Multimodal selection of objects in a dense virtual environment: Ray casting and voice commands

José Rodolfo Mondragón Zenteno, *Fellow, UGA,* and Hussam Tariq, *Fellow, UGA*

***Abstract*—Object selection in dense virtual environments remains an open challenge in Human-Computer Interaction (HCI). While multimodal techniques have been explored for object manipulation, their use in selection tasks is limited. This work introduces a multimodal technique combining ray casting and voice commands to enable precise and fast object selection using a Wizard of Oz (WoZ) experiment design. Results demonstrate a 17.27% reduction in selection time and a 32.72% decrease in the overall average selection errors.**

***Index Terms*—Virtual Reality, Object Selection, Dense Virtual Environment, Multimodal selection**

I. Introduction

In the last decade, Virtual and Augmented Reality (VR and AR, respectively) have gained popularity because of their ability to simulate our world and take us into immersive digital experiences or interactive virtual environments. This has led to their application in several fields such as education, engineering and architectural design, marketing, and healthcare, among others.

An open challenge in the field arises within the interaction between users and virtual environments. When users try to point and select a target in a scenario where multiple objects are close and partially occluding each other, the task becomes difficult, tiring, and occasionally frustrating. This kind of scenario is what we call a **dense virtual environment**.

Ray casting is the most popular strategy for pointing and selecting objects in virtual environments since it is intuitive and similar to the natural task of finger pointing, nevertheless, as Marc Baloup et al. explain, its precision and efficiency depend on the quality of the device, the user's motor skills and natural pulse tremble. These metrics drop especially when the objects are away, in dense environments, and appear to be small [1]. In recent years, multiple techniques with novel approaches have been proposed to address this issue, we will review some of them in the next section.

In this work, we propose a two-stage multimodal technique that uses ray casting and voice commands to achieve a precise and fast object selection in dense virtual environments.

II. Related Work

As previously mentioned, several object selection techniques for virtual dense environments have been proposed in recent years, we will focus our attention on those that use a two-stage process like ours.

In 2023, Chaffangeon et al. [2] proposed an object selection method called "look and midair", this two-stage process uses ray casting during the first stage to point to the desired object, and a zoom-in window to expand the objects within a predefined group selection frame, once the area is selected, a transparency filter to view behind occluding objects can be moved within the preselected space with an additional controller, and the pointer is used to select the final target.

Sheldon Dobbs et al. presented a novel system for architectural design in virtual environments, with a multimodal tool function that allows the user to use voice commands to make changes to the selected object properties such as color, material, size, orientation, and position. Despite good results, the interoperability of the software brought communication issues such as latency and sometimes misunderstandings of the commands even in controlled environments. Although the commands were not applied in the object selection process, the study sheds light on the potential of using both modalities to improve the virtual experience and allow users to interact with 3D environments more naturally. [3]

In 2013, William Delamere et al. presented two different selection strategies, both use a volume pointing stage first and secondly a disambiguation selection stage using hand gestures. In the first one, a conical volume around the pointing ray is computed, and the group of objects that fall into it are "preselected". In the second stage, they use wrist rolling or sliding gestures to help the user select a specific object from that group. As a result, they found that these methods outperformed other techniques that required the user to lose focus on the disambiguation techniques. However, the user needs to keep pointing to the preselected object group which can lead to fatigue. [4]

III. Methodology and Experimental Framework

As discussed, selecting targets in dense virtual environments remains an open problem in the VR Human-Computer Interaction research field. In this work, we propose using a combination of ray casting and voice commands in a two-stage process when selecting an object in such scenarios. Our goal is to evaluate if this combination of modalities can be helpful for the user in the final selection task since humans naturally use the voice to communicate.

A. Method

Our approach comprises two different stages, a volume or sub-group selection step, and a disambiguation step:

1) Fist Stage: Ray cast for sub-group selection: In this step, we use a controller to point to an object (of radius $r_{object} = 0.5m$) and select a spherical volume with a radius of $r_{volume} = 1m$ around it using ray casing. Then, the objects that are touched or reached by this volume are pre-selected. The spherical volume is not visible to the user, it is used only to compute the closest objects to the initial selection.

