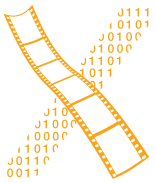

Chair for Computer Science 10 (Media Computing and Human-Computer Interaction)

RWTH AACHEN UNIVERSITY

ShARePen: Enabling Spectators to Understand Pointing Operations in Co-Located Handheld AR

Master's Thesis
submitted to the
Media Computing Group
Prof. Dr. Jan Borchers
Computer Science Department
RWTH Aachen University

by
Marvin Bruna

Thesis advisor:
Prof. Dr. Jan Borchers

Second examiner:
Dr. Simon Völker

Registration date: 23.05.2022
Submission date: 07.12.2022

Zentrales Prüfungsamt/Central Examination Office

Eidesstattliche Versicherung
Statutory Declaration in Lieu of an Oath

Name, Vorname/Last Name, First Name

Matrikelnummer (freiwillige Angabe)
Matriculation No. (optional)

Ich versichere hiermit an Eides Statt, dass ich die vorliegende Arbeit/Bachelorarbeit/ Masterarbeit* mit dem Titel
I hereby declare in lieu of an oath that I have completed the present paper/Bachelor thesis/Master thesis* entitled

selbstständig und ohne unzulässige fremde Hilfe (insbes. akademisches Ghostwriting) erbracht habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt. Für den Fall, dass die Arbeit zusätzlich auf einem Datenträger eingereicht wird, erkläre ich, dass die schriftliche und die elektronische Form vollständig übereinstimmen. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.
independently and without illegitimate assistance from third parties (such as academic ghostwriters). I have used no other than the specified sources and aids. In case that the thesis is additionally submitted in an electronic format, I declare that the written and electronic versions are fully identical. The thesis has not been submitted to any examination body in this, or similar, form.

Ort, Datum/City, Date

Unterschrift/Signature

*Nichtzutreffendes bitte streichen

*Please delete as appropriate

Belehrung:
Official Notification:

§ 156 StGB: Falsche Versicherung an Eides Statt
Wer vor einer zur Abnahme einer Versicherung an Eides Statt zuständigen Behörde eine solche Versicherung falsch abgibt oder unter Berufung auf eine solche Versicherung falsch aussagt, wird mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft.
Para. 156 StGB (German Criminal Code): False Statutory Declarations
Whoever before a public authority competent to administer statutory declarations falsely makes such a declaration or falsely testifies while referring to such a declaration shall be liable to imprisonment not exceeding three years or a fine.
§ 161 StGB: Fahrlässiger Falscheid; fahrlässige falsche Versicherung an Eides Statt
(1) Wenn eine der in den §§ 154 bis 156 bezeichneten Handlungen aus Fahrlässigkeit begangen worden ist, so tritt Freiheitsstrafe bis zu einem Jahr oder Geldstrafe ein.
(2) Straflosigkeit tritt ein, wenn der Täter die falsche Angabe rechtzeitig berichtigt. Die Vorschriften des § 158 Abs. 2 und 3 gelten entsprechend.
Para. 161 StGB (German Criminal Code): False Statutory Declarations Due to Negligence
(1) If a person commits one of the offences listed in sections 154 through 156 negligently the penalty shall be imprisonment not exceeding one year or a fine.
(2) The offender shall be exempt from liability if he or she corrects their false testimony in time. The provisions of section 158 (2) and (3) shall apply accordingly.

Die vorstehende Belehrung habe ich zur Kenntnis genommen:
I have read and understood the above official notification:

Ort, Datum/City, Date

Unterschrift/Signature

Contents

List of Figures

List of Tables

Abstract

Research into the collaborative usage of Augmented Reality (AR) has been an ongoing topic over the last decades. Modern hand-held systems, such as smartphones, allow the user an augmented view of the real-world, changing the way we interact with virtual information. It is essential for two or more people who are working together to have constant knowledge about what the other person(s) is/are working on or referring to. In everyday communication, we often achieve this by pointing at objects, persons, animals etc., such that another person can be sure what we are referring to. In handheld AR however, this simple method of specifying something is often difficult to understand, due to the different viewports each user has, that create issues where certain objects might be occluded or not in the current field of view (FoV) for some users. Additionally, depth perception makes it hard to accurately point at something, therefore a ray cast approach is often chosen to alleviate this issue, but this makes it harder to follow the pointing again.

To investigate how we can help a spectator understand such ray cast pointing operations with the ARPen, a bimanual system using a trackable pen and a smartphone, we developed four different visualization techniques based on the research with similar systems. These four techniques were a simple highlighting (*Baseline*), the representation of the ray cast (*Ray*), giving the scene a see-through look by decreasing the opacity of currently not pointed at objects (*Opacity*) and a picture-in-picture (PiP) live video stream showing the perspective as seen from the presenter's device (*Video*).

The conducted user study, that evaluated and compared these methods in different user to user positions over two linked tasks, recognizing pointed at objects and then relocating them, revealed that the simple highlighting already offers a good visual representation that can easily be understood by the spectator. However, they preferred having the chance to additionally add the ray visualization or use the opacity mode if they needed them specifically or if they felt generally helpful to them. The perspective switch via the video had a generally negative impact on the performance in the tasks and was also disliked the most, which also showed in its

low amount of usage. Our findings provide a starting point for research into the collaborative possibilities the ARPen has to offer.

Überblick

Wie man Augmented Reality (AR) in gemeinsamen Tätigkeiten nutzen kann ist ein fortlaufendes Forschungsthema in den letzten Jahrzehnten. Moderne Handheld Systeme, wie z.b. Smartphones, erlauben es dem Nutzer eine augmentierte Sicht auf die reale Welt zu haben und verändern somit die Art der Interaktion mit virtuellen Informationen. Es ist essentiell für zwei oder mehr Personen die zusammenarbeiten zu jedem Zeitpunkte zu wissen, was die andere(n) Person(en) gerade macht/machen oder auf was sie sich bezieht/beziehen. In täglicher Kommunikation erreichen wir dies oft, indem wir auf Objekte, Personen, Tiere etc. zeigen, damit eine weitere Person sich sicher seien kann auf was wir verweisen. In Handheld AR ist diese einfache Methode etwas zu spezifizieren allerdings häufig schwer zu verstehen, da jeder Nutzer sein eigenes Sichtfenster hat, was zu Problemen führen kann, da bestimmte Objekte für einzelne Nutzer verdeckt sein können oder sie sich nicht im aktuellen Sichtfeld befinden können. Zudem macht es die Tiefenwahrnehmung schwierig mit hoher Genauigkeit auf etwas zu zeigen, weshalb häufig eine raycasting Herangehensweise gewählt wird um diesem Problem entgegenzuwirken, allerdings macht diese Herangehensweise es wieder schwerer der Geste zu folgen.

Um zu untersuchen wie man einem Zuschauer helfen kann solche raycasting Zuweisungen mit dem ARPen, einem zweihändigem System bestehend aus einem verfolgbarem Stift und einem Smartphone, zu verstehen, haben wir vier verschiedene Visualisierungen, basierend auf Forschungen mit ähnlichen Systemen, entwickelt. Diese vier Techniken waren ein einfaches hervorheben (*Grundlinie*), die Repräsentation des Raycasts (*Strahl*), die Möglichkeit der Szene ein transparentes Aussehen zu geben indem die Opazität aller Objekte auf die nicht gezeigt wird reduziert wird (*Opazität*) und einer Bild in Bild (BiB) live Videoübertragung, welches die Perspektive aus Sicht von dem Gerät des Präsentators zeigt (*Video*).

Die durchgeführte Nutzerstudie, in der diese verschiedenen Methoden in unterschiedlichen Nutzer zu Nutzer Positionen mittels zwei miteinander verbundenen Aufgaben, die gezeigten Objekte erkennen und sie danach wiederfinden, evaluiert

und verglichen wurden haben gezeigt, dass das einfache Hervorheben bereits eine gute Visualisierung bietet, welche einfach von einem Zuschauer verstanden werden kann. Allerdings wurde die Möglichkeit sich zusätzlich den Strahl anzeigen zu lassen oder der Opazitätsmodus bevorzugt, entweder wenn sie explizit gebraucht wurden oder wenn sie im allgemeinen als hilfreich betrachtet wurden. Der Perspektivenwechsel mittels des Videos hatte generell einen negativen Einfluss auf die Leistungen in den Aufgaben und war am unbeliebtesten, was sich auch in der geringen Verwendung von dieser Methode zeigt. Unsere Ergebnisse stellen einen Startpunkt für die Forschung in die kollaborativen Möglichkeiten die der ARPen bietet dar.

Acknowledgements

This thesis would not have been possible without the support of many people.

First, I want to thank my supervisor Dr. Philipp Wacker, who was always there to give me his invaluable feedback and advice throughout the time spend on this thesis, providing me with valuable lessons on where to improve. As he is leaving our chair at the end of the year, I also want to take this chance to wish him only the best in the future.

My gratitude also goes to all the people who joined my user study, offering their valuable time to help me and giving me greatly appreciated feedback and data to work with.

A special thanks to the four friends I only made last year, while we worked on the master practicum together. The regular evenings we since have spend together provided a welcome distraction.

Even greater thanks go to the two childhood friends who came to visit me after quite a long time of not seeing each other, participating in my user study and then spending a nice evening together.

Last but not least, I want to thank my whole family for their continuous support throughout this thesis, but also throughout my whole time spend as a student in which I was often away from them.

Conventions

Throughout this thesis we use the following conventions:

- The whole thesis is written in American English.
- The first person is written in plural form.
- Unidentified third persons are described in male form.

Definitions of technical terms or short excursus are set off in colored boxes.

EXCURSUS:
Excursus are detailed discussions of a particular point in a book, usually in an appendix, or digressions in a written text.

Definition:
Excursus

Conditions in the user study and other important words are written in italic-style text.

myCondition or *myImportantWord*

Source code and implementation symbols are written in typewriter-style text.

`myClass`

Download links are set off in colored boxes.

File: myFile[a]

[a] http://hci.rwth-aachen.de/public/users/bruna/file_number.file

Chapter 1

Introduction

Research into AR and VR started back in the late 1960s

The beginnings of research into Virtual Reality (VR) and Augmented Reality (AR) date back nearly 55 years into the late 1960s, when Sutherland [1968] developed the first prototype of a Head-mounted display (HMD), giving the world an initial glimpse into the possibilities that VR and AR would offer in the future. While research continued over the following decades, with improvements in a lot of different areas as well as the introduction of new systems, like the CAVE [Cruz-Neira et al., 1992], it took until 2008 for the first commercial use of AR to occur. The Bayerische Motoren Werke (BMW) released a virtual 3D-Model alongside their new campaign for the latest Mini Cabrio [BMW, December 1st, 2008]. However, over the last few years VR and AR had a rapid increase in popularity, caused by systems like the HTC Vive[1] and games, such as Pokémon GO[2], as they got more affordable for the consumers market. A similar increase also applies to the professional sector, where the possibilities of AR are getting explored more often, e.g., Live BIM models[3] or Lenovo's latest plans for the future of the Metaverse presented in their Tech World [Lenovo, October 18th, 2022].

[1] https://www.vive.com/de/

[2] https://www.pokemon.com/de/app/pokemon-go/

[3] https://www.linkedin.com/posts/michaelwilliams1984_bim-collaboration-consultants-activity-6971067023149727744-8DXi

VR immerses the user, while AR only overlays virtual data

VR and AR both allow the user to interact with virtual objects. VR does this by fully emerging the user in the experience, whereas AR places the virtual content on top of the real-world. AR therefore creates a "middle ground" between VR and telepresence (completely in the real-world) [Azuma, 1997]. Azuma further defined AR as systems that follow three key characteristics:

1. Combines real and virtual.
2. Interactive in real time.
3. Registered in 3D.

This makes VR and AR systems appealing for research on collaboration, as they offer the possibility for remote interaction, while also potentially enhancing co-located collaborative work. In both cases this happens by supplementing the scene with additional information, giving the users a higher chance to establish a common ground of knowledge needed to work together or other additional information that enhance the experience or work.

Research on remote collaboration has been intensively explored

The research on how to use AR in a collaborative environment has been an ongoing topic. However, most of it focuses on remote collaboration, where an expert guides a layman through a problem solving or learning process. The collaborative guiding process has been done through a variety of techniques, e.g., showing the local worker a point-cloud representation of the remote user's hands [Gao et al., 2016], using a virtual laser-pointer [Hoppe et al., 2018] or using other forms of spatial annotations [Thoravi Kumaravel et al., 2019] and [Ludwig et al., 2021], to refer to and point at different objects.

Lack of research on co-located scenarios

While there has been a large amount of research conducted in the remote setting, there exists a distinct lack of research that focuses on co-located collaboration, meaning that users share the same physical space. In the remote setting HMDs, e.g., the Microsoft HoloLens[4] are the preferred choice, however for co-located AR, hand-held devices, such as mod-

[4] https://www.microsoft.com/de-de/hololens

ern smartphones, offer a "minimally intrusive, socially acceptable, readily available and highly mobile" [Zhou et al., 2008] way to display and interact with AR content, making them an accessible alternative to HMDs. However, this means that every user has their own unique view at the content, as opposed to only seeing what the local worker is seeing in the remote setting. Since every person uses their own device, each user has their own unique view and unique position inside the scene and towards other users. This amplifies problems, like the occlusion of certain objects by others, which can lead to issues regarding "referential awareness", where a set of objects is only visible to specific, but not all, users at any given time (section 2.2. in [Argelaguet et al., 2011]) and can cause people to tend to position themselves in a way, such that their view matches that of others [Poretski et al., 2021]. This makes collaboration difficult, as the establishment of a common ground of knowledge becomes complicated, which is why deictic gestures are even more important when working in co-located handheld AR.

Deictic gestures are even more important when being co-located

In everyday communication we often use pointing, or deictic gestures as they are more commonly referred to in literature, as means to specify objects that are part of a conversation or to shift the focus of a listener to enhance or even enable collaboration in the first place.

Deictic gestures are an everyday communication tool

DEICTIC GESTURES:
Deictic gestures are generally understood as 'pointing gestures' that indicate real, implied or imaginary persons, objects, directions, etc., and are strongly related to their environment or 'gesture space', including their point of origin (origo) and occur with or without talk. - [Price and Jaworski, 2012]

Definition:
Deictic gestures

In a remote scenario, there has to be a way to convey a gesture over to a local worker and when looking at virtual objects, perspective and especially depth perception can be an issue [Kruijff et al., 2010]. While these issues can be worked on by both parties in the remote setting, e.g., the expert instructing the local worker to shift his head to adjust the camera feed, in a co-located handheld scenario, where ev-

ery person has their own non-shared field of view (FoV), it is harder to do so, since we cannot see what another person is exactly viewing. This makes it even more difficult to perceive and convey deictic gestures, as they might not be visible to all users, due to their individual perspectives. A common solution to alleviate perspective and depth perception problems is the use of the ray cast method. Instead of using the physical hands to perform the pointing, a ray is send out from a device, e.g., a controller, to highlight objects it intersects with. This method has proven to be a reliable and preferred method with regards to pointing or selection operations and tasks ([Hartmann and Vogel, 2018], [Mifsud et al., 2022] and [Wacker et al., 2019]).

1.1 ARPen

The ARPen system allows for hand-held mid-air sketching and manipulation

Wacker et al. [2019] developed a bimanual system, consisting of a smartphone and a trackable pen, which is the cornerstone of this thesis. The system allows users to sketch 3D-objects in mid-air and to manipulate other virtual objects with the pen. Wacker et al. also already investigated different selection techniques with the ARPen[5], therefore we felt that the pen could be used in a collaborative setting as a pointing device and maybe as a tool for simultaneous collaborative sketching in the future. In the chapter 2 "Related Work" we take a closer look at how the ARPen works and what the results of the selection techniques study were in detail.

1.2 Outline

The goal of this thesis is to compare different visualization techniques for pointing operations using the ARPen in a co-located collaborative environment, where one user performs the pointing with the ARPen and another user has to recognize and relocate the pointed at objects, showing that he in fact understood the pointing. This allows us to

[5] https://github.com/i10/ARPen

establish some guidelines for the future use of the ARPen in a collaborative work environment, where pointing operations will play an important role to be able to perform and excel in collaborative tasks. First, we look at a process called "grounding", why pointing is relevant for that "grounding" and why it is an important part of being able to collaborate together. After that, we begin looking at the related work done in the remote and the co-located setting. Next, we take a short digression explaining how the ARPen works, what it's current functions are and the results obtained from prior studies on and with it. After that, we discuss how we arrived at the different visualization techniques researched in this thesis and how we expanded the ARPen app to support a shared AR session, highlighting certain problems we encountered along the way and how they were eventually solved. We then compare the different visualization techniques against each other in a user study, whose setup, procedure and results are presented. These results are used to give recommendations on how to proceed with the use of the ARPen in a co-located collaborative environment and what techniques can enable and facilitate the understanding of pointing operations allowing for a faster, better and less stressful establishment of common knowledge about a shared AR scene. A summary and suggestions for future work finalize this thesis.

Chapter 2

Related Work

"It takes two people working together to play a duet, shake hands, play chess, waltz, teach or make love. To succeed, the two of them have to coordinate both the content and the process of what they are doing."

—Herbert H. Clark and Susan E. Brennan in "Grounding in communication" published in the book "Perspectives on socially shared cognition"

2.1 Overview

We start, by giving a short introduction on why a common ground of knowledge is important for collaboration and how it is normally established in everyday communication. Following this, we look at some of the research done on collaborating in a remote environment, as it still gives good indication what techniques can be viable and why they might have failed in the past. We look at how pointing operations are conveyed from a remote expert to a local worker and why view independence is desired by the remote expert. After that, we look at studies that focused on co-located collaboration. More specifically, we look at four different papers, two using HMDs and two using handheld AR, to see what they discovered about collaboration

while being co-located. Next, we discuss one more paper on co-located collaboration, that compared a physical to an AR approach, showing key differences when it came to the position of users towards each other. After that, we look at how the ARPen works, what it's current functionalities are and what previous studies using it have shown so far. Finally, by summarizing what we have learned from the related work, we create the bridge to the main part of this thesis.

2.2 Grounding in Communication

Establishing a common ground is vital for collaboration

Clark and Brennan [1991] argued, that in order to coordinate on content two people need a common ground to work on. They described this common ground as "mutual knowledge, mutual beliefs and mutual assumptions" and in order to coordinate on a process, this common ground had to be updated moment by moment. Within their work, they also looked at "Grounding References" [Clark and Brennan, 1991] and they explained that often "conversations focus on objects and their identities" and that it is crucial, that people are able identify those objects fast and correct. They further explained, that those kind conversations often arise when an "expert is teaching a novice" and that the "purpose of interest" is to establish a common ground, in which the addressees are able to correctly identify a referent. One of the common techniques Clark and Brennan [1991] described are "indicative gestures", that give positive evidence that an object has been identified by "pointing, looking or touching" it. However, these gesture of course can also be used the other way, where the speaker uses them to further specify something. They concluded, that the process of grounding, i.e., establishing a common ground of knowledge, is essential with regards to communication and therefore collaboration and that the techniques used to help with the process of grounding may vary from medium to medium.

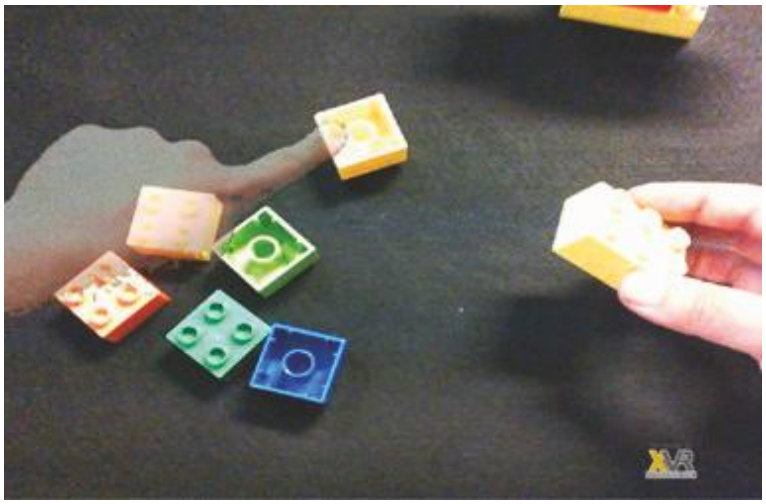

Figure 2.1: Augmented view seen from the workers perspective ([Huang and Alem, 2011]).

2.3 Remote Collaboration

2.3.1 Hand Gestures in Mobile Remote Collaboration

Hand gestures in mobile remote collaboration can help users

Huang and Alem [2011] developed a system, that supports hand gestures in a mobile remote collaboration setting. The camera feed from the workers side was sent to the helpers area, to be displayed on a shared visual space. The expert could then perform gestures above this shared visual space, that were captured by another camera and then send back with the background scenes to the worker. This video feed was visible to the worker via a near-eye display (see Figure 2.1), allowing the expert to perform pointing gestures with their hands, that the worker could then see. The authors had their users perform two tasks using their system, a LEGO™assembly and a PC repair one.

Results

Results of their user study ”confirmed the usability and usefulness” of the system. The participants were able to complete the mentioned tasks ”with quality and satisfaction in a reasonable time”. The answers given in the questionnaire also indicated that users liked the system and

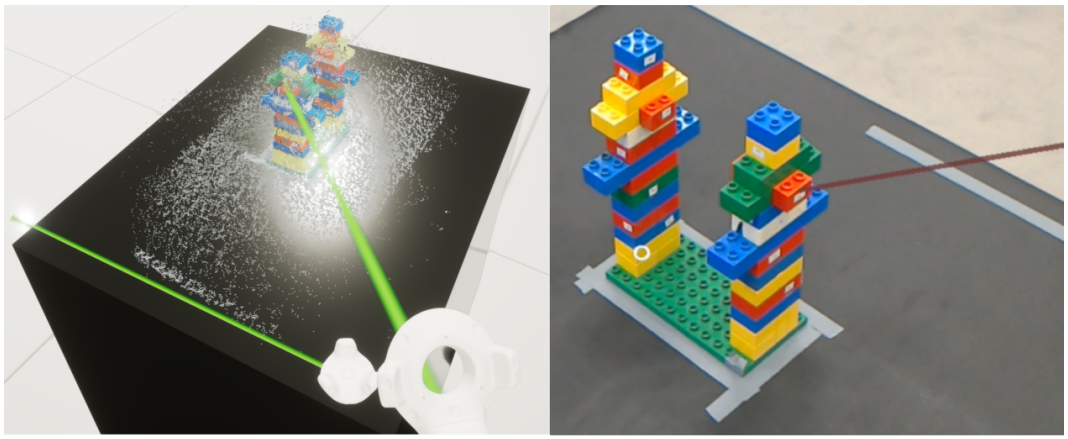

Figure 2.2: VR view seen by the expert using a HTC Vive, highlighting a red block using the laser pointer (left). AR view seen by the local worker using a Microsoft HoloLens, the laser pointer highlighting the red block (right) ([Hoppe et al., 2018]).

found it useful. However, the authors also noticed some problems, mainly that some participants had issues with spatial awareness when using the near-eye display and a certain degree of asynchronicity, caused by the video feed lagging behind.

2.3.2 Multi-User Collaboration on Complex Data in Virtual and Augmented Reality

AR collaboration on complex data

With more modern VR/AR enabling systems, like the Kinect, the HTC Vive[1] and the Microsoft HoloLens[2] Hoppe et al. [2018] were able to build a system, that let a remote expert see a "high fidelity point cloud representation of a real world object in Virtual Reality". The expert could then indicate points of interest via a laser pointer coming from a tracked controller. The laser pointer was shown via AR to the local worker (see Figure 2.2). The authors compared their system to another one that "contained pre-recorded images for the expert and a live video stream, as well as speech communication". They tasked the expert to first locate a specific block and then relay information regarding that block to the local worker, who then had to read a text

[1] https://www.vive.com/de/
[2] https://www.microsoft.com/de-de/hololens

label printed on it to confirm the finding of the block.

Results

While the results for the task time were not significant, they still showed a tendency, where the VR/AR system was slightly faster. Furthermore, the questionnaires indicated that "attractiveness, stimulation and novelty are ranked higher for the VR/AR setup". During the task they also noted that subjects used pointing with the finger to confirm the blocks. They concluded, that the insignificance in their data might be due to issues with the visibility of the point cloud and the calibration between the VR and AR system. They noticed, that a small distance measurement error was present that lead to the beam of the laser pointer being slightly shifted, which might have had a negative impact on the performance of the participants.

2.3.3 The Effect of View Independence in a Collaborative AR System

View independence is preferred for remote collaboration

Another part of research on AR collaboration focuses on view independence as an important factor in remote collaboration, as it allows for a complementary way to work in separated workspaces, while also giving the remote user a greater situational awareness [Fussell et al., 2003]. Previous research has shown that view independent AR systems are preferred for collaboration over a fixed video view [Gauglitz et al., 2014]. Therefore, Tait and Billinghurst investigated how different levels of view independence would affect a remote collaboration task. They build a HMD based system, where the remote user would send information for an object placement task to the local worker via a desktop user interface and where the remote expert could "see a virtual copy of the local users real environment". The authors then created four different degrees of view independence: *Video Only (No Independence)*, *Fixed(No Independence)*, *Freeze (Semi-independent)*, *Independent (Fully-Independent)* and compared them against each other, measuring variables like accuracy and time. Additionally, they asked their users to fill out the "System Usability Scale Questionnaire", to rank the system based on "preference, confidence, perceived speed and perceived accuracy". Fur-

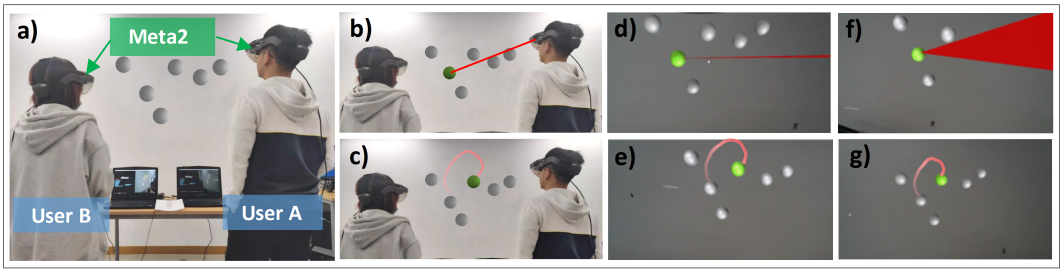

Figure 2.3: "(a) A picture of the experimental setup; Two users are locating a target using (b) Pointing Line (PL) cues and (c) Moving Track (MT) cues; The view of User A who is locating a target using (f) PL and (g) MT; The view of User B who is now looking the target that User A has selected using (d) PL and (e) MT" ([Chen et al., 2021]).

ther analysis was also done on the recorded video and comments made by the users.

