Disambiguation Techniques for Freehand Object Manipulations in Virtual Reality

Di Laura Chen* Ravin Balakrishnan* Tovi Grossman*
University of Toronto, Toronto, ON, Canada

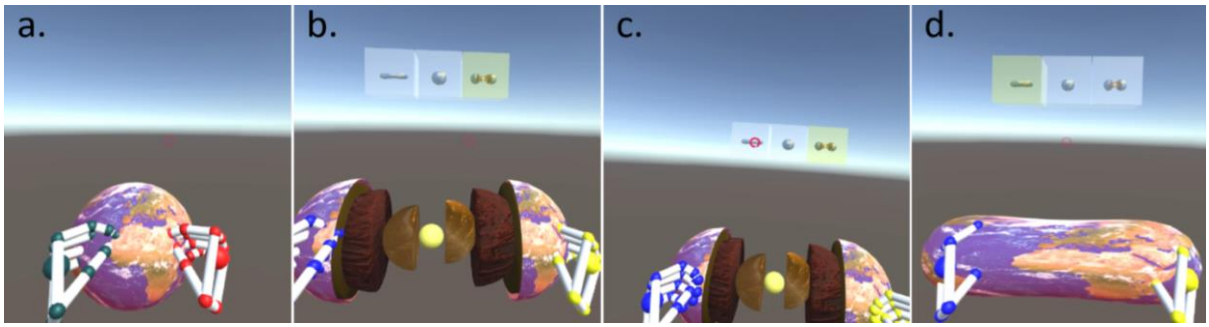

Figure 1. Our system for resolving ambiguities of freehand gestures in VR. In this example, we show how head gaze can be used to disambiguate a freehand manipulation during the gesture. a) The user begins a freehand gesture. b) Ambiguity in the intended operation is detected, and a menu showing live previews of operation alternatives is shown. c) Without stopping the gesture, the user directs their head gaze at the desired operation. d) The user continues the gesture with the desired operation.

ABSTRACT

Manipulating virtual objects using bare hands has been an attractive interaction paradigm in virtual and augmented reality due to its intuitive nature. However, one limitation of freehand input lies in the ambiguous resulting effect of the interaction. The same gesture performed on a virtual object could invoke different operations on the object depending on the context, object properties, and user intention. We present an experimental analysis of a set of disambiguation techniques in a virtual reality environment, comparing three input modalities (head gaze, speech, and foot tap) paired with three different timings in which options appear to resolve ambiguity (before, during, and after an interaction). The results indicate that using *head gaze* for disambiguation *during* an interaction with the object achieved the best performance.

Keywords: Freehand gestures, VR, uncertainty, ambiguity.

Index Terms: Human-centered computing ~ Human computer interaction (HCI) ~ Interaction techniques ~ Gestural input

1 INTRODUCTION

Freehand interaction allows users to manipulate virtual objects with their bare hands, similar to manipulating real-world objects, through physically realistic behaviours [10]. The concept of freehand interaction has been explored in various domains, both in academia [2], [8], [10], [19], [27], [49] and industry [14], [41]. In particular, it is a promising interaction model in virtual reality (VR) and augmented reality (AR) because it is familiar to users and takes advantage of the 3D space. For example, the HoloLens 2 [41] introduced *instinctual interaction*, allowing users to touch, grasp, and move holograms as if they were real objects.

One significant challenge of freehand input is handling the ambiguous resulting effect of the interaction. Freehand input allows users to naturally interact with a virtual object by leveraging their prior experiences with real-world objects and the expressiveness of their hands. However, in real life, often a single gesture can be used to perform multiple tasks. Similarly, in VR and AR, the same gesture could activate different operations on a virtual object depending on the context, object properties, and user intention. For instance, in a shape modeling application, when a user sees a virtual rod, perhaps the user's first natural reaction would be to grab and bend the rod to see what will happen. Since virtual objects do not have to follow the rules of physics, a virtual rod could behave as if it were both brittle and elastic. Thus, performing the same bending gesture on the rod could lead to curving, breaking, or creasing the rod, resulting in operation ambiguity.

Gestural operation ambiguity is common in user-elicited gestures [5], [38], [54]. For example, Arora et al. [5] have explored using freehand gestures for authoring VR animations and have observed that users would sometimes specify the same gestures for different animation effects. One straightforward solution to eliminate ambiguity would be to use a pre-defined mapping that assigns each gestural manipulation to a unique operation [16]. However, this method undermines the natural properties of freehand interactions, as users are forced to remember the gestural commands, and the vast possibilities of operations makes the method unscalable. In prior literature, techniques for interacting under uncertainty have been explored [43], but such techniques have not been adapted to VR or freehand interaction.

In this paper, we investigate a set of techniques for disambiguating the effect of freehand manipulations in VR, and present empirical evidence for the performance of these techniques. Grounded by previous research, e.g. [9], [11], [12], [21], [33], [50], we identified two factors that may be crucial for resolving ambiguity: modality and timing. We compared three input modalities (head gaze, speech, and foot tap) paired with three timings (before, during, and after an interaction) in which options become available to resolve ambiguity. We tested these disambiguation techniques on two sets of operations: operations that are highly distinguishable from each other and operations that look very similar to each other during a manipulation. The results indicated that using *head gaze* for disambiguation *during* object

* {chendi, ravin, tovi}@dgp.toronto.edu

2642-5254/20/$31.00 ©2020 IEEE
DOI 10.1109/VR46266.2020.00048

manipulation achieved the best performance (Figure 1). The insights derived from our results can help to inform the design of freehand spatial interactions with ambiguous outcomes.

2 Related Work

Our work is guided by prior research in three areas of HCI: ambiguity resolution techniques; input modalities for VR; and feedforward and suggestive user interfaces.

2.1 Ambiguity Resolution Techniques

Prior work has developed techniques to resolve input ambiguity in various areas of research. Mankoff et al. [31], [32] focused on resolving uncertainty in recognition-based interfaces through mediators. Schwarz et al. [43] established a framework for handling and dispatching uncertain input in a lazy fashion. Work has also been done on interpreting ambiguous input and giving systematic feedback to reflect the ambiguity [44]. Kaiser et al. [23] implemented an architecture that disambiguates actions in a system where uncertainty arises from multiple input sources (speech, gesture, and the environment) working synergistically. Hincapié-Ramos et al. [20] introduced a raycasting technique for AR head-mounted displays (HMD) and explored disambiguation among possible targets in the ray's path. This body of research addresses the problem of ambiguous *input*, whereas in our present work, the ambiguity lies in the intended outcome of performing identical gestures.