The user must point to an object that is close to the desired target, and then to select the object sub-group, the user should click a button while pointing. An exemplification of this can be seen in figure 1.

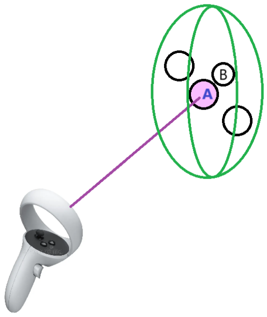

Fig. 1. Visual representation of the volume selection. B is the target, and the user selects a close object A.

2) Second Stage: Disambiguation step: To enter the disambiguation stage, the user has to take his pointing arm to the resting position while keeping the selection button pressed. Once the system detects that the controller's orientation is pointing to the ground while an object is selected, the user can use voice commands to move the selection to the desired target.

We introduce six voice commands, five to control the position of the selection: *"behind, up, down, left, right"*, and one for the final object selection *"select"*. The figure 2, shows a visual representation of the process. To abort the second stage because the desired target was not within the pre-selected group, the user has to release the selection button on the controller, this action eliminates the pre-selected group, and the user can start the process all over again.

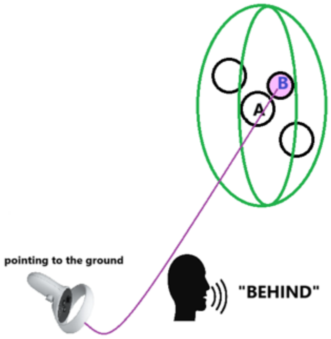

Fig. 2. Visual representation of the voice command functionality. The user points to the ground while holding the selection and says a voice command to move the selection to the desired object B.

B. Prototype Development

This system was implemented in a Meta Quest 3 and Touch headset, the virtual world and scenes for experiments were created in the Unity VR IDE, and the custom software for the multimodal selection was written in C# using VisualStudio 2022 IDE. To achieve voice command recognition, a WoZ design was implemented: one of the authors moves the selection in the spoken direction with a hidden controller to simulate that the system has heard the voice command, but the users are not aware of this procedure.

C. Experimental Framework

1) Experimental Set-up: To conduct experiments, 6 different dense virtual scenarios were built. Every scenario contains a cloud of twenty spherical objects floating close to each other, producing some occlusions.

A single target object was placed in each scene. To identify it, it was made green color, and the rest (19 objects) were made red. All the spheres have the same size, a radius of $r_{object} = 0.5m$ (See figure 3).

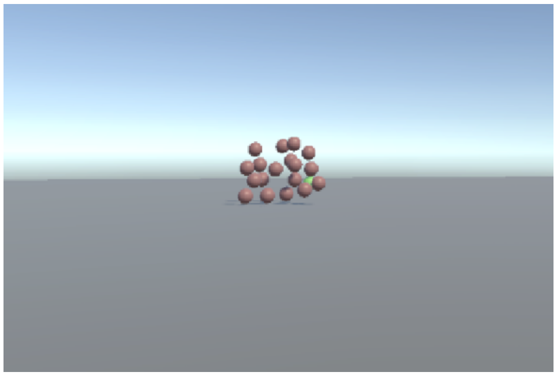

Fig. 3. Sample of a test scene in the virtual world

The pointing ray was made visible to the user, it has a red color when no selectable object is being pointed at and blue when the opposite. This visual feedback is introduced to help the user notice when he is pointing to an intractable object. The figure 5, shows an example of how the user sees the scene when pointing to an object.

The user's point of view was then placed at a far distance where the objects seemed small. To select it, a pilot test was carried on with 3 participants. For this process, we used only 3 scenarios and ray casting to select objects. The user view was placed at 2 different positions: at $15m$ and $20m$ away from the objects. All of the participants' performance dropped at the second option, going from making 0.33 average selection errors per scene to 0.78. Thus, we selected the second distance ($d = 20m$) as the final for the system testing.

The testing scenes were connected so that when the target was selected, the next was loaded, allowing continuity in the experiments. To interact with the virtual environment, the user has to wear the headset and use the right controller to point and select the objects.