Results

The results of their study showed, that overall the *Independent* view was the fastest, but the accuracy between the different setups showed no significant differences. Furthermore, the *Independent* view was ranked highest by the remote users and they felt it offered the highest accuracy to them. It also scored highest in the user rankings with regards to the quality of collaboration. They also noticed, that less adjustment instructions where needed when using the *Independent* view. So overall, they showed that a higher view independence resulted in better performance and is generally preferred by the users.

2.4 Co-located Collaboration

2.4.1 Effect of Visual Cues on Pointing Tasks in Co-located Augmented Reality Collaboration

The effect of visual cues on pointing tasks

After having looked at remote collaboration, we now take a closer look at co-located collaboration. Chen et al. [2021] analyzed the effect of different visual cues on a pointing task. To be more precise, they compared a "Pointing Line (PL)" technique to a "Moving Track (MT)" one (see Figure

2.3). The PL is basically a gaze ray, that is cast from the user's head to an AR object. The MT on the other hand shows "a trail that follows the cursor" with a limited length and a gradual fading that gets weaker at the end of the trail. A highlighting of the currently selected object was present in both techniques. They compared these two setups in two different object states ("*Static* and *Dynamic*"), as well as three different "*Density*" states, ranging from "*Low* (6 objects with no occlusions)" over "*Medium* (12 objects with slight occlusions)" to "*High* (18 objects with severe occlusion)" [Chen et al., 2021]. They measured the time it took the users to complete the task, their success rate, had them fill out a "Social Presence" and a "Usability" questionnaire and performed a small interview at the end of the study.

Results

The results of the study showed, that the users took significantly less time using the PL in the *Dynamic* setting, in the *Medium* setting and the combined *Dynamic* $\times$ *Medium* one. In other combinations PL was still faster, but the differences were not significant. Both systems displayed a high accuracy rate for all conditions and no significant interaction effects were found. However, when they looked at the descriptive data, they saw that PL achieved higher rates than MT and "for mean results of accuracy rate, PL performed better than MT in static trials, while MT performed better in dynamic trials". They further found out, that *Technique* had a significant effect on accuracy rating and the performed post-hoc test revealed that participants with PL got higher accuracy rate than the MT participants. The results of the questionnaires and interview showed that for the "Usability" section the users preferred the PL over the MT, but for the "Social Presence" and "User Preference" ones they instead preferred MT over the PL.

2.4.2 See-through Techniques for Referential Awareness in Collaborative Virtual Reality

See-through techniques to help with occlusion

Argelaguet et al. [2011] investigated two techniques, both focused around resolving the occlusion problem. The techniques define a cutting volume and objects that fall inside that volume are either fully removed or displayed semi-

transparently. Their system used a projection based two-user VR setup using LC-shutter technology, therefore both users had to wear trackable shutter glasses. To test their system, they designed a pointing task performed between the two users and compared the techniques to a baseline, where user had to walk around to obtain view of otherwise occluded objects. The task consisted of one user, the "presenter", who had to point at certain objects inside a model and a second user, the "observer", who had to identify and locate the object. The tasks were put into blocks, consisting of 5 trials each, to see if learning effects were present and the authors further had the participants score the techniques with regards to "spatial understanding", "collaboration" and "comfort".

Results

The authors found a significant decrease of "discovery time" over the blocks, however no significant differences were found regarding *technique*. They concluded, that this mainly came down to users following "the presenter to ensure a similar point of view", whereas they did not have to move around when using the techniques, something that also showed significantly in the data for "covered distance". The data for "retrieval time" showed the same results as the ones for "discovery time". The results of the scoring showed a significant preference for the two techniques over the baseline, with regards to the comfort and the collaboration. However, those results did not show up for the spatial understanding. They further noted that the techniques reduced "the number of cases in which user needed to get very close or bump into each other".

2.4.3 CollabAR - Investigating the Mediating Role of Mobile AR Interfaces on Co-Located Group Collaboration

Effects of mobile AR interfaces on co-located group collaboration

Wells and Houben [2020] researched how a mobile AR interface might effect co-located group collaboration. They build a web application, which allowed the users to render a 3D object by pointing the camera at a marker. After the initial rendering, users were able to manipulate the object in a turn-based fashion, where only one user was able to ma-

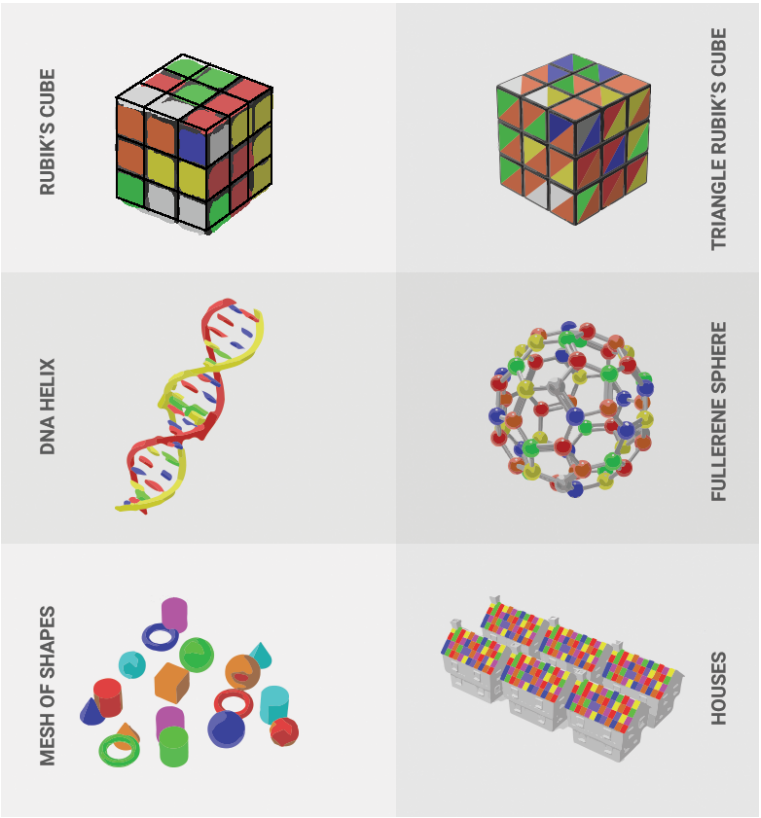

Figure 2.4: The models created by Wells and Houben [2020].

nipulate it at a given time, while the others were locked out of doing so. This lockout was indicated by a border around the view seen on the device. The manipulation information was then send to a database, which allowed the corresponding update to happen on other devices as well. The authors created different models, that represented abstract examples of "complex virtual data" typically found in different AR domains. The models were designed in a way, such that they had different levels of complexity (see Figure 2.4). For their study they had the users perform three tasks for each difficulty level, two inspection and one comparing task, e.g., a question would be "count how many red tiles are on this cube".

The main takeaway from their study is, that there were "ex- Results

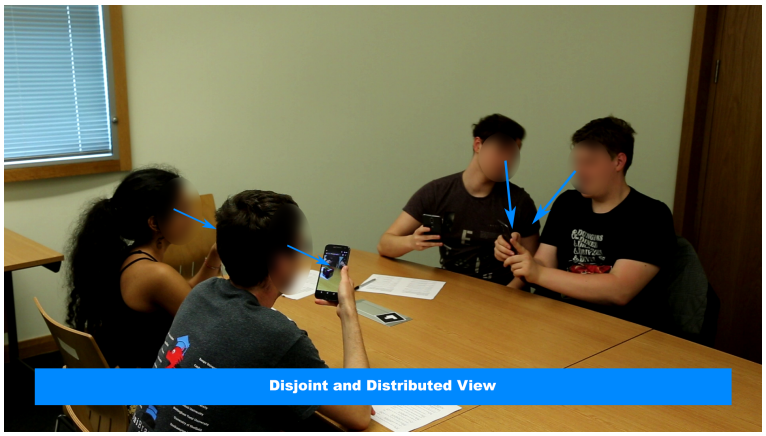

Figure 2.5: Depiction of the "Disjoint and Distributed View" ([Wells and Houben, 2020]).

treme amounts of context switches", where a context switch is "an instance in which the participant looks away from the virtual content". They also experienced different collaboration styles, two of which included users looking at one singular device instead of looking at the scene through their own one (see Figure 2.5). They further observed an overall high cognitive load, physical load as well as overall effort. The authors provided some design recommendations, including improvements on visual guidance, e.g., offering various views and the addition of better awareness cues to indicate where users are looking at and where the focus of attention should currently be.

2.4.4 Collaborative Programming Problem-solving in Augmented Reality

Using AR to investigate how it affects collaborative problem-solving (CPS)

Chung et al. [2021] researched how AR might affect a collaborative problem-solving task in the field of programming. They designed an AR-enabled app and a CPS programming task, where two people had to collaboratively write a computer program. Their app "could present the same content in AR and non-AR visualizations". This allowed them to compare the two visualization techniques in a within-subject experiment. They further used a hidden

Figure 2.6: Participant looking at his partners screen ([Chung et al., 2021]).

Markov model to analyze the evaluation of the code and communications topics quantitatively. Furthermore, they used a semi-structured interview and a questionnaire to asses "the user's attitudes and experience after the experiment".

Results

Chung et al. experiment showed, that the implementations that involved AR shared a higher level of similarity to the standard solution, indicating that "participants solved the task more effectively in AR version than in the non-AR version". The results obtained from the hidden Markov model showed, that in the AR setting users were able to edit their code much more steadily and productively than in the non-AR one. The model also indicated, that AR improved the efficiency of communication between the users. Furthermore, the results of the questionnaire showed a more positive attitude towards the AR content and a higher engagement. However, "due to the state of networking", their app sometimes had delays while synchronizing the code of the participants. This raised some concerns with regards to usability, but the authors firmly believe that these issue could be solved by better hardware. Finally in the interview 19 out of 24 participants stated that they preferred the AR setting, despite the aforementioned concerns.

2.5 Physicality As an Anchor for Coordination: Examining Collocated Collaboration in Physical and Mobile Augmented Reality Settings

Comparing a physical to a mobile AR approach

Poretski et al. [2021] compared a pure physical approach to a mobile AR one with regards to a co-designing and co-building of a structure tasks. In both settings the users had to place blocks, either physical or virtual ones, to build a *city hall*, a *house* or a *castle*. They gathered data on the user experience, their subjective assessment of collaboration, the perception of structure quality and work creativity. Additionally, they used their video recordings for further analysis regarding "body positions, gestures, and position of the participants" and the way they "talked, moved, and interacted with each other".

Results

Their results showed, that the participants chose between three different positions when working on the task (*Near, Adjacent and Opposing*). They further noticed, that for the AR condition the participants spend a pretty balanced amount in each position. However, when compared to the physical approach, the time spend in the *Near* position was significantly higher in the AR condition. They concluded, that this was due to the necessity to better understand "each other's perspective on the artifact of work". They also noticed a "rich deictic behavior in the physical condition", which contrasted the verbal character of the AR condition. They believed, that the primary reason for these results are the limitations AR places on the "participants' mutual awareness". The participants also rated the AR system similar or higher than the physical one with regards to to collaboration, user experience, creativity of work and quality of output.

2.6 The ARPen and its App

ARPen allows for mid-air object manipulation using a bimanual AR system

Wacker et al. [2019] presented the bimanual AR system, combining a pen and smartphone, on which this thesis is

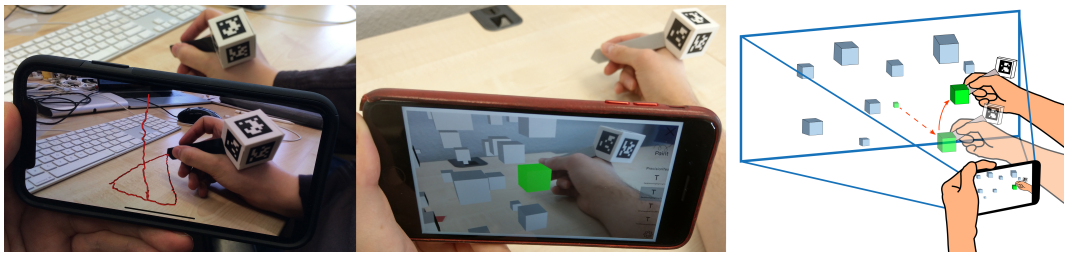

Figure 2.7: Mid-air sketching using the smartphone app and the ARPen (left). Image from the mid-air selection technique study (middle). Preferred technique for the translations, which was the *pen ray pickup* (right) ([Wacker et al., 2019]).

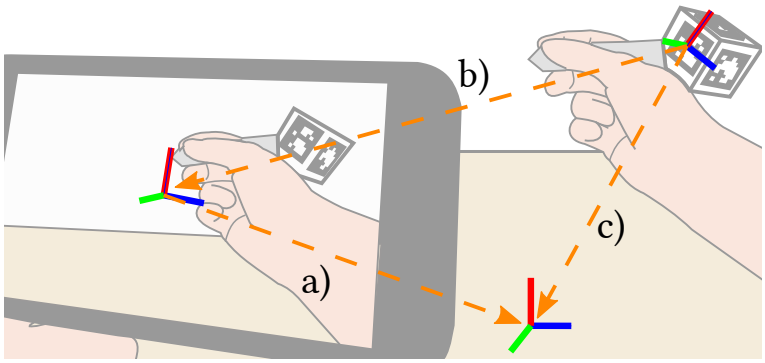

Figure 2.8: a) Camera's position relative to the surface. b) arUco tracking the marker relative to the camera. c) By combining these two calculations the position of the marker relative to the surface can be computed ([Wacker et al., 2019]).

build. They used ARKit[3] to create an iOS app, that allowed the tracking of an arUco marker to determine the mid-air position of the pen (see Figure 2.8). They then used SceneKit[4] to render a ball at the tip of the pen and to implement different functionalities, e.g., allowing the user to draw a path mid-air (see Figure 2.7, left) or create a cube. They conducted a pre-study to determine what the preferred orientation and size of the phone would be going forward. Their results led them to use a "*big* iPhone and the *pinkie* grasp in the *camera right* orientation" for their two other studies, one of which focused on different selec-

[3]https://developer.apple.com/documentation/arkit
[4]https://developer.apple.com/documentation/scenekit

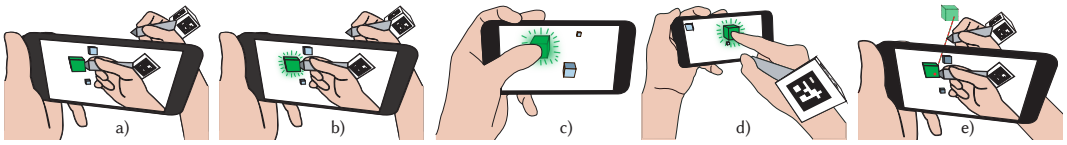

Figure 2.9: a) Without highlight, b) With highlight, c) One-handed, d) Two-handed, e) Pen ray ([Wacker et al., 2019]).

tion techniques, that we take a closer look at. The authors compared five different selection techniques and measured their "success rate, selection time, deviation from the target, and the size of the object on the screen during the selection". Those five selection techniques were: "*Pen Selection Without Highlighting, Pen Selection With Highlighting, One-handed Touch Selection, Two-handed Touch Selection and Pen Ray Selection*" (see Figure 2.9).

Results

The study on the selection techniques showed, that *Pen ray* seemed to be the best solution for a selection task. It had the highest *success* rating, together with a fast *selection time*. Additionally the *projected size* was also small, which indicates that the device was not moved a lot to select a given target. *Pen ray* also ranked highest together with *two-handed touch*, with regards to which techniques the users preferred. *With highlight* also had a good *success* rate, but due to participants having to adjust their position to find the correct depth the *selection time* was the slowest. Both touch techniques performed well and *Without Highlight* had the worst performance regarding *success* and was also the least preferred setting.

2.7 Summary

Grounding as essential part that is needed for collaboration

We started, by looking at why a common ground of knowledge is essential for collaboration and how it is established in everyday communication using a process Clark and Brennan [1991] called grounding. We saw, that one important technique that can be used during the grounding process is the use of what they called "indicative gestures", today more often referred to as deictic gestures.

Remote collaboration

After that, we looked at research done on remote collaboration, where an expert would guide a layman through a task or a problem-solving process. Here, we saw that AR can be helpful when working on those tasks and that the AR systems were generally liked by the participants of their respective studies ([Huang and Alem, 2011],[Hoppe et al., 2018]). However, we also observed a few issues when using AR for collaboration, mainly synchronization issues of the video feed and the calibration between the systems. We also saw, that users performed faster using an independent view and that this was also the most liked technique when compared to other levels of view independence [Tait and Billinghurst, 2015].

Co-located collaboration

We then looked at co-located collaboration. Chen et al. [2021] showed, that with a basic ray representation, they called it "Pointing Line", users performed faster and more accurate. Their users also rated the "Pointing Line" higher with regards to *Usability*, but lower for *Social Presence* and *User Preference* when compared to their other technique called "Moving Track". Next, we looked at two see-through techniques researched by Argelaguet et al. [2011]. They showed, that their techniques reduced the amount of movement required to be performed by the participants to find the object the presenter was pointing at. And although there were no significant difference in the time it took the user to discover and locate the object, they concluded that this was due to the user following the presenter. This is something we believe is not an accurate representation of a real scenario, where either physical space or social aspects often can be limiting factors of how close one person can or will follow another. Argelaguet et al. [2011] also noted, that the techniques reduced "the number of cases in which user needed to get very close or bump into each other", which shows that those techniques could help to keep a certain socially appropriate distance, e.g., between a worker and his superior. The results also showed a significant preference for the two techniques over the baseline regarding comfort and collaboration. After this, we looked at two hand-held AR projects. The first one focused on the meditating role of AR interfaces [Wells and Houben, 2020]. Here, results indicated that there were "extreme amounts of context switches", that sometimes users tended to look to-

gether at a singular device, instead of each looking at their own one and that in general users felt a high cognitive and physical load as well as overall effort. They also explained various improvement recommendations, such as offering various views or better awareness cues. Next, Chung et al. [2021] research on collaborative programming showed, that participants once again performed better with the AR system compared to the non-AR one. Their participants were able to perform faster, more steadily and more productively when using the AR system. These results were also reflected in the answers given in the questionnaires. However, similar to the remote setting, there were some concerns and issues regarding the synchronization of the system and similar to Wells and Houben [2020] users sometimes tended to look at their partner's device instead of looking at their own to check the AR scene. After that, we discussed the research done by Poretski et al. [2021], who showed that users in AR condition spend more near each other when compared to the non-AR one and had an overall more balanced positioning to each other. They believed, that users tended to do this in order to be able to understand the perspective of the other person. Additionally, they saw a decline in deictic gestures/behavior in the AR condition, as AR places limitations on the mututal awareness between participants. Finally, we looked at the cornerstone of this thesis, the ARPen. A system designed by Wacker et al. [2019], that combined a smartphone and a trackable pen to create a bimanual AR system for mid-air object manipulation. Their research suggests, that for a selection task a ray selection is the preferred choice, together with highlighting the pointed at object, which coincides with Chen et al. [2021] "Pointing Line", as well as with Hoppe et al. [2018] use of an AR laser-pointer to highlight a physical object remotely.

Transition to the main part of the thesis

Overall, we learned about "grounding", a process essential to being able to collaborate together. We saw, that there are clear indications that AR can be helpful and exciting to use in a collaborative manner, be it remotely or co-located. However, we also described some of the more common issues, such as synchronization, occlusion and the tendency of users to look away towards another person's device or get close to a person to achieve a similar point of view in

a hand-held co-located scenario, in order to reduce the impact of the mutual awareness problem. Additionally, we saw at least one example where AR collaboration was perceived as a mentally and physically taxing task. We also saw, that AR limits the possibilities for deictic gesture to be conveyed and perceived. In this thesis we wanted to find out, how we can bring collaboration to the ARPen app and how we can help users to understand pointing operations (deictic gestures) with the ARPen, in order to alleviate the problems caused by a potential lack of mutual awareness and therefore the lack of a common ground of knowledge to collaborate upon.

Chapter 3

ShARePen

Overview

In the following chapter we discuss our motivation and intentions for developing and investigating pointing operations with the ARPen. We explain our choice for the techniques used in the user study and cover core parts of the implementation, such as synchronization, the two plugins developed over the course of this thesis and the sharing of video data as part of one of the four techniques. Finally, we briefly touch upon the feedback given after a test run of the user study with one person and the actions taken after this feedback was considered.

3.1 Motivation and Intentions

Recap of the problems

In the previous chapter 2 “Related Work” we showed, that using AR for collaborative tasks can be helpful and often outperforms more traditional approaches. However, previous systems often suffered from technological issues, such as problems with the synchronization or with real-time video data transfer. We also saw indications, that users tried to solve awareness problems by shifting their position, such that it nearly matches that of another participant or even look away from their own to their peer’s device and that AR often limits the natural possibilities to perform and perceive deictic gestures.

Motivation and intentions

For this thesis, we wanted extended the functionality of the ARPen app, by allowing two users to look at a single shared AR experience. We emphasized on creating a stable and highly synchronous system, that would also offer a good performance with regards to sharing video data. Using this system, we wanted to investigate if a basic awareness cue would be enough to understand pointing operations with the ARPen or if additional visualization techniques could improve the overall user performance and experience with the system, as having good awareness cues is essential for collaboration. Additionally, we also wanted to see if different user to user positions, as discussed by Poretski et al. [2021], would have an impact on the performance and the preferences of the participants. Overall, we wanted to investigate how users can comprehend pointing operations performed with the ARPen and how we potentially can aid them, so that referential awareness is not an issue and a common ground of knowledge, that is essential for collaboration, can be established fast and correctly.

3.2 Discussion of the Pointing Method and the Four Visualization Techniques

Ray cast approach for pointing

As Hartmann and Vogel [2018], Mifsud et al. [2022], Wacker et al. [2019] and others have previously shown, the ray casting approach is the preferred choice for pointing and selection tasks. Therefore, we decided to also use this technique for our pointing operation. We would cast a ray from the camera through the pen tip into the scene and see if it intersects with a virtual object. Based on this decision and the research presented in chapter 2 “Related Work”, we propose the following four techniques for the comparison in this thesis:

1. *Baseline*.
2. *Ray*.
3. *Opacity*.
4. *Video*.

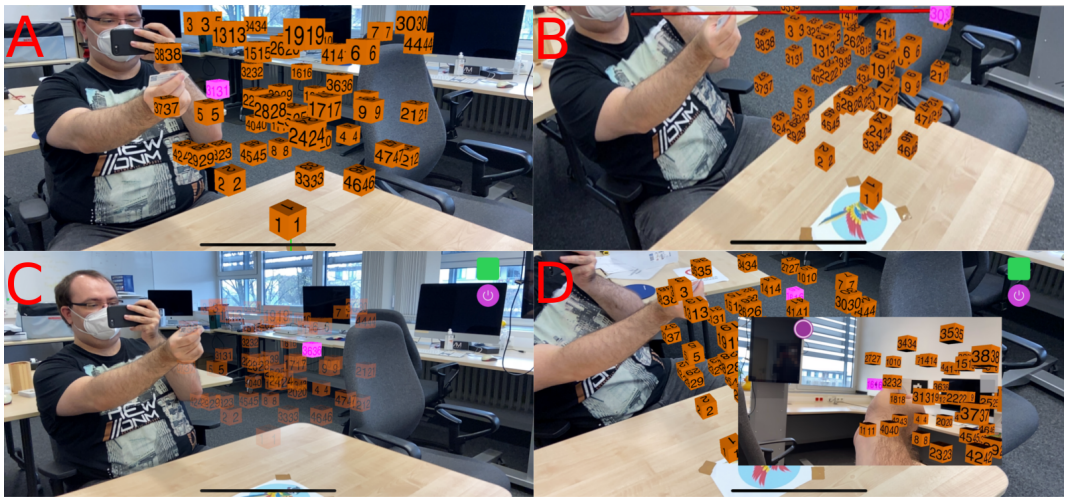

Figure 3.1: A) Baseline, **B)** Ray, **C)** Opacity, **D)** Video.

Explaining the choice of techniques

Wacker et al. [2019] showed, that users preferred a highlighting of the selected/pointed at object over a setting without additional highlighting. Therefore, we decided to choose this as our *Baseline* (see Figure 3.1, A). We also could have chosen a version with no highlighting as the baseline. However, we believe that it would be close to impossible to correctly understand a gesture performed through a ray cast without any information except the position of the pen. Since users already struggled in the selection task when they handled the pen themselves and did not have any highlighting [Wacker et al., 2019], we firmly believe this would be even harder, if the person is not handling the pen. This basic highlighting would serve as a standalone condition, but it would also be part of every other condition, as we went for a study approach where users were not forced to use a specific aid presented to them, but rather could opt-in to use it if they wanted to. Since a ray cast would be used for the pointing, we propose a rendered version of this ray as the next technique (see Figure 3.1, B). We felt, that this would give the users a better way to follow the motion of the pointing operation(s) and we have seen good results using this visualization in the past, e.g., Hoppe et al. [2018] or Chen et al. [2021]. Since Argelaguet et al. [2011] had good success with their see-through techniques, that helped fighting occlusion in a pointing task and occlusion being an issue in general for AR [Kruijff et al., 2010], we

decided to adopt their transparency based approach (see Figure 3.1, C). By reducing the opacity of all objects currently not pointed at, we thought we can help to enhance the perception of the pointed at object, while also helping to reduce problems caused by occlusion. Lastly, we suggest a live video feed condition (see Figure 3.1, D). Based on the research by Wells and Houben [2020] and Chung et al. [2021], we think that there is a good indication that people tend to look away from their own, to another person's device. This is further supported by the findings of Poretski et al. [2021], that showed a higher time spend close to each other in an AR setting compared to a non-AR one. We also saw, that an independent view is preferred in a remote scenario [Tait and Billinghurst, 2015]. Therefore, we thought it might be beneficial to offer the user the possibility to switch to another person's point of view (PoV) on demand, basically inverting the idea of an independent view from the remote setting. Instead of offering this independent view, which already exists in a shared setting since every person has their own device and therefore their own unique view, we would instead offer the user a semi-dependent approach via a picture-in-picture (PiP) view from another device's perspective. This should allow the users to easily see the pointed at object, even if it is occluded from their PoV, while also giving them the option to hopefully understand the pointing not only from their own, but also another person's perspective.

