Look Over Here! Comparing Interaction Methods for User-Assisted Remote Scene Reconstruction

Carina Liebers
carina.liebers@uni-due.de
University of Duisburg-Essen
Essen, Germany

Niklas Pfützenreuter
niklas.pfuetzenreuter@uni-due.de
University of Duisburg-Essen
Essen, Germany

Marvin Prochazka
marvin.prochazka@stud.uni-due.de
University of Duisburg-Essen
Essen, Germany

Pranav Megarajan
pranav.megarajan@offis.de
OFFIS - Institute for IT
Oldenburg, Germany

Eike Furuno
eike.furuno@offis.de
OFFIS - Institute for IT
Oldenburg, Germany

Jan Löber
jan.loeber@uni-due.de
University of Duisburg-Essen
Essen, Germany

Tim C. Stratmann
tim.stratmann@offis.de
OFFIS - Institute for IT
Oldenburg, Germany

Jonas Auda
jonas.auda@uni-due.de
University of Duisburg-Essen
Essen, Germany

Donald Degraen
donald.degraen@uni-due.de
University of Duisburg-Essen
Essen, Germany

Uwe Gruenefeld
uwe.gruenefeld@uni-due.de
University of Duisburg-Essen
Essen, Germany

Stefan Schneegass
stefan.schneegass@uni-due.de
University of Duisburg-Essen
Essen, Germany

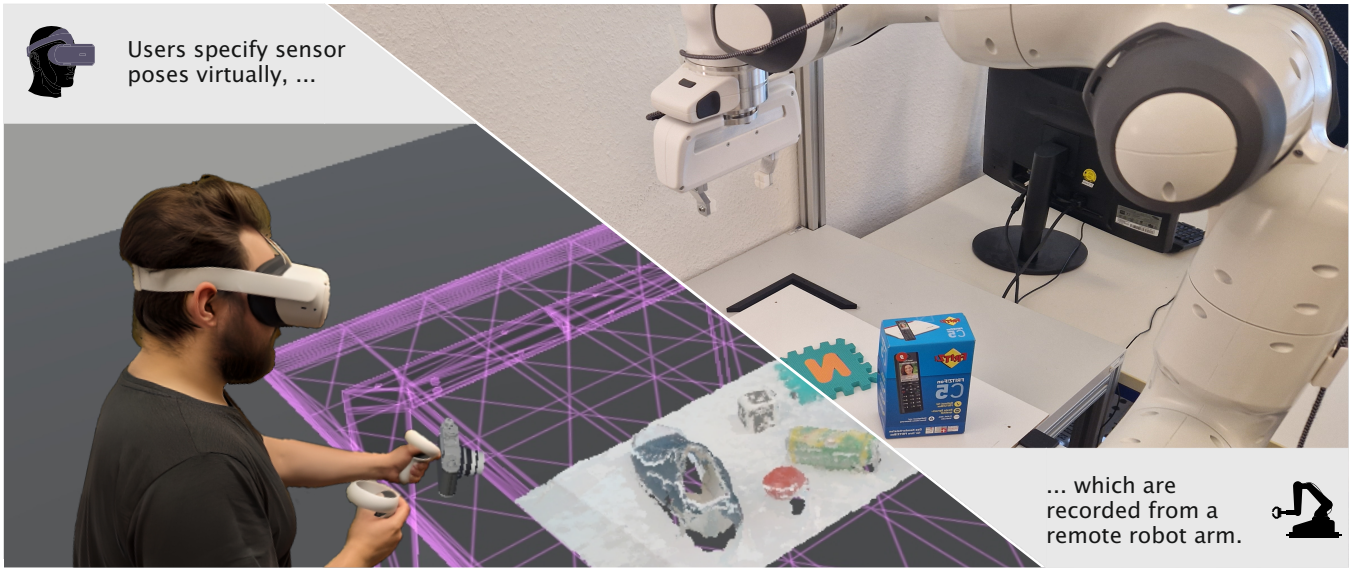

Figure 1: We enabled human assistance for scene reconstruction via the teleoperation of a static robot arm utilizing Virtual Reality. Our approach visualizes the current scan process, enabling users to provide new sensor poses for the capture (left). The robot executes the task (right), leading to an updated virtual representation.

CHI EA '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0331-7/24/05
https://doi.org/10.1145/3613905.3650982

ABSTRACT

Detailed digital representations of physical scenes are key in many cases, such as historical site preservation or hazardous area inspection. To automate the capturing process, robots or drones mounted with sensors can algorithmically record the environment from different viewpoints. However, environmental complexities often lead

to incomplete captures. We believe humans can support scene capture as their contextual understanding enables easy identification of missing areas and recording errors. Therefore, they need to perceive the recordings and suggest new sensor poses. In this work, we compare two human-centric approaches in Virtual Reality for scene reconstruction through the teleoperation of a remote robot arm, i.e., directly providing sensor poses (direct method) or specifying missing areas in the scans (indirect method). Our results show that directly providing sensor poses leads to higher efficiency and user experience. In future work, we aim to compare the quality of human assistance to automatic approaches.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**.

KEYWORDS

RGBD sampling, manual sampling, teleoperation, human-robot interaction, virtual reality

ACM Reference Format:
Carina Liebers, Niklas Pfützenreuter, Marvin Prochazka, Pranav Megarajan, Eike Furuno, Jan Löber, Tim C. Stratmann, Jonas Auda, Donald Degraen, Uwe Gruenefeld, and Stefan Schneegass. 2024. Look Over Here! Comparing Interaction Methods for User-Assisted Remote Scene Reconstruction. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3613905.3650982

1 INTRODUCTION

Virtual reconstructions of physical scenes have various use cases ranging from robot training [19], telesurgery, or documentation of archaeological excavation sites [6] and the surface of Mars [38]. A sensor system mounted to a robot or drone is a common approach to sampling its surroundings automatically [11, 30]. Scene reconstruction has many challenges, which current automatic approaches have not been able to solve entirely (i.e., scan completeness) [39, 45]. We hypothesize that humans can fill these gaps when quality is more important than shorter reconstruction times.

