and hence precision in depth) we iteratively adjusted a scale factor between the VE and the Manomotion-provided depth coordinates. In the end, a distance of 2 m in the VEs mapped to approximately 70 cm of actual hand motion in reality. This ensured that the farthest targets (2 m into the screen) were still reachable from a seated position with air touch.

3.3 Procedure

After describing the experiment and obtaining informed consent, participants completed a demographic questionnaire. Next, they sat down and put on the HMD. We then gave them instructions about how to control each selection technique and gave them about a

[1] www.manomotion.com.

minute to practice using the system. These practice trials were not recorded. The first target sphere would appear at the center of the viewport. After selecting the first target, the formal test began. Participants confirmed each selection by using the current selection indication method. Targets appeared in the VE following the ring pattern common to ISO 9241-9 evaluations, as described above [9]. Upon completing one condition, which consisted of 72 trials ($12 \times 2 \times 3$), participants were given approximately 1–2 min to rest before beginning the next condition. After all conditions were completed, they filled out a preference questionnaire.

3.4 Design

The dependent variables in both experiments included movement time (s), error rate (%) and throughput (bit/s). Movement time was calculated from the beginning of a selection trial when the target appears, to the time when the participant confirms the selection by pressing the button on the secondary touchpad. Error rate was calculated as the percentage of trials where the participant missed the target in a given block. Throughput was calculated according to Eq. 2 presented in Sect. 2.3. Finally, we also collected subjective data using questionnaires and interviews after each participant completed the experiment.

4 Experiment 1: Selection Techniques

This experiment compared performance of the selection techniques described above: air touch, finger ray, and head ray. Experiment-specific details that differ from the general methodology sections are now described.

4.1 Experiment 1 Apparatus

For experiment 1, we also used a secondary touchpad device: a Xiaomi cellphone to provide an external selection indication method. This device was connected by Bluetooth to the Galaxy S8. To indicate selection, participants tapped the thumb-sized "A" button displayed on the device's touchscreen (see Fig. 7).

Fig. 7. Xiaomi cellphone used as secondary selection device in experiment 1.

While we note that this is not a realistic selection indication mechanism, we decided to include it to ensure consistency between the selection techniques. This avoids conflating selection technique and selection indication method, for example, using gestures to indicate selection with air touch and finger ray, and the Cardboard button with head ray. We considered using the Cardboard button instead, but this would prevent using the right hand with both finger ray and air touch. Since most people are right-handed, using the left hand to perform hand postures while using the right hand to press the button would certainly provide unrealistic performance results.

4.2 Experiment 1 Design

The experiment employed a 3 × 2 × 3 within-subjects design with the following independent variables and levels:

Technique: Head ray (HR), finger ray (FR), air touch (AT);
Object depth: Close (1.3 m), medium (1.7 m) and far (2 m);
Object size: Large (0.7 m) and small (0.4 m);

For each selection technique, participants completed 6 blocks (3 object depths × 2 object sizes) of 12 selection trials, for a total of 72 selections with each selection technique per participant. Across all 12 participants this yielded 2592 trials in total. Each selection trial required selecting one target sphere. Within a block, both target depth and target size were constant. Target size and selection technique order was counterbalanced according to a Latin square, while depth increased with block.

4.3 Result and Discussion

Movement Time. Movement time (*MT*) is the average time to select targets. Mean movement time for each interaction method is shown in Fig. 8.

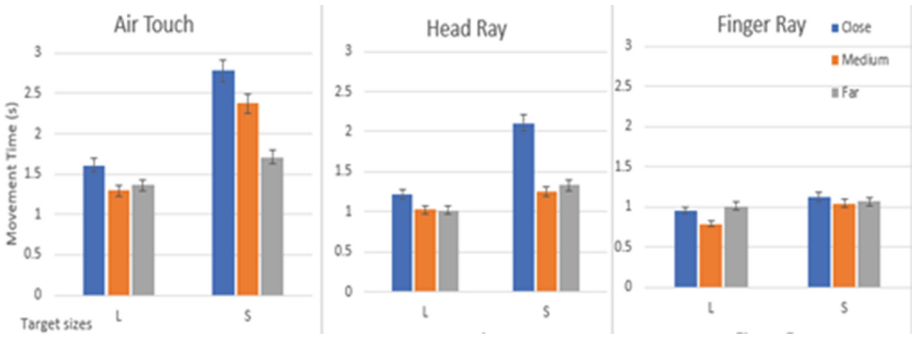

Fig. 8. Movement time by target size, depth, and technique. Error bars show ±1 *SD*.