3.3 Implementation

The project can be downloaded from Oliver (RWTH i10 fileserver). Please read the "README SHAREPEN.md" file contained in the "SharedAR Container" folder on how to setup the App container, such that a device can load all required files.

ShARePen.zip[a]

[a] http://hci.rwth-aachen.de/public/users/bruna/ShARePen.zip

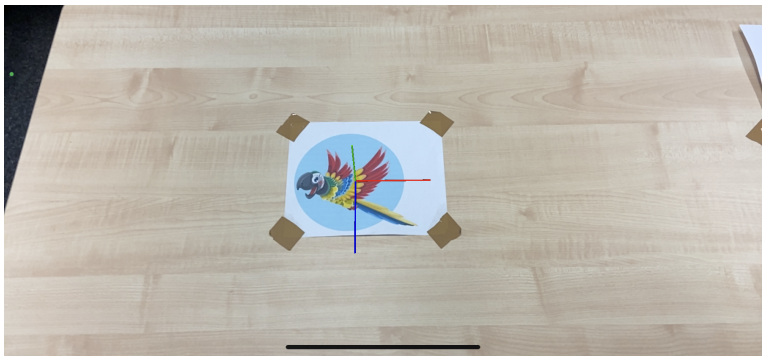

Figure 3.2: A 3D-coordinate system is displayed when the image is tracked, indicating that the world position has been reset and the system should run synchronous again. Original image of the parrot taken from Pixabay.

3.3.1 Synchronizing the Content Between the Devices

Using the Multipeer Connectivity framework to generate a shared AR session

We used Apple's Multipeer Connectivity[1] framework to create a shared `MCSession`, which both devices had to join. Using this connection, we share content, e.g., `SCNNodes` and commands, e.g., via `Strings` between the two devices, allowing us to manipulate the scene and session in real-time. It is important to note, that all data transferred between the devices needs to conform to the *Codable* and *Decodable* protocols and that future extensions need to keep this in mind.

Synchronization via an ARImageAnchor

To synchronize the `ARWorldMap` and therefore the content displayed between the two devices we used an `ARImageAnchor` (see Figure 3.2). Every time this *ARImageAnchor* switches from being not tracked to being tracked, we reset the origin of the *ARWorldMap* to match the transformation of the anchor. This allows us to have the exact same coordinate system on both devices, as long as they both perform this synchronization at the beginning and whenever something gets asynchronous afterwards. This means, we do not have to do additional transformations to calculate properties, like the position of objects, on the de-

[1] https://developer.apple.com/documentation/multipeerconnectivity

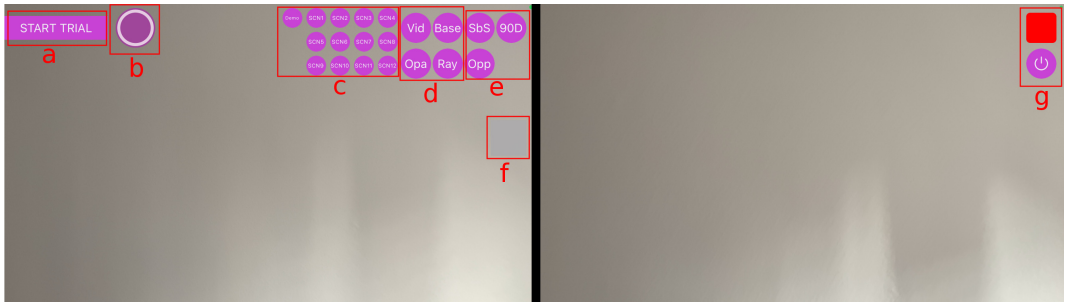

Figure 3.3: Left: SharedARPlugin, a) start trial, b) start/stop measurement during trial, c) switch scenes, d) switch modes, e) switch position (for logging only), f) indicator for measurement during trial. **Right:** SpectatorSharedARPlugin, g) indicator and on/off button for an additional visualization aid.

vices. We achieved the necessary accuracy required for the user study with just this singular anchor point, even when the anchor was not permanently being tracked. However, this of course could and should be extended in the future especially if the AR space is larger in size, as moving further away from the anchor would increase any existing unwanted offset.

3.3.2 The SharedARPlugin and SpectatorSharedARPlugin

For the purpose of the user study, we split the required functionalities into two plugins:

1. The `SharedARPlugin`.
2. The `SpectatorSharedARPlugin`.

Since both plugins are similar by nature, we compare them to each other, highlighting the main similarities and the key differences.

Each plugin uses their own UI layout

Both plugins use different user interface (UI) layouts (see Figure 3.3), as the person that handles the ARPen needs to be able to switch modes, the positional setting and the scene during runtime, while also being able to start a trial

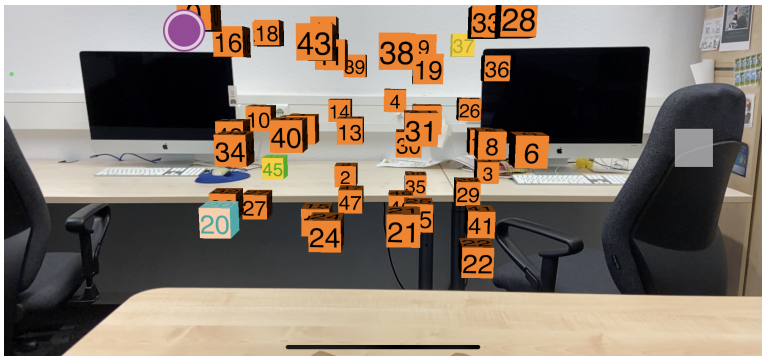

Figure 3.4: Cube are colored during trial for the presenter. 1st (blue), 2nd (yellow) and 3rd (green).

and indicate when he arrives at the correct object in such a trial. The participant on the other hand only needs a very minimalistic UI, as he does not perform any live changes on the AR scene, except activating or deactivating an additional visualization.

Live switching of scenes, modes and user positions

We stored CSV-files containing all the necessary positional information for the different scenes used in the user study on the devices beforehand. Therefore, we could easily load them during runtime, without having to restart the app with different settings each time. We also did this, so that we would be familiar with the scenes during the user study, as we needed to perform the correct pointing operations consistently. Additionally, the `currentMode` and the `userPosition` could also be set during runtime via simple button presses (see Figure 3.3, c and d).

Additional guidance through coloring

In addition to being familiar with the scenes, we also build a color based guidance (see Figure 3.4) to assist us during the user study, so that we would be able to perform with the same level of quality for each participant. The corresponding data required for this coloring is stored in a JSON-file. During the user study, we could then activate it by tapping the screen, as long as we are using the `SharedARPlugin`, to color the objects for the current trial. This coloring would persist through the trial in all modes except the video one, because there the other person should of course not be able to simply see the correct objects in the video feed with the

coloring applied.

3.3.3 Live Video Feed

For the live video feed, we followed Apple's AR Streaming example[2]. However, we noticed severe issues with the frame rate of the video, which would occur after a low amount of time when using this setting. To fix this problem, we set up a Raspberry Pi as a makeshift router, because these issues were seemingly caused by the 5 GHz frequency WLAN band and since the Raspberry only has a 2.4 GHz band the problems disappeared, once the devices started communicating over this lower frequency band. We have no definitive explanation as to why the problems occurred in the first place and why using a lower frequency band fixed them, but we believe it might be due to some *Quality of Service* (QoS) being applied only at the higher frequency band, which would interfere with the video transfer. Something similar also happens with Apple's build in screen share between devices, where the solution is the same. This also gave us the initial idea to use this Raspberry Pi approach.

3.4 Pre-study

Running a single person pre-study

We ran a single user pre-study to test the overall stability of the application and to gather initial feedback on what could be improved. During this initial test run, no stability issues occurred. The feedback was mainly positive, however some points of criticism were uttered and noticed. Firstly, the study took too long, both from the perspective of the participant and ourselves, as in the end we simply ran out of battery on one device. Furthermore, the user complained that having to opt-out of the modes was tiresome and stressful. Even though we only recruited a single user, we still felt that those points had a certain degree of validity, so we made some adjustments.

[2] https://developer.apple.com/documentation/arkit/streaming_an_ar_experience

3.4.1 Adjustments after the Pre-study

Based on the feedback and data gathered from the pre-study, we lowered the overall number of trials each user had to perform, so that we would not run out of battery and to keep the fatigue for the participant at a lower level. We also changed the opt-out approach to the opt-in one, meaning that a participant had to turn on the additional visualizations, i.e., *Ray*, *Opacity* and *Video*, during the trials, giving us a higher external validity and therefore allowing us to work with a representation of the software, that more closely fits the everyday usage.

Chapter 4

Evaluation

Overview

In the following chapter we discuss our choice to use a two task setup. We detail the overall design of the user study, followed by giving the hypotheses and the reasoning of why we chose them. After that, we take a look at the results, both quantitatively and qualitatively for each of the two tasks. Then, we discuss those results with regards to the aforementioned hypotheses and what potential design recommendations can be derived from them.

4.1 Recognition and Relocation Task

We decided to use a two task setup for the user study, since we believe that **"understanding"** a pointing operation means two things. First, the user should be able to recognize the pointed at object and should be able to give verbal feedback once he has recognized an ongoing pointing operation and second, the user should be able to relocate the object in the scene, showing that he indeed did not only recognize that a pointing was happening, but also that he understood the position of the object inside the scene. Therefore, we used a two task setup, where first a sequence of three objects was shown to the user, waiting for verbal confirmation that the position/number of the cube had been recognized and sufficiently memorized at each object

(recognition), followed by the relocation part, where the user then had to tap the shown cubes via touch-input in the correct order, showing that he did in fact understood the pointing operation.

4.2 Design of the User Study

Overview

In this section we discuss the overall design of the user study. We explain the different user to user positions, the environment in which the study took place and what devices were used. Additionally, we shortly describe how the scenes were constructed and how the conditions and scenes were counterbalanced to prevent potential bias and minimize learning effects. We then give a detailed report on the procedure of the study and what measurements were taken during the study.

4.2.1 User to User Positions

Based on the results by Poretski et al. [2021], we also wanted to investigate if different user positions would affect the tasks with regards to performance or the preferred choice of the users for the visualizations. Therefore, the study was conducted in three different user to user positions (*Position)*: *Opposite*, *90-Degree* and *Side-by-side* (see Figure 4.1), giving us a more controlled approach to what the participants naturally did during Poretski et al. [2021] user study.

4.2.2 Environment and Devices

The study took place at a table in the HiWi room of the i10 chair at RWTH-Aachen University. The marker for the `ARImageAnchor` used for synchronization (see section 3.2 “Discussion of the Pointing Method and the Four Visualization Techniques”) was placed in way, such that the AR objects would approximately be equally far away from the

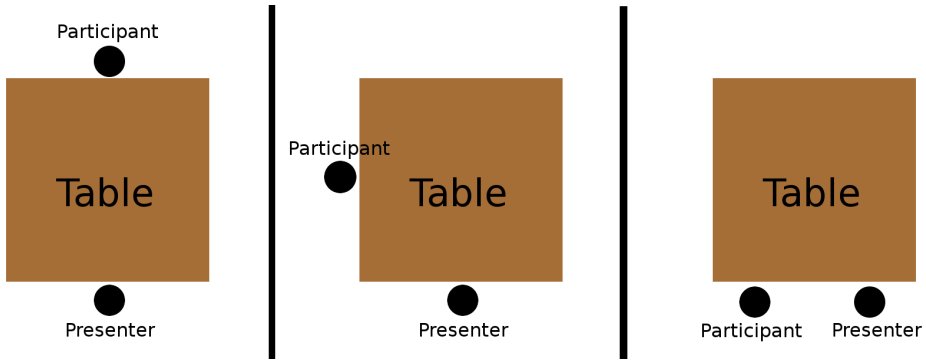

Figure 4.1: Left: Opposite, **Middle:** 90-Degree, **Right:** Side-by-side.

user in every positional setup. A chair was provided for every position, however users were not required to keep sitting for the whole study and were allowed to stand up during trials, if they felt it would give them an advantageous perspective. Each participant was given an iPhone 11 and the conductor, henceforth also referred to as presenter, used an iPhone SE 2nd generation. A Raspberry Pi 3 Model B V1.2 was used as the makeshift router to fix the video lag. We did not use a Bluetooth ARPen, instead we settled with the cardboard variant, as it should not affect the pointing in a meaningful way and it helped to reduce the computational complexity required to track the pen, therefore allowing the device to keep more resources for the data transfers and video rendering. Additionally, since we performed every pointing operation ourselves and performed extensive testing during the implementation process, we felt confident and familiar with the handling of the pen, such that we could keep the singular marker always in our FoV with ease.

4.2.3 Scenes Construction

Twelve scenes were constructed for the study, such that every combination of visualization and position could have a unique scene for each user, which would be repeated after the twelfth participant. Each scene consisted of 48 cubes, numbered from 0 to 47 (see Figure 4.2). These cubes were always arranged in a 4x3x4 (Width x Height x Depth) grid.

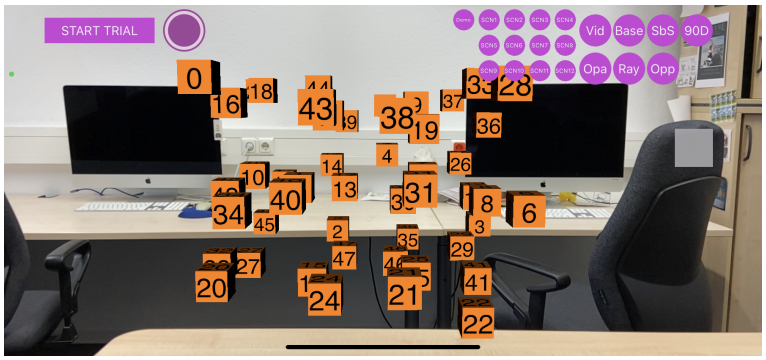

Figure 4.2: Example of one scene used in the user study.

In each scene the offset between the cubes was slightly different, creating similar but yet unique scenes. Additionally, the offset range was selected in a way, that created dense scenes where the occlusion of objects by others would occur frequently, giving us complex and challenging arrangements.

4.2.4 Study Procedure

Counterbalancing conditions

The study focused on comparing different visualization techniques in different user to user positions. Therefore, we combined a 4x4 Latin square for the techniques, a 3x3 Latin square for the positions and a 12x12 Latin square for the scenes to counterbalance any learning effects and eliminate a potential bias (see Appendix A). Additionally, all objects pointed at were also randomized before the user study and placed in the JSON-file as explained in 3.3.2 “The SharedARPlugin and SpectatorSharedARPlugin”, however it was made sure that after 8, 16 and 24 users every object would have been visited equally often, further removing any bias that might exist otherwise. We also made sure, that no number would be repeated in the same scene for the same user, e.g., if user ’0’ had seen the number ’10’ as part of one sequence, the ’10’ would not appear again until after the scene, and therefore the mode and/or position changed and the ’10’ would be in a new position.

Procedure

On arrival, the participant filled out the consent form

(see Appendix A), after which he was handed the iPhone 11 with the app running and the correct plugin (`SpectatorSharedARPlugin`) selected and instructed to keep it in landscape mode throughout the study. We then explained, how the synchronization of the devices works and that if objects would shift or move during the study, the user would have to re-synchronize the device, by pointing the camera at the parrot image again. After that, the user was introduced to the cubes and the overall chaining of the trials, explaining that there would always be a recognition task, followed by a relocation one and that each trial would consists of a sequence of three cubes showed one after another, waiting for verbal confirmation that the cube had been recognized and memorized. We then ran a demo trial, where the user could familiarize himself with the tasks. After this initial run, we did further demos showcasing each additional visualization (*Mode*), such that the user could also familiarize with them. The participant was then informed to strife for the best results in both tasks, meaning that while time would be an important factor, so would be the correctness of the selected cubes during the relocation. They were also told, that it was okay to stand up if they felt like it, but that they were not allowed to move around to another position, slight leaning however was explicitly allowed as well. We also made sure, that the participants understood that there would be six trials for every combination of *Mode* and *Position*, that they were not required to activate an additional visualization if they did not want to and that they should keep any verbal feedback until after the trials were completed, as to not cause distraction mid trial. As the presenter, we also made sure to start every trial with the pen in a similar position, namely below the middle point of the bottom cube row. After each set of trials, the participant was also asked to fill out a question regarding the *Perceived Stress* and another one regarding their *Perceived Performance*. After all *Mode*s were completed in a single *Position*, the participant was further asked to rank each *Mode* for both of the tasks separately. Finally, after all trials were completed, the user ranked the *Mode*s one last time without specific regards to either *Position* or task and they were asked a handful of questions in an interview style, e.g., if the tasks were appropriately difficult or if there were any synchronization or video issues. Overall, each

participant was shown 216 objects (18 cubes × 3 *Position*s × 4 *Mode*s) over the course of 6 trials per *Position* × *Mode* combination, therefore each user performed 72 trials in each of the two tasks, for a total of 144 trials.

4.2.5 Measurements

We recorded the *Recognition Time*, as the time it took the user to recognize and memorize the cubes without the time it took the presenter to move from cube to cube (manual start/stop performed by the presenter). We also recorded the *Relocation Time*, as the time it took users to complete each relocation task. The *Relocation Time* was started once the participant pressed the start button for the task and was automatically stoped once he finished. For both tasks, we logged the summed up *Absolute Translation* (in meters) and the summed up *Absolute Rotation* (in degrees) in X-, Y- and Z-direction, as the change of movement between frames with a frequency of 30Hz. For the recognition task, we also stored the *Help Time* (time an additional visualization was active), the *Button Presses* (amount of times the button to activate/deactivate an additional visualization was pressed) and the *Trial Time* (the *Recognition Time* plus the time it took the presenter) to get the *Help Time Percent* an additional visualization was active (*Help Time* ÷ *Trial Time*). During the relocation task, we further recorded if a trial was successful and if not, how many mistakes were made. A trial was only considered successful if none of the three cubes was selected wrongfully. We also collected the subjective ratings for *Perceived Stress* and *Perceived Performance* for each *Position* × *Mode* × *Task* on a 5-point Likert-Scale, a ranking of each *Mode* in each *Position* × *Task* and a final ranking of the techniques, without regards to *Task* or *Position*. For all of these subjective ratings and rankings, a higher value would correspond to a better result, e.g., a '5' on the *Perceived Stress* scale would be equivalent to "Not stressed at all" and a '5' on the *Perceived Performance* scale would be equal to "Very good" and a '4' on the ranking would mean the highest number of points given. For the subjective rankings, every amount of points (1-4) could also only be given once for each setting. The full questionnaire can be found

in Appendix A.

4.3 Hypotheses

As we saw in the previous chapters, having good and clear awareness cues is one of the key factors when it comes to recognizing pointing operations in AR. We introduced four different techniques, a baseline and three additional visualizations to investigate which would provide the best awareness cues. We expected participants to be able to faster recognize objects with the additional visual cues provided by the three non baseline techniques. We also assumed, that this might carry over to the relocation task, as users would have a better spatial understanding of the scene when they had an additional visual aid during the recognition phase. We also felt, that this improved spatial understanding might positively affect the *Success Rate* for the relocation task. However, we also believe that the *Video* mode is rather complex and that users might find it difficult to mentally combine the two perspective, therefore we think that users will engage with it the least out of all the additional visualizations. We further assumed, that users would need to move their device less with the additional techniques during recognition and that this improved recognition would again carry over to the relocation task, where less "searching" would then be required and therefore the movement should also be lower for the relocation of objects. Having a potentially better spatial understanding with the additional visual aids, we thought that the self reported stress of the user would decrease in both tasks and that their perceived performance would instead increase. Lastly, we believe that users would rate the additional visualizations higher in both tasks and also in the final overall ranking with no regards to position or task, as all additional techniques should improve upon the baseline, which would still be present during the additional visualization. However, we acknowledge that this might not hold true for the *Video*, especially for the relocation task, as its complexity might be detrimental to the experience of the users and therefore might result in lower rankings. Based on these considerations and assumptions, we derived the

following hypotheses about the results of the user study, where additional visualizations would always mean *Ray*, *Opacity* and *Video* and decreases and increases would always be compared to the *Baseline*:

- **H1.1:** The additional visualizations will decrease the mean of the *Recognition Time*.
- **H1.2:** The additional visualizations will decrease the mean of the *Relocation Time*.
- **H2.1:** The additional visualizations will increase the the mean *Success Rate* for the relocation.
- **H3.1:** The additional visualizations will decrease the mean of the *Absolute Translation* and the *Absolute Rotation* in the recognition task.
- **H3.2:** The additional visualizations will decrease the mean of the *Absolute Translation* and the *Absolute Rotation* in the relocation task.
- **H4.1:** The participants will engage less with the *Video* in the recognition task, compared to the other two additional visualizations.
- **H5.1:** The additional visualizations will decrease the self reported *Perceived Stress* level in the recognition task.
- **H5.2:** The additional visualizations will decrease the self reported *Perceived Stress* level in the relocation task.
- **H6.1:** The additional visualizations will increase the self reported *Perceived Performance* in the recognition task.
- **H6.2:** The additional visualizations will increase the self reported *Perceived Performance* in the relocation task.
- **H7.1:** The additional visualizations will be ranked higher by the users in the recognition task.
- **H7.2:** The additional visualizations will be ranked higher by the users in the relocation task.

- **H7.3:** The additional visualizations will also be ranked higher by users in the final ranking (regardless of position and task).
- **H7.4:** The *Video* will overall be ranked lowest of the three additional visualizations in the relocation task.
- **H7.5:** The *Video* will be ranked lowest of the three additional visualizations in the final ranking.

4.4 Participants

We manged to recruit 24 people (aged from 21 to 29, M = 25 years, SD = 2.4 years, 8 female and 1 non-binary) giving us the perfect balancing of *Mode*, *Position*, the scenes and objects shown as discussed earlier. All participants were either students or research assistants in the field of computer science. Additionally, all of them knew what AR was, but only 13 had previous experience with it. Of these 13 people, two rated their proficiency with AR apps as ”very bad”, two as ”bad”, four as ”neutral”, three as ”good” and two as ”very good”. Additionally, out of these 13 people, seven had prior experience with the ARPen and its app. When asked about their proficiency again, two responded with ”bad”, three responded with ”good” and the remaining two answered ”very good”.

4.5 Results

To the best of our knowledge, no mistakes with regards to the sequences shown made it into the data. We considered it as a mistake if we either swapped the order in a sequence by accident or pointed at a completely wrong cube that was not part of the sequence. If such a mistake was made and recognized by us, we would simply repeat the trial. The same is true for the only instance in which the app crashed, here all trials were also repeated for the setting in which it crashed. Overall, we recorded 1728 trials each for the recognition and the relocation task, for a total of 3456. We

included the split into the different positions for the graphs, when *Position* had a significant effect or if it was interesting in other ways, if the effect was not significant a split graph can be found in Appendix B. The same is true for all results where we did not include a graph at all in this section. As we went for a descriptive analysis regarding the user rankings, we also always included the split in these graphs.

4.5.1 Quantitative Results for the Recognition Task

For every participant we averaged their time (*Recognition Time*), their performed movement with the device (*Absolute Translation* and *Absolute Rotation*), the time an additional visualization was active (*Help Time*), the amount of times they activated and deactivated the visualization (*Button Presses*) and the overall time for each trial, which included the time it took the presenter to move the pen from object to object (*Trial Time*) for every condition combination of *Position* $\times$ *Mode*. We then used the *Help Time* divided by the *Trial Time*, to get the percent of time an additional aid was active when available (*Help Time Percent*). For the *Help Time Percent* and the *Button Presses* we excluded the *Baseline* condition from data analysis, as both values would always be zero, since these measurements only applied for the additional visual cue conditions (*Ray*, *Opacity* and *Video*). When the data was decently normally distributed, we analyzed the effect of *Mode*, *Position* and *Position* $\times$ *Mode* via mixed-effect ANOVAs with the user as a random variable. We Log-transformed *Recognition Time*, *Absolute Translation* and the *Absolute Rotation* before the evaluation, to further improve the distribution. If the data was not normally distributed, i.e., *Help Time Percent*, we would use a Generalized Linear Mixed Model (GLMM). All post-hoc pairwise comparisons were performed using Tukey HSD tests. The subjective Likert-Scale ratings were analyzed via Generalized Estimating Equations (GEEs) and the ratings were averaged for the corresponding main effect if one was present. They were then post-hoc tested via pairwise Friedman tests with a Bonferroni correction. All tests were run at the $\alpha = 0.05$ level. This was done with a combination of *JMP* and *IBM SPSS*, since only *IMB SPSS* offers GEEs at the time of

writing. For the rankings we inverted the points given to make it visually clearer, meaning if a user gave '4' points to a a technique it would receive rank '1' in the graphical representations regarding the ranks found throughout this section. These graphs were only analyzed in a descriptive fashion, as it fits them better than a strict analysis via models.

Neither the *Position* ($F_{2,275}$ = 1.0991, p = 0.3346), nor the *Mode* ($F_{3,275}$ = 0.6562, p = 0.5796), nor the *Position* $\times$ *Mode* ($F_{6,275}$ = 0.6650, p = 0.6780) showed a significant effect on the *Recognition Time*.

Absolute Translation (Recognition)

Mode	Significance		Mean	SD
Baseline	A		1.27 m	0.42 m
Ray	A		1.23 m	0.41 m
Opacity	A	B	1.15 m	0.48 m
Video		B	1.01 m	0.44 m

Table 4.1: Means and standard deviations of *Absolute Translation* for the main effect of *Mode* in the recognition task. Rows not connected by the same letter are significantly different.

Absolute Rotation (Recognition)

Mode	Significance		Mean	SD
Baseline	A	B	107.34 degree	46.26 degree
Ray	A		111.40 degree	49.43 degree
Opacity	A	B	99.30 degree	53.63 degree
Video		B	87.01 degree	44.58 degree

Table 4.2: Means and standard deviations of *Absolute Rotation* for the main effect of *Mode* in the recognition task. Rows not connected by the same letter are significantly different.

The *Mode* had a significant effect on the *Absolute Translation* ($F_{3,275}$ = 7.2092, $p < 0.001$), however *Position* ($F_{2,275}$ = 2.4904, p = 0.0847) and *Position* $\times$ *Mode* ($F_{6,275}$ = 0.8240, p = 0.5521) did not show a significant effect. The means and results of the post-hoc tests can be seen in Table 4.1.

	Help Time Percent		
Mode	Significance	Mean	SD
Ray	A	61.44 %	35.31 %
Opacity	A B	52.89 %	34.27 %
Video	B	41.63 %	35.87 %

Table 4.3: Means and standard deviations of *Help Time Percent* for the main effect of *Mode*. Rows not connected by the same letter are significantly different.