Also related to our work are gesture elicitation studies, which allow users to assign gestures to different tasks, which can lead to ambiguities [5], [38], [54]. To resolve a gesture ambiguity, researchers have suggested assigning the gesture to the task with the higher consensus score among users [38], [54], separating the tasks into unique selection spaces [38], applying mode-switching techniques [5], or relying on the context in which the gesture was performed [54].

2.2 VR Input Modalities

Prior research has shown evidence that using hands for interactions in VR results in a higher level of embodiment than using controllers [3]. Freehand manipulation of virtual objects is intuitive and requires little memorization, making it ideal for creating a convincing spatial interaction experience [13], [24], [52].

However, during freehand manipulations, the user's hands may not be available to perform any form of manual disambiguation procedure. As such, other modalities of disambiguation are required. Prior work has developed numerous modalities that may be relevant when the hands are occupied.

Speech is arguably one of the most common modalities to interact with user interfaces when the hands are occupied. Researchers have explored using a combination of speech and gesture for interacting with virtual content and graphical elements [11], [22], [34], [37]. In an AR multimodal interface using speech and paddle gestures, Irawati et al. [22] concluded that speech is useful for system control while gestures are suitable for directly manipulating the virtual objects.

Gaze is another common modality used in combination with gesture or touch for interaction. Pfeuffer et al. [36] experimented with gaze and pinch interactions in VR, allowing indirect freehand manipulations of both near and far objects that the user's gaze falls upon. Chatterjee et al. [12] found that combining gaze and gesture for common digital tasks can outperform interactions using gaze or gesture alone. Simeone et al. [46] expanded two-finger touch interactions with a third gaze input channel. Others have examined using head gaze, eye gaze, or a combination of head and eye movements to enhance interactions [7], [25], [48] or share awareness cues for collaborative tasks [39].

Foot-based interaction techniques have also been widely explored. Müller et al. [33] compared direct and indirect foot-tap interactions in HMDs, and found that direct interfaces are suitable for short-term and fine-grained actions while indirect input is suitable for interactions that require less accuracy. This corresponds to findings by Pakkanen and Raisamo [35] that foot-based interactions are appropriate for non-accurate spatial tasks. Foot-based input has also been used as an input modality for 3D interaction tasks such as navigation, selection, manipulation, and system control [47], [58], and has been used together with hand gestures for interactions [28], [29].

Previous research have proposed other input modalities, such as using bare hands for mode-switching in VR [49] and on-body interfaces that require touching various parts of the body, such as skin [17], [53] and face [45], [56]. However, for ambiguity resolution, we only considered modalities that do not involve hands, since the hands are already being used for the object manipulations.

2.3 Feedforward and Suggestive User Interfaces

To resolve ambiguity, a system may need to suggest alternative options to the user. This concept is related to prior HCI work that provides feedforward or suggestive user interfaces.

For example, Side Views [50] introduced a user interface mechanism that provides on-demand previews of commands through pop-up windows. Lafreniere et al. [26] designed software command disambiguation techniques that provide support before, during, and after execution of an incorrect command. Medusa [4], a proximity-aware multi-touch tabletop system, displayed a freehand gesture preview guide when the user's hand hovered over the proximity sensors. ShadowGuides [15] provided a similar gesture preview menu for registration poses. Xu et al. [55] resolved ambiguity of user input in layout beautification by presenting a preview of the beautification with inferred constraints.

Research on sketch-based drawing interfaces have looked at providing immediate, continuous feedback during freehand drawing to morph raw input strokes into ideal geometric shapes [1], [6]. Similarly, OctoPocus [9] dynamically guides the user in drawing command gestures, giving real-time feedback showing the path that is already drawn and feedforward indicating the path to follow. The framework proposed by Schwarz et al. [43] for handling uncertain input employs *lazy interpretation*, allowing interactors to provide feedback about multiple possible ambiguous inputs and delaying ambiguity resolution until necessary.

Suggestive interfaces [21], [51] are another promising category of work which could be applied to ambiguity resolution. In a 3D drawing task [21], the user draws simple geometric elements, thereby giving hints about a desired operation. The system then suggests possible operations as thumbnails that the user can select. Other suggestion-based interfaces include VirtualGrasp [57] and SceneSuggest [42], where the system gives suggestions about object candidates after the user provides hints to the system about the user's intention.

This set of prior work indicates that the timing at which ambiguity is resolved and the types of feedback provided to users could be important factors for disambiguation. Previews are generally shown *before* performing an operation; real-time continuous feedback and feedforward are delivered *during* an operation; and suggestions for possible alternatives are usually given *after* an operation. Guided by this prior research, we explore the possibility of disambiguating freehand input before, during, and after interaction with a virtual object.

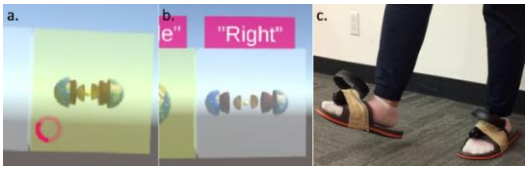

Figure 2. Input modalities for disambiguation. a) Head gaze. b) Speech. c) Foot.

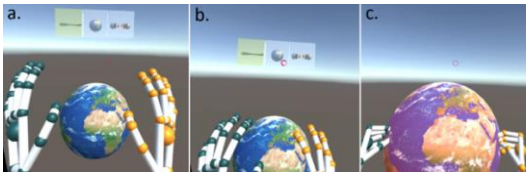

Figure 3. Disambiguating before a gesture. a) The disambiguation menu appears when the user's hands approach. b) The user selects the desired operation. c) The user performs the gesture.

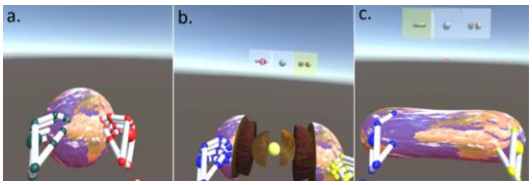

Figure 4. Disambiguating during a gesture. a) The user begins an ambiguous gesture. b) The disambiguation menu is displayed with live previews. c) The user selects the desired operation and continues the gesture.

