Stereoscopic Viewing and Monoscopic Touching: Selecting Distant Objects in VR Through a Mobile Device

Joon Hyub Lee
Dept. of Industrial Design, KAIST
Republic of Korea
joonhyub.lee*

Taegyu Jin
Dept. of Industrial Design, KAIST
Republic of Korea
taegyu.jin*

Sang-Hyun Lee
Dept. of Industrial Design, KAIST
Republic of Korea
sang-hyun.lee*

Seung-Jun Lee
Dept. of Industrial Design, KAIST
Republic of Korea
seung-jun.lee*

Seok-Hyung Bae
Dept. of Industrial Design, KAIST
Republic of Korea
seokhyung.bae*

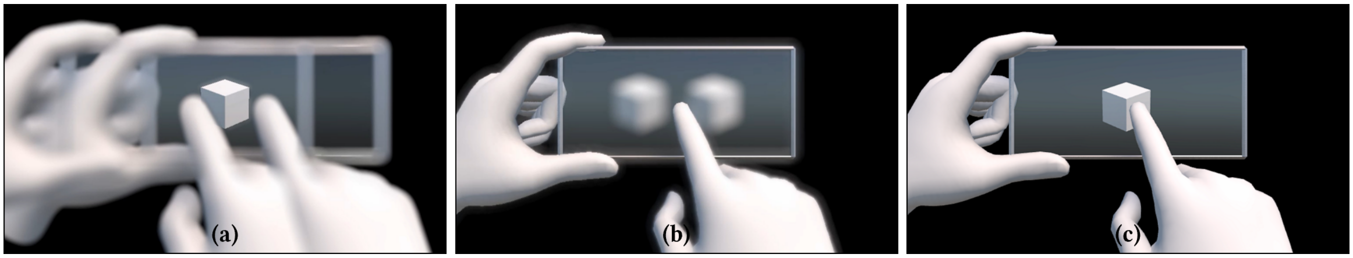

Figure 1: We propose a novel VR interaction technique that enables using a smartphone as a small, transparent window and selecting distant objects seen through it by touching on it. As the above simulations show, binocular parallax can cause double vision even in VR, meaning when the user (a) focuses on the distant object, the near finger appears as two, or (b) focuses on the near finger, the distant object appears as two. (c) We solve this problem by introducing two modes: stereoscopic viewing mode and monoscopic touching mode.

ABSTRACT

In this study, we explore a new way to complementarily utilize the immersive visual output of VR and the physical haptic input of a smartphone. In particular, we focus on interacting with distant virtual objects using a smartphone in a through-plane manner and present a novel selection technique that overcomes the binocular parallax that occurs in such an arrangement. In our proposed technique, when a user in the stereoscopic viewing mode needs to perform a distant selection, the user brings the fingertip near the screen of the mobile device, triggering a smoothly animated transition to the monoscopic touching mode. Using a novel proof-of-concept implementation that utilizes a transparent acrylic panel, we conducted a user study and found that the proposed technique is significantly quicker, more precise, more direct, and more intuitive compared to the ray casting baseline. Subsequently, we created VR applications that explore the rich and interesting use cases of the proposed technique.

CCS CONCEPTS

• **Human-centered computing → Interaction techniques**.

KEYWORDS

Virtual reality, smartphone, transparency, distant selection, binocular parallax

ACM Reference Format:
Joon Hyub Lee, Taegyu Jin, Sang-Hyun Lee, Seung-Jun Lee, and Seok-Hyung Bae. 2023. Stereoscopic Viewing and Monoscopic Touching: Selecting Distant Objects in VR Through a Mobile Device. In *The 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23), October 29-November 1, 2023, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3586183.3606809

1 INTRODUCTION

In the near future, immersive VR will become more commonplace; however, it is likely that mobile devices such as smartphones will not be completely replaced and will continue to be used alongside VR. This is similar to the manner in which mobile devices became commonplace yet desktop PCs were not completely replaced and continue to be used alongside mobile devices. Thus, researchers are exploring various mixed modality interactions that aim to combine the strength of each modality involved.

Among the many peripheral devices for VR, the familiar smartphone offers versatility in a wide range of common VR scenarios. It is small and light, meaning it can be used as a handle for moving 3D objects (as-plane interactions) [6, 11, 15, 17, 22, 28, 36, 39], and it works with multi-touch and pen, meaning it can be used as an input surface where UI elements such as menus and buttons can be touched, and where writings and drawings can be made (on-plane interactions) [2, 5, 6, 9–11, 17, 21, 29, 31, 38].

In particular, using the screen of the mobile device as an image plane and projecting a 3D ray that originates from the user's viewpoint through the touch point can lead to an intuitive and effective

[*] Authors' email addresses are { firstname.lastname }@kaist.ac.kr.

UIST '23, October 29-November 1, 2023, San Francisco, CA, USA

ACM ISBN 979-8-4007-0132-0/23/10…$15.00
https://doi.org/10.1145/3586183.3606809

way of dealing with distant objects in large VR spaces (through-plane interactions). We introduce two distinct modes, stereoscopic viewing and monoscopic touching, as well as an animated transition between the two, to overcome the usability issues associated with binocular parallax that occur in this arrangement (Figure 1).

2 RELATED WORK

In this section, we contextualize our contribution in relation to previous studies that utilize mobile devices such as smartphones as an input modality in immersive environments. Moreover, we draw comparisons to studies on distant interactions utilizing transparent panels and problems associated with the binocular parallax in such an arrangement. Since many available VR headsets also support optical passthrough AR [26], we do not distinguish between AR and VR in this section. However, we exclude mobile AR, which overlays virtual content on top of the image of the physical space displayed on the small screen of the mobile device, and only include headsets that provide immersive visual experiences.