The baseline of comparison is the ray casting selection technique, and to compare the performance we use the same 6 scenarios for each strategy.

2) Participants: Twelve subjects were recruited to participate in the experiments, their ages ranged from 20 to 30 with an average age of 24.8 years old. Regarding their academic background: 7 of them belong to computer science, 3 to social sciences, 1 to engineering, and one to mathematics. From the twelve, 8 had never used VR headsets before.

3) Process: At the start, the participants are placed in a training scene where they can familiarize themselves with wearing the VR device and the selection technique that they will use for the first part of the experiment. Six participants started with the ray casting technique and the rest with the multimodal technique. In this first scene, they receive instructions about how to use the Quest controller to point at the objects and which button to press to select them. If they start with the multimodal system, they are also explained which commands they have, and how to enter and exit the disambiguation stage. Figure 4 shows the training scenes for the participants.

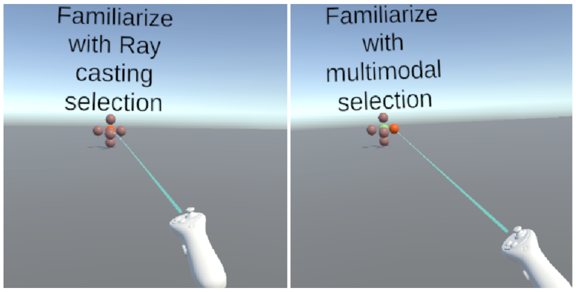
Familiarize with Ray casting selection
Familiarize with multimodal selection

Fig. 4. Training scenes for the experiment participants.

Once the participant confirms that feels comfortable, is taken to the first test environment. When the participant crosses all 6 scenarios by correctly selecting the green target, is given a break of 5 minutes. Figures 5, 6 and 7 display examples of the experimental set-up in use. After the break, the participant is placed into a second training scene where receives instructions on how to use the second selection technique. Again, when the user confirms that is ready, is taken into a first testing environment, and the experiment ends when all the 6 scenarios are crossed. For both techniques, the order of the testing scenarios is randomized, however, the content of the scenes remains the same to have a fair comparison.

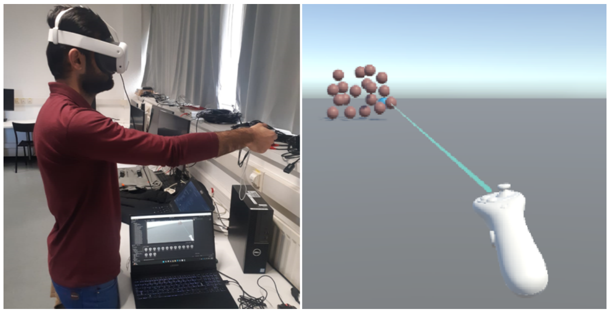

Fig. 5. The experimental set-up in use. On the left, a person uses the ray casting technique to point and select a target. On the right, the user's point of view during the task.

Finally, the participants are asked to answer a form regarding their experience using both selection methods, this data is used to make a qualitative evaluation that is discussed in the next section.

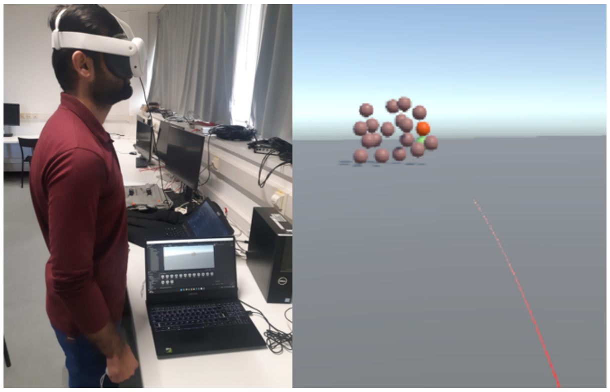

Fig. 6. The experimental set-up in use. On the left, a person uses the disambiguation stage to select a target. On the right, the user's point of view during the task.