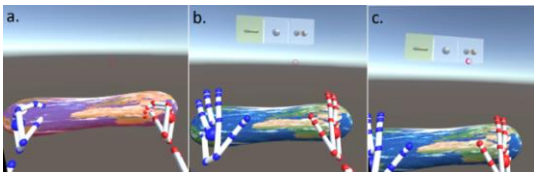

Figure 5. Disambiguating after a gesture. a) The user completed an ambiguous gesture. b) When the grasp is released, the disambiguation menu is displayed with alternative operations. c) The user selects the desired operation.

3 Disambiguation System and Techniques

Our system is developed within a VR environment which supports freehand manipulation of 3D objects. Grounded by our review of related literature, we have identified that modality and timing are two key considerations for how disambiguation should be carried out. In this section, we describe our system and disambiguation techniques.

3.1 VR Disambiguation Menu

Within our developed system, each technique utilizes a VR disambiguation menu (Figure 1b). When a gesture could result in an ambiguous operation, a pop-up menu appears above the object being manipulated. The pop-up menu consists of preview cubes, each containing a 3D preview of the operation alternatives. In our descriptions below, we considered operations that would have three possible alternatives, and as such, the menu would consist of three preview cubes. The timing of when the menu appears, and the modality used to select from the menu, varies based on the techniques, as described below.

3.2 Input Modalities

We explored three promising input modalities for hands-free operation in VR: head gaze, speech, and foot (Figure 2).

For *head gaze-based* disambiguation, a circular cursor is rendered in VR and follows the user's head direction. When the cursor falls on a preview cube, the cursor starts to fill up like a circular progress bar (Figure 2a). If the cursor dwells on a preview cube for more than 500 ms, the associated operation is selected. The 500 ms dwell time was shown in previous research to achieve the best performance for dwell-based techniques [30].

For *speech-based* disambiguation, users vocalize their desired option from the menu. Labels are placed above each preview cube, showing the speech command that the user must say in order to select the corresponding cube (Figure 2b). The speech command could be the name of the operation (e.g. "twist", "bend") or could be a logical label (e.g. "left", "A", "one").

For *foot-based* disambiguation, the ground in front of the user's feet is divided into three rectangular selection areas, corresponding to the three preview cubes. When the user places the right foot in one selection area, the complementary preview cube is highlighted in transparent green. The user can select a preview cube by tapping the foot (Figure 2c) in the corresponding selection area. Based on prior design recommendations [33], we chose an indirect foot-tap interface and divided the areas into columns rather than rows. Our algorithm used relative positions of the left and right foot to determine the selection area. When the right foot steps forward, the green highlight becomes active.

3.3 Timing of Disambiguation

We also explored three possibilities for when the disambiguation menu appears: before, during, and after a gesture.

Disambiguating *before* a gesture: when the user's hands hover near an object that can be manipulated in multiple ambiguous ways, the menu is displayed (Figure 3). The preview cubes show a static preview of the effect that each manipulation would have on the object. The user can select an option before interaction begins. The menu disappears when the manipulation begins, or if the hands move away from the object.

Disambiguating *during* a gesture: the menu is shown after the user has grasped an object with their hands and starts the gesture (Figure 4). As the user performs the gesture, the preview cubes are continuously updated to allow the user to see what each alternative operation would look like in real-time. While the gesture continues, the user can make selections from the menu, and the effect on the object will change accordingly. The menu stays visible, and alternative choices can be made, until the operation is completed.

Disambiguating *after* a gesture: the menu becomes visible after the user has completed a manipulation and has released the object (Figure 5). The preview cubes display static results of alternative operations, from which the user can make a selection. If the user grasps the object again or the hands move away from the object, the menu disappears.

3.4 Apparatus and System Implementation

We used an Oculus Rift as the VR HMD device. For hand tracking, we mounted a Leap Motion controller at the center of the HMD. The system was written in Unity 2018.4.3f1 and was run on a Windows 10 machine with NVIDIA GeForce GTX 1070 GPU and 2.90 GHz Intel Core i7-7820HK CPU. Grasp and release motions were automatically recognized by the Leap Motion controller, while changes in hand movements were calculated by comparing palm positions between two frames. For speech recognition, we used Unity's Windows.Speech API, which listens for voice input

and attempts to match the phrases to a list of registered keywords. For the foot modality, we strapped two Oculus Touch controllers onto a pair of slippers to track the positions of the feet.

4 EXPERIMENT

The goal of our study is to empirically evaluate the performance of the disambiguation techniques, composed of the *head gaze* (*h-gaze* for short), *speech*, and *foot* modalities paired with the *before*, *during*, and *after* timings. We designed an object matching task to measure the task completion time (TCT) and error rate of each technique. In our pilot study, it seemed that another important factor which determined the usability of the techniques was how visually distinguishable the set of alternatives were – selecting the appropriate alternative was more difficult when the operations were similar. Thus, in the experiment, we also included a condition of distinguishability.

4.1 Participants

We recruited 12 participants (6 female, 1 left-handed), 18 to 40 years in age ($\mu = 27.33$, $\sigma = 6.95$), through social networks and mailing lists. Nine participants had prior experience with using a VR device, although four out of the nine participants were not sure about the kind of VR device that they have tried. Compensation for each participant was $25.

4.2 Procedure

4.2.1 Experiment Task and Operations

The experiment consisted of a 3D object matching task in an HMD VR environment. A trial started with a 3D globe, 202 mm in diameter, displayed in the virtual environment in front of the user. The user could interact with the globe by using two hands to grasp it and then pulling their hands away from each other or pushing them back together. When the user performed the pulling or pushing gesture, there would be three possible resulting operations performed on the globe. These operations represented the output ambiguity that needed to be resolved. Experimenting with three options served as a basic testbed for our disambiguation techniques. During a trial, a semi-transparent overlay was displayed, centered at the same fixed location as the globe. The overlay showed the required outcome after applying one of the possible effects on the globe (Figure 6). The objective of the task was to pull or push the globe until it matched the overlay.