Movement time was analyzed with repeated-measures ANOVA. The result ($F_{2,18}$ = 33.98, $p < .001$) revealed that technique had a significant main effect on MT. Post

hoc testing with the Bonferroni test (at the $p < .05$ level) revealed that the difference between all techniques was significant. Overall, air touch was worst in terms of speed. Finger ray was faster than both air touch and head ray. Air touch offered significantly higher movement times than the other selection techniques at 1.86 s, about 50% slower than head ray and almost twice as long as finger ray. Head ray was, on average, slower than finger ray.

There were also significant interaction effects between techniques and target size ($F_{2,18} = 3.91$, $p < .05$), as well as technique and target depth ($F_{4,36} = 2.98$, $p < .05$). Predictably, smaller size targets generally took longer for all techniques. However, the interaction effect indicates that this was most pronounced with air touch, where small targets were about 60% worse than large targets. Target size only affected the ray-based techniques when the distance was close or medium ($p < 0.05$).

We had expected that camera noise would affect movement time for both camera-based techniques. It is encouraging that despite camera noise, finger ray still offered faster selection than head ray. As noted earlier, the combination of large target size and closer target distances (close and medium) resulted in a significant difference in movement time between head ray and finger ray. This is likely because head ray required more head motion than finger ray, which allowed subtle finger or arm movements. This is consistent with previous research that also found that finger-based selection was faster than the head when reaching a target [18].

Error rate. A selection error was defined as missing the target, i.e., performing the selection while the cursor is outside the target. Error rate is thus calculated as the percentage of targets missed for each experiment block (i.e., 12 selections). Average error rates for each condition are seen in Fig. 9.

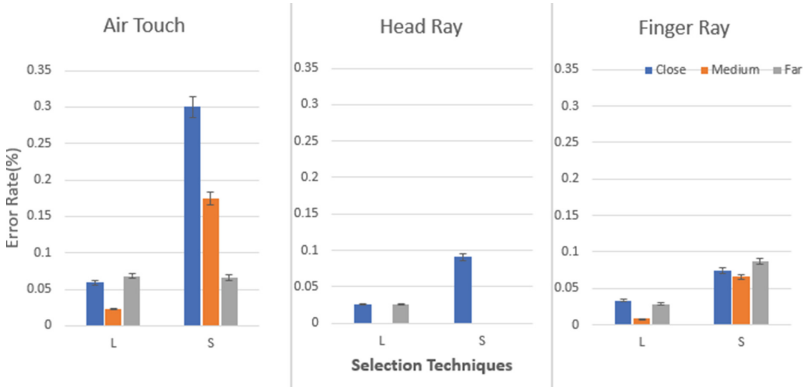

Fig. 9. Error rate by target size, depth, and technique. Error bars show ±1 *SD*.

A repeated measures ANOVA revealed that there was a significant main effect for technique ($F_{2,18} = 385.52$, $p < .001$). Post hoc testing with the Bonferroni test (at the $p < .05$ level) revealed significant differences between each technique. Air touch had a significantly higher error rate than other two selection techniques, five times that of

head ray, and around double that of finger ray. Moreover, there were significant two-way interactions between selection technique and both target size ($F_{2,18} = 4.28$, $p < .001$) and target depth ($F_{4,36} = 5.58$, $p < .001$).

Specifically, post hoc testing with the Bonferroni test (at the $p < .05$ level) revealed that target depth did not yield a significant difference with finger ray or head ray. However, there were significant differences between target depth with air touch. Overall, medium distance (1.7 m) had the lowest error rate across all selection techniques. Close targets had a worse error rate, likely due to ergonomic reasons; close targets falling partially outside the field of view required more movement to select them and often required an unnatural arm pose. Air touch and head ray performed significantly worse at close distance than other two depths. Notably, as seen in Fig. 9 the single worst condition was the combination of close small targets with air touch.

Finally, post hoc testing revealed a significant difference ($p < .001$) between target sizes with air touch, but not the other two selection techniques. The error rate increased dramatically with the smaller target size when using the air-touch interaction method. Generally, smaller target sizes had higher error rate for each selection technique. Air touch was much less precise and had a higher error rate than the other two selection techniques, especially with different target depths. This indicates that participants had difficulty identifying the target depths when using direct touch. As expected, this was not a problem with the ray techniques.