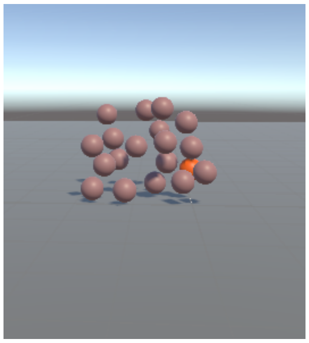

Fig. 7. The user's point of view after using the voice command *"down"*. The selection has moved towards the target.

IV. Results

A. Quantitative Results

To evaluate and compare our proposed method to the baseline we selected the following metrics:

- Average number of errors per scene (AES): It's the average number of errors made by the participants through six scenes, it is calculated for each technique.
- Overall average number of errors per scene (OAES): As the name suggests, it averages the previous measurement over the twelve participants.
- Average time to complete the experiment: The time that the user took to complete the experiment with each technique.

Figure 8 displays the user's average number of errors when using ray casting and multimodal selection. One can see that our method helps must of the user to reduce the number of errors.

Figure 9 displays the amount of time (in seconds) that the user needed to complete the experiments when using ray casting and multimodal selection. One can see that the proposed technique helped to the majority of the users to reduce the number of errors.

Finally, Tables I and II show the overall average number of errors and average time to complete the experiments with each method respectively.

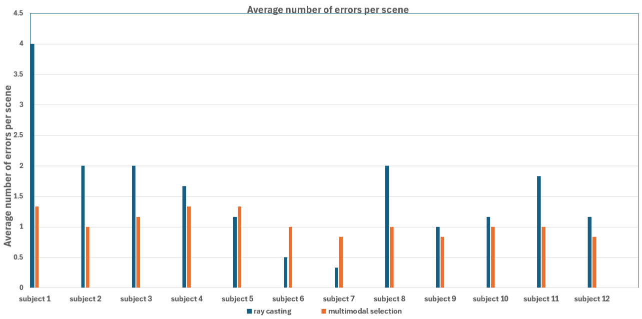

Fig. 8. This graphic shows the average number of errors made by the participants with each selection technique

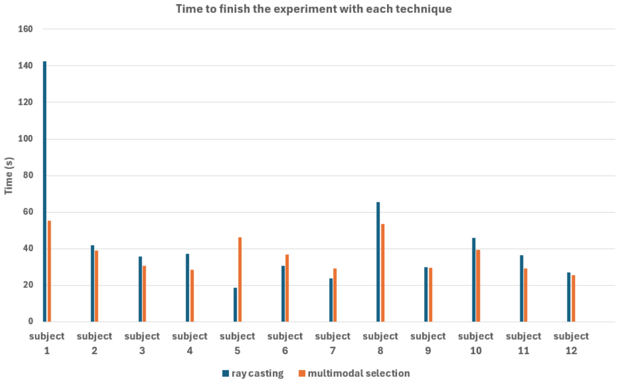

Fig. 9. This graphic shows the time taken by each participant to finish the experiment with each selection technique

Overall Average number of selection errors per scene	
Ray casting	**Multimodal Selection**
1.569444444	**1.055555556**
Error reduction (%)	**32.727**

TABLE I
OVERALL AVERAGE NUMBER OF ERRORS PER SCENE FOR RAY CASTING AND MULTIMODAL SELECTION METHODS.

Average time to complete the experiment	
Ray casting	**Multimodal Selection**
44.595 s	**36.92 s**
Time reduction (%)	**17.21**

TABLE II
AVERAGE TIME TO COMPLETE THE EXPERIMENT USING RAY CASTING AND MULTIMODAL SELECTION METHODS.

B. Qualitative Results

The participants were asked to answer a form about their experience during the experiments and to evaluate the techniques.

First, they had to evaluate how easy was to select the targets with each technique in a scale from one to five (where one is difficult and five is easy). Figures 10 and 11 display the results, showing that the users perceive the multimodal technique as easier to use.