We designed two sets of operation alternatives with either high distinguishability (*HD*) or low distinguishability (*LD*). For *HD*, the operations were all highly distinct from one another, consisting of: stretching the globe; uniformly scaling the globe; and, pulling the globe apart to reveal the Earth's inner layers (Figure 7). For *LD*, the operations all looked similar to one another. All the operations would stretch the globe, but at different thickness levels (Figure 8).

When the disambiguation menu was triggered, the three alternative operations were shown in preview cubes. Prior to a gesture beginning, one of the three operations was already selected by default. If the default operation did not match the operation needed for a trial, the user would need to select the appropriate operation from the preview cubes and manipulate the globe to match the overlay. We defined matching as when the manipulated globe was within ±10 mm of the overlay margin. The currently selected preview cube was colored transparent yellow, while the unselected cubes were transparent white. The globe emitted a glowing red color when it was grasped with two hands and emitted a green color when it matched the overlay.

In the pilot study, we tried different combinations of speech commands, such as "one", "two", "three", or "stretch", "scale", "open". We found the commands "left", "middle", "right" to be the most easily recognized by the speech engine and used this set of commands in the study.

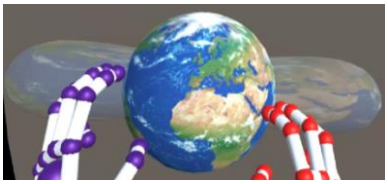

Figure 6. A semi-transparent overlay showed the required operation.

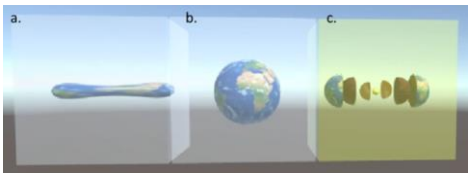

Figure 7. HD operations. a) Stretch. b) Uniform scale. c) Pull apart.

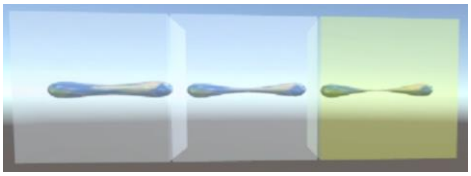

Figure 8. LD operations stretch the globe at different thicknesses.

4.2.2 Study Setup

For each participant, we explained the purpose of the study and the procedure involved. The participant filled out a pre-study questionnaire. Next, we asked the participant to adjust the HMD until it fit comfortably on the head. All experiments were conducted with the participant in a standing position. Before starting each block, we asked the participant to wear the VR headset and look straight ahead. The system calibrated the position of the objects in the scene according to the participant's head position and placed the globe at a comfortable distance in front of the participant's chest. We verified through the pilot study that this positioning allowed participants to easily manipulate the globe using two hands for long periods of time. Before the start of each trial, the user's hands were in a resting position, with the hands and arms naturally resting along the side of the body. The system displayed a 2-second countdown to start a trial. We asked the participants to complete the task as fast and as accurately as possible. Short breaks could be taken between blocks to avoid fatigue.

After every three techniques (grouped by modality), the participant answered a Raw NASA-TLX (RTLX) [18] questionnaire focusing on the modality. When all techniques were completed, the participant filled out a longer questionnaire regarding their experiences. Finally, we conducted a brief interview to obtain further insights. The entire study took approximately 90 minutes.

4.3 Experimental Design

The experiment employed a within-subjects design, where the primary factors were the disambiguation *modality* (*h-gaze, speech, foot*) and *timing* (*before, during, after*). For each combination, we conducted four blocks of trials, with two for each level of *distinguishability* (*HD, LD*). Each block consisted of six trials of matching tasks. The order of the *HD* and *LD* blocks were randomized. We varied the size of the overlay to be 30%, 60%, and

90% of the effect's maximum size. The goal operation in each trial was randomized. Half of the trials used the goal operation as the default operation, meaning the user would not need to switch operations. Thus, in each block of six trials, three trials started with matching operations at 30%, 60%, and 90% overlay sizes, while the other three trials started with non-matching operations at 30%, 60%, and 90% overlay sizes. We randomized the order of these trials within each block. In total, each participant performed 3 modalities × 3 timings × 4 blocks × 6 trials = 216 matching trials, and we collected 216 × 12 participants = 2592 trials overall. The ordering of modality and timing were counterbalanced across participants. Before testing each technique, we gave the participants time to become familiar with the technique, up to a maximum of two blocks.

4.3.1 Dependent Variables

For each trial, we measured the *TCT,* starting from when the globe and the overlay appeared, and ending when the globe's shape matched the overlay and the participant has released the grasp. We counted two types of errors in each trial. A *selection error* was counted if the participant selected an option that did not match the overlay. A *manipulation error* occurred if the participant released the grasp on the globe before the match was completed. For both types of errors, we counted the number of times each error occurred in a trial.

5 Results

We calculated the mean TCTs for each combination of factors (*modality* × *timing* × *distinguishability*) and removed outliers that were more than 3 standard deviations away from the mean. During the experiments, we also skipped seven trials due to technical problems. In total, we removed 1.9% of the completed trials. We analyzed the remaining collected data using a three-way repeated measures ANOVA and performed Tukey's HSD test to find significant pairwise differences at the $p < .05$ level.

5.1 Task Completion Time

The analysis revealed that all of the factors, *modality* ($F_{2,22} = 8.9$, $p < .0005$), *timing* ($F_{2,22} = 14.2$, $p < .0001$) and *distinguishability* ($F_{1,11} = 36.6$, $p < .0001$), had a significant effect on TCT (Figure 9). Furthermore, we found a significant interaction effect between *timing* and *distinguishability* ($F_{2,22} = 8.7$, $p < .0005$). There were no significant interaction effects between other combinations of factors. Post-hoc tests reported that *h-gaze* ($\mu = 6.32$ s) and *speech* ($\mu = 6.40$ s) were significantly faster than *foot* ($\mu = 7.76$s), while *during* ($\mu = 5.93$ s) and *after* ($\mu = 6.63$ s) were significantly faster than *before* ($\mu = 7.94$ s). Additionally, the TCT of *HD* tasks ($\mu = 5.89$ s) was significantly lower than that of the *LD* tasks ($\mu = 7.77$ s). We did not find significant differences between *h-gaze* and *speech*, or between *during* and *after*.