The *Mode* also showed a significant effect on *Absolute Rotation* ($F_{3,275} = 3.1155$, $p < 0.03$), while *Position* ($F_{2,275} = 1.6749$, $p = 0.1892$) and *Position* $\times$ *Mode* ($F_{6,275} = 0.3656$, $p = 0.9003$) again had no significant effect. Again, the means and results of the post-hoc tests can be seen in Table 4.2.

For the *Help Percent Time*, the *Mode* again had a significant effect ($F_{2,206.1} = 6.6627$, $p < 0.01$) and *Position* ($F_{2,206} = 0.3833$, $p = 0.6821$), as well as *Position* $\times$ *Mode* ($F_{4,206} = 0.1588$, $p = 0.9588$) did not have a significant effect. Table 4.3 shows the means and results of the post-hoc tests. Additionally, Figure 4.3 shows the mean number of *Button Presses* and the mean *Help Time Percent* in each of the *Mode*s. Despite users pressing the button on average more often in the *Video* condition, compared to the other two conditions, it has the lowest average *Help Time Percent*.

We found that *Position* ($\chi^2(2) = 10.118$, $p < 0.01$) and *Mode* ($\chi^2(3) = 34.092$, $p < 0.001$) had a significant effect on the *Perceived Stress* ratings, while *Position* $\times$ *Mode* ($\chi^2(6) = 3.714$, $p = 0.715$) did not. Post-hoc tests showed, that the effect held true for the *Position* and that users felt the least amount of stress in the *Side-by-side* (M: 4.32, SD: 0.85) position. However, the effect was only significant in comparison to *90-Degree* (M: 3.99, SD: 1.03), but not to *Opposite* (M: 4.01, SD: 1.19). Both *90-Degree* and *Opposite* additionally also fell into the same significance category. The tests also showed, that *Video* (M: 3.44, SD: 1.29) was rated significantly more stressful than *Ray* (M: 4.43, SD: 0.80) and *Opacity* (M: 4.46, SD: 0.79), but not *Baseline* (M: 4.10, SD: 0.89). Further comparisons between all non *Video* conditions showed no significant differences.

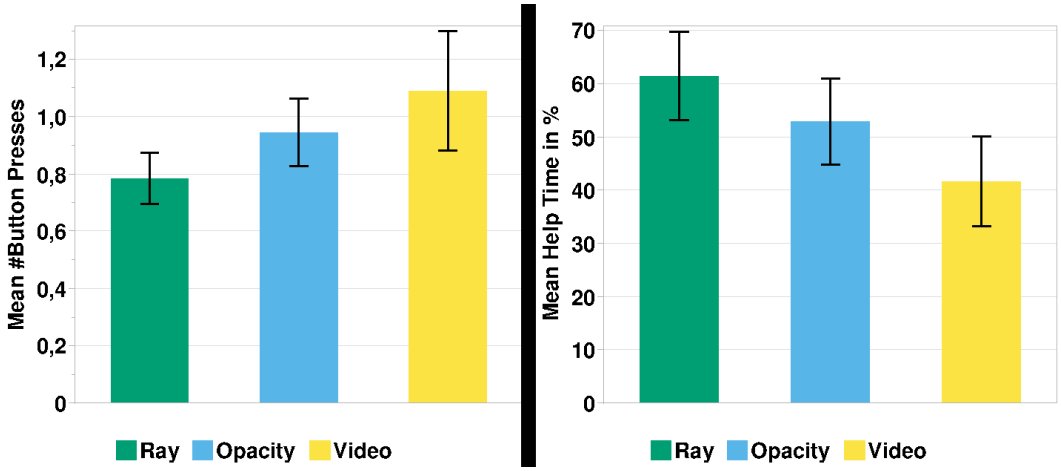

Figure 4.3: Effect of *Mode* on the *Button Presses* and the *Help Time Percent*. The percent of time the additional visual aid was active is higher for *Ray* and *Opacity*, even though the button to activate and deactivate it was pressed less often. Whiskers denote the 95% CI.

For the *Perceived Performance* ratings we again had a significant effect of *Position* ($\chi^2(2) = 7.464$, $p < 0.03$) and *Mode* ($\chi^2(3) = 12.418$, $p < 0.01$), but no significant effect of *Position* $\times$ *Mode* ($\chi^2(6) = 0.525$, $p = 0.998$). The post-hoc tests did not hold true for the effect of *Position* (*Opposite* M: 4.21, SD: 0.95; *90-Degree* M: 4.15, SD: 0.93; *Side-by-side* M: 4.39, SD: 0.85) or the *Mode* (*Baseline* M: 4.24, SD: 0.81; *Ray* M: 4.47, SD: 0.80; *Opacity* M: 4.42, SD: 0.80; *Video* M: 3.86, SD: 1.09). However, it seems that the participants felt a bit less confident when using the *Video*.

The inverted point-based rankings ('4' points given in the questionnaire is equal to rank '1' in the graph) show that users liked the *Ray* and *Opacity* in all positions the most, followed by the *Baseline* and that *Video* was generally disliked. Also it seems like users had a more concise preference for the *Opacity* when positioned opposite to the presenter holding the pen (see Figure 4.4).

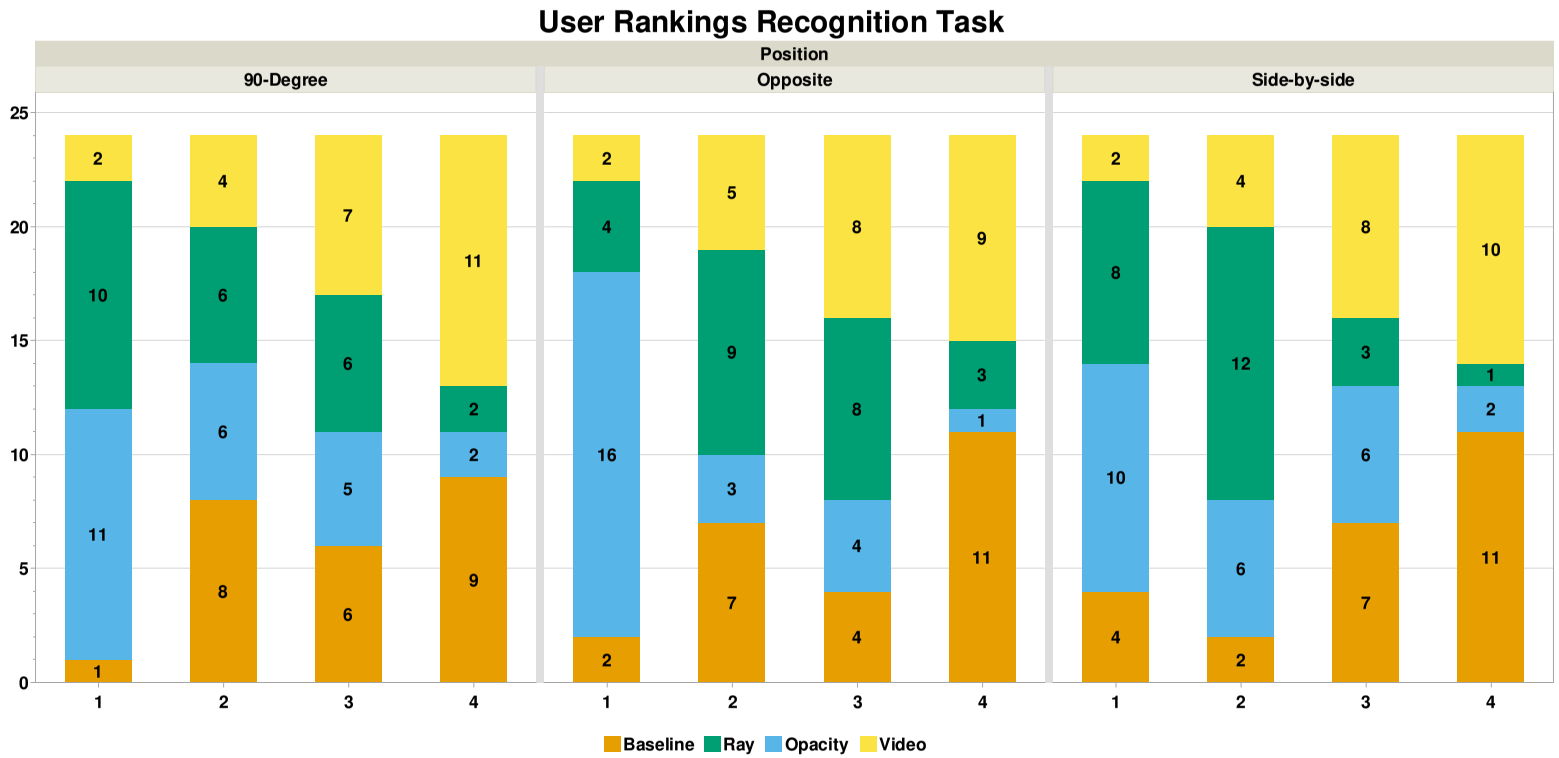

Figure 4.4: Subjective Ranking of *Mode* in the different *Position*s for the recognition task. *Ray* and *Opacity* are in the first two places, followed by *Baseline* and *Video* comes in last. *Opacity* also seems to be more heavily preferred in the *Opposite* position, whereas it is more balanced in the other two *Position* settings.

4.5.2 Qualitative Results for the Recognition Task

Participants often remarked, that while the *Video* is very good for recognizing the number on the shown cube, figuring out the corresponding position in the scene was "difficult" and "confusing". Some users also said, that the video frame was either too large or too small. When asked about it, nearly all participants would prefer to be able to scale the video to their own liking, with some stating they would potentially have preferred a "Split Screen" over the "Picture-in-picture" approach. Users also reported, that it sometimes was difficult to distinguish two cubes in the *Opacity* setting when one was directly behind, or in front, of another. The participants further pointed out, that they either had a strong preference for the *Ray* or for the *Opacity*. Users who had a strong preference for the *Ray* also often remarked, that they could keep it permanently active as it did not impact the overall look of the scene and therefore it caused the least amount of distraction. However, some participants would have preferred a more 3D feel for the ray, meaning it should get smaller in the distance and that they should have the option to adjust the transparency of the ray itself, as it sometimes could occlude the numbers. Two users also reported that they would have liked another highlighting color, that would work complementary with the color of the cubes, this was not related to any colorblindness. A few users also reported that the message displayed to them, indicating they should inform the presenter once they had found/memorized the current cube, was obstructing their view from time to time. As expected, some participants only activated the additional visualizations if they needed them, whereas others did not activate them at all due to disliking or not needing them and some did always activate them and kept them on, either because they liked them or because they found them "interesting" and wanted to test them more. Participants would often repeat the shown sequence aloud, to keep track of the order and to be able to search for a cube if they forgot the position, but still had the number in mind. Only two of the participants regularly stood up from the chair to get a different view, whereas all participants had the tendency to follow the pointing by leaning to the sides to keep a clear view.

4.5.3 Quantitative Results for the Relocation Task

For every participant we averaged their time (*Relocation Time*), their performed movement with the device (*Absolute Translation* and *Absolute Rotation*), the percent of correctly performed trials (*Success Rate*) and the number of wrongfully selected cubes (*Wrong Nodes*). Again, to analyze the effect of *Mode*, *Position* and *Position* $\times$ *Mode*, we performed mixed-effect ANOVAs with the user as a random variable, when the data was decently normally distributed. A Log-transformation was applied to the *Relocation Time*, the *Absolute Translation* and the *Absolute Rotation* before the evaluation. All post-hoc pairwise tests for the normally distributed data were performed using Tukey HSD test with $\alpha = 0.05$. The subjective Likert-Scale ratings were again analyzed using GEEs and post-hoc tested via pairwise Bonferroni corrected Friedman tests, where the ratings were averaged for each user per significant condition. For the rankings we again inverted the points given to make it visually clearer and only analyzed them descriptively, as seen in the recognition results already.

	Relocation Time		
Mode	Significance	Mean	SD
Baseline	A	6.67 s	1.75 s
Ray	A	6.98 s	2.50 s
Opacity	B	8.56 s	3.96 s
Video	C	11.05 s	5.49 s

Table 4.4: Means and standard deviations of *Relocation Time* for the main effect of *Mode*. Rows not connected by the same letter are significantly different.

We found, that the *Position* ($F_{2,275} = 1.6385$, $p = 0.1962$) and *Position* $\times$ *Mode* ($F_{6,275} = 0.2147$, $p = 0.9720$) did not have a significant effect on the *Relocation Time*, but *Mode* ($F_{3,275} = 22.3197$, $p < 0.0001$) did have a significant effect. The means results of the post-hoc tests can be seen in Table 4.4.

For the *Absolute Translation* and *Absolute Rotation*, we again saw a significant effect of *Mode* ($F_{3,275} = 14.9719$, $p < 0.0001$ and $F_{3,271.8} = 10.4271$, $p < 0.0001$). *Position* ($F_{2,275} = 2.2452$,

p = 0.1078 and $F_{2,271.8}$ = 1.3937, p = 0.2499) and *Position* × *Mode* ($F_{6,275}$ = 0.2493, p = 0.9593 and $F_{6,271.8}$ = 0.2782, p = 0.9469) did not display any significant effect. The respective pos-hoc tests can be seen in Table 4.5 and Table 4.6.

Absolute Translation (Relocation)

Mode	Significance			Mean	SD
Baseline	A			0.88 m	0.36 m
Ray	A	B		0.93 m	0.50 m
Opacity		B		1.21 m	0.73 m
Video			C	1.65 m	1.14 m

Table 4.5: Means and standard deviations of *Absolute Translation* for the main effect of *Mode* in the relocation task. Rows not connected by the same letter are significantly different.

Absolute Rotation (Relocation)

Mode	Significance		Mean	SD
Baseline	A		83.75 degree	35.84 degree
Ray	A		90.43 degree	52.02 degree
Opacity	A		115.90 degree	76.18 degree
Video		B	151.40 degree	107.07 degree

Table 4.6: Means and standard deviations of *Absolute Rotation* for the main effect of *Mode* in the relocation task. Rows not connected by the same letter are significantly different.

A GLMM showed a significant effect of *Mode* on the *Success Rate* ($F_{3,276}$ = 5.4040, $p < 0.0013$), while *Position* ($F_{2,276}$ = 0.2609, p = 0.7706) and *Position* × *Mode* ($F_{6,276}$ = 0.5378, p = 0.7793) again did not show any significant effect. The results of the post-hoc test can be seen in Table 4.7. Since *Mode* had a significant effect on the *Success Rate*, we also looked at the number of mistakes made (see Figure 4.5).

The GEE showed a significant effect of both *Position* ($\chi^2(2)$ = 7.031, $p < 0.030$) and *Mode* ($\chi^2(3)$ = 37.918 , $p < 0.0001$) on the *Perceived Stress*, whereas the *Position* × *Mode* ($\chi^2(6)$ = 4.236, p = 0.645) did not display a significant effect. For the *Position* the post-hoc Friedman test could not confirm the significance of the effect ($\chi^2(2)$ = 4.651, p = 0.098). The post-hoc comparison for the *Mode*s ($\chi^2(2)$ = 4.651, $p < 0.001$)

	Success Rate			
Mode	Significance		Mean	SD
Baseline	A		98.60 %	5.44 %
Ray	A	B	95.35 %	10.55 %
Opacity		B	91.39 %	13.11 %
Video		B	91.60 %	11.88 %

Table 4.7: Means and standard deviations of *Success Rate* for the main effect of *Mode* in the relocation task. Rows not connected by the same letter are significantly different.

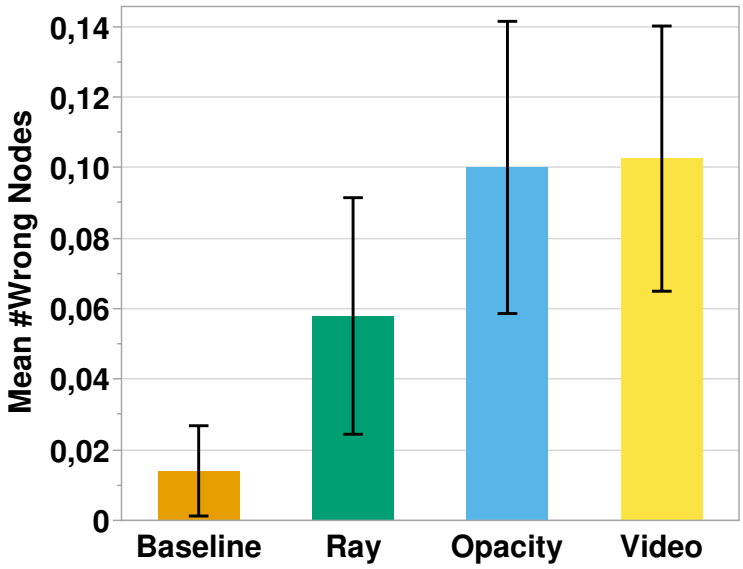

Figure 4.5: The users selected over more than three times the amount of wrong nodes when they had the *Ray* during the recognition phase and over five times if they had the *Opacity* or *Video* available to them during recognition, compared to the *Baseline*. However, the amount of mistakes made overall is still low as a value of 0.05 relates to '1' mistake made over all 18 trials performed with each *Mode* (1 out of 54 cubes). Whiskers denote the 95% CI.

however showed, that users rated the *Video* (M: 2.95, SD: 1.19) significantly lower compared to all other techniques. No other significant differences were found between the remaining three techniques (*Baseline* M: 4.28, SD: 0.83; *Ray* M: 4.17, SD: 0.90; *Opacity* M: 3.82, SD: 1.15).

The *Mode* ($\chi^2(3) = 28.095$, $p < 0.001$) displayed a significant effect on the *Perceived Performance*, whereas *Position* ($\chi^2(2) = 5.809$, $p = 0.055$) and *Position* $\times$ *Mode* ($\chi^2(6) = 2.994$, $p = 0.810$) did not. Post-hoc tests showed, that participants rated their *Perceived Performance* higher for all non *Video* (M: 3.11, SD: 1.13) techniques (*Baseline* M: 4.32, SD: 0.73; *Ray* M: 4.29, SD: 0.86; *Opacity* M: 4.07, SD: 1.04). Comparing the higher rated three techniques did not show any further significant differences between them.

The rankings show, that user overall liked the *Ray* and *Opacity* setting the most, followed closely by the *Baseline* and that *Video* was heavily disliked (see Figure 4.6). It also seems like users had a higher preference for the *Opacity* setting when sitting face to face with the instructor, whereas they preferred the *Ray* in the *90-Degree* scenario. Both *Ray* and *Opacity* also are close to being even in the *Side-by-side* position.

4.5.4 Qualitative Results for the Relocation Task

Participants often remarked, that they had issues to relocate the cubes when having used the *Video* and that they were confused by having to combine the two perspectives inside their head. Some users also reported, that they would have liked to been able to also have the *Opacity* during the relocation, as they had memorized and visualized the scene with it and that finding the position of the cube was therefore difficult without it. Users also said, that they sometimes had trouble remembering the position and/or number of the cubes when they had used an additional visual aid during the recognition phase. Most people responded, that they would often relocate the cubes by looking for their number instead of strictly remembering the exact position. The two participants who stood up during the recognition

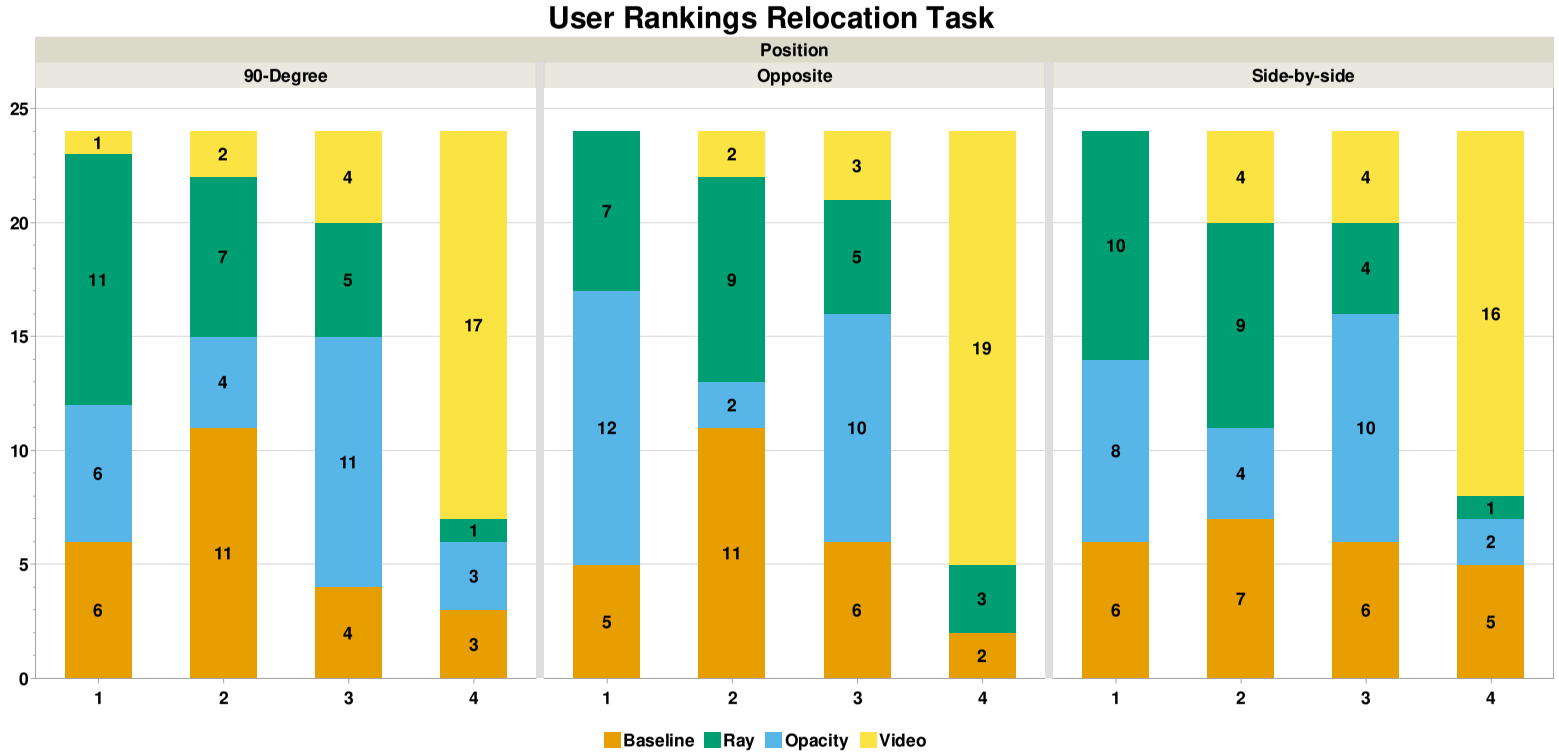

Figure 4.6: Subjective Ranking of *Mode* in the different *Position*s for the relocation task. *Ray* and *Opacity* are in the first two places, with *Baseline* trailing closely behind. *Video* is rated lowest by far. *Opacity* also seems to be preferred in the *Opposite* position, whereas *Ray* seems to be preferred in the *90-Degree* position and both seem to be close to evenly in the *Side-by-side* scenario.

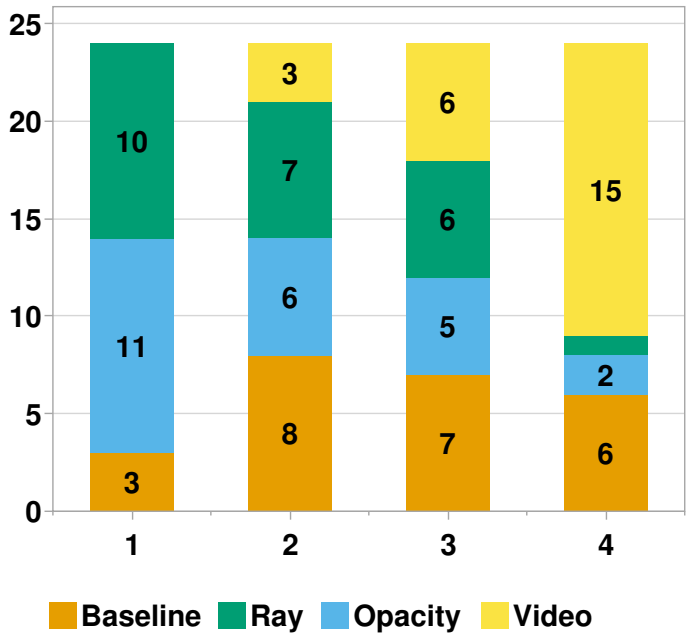

Figure 4.7: The final ranking regardless of *Position* or the *Task*. *Ray* and *Opacity* seem to be leading again, with *Baseline* being third and *Video* rated lowest.

task also did so for the relocation task. Most people tried to keep a static view during relocation, meaning they only moved their phone if they had to, to get the object into view or if they lost track of the position and had to actively search for a cube. Three participants also went close to the scenes with their phones during the relocation task and when asked about it, they responded with "it is easier this way to tap the correct cubes", even though tapping a wrong one would not result in failure directly as users had to confirm their choice.

4.5.5 Quantitative Results for the Final Ranking & General Feedback from the Interview

For the final ranking (without regards to either *Position* or the *Task*) the users rated the techniques similar to what we have seen before, with *Ray* and *Opacity* in the leading spots,

followed by *Baseline* and *Video* coming in last (see Figure 4.7). The users also provided general feedback regarding the difficulty of the task, where nearly all agreed that it was appropriate. The participants also did not report any technical difficulties with the system, neither with regards to the synchronization nor with any potential video lag. A few users suggested, that the *Video* might be more beneficial in a more semantic task setting.