The number of takes and the amount of energy during recording can be a critical aspect. In telesurgery, where a remote surgeon steers a robot operating on a patient, reconstructing the exact spot of interest is crucial. While existing algorithms aim to mitigate reconstruction issues by predicting required viewpoints, they often overestimate their amount [35]. However, contrary to automatic algorithms, human experts have a better contextual understanding. Another example is archaeological excavation sites, where the high quality of the artifact is more important than random objects in the environment. Moreover, each take can decrease the robots' overall lifetime when robots are unreachable with a limited amount of energy, like the Mars rover's robotic arm [38], or environments cause physical harm, like in nuclear reactors [5]. Humans intuitively understand object appearance, usage, and the required task. In environmental scans, they can assist in identifying and assessing errors, like missing, discolored, or misplaced points [20].

In this work, we assess two human approaches for scene reconstruction in a teleoperation scenario. The first one focuses on specifying sensor poses (direct method), with the second users can indicate areas with missing information (indirect method). For both approaches, humans need to understand the robot's properties and perception. We utilized Virtual Reality (VR) for our approach (outlined in the accompanying video figure), as it offers a natural view and spatial adjustments in a three-dimensional (3D) space [41], while enhancing environment [21] and data understanding [34]. We evaluate the methods in a user study ($N = 16$) by analyzing scan completeness, efficiency, and user experience. It reveals that both methods yield similar results for scan completeness. However, the direct method significantly outperforms the indirect method in processing time and user experience. Thus, specifying new sensor poses directly is more suitable for remote user assistance in scene reconstruction.

> **Contribution Statement.** Our contribution is twofold. First, we create and introduce an approach to enable human assistance for scene reconstruction in VR, providing two interaction methods, the direct method and indirect method, for specification. Second, we present insights into both human-assisted interaction methods by evaluating scan completeness, efficiency, and user experience in a user study ($N = 16$).

2 RELATED WORK

Following, we describe autonomous view planning, as the current standard for scene reconstruction using robots. Since our work follows a human-in-the-loop approach, we outline such works in robotics utilizing Mixed Reality (MR).

2.1 Autonomous Scene Reconstruction Utilizing Robots

In robotics, there are two approaches that are commonly used to autonomously explore environments for scene reconstruction, frontier-based or next-best view (NBV) based methods. Using frontier-based approaches, robots navigate between explored and unexplored places in the environment [43], drawing maps for planning the next movements [2, 3]. Contrasting, NBV approaches aim to find the most efficient sequence of robot sensor viewpoints in the environment before execution. They sample potential viewpoints near the frontier of the explored environment or randomly and evaluating their potential information gain [45]. NBV approaches are applied in various applications to scan objects [11, 16], and entire environments for reconstruction, in indoor [22], and outdoor scenarios, with mobile robots [30, 40], or drones [4]. Newer approaches apply machine learning for planning [26], or filling gap regions of the received reconstructions [10]. However, selecting optimal sensor poses remains challenging. We assume humans to benefit the sampling with their knowledge of object appearance and existence, particularly when familiar with the environment. Thus, we focus on human-assisted sensor placement for scene reconstruction.

2.2 Human Assistance in Robotics Using Mixed Reality

When domain knowledge is required for robot programming, potentially leading to error-prone tasks and faulty robot execution

when missing [44], approaches involving human assistants are increasingly used. Including experts and non-professionals in complex robot tasks can enrich the current processes and reduce the overall complexity and entry barriers associated with robot programming [44]. Conversely, robotic systems can learn from human assistants [29]. To facilitate interaction with automatic systems, MR systems display virtual information in a spatial representation. Augmented Reality (AR) systems enhance the real world with virtual information regarding the robot [37]. Related works displayed a robot's code, its virtual representation, detected objects, and movement goals in the real-world [44], or enabled human assistants to direct a robot by drawing paths or selecting waypoints for path-planning [17]. VR was shown to actively engage users, offering a more immersive experience that facilitates natural viewing and seamless adjustment of the virtual reconstruction [21]. Using VR, human users can remotely explore an environment by teleoperating a mobile robot with RGBD cameras [33]. Its interactive, immersive exploration induced a higher situational awareness and precise navigation in challenging environments [33]. Next to 3D models retrieved from sensor data, MR approaches often include point clouds directly [8, 9, 21, 31]. Utilizing point clouds, users could explore recordings of a terrestrial environment from a remote mobile robot [8], or 3D reconstructions of rooms, including people, furniture, and objects, obtained from depth sensors [31]. They have been utilized for real-time collaboration, such as in virtual multi-user holoconferencing systems to display remote participants [9]. Based on these works, we choose VR as technology to enable remote assistance for scene reconstruction to display real-world scans as point clouds and the robot's operational range, in our application. Similarly to Krings et al. [17], we facilitate robot control, focusing on sensor placement.

3 THE HUMAN GUIDANCE

Following, we present our approach for scan perception and proposing new sensor poses for human assistants. Our video figure displays the application with the interaction methods in detail.