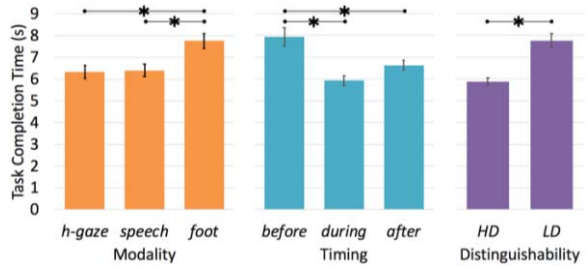

Figure 9. Effect of *modality*, *timing*, and *distinguishability* on TCT. Asterisks indicate significantly different pairs.

5.2 Selection Errors

Modality ($F_{2,22} = 5.9$, $p < .005$), *timing* ($F_{2,22} = 5.7$, $p < .005$) and *distinguishability* ($F_{1,11} = 51.2$, $p < .0001$) had significant effects on selection errors. There was also a significant interaction effect between timing and distinguishability ($F_{2,22} = 3.6$, $p < .05$). Post-hoc tests showed that the *speech* modality ($\mu = 0.10$) had significantly lower selection error counts than *h-gaze* ($\mu = 0.24$). No significant differences were detected when comparing *foot* ($\mu = 0.19$) to *h-gaze* and *speech*. The *after* ($\mu = 0.11$) timing was significantly lower in selection error than *before* ($\mu = 0.24$), while differences between *during* ($\mu = 0.17$) and the other timings were not significant. Participants made significantly fewer selection errors on *HD* tasks ($\mu = 0.05$) compared to *LD* tasks ($\mu = 0.29$).

5.3 Manipulation Errors

Similar to selection error, the ANOVA confirmed significant main effects of *modality* ($F_{2,22} = 3.2$, $p < .05$), *timing* ($F_{2,22} = 21.4$, $p < .0001$) and *distinguishability* ($F_{1,11} = 14.211$, $p < .0005$) on manipulation errors, as well as significant interaction effects between timing and distinguishability ($F_{2,22} = 7.118$, $p < .005$). Post-hoc tests found no significant differences between pairs of modalities. The means for *h-gaze*, *speech* and *foot* were 1.10, 1.16 and 1.36 respectively. In terms of *timing*, *during* ($\mu = 0.80$) had significantly lower manipulation errors than *before* ($\mu = 1.38$) and *after* ($\mu = 1.44$). For *distinguishability*, *HD* ($\mu = 1.04$) produced significantly less manipulation errors than *LD* ($\mu = 1.37$).

5.4 Subjective Questionnaire Ratings

After the study, we asked participants four questions on a 5-point Likert scale (1: strongly disagree, 5: strongly agree) for each *modality* and *timing*. We examined the following metrics, asking participants to give a rating for each of the statements below:

a. *(Ease of Learning)*: This modality/timing was easy to learn.
b. *(Ease of Use)*: It was easy to use this modality/timing to complete the task.
c. *(Intuitiveness)*: This modality/timing felt natural or intuitive for completing the task.
d. *(Willingness to Use)*: I would like to use this modality/timing for resolving ambiguity when interacting with HMDs.

A Friedman test with correction for ties saw that modality had a significant effect on all four metrics ((a) $\chi^2(2) = 13.862$, $p < .005$; (b) $\chi^2(2) = 18.488$, $p < .0005$; (c) $\chi^2(2) = 19.436$, $p < .0005$; (d) $\chi^2(2) = 21.415$, $p < .0001$). Post-hoc tests confirmed that in every case, *foot* was rated significantly lower than *h-gaze* and *speech*. We also found a significant effect of timing on all four metrics ((a) $\chi^2(2) = 13.714$, $p < .005$; (b) $\chi^2(2) = 15.730$, $p < .0005$; (c) $\chi^2(2) = 17.897$, $p < .0005$; (d) $\chi^2(2) = 16.800$, $p < .0005$). For every metric, *during* had a significantly higher rating than *before* and *after*. Additionally, for metrics (c) and (d), *before* was rated significantly higher than *after*.

5.5 RTLX

Analysis of the RTLX scores and the workload factors informed us that modality significantly influenced task workload ($F_{2,33} = 5.9$, $p < .01$). Post-hoc tests showed significant differences in workload between *foot* ($\mu = 56.11$) and *h-gaze* ($\mu = 36.88$) as well as *foot* and *speech* ($\mu = 37.92$). For the workload dimensions, modality had significant main effects on physical demand ($F_{2,33} = 7.1$, $p < .005$), performance ($F_{2,33} = 6.9$, $p < .005$), and frustration ($F_{2,33} = 4.836$, $p < .05$). *Foot* ($\mu_{phy} = 64.17$, $\mu_{perf} = 41.67$) required more physical effort and had poorer performance than both *h-gaze* ($\mu_{phy} = 28.75$,

μ_{perf} = 21.67) and *speech* (μ_{phy} = 31.25, μ_{perf} = 22.08). *Foot* (μ = 58.75) was also found to be more frustrating than *h-gaze* (μ = 27.92). No significant effects were detected for mental demand, temporal demand, or effort.

5.6 Qualitative Interview Feedback

Most participants preferred *h-gaze* for disambiguation because it was natural, fast, and relaxing (P1, P2, P5, P9). However, *h-gaze* was sometimes oversensitive to head movements (P4, P5, P6). Several participants liked *speech* for its naturalness, accuracy, and speed (P3, P4, P12). Participants did not like the *foot* modality, mostly because it was physically and mentally demanding and uncomfortable. For timing, participants favored *during* due to the smooth input flow, the freedom to choose when to disambiguate, and the continuous visual feedback (P2, P3, P10, P11). A few participants expressed that *before* was efficient for *HD* tasks because *"right away you could tell which [option] you needed"* (P12). However, *before* was difficult to use for *LD* tasks because it was hard to identify which operation was correct before interacting (P1, P4, P6, P9). Participants generally did not like *after*.