4.6 Discussion

Since the results showed no significant effect on the *Recognition Time* we reject **H1.1**. We assume, this might be due to the still rather simple scene setup, if we would have used even denser and even more complex scenes, we think the results might have been different. We also saw no time improvement for the relocation task, instead *Opacity* and *Video* performed significantly worse than *Baseline* and *Ray*. The qualitative remarks given for those two techniques give a good indication, that this might be due to the added complexity having to combine two views (*Video*) and the impact on the overall look and feel of the scene, as it switches from the see-through look back to the solid one for the relocation task (*Opacity*). We therefore also reject **H1.2**. While all techniques had an over 90% *Success Rate* for the relocation task, it was again the *Baseline* that outperformed the additional techniques, so we also have to reject **H2.1**. Again the qualitative remarks give a good indication as to why this happened, as people reported they somehow struggled with remembering the position or order from time to time if they used an additional visualization. This means that they probably got a false sense of confidence when memorizing the sequence with the additional visual cues turned on. However, the amount of mistakes made was generally low, where often user would make only one or two errors over all 18 trials conducted for each *Mode* (1 or 2 cubes selected incorrectly out of 54 total cubes). **H3.1** can be partially accepted, as the *Video* indeed did reduce the the mean of the *Absolute Translation* in the recognition task compared to the *Baseline*, the other two additional techniques however could not significantly improve upon it, but the means

were still slightly lower than in the *Baseline*. For the *Absolute Rotation* again only a insignificant improvement by the *Video* and *Opacity* over the *Baseline* could be seen. Both of these results are also probably caused by the users still trying to follow the overall movement of the presenter, even in the *Video* scenario as they would often turn it on initially, but then turn it off to get a better view on the overall scene, where they then had to move the device to bring the cube back into their actual view. We also have to reject **H3.2** as none of the techniques outperformed the *Baseline* in the relocation task with regards to the movement amount. *Opacity* and *Video* even performed significantly worse with regards to *Absolute Translation* and *Video* also performed worse in the *Absolute Rotation* category. This ties back to the added complexity and overall change of look of the scenes mentioned earlier, that already impacted the *Relocation Time* negatively. Users had to actually search the cubes as they could often only remember the number, but not the position and sometimes they even completely forgot a number and had to take a guess. The results of the analysis on the *Help Time Percent* showed that despite users pressing the on/off button more often for the *Video* it still had the lowest percent of active time for the three additional visual cues. This is further supported by the qualitative remarks, as users often pointed out they only engaged with it sporadically, if it all, and then they often switched between on and off, with one user even implementing a "rapid toggle" tactic. Based on these observations we accept **H4.1**. Another interesting point is, that the *Position* did affect the *Perceived Stress* in the recognition task, something that did not happen on the more traditional hard data measurements, e.g., time and movement. It was expected that users would feel the most comfortable in the *Side-by-side* scenario, as it would offer a similar PoV to that of the presenter. We assume, that users felt significantly more stressed in the *90-Degree* position, because it is a more uncommon position to work together for students, as they often either sit face to face or next to each other in traditional real-world scenarios based on the tables in the "RWTH-Informatikzentrum" and that they therefore struggled more with combining their view with that of the presenter in this unfamiliar position, e.g., the furthest right cube in the bottom row for the presenter would be the cube with the highest depth in the rightmost column in the bot-

tom row. Additionally, the users felts the most stress using the *Video* and since *Ray* and *Opacity* both did not significantly outperform the *Baseline* we reject **H5.1**. It is not fully clear, as to why users felt more stress using the *Video* in the recognition task. However, we assume it is because the tasks are linked closely together and that potential negative experience during relocation also impacted their perception for the recognition task. The qualitative remarks also indicated that users felt, that it was really good to "see the number, but not the actual position", further supporting the idea that the negative experience for the relocation task also affected the ratings for the recognition portion. *Video* was also rated even lower with regards to *Perceived Stress* in the relocation task (Relocation M: 2.95, SD: 1.19; Recognition M: 3.44, SD: 1.29). As no other differences were found, we reject **H5.2**. We have to reject **H6.1** and **H6.2** as either no significant differences were found or the *Video* was even rated significantly lower than the other three conditions (relocation task). The user rankings for both tasks and the final ranking showed, that users preferred either *Ray* or *Opacity* over the *Baseline* and that *Video* was generally the least preferred option. Therefore, we partially accept **H7.1**, **H7.2**, **H7.3** and fully accept **H7.4** and **H7.5**. We believe that *Video* was rated lowest as it was the most complex one and users struggled to mentally combine the two perspective into one coherent image and that while it is really good to see an object, based on the qualitative remarks, it is not well suited for this task setup, where remembering a position also plays an important role.

4.7 Design Recommendations

Using the quantitative and qualitative data derived from the study, we can now give some design recommendations on what worked already and what might need some further improvements.

Baseline is already quite good

The results showed, that the *Baseline* highlighting already provides a good visual cue and that users even were able to perform better using it, than using any of the three additional visualizations. We therefore suggest, to keep it as

the overall base setting for the future. However, as two users suggested it might be beneficial to adjust the highlighting color based on the color of the pointed at object, to create more drastic differences, e.g., the use of complementary colors.

Ray and *Opacity* as viable additions

Task performance wise *Ray* and *Opacity* were nearly always closely behind the *Baseline*, but they were generally preferred by the users, based on the rankings presented. We suggest to keep them as additional techniques going forward, where users can turn them on if they need them or if they feel helpful to them. Based on the qualitative remarks, we feel that it could be good to perform some further visualization improvements for the *Ray* that would give it more of a 3D feel and that would stop it from potentially occluding important information, such that it not solves one issue by introducing another. *Opacity* could also see improvements, such that it is no longer an issue if two objects are directly and closely behind one and another, either by further decreasing the opacity of other objects or by having a stronger visual difference, e.g., usage of complementary colors again.

Video needs improvements and different tasks

With the generally negative feel about the *Video* and its subpar performance in the relocation task, it is hard to justify keeping it, but we believe that the tasks, especially the relocation one, might not have been the strongest suite for it and that more semantic tasks would probably be better to assess its potential value for collaboration. Nevertheless, we still suggest that its current iteration needs some improvements. Users should be able to customize the size to their liking and a potential "Split Screen" view could be introduced as an alternative to the PiP approach.

Message label occluding objects

For this study, we added label that tells the participant to inform us when they had recognized and memorized a shown cube and that a trial was started, but this label would sometimes occlude the actual objects. Since this kind of a label would not be used in an actual version, we do not believe this to be an issue. However, in potential future studies it might be more beneficial for example to use a colored border approach, e.g., a red border around the view could indicate no ongoing trial and a green border would

indicate that a trial is in progress.

Automatic adjustments

All of the described adjustments, e.g., transparency of the ray or customization of the video, of course could also be automated to only occur in specific scenarios, e.g., when the ray occludes important information. While this is an interesting approach, we are not sure whether users would prefer this approach over being able to manipulate these settings themselves, as the automatic approach might not fit their desire.

Chapter 5

Summary and Future Work

Overview

This final chapter concludes this thesis. We summarize our work on taking the first step towards being able to collaborate with the ARPen in a co-located setting. Furthermore, we sketch out potential future work as this thesis is just the entry point, that will most likely open up many more research topics on collaboration with the ARPen.

5.1 Summary and contributions

We aimed to create an entry point for collaborative work with the ARPen

Research into how viable AR is for collaborative work has already been heavily explored, yet it continues to be an interesting and ongoing field of research, especially when it comes to co-located collaboration and how visual cues should be designed, such that they are easy to understand for others. With our work we focused on creating an entry point for the usage of the ARPen in a co-located collaborative scenario. We aimed to build an experimental extension for the ARPen app, that would allow users to experience the same AR content with high stability and a high level of synchronization. As pointing operations play an important role in everyday communication and in many forms of collaborative work, it seemed to be a good idea to look at how

to visualize a pointing operation performed via the ARPen, such that another user has an easy time understanding it.

We implemented the sharing of AR content and four visualization techniques

For this thesis, we therefore extended the existing ARPen app with the functionality to synchronize the AR content between two devices. We used the `Multipeer Connectivity Framework` and an `ARImageAnchor` to perform the synchronization process with a high accuracy and stability. Based on previous research, we decided to implemented four different visualization techniques, a basic highlighting (*Baseline*) which would always be present, the rendered version of the ray cast used for the pointing (*Ray*), the reduction of the opacity of not pointed at objects to better indicate the pointed at object, while also helping with potential occlusion (*Opacity*) and a new approach, in the co-located scenario, with the possibility to switch to the view of another person's device (*Video*).

Raspberry Pi as router

To fix issues with the live video feed, we had to convert a Raspberry Pi into a makeshift router, such that all data would be transferred only over a 2.4 GHz WLAN band.

Evaluation showed promising results for *Ray* and *Opacity*

We then compared the different techniques in a user study, which consisted of two linked tasks. First, users had to recognize and memorize a shown sequence of three numbered cubes with one of the techniques available to them, after which they then had to relocate the shown cubes, by tapping them in the correct order via touch input on their device. This was done in three different user to user positions and users were not required to use any of the three additional visualizations when available. We evaluated quantitative (task times, required movement, time additional visualization were active, number of button presses to turn an additional help on/off, success rate in the relocation task, mistakes made in the relocation task, Likert-Scale questions regarding stress and performance and rankings) and qualitative (remarks after the trials, additional comments given at the end of the study and answers to the short interview questions) data. This evaluation showed, that while the *Baseline* would often be the best performance wise, e.g., task time or success rate, it were *Ray* and *Opacity* that were preferred by the participants. *Video* on the other hand would generally be the most disliked by the users. All of this data

was presented and discussed in 4.5 “Results” and 4.6 “Discussion” and allowed us to derive some design recommendations, such as further improvements to the visualizations of *Baseline*, *Ray* and *Opacity* with the aim to eliminate some of their issues, e.g., introducing complementary colors for the highlighting, allowing users to adjust the transparency of the ray or further decreasing the opacity of not pointed at objects in the *Opacity* setting. For the *Video* we suggested to make it more customizable and to potentially include a ”Split Screen” view as an alternative to the PiP approach currently present.

Contributions of the thesis

This work contributes to the ongoing research of collaborative work using AR in a co-located setting. Some of the proposed additional techniques proved to be significantly more liked over the baseline by the participants of the user study, while not suffering from a high performance deficit with regards to, e.g., the time it took to perform the tasks. This is a good indication, that these techniques can provide viable additional visual cues, that can be used to understand pointing operations in an co-located AR environment using the ARPen.

5.2 Future work

Overview

As this thesis serves as an entry point to exploring the collaborative possibilities of the ARPen, there are many other interesting research topics linked to it, some of which we discuss in this section.

Re-evaluate the *Video* with different tasks

Since the video generally performed worse and was disliked by the users, we suggest to re-evaluate it with different tasks. We believe, that this technique would show better results when used in semantic tasks, similar to what Wells and Houben [2020] tested, e.g., asking a user how many red tile are on a specific side of a virtual Rubik’s Cube. Here, a comparison could be made between being able to manipulate it on one’s own device, being able to switch the view to that of another person as we presented it in this thesis and having to walk around the cube without additional techniques. For example, here it could also

be possible to compare the two video approaches, PiP as presented in this thesis and "Split Screen" as suggested by some users.

Evaluate *Ray* and *Opacity* in semantic tasks

Similar to the aforementioned re-evaluation of the video, we think it might be beneficial to also test the *Ray* and the *Opacity* techniques in semantic tasks, as these tasks offer a broader spectrum for collaborative interactions.

Evaluate the techniques in a group sketching task

Since the ARPen was also build for mid-air sketching and designing, we feel that all the techniques should be evaluated in a group-based scenario with multiple ARPens in use, where the group is collaboratively working on a design task, e.g., a virtual car. This should be helpful to get a better understanding of collaborative group dynamics with the ARPen and it might give further information on what improvements are needed for the visual pointing cues, e.g, Person A points at the rear of a virtual car to talk about a certain design decision, what help can we give to Person B and Person C to understand it without obstructing Person D who is currently sketching some other design ideas.

Evaluate the visualizations with different pointing techniques

Finally, we only evaluated the visualization with a ray cast based pointing, as it is the most common and the most preferred approach. However, there are different pointing techniques, e.g., having to move the tip of the pen inside a cube, and it might be interesting to investigate how these could affect our shown visualization techniques.

Appendix A

User Study Consent Form and Questionnaire

The following consent form and questionnaire were handed out to the participants of the study. The consent form was filled out before the study started and the questionnaire given to the participant after each of the six trials with the corresponding part marked. Additionally, the fully counterbalanced Latin square is shown here.

Informed Consent Form

Helping Users Understand Pointing Operations with the ARPen

PRINCIPAL INVESTIGATOR Marvin Bruna
Media Computing Group
RWTH Aachen University
Phone: 01525/1908025
Email: marvin.bruna@rwth-aachen.de

Purpose of the study: The goal of this study is to analyze different aiding visualization techniques with regards to pointing operations using the ARPen. The users will be asked to perform two different tasks, first a recognition of a pointing operation, followed by a relocation task of the shown AR objects. Each of these tasks will be performed in three different positions(Opposite of the presenter, in a 90-degree angle towards the presenter and side-by-side with the presenter) with four different techniques (three + the baseline). Additionally, users will be asked to fill out a questionnaire. I will use these questionnaires as well as the multiple other data points (e.g., movement or time) to analyze the techniques. Furthermore, I will look at how the user behaves during the task (e,g. do they look over to the presenter's device when side-by-side).

Procedure: Participation in the study involves two phases. In the first phase, you will be introduced to the general usage of the app and the way the tasks will be chained. In the second phase, you will perform the two tasks in the mentioned settings from above. After each combination of technique and position you will be asked to fill-out the corresponding part of the questionnaire regarding that position and technique.

Risks/Discomfort: You may become fatigued during your participation in the study. There are no other risks associated with participation in the study. Should completion of either the tasks or the questionnaire become distressing to you, it will be terminated immediately.

Benefits: The results of this study will be useful for future research and work on the ARPen and how to use it in a shared space collaborative setting.

Alternatives to Participation: Participation in this study is voluntary. You are free to withdraw or discontinue the participation.

Cost and Compensation: Participation in this study will involve no cost to you. There will be snacks and drinks for you during and after the participation.

Confidentiality: All information collected during the study period will be kept strictly confidential. You will be identified through identification numbers. No publications or reports from this project will include identifying information on any participant. If you agree to join this study, please sign your name below.

_____ I have read and understood the information on this form.
_____ I have had the information on this form explained to me.

Participants' Name	Participants' Signature	Date
	Principal Investigator	Date

If you have any questions regarding this study, please contact Marvin Bruna at 01525/1908025 , email: marvin.bruna@rwth-aachen.de

Figure A.1: The consent form handed out before the user study.

ID:

General Questions:

How old are you: _____

Gender: _____

Do you have prior experience with AR?: Yes ☐ No ☐

If you have prior experience how would you rate your proficiency with AR Apps (1 = very bad, 5 = very good):

1	2	3	4	5

Do you have prior experience with the ARPen?: Yes ☐ No ☐

If you have prior experience how would you rate your proficiency with the ARPen and its App:

1	2	3	4	5

Opposite Position Recognition (Base):

1 = Very stressed/Very stressful 5 = Not stressed at all/Not stressful at all (for questions regarding stress)

1 = very bad/low 5 = very good/high (for questions regarding ease of recognition/performance)

How stressed did you feel having no additional help to recognize and memorize the shown sequences of cubes?

1	2	3	4	5

How do you rate your chance/performance to recognize the pointing without additional help?

1	2	3	4	5

Opposite Position Relocation (Base):

1 = Very stressed/Very stressful 5 = Not stressed at all/Not stressful at all (for questions regarding stress)

1 = very bad/low 5 = very good/high (for questions regarding ease of relocation/performance)

How stressed did you feel relocating the sequences of cubes after having no additional help in the recognition/memorization part?

1	2	3	4	5

How to you rate your chance/performance to relocate the sequences of cubes after having no additional help in the recognition/memorization part?

1	2	3	4	5

Opposite Position Recognition (Ray):

How stressed did you feel having the ray as additional help to recognize and memorize the shown sequences of cubes?

1	2	3	4	5

How do you rate your chance/performance to recognize the pointing with the ray as additional help?

1	2	3	4	5

Figure A.2: The first page of the questionnaire.

Opposite Position Relocation (Ray):

How stressed did you feel relocating the sequences of cubes after having the ray as additional help in the recognition/memorization part?

1	2	3	4	5

How to you rate your chance/performance to relocate the sequences of cubes after having the ray as additional help in the recognition/memorization part?

1	2	3	4	5

Opposite Position Recognition (Opacity):

How stressed did you feel having the opacity reduction as additional help to recognize and memorize the shown sequences of cubes?

1	2	3	4	5

How do you rate your chance/performance to recognize the pointing with the opacity reduction as additional help?

1	2	3	4	5

Opposite Position Relocation (Opacity):

How stressed did you feel relocating the sequences of cubes after having the opacity reduction as additional help in the recognition/memorization part?

1	2	3	4	5

How to you rate your chance/performance to relocate the sequences of cubes after having the opacity reduction as additional help in the recognition/memorization part?

1	2	3	4	5

Opposite Position Recognition (Video):

How stressed did you feel having the video as additional help to recognize and memorize the shown sequences of cubes?

1	2	3	4	5

How do you rate your chance/performance to recognize the pointing with the video as additional help?

1	2	3	4	5

Opposite Position Relocation (Video):

How stressed did you feel relocating the sequences of cubes after having the video as additional help in the recognition/memorization part?

1	2	3	4	5

How to you rate your chance/performance to relocate the sequences of cubes after having the video as additional help in the recognition/memorization part?

1	2	3	4	5

Please generally rank the settings from worst = 1 to best = 4, based on how you felt using them etc.

Recognition Task

No Help	
Rendered Ray	
See-Through	
Video	

Relocation Task

No Help	
Rendered Ray	
See-Through	
Video	

Figure A.3: The second page of the questionnaire.

90-degree Position Recognition (Base):

How stressed did you feel having no additional help to recognize and memorize the shown sequences of cubes?

1	2	3	4	5

How do you rate your chance/performance to recognize the pointing without additional help?

1	2	3	4	5

90-degree Position Relocation (Base):

How stressed did you feel relocating the sequences of cubes after having no additional help in the recognition/memorization part?

1	2	3	4	5

How to you rate your chance/performance to relocate the sequences of cubes after having no additional help in the recognition/memorization part?

1	2	3	4	5

90-degree Position Recognition (Ray):

How stressed did you feel having the ray as additional help to recognize and memorize the shown sequences of cubes?

1	2	3	4	5

How do you rate your chance/performance to recognize the pointing with the ray as additional help?

1	2	3	4	5

90-degree Position Relocation (Ray):

How stressed did you feel relocating the sequences of cubes after having the ray as additional help in the recognition/memorization part?

1	2	3	4	5

How to you rate your chance/performance to relocate the sequences of cubes after having the ray as additional help in the recognition/memorization part?

1	2	3	4	5

90-degree Position Recognition (Opacity):

How stressed did you feel having the opacity reduction as additional help to recognize and memorize the shown sequences of cubes?

1	2	3	4	5

How do you rate your chance/performance to recognize the pointing with the opacity reduction as additional help?

1	2	3	4	5

90-degree Position Relocation (Opacity):

How stressed did you feel relocating the sequences of cubes after having the opacity reduction as additional help in the recognition/memorization part?

1	2	3	4	5

How to you rate your chance/performance to relocate the sequences of cubes after having the opacity reduction as additional help in the recognition/memorization part?

1	2	3	4	5

Figure A.4: The third page of the questionnaire.

90-degree Position Recognition (Video):

How stressed did you feel having the video as additional help to recognize and memorize the shown sequences of cubes?

1	2	3	4	5

How do you rate your chance/performance to recognize the pointing with the video as additional help?

1	2	3	4	5

90-degree Position Relocation (Video):

How stressed did you feel relocating the sequences of cubes after having the video as additional help in the recognition/memorization part?

1	2	3	4	5

How to you rate your chance/performance to relocate the sequences of cubes after having the video as additional help in the recognition/memorization part?

1	2	3	4	5

Please generally rank the settings from worst = 1 to best = 4, based on how you felt using them etc.

Recognition Task

No Help	
Rendered Ray	
See-Through	
Video	

Relocation Task

No Help	
Rendered Ray	
See-Through	
Video	

Side-by-side Position Recognition (Base):

How stressed did you feel having no additional help to recognize and memorize the shown sequences of cubes?

1	2	3	4	5

How do you rate your chance/performance to recognize the pointing without additional help?

1	2	3	4	5

Side-by-side Position Relocation (Base):

How stressed did you feel relocating the sequences of cubes after having no additional help in the recognition/memorization part?

1	2	3	4	5

How to you rate your chance/performance to relocate the sequences of cubes after having no additional help in the recognition/memorization part?

1	2	3	4	5

Side-by-side Position Recognition (Ray):

How stressed did you feel having the ray as additional help to recognize and memorize the shown sequences of cubes?

1	2	3	4	5

How do you rate your chance/performance to recognize the pointing with the ray as additional help?

1	2	3	4	5

Figure A.5: The fourth page of the questionnaire.

Side-by-side Position Relocation (Ray):

How stressed did you feel relocating the sequences of cubes after having the ray as additional help in the recognition/memorization part?

1	2	3	4	5

How to you rate your chance/performance to relocate the sequences of cubes after having the ray as additional help in the recognition/memorization part?

1	2	3	4	5

Side-by-side Position Recognition (Opacity):

How stressed did you feel having the opacity reduction as additional help to recognize and memorize the shown sequences of cubes?

1	2	3	4	5

How do you rate your chance/performance to recognize the pointing with the opacity reduction as additional help?

1	2	3	4	5

Side-by-side Position Relocation (Opacity):

How stressed did you feel relocating the sequences of cubes after having the opacity reduction as additional help in the recognition/memorization part?

1	2	3	4	5

How to you rate your chance/performance to relocate the sequences of cubes after having the opacity reduction as additional help in the recognition/memorization part?

1	2	3	4	5

Side-by-side Position Recognition (Video):

How stressed did you feel having the video as additional help to recognize and memorize the shown sequences of cubes?

1	2	3	4	5

How do you rate your chance/performance to recognize the pointing with the video as additional help?

1	2	3	4	5

Side-by-side Position Relocation (Video):

How stressed did you feel relocating the sequences of cubes after having the video as additional help in the recognition/memorization part?

1	2	3	4	5

How to you rate your chance/performance to relocate the sequences of cubes after having the video as additional help in the recognition/memorization part?

1	2	3	4	5

Please generally rank the settings from worst = 1 to best = 4, based on how you felt using them etc.

Recognition Task

No Help	
Rendered Ray	
See-Through	
Video	

Relocation Task

No Help	
Rendered Ray	
See-Through	
Video	

Figure A.6: The fifth page of the questionnaire.

Please now generally rank the settings from worst = 1 to best = 4 without specific regard to either task or position.

No Help	
Rendered Ray	
See-Through	
Video	

Additional comments:

Figure A.7: The sixth page of the questionnaire.

ID: 1 + 13	Opposing Position: Base + Scene1 -> Ray + Scene2 -> Opacity + Scene12 -> Video + Scene3 90Degree Position: Ray + Scene11 -> Video + Scene4 -> Base + Scene10 -> Opacity + Scene5 Side By Side Position: Video + Scene9 -> Opacity + Scene6 -> Ray + Scene8 -> Base + Scene7
ID: 2 + 14	Opposing Position: Opacity + Scene2 -> Base + Scene3 -> Video + Scene1 -> Ray + Scene4 Side By Side Position: Base + Scene12 -> Ray + Scene5 -> Opacity + Scene11 -> Video + Scene6 90 Degree Position: Ray + Scene10 -> Video + Scene7 -> Base + Scene9 -> Opacity + Scene8
ID: 3 + 15	Side by Side Position: Video + Scene3 -> Opacity + Scene4 -> Ray + Scene2 -> Base + Scene5 Opposing Position: Opacity + Scene1 -> Base + Scene6 -> Video + Scene12 -> Ray + Scene7 90 Degree Position: Base + Scene11 -> Ray + Scene8 - > Opacity + Scene10 -> Video + Scene9
ID: 4 + 16	Side by Side Position: Ray + Scene4 -> Video + Scene5 -> Base + Scene3 -> Opacity + Scene6 90 Degree Position: Video + Scene2 -> Opacity + Scene7 -> Ray + Scene1 -> Base + Scene8 Opposing Position: Opacity + Scene12 -> Base + Scene9 -> Video + Scene 11 -> Ray + Scene10
ID: 5 + 17	90 Degree Position: Base + Scene5 -> Ray + Scene6 -> Opacity + Scene4 -> Video + Scene7 Side By Side Position: Ray + Scene3 -> Video + Scene8 -> Base + Scene2 -> Opacity + Scene9 Opposing Position: Video + Scene1 -> Opacity + Scene10 -> Ray + Scene12 -> Base + Scene11
ID: 6 + 18	90 Degree Position: Opacity + Scene6 -> Base + Scene7 -> Video + Scene5 -> Ray + Scene8 Opposing Position: Base + Scene4 -> Ray + Scene9 -> Opacity + Scene3 -> Video + Scene10 Side By Side Position: Ray + Scene2 -> Video + Scene11 -> Base + Scene1 -> Opacity + Scene12
ID: 7 + 19	Opposing Position: Video + Scene7 -> Opacity + Scene8 -> Ray + Scene6 -> Base + Scene9 90Degree Position: Opacity + Scene5 -> Base + Scene10 -> Video + Scene4 -> Ray + Scene11 Side By Side Position: Base + Scene3 -> Ray + Scene 12 -> Opacity + Scene2 -> Video + Scene1
ID: 8 + 20	Opposing Position: Ray + Scene8 -> Video + Scene9 -> Base + Scene7 -> Opacity + Scene10 Side By Side Position: Video + Scene6 -> Opacity + Scene11 -> Ray + Scene5 -> Base + Scene12 90 Degree Position: Opacity + Scene4 -> Base + Scene1 -> Video + Scene3 -> Ray + Scene2
ID: 9 + 21	Side By Side Position: Base + Scene9 -> Ray + Scene10 -> Opacity + Scene8 -> Video + Scene11 Opposing Position: Ray + Scene7 -> Video + Scene12 -> Base + Scene6 -> Opacity + Scene1 90Degree Position: Video + Scene5 -> Opacity + Scene2 -> Ray + Scene4 -> Base + Scene3
ID: 10 + 22	Side By Side Position: Opacity + Scene10 -> Base + Scene11 -> Video + Scene9 -> Ray + Scene12 90 Degree Position: Base + Scene8 -> Ray + Scene1 -> Opacity + Scene7 -> Video + Scene2 Opposing Position: Ray + Scene6 -> Video + Scene3 -> Base + Scene5 -> Opacity + Scene4
ID: 11 + 23	90 Degree Position Video + Scene11 -> Opacity + Scene12 -> Ray + Scene10 -> Base + Scene1 Side By Side Position: Opacity + Scene9 -> Base + Scene2 -> Video + Scene8 -> Ray + Scene3 Opposing Position: Base + Scene7 -> Ray + Scene4 -> Opacity + Scene6 -> Video + Scene5
ID: 12 + 24	90 Degree Position: Ray + Scene12 -> Video + Scene1 -> Base + Scene11 -> Opacity + Scene2 Opposing Position: Video + Scene10 -> Opacity + Scene3 -> Ray + Scene9 -> Base + Scene4 Side By Side Position: Opacity + Scene8 -> Base + Scene5 -> Video + Scene7 -> Ray + Scene6

Figure A.8: The counterbalanced combination of the different Latin squares.