Throughput. The average throughput for each technique is seen in Fig. 10. ANOVA revealed that selection techniques had a significant main effect on throughput ($F_{2,18} = 90.63$, $p < .001$). Post hoc testing with the Bonferroni test (at the $p < .05$ level) showed that all three techniques were significantly different from one another. Finger ray had the highest throughput at 2.6 bit/s, followed by head ray (1.8 bit/s) and then air touch (1.0 bit/s). Throughput for head ray was in line with recent work reporting about 1.9 bit/s when using a similar head-based selection method [18]. Surprisingly, finger ray also offered higher throughput than using a mouse in a HMD-based VR environment in the same study [18]. This suggests there is merit to using camera-based finger tracking as an alternative to common head-based selection in mobile VR.

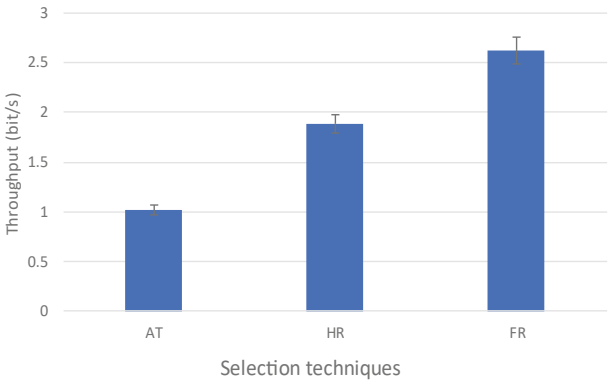

Fig. 10. Throughput for each technique. Error bars show ±1 SD.

Subjective Data and Interview. Participants completed a questionnaire ranking the techniques. Overall, air touch ranked worst of the three (see Fig. 11).

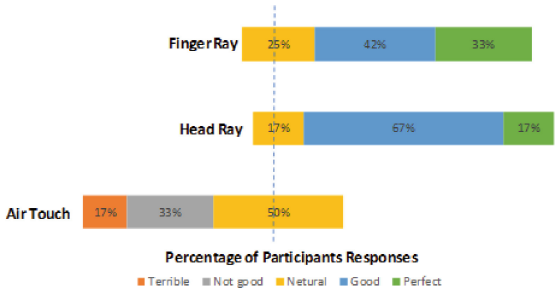

Fig. 11. Participants preference for each technique.

We interviewed each participant after Experiment 1 to solicit their feedback about each selection technique. Most participants mentioned physical fatigue with air touch. They found it difficult to hit targets because they needed to adjust their hand position forward and backward constantly to find the target position in depth. This result is similar to previous research on visual feedback in VR [21], which reported the same "homing" behavior, and found that highlighting targets on touch increased movement time but decreased the error rate. As reported in previous work [20, 21], stereo viewing appeared to be insufficient for participants to reliably detect the target depth with air touch, necessitating the use of additional visual feedback. Although we added colour change upon touching a target and used a room environment to help facilitate depth perception, it seemed participants still had difficulty determining target depth.

Fatigue was high at the largest target depth; participants had to stretch their arms further to reach the targets, which made their upper arms and shoulder even more tired. Head ray also yielded some neck fatigue, especially for close targets, as these increased the amount of required head motion compared to farther targets; targets could be potentially partially outside the field of view.

Despite poor performance, two participants reported that they preferred air touch because the found "it is interesting to really use my finger to touch the targets rather than just turning around my head and touching the button." They found the latter "very boring." All participants found that head ray was most efficient. "It is very easy to control and fast." However, 10 participants said they would choose finger ray as their favorite because "it is convenient to move my fingers slightly to hit the target. I did not even need to move my head and arms."

5 Experiment 2: Selection Indication

This experiment compared different methods of indicating selection, or phrased differently, methods of "clicking" a target. The first experiment showed that finger ray and head ray can offer higher selection performance than air touch in mobile VR scenarios.

However, for reasons described earlier, we used an external touchpad as the selection indication mechanism. To address this limitation, Experiment 2 compared two alternative methods of selection indication, including hand gestures and buttons. We also included the Experiment 1 data for the head and finger ray conditions to compare the touchpad selection indication mechanism as a baseline.

In consideration of participants' time, and to maintain a manageable experiment size, we decided to remove one selection technique from Experiment 1. We ultimately decided to exclude air touch from this experiment for two reasons: 1) our informal observations prior to conducting Experiment 1 suggested that air touch would be less effective than the other techniques (as confirmed by our results), and 2) air touch was found to work less reliably with the gesture selection indication method detailed below than finger ray and head ray.