By extending the focus-plus-context concept [4] to spatial interactions, Grubert et al. were the first to show the visionary concepts of mixed modality interactions [13] that were further developed in later studies. Some researchers proposed displaying auxiliary 2D contents in the 3D space around a mobile device and interacting with the main one by bringing it onto the touchscreen [6, 11, 17, 29, 31]. Others proposed using a mobile device as a pointing device to cast a ray [8, 23, 24, 35, 36, 39] or map a trackpad [6, 8, 15, 18, 35], a menu bar to display contextually relevant items [32], a handle to manipulate 3D contents such as 3D CAD models and 3D charts fixed to it [6, 11, 15, 17, 22, 36, 39], a cutting tool to reveal the cross sections of 3D volumes [6, 22, 28], and a flat surface to write and draw immersively [2, 5, 9, 10, 21, 38].

These previous studies mainly focused on interactions that utilize the thin, flat, and rigid form of mobile devices (as-plane), their touch and pen input capabilities (on-plane), or a combination of the two. Relatively unexplored is the metaphor of a transparent window for interacting with distant objects seen through it (through-plane), rendering the viewport in the user perspective rather than the device perspective [3, 37] and treating the mobile device as a graspable and touchable image plane to enable intuitive interactions comparable to direct manipulation, even for objects that lie beyond the user's reach [30]. Similar approaches have been tried in tabletop setups [33, 34, 37] or a virtual viewport attached to a wand-type VR controller [3], but never on a mobile device in VR in a way that takes advantage of both. Although conceptually discussed in previous studies [20, 39], this idea had yet to be implemented and formally evaluated for a common task, such as selection, against a common baseline, such as ray casting, until this study.

Binocular parallax inevitably occurs when the user tries to interact with distant objects stereoscopically seen through a transparent panel [20], as is the case in our technique. Depending on the size and distance of the object from the panel relative to the distance of the panel from the eyes, the selection performance may be severely degraded or even impossible [19, 20]. The previous solution relied on the duplicating phenomenon [19], was limited in that the selecting finger still appeared duplicated, and only worked for one target at one location at a time. The fundamental solution to this is to render the affected area monoscopically. Doing so on a large wall has been found to be effective at resolving any analogous issue that arises when selecting distant targets by casting rays through the wall [1]. We are the first to apply such an approach to the mobile device form factor in VR, devise a transformation that interpolates between stereoscopic and monoscopic rendering for back-and-forth transitions between the two, and evaluate its usability.

3 TECHNIQUE

We introduce a novel selection technique that utilizes a mobile device, such as a smartphone, of which the 6DOF movement is tracked in real time within the immersive environment of VR. Our technique consists of two modes, stereoscopic viewing and monoscopic touching, and an animated transition to make the switch between the two modes visually comfortable.

3.1 Stereoscopic Viewing Mode

The stereoscopic viewing mode is activated by default when the user simply holds the mobile device in the non-dominant hand. In this mode, the mobile device is rendered as a transparent glass panel, allowing the user to view a distant object through it in stereoscopic 3D. The object appears visually identical inside and outside of the frame of the mobile device, ensuring that the user can first gaze at a distant object with both eyes and then lift the mobile device to eye level without taking the eyes off the object (Figure 2a).

3.2 Monoscopic Touching Mode

In the stereoscopic viewing mode, binocular parallax occurs when the user brings the index finger of the dominant hand close to the screen of the mobile device to select a distant object viewed through the device, hindering precise selection (Figure 1a, b). Therefore, when the fingertip closely approaches the screen, the monoscopic touching mode is activated. In this mode, the viewport is rendered monoscopically in the perspective of the user, specifically the midpoint between the two eyes (mid-eye), preventing binocular parallax from occurring (Figure 2b).

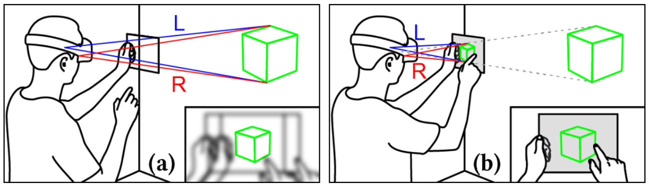

Figure 2: (a) When simply holding the smartphone, it acts like a clear glass panel and the user can stereoscopically view through it. (b) When the fingertip comes closer to the smartphone, it acts like an opaque perspective-corrected monoscopic viewport that the user can touch without binocular parallax.

3.3 Animated Flattening Transition

Although the perspective is unchanged, switching between the stereoscopic viewing and monoscopic touching modes requires a sudden shift in the vergence depth from the distant object to the near smartphone screen. This can be visually perplexing and cause eyestrain [16]. Therefore, the transition between the two modes may be smoothly animated depending on the distance between the fingertip and the screen surface and gradually guides the vergence shift (Figure 3).

In the stereoscopic viewing mode, the distant object retains its original volume, but in the monoscopic touching mode, it is reprojected homographically flat onto the screen of the mobile device. This *flatness* can be interpolated.