3.1 Approach

A robot placed in a new environment knows nothing about it. Therefore, it needs to capture its environment. However, initial scans can not capture an environment completely, as object occlusion leads to artifacts and missing information. Thus, to complete an entire recording, the robot needs to capture the unknown areas. Given an initial incomplete scan, two elemental strategies exist to complete it: One strategy is to specify optimal sensor poses covering missing areas. The other strategy is to indicate areas with missing information. We focus on these two strategies for interaction. The sensor-focused approach requires users to provide the exact placement location of a sensor. Thereby, users place the robot's sensors directly in the virtual environment and see by the frustum how the missing space is covered (*direct method*). The *indirect method* focuses on the missing area. Here, users provide information about inadequate areas by creating a plane covering them. Subsequently, a sensor pose recording the specified area is calculated.

3.2 Implementation

Following, we describe our implementation of the interaction methods and how we enabled teleoperation with a static robot arm. The interaction methods are integrated into a VR application, visualizing the environment scan.

3.2.1 Data Acquisition Loop. Our application enables teleoperating a static robot arm, a Panda from Franka Emika[1] (see Figure 2), in a geographically remote location. The arm, mounted on a table, was equipped with an Intel RealSense D435 RGBD camera, and was controlled using ROS1 and MoveIT. From the sensor's RGBD images, we generate a TSDF mesh of the scene with Open3D[2] to calculate a single point cloud. Combined with the robot's operational area, the point cloud is stored in a Collada file. Utilizing one input method, users provide new sensor positions uploaded to the server in Collada format. The robot takes new images when a new file with poses is found. The process iterates until the human assistant confirms scene completion.

3.2.2 VR Application. In VR, we visualize the point cloud on a virtual table. Its height is derived from the vertical offset of the robot arm. To represent the robot's sensors, we use virtual cameras. Their field of view (FoV) is indicated with rays. We display the robot's operational area as a blue cuboid if a pose is unreachable. In such cases, the camera's color changes to red to indicate the inaccessibility. However, the robot unit might still not capture all suggested poses. To communicate such errors, we visually represent erroneous poses as static red cameras, requiring deletion before sending new poses. This ensures human assistants recognize which positions are infeasible.

3.2.3 Interaction Methods. The direct method (see Figure 2A) facilitates camera creation through a button press. Users can grab or select a camera by casting a ray, which then allows them to adjust its position and orientation. When positioned outside the robot's operational area, the camera changes its color to red to communicate the pose is not feasible. Users can delete cameras by selecting them and pressing the delete button.

The indirect method (see Figure 2B) enables users to indicate an area for recording. To do so, users define the corners of the recording plane by placing three spheres using a button on the right controller. A fourth sphere appears, completing the rectangular plane. Users can adjust it by interacting with the spheres. An arrow indicates the sensor's side, changeable by dragging it to the other side. When users confirm the plane's placement, a camera covering the plane appears. It is colored red when outside the robot's operational area and can be deleted.

4 EVALUATION

We conducted a user study to investigate "which input method, the direct method or indirect method, is most suitable for specifying new sensor poses regarding the scene completeness, efficiency, and user experience?" (**RQ1**).

[1]Franka Emika. https://www.franka.de/, last accessed: March 8, 2024
[2]Open3D. https://www.open3d.org/docs/release/tutorial/pipelines/rgbd_integration.html, last accessed: March 8, 2024

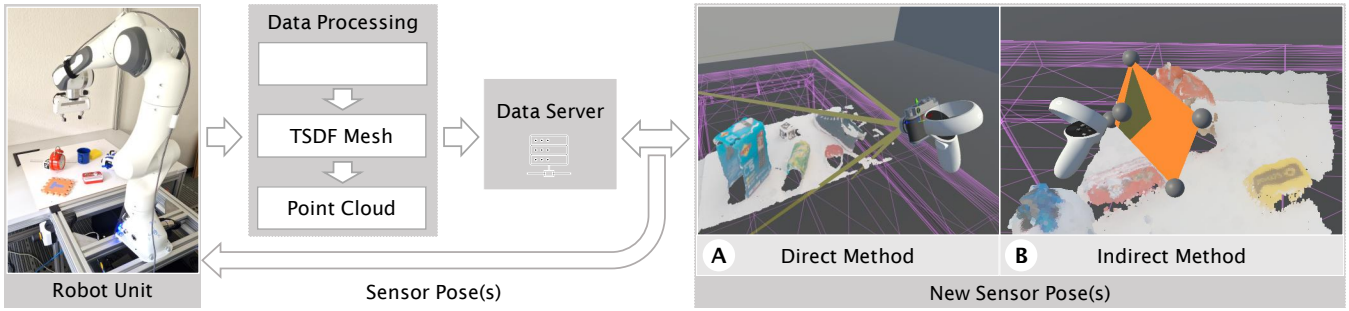

Figure 2: A robot unit first executes an initial scan. The RGBD images are transformed into a mesh of the scene and, finally, a point cloud. The cloud is uploaded to our server as a Collada file. Users either utilize the direct method (A) or the indirect method (B) to suggest sensor poses. Upon receiving the poses, the robot makes new scans. The process iterates until the user confirms completion.

4.1 Study Design

We conducted a controlled laboratory study with a within-subjects design to compare the interaction methods in VR. Our independent variables were the *interaction methods* with two levels: the direct method and the indirect method. We employed a Latin square design, resulting in two configurations, whereas we used two different scene compositions, similar to related studies [11, 16, 26], to prevent recognition. We measured the dependent variables: *task completion time (TCT)*, *camera poses*, *number of recordings*, *spawned cameras*, *deleted cameras*, *user experience*, *workload*, *easiness of placement*, *precision of placement*, *perception of recordings*, *scene control*, *scene completeness*, *process engagement*, *perceived time*, *perceived efficiency*, and *scene understanding*. Table 1 lists details of the variables. We captured the participants' interaction times and resulting point cloud scans for later analysis [1]. Finally, we conducted semi-structured interviews for qualitative feedback.