5.1 Experiment 2 Apparatus

This experiment used only the head ray and finger ray techniques from Experiment 1. The separate Xiaomi smartphone, used to indicate selection in Experiment 1, was not used in Experiment 2. Instead, Experiment 2 used two new selection indication methods: the Google Cardboard button (CB) and the *tap* hand gesture (HG).

The tap gesture is similar to that performed when selecting an icon on a touchscreen device, see Fig. 12 (left) [17]. It requires bending the finger at the knuckle to indicate a selection. We originally considered a pinch gesture instead, which involved bending index finger and thumb like a "C" shape, then closing the fingertips. However, the tracking SDK was unable to reliably detect the pinch gesture, yielding longer selections times than tap.

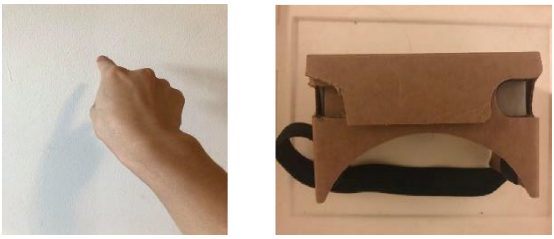

Fig. 12. Left: Tap gesture. Right: Modified Google Cardboard with both left and right-sided button

Participants used their dominant hand to perform the tap gesture. When using the finger ray selection method, they had to keep the pointing finger stable until the tap gesture was performed. In the Cardboard button condition, participants pressed the capacitive button built into the cardboard frame. Since we had expected most participants would be right-handed, we added a second button on the left side of Google Cardboard (see Fig. 12 right). Adding the left-side Cardboard button ensured that participants could always select with their dominant hand, and indicate selection using their non-dominant hand. This ensured consistency with the hand gesture condition, which also always used the dominant hand.

5.2 Experiment 2 Design

To investigate interactions between selection technique and selection indication, this experimented included both finger ray and head ray, and the two selection indication methods described above. The experiment employed a 2 × 2 × 3 × 2 within-subject design with the following independent variables:

Technique: Head Ray (HR), Finger Ray (FR);
Indication: Cardboard button (CB), hand gesture (HG);
Object depth: close (1.3 m), medium (1.7 m) and far (2 m);
Object size: large (0.7 m) and small (0.4 m);

Participants completed Experiment 2 immediately following Experiment 1. As noted previously, our analysis also includes the data for the touchpad selection indication method from Experiment 1 as a comparison point for the two new selection indication mechanisms. Each block consisted of 12 selection trials for each combination of target size and target depth. Selection techniques and selection indication mechanisms were counterbalanced according to a Latin square. Overall, there were 12 participants × 2 pointing methods × 2 selection indication mechanisms × 2 target sizes × 3 target depths × 12 selections = 3456 trials in total.

5.3 Results and Discussion

Results were analyzed by using a repeated-measures ANOVA. Since Experiment 1 exclusively used the touchpad as a selection indication mechanism, we included this data as a basis of comparison with the two new selection indication mechanisms. Specifically, our analysis includes data from Experiment 1 for the finger ray and head ray, for all dependent variables. These are depicted as "FR + TP" and "HR + TP" in the results graphs below (i.e., TP indicates "touchpad" selection indication). On all results graphs, two-way arrows (← →) indicate a pairwise significant difference with post hoc test at 5% significance level. The best performing conditions is highlighted in red.

Completion Time. Mean completion time for each condition is seen in Fig. 13 Completion time was analyzed using a repeated-measures ANOVA, which revealed a significant interaction effect ($F_{5,36} = 22.44$, $p < .001$) between selection technique and selection indication method. Post hoc testing with the Bonferroni test (at the $p < 0.05$ level) revealed a significant difference between head ray and finger ray when using the Cardboard button.

There was no significant difference between head ray and finger ray when using hand gestures, nor when using the touchpad. There were significant differences between touchpad and both the Cardboard button and hand gesture when using finger ray as the selection technique. Finger ray with either the Carboard button or hand gesture yielded higher times compared to the Experiment 1 touchpad. Finger ray also took longer with Carboard button than head ray with Carboard button. This surprised us, given how fast finger ray was in Experiment 1. This highlights the importance of investigating selection indication mechanisms in conjunction with pointing techniques. Results separated by target depth and size are seen in Fig. 14.