Specifically, vertices of meshes of VR objects are either brought forth to their on-screen positions on the surface or sent back to their original in-space positions along the line of sight of the midpoint between the two eyes (mid-eye). When the vertices move at a linear rate in-space (Equation 1), they appear to move at a nonlinear rate on-screen from the perspectives of the left and right eyes, leading to distracting distortions that are visually similar to the hyperspace jump effect in *Star Wars*. Therefore, we move the vertices at a nonlinear rate in-space (Equation 2) to ensure that their on-screen reprojections move at a linear rate.

$$\mathbf{v}(s) = \mathbf{v}_0 + s(\mathbf{v}_1 - \mathbf{v}_0) \tag{1}$$

$$\mathbf{v}(s) = \mathbf{v}_0 + \frac{(1+k)s}{1+ks}(\mathbf{v}_1 - \mathbf{v}_0) \tag{2}$$

$$\text{where } k = \frac{(\mathbf{v}_0 - \mathbf{c}) \cdot \mathbf{n}}{(\mathbf{c} - \mathbf{m}) \cdot \mathbf{n}}$$

The terms of the equations are $\mathbf{v}$: vertex position during transition, $\mathbf{v}_0$: original vertex position, $\mathbf{v}_1$: completely flattened vertex position calculated as the intersection between the line connecting the mid-eye position and the original vertex, and the screen plane, $\mathbf{m}$: mid-eye position, $\mathbf{c}$: screen plane center position, $\mathbf{n}$: screen plane normal, and s: *flatness*, a scalar inversely proportional to the fingertip-to-surface distance that ranges from 0 (completely unflat) to 1 (completely flat). As a result, when the fingertip approaches a certain distance from the screen (e.g. 10 cm), the transition from stereoscopic viewing to monoscopic touching will start, and when it nears the screen (e.g. 5 cm), the transition will be complete. The process is reversed when the fingertip departs from the screen.

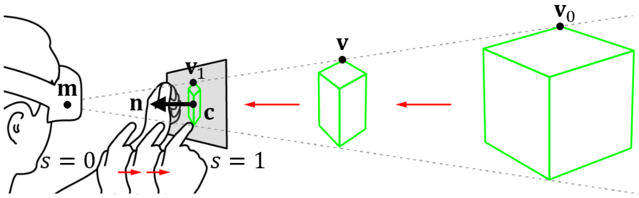

Figure 3: As the fingertip approaches the screen surface to perform a selection, the transition is smoothly animated by interpolating the intermediate 3D volume (middle) between the stereoscopic viewing mode, where the object retains its original 3D volume (right), and the monoscopic touching mode, where the object is flattened to a 2D picture by reprojection onto the screen surface as seen by the midpoint between the two eyes (left).

4 PROOF OF CONCEPT

Implementing the proposed technique in VR requires quick and precise acquisition of the real-time position and orientation of the physical smartphone, the posture of the non-dominant hand holding it and the dominant hand touching it, and the occurrence and position of the touch.

However, commonly used external trackers can add substantial volume [31, 39] and weight [3, 15, 24, 28] when attached to the small and light smartphone, which can interfere with natural gripping or moving and cause fatigue. Although displaying AR markers on the smartphone screen can deliver the most unaltered experience of using a smartphone in VR, the tracking latency is high and can fail when moving too fast [23, 36].

Therefore, to implement a proof of concept that is minimalistic and lightweight, we instead used the Meta Quest 2 VR headset and its optical hand tracking capability. For the smartphone, we used an acrylic panel weighing 50g that was the size and shape of an iPhone 14. As the panel was transparent, the hand tracking camera on the VR headset could see through it to track the finger posture, from which the position and orientation of the panel being held could be calculated (Figure 4a, b). In addition, we taped a small piece of metal to the tip of the finger and determined contact from the sound of the metal touching the panel.

The same implementation was used for both the ray casting (Figure 4c) and the proposed technique (Figure 4d) to avoid confounding factors such as differences in device weight and tracking precision that can affect selection performances.

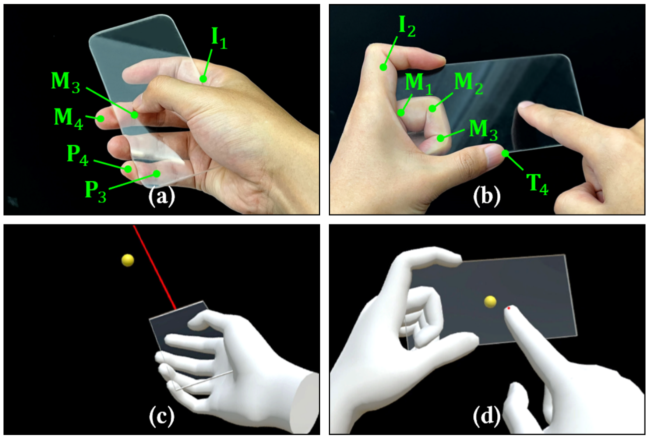

Figure 4: The position and orientation of the transparent acrylic panel was calculated from the real-time tracked positions of the joints of the fingers that held it: T (thumb), I (index finger), M (middle finger), and P (pinkie). (a) For the ray casting technique, it was calculated as the plane that contained I_1 and the midpoints between M_3 & M_4 and P_3 & P_4, with the midpoint of the left edge being the midpoint between M_3 & M_4 and the left edge containing the midpoint between P_3 & P_4. (b) For the proposed technique, it was calculated as the plane that contained M_1, M_2, and M_3, with the upper left corner being I_2 and the lower edge containing T_4. Performing (c) the ray casting and (d) the proposed technique in VR with the proof-of-concept implementation.

The software was written using the Unity 3D engine version 2021.3.10 and run on an Alienware 15 gaming laptop with a quad-core Intel Core i7 CPU clocked at 2.90GHz, an Nvidia GeForce GTX 1080 Max-Q GPU, and 16GB of RAM.

5 EVALUATION

We conducted a quantitative experiment to evaluate the performance and usability of our technique, where the participants were instructed to select distant targets using our proposed technique (SVMT) and the ray casting baseline (RC), and fill a qualitative survey based on NASA TLX [14].