4.2 Procedure

Before starting the study, we introduced participants to the study procedure and objectives. We addressed all open questions and informed them about recording the TCT and final scan results, as well as their right to withdraw without drawbacks. We conducted the study after obtaining their written consent. We used the *Meta Quest 2* head-mounted display (HMD) as a device for *VR* and preconfigured a virtual safety guard onto a $4.07m \times 4.05m$ open space. The experimenter monitored participants to ensure their safety. First, participants entered a tutorial to familiarize themselves with the *interaction methods* and recording process. Since the robot needs approximately 1.5 minutes to take images, we offered a chair during waiting times to ensure the participants' comfort. Communication between the study and remote robot location was established before participants entered the first study scene. For each *interaction method*, their task was to perform a scene scan as complete as possible. Participants were not subjected to a time limit. They could request the unlock of the *Complete* button to finish. After each scene, they answered the NASA-Task Load Index (TLX), User Experience Questionnaire (UEQ), and additional Likert items. At the end of both scenes, we conducted a semi-structured interview for qualitative feedback. The study took approximately 60 minutes per participant.

4.3 Participants

We recruited 16 volunteers (2 female, 14 male, 0 diverse), between 23 and 35 years ($M = 28.69, SD = 3.36$). Only 1 was left-handed. We inquired participants about their experience with VR on a 6-point Likert scale, corresponding to "I have never used VR before" (1), "I used VR once" (2), "I use VR yearly" (3), "I use VR monthly" (4), "I use VR weekly" (5) and "I use VR every day" (6). They responded with a medium value of 4.68 ($IQR = 2$) and considered their experience with robots with a medium value of 3.19 ($IQR = 2$) on the same scale. To ensure participants' privacy, we only recorded pseudonymized data. Our local ethics committee approved the study.

5 RESULTS

Following, we outline the results of the evaluation. We present the quantitative data before reporting the subjective insights of our participants' feedback. Figure 3 displays our participant's responses regarding the input methods to the Likert items.

5.1 Quantitative Analysis

Our quantitative data includes measurements throughout the study and participants' ratings. We found that most of our data is not normally distributed. Therefore, we applied non-parametric tests and performed a Wilcoxon Signed-rank test directly. We only report the significant results listed as median (interquartile range).

Scan Completeness We compared the participant's results to a ground truth file, our best coverage of the scenes, with the *Chamfer Distance (CD)* and *Earth Mover's Distance (EMD)* as the two broadly utilized metrics to measure point cloud similarity [42]. The CD is 0.0053 (IQR=0.0014) for the direct method and 0.0053 (IQR=0.0018) for the indirect method, respectively. The EMD is 0.0101 (IQR=0.0032) for the direct method and 0.0109 (IQR=0.0109) for the indirect method.

Task Completion Time (TCT) Our participants spent 225.06s (IQR131.24s) in the direct method and 364.19s (IQR= 207.17s) in the indirect method. We found a significant difference ($W = 70.0, Z = -2.19, p < .05, r = .39$) between the methods. Thus, our participants

Table 1: Variables and Their Measurement for Evaluating the *Input Modalities*.

Measurement	Calculation	Measurement	7-Point Likert Items
Task Completion Time (TCT)	Interaction time in seconds to complete the scene	Easiness of Placement	*"It was very easy to place cameras in the scene."*
		Precision of Placement	*"I was able to specify the camera position very" "precisely."*
Camera Poses	Overall number of cameras placed		
Number of Recordings	Recording cycles in one session	Perception of Recordings	*"The robot unit recordings matched my vision of the" "scene very well."*
Spawned Cameras	Number of cameras spawned in one session	Scene Control	*"I had the impression of having complete control over" "the recordings of the scene."*
Deleted Cameras	Number of cameras deleted in a session		
User Experience	User Experience Questionnaire (UEQ) Short [18, 36]	Scene Completeness	*"I was able to achieve a complete portrayal of the scene."*
		Process Engagement	*"I felt very engaged in the process of capturing the" "scenes."*
Task Load Index	NASA-Task Load Index (TLX) [13]		
Chamfer Distance (CD)	Sum of the square distance of the scan results to the ground truth	Perceived Time	*"It took me not much time to complete the whole scene."*
		Perceived Efficiency	*"I find this method of interaction to be very efficient."*
Earth Mover's Distance (EMD)	Average distance between point pairs of the scan results to the ground truth	Scene Understanding	*"I had a very good understanding of the scene compo-" "sition."*

needed more time in the indirect method to finalize a scene.

Spawned Cameras The number of spawned cameras were 16.0 (IQR=10.0) in the direct method and 25.5 (IQR= 19.5) in the indirect method. We found a significant difference ($W = 67.5, Z = -2.29, p =< .05, r = .40$) between the methods and can conclude that more cameras were spawned in the indirect method.

Deleted Cameras The number of deleted cameras were 8.0 (IQR=5.25) in the direct method and 20.5 (IQR=20.75) in the indirect method. We found a significant difference between the methods ($W = 42.5, Z = -3.23, p =< .001, r = .58$). Thus, more cameras were deleted in the indirect method.

User Experience We received a total score of 1.75 (IQR=1.06) for the direct method and 0.63 (IQR=1.13) for the indirect method. A comparison revealed a significant difference of the input methods ($W = 187.0, Z = 2.23, p < .05, r = .39$). For the pragmatic scores, the user experience was rated as 1.5 (IQR=1.25) for the direct method and -0.75 (IQR=2.06) for the indirect method, revealing a significant difference between the methods ($W = 202.0, Z = 2.8, p =< .01, r = .49$). Thus, our participants perceived a higher user experience using the direct method overall and in the pragmatic scores.