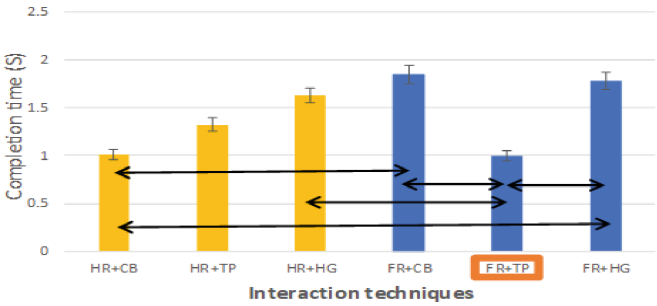

Fig. 13. Average completion time for combinations of techniques. Error bars show ±1 SD (Color figure online)

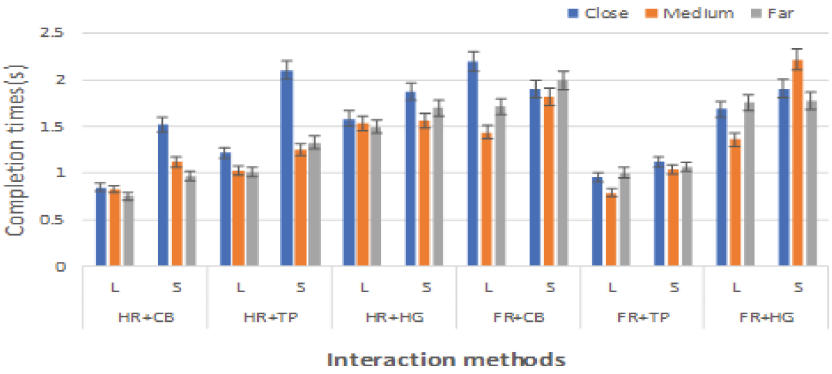

Fig. 14. Completion time in depths, sizes, and combination techniques. Error bars show ±1 SD

Generally, smaller target size yielded slower selections. Like Experiment 1, the medium target depth yielded faster completion times compared to the far and close target distances. The finger ray + touchpad still offered the fastest selection times overall, for every target size and depth combination. This suggests that finger ray itself is a promising technique, if a suitable selection indication method is used with it. However, as discussed earlier, the touchpad is an impractical solution.

Error rate. Average error rates for each technique combination is shown in Fig. 15. Repeated-measures ANOVA revealed a significant interaction effect between selection technique and selection indication method ($F_{5,36} = 16.57$, $p < .001$). Post hoc testing with the Bonferroni test (at the p < . 05 level) revealed significant differences between finger ray + Carboard button and using head ray with all three selection indication mechanisms. The highest error rate was with the finger ray + Cardboard button condition. Notably, hand gestures worked better with both finger ray and head ray, than either selection method worked with Cardboard button, which had highest error rate for both selection methods.

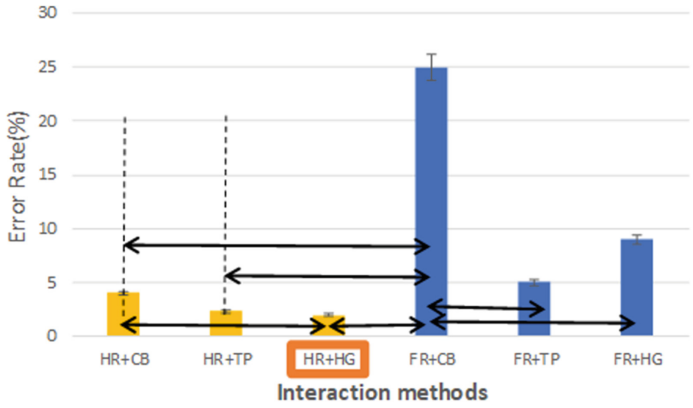

Fig. 15. Average error rate for combinations of techniques. Error bars show ±1 SD

According to a Bonferroni post hoc test (at the $p = .05$ level), target depth did not have significant effects on error rate with the exception of the finger ray + hand gesture combination. Finger ray + Cardboard combination had a higher error rate in each combination of target depths and sizes than other interaction methods.

Throughput. Average throughput for each technique combination is shown in Fig. 16. Repeated measures ANOVA ($F_{5,36} = 70.08, p < .001$) indicated a significant interaction effect between selection technique and selection indication method. Post hoc testing with the Bonferroni test (at the $p < .05$ level) revealed a significant difference in throughput between each selection indication method when using head ray. With head ray, the hand gesture performed worst, and Cardboard button performed best. With finger ray, throughput was much lower with the hand gesture and Cardboard button than with the touchpad.