We chose selection as the task because it is the most frequently performed task in VR and is the basis of more complex spatial tasks. We chose RC as the baseline for the performance comparison because it is one of the oldest [27] and most widely used yardsticks for measuring performance benefits of novel selection techniques in VR, not only in commercial systems, but also in research.

5.1 Participant

We recruited 12 participants (3 females, 9 males, 19-26 years old), all of whom were right-handed. All but 3 had used VR before.

5.2 Procedure

The participants used both techniques in a counterbalanced order. During a warm-up period, they could practice each technique repeatedly until they felt confident using it. They were each seated on a comfortable stool and given a 15-second break after every 10 completed tasks to prevent fatigue. After completing all tasks, they filled out a survey and were interviewed for comments. Each session lasted approximately 45 minutes, including the warm-up.

5.3 Task

To begin each task, the participants selected the start button in the form of a yellow sphere in the forward direction at eye level, ensuring that each task could be started with the same gaze direction and hand position. Immediately after selecting the start button, an actual target appeared in the form of a green sphere, the position of which was determined as a unique combination of 24 angular displacements and 3 distances (Figure 5) in a predetermined scrambled order, once for each of the two techniques. This order was identical for all participants in all sessions. The target size roughly corresponded to that of a Post-it note, whereas the target distances corresponded to those of a personal office, a small meeting room, and a large lecture hall. The participants were instructed to select the target's center as quickly and precisely as possible.

5.4 Technique

For RC, the participants held the smartphone in the dominant hand with the screen directed up, similar to a TV remote control, aimed a ray visualized as a 2mm-thick infinite line, and touched anywhere on the screen with the thumb to select (Figure 4c). For SVMT, they held the smartphone in the non-dominant hand with the screen directed toward the face, similar to taking a landscape photo, and touched the screen with the tip of the index finger of the dominant hand visualized as a 2mm-wide dot to mark a point and select (Figure 4d).

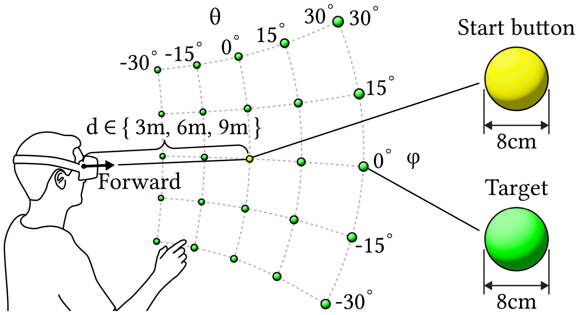

Figure 5: The selection targets, rendered as green spheres with a diameter of 8cm, appeared on a grid of spherical coordinate system centered on the participant's head position, at azimuth angles $\theta \in \{-30°, -15°, 0°, 15°, 30°\}$, elevation angles $\varphi \in \{-30°, -15°, 0°, 15°, 30°\}$, and distances $\mathrm{d} \in \{3\mathrm{m}, 6\mathrm{m}, 9\mathrm{m}\}$. The targets along the center direction ($\theta = 0°$ and $\varphi = 0°$) were excluded because each of them could be mistaken as the start button, rendered as a yellow sphere with a diameter of 8cm, which also appeared in the same direction.

5.5 Measurement

The independent variables were the technique type and the selection distance, and the dependent variables were the selection time (t) and the selection error (e). The selection time was measured as the time span between the selection of the start button and the selection of the target. For RC, the selection error was measured as the shortest distance between the surface of the target sphere and the ray. For SVMT, it was measured as the shortest distance between the surface of the target sphere and a ray originating from the mid-eye position going through the marked point on the screen.

5.6 Result

In total, 1,728 data points were collected (participants × angular displacements × target distances × technique types = 12 × 24 × 3 × 2 = 1,728). Paired t-tests between the techniques showed significant differences between the mean selection times, t_{RC} (2.79s) and $\mathrm{t}_{\mathrm{SMVT}}$ (2.20s) ($t_{863} = 11.9$, $p < 0.01$), and between the mean selection errors, e_{RC} (12.6cm) and $\mathrm{e}_{\mathrm{SMVT}}$ (8.70cm) ($t_{863} = 8.72$, $p < 0.01$).

The two-way within-subjects ANOVA showed a significant main effect of technique type ($F_{1,287} = 101$, $p < 0.01$) and target distance ($F_{2,574} = 43.3$, $p < 0.01$) on selection time. Post hoc analysis with paired t-tests between the techniques showed significant differences between the mean selection times at 3m ($t_{287} = 5.03$, $p < 0.01$), 6m ($t_{287} = 7.99$, $p < 0.01$), and 9m ($t_{287} = 7.47$, $p < 0.01$) (Figure 6).

The two-way within-subjects ANOVA showed a significant main effect of technique type ($F_{1,287} = 61.6$, $p < 0.01$) and target distance ($F_{2,574} = 328$, $p < 0.01$) on selection error. Post hoc analysis with paired t-tests between the techniques showed significant differences between the mean selection errors at 3m ($t_{287} = 4.06$, $p < 0.01$), 6m ($t_{287} = 4.65$, $p < 0.01$), and 9m ($t_{287} = 6.40$, $p < 0.01$) (Figure 7).

There were significant interaction effects of technique type × target distance on selection time ($F_{2,574} = 8.02$, $p < 0.01$) and selection error ($F_{2,574} = 13.2$, $p < 0.01$). In both, larger distances corresponded to increased differences between the marginal means of the techniques.