Task Load Index We received an overall score of 53.5 (IQR=19.5) for the direct method and 69.0 (IQR=20.0) for the indirect method and found a significant difference ($W = 64.5, Z = -2.39, p < .05, r = .39$). Further, the direct method was rated as 7.0 (IQR=8.0), the indirect method as 15.0 (IQR=4.0) for the mental demand, revealing a significant difference ($W = 55.0, Z = -2.76, p < .01, r = .49$). For the effort, we received 7.5 (IQR=5.5) for the direct method and 14.0 (IQR=6.75) for the indirect method and found significant difference between both methods ($W = 56.5, Z = -2.7, p < .01, r = .48$). Using the direct method, our participants perceived a lesser overall workload, mental demand, and effort.

Easiness of Placement Our participants rated the direct method with 6.5 (IQR=1.0) and the indirect method with 3.0 (IQR=1.25). We found a significant difference between the methods ($W = 251.5, Z = 4.75, p < .001, r = .84$). Therefore, the direct method was perceived as easier for placing cameras.

Precision of Placement Our participants rated the direct method with 6.5 (IQR=1.0) and the indirect method with 2.0 (IQR=1.25). We found a significant difference between the methods ($W = 229.0, Z = 3.89, p < .001, r = .69$). Our participants perceived the camera placement as more precise using the direct method.

Scene Control Our participants rated the direct method with 5.0 (IQR=1.25) and the indirect method with 3.0 (IQR=2.0). We found a significant difference between the methods ($W = 210.0, Z = 3.15, p < .01, r = .56$). Thus, we can conclude that the participants perceived a higher level of scene control when using the direct method.

Perceived Time Our participants rated the direct method with 3.5 (IQR=3.0) and the indirect method with 2.0 (IQR=2.0). We found a significant difference between the methods ($W = 181.0, Z = 2.04, p < .05, r = .36$). We can conclude that participants perceived less time for the direct method.

Perceived Efficiency Our participants rated the direct method with 5.0 (IQR=2.0) and the indirect method with 2.0 (IQR=1.25). We found a significant difference between the methods ($W = 228.0, Z = 3.82, p < .001, r = .68$). We can conclude that participants perceived the direct method as more efficient.

5.2 Qualitative Analysis

To analyze the interview data, we combined all answers and conducted a thematic analysis after Clarke and Braun [7] by employing open coding involving three researchers, all authors. First, we identified 509 atomic statements and coded them, resulting in 59 codes before identifying 12 categories and 4 themes, using the Miro Whiteboard Tool.

Efficiency Most participants (11) found the indirect method less efficient for larger areas, requiring the creation of multiple small planes, leading to an increased scan duration. They reported frequent adjustments when cameras appeared outside the robot's range (P7). However, the indirect method was favored for its precision for small areas (6), benefiting understanding of the recording's coverage (P3). For the direct method, participants emphasized the ease of camera pose adjustments (P17) and visually perceiving if the camera is valid (P13).

Intuitivity and Usability The direct method was well-received for its expectation-compliant behavior (14), often referred to as similar to photography in the real world (P12) The participants

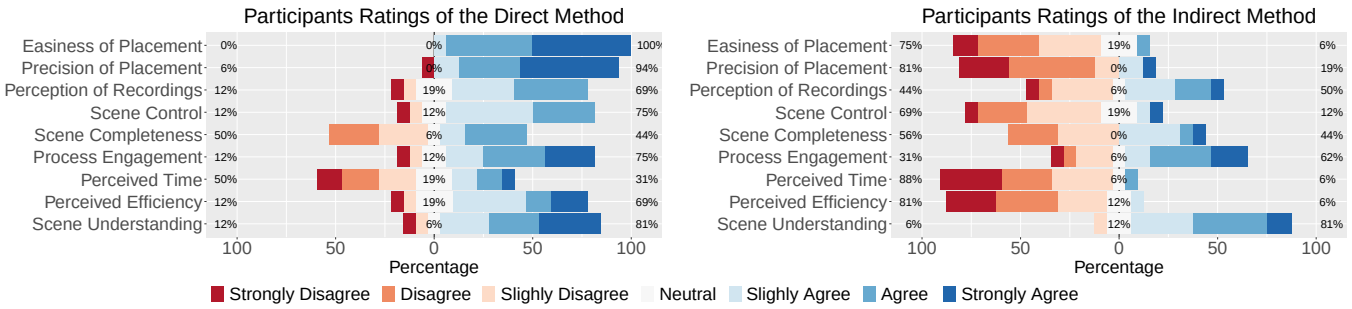

Figure 3: Participant's responses: Comparison between the direct method (right) and the indirect method (left).

encountered difficulties adjusting the camera's distance and angle in the indirect method, leading to frustration, when further limited by the robot's operating area (P11). These challenges required a higher learning effort, reducing the created plane size throughout the study.

Cognitive Effort Most participants (14) found the point cloud representation effective. However, a few (3) observed a discrepancy between the expected and actual scan results. Our participants reported being unaware of the robot during execution. While P8 stated that not seeing the robot increased uncertainty, P14 found the lack of information beneficial, allowing him to focus on the input. Some (5) wished for a visual progress representation during the robot execution.

Experiences and Improvement Suggestions Overall, our participants described the application as engaging and enjoyable (13). However, they sometimes (8) faced challenges covering all desired points due to the operation area's design (P2). Some suggested continuously visualizing (5) or providing a more precise operational area (2) to ease the estimation of poses lying within the robot's reach. Moreover, some suggested live-steer the robot unit (5) to enable direct control of its movements, avoiding invalid placements and unnecessary recordings. To improve the indirect method, a "best-fit" feature was suggested, recommending alternate camera positions covering the specified area. Participants further suggested a camera preview, facilitating immediate adjustments to align the camera placement with their vision.