Appendix B

Scenes and Additional Graphs

The following images depict all thirteen scenes (demo + 12 actual scenes) used in the study. Additionally, graphs that had no positional split and the ones where we only described the results, but did not include a graphical representation, are shown here with the positional split included.

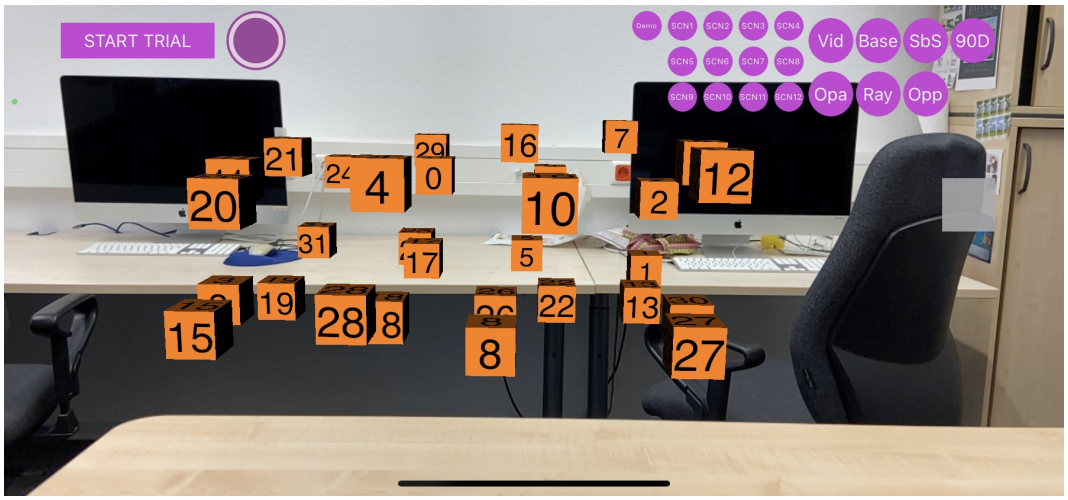

Figure B.1: The demo scene.

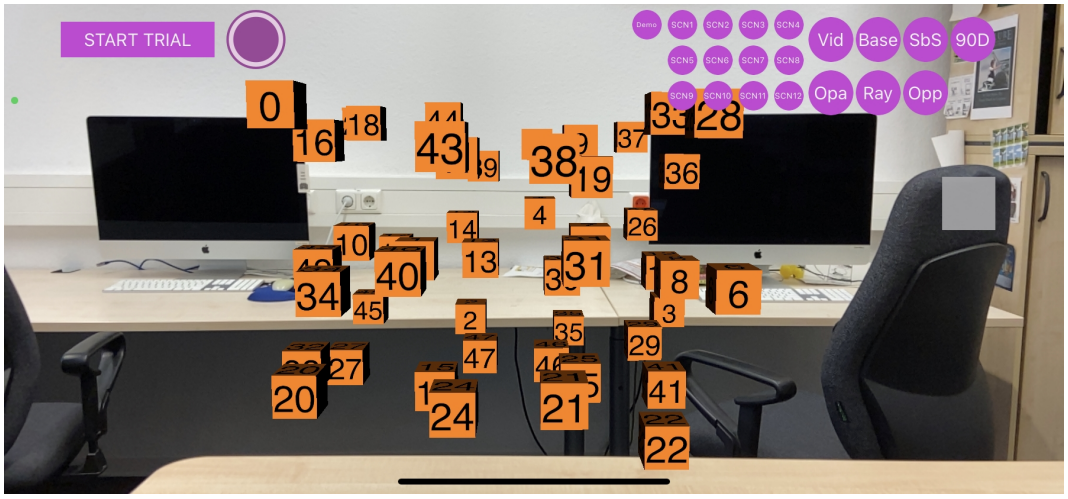

Figure B.2: Scene numbered as 1 in the user study.

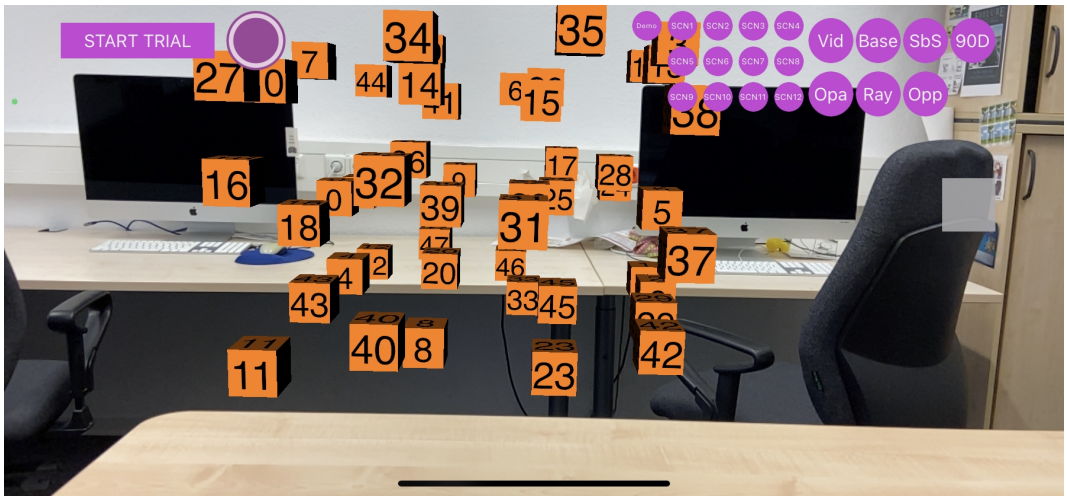

Figure B.3: Scene numbered as 2 in the user study.

Figure B.4: Scene numbered as 3 in the user study.

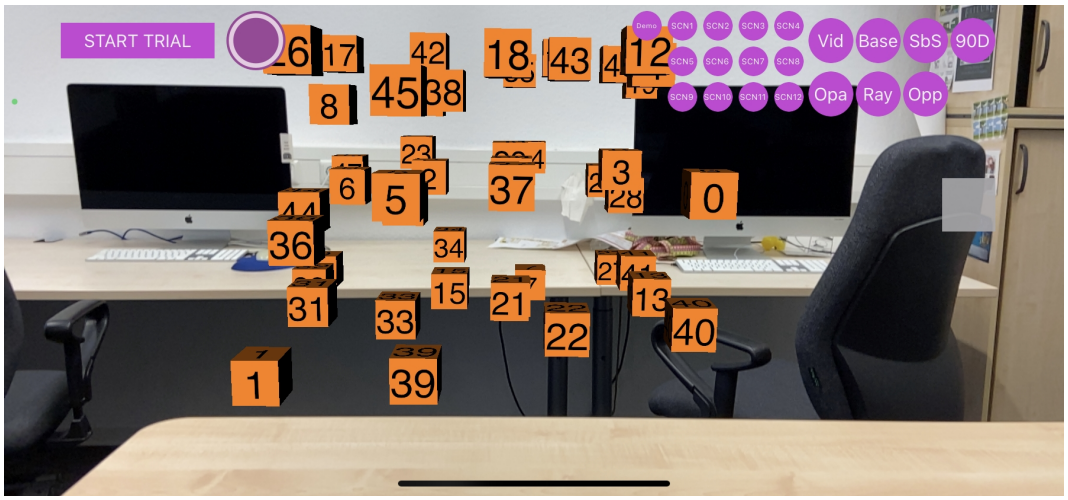

Figure B.5: Scene numbered as 4 in the user study.

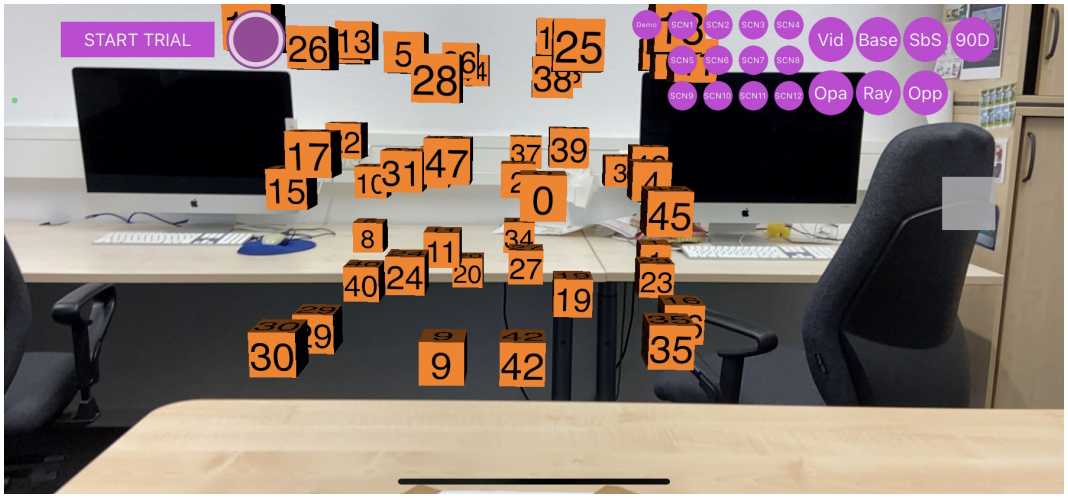

Figure B.6: Scene numbered as 5 in the user study.

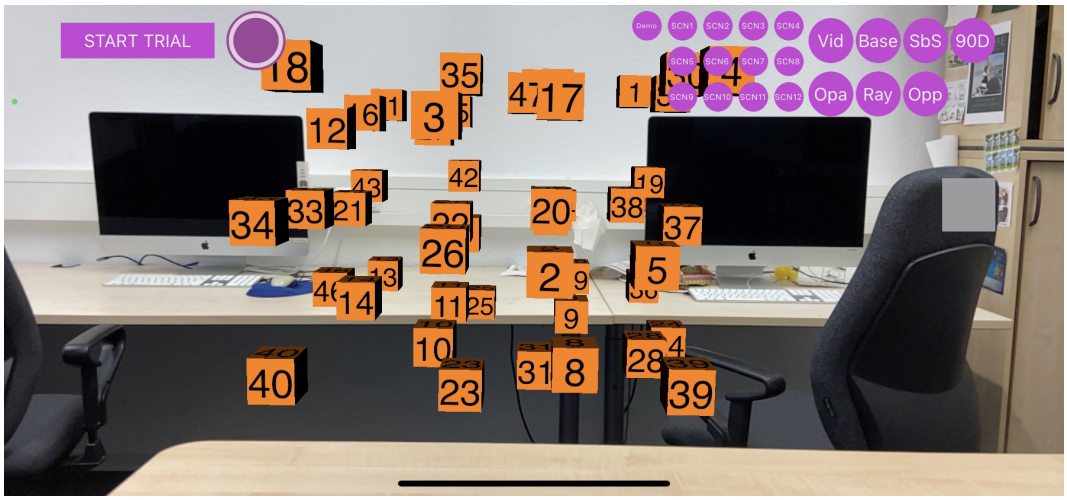

Figure B.7: Scene numbered as 6 in the user study.

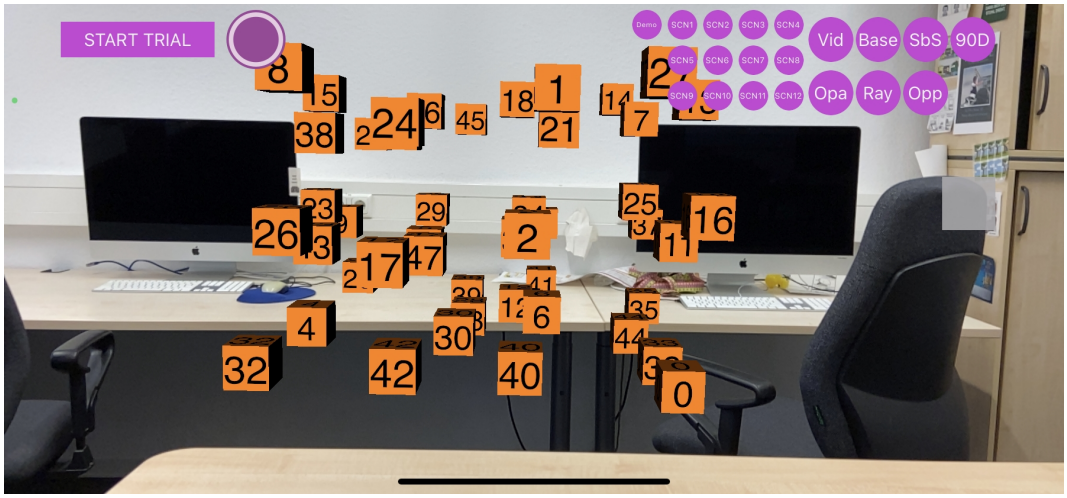

Figure B.8: Scene numbered as 7 in the user study.

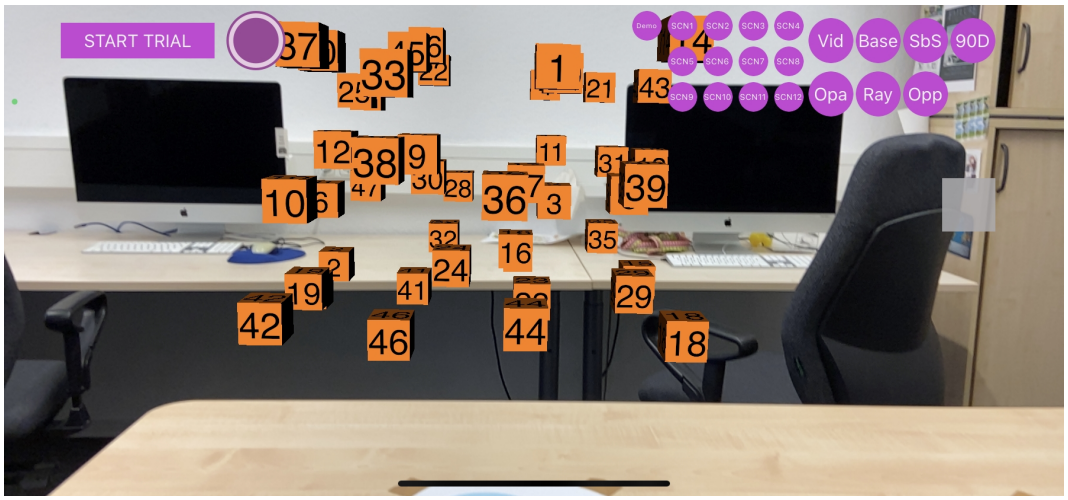

Figure B.9: Scene numbered as 8 in the user study.

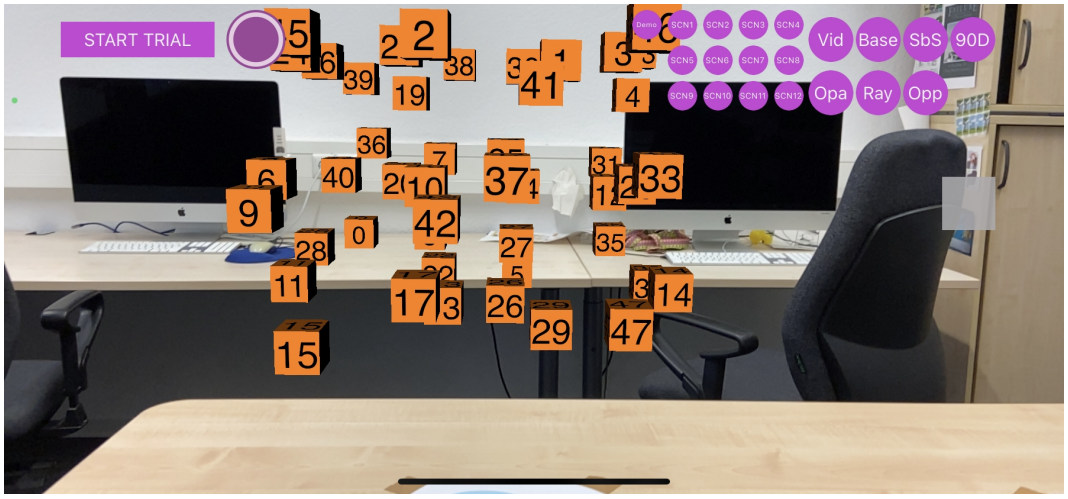

Figure B.10: Scene numbered as 9 in the user study.

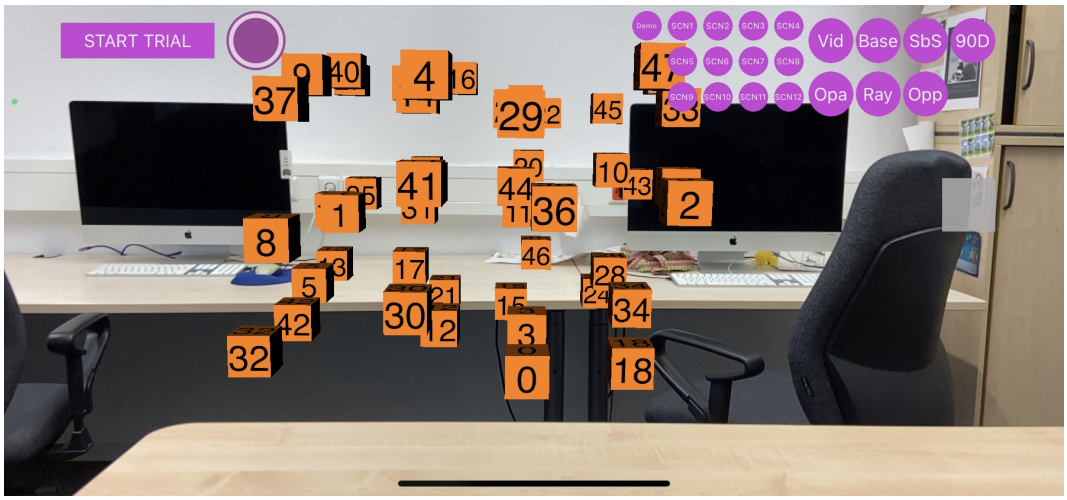

Figure B.11: Scene numbered as 10 in the user study.

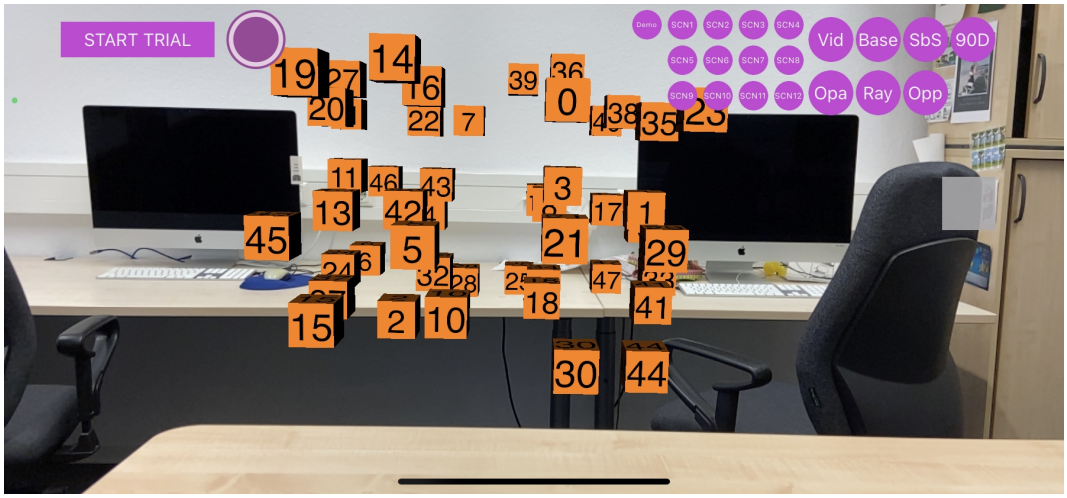

Figure B.12: Scene numbered as 11 in the user study.

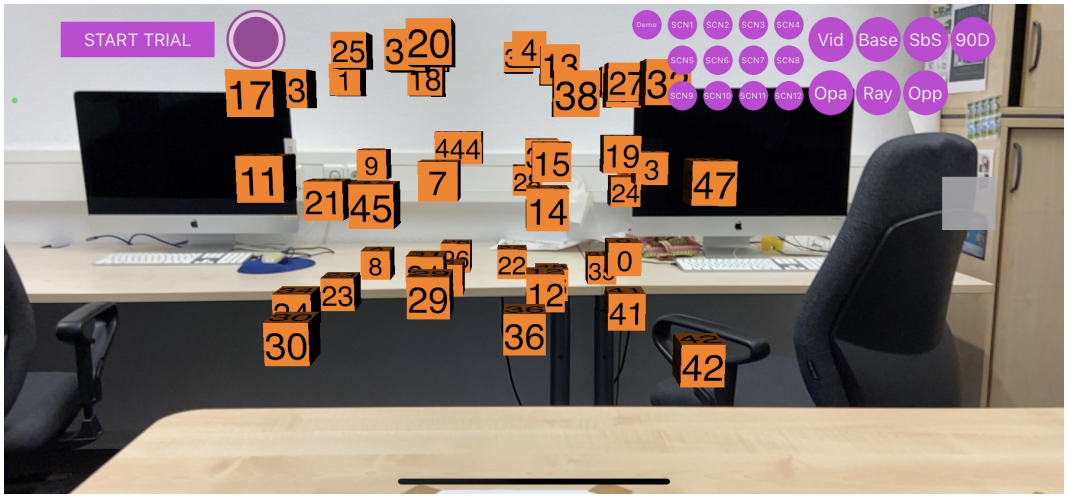

Figure B.13: Scene numbered as 12 in the user study.

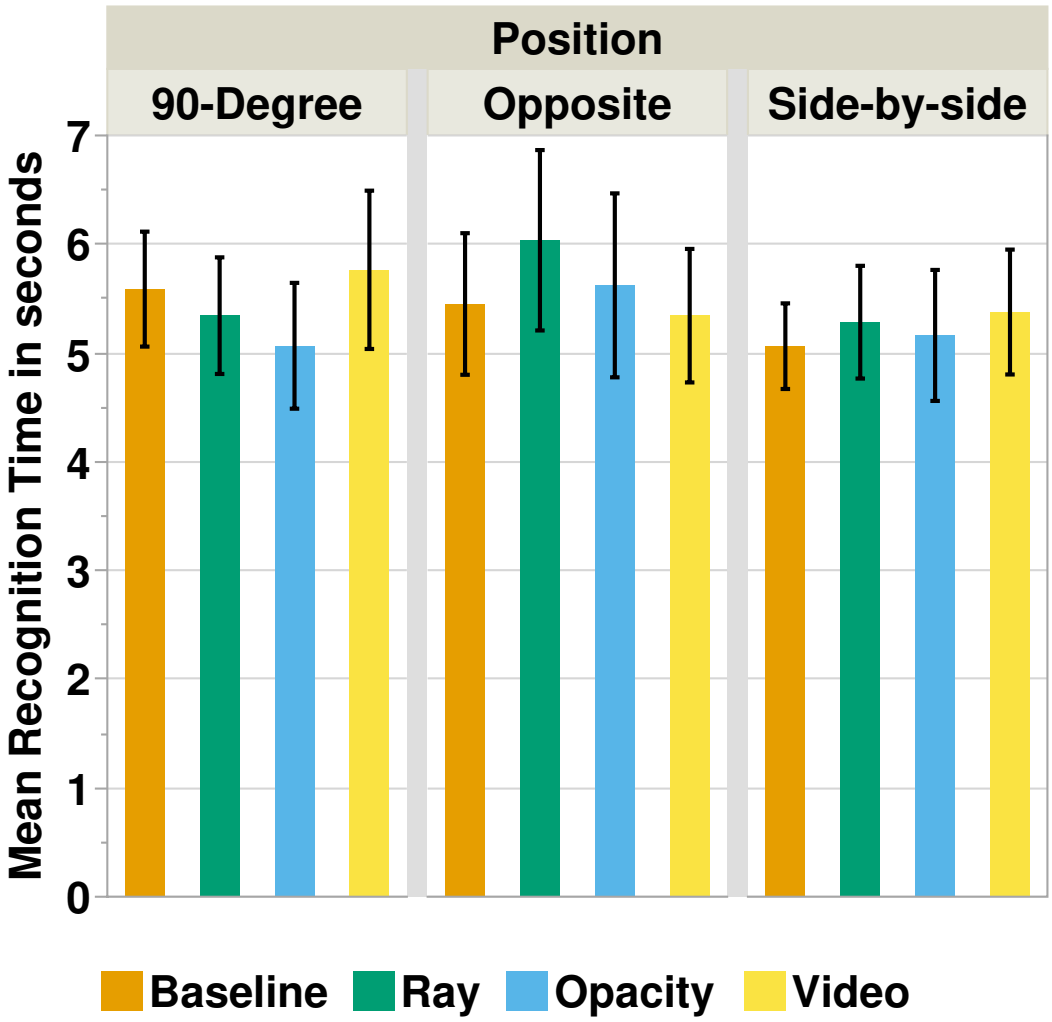

Figure B.14: The effect of *Mode* and *Position* on the mean *Recognition Time*. Whiskers denote the 95% CI.