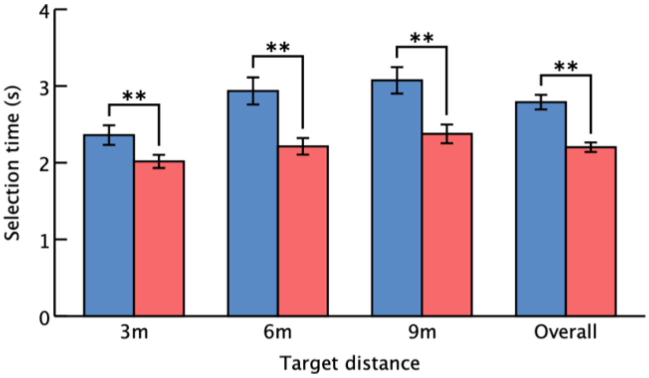

Figure 6: Selection times by target distances. Blue: ray casting (RC), red: stereoscopic viewing and monoscopic touching (SVMT), bridges: statistical significance, *: $p < 0.05$, **: $p < 0.01$, and error bars: ±2 SE.

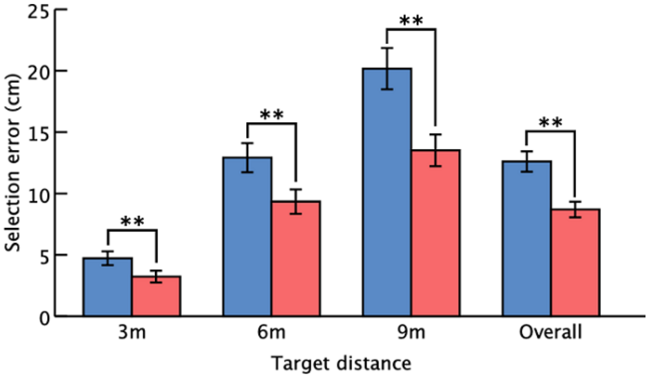

Figure 7: Selection errors by target distances. Blue: RC, red: SVMT, bridges: statistical significance, *: $p < 0.05$, **: $p < 0.01$, and error bars: ±2 SE.

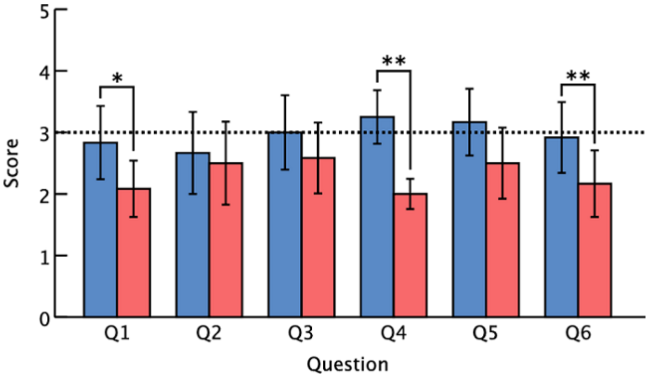

Figure 8: 5-point Likert scale scores of questions based on NASA TLX; Q1: mental demand, Q2: physical demand, Q3: temporal demand, Q4: performance, Q5: effort, and Q6: frustration. All prompts of questions except Q4: "__" is low (1) - high (5); prompt of Q4: "__" is high (1) - low (5). Dotted line: neutral (3.0, lower is better), blue: RC, red: SVMT, bridges: statistical significance, *: $p < 0.05$, **: $p < 0.01$, and error bars: ±2 SE.

5.7 Survey

On the 5-point Likert scale questions asking about different types of demand felt during selection tasks (Q1: mental demand, Q2: physical demand, Q3: temporal demand, Q4: performance, Q5: effort, and Q6: frustration), SVMT received scores below neutral on all questions (lower is better), whereas RC received scores above neutral on Q4 and Q5. The Friedman test showed a significant main effect of technique type on score ($\chi^2_{1,72} = 19.0$, $p < 0.01$). Post hoc analysis with paired t-tests between the techniques showed significant differences between the mean scores on Q1 ($t_{11} = 2.46$, $p < 0.05$), Q4 ($t_{11} = 5.75$, $p < 0.01$), and Q6 ($t_{11} = 3.45$, $p < 0.01$), where SVMT received better scores (Figure 8).

6 DISCUSSION & FUTURE WORK

In this section, we discuss the performance advantages, usability characteristics, limitations of the implementation, and applications of our proposed technique, and directions for future work based on the results of the evaluation.

Our technique is quicker and more precise. Overall, selections using SVMT were 27% quicker and 45% more precise than the widely used RC baseline. More specifically, the longer the target distance, the quicker and more precise were the selections, being 33% quicker at 6m and 49% more precise at 9m. This comes as no surprise, as the image plane interaction that allows the user to handle large spaces or distant objects as if they were a flat picture [30], and shows that the proposed technique will help support a larger VR space filled with smaller interactive objects.

Our technique is more direct and intuitive. SVMT was considered less mentally demanding (Q1), more performant (Q4), and less frustrating (Q6) than RC. When the participants were asked how they would explain the two techniques to their friends and what the pros and cons are, they likened RC to *"a TV remote"* (P1, 3, 6, 12) and *"a laser pointer"* (P2, 4, 9, 10, 11), and noted that it was *"comfortable to use with one hand"* (P3, 5, 10) because *"small hand movement resulted in a large ray movement"* (P2, 4, 12). However, others felt that *"using only one hand made it less stable"* (P1, 3) and *"prone to shake"* (P2, 3, 4, 5, 6, 7, 10, 12). As a result, many felt that *"aiming was stressful"* (P1, 5, 7, 8), as if they were *"trying to shoot down a fly with a laser pointer"* (P11).