6 DISCUSSION

The direct method outperformed the indirect method Our participants preferred the direct method of placing sensor poses directly regarding workload (15.5% better as indirect method), perceived efficiency (42.9%), easiness of placement (50%), precision of placement (64.3%), and scene control (28.6%). Based on subjective feedback, we infer the task load index disparity stems from estimating the final camera position in the indirect method, where users specified areas to record, redundant in the direct method. While the indirect method required participants to provide the coverage areas, they were more focused on the final camera placement. Moreover, the direct method was significantly faster and outperformed the indirect method in efficiency. It showed higher user experience in overall (18.7%) and pragmatic scores (37.5%) than the indirect method. Potentially, this was influenced by a preference for familiar experiences [21, 28], as real-life photography was named a familiar analogy. We believe user experience is the key to the discrepancy of the input methods in our participant's ratings, as it encompasses all interaction aspects [32].

The interaction method did not affect scan completeness Our study revealed both input methods to receive similar scan coverage. They received the same median values for the CD, only differing in the IQR with a difference of 0.0004. The EMD differed by 0.0008 in the median and 0.0077 in the IQR. All favorable for the direct method. Thus, both methods are applicable for scene reconstruction, resulting in similar scan coverage. We acknowledge that the participants' unfamiliarity with the study scenes may serve as confounding factors, potentially negatively impacting their results. However, since the direct specification of sensor poses was found more suitable regarding efficiency and user experience, we recommend it for facilitating user assistance in scene reconstruction.

How to implement the direct method In our study, we used a digital representation of a camera rather than a precise visual of the robot's sensor, assuming that their visuals aren't essential for accurate specification. We highlighted key features like camera orientation and frustum, which participants found useful in determining coverage. However, some participants noticed minor discrepancies potentially arising from positional inaccuracies [15], which might be communicated using uncertainty visualization. We used VR controllers in our application to facilitate interaction. When utilizing other input devices, we assume an influence on the provided pose precision. For instance, hand interaction in VR is less precise than controller [21, 25]. Furthermore, in our study scenes incorporated everyday objects, aligning the participants' reach almost identically with the robot's operational area. In larger-scale applications, such as drone surveillance over extensive areas like mountain ranges or plantations, this methodology would need adaptation to ensure the mapping of potential poses (direct method) or the scans (indirect method) aligns with the participants' reach.

7 OUTLOOK

As the indirect method performed worse than the direct method, it could be enhanced to improve user assistance. Following feedback from our study participants, these improvements could include adding a preview feature and recommending multiple camera angles or suitable poses close to the current estimated location.

By introducing these methods, we aim to enhance scene reconstruction through human assistance. As our setup differs, we can not compare our approaches directly to existing view-planning

approaches. However, our results indicate a similar or increased scan completeness [14, 23, 27]. Thus, we aim to extend our comparison to automatic approaches, like frontier-based [12, 24] or NBV-based [45] approaches, in future work. We believe that involving humans can improve the quality of scans by leveraging their expert knowledge [20]. However, there might also be situations where automated methods perform better than human-assisted ones or where the difference in performance does not justify the extra effort required. Therefore, we see a potential for future research to identify for which use case each method performs best.

8 CONCLUSION

In this work, we compared two interaction methods in VR facilitating human assistance in the remote scene reconstruction via teleoperating a robotic arm. Users either steered the robotic arm by proposing new sensor poses directly or indicating areas with missing information. Our user study ($N = 16$) revealed that directly inserting sensor poses is more efficient and received higher user experience. Both methods received similar values in scan completeness, only yielding differences in the fourth decimal place measured with the EMD. Thus, providing sensor poses directly is more suitable for human assistance in scene reconstruction. In future work, we aim to compare this approach to the current automatic solutions to provide insights into the quality of human assistance.

ACKNOWLEDGMENTS

This work is funded by the German Federal Ministry of Education and Research (01IS21068B). We used ChatGPT for text editing tasks.

REFERENCES

[1] Shivam Agarwal, Jonas Auda, Stefan Schneegaß, and Fabian Beck. 2020. A Design and Application Space for Visualizing User Sessions of Virtual and Mixed Reality Environments. In *Vision, Modeling, and Visualization*, Jens Krüger, Matthias Niessner, and Jörg Stückler (Eds.). The Eurographics Association, Norrköping, Sweden, 117–126. https://doi.org/10.2312/vmv.20201194

[2] Mohammad Al-khawaldah and Andreas Nüchter. 2012. Multi-Robot Exploration and Mapping with a rotating 3D Scanner. *IFAC Proceedings Volumes* 45, 22 (2012), 313–318. https://doi.org/10.3182/20120905-3-HR-2030.00025 10th IFAC Symposium on Robot Control.

[3] Abraham Bachrach, Ruijie He, and Nicholas Roy. 2009. Autonomous Flight in Unknown Indoor Environments. *International Journal of Micro Air Vehicles* 1, 4 (2009), 217–228. https://doi.org/10.1260/175682909790291492 arXiv:https://doi.org/10.1260/175682909790291492

[4] Andreas Bircher, Mina Kamel, Kostas Alexis, Helen Oleynikova, and Roland Siegwart. 2016. Receding Horizon "Next-Best-View" Planner for 3D Exploration. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, Stockholm, Sweden, 1462–1468. https://doi.org/10.1109/ICRA.2016.7487281

[5] Robert Bogue. 2011. Robots in the nuclear industry: a review of technologies and applications. *Industrial Robot: An International Journal* 38, 2 (2011), 113–118.