Going deeper in their personal experience, 75% of the participants stated that they believe this method helped them

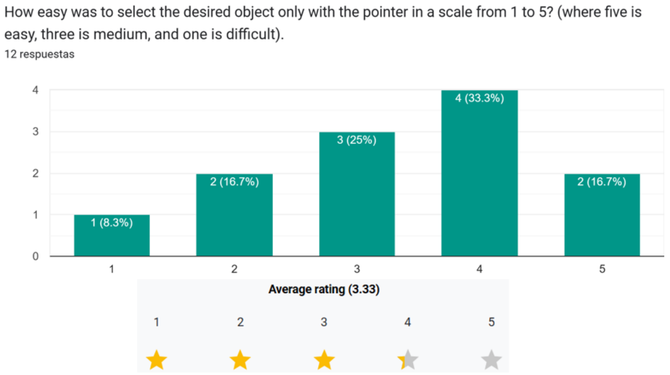
How easy was to select the desired object only with the pointer in a scale from 1 to 5? (where five is easy, three is medium, and one is difficult).
12 respuestas

Fig. 10. User perception of the difficulty to select objects using ray casting. The scale is from one to five, where five is easy and 1 is difficult

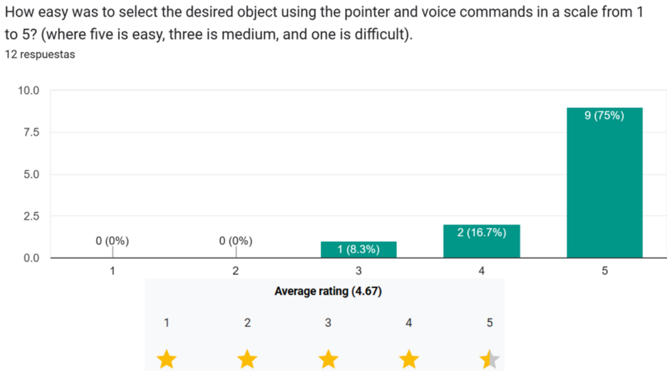
How easy was to select the desired object using the pointer and voice commands in a scale from 1 to 5? (where five is easy, three is medium, and one is difficult).
12 respuestas

Fig. 11. User perception of the difficulty to select objects using multimodal selection. The scale is from one to five, where five is easy and 1 is difficult

to achieve that goal faster. Also, 83.3% of them said that the multimodal technique was less tiring for their arm and posture, from this assessment, we could say that the proposed system seems to be more comfortable and friendly to the user.

V. CONCLUSIONS

The results show that the proposed multimodal system outperforms the baseline, reducing average selection time by 17.21% and the overall average selection errors by 32.72%. However, the error metric may be biased, as most participants preferred the two-stage selection process (pointing and voice command), potentially inflating the method's performance. Additionally, some participants restarted the selection process instead of repeating misheard commands, impacting the comparison's fairness.

Qualitative evaluations align with the quantitative results, highlighting improved user experience. However, participants noted issues with the ”behind” command in ambiguous scenarios where the target's position combines ”behind” with another direction (e.g., ”behind and right”). This limitation stems from the design, which associates objects with the closest coordinate axis in the selection volume.

Participants also suggested offering the system as an optional feature rather than an always-on tool, citing the learning curve and potential annoyance of repeated voice commands.

These insights emphasize the need to minimize user effort and explore strategies to streamline the selection process in future work.

REFERENCES

[1] M. Baloup, T. Pietrzak, and G. Casiez, “Raycursor: A 3d pointing facilitation technique based on raycasting,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, (New York, NY, USA), p. 1–12, ACM, 2019.

[2] A. Chaffangeon Caillet, A. Goguey, and L. Nigay, “3d selection in mixed reality: Designing a two-phase technique to reduce fatigue,” in *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, p. 800–809, IEEE, 2023.

[3] A. Sheldon, T. Dobbs, A. Fabbri, N. Gardner, H. Haeusler, C. Ramos, and Y. Zavoleas, “Putting the ar in (ar)chitecture - integrating voice recognition and gesture control for augmented reality interaction to enhance design practice,” in *CAADRIA proceedings*, CAADRIA, 2019.

[4] W. Delamare, C. Coutrix, and L. Nigay, “Mobile pointing task in the physical world: balancing focus and performance while disambiguating,” *International Conference on Human-Computer Interaction with Mobile Devices and Services*, vol. 27, p. 89–98, 2013.