5.7 Discussion of Results

From the quantitative analysis and the qualitative feedback, we saw that *h-gaze* and *during* was the most preferred combination and achieved the best performance. Although there was minimal difference between the TCTs of *h-gaze* and *speech*, more participants favoured *h-gaze*. In addition, even though the TCT for *after* was comparable to *during*, participants generally preferred *during*. We observed a difference between *HD* and *LD* both in terms of TCTs and selection errors. The interaction effect between timing and distinguishability indicated that for *HD*, TCTs and selection errors were similar across all timings; but for *LD*, *before* had much higher TCTs and selection errors than other timings. As explained by participants, for *LD* it was more difficult to judge which operation to choose until the gesture had started. This highlights a key advantage of performing disambiguation *during* a gesture. The results further showed that *speech* had the lowest selection error while *h-gaze* had the highest. This was due to *h-gaze* being more prone to accidental activations. Manipulation errors were the lowest for *during* because of the continuous flow of input. For *before* and *after*, manipulation errors were high because participants needed to release and reapply the grasp when they wished to manipulate a different effect.

Regarding workload, *foot* demanded the highest physical effort and was more frustrating. Although the RTLX did not report any significant differences for mental demand between the modalities, participant comments indicated that *foot* may have required more mental effort. *H-gaze* and *speech* produced similar mental and physical workloads.

6 Discussion and Future Work

We now discuss the contribution of our work more broadly, along with its limitations and possible lines of future work.

6.1 Designing for Timing

We saw that participants preferred the *during* timing over *before* and *after*. *During* was especially useful for *LD* tasks, since it was hard to determine the desired operation until the manipulation began. On the other hand, real-time updates were less critical for *HD* tasks, since the user could tell which operation was the correct one at the beginning of the task. Thus, although *during* is preferred overall, the *before* timing is also a feasible design when operation alternatives have high visual distinguishability.

6.2 Scalability

In this study, we explored disambiguation techniques for a single gestural interaction with a single virtual object. However, often user's actions are part of larger input workflows, and ambiguity could occur at any point. Understanding how disambiguation could be applied to sequenced or hierarchal interactions would be an interesting topic for future work.

A related issue is looking into how our technique can scale to larger sets of alternatives. In our work, we investigated disambiguation techniques when there were only three alternative operations. The number of operations presented to the user may affect the performance of the different modalities and timings. Further work on this topic would be beneficial to inform the design of more complex freehand manipulations of virtual objects.

6.3 Exploring Other Techniques

During the design process, we considered other techniques that would be compelling to explore. For instance, while the hands are grasping the object, the user could move the object towards the preview cube containing the desired option. The object in the user's hands could then switch with the object in the preview cube. Alternatively, the camera could smoothly zoom into the object in a selected preview cube, which would provide a potentially less jarring transition. Designers could also consider using eye gaze techniques proposed in prior research (e.g.[40]) instead of head gaze to reduce head movements, making the system less prone to false activations. Investigating other visual effects and design choices would be valuable for understanding how they might affect performance of the explored factors.

6.4 Limitations

The design and implementation of our study imposed some limitations. First, the speech recognition system sometimes had difficulties in recognizing phrases, particularly with different accents. Occasionally the participants had to repeat a command more than once for it to be recognized. We tried to improve recognition by choosing longer words, after discovering that the speech engine was more accurate with longer-syllable phrases. Second, the *foot* modality only supported right foot-tap interactions, which was physically tiring for participants since they had to balance on their left foot. It was also more difficult to select the leftmost preview cube with the right foot. Moreover, our sample size was relatively small for an experiment involving three factors; additional studies may be necessary to further strengthen the results. Finally, our experiment was in an abstract environment and consisted of manipulations with a single object. In the future, it would be important to understand how our results generalize to other objects, as well as more realistic application environments where multiple objects are present.

7 Conclusion

We have presented a system and set of techniques for disambiguating freehand manipulations in VR. We conducted an empirical analysis of a combination of input modalities and timings, and examined the effects of operation distinguishability on their associated performance time and error rates. Our results showed that overall, head gaze was the most preferred input modality, followed by speech and foot-tap. Disambiguating *during* a gesture was the most compelling timing, due to the real-time feedback and continuous flow of interaction. We hope our results could serve as a guide for ambiguity resolution in bare hand spatial interactions for VR and beyond.

REFERENCES

[1] P. Agar and K. Novins. "Polygon Recognition in Sketch-based Interfaces with Immediate and Continuous Feedback." in *Proceedings of the 1st International Conference on Computer Graphics and Interactive Techniques in Australasia and South East Asia*, New York, NY, USA. 2003. pp. 147–150.

[2] M. Al-Kalbani, I. Williams, and M. Frutos-Pascual. "Analysis of Medium Wrap Freehand Virtual Object Grasping in Exocentric Mixed Reality." in *2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 2016. pp. 84–93.

[3] A. Alzayat, M. Hancock, and M. A. Nacenta. "Quantitative Measurement of Tool Embodiment for Virtual Reality Input Alternatives." in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA. 2019. pp. 443:1–443:11.

[4] M. Annett, T. Grossman, D. Wigdor, and G. Fitzmaurice. "Medusa: a Proximity-Aware Multi-Touch Tabletop." in *Proceedings of the 24th annual ACM symposium on User interface software and technology - UIST '11*, Santa Barbara, California, USA. 2011. p. 337.

[5] R. Arora, R. H. Kazi, D. M. Kaufman, W. Li, and K. Singh. "MagicalHands: Mid-Air Hand Gestures for Animating in VR." in *Proceedings of the 32Nd Annual ACM Symposium on User Interface Software and Technology*, New York, NY, USA. 2019. pp. 463–477.

[6] J. Arvo and K. Novins. "Fluid Sketches: Continuous Recognition and Morphing of Simple Hand-Drawn Shapes." in *Proceedings of the 13th annual ACM symposium on User interface software and technology - UIST '00*, San Diego, California, United States. 2000. pp. 73–80.

[7] R. Atienza, R. Blonna, M. I. Saludares, J. Casimiro, and V. Fuentes. "Interaction Techniques Using Head Gaze for Virtual Reality." in *2016 IEEE Region 10 Symposium (TENSYMP)*. 2016. pp. 110–114.

[8] H. Bai, G. A. Lee, M. Ramakrishnan, and M. Billinghurst. "3D Gesture Interaction for Handheld Augmented Reality." in *SIGGRAPH Asia 2014 Mobile Graphics and Interactive Applications*, New York, NY, USA. 2014. pp. 7:1–7:6.