[6] D. Borrmann, R. Heß, H. R. Houshiar, D. Eck, K. Schilling, and A. Nüchter. 2015. ROBOTIC MAPPING OF CULTURAL HERITAGE SITES. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XL-5/W4 (2015), 9–16. https://doi.org/10.5194/isprsarchives-XL-5-W4-9-2015

[7] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[8] Gerd Bruder, Frank Steinicke, and Andreas Nuchter. 2014. Poster: Immersive point cloud virtual environments. In *2014 IEEE Symposium on 3D User Interfaces (3DUI)*. IEEE, Minneapolis, MN, USA, 161–162. https://doi.org/10.1109/3DUI.2014.6798870

[9] Gianluca Cernigliaro, Marc Martos, Mario Montagud, Amir Ansari, and Sergi Fernandez. 2020. PC-MCU: Point Cloud Multipoint Control Unit for Multi-User Holoconferencing Systems, In Proceedings of the 30th ACM Workshop on Network and Operating Systems Support for Digital Audio and Video. *Proceedings of the 30th ACM Workshop on Network and Operating Systems Support for Digital Audio and Video*, 47–53. https://doi.org/10.1145/3386290.3396936

[10] Kunyao Chen, Fei Yin, Baichuan Wu, Bang Du, and Truong Nguyen. 2021. Mesh Completion with Virtual Scans. In *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, Anchorage, AK, USA, 3303–3307. https://doi.org/10.1109/ICIP42928.2021.9506612

[11] Chenming Wu, Rui Zeng, Jia Pan, Charlie C. L. Wang, and Yong-Jin Liu. 2019. Plant Phenotyping by Deep-Learning-Based Planner for Multi-Robots. *IEEE Robotics and Automation Letters* 4, 4 (2019), 3113–3120. https://doi.org/10.1109/LRA.2019.2924125

[12] Titus Cieslewski, Elia Kaufmann, and Davide Scaramuzza. 2017. Rapid exploration with multi-rotors: A frontier selection method for high speed flight. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, Vancouver, BC, Canada, 2135–2142. https://doi.org/10.1109/IROS.2017.8206030

[13] Sandra G. Hart. 1986. TASK LOAD INDEX (NASA-TLX). *NASA Technical Report* (1986), 1–26.

[14] Hao Hu, Sicong Pan, Liren Jin, Marija Popović, and Maren Bennewitz. 2023. Active Implicit Reconstruction Using One-Shot View Planning. arXiv:2310.00685 [cs.RO]

[15] Yizhou Jiang, Liandong Yu, Huakun Jia, Huining Zhao, and Haojie Xia. 2020. Absolute Positioning Accuracy Improvement in an Industrial Robot. *Sensors* 20, 16 (2020), 14 pages. https://doi.org/10.3390/s20164354

[16] Michael Krainin, Brian Curless, and Dieter Fox. 2011. Autonomous generation of complete 3D object models using next best view manipulation planning. In *2011 IEEE International Conference on Robotics and Automation*. IEEE, Shanghai, China, 5031–5037. https://doi.org/10.1109/ICRA.2011.5980429

[17] Sarah Claudia Krings, Enes Yigitbas, Kai Biermeier, and Gregor Engels. 2022. Design and Evaluation of AR-Assisted End-User Robot Path Planning Strategies. In *Companion of the 2022 ACM SIGCHI Symposium on Engineering Interactive Computing Systems*, Marco Winckler and Aaron Quigley (Eds.). ACM, New York, NY, USA, 14–18. https://doi.org/10.1145/3531706.3536452

[18] Bettina Laugwitz, Theo Held, and Martin Schrepp. 2008. Construction and evaluation of a user experience questionnaire. *HCI and Usability for Education and Work: 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society* 5298, 4 (2008), 63–76.

[19] Gal Leibovich, Guy Jacob, Shadi Endrawis, Gal Novik, and Aviv Tamar. 2022. Validate on Sim, Detect on Real - Model Selection for Domain Randomization, In 2022 International Conference on Robotics and Automation (ICRA). *2022 International Conference on Robotics and Automation (ICRA)*, 7528–7535. https://doi.org/10.1109/ICRA46639.2022.9811621

[20] Carina Liebers, Pranav Megarajan, Jonas Auda, Tim C. Stratmann, Max Pfingsthorn, Uwe Gruenefeld, and Stefan Schneegass. 2024. Keep the Human in the Loop: Arguments for Human Assistance in the Synthesis of Simulation Data for Robot Training. *Multimodal Technologies and Interaction* 8, 3 (2024), 18. https://doi.org/10.3390/mti8030018

[21] Carina Liebers, Marvin Prochazka, Niklas Pfützenreuter, Jonathan Liebers, Jonas Auda, Uwe Gruenefeld, and Stefan Schneegass. 2023. Pointing It out! Comparing Manual Segmentation of 3D Point Clouds between Desktop, Tablet, and Virtual Reality. *International Journal of Human–Computer Interaction* 0, 0 (2023), 1–15. https://doi.org/10.1080/10447318.2023.2238945

[22] Ligang Liu, Xi Xia, Han Sun, Qi Shen, Juzhan Xu, Bin Chen, Hui Huang, and Kai Xu. 2018. Object-Aware Guidance for Autonomous Scene Reconstruction. *ACM Trans. Graph.* 37, 4, Article 104 (jul 2018), 12 pages. https://doi.org/10.1145/3197517.3201295

[23] Federico Magistri, Elias Marks, Sumanth Nagulavancha, Ignacio Vizzo, Thomas Läebe, Jens Behley, Michael Halstead, Chris McCool, and Cyrill Stachniss. 2022. Contrastive 3D shape completion and reconstruction for agricultural robots using RGB-D frames. *IEEE Robotics and Automation Letters* 7, 4 (2022), 10120–10127.