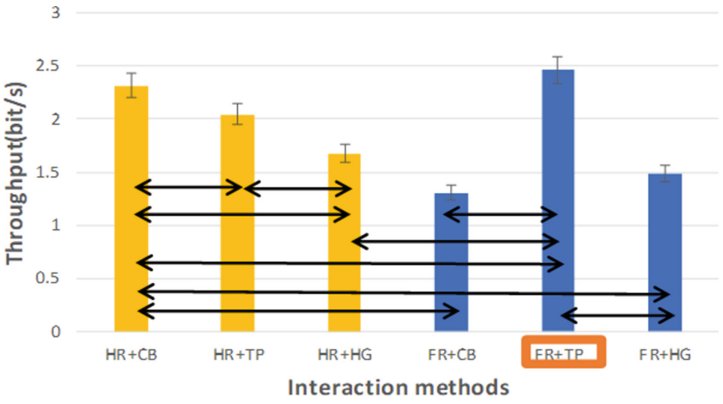

Fig. 16. Throughput by selection technique and selection indication method.

Subjective Results. Based on our questionnaire results (Fig. 17), participants rated finger ray + Cardboard button worst, and head ray + Cardboard button best. Both finger ray + touchpad and head ray + touchpad were ranked positively. During post-experiment interviews, participants indicated that hand gestures were more convenient than pressing the button or touchpad. Several indicated that they found the Cardboard button was sometimes a bit unresponsive, requiring them to press it harder. Some also mentioned that the HMD was not tight enough when they pressed down on the button, requiring them to hold the HMD with their other hand at times. No participants mentioned any physical fatigue in Experiment 2, likely due to the absence of air touch.

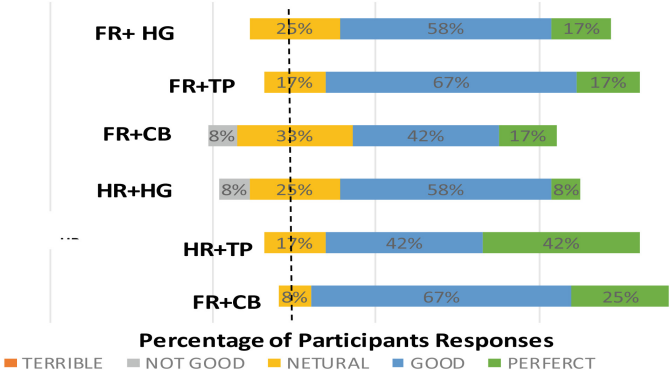

Fig. 17. Questionnaire score from each participant for each combination of techniques

5.4 Discussion

When using the finger ray with either hand gestures or the Carboard button, selection performance was notably worse. This was a surprising result, given that in Experiment 1, finger ray was faster than the head ray. This suggests that neither hand gesture nor the Cardboard button are suitable selection indication mechanisms when using the finger ray selection technique.

Nevertheless, the finger ray selection technique shows promise, despite the poor tracking resolution of a mobile device camera. When used with the touchpad, the finger ray was the best technique overall. This is likely because the touchpad was more reliable than either the Cardboard button or using hand gestures recognized by the built-in camera. As mentioned earlier, participants indicated that the Cardboard button sometimes felt unresponsive. Similarly, participant hand gestures were not always recognized by the tracking SDK. On the other hand, using head ray with the Cardboard button yielded significantly lower completion times, in line with finger ray + touchpad. We suspect this is because head and neck movements resulted in more whole-body movement, unlike finger ray which only required finger movement. For example, when using the head ray, participants had to turn their bodies slightly to face the target. During such movement,

it is faster to press the Cardboard button (since it is positioned on the HMD) rather than tapping the touchpad (which is fixed on the table).

Finger ray + hand gesture took longer than head ray + hand gesture. This result surprised us, as we had expected the hand gesture would be a "natural fit" with the finger ray technique. After all, the hand gesture was performed with the same hand being used to point at targets with finger ray. Our expectation was that participants could perform the hand gesture as soon as the ray intersected the target, which may thus be faster than pressing the Cardboard button. This result can likely be explained by the comparative lack of camera sensitivity. The participants' tap gestures were frequently not recognized on the first try; multiple hand gestures thus increased the time required to select targets. It is possible that a better camera or a different gesture may improve this result.