[9] O. Bau and W. E. Mackay. "OctoPocus: A Dynamic Guide for Learning Gesture-based Command Sets." in *Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology*, New York, NY, USA. 2008. pp. 37–46.

[10] H. Benko, R. Jota, and A. Wilson. "MirageTable: Freehand Interaction on a Projected Augmented Reality Tabletop." in *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*, Austin, Texas, USA. 2012. p. 199.

[11] R. A. Bolt. "'Put-that-there': Voice and Gesture at the Graphics Interface." in *Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques*, New York, NY, USA. 1980. pp. 262–270.

[12] I. Chatterjee, R. Xiao, and C. Harrison. "Gaze+Gesture: Expressive, Precise and Targeted Free-Space Interactions." in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, New York, NY, USA. 2015. pp. 131–138.

[13] L. Dipietro, A. M. Sabatini, and P. Dario. "A Survey of Glove-Based Systems and Their Applications." *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*. vol. 38, no. 4. pp. 461–482. Jul. 2008.

[14] J. Feltham. "Leap Motion's Interaction Engine puts the handy into hand-tracking." *VentureBeat*. 28-Aug-2016. .

[15] D. Freeman, H. Benko, M. R. Morris, and D. Wigdor. "ShadowGuides: Visualizations for In-situ Learning of Multi-touch and Whole-hand Gestures." in *Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces*, New York, NY, USA. 2009. pp. 165–172.

[16] E. Ghomi, S. Huot, O. Bau, M. Beaudouin-Lafon, and W. E. Mackay. "Arpège: Learning Multitouch Chord Gestures Vocabularies." in *Proceedings of the 2013 ACM international conference on Interactive tabletops and surfaces - ITS '13*, St. Andrews, Scotland, United Kingdom. 2013. pp. 209–218.

[17] C. Harrison, D. Tan, and D. Morris. "Skinput: Appropriating the Body As an Input Surface." in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA. 2010. pp. 453–462.

[18] S. G. Hart. "Nasa-Task Load Index (NASA-TLX); 20 Years Later." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. vol. 50, no. 9. pp. 904–908. Oct. 2006.

[19] O. Hilliges, D. Kim, S. Izadi, M. Weiss, and A. Wilson. "HoloDesk: Direct 3D Interactions with a Situated See-Through Display." in *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*, Austin, Texas, USA. 2012. p. 2421.

[20] J. D. Hincapié-Ramos, K. Ozacar, P. P. Irani, and Y. Kitamura. "GyroWand: IMU-based Raycasting for Augmented Reality Head-Mounted Displays." in *Proceedings of the 3rd ACM Symposium on Spatial User Interaction*, Los Angeles, California, USA. 2015. pp. 89–98.

[21] T. Igarashi and J. F. Hughes. "A Suggestive Interface for 3D Drawing." in *Proceedings of the 14th Annual ACM Symposium on User Interface Software and Technology*, New York, NY, USA. 2001. pp. 173–181.

[22] S. Irawati, S. Green, M. Billinghurst, A. Duenser, and H. Ko. "'Move the couch where?': Developing an Augmented Reality Multimodal Interface." in *2006 IEEE/ACM International Symposium on Mixed and Augmented Reality*. 2006. pp. 183–186.

[23] E. Kaiser *et al.* "Mutual Disambiguation of 3D Multimodal Interaction in Augmented and Virtual Reality." in *Proceedings of the 5th International Conference on Multimodal Interfaces*, New York, NY, USA. 2003. pp. 12–19.

[24] D. Kim *et al.* "Digits: Freehand 3D Interactions Anywhere Using a Wrist-worn Gloveless Sensor." in *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology*, New York, NY, USA. 2012. pp. 167–176.

[25] M. Kytö, B. Ens, T. Piumsomboon, G. A. Lee, and M. Billinghurst. "Pinpointing: Precise Head- and Eye-Based Target Selection for Augmented Reality." in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Montreal QC, Canada. 2018. pp. 1–14.

[26] B. Lafreniere, P. K. Chilana, A. Fourney, and M. A. Terry. "These Aren'T the Commands You'Re Looking For: Addressing False Feedforward in Feature-Rich Software." in *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*, New York, NY, USA. 2015. pp. 619–628.

[27] S.-S. Lee, J. Chae, H. Kim, Y. Lim, and K. Lee. "Towards More Natural Digital Content Manipulation via User Freehand Gestural Interaction in a Living Room." in *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, New York, NY, USA. 2013. pp. 617–626.

[28] Z. Lv. "Wearable Smartphone: Wearable Hybrid Framework for Hand and Foot Gesture Interaction on Smartphone." in *2013 IEEE International Conference on Computer Vision Workshops*. 2013. pp. 436–443.

[29] Z. Lv, A. Halawani, S. Feng, H. Li, and S. U. Réhman. "Multimodal Hand and Foot Gesture Interaction for Handheld Devices." *ACM Trans. Multimedia Comput. Commun. Appl.* vol. 11, no. 1s. pp. 10:1–10:19. Oct. 2014.

[30] I. S. MacKenzie. "Evaluating Eye Tracking Systems for Computer Input." *Gaze Interaction and Applications of Eye Tracking: Advances in Assistive Technologies*. pp. 205–225. Jan. 2011.

[31] J. Mankoff, S. E. Hudson, and G. D. Abowd. "Providing Integrated Toolkit-level Support for Ambiguity in Recognition-based Interfaces." in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA. 2000. pp. 368–375.

[32] J. Mankoff, S. E. Hudson, and G. D. Abowd. "Interaction Techniques for Ambiguity Resolution in Recognition-based Interfaces." in *Proceedings of the 13th Annual ACM Symposium on User Interface Software and Technology*, New York, NY, USA. 2000. pp. 11–20.

[33] F. Müller, J. McManus, S. Günther, M. Schmitz, M. Mühlhäuser, and M. Funk. "Mind the Tap: Assessing Foot-Taps for Interacting with Head-Mounted Displays." in *Proceedings of the 2019 CHI*

Conference on Human Factors in Computing Systems, New York, NY, USA. 2019. pp. 477:1–477:13.