On the other hand, they likened SVMT to *"adjusting the focal length of a camera"* (P3, 10), where *"the object initially appeared as two but became one as the finger approached it"* (P2, 3, 10), or *"bringing the faraway object to the screen"* (P4, 9, 11, 12) and then *"directly touching it"* (P2, 3, 4, 7, 9, 10, 11, 12). Some complained that *"using two hands required more effort"* (P3, 4, 9, 10), *"focusing on the object was difficult"* (P6), and *"the finger blocked the view"* (P5). However, others felt that *"using two hands made it more stable"* (P3, 8), and they *"became used to it soon"* (P2, 10, 11, 12) and could perform selections *"quickly"* (P1), *"precisely"* (P3, 5, 6, 7, 9, 10), and *"intuitively"* (P2). As a result, some felt that it was *"similar to a smartphone AR game"* (P9) or even *"the Fruit Ninja game"* (P8).

Our technique will become even better. Because the position and orientation of the panel were calculated from the positions of the finger joints that held it, participants could hold the panel in only two specific postures. However, all participants with varying hand sizes could comfortably hold the panel in the two postures

(Figure 4a, b), so we did not encounter any calibration issues. Moreover, the Quest 2's hand tracking could adapt to varying hand sizes, ensuring the panel pose could be calculated accurately. No participant mentioned discomfort when holding and using the panel.

Another limitation of our implementation was the need to infer touches based on the sound they produced, which resulted in instances of both false negative (the produced sound failed to trigger) and false positive (outside noises triggered) activations. In those instances, the single affected selection was repeated from the start. These instances were rare (approximately 1 out of 20 times), remedied, and appeared in both techniques, meaning the performance comparison between the techniques was unaffected.

Recent developments include optically tracking peripheral devices such as keyboards from VR headsets [26], so it is likely that actual smartphones will be tracked in VR in the near future, enabling arbitrary holding postures and accurate touch registrations.

Our technique can be versatile. We created three VR applications that show that the proposed technique can work seamlessly with the 2D interface displayed on the smartphone screen (Figure 9) and enables quicker and more precise selection of challenging targets, such as a moving target (Figure 10), or multiple targets that appear close together (Figure 11).

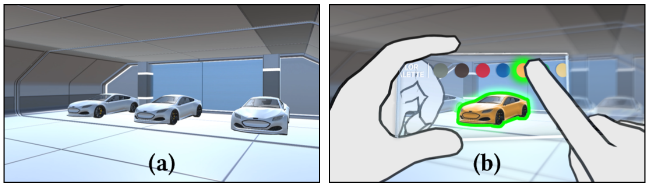

Figure 9: A VR application depicting seamless interoperation with on-screen 2D interface. (a) In a VR car design studio, (b) the user can select one of the concept cars and change the body color with the 2D color palette overlay.

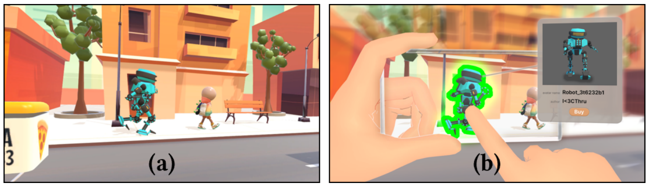

Figure 10: A VR application depicting selection of a moving target. (a) When walking down a street in a metaverse city, (b) the user can obtain information about an attractive robot avatar across the street before it walks away.

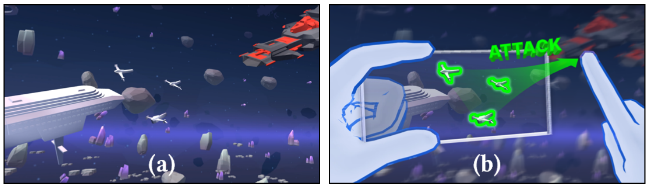

Figure 11: A VR application depicting selection of multiple targets. (a) In a VR space battle game, (b) the user can command multiple spaceships that appear together within the frame of the smartphone with quick, consecutive taps.

For future work, while most participants completed the 72 selections without reporting any eye strains, ergonomics issues regarding prolonged, repeated tasks, vergence-accommodation conflict, and the quantitative impact of animations on reducing visual fatigue deserve further investigation. It would also be interesting and useful to model the selection performances of the proposed technique with Fitts' Law, as it has both 2D [7] and 3D [12] characteristics.

In addition, the technique could be extended not only to mobile devices of various sizes such as portable iPads and bigger Wacom tablets, but also to a large wall, as long as there is a physical surface to touch and a way to register the touch [25]. Finally, the technique could be extended to pen input to support usage scenarios such as 3D sketching [9] for quick and precise spatial pen drawing with the help of passive haptic feedback from the screen surface.

7 CONCLUSION

In this study, we focused on the thin, flat and rigid form of a smartphone and reimagined it as a portable glass panel that can be used to select distant objects in VR by directly touching them while looking at them through the device. The binocular parallax that occurs in such an arrangement was successfully overcome through the introduction of two modes, stereoscopic viewing and monoscopic touching, and a smoothly animated transition between them.

For proof of concept, we used a transparent acrylic panel in the shape and size of a regular smartphone to enable unoccluded tracking of finger poses from the VR headset, which led to a minimal yet precise implementation of the proposed technique without the need for additional tracking devices attached to the smartphone. Under controlled conditions, our novel selection technique was up to 33% quicker and 49% more precise than the ray casting baseline, thanks to the intuitiveness of touching the image plane directly.

Our approach opens the possibility of direct, touch-like techniques that can be used for remote spatial interactions. We expect that the use of the intuitive transparency metaphor will further enrich the spatial interaction vocabulary and help create an ecosystem of mobile devices and immersive VR headsets in the near future.

ACKNOWLEDGMENTS

This research was supported by the KAIST Convergence Research Institute Operation Program.