We were also surprised by the significantly higher error rate for finger ray with the Cardboard button. This may be because of the so-called "Heisenberg" effect [3] in 3D selection, where the selection indication mechanism sometimes moves the pointing device at the instant of selection, which results in missing the target. In this case, when pressing the Cardboard button, the HMD often moved slightly, which moved the selection ray. However, when using head ray, participants frequently used their other hand to hold the Google Cardboard, so the error rate was notably better. In contrast, when using the finger ray, participants used one hand to direct the ray, and the other to press the button. As a result, the error rate increased in that condition. Due to the overall better movement time and accuracy with the Cardboard button, throughput was also higher with the finger ray condition.

In terms of subjective preference, the head ray + Cardboard button was rated best. This indicates that smooth operation during pointing is an important factor for users. Further, there was no physical fatigue was reported during Experiment 2; it seems the air touch technique used in Experiment 1 was the primary cause of the physical fatigue reported by participants. This suggests that finger ray yields much lower fatigue than air touch. In general, head ray + Cardboard button still had the advantage on both throughput, error rate and completion time. However, both head ray and finger ray with hand gestures were not far off, and could be a potential alternative in the future, especially with advances in camera-based tracking.

5.5 Limitations

Our experiments were conducted in "idealized" lab settings with specific lighting levels and background colour chosen to increase contrast. This provided optimal conditions for the camera tracking SDK, despite which, we still observed constant cursor jitter during the experiment. In real-world usage scenarios, there would clearly be worse tracking interference. Hence, our results should be viewed as a "best-case" with current technologies.

Due to the jitter observed in the camera-based selection techniques, target size was also limited. It would be impossible to hit very small targets; previous work has shown that once jitter approaches half the target size, selection accuracy falls dramatically [24]. Through pilot testing, we modified the experiment conditions to account for this problem, but clearly these results would not generalize to smaller targets that would be possible in standard VR systems.

Finally, we note that the depth of the virtual hand in the air touch condition was calculated using a scale factor between the Manomotion SDK and the VE coordinate system. However, the scale factor in the z-axis was fixed in our study, which was not ideal for all participants. For example, one participant with shorter arms had difficulty in reaching the farthest targets. Customizing this ratio for each participant would provide a better user experience.

6 Conclusions

In this paper, we compared potential selection techniques for low-cost mobile VR. Our objective was to assess if alternatives to common head-based selection methods were feasible with current technology, employing computer vision tracking approaches on mobile devices. To this end, our study employed only a smartphone and a cardboard HMD. In the first experiment, we compared air touch, head ray and finger ray in selection tasks. Overall, air touch performed worst, and finger ray performed best. However, since this experiment used an unrealistic selection indication mechanism (to improve experimental internal validity), we conducted a second experiment to compare selection indication methods. Results of Experiment 2 indicated that the secondary touchpad worked very well with finger ray, despite its impracticality. The built-in Cardboard button worked well with head ray.

Our results suggest that finger ray is promising for mobile VR, even when tracked by a single camera. Despite tracking imprecision, the technique performed well when used with an external touchpad. Future research could focus on further investigating potential selection indication methods to use with finger ray. For example, different gestures that are more reliably detectable may yield better performance than the tap gesture used in our experiment. Such gestures would also work in practical contexts, unlike the touchpad used in our first experiment.

In contrast, direct touch techniques like air touch performed very poorly; single-camera hand tracking seems to be out of reach for current mobile device cameras. Our results indicate that higher DOF techniques yield lower performance, consistent with previous results [2, 20]. Direct touch might be possible with more powerful future mobile devices supporting more robust vision-based tracking software, or if depth cameras become common on mobiles. Overall, from Experiment 1, there seems to be greater promise for ray-based selection techniques employing mobile camera tracking than virtual hand techniques.

References

1. Argelaguet, F., Andujar, C.: A survey of 3D object selection techniques for virtual environments. Comput. Graph. **37**(3), 121–136 (2013)
2. Arsenault, R., Ware, C.: The importance of stereo and eye-coupled perspective for eye-hand coordination in fish tank VR. Presence Teleoperators Virtual Environ. **13**(5), 549–559 (2004)
3. Bowman, D., Wingrave, C., Campbell, J., Ly, V.: Using pinch gloves (tm) for both natural and abstract interaction techniques in virtual environments. In: Proceedings of the HCI International. Springer, New York (2001)