[34] A. Olwal, H. Benko, and S. Feiner. "SenseShapes: Using Statistical Geometry for Object Selection in a Multimodal Augmented Reality." in *The Second IEEE and ACM International Symposium on Mixed and Augmented Reality, 2003. Proceedings.* 2003. pp. 300–301.

[35] T. Pakkanen and R. Raisamo. "Appropriateness of Foot Interaction for Non-accurate Spatial Tasks." in *CHI '04 Extended Abstracts on Human Factors in Computing Systems*, New York, NY, USA. 2004. pp. 1123–1126.

[36] K. Pfeuffer, B. Mayer, D. Mardanbegi, and H. Gellersen. "Gaze + Pinch Interaction in Virtual Reality." in *Proceedings of the 5th Symposium on Spatial User Interaction*, New York, NY, USA. 2017. pp. 99–108.

[37] T. Piumsomboon, D. Altimira, H. Kim, A. Clark, G. Lee, and M. Billinghurst. "Grasp-Shell vs Gesture-Speech: A Comparison of Direct and Indirect Natural Interaction Techniques in Augmented Reality." in *2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, Munich, Germany. 2014. pp. 73–82.

[38] T. Piumsomboon, A. Clark, M. Billinghurst, and A. Cockburn. "User-Defined Gestures for Augmented Reality." in *Human-Computer Interaction – INTERACT 2013*. 2013. pp. 282–299.

[39] T. Piumsomboon, A. Dey, B. Ens, G. Lee, and M. Billinghurst. "The Effects of Sharing Awareness Cues in Collaborative Mixed Reality." *Front. Robot. AI*. vol. 6. 2019.

[40] T. Piumsomboon, G. Lee, R. W. Lindeman, and M. Billinghurst. "Exploring natural eye-gaze-based interaction for immersive virtual reality." in *2017 IEEE Symposium on 3D User Interfaces (3DUI)*. 2017. pp. 36–39.

[41] J. Roach. "Making the HoloLens 2: Advanced AI built Microsoft's vision for ubiquitous computing." *Innovation Stories*. 07-Nov-2019. [Online]. Available: https://news.microsoft.com/innovation-stories/hololens-2-shipping-to-customers/. [Accessed: 12-Nov-2019].

[42] M. Savva, A. X. Chang, and M. Agrawala. "SceneSuggest: Context-Driven 3D Scene Design." *arXiv:1703.00061 [cs]*. Feb. 2017.

[43] J. Schwarz, S. Hudson, J. Mankoff, and A. D. Wilson. "A Framework for Robust and Flexible Handling of Inputs with Uncertainty." in *Proceedings of the 23nd annual ACM symposium on User interface software and technology - UIST '10*, New York, New York, USA. 2010. p. 47.

[44] J. Schwarz, J. Mankoff, and S. Hudson. "Monte Carlo Methods for Managing Interactive State, Action and Feedback Under Uncertainty." in *Proceedings of the 24th annual ACM symposium on User interface software and technology - UIST '11*, Santa Barbara, California, USA. 2011. p. 235.

[45] M. Serrano, B. M. Ens, and P. P. Irani. "Exploring the Use of Hand-to-face Input for Interacting with Head-worn Displays." in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA. 2014. pp. 3181–3190.

[46] A. L. Simeone, A. Bulling, J. Alexander, and H. Gellersen. "Three-Point Interaction: Combining Bi-manual Direct Touch with Gaze." in *Proceedings of the International Working Conference on Advanced Visual Interfaces*, New York, NY, USA. 2016. pp. 168–175.

[47] A. L. Simeone, E. Velloso, J. Alexander, and H. Gellersen. "Feet Movement in Desktop 3D Interaction." in *2014 IEEE Symposium on 3D User Interfaces (3DUI)*. 2014. pp. 71–74.

[48] O. Špakov and P. Majaranta. "Enhanced Gaze Interaction Using Simple Head Gestures." in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing - UbiComp '12*, Pittsburgh, Pennsylvania. 2012. p. 705.

[49] H. B. Surale, F. Matulic, and D. Vogel. "Experimental Analysis of Barehand Mid-air Mode-Switching Techniques in Virtual Reality." in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA. 2019. pp. 196:1–196:14.

[50] M. Terry and E. D. Mynatt. "Side Views: Persistent, On-demand Previews for Open-ended Tasks." in *Proceedings of the 15th Annual ACM Symposium on User Interface Software and Technology*, New York, NY, USA. 2002. pp. 71–80.

[51] S. Tsang, R. Balakrishnan, K. Singh, and A. Ranjan. "A Suggestive Interface for Image Guided 3D Sketching." in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA. 2004. pp. 591–598.

[52] R. Wang, S. Paris, and J. Popović. "6D Hands: Markerless Hand-Tracking for Computer Aided Design." in *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, New York, NY, USA. 2011. pp. 549–558.

[53] M. Weigel, V. Mehta, and J. Steimle. "More Than Touch: Understanding How People Use Skin As an Input Surface for Mobile Computing." in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA. 2014. pp. 179–188.

[54] J. O. Wobbrock, M. R. Morris, and A. D. Wilson. "User-Defined Gestures for Surface Computing." in *Proceedings of the 27th international conference on Human factors in computing systems - CHI 09*, Boston, MA, USA. 2009. p. 1083.

[55] P. Xu, H. Fu, T. Igarashi, and C.-L. Tai. "Global Beautification of Layouts with Interactive Ambiguity Resolution." in *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, New York, NY, USA. 2014. pp. 243–252.

[56] K. Yamashita, T. Kikuchi, K. Masai, M. Sugimoto, B. H. Thomas, and Y. Sugiura. "CheekInput: Turning Your Cheek into an Input Surface by Embedded Optical Sensors on a Head-mounted Display." in *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*, New York, NY, USA. 2017. pp. 19:1–19:8.

[57] Y. Yan, C. Yu, X. Ma, X. Yi, K. Sun, and Y. Shi. "VirtualGrasp: Leveraging Experience of Interacting with Physical Objects to Facilitate Digital Object Retrieval." in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, Montreal QC, Canada. 2018. pp. 1–13.

[58] K. Yin and D. K. Pai. "FootSee: An Interactive Animation System." in *Proceedings of the 2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, Aire-la-Ville, Switzerland, Switzerland. 2003. pp. 329–338.