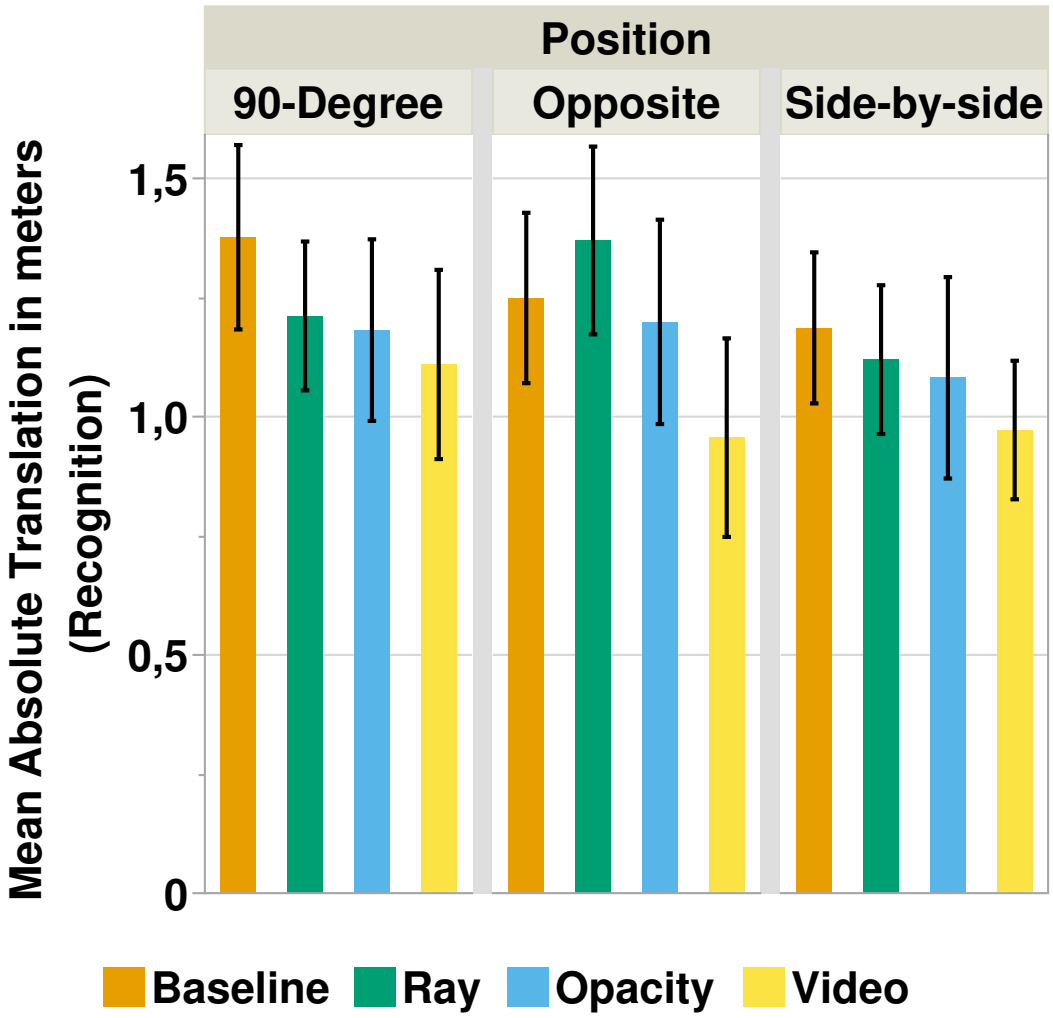

Figure B.15: The effect of *Mode* and *Position* on the mean *Absolute Translation* (Recognition). Whiskers denote the 95% CI.

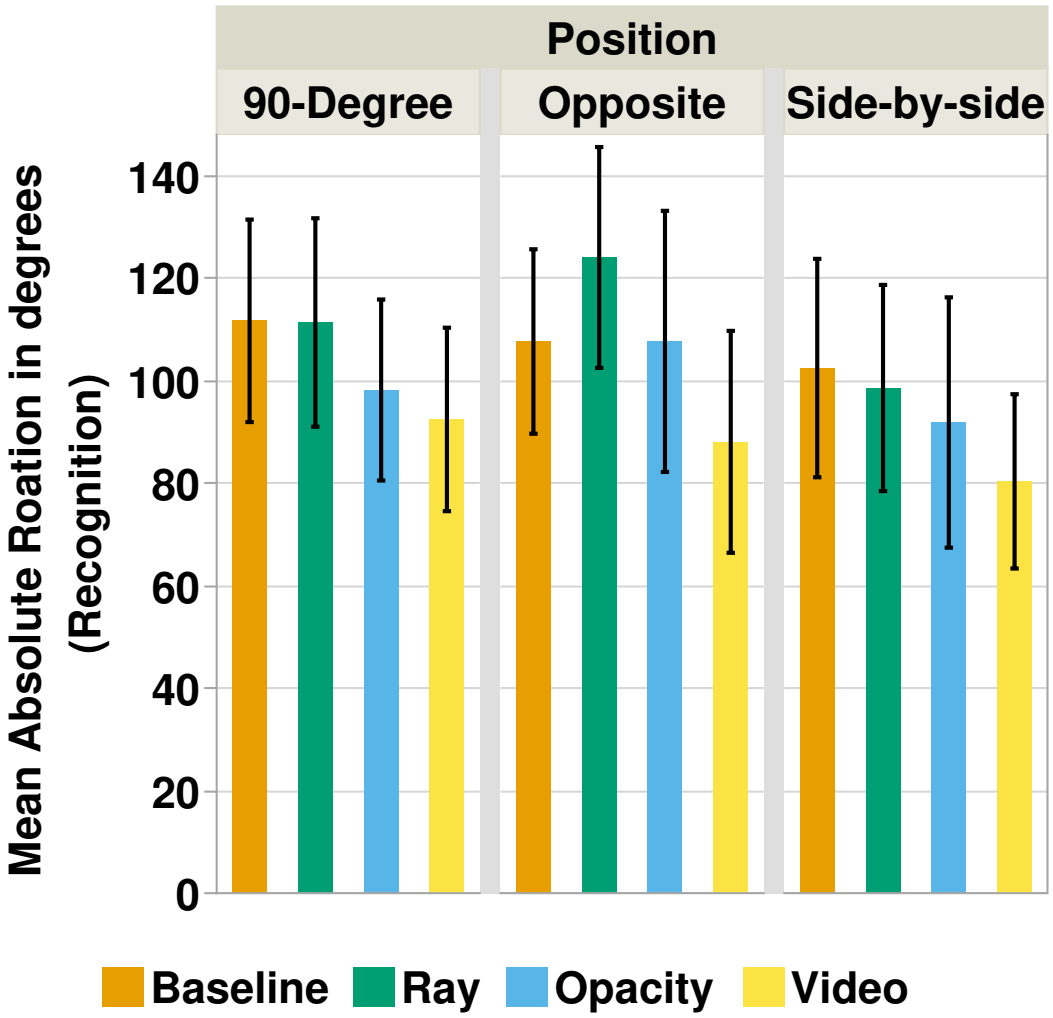

Figure B.16: The effect of *Mode* and *Position* on the mean *Absolute Rotation* (Recognition). Whiskers denote the 95% CI.

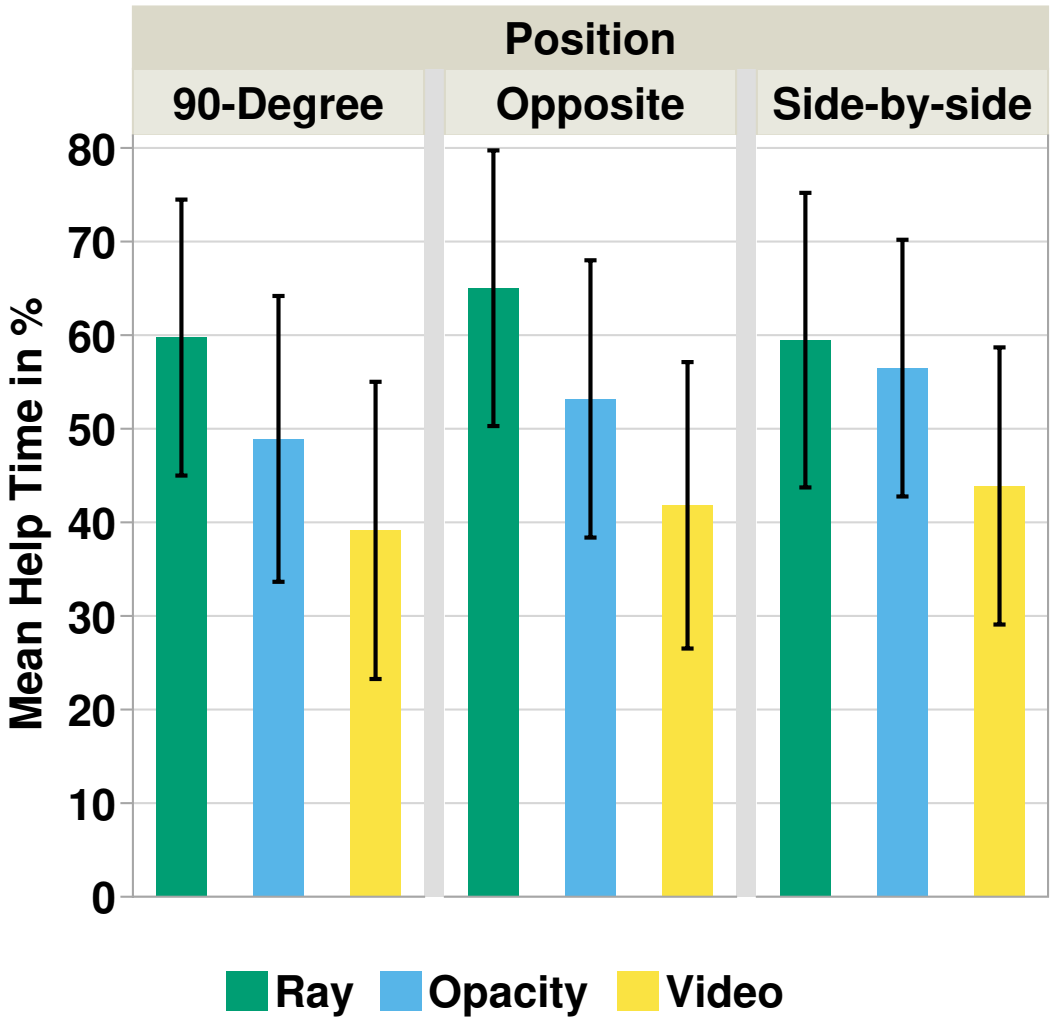

Figure B.17: The effect of *Mode* and *Position* on the mean *Help Time Percent*. Whiskers denote the 95% CI.

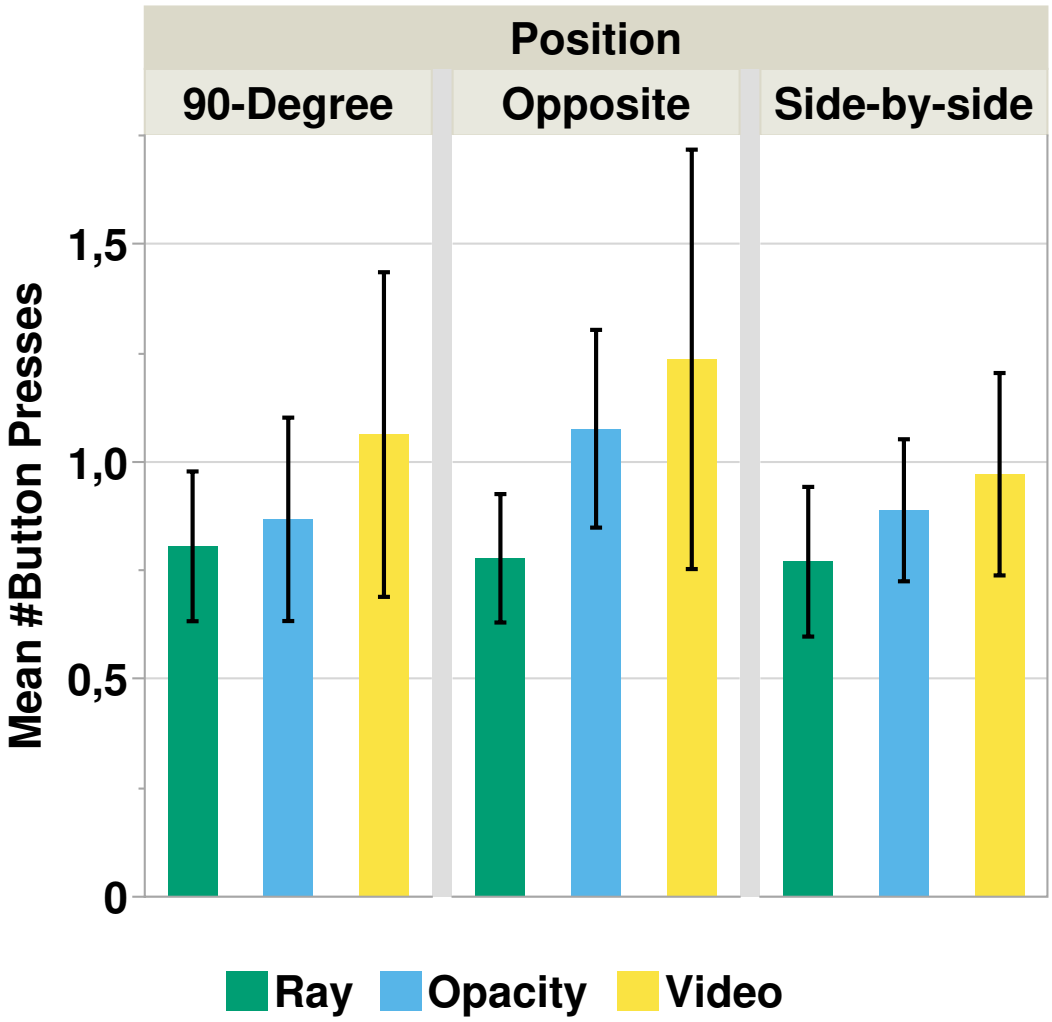

Figure B.18: The effect of *Mode* and *Position* on the mean *Button Presses*. Whiskers denote the 95% CI.

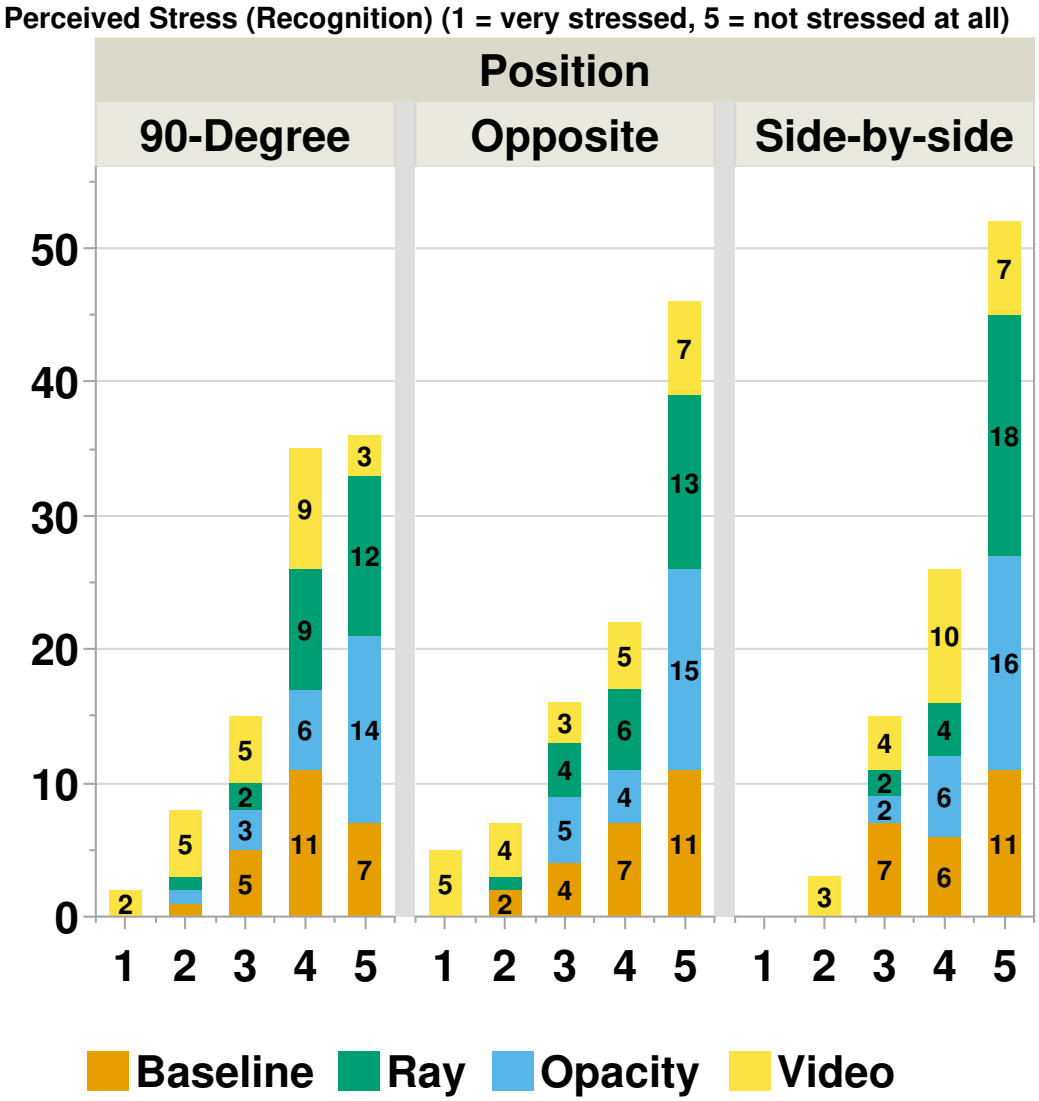

Figure B.19: The effect of *Mode* and *Position* on the mean *Perceived Stress* (Recognition). Whiskers denote the 95% CI.

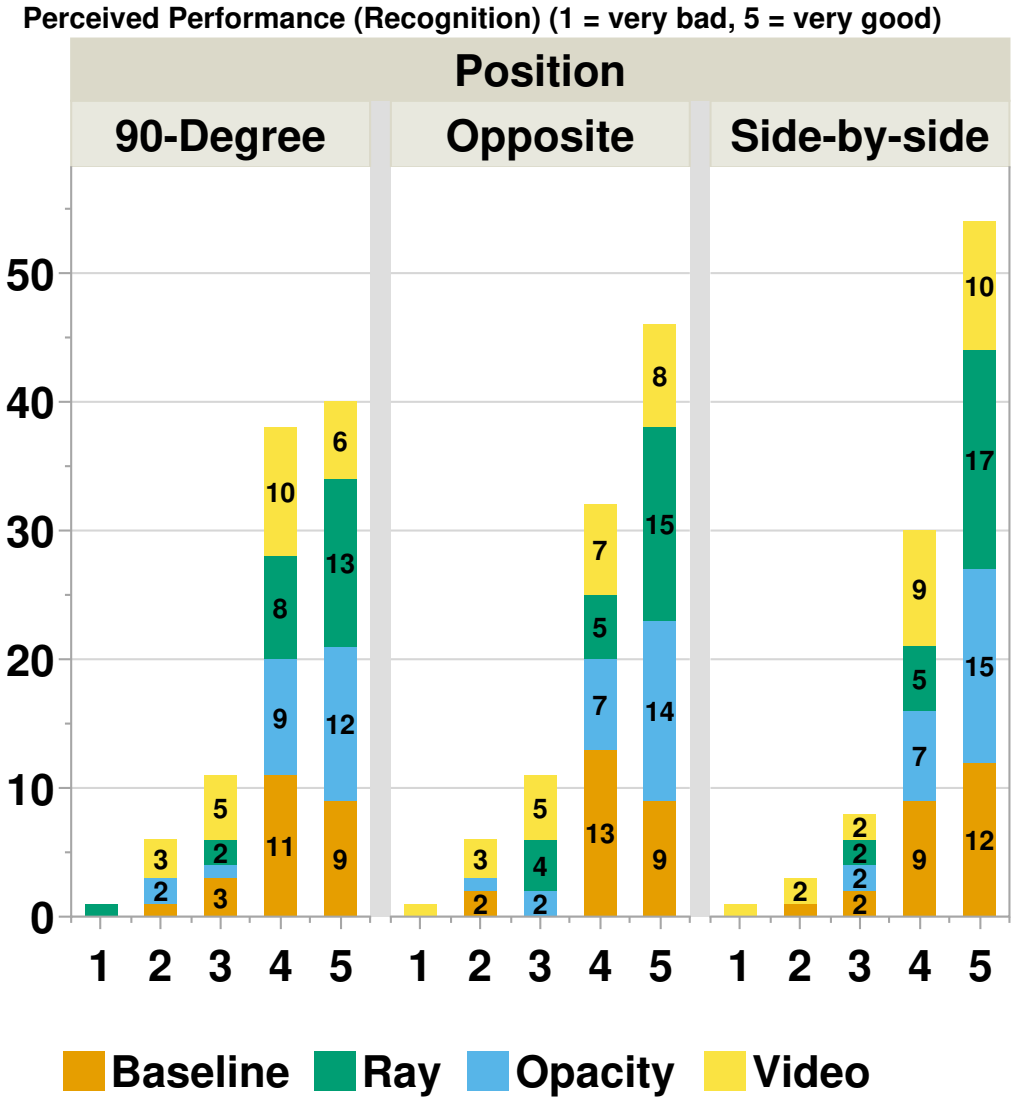

Figure B.20: The effect of *Mode* and *Position* on the mean *Perceived Performance* (Recognition). Whiskers denote the 95% CI.

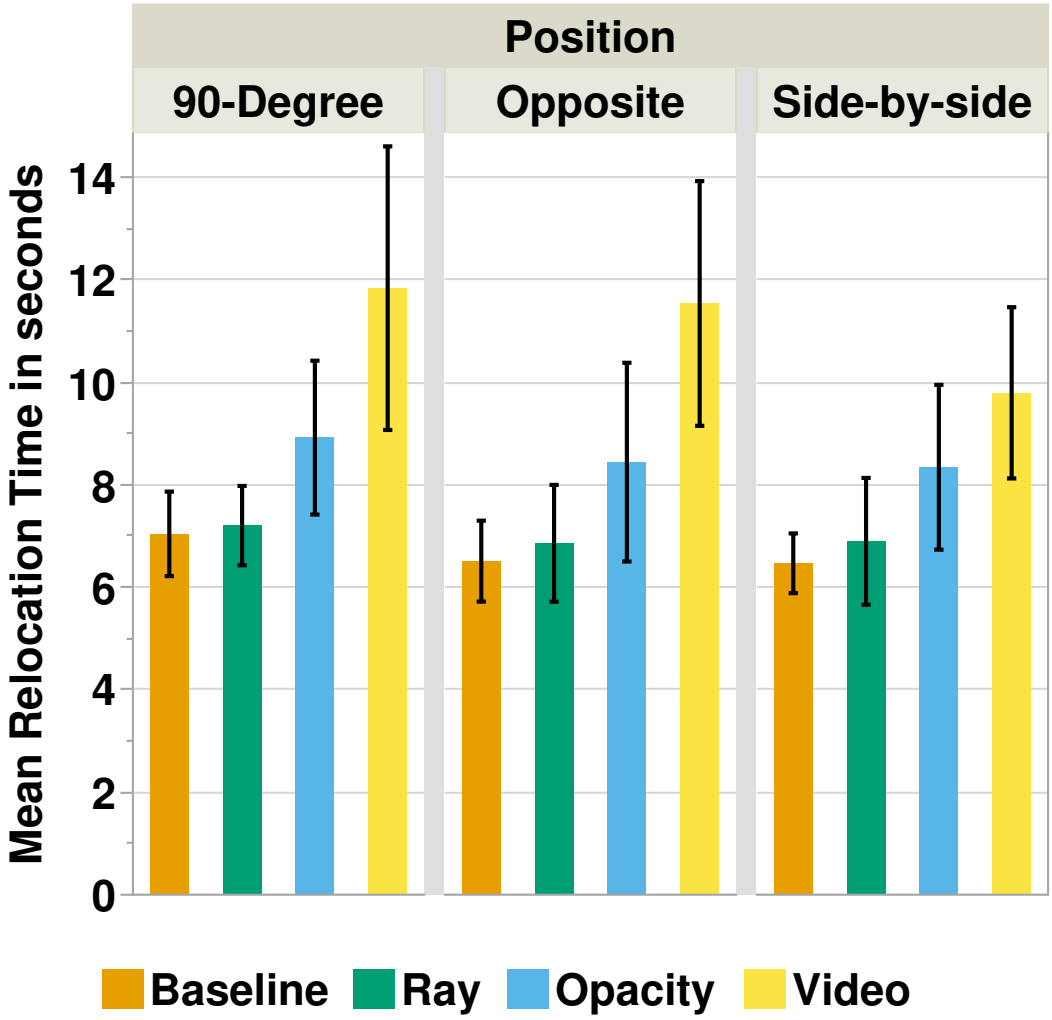

Figure B.21: The effect of *Mode* and *Position* on the mean *Relocation Time*. Whiskers denote the 95% CI.

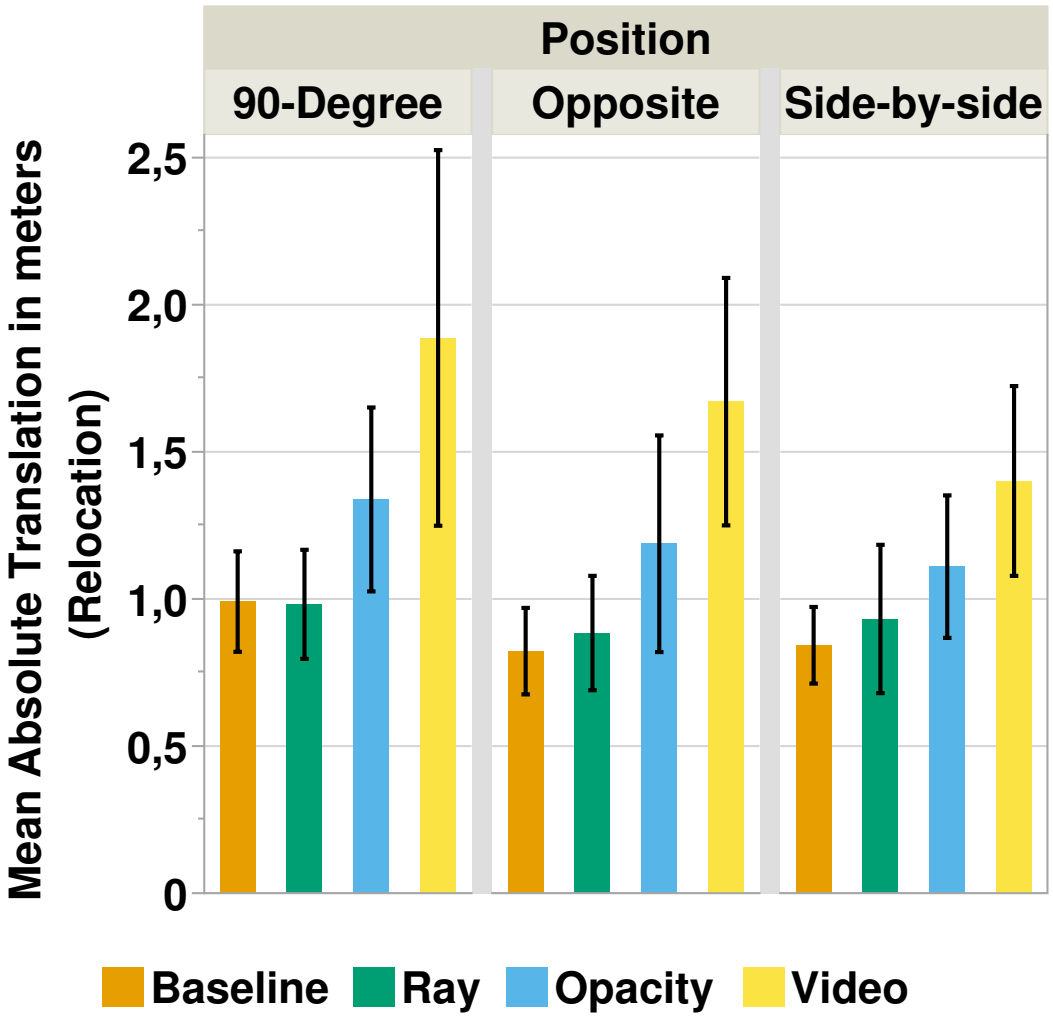

Figure B.22: The effect of *Mode* and *Position* on the mean *Absolute Translation* (Relocation). Whiskers denote the 95% CI.