4. Bowman, D.A., Hodges, L.F.: An evaluation of techniques for grabbing and manipulating remote objects in immersive virtual environments. In: Proceedings of the ACM Symposium on Interactive 3D graphics - SI3D 1997, pp. 35–38. ACM, New York (1997)
5. Erol, A., Bebis, G., Nicolescu, M., Boyle, R.D., Twombly, X.: Vision-based hand pose estimation: a review. Comput. Vis. Image Underst. **108**(1–2), 52–73 (2007)
6. Fitts, P.M.: The information capacity of the human motor system in controlling the amplitude of movement. J. Exp. Psychol. **47**(6), 381–391 (1954)
7. Hand, C.: A survey of 3D interaction techniques. Comput. Graph. Forum **16**(5), 269–281 (1997)
8. Hernández, B., Flores, A.: A bare-hand gesture interaction system for virtual environments. In: Proceedings of the International Conference on Computer Graphics Theory and Applications (GRAPP), pp. 1–8. IEEE, New York (2014)
9. ISO, ISO: 9241–9 Ergonomic requirements for office work with visual display terminals (VDTs) - part 9: requirements for non-keyboard input devices international standard. Int. Organ. Stand. (2000)
10. Kopper, R., Bowman, D.A., Silva, M.G., McMahan, R.P.: A human motor behavior model for distal pointing tasks. Int. J. Hum. Comput. Stud. **68**(10), 603–615 (2010)
11. Kulshreshth, A., LaViola Jr, J.J.: Evaluating performance benefits of head tracking in modern video games. In: Proceedings of the ACM Symposium on Spatial User Interaction - Sui 2013, pp. 53–60, ACM, New York (2013)
12. LaViola Jr., J.J., Kruijff, E., McMahan, R.P., Bowman, D., Poupyrev, I.P.: 3D User Interfaces: Theory and Practice. Addison-Wesley Professional, USA (2017)
13. Lin, C.J., Ho, S.-H., Chen, Y.-J.: An investigation of pointing postures in a 3D stereoscopic environment. Appl. Ergon. **48**, 154–163 (2015)
14. Mine, M.R., Frederick P., Brooks, J., Sequin, C.H.: Moving objects in space: exploiting proprioception in virtual-environment interaction. In: SIGGRAPH 1997: Proceedings of the ACM Conference on Computer Graphics and Interactive Techniques, pp. 19–26. ACM, New York (1997)
15. Ni, T., Bowman, D.A. Chen, J.: Increased display size and resolution improve task performance in information-rich virtual environments. In: Proceedings of Graphics Interface 2006, pp. 139–146. CIPS, Toronto (2006)
16. Pierce, J.S., Forsberg, A.S., Conway, M.J., Hong, S., Zeleznik, R.C., Mine, M. R.: Image plane interaction techniques in 3D immersive environments. In: Proceedings of the Symposium on Interactive 3D Graphics - SI3D 1997, pp. 39–43. ACM, New York (1997)
17. Powell, W., Powell, V., Brown, P., Cook, M., Uddin, J.: Getting around in google cardboard–exploring navigation preferences with low-cost mobile VR. In: Proceedings of the IEEE VR Workshop on Everyday Virtual Reality (WEVR) 2016, pp. 5–8. IEEE, New York (2016)
18. Ramcharitar, A., Teather, R.J.: EZCursorVR: 2D selection with virtual reality head-mounted displays. In: Proceedings of Graphics Interface 2018, pp. 114–121. CIPS, Toronto (2018)
19. Teather, R.J., Stuerzlinger, W.: Pointing at 3D targets in a stereo head-tracked virtual environment. In: Proceedings of the IEEE Symposium on 3D User Interfaces, pp. 87–94. IEEE, New York (2011)
20. Teather, R.J., Stuerzlinger, W.: Pointing at 3D target projections using one-eyed and stereo cursors. In: Proceedings of the ACM Conference on Human Factors in Computing Systems - CHI 2013, pp. 159 – 168. ACM, New York (2013)
21. Teather, R.J., Stuerzlinger, W.: Visual aids in 3D point selection experiments. In: Proceedings of the ACM Symposium on Spatial User Interaction - SUI 2014, pp. 127–136. ACM, New York (2014)
22. Vanacken, L., Grossman, T., Coninx, K.: Exploring the effects of environment density and target visibility on object selection in 3D virtual environments. In: Proceedings of the IEEE Symposium on 3D User Interfaces - 3DUI 2007, pp. 117–124. IEEE, New York (2007)

23. Yoo, S., Parker, C.: Controller-less interaction methods for Google cardboard. In: Proceedings of the ACM Symposium on Spatial User Interaction - SUI 2015, p. 127. ACM, New York (2015)
24. Zeleznik, R.C., Forsberg, A.S., Schulze, J.P.: Look-that-there: exploiting gaze in virtual reality interactions. Technical report, Technical Report CS-05 (2005)