REFERENCES

[1] Ferran Argelaguet and Carlos Andujar. 2009. Visual feedback techniques for virtual pointing on stereoscopic displays. In *Proc. VRST '09*. 163–170. https://doi.org/10.1145/1643928.1643966
[2] Rahul Arora, Rubaiat Habib Kazi, Tovi Grossman, George Fitzmaurice, and Karan Singh. 2018. SymbiosisSketch: combining 2D & 3D sketching for designing detailed 3D objects in situ. In *Proc. CHI '18*. Article 185, 15 pages. https://doi.org/10.1145/3173574.3173759
[3] Domagoj Baričević, Cha Lee, Matthew Turk, Tobias Höllerer, and Doug A. Bowman. 2012. A hand-held AR magic lens with user-perspective rendering. In *Proc. ISMAR '12*. 197–206. https://doi.org/10.1109/ISMAR.2012.6402557
[4] Patrick Baudisch, Nathaniel Good, and Paul Stewart. 2001. Focus plus context screens: combining display technology with visualization techniques. In *Proc. UIST '01*. 31–40. https://doi.org/10.1145/502348.502354
[5] Verena Biener, Travis Gesslein, Daniel Schneider, Felix Kawala, Alexander Otte, Per Ola Kristensson, Michel Pahud, Eyal Ofek, Cuauhtli Campos, Matjaž Kljun, Klen Čopič Pucihar, and Jens Grubert. 2022. PoVRPoint: authoring presentations in mobile virtual reality. *TVCG* 28, 5 (May 2022), 2069–2079. https://doi.org/10.1109/TVCG.2022.3150474

[6] Verena Biener, Daniel Schneider, Travis Gesslein, Alexander Otte, Bastian Kuth, Per Ola Kristensson, Eyal Ofek, Michel Pahud, and Jens Grubert. 2020. Breaking the screen: interaction across touchscreen boundaries in virtual reality for mobile knowledge workers. *TVCG* 26, 12 (Dec 2020), 3490–3502. https://doi.org/10.1109/TVCG.2020.3023567

[7] Xiang Cao, Jacky Jie Li, and Ravin Balakrishnan. 2008. Peephole pointing: modeling acquisition of dynamically revealed targets. In *Proc. CHI '08*. 1699–1708. https://doi.org/10.1145/1357054.1357320

[8] Rajkumar Darbar, Arnaud Prouzeau, Joan Odicio-Vilchez, Thibault Lainé, and Martin Hachet. 2021. Exploring smartphone-enabled text selection in AR-HMD. In *Proc. GI '21*. 117–126. https://doi.org/10.20380/GI2021.14

[9] Tobias Drey, Jan Gugenheimer, Julian Karlbauer, Maximilian Milo, and Enrico Rukzio. 2020. VRSketchIn: exploring the design space of pen and tablet interaction for 3D sketching in virtual reality. In *Proc. CHI '20*. Article 501, 14 pages. https://doi.org/10.1145/3313831.3376628

[10] Danilo Gasques, Janet G. Johnson, Tommy Sharkey, and Nadir Weibel. 2019. What you sketch is what you get: quick and easy augmented reality prototyping with PintAR. In *CHI '19 Extended Abstracts*. Article LBW1416, 6 pages. https://doi.org/10.1145/3290607.3312847

[11] Travis Gesslein, Verena Biener, Philipp Gagel, Daniel Schneider, Per Ola Kristensson, Eyal Ofek, Michel Pahud, and Jens Grubert. 2020. Pen-based interaction with spreadsheets in mobile virtual reality. In *Proc. ISMAR '20*. 361–373. https://doi.org/10.1109/ISMAR50242.2020.00063

[12] Tovi Grossman and Ravin Balakrishnan. 2004. Pointing at trivariate targets in 3D environments. In *Proc. CHI '04*. 447–454. https://doi.org/10.1145/985692.985749

[13] Jens Grubert, Matthias Heinisch, Aaron Quigley, and Dieter Schmalstieg. 2015. MultiFi: multi fidelity interaction with displays on and around the body. In *Proc. CHI '15*. 3933–3942. https://doi.org/10.1145/2702123.2702331

[14] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (task load index): results of empirical and theoretical research. Adv. Psychol., Vol. 52. 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

[15] Sebastian Hubenschmid, Johannes Zagermann, Simon Butscher, and Harald Reiterer. 2021. STREAM: exploring the combination of spatially-aware tablets with augmented reality head-mounted displays for immersive analytics. In *Proc. CHI '21*. Article 469, 14 pages. https://doi.org/10.1145/3411764.3445298

[16] Ernst Kruijff, J. Edward Swan, and Steven Feiner. 2010. Perceptual issues in augmented reality revisited. In *Proc. ISMAR '10*. 3–12. https://doi.org/10.1109/ISMAR.2010.5643530

[17] Ricardo Langner, Marc Satkowski, Wolfgang Büschel, and Raimund Dachselt. 2021. MARVIS: combining mobile devices and augmented reality for visual data analysis. In *Proc. CHI '21*. Article 468, 17 pages. https://doi.org/10.1145/3411764.3445593

[18] Khanh-Duy Le, Tanh Quang Tran, Karol Chlasta, Krzysztof Krejtz, Morten Fjeld, and Andreas Kunz. 2021. VXSlate: exploring combination of head movements and mobile touch for large virtual display interaction. In *Proc. DIS '21*. 283–297. https://doi.org/10.1145/3461778.3462076

[19] Joon Hyub Lee and Seok-Hyung Bae. 2013. Binocular cursor: enabling selection on transparent displays troubled by binocular parallax. In *Proc. CHI '13*. 3169–3172. https://doi.org/10.1145/2470654.2466433