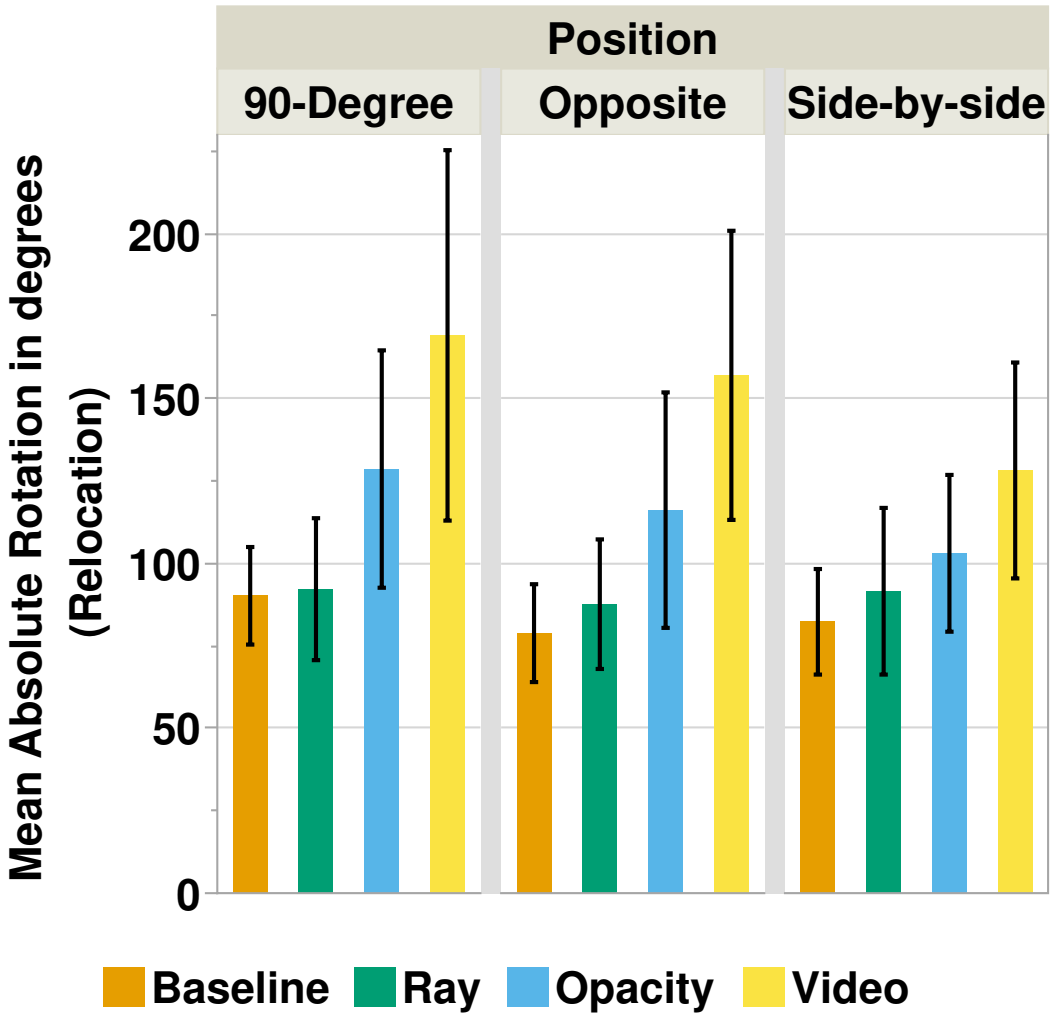

Figure B.23: The effect of *Mode* and *Position* on the mean *Absolute Rotation* (Relocation). Whiskers denote the 95% CI.

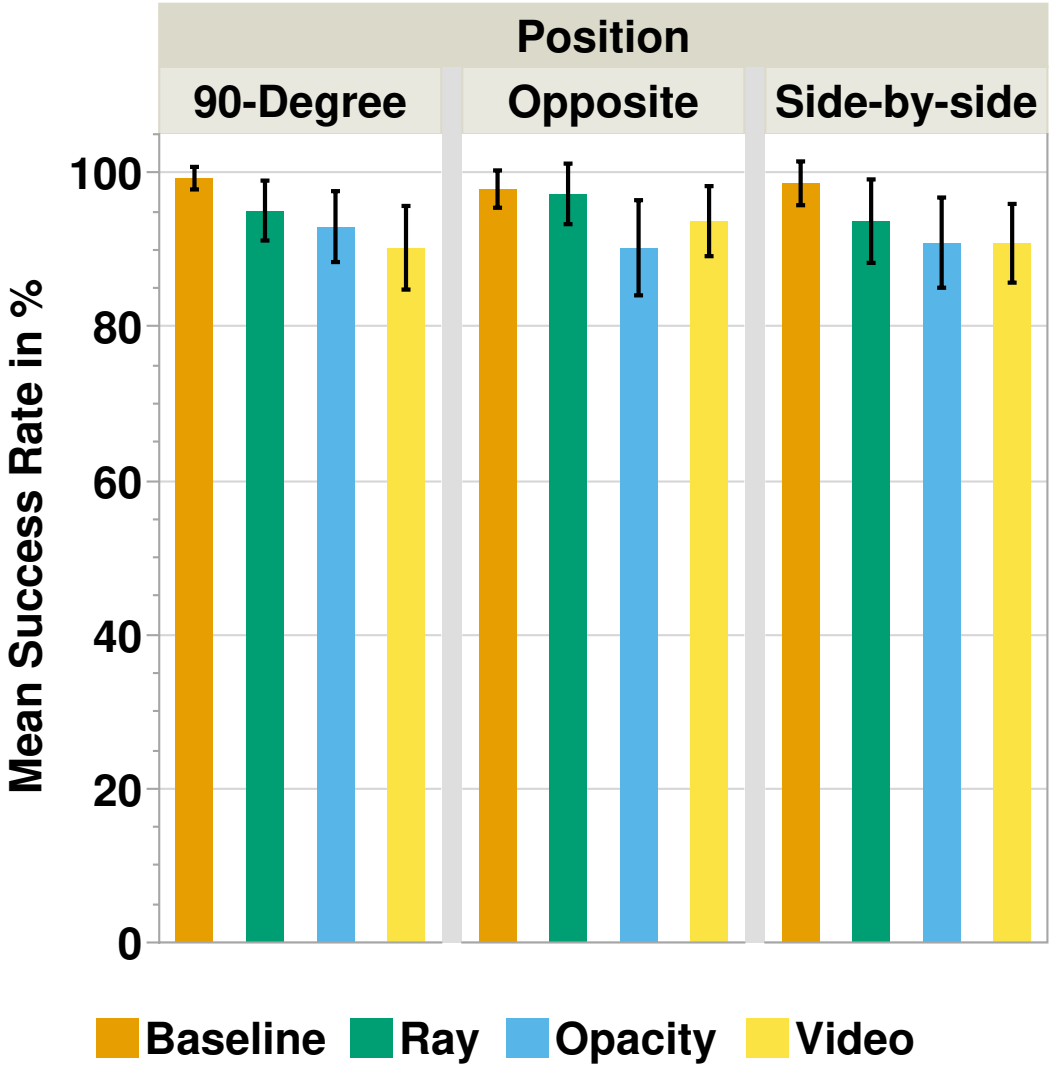

Figure B.24: The effect of *Mode* and *Position* on the mean *Success Rate*. Whiskers denote the 95% CI.

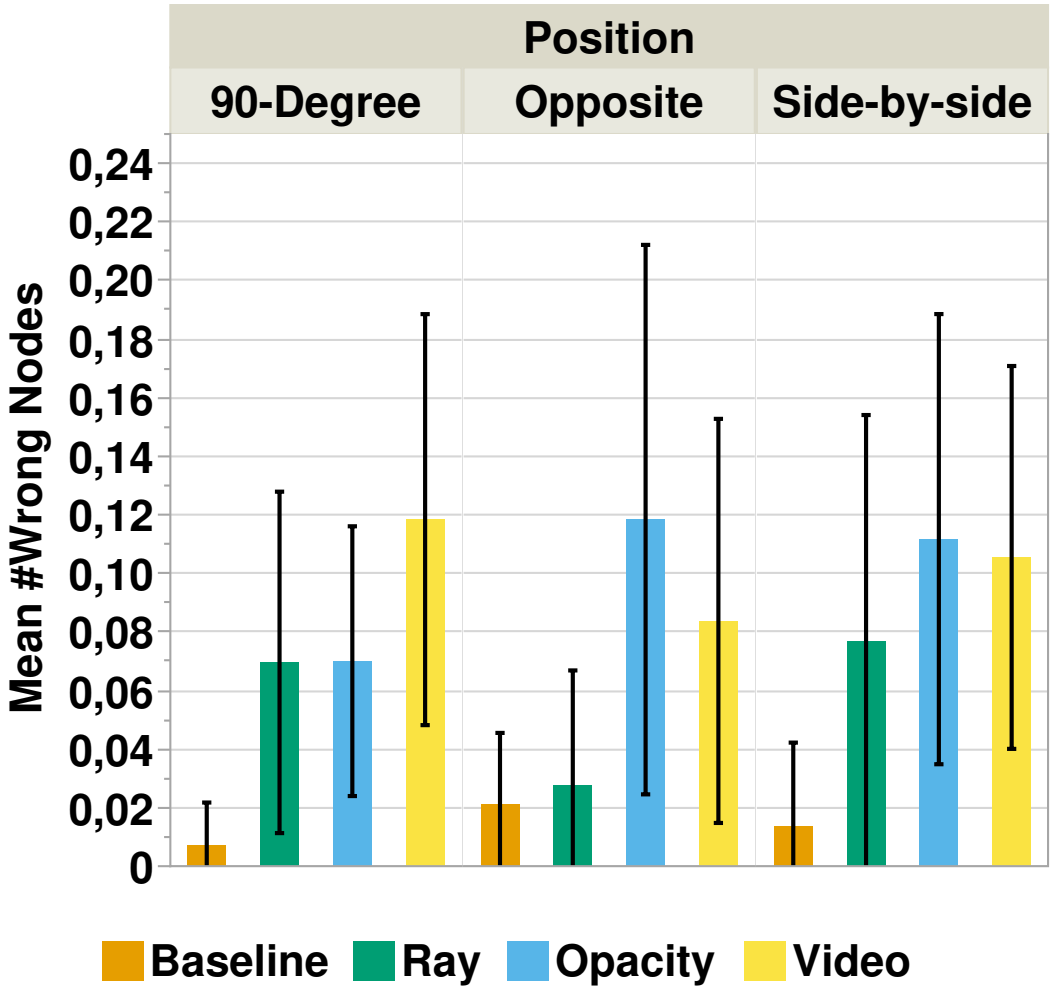

Figure B.25: The effect of *Mode* and *Position* on the mean number of *Wrong Nodes*. Whiskers denote the 95% CI.

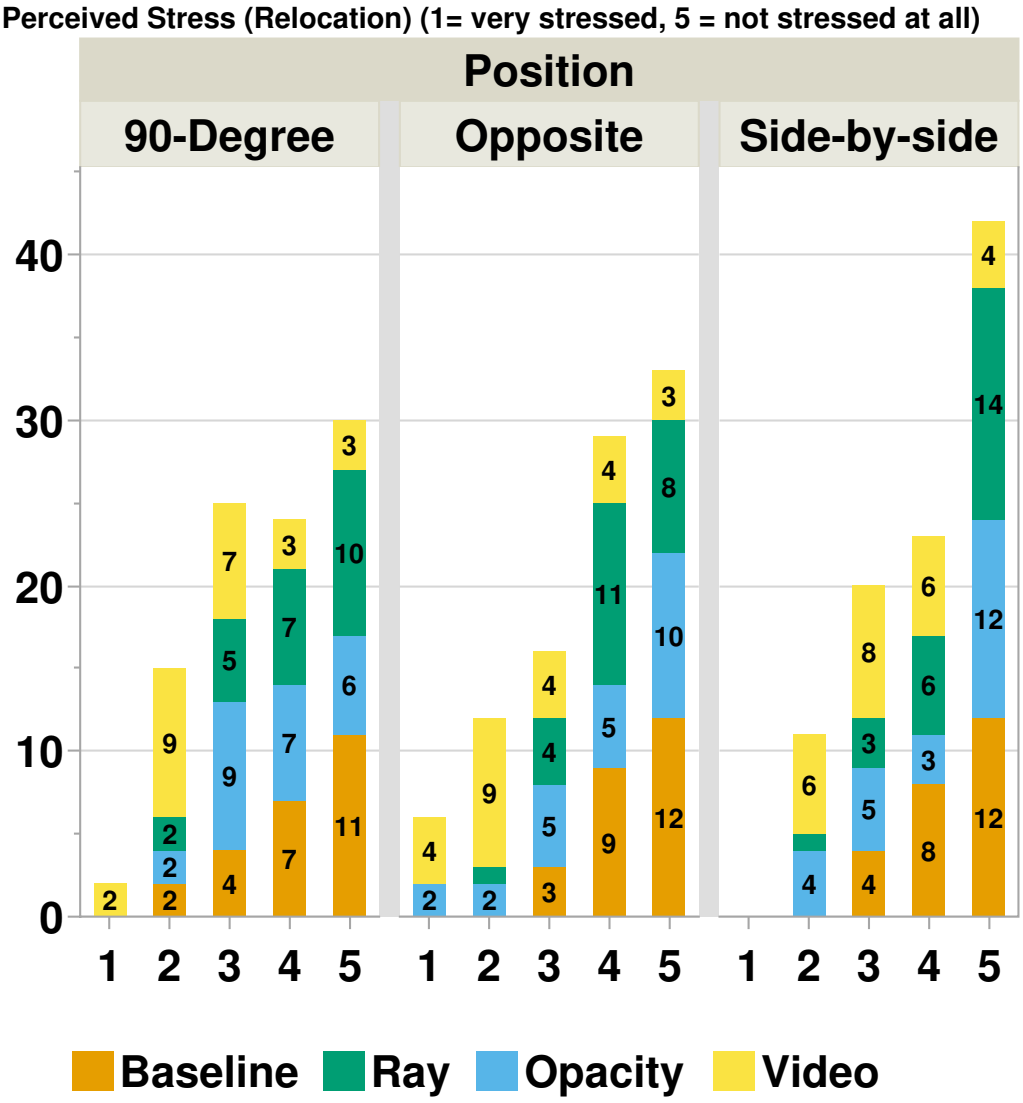

Figure B.26: The effect of *Mode* and *Position* on the mean *Perceived Stress* (Relocation). Whiskers denote the 95% CI.

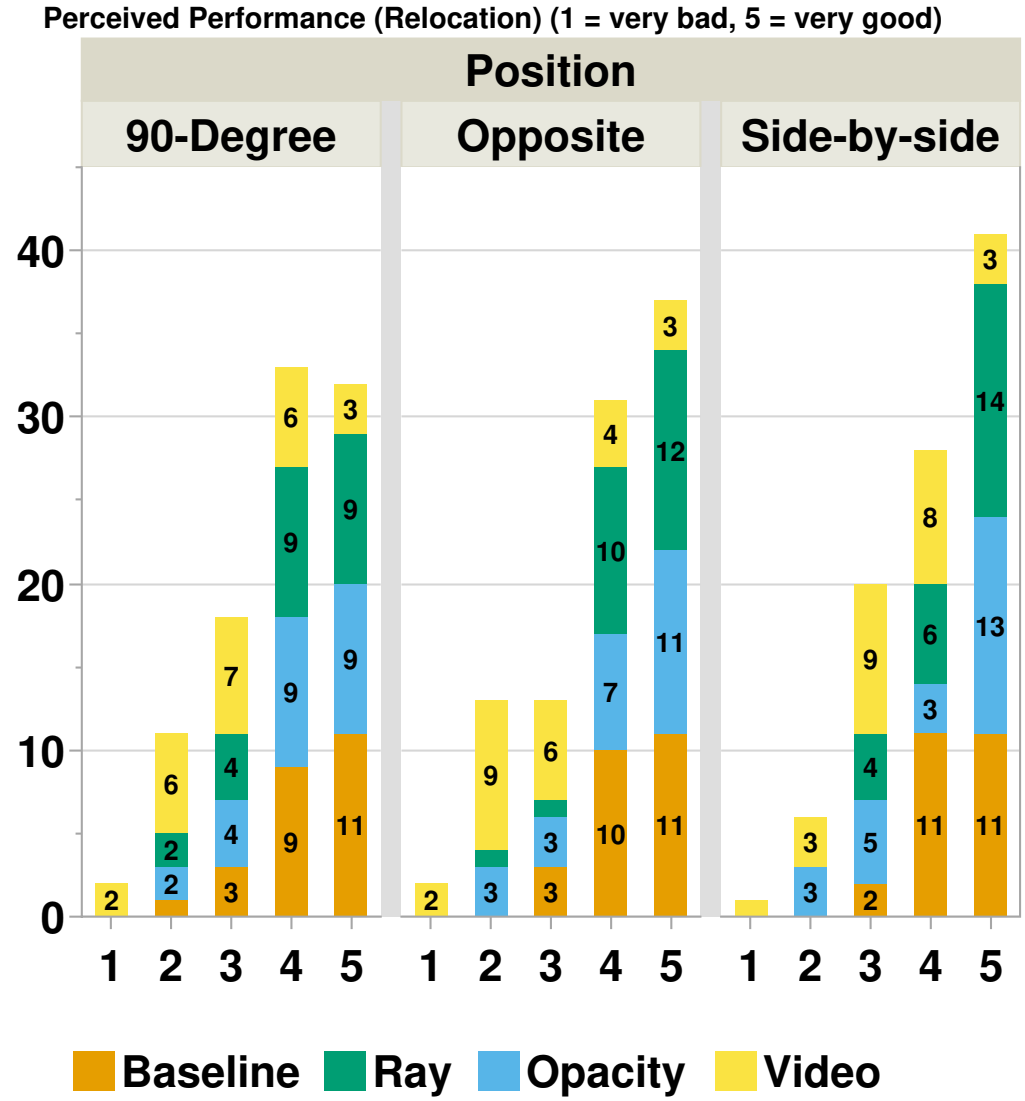

Figure B.27: The effect of *Mode* and *Position* on the mean *Perceived Performance* (Relocation). Whiskers denote the 95% CI.

Bibliography

Ferran Argelaguet, Alexander Kulik, André Kunert, Carlos Andujar, and Bernd Froehlich. See-through techniques for referential awareness in collaborative virtual reality. *International Journal of Human-Computer Studies*, 69(6):387–400, 2011. ISSN 1071-5819. doi: https://doi.org/10.1016/j.ijhcs.2011.01.003. URL https://www.sciencedirect.com/science/article/pii/S107158191100005X.

Ronald T. Azuma. A survey of augmented reality. *Presence: Teleoper. Virtual Environ.*, 6(4):355–385, aug 1997. ISSN 1054-7460. doi: 10.1162/pres.1997.6.4.355. URL https://doi.org/10.1162/pres.1997.6.4.355.

BMW. Das cabrio in ihren händen! die launchkampagne des neuen mini cabrios überrascht mit neuen kommunikationsmaßnahmen. https://www.press.bmwgroup.com/deutschland/article/detail/T0005444DE, December 1st, 2008. [Online; accessed October-12th-2022].

Lei Chen, Yilin Liu, Yue Li, Lingyun Yu, BoYu Gao, Maurizio Caon, Yong Yue, and Hai-Ning Liang. Effect of visual cues on pointing tasks in co-located augmented reality collaboration. In *Symposium on Spatial User Interaction*, SUI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450390910. doi: 10.1145/3485279.3485297. URL https://doi.org/10.1145/3485279.3485297.

Cheng-Yu Chung, Nayif Awad, and I-Han Hsiao. Collaborative programming problem-solving in augmented reality: Multimodal analysis of effectiveness and group col-

laboration. *Australasian Journal of Educational Technology*, 37(5):17–31, 2021.

Herbert H. Clark and Susan E. Brennan. Grounding in communication. In Lauren Resnick, Levine B., M. John, Stephanie Teasley, and D., editors, *Perspectives on Socially Shared Cognition*, pages 127–149. American Psychological Association, 1991.

Carolina Cruz-Neira, Daniel J Sandin, Thomas A DeFanti, Robert V Kenyon, and John C Hart. The cave: audio visual experience automatic virtual environment. *Communications of the ACM*, 35(6):64–73, 1992.

Susan R. Fussell, Leslie D. Setlock, and Robert E. Kraut. Effects of head-mounted and scene-oriented video systems on remote collaboration on physical tasks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '03, page 513–520, New York, NY, USA, 2003. Association for Computing Machinery. ISBN 1581136307. doi: 10.1145/642611.642701. URL https://doi.org/10.1145/642611.642701.

Lei Gao, Huidong Bai, Gun Lee, and Mark Billinghurst. An oriented point-cloud view for mr remote collaboration. In *SIGGRAPH ASIA 2016 Mobile Graphics and Interactive Applications*, SA '16, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450345514. doi: 10.1145/2999508.2999531. URL https://doi.org/10.1145/2999508.2999531.

Steffen Gauglitz, Benjamin Nuernberger, Matthew Turk, and Tobias Höllerer. World-stabilized annotations and virtual scene navigation for remote collaboration. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, UIST '14, page 449–459, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450330695. doi: 10.1145/2642918.2647372. URL https://doi.org/10.1145/2642918.2647372.

Jeremy Hartmann and Daniel Vogel. An evaluation of mobile phone pointing in spatial augmented reality. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI EA '18, page 1–6,

New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356213. doi: 10.1145/3170427.3188535. URL https://doi.org/10.1145/3170427.3188535.

Adrian Hoppe, Kai Westerkamp, Sebastian Maier, Florian Camp, and Rainer Stiefelhagen. *Multi-user Collaboration on Complex Data in Virtual and Augmented Reality*, pages 258–265. 06 2018. ISBN 978-3-319-92278-2. doi: 10.1007/978-3-319-92279-9_35.

Weidong Huang and Leila Alem. Supporting hand gestures in mobile remote collaboration: A usability evaluation. In *Proceedings of the 25th BCS Conference on Human-Computer Interaction*, BCS-HCI '11, page 211–216, Swindon, GBR, 2011. BCS Learning & Development Ltd.

Ernst Kruijff, J. Edward Swan, and Steven Feiner. Perceptual issues in augmented reality revisited. In *2010 IEEE International Symposium on Mixed and Augmented Reality*, pages 3–12, 2010. doi: 10.1109/ISMAR.2010.5643530.

Lenovo. Lenovo tech world 2022, press release. https://news.lenovo.com/pressroom/press-releases/smarter-tech-innovations-define-future-of-digital-world/, October 18th, 2022. [Online; accessed October-28th-2022].

Thomas Ludwig, Oliver Stickel, Peter Tolmie, and Malte Sellmer. share-it: Ad hoc remote troubleshooting through augmented reality. *Computer Supported Cooperative Work (CSCW)*, 30, 02 2021. doi: 10.1007/s10606-021-09393-5.

Domenick M. Mifsud, Adam S. Williams, Francisco Ortega, and Robert J. Teather. Augmented reality fitts' law input comparison between touchpad, pointing gesture, and raycast. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 590–591, 2022. doi: 10.1109/VRW55335.2022.00146.

Lev Poretski, Joel Lanir, Ram Margalit, and Ofer Arazy. Physicality as an anchor for coordination: Examining collocated collaboration in physical and mobile augmented reality settings. *Proceedings of the ACM on Human-Computer Interaction*, 5, 08 2021. doi: 10.1145/3479857.

Sara Price and Adam Jaworski. Mode (2012). glossary of multimodal terms. https://multimodalityglossary.wordpress.com/gesture/, 2012. [Online; accessed December-2nd-2022].

Ivan E. Sutherland. A head-mounted three dimensional display. In *Proceedings of the December 9-11, 1968, Fall Joint Computer Conference, Part I*, AFIPS '68 (Fall, part I), page 757–764, New York, NY, USA, 1968. Association for Computing Machinery. ISBN 9781450378994. doi: 10.1145/1476589.1476686. URL https://doi.org/10.1145/1476589.1476686.

Matthew Tait and Mark Billinghurst. The effect of view independence in a collaborative ar system. *Comput. Supported Coop. Work*, 24(6):563–589, dec 2015. ISSN 0925-9724. doi: 10.1007/s10606-015-9231-8. URL https://doi.org/10.1007/s10606-015-9231-8.

Balasaravanan Thoravi Kumaravel, Fraser Anderson, George Fitzmaurice, Bjoern Hartmann, and Tovi Grossman. Loki: Facilitating remote instruction of physical tasks using bi-directional mixed-reality telepresence. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, UIST '19, page 161–174, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450368162. doi: 10.1145/3332165.3347872. URL https://doi.org/10.1145/3332165.3347872.

Philipp Wacker, Oliver Nowak, Simon Voelker, and Jan Borchers. Arpen: Mid-air object manipulation techniques for a bimanual ar system with pen & smartphone. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–12, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300849. URL https://doi.org/10.1145/3290605.3300849.

Thomas Wells and Steven Houben. Collabar - investigating the mediating role of mobile ar interfaces on co-located group collaboration. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA, 2020. Association for

Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376541. URL https://doi.org/10.1145/3313831.3376541.

Feng Zhou, Henry Been-Lirn Duh, and Mark Billinghurst. Trends in augmented reality tracking, interaction and display: A review of ten years of ismar. In *2008 7th IEEE/ACM International Symposium on Mixed and Augmented Reality*, pages 193–202, 2008. doi: 10.1109/ISMAR.2008.4637362.

Index

Typeset December 5, 2022