[20] Joon Hyub Lee, Seok-Hyung Bae, Jinyung Jung, and Hayan Choi. 2012. Transparent display interaction without binocular parallax. In *UIST '12 Adjunct*. 97–98. https://doi.org/10.1145/2380296.2380340

[21] Joon Hyub Lee, Donghyeok Ma, Haena Cho, and Seok-Hyung Bae. 2021. Post-Post-It: a spatial ideation system in VR for overcoming limitations of physical Post-it notes. In *CHI '21 Extended Abstracts*. Article 300, 7 pages. https://doi.org/10.1145/3411763.3451786

[22] Weizhou Luo, Eva Goebel, Patrick Reipschläger, Mats Ole Ellenberg, and Raimund Dachselt. 2021. Exploring and slicing volumetric medical data in augmented reality using a spatially-aware mobile device. In *ISMAR '21 Adjunct*. 334–339. https://doi.org/10.1109/ISMAR-Adjunct54149.2021.00076

[23] Akhmajon Makhsadov, Donald Degraen, André Zenner, Felix Kosmalla, Kamila Mushkina, and Antonio Krüger. 2022. VRySmart: a framework for embedding smart devices in virtual reality. In *CHI '22 Extended Abstracts*. Article 358, 8 pages. https://doi.org/10.1145/3491101.3519717

[24] Fabrice Matulic, Aditya Ganeshan, Hiroshi Fujiwara, and Daniel Vogel. 2021. Phonetroller: visual representations of fingers for precise touch input with mobile phones in VR. In *Proc. CHI '21*. Article 129, 13 pages. https://doi.org/10.1145/3411764.3445583

[25] Manuel Meier, Paul Streli, Andreas Fender, and Christian Holz. 2021. TapID: rapid touch interaction in virtual reality using wearable sensing. In *Proc. VR '21*. 519–528. https://doi.org/10.1109/VR50410.2021.00076

[26] Meta. 2020. *Infinite Office*. Retrieved Sep 1, 2022 from https://youtu.be/5_bVkbG1ZCo

[27] Mark R. Mine. 1995. *Virtual environment interaction techniques*. Technical Report TR95-018. UNC Chapel Hill CS Dept. 1–18 pages.

[28] Roberto A. Montano-Murillo, Cuong Nguyen, Rubaiat Habib Kazi, Sriram Subramanian, Stephen DiVerdi, and Diego Martinez-Plasencia. 2020. Slicing-volume: hybrid 3D/2D multi-target selection technique for dense virtual environments. In *Proc. VR '20*. 53–62. https://doi.org/10.1109/VR46266.2020.00023

[29] Erwan Normand and Michael J. McGuffin. 2018. Enlarging a smartphone with AR to create a handheld VESAD (virtually extended screen-aligned display). In *Proc. ISMAR '18*. 123–133. https://doi.org/10.1109/ISMAR.2018.00043

[30] Jeffrey S. Pierce, Andrew S. Forsberg, Matthew J. Conway, Seung Hong, Robert C. Zeleznik, and Mark R. Mine. 1997. Image plane interaction techniques in 3D immersive environments. In *Proc. I3D '97*. 39–ff. https://doi.org/10.1145/253284.253303

[31] Carolin Reichherzer, Jack Fraser, Damien Constantine Rompapas, and Mark Billinghurst. 2021. SecondSight: a framework for cross-device augmented reality interfaces. In *CHI '21 Extended Abstracts*. Article 234, 6 pages. https://doi.org/10.1145/3411763.3451839

[32] Houssem Saidi, Emmanuel Dubois, and Marcos Serrano. 2021. HoloBar: rapid command execution for head-worn AR exploiting around the field-of-view interaction. In *Proc. CHI '21*. Article 745, 17 pages. https://doi.org/10.1145/3411764.3445255

[33] Dieter Schmalstieg, L. Miguel Encarnação, and Zsolt Szalavári. 1999. Using transparent props for interaction with the virtual table. In *Proc. I3D '99*. 147–153. https://doi.org/10.1145/300523.300542

[34] Martin Spindler, Wolfgang Büschel, and Raimund Dachselt. 2012. Use your head: tangible windows for 3D information spaces in a tabletop environment. In *Proc. ITS '12*. 245–254. https://doi.org/10.1145/2396636.2396674

[35] Hemant Bhaskar Surale, Aakar Gupta, Mark Hancock, and Daniel Vogel. 2019. TabletInVR: exploring the design space for using a multi-touch tablet in virtual reality. In *Proc. CHI '19*. Article 13, 13 pages. https://doi.org/10.1145/3290605.3300243

[36] Arda Ege Unlu and Robert Xiao. 2021. PAIR: phone as an augmented immersive reality controller. In *Proc. VRST '21*. Article 27, 6 pages. https://doi.org/10.1145/3489849.3489878

[37] Klen Čopič Pucihar, Paul Coulton, and Jason Alexander. 2014. The use of surrounding visual context in handheld AR: device vs. user perspective rendering. In *Proc. CHI '14*. 197–206. https://doi.org/10.1145/2556288.2557125

[38] Shengzhi Wu, Daragh Byrne, and Molly Wright Steenson. 2020. "Megereality": leveraging physical affordances for multi-device gestural interaction in augmented reality. In *CHI '20 Extended Abstracts*. Article INT008, 4 pages. https://doi.org/10.1145/3334480.3383170

[39] Fengyuan Zhu and Tovi Grossman. 2020. BISHARE: exploring bidirectional interactions between smartphones and head-mounted augmented reality. In *Proc. CHI '20*. Article 106, 14 pages. https://doi.org/10.1145/3313831.3376233