[24] A. Mannucci, S. Nardi, and L. Pallottino. 2018. Autonomous 3D exploration oflarge areas: A. cooperative frontier-based approach. *in Modelling Simul. Auton. Syst.* 1075 (2018), 18–39.

[25] Alexander Masurovsky, Paul Chojecki, Detlef Runde, Mustafa Lafci, David Przewozny, and Michael Gaebler. 2020. Controller-Free Hand Tracking for Grab-and-Place Tasks in Immersive Virtual Reality: Design Elements and Their Empirical Study. *Multimodal Technologies and Interaction* 4, 4 (Dec. 2020), 91. https://doi.org/10.3390/mti4040091 Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

[26] Miguel Mendoza, J. Irving Vasquez-Gomez, Hind Taud, L. Enrique Sucar, and Carolina Reta. 2020. Supervised learning of the next-best-view for 3d object reconstruction. *Pattern Recognition Letters* 133 (2020), 224–231. https://doi.org/10.1016/j.patrec.2020.02.024

[27] Rohit Menon, Tobias Zaenker, Nils Dengler, and Maren Bennewitz. 2023. NBV-SC: Next best view planning based on shape completion for fruit mapping and reconstruction. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, IEEE, Detroit, MI, USA, 4197–4203.

[28] Michael Minge. 2008. Dynamics of User Experience. *Minge, Michael. "Dynamics of user experience." Proceedings of the Workshop on Research Goals and Strategies for Studying User Experience and Emotion, NordiCHI* 8 (2008).

[29] Reuth Mirsky, Kim Baraka, Taylor Kessler Faulkner, Justin Hart, Harel Yedidsion, and Xuesu Xiao. 2022. Human-Interactive Robot Learning (HIRL). *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (2022),

1278–1280. https://doi.org/10.1109/HRI53351.2022.9889551

[30] Menaka Naazare, Francisco Garcia Rosas, and Dirk Schulz. 2022. Online Next-Best-View Planner for 3D-Exploration and Inspection With a Mobile Manipulator Robot. *IEEE Robotics and Automation Letters* 7, 2 (2022), 3779–3786. https://doi.org/10.1109/LRA.2022.3146558

[31] Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L. Davidson, Sameh Khamis, Mingsong Dou, Vladimir Tankovich, Charles Loop, Qin Cai, Philip A. Chou, Sarah Mennicken, Julien Valentin, Vivek Pradeep, Shenlong Wang, Sing Bing Kang, Pushmeet Kohli, Yuliya Lutchyn, Cem Keskin, and Shahram Izadi. 2016. Holoportation: Virtual 3D Teleportation in Real-Time. *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (2016), 741–754. https://doi.org/10.1145/2984511.2984517

[32] Jaehyun Park, Sung H. Han, Hyun K. Kim, Youngseok Cho, and Wonkyu Park. 2013. Developing Elements of User Experience for Mobile Phones and Services: Survey, Interview, and Observation Approaches. *Human Factors and Ergonomics in Manufacturing & Service Industries* 23, 4 (2013), 279–293. https://doi.org/10.1002/hfm.20316

[33] Patrick Stotko, Stefan Krumpen, Max Schwarz, Christian Lenz, Sven Behnke, Reinhard Klein, and Michael Weinmann. 2019. A VR System for Immersive Teleoperation and Live Exploration with a Mobile Robot. *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2019), 3630–3637. https://doi.org/10.1109/IROS40897.2019.8968598

[34] Harrison Pearl, Hillary Swanson, and Michael Horn. 2019. Coordi: A Virtual Reality Application for Reasoning about Mathematics in Three Dimensions. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, Stephen Brewster, Geraldine Fitzpatrick, Anna Cox, and Vassilis Kostakos (Eds.). ACM, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3312931

[35] Christian Potthast and Gaurav S Sukhatme. 2014. A probabilistic framework for next best view estimation in a cluttered environment. *Journal of Visual Communication and Image Representation* 25, 1 (2014), 148–164.

[36] Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. 2017. Design and evaluation of a short version of the user experience questionnaire (UEQ-S). *International Journal of Interactive Multimedia and Artificial Intelligence, 4 (6), 103-108.* 4, 6 (2017), 103–108.

[37] Ryo Suzuki, Adnan Karim, Tian Xia, Hooman Hedayati, and Nicolai Marquardt. 2022. Augmented Reality and Robotics: A Survey and Taxonomy for AR-enhanced Human-Robot Interaction and Robotic Interfaces. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022), 1–33. https://doi.org/10.1145/3491102.3517719

[38] Edward Tunstel, Mark Maimone, Ashitey Trebi-Ollennu, Jeng Yen, Rich Petras, and Reg Willson. 2005. Mars exploration rover mobility and robotic arm operational performance. In *2005 IEEE International Conference on Systems, Man and Cybernetics*, Vol. 2. IEEE, IEEE, Waikoloa, HI, USA, 1807–1814.

[39] J. Irving Vasquez-Gomez, L. Enrique Sucar, Rafael Murrieta-Cid, and Efrain Lopez-Damian. 2014. Volumetric Next-best-view Planning for 3D Object Reconstruction with Positioning Error. *International Journal of Advanced Robotic Systems* 11, 10 (2014), 159. https://doi.org/10.5772/58759

[40] Yiduo Wang, Miland Ramezani, and Maurice Fallon. 2020. Actively Mapping Industrial Structures with Information Gain-Based Planning on a Quadruped Robot. *2020 IEEE International Conference on Robotics and Automation (ICRA)* (2020), 8609–8615. https://doi.org/10.1109/ICRA40945.2020.9197153

[41] Matt Whitlock, Stephen Smart, and Danielle Albers Szafir. 2020. Graphical Perception for Immersive Analytics. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, Atlanta, GA, USA, 616–625. https://doi.org/10.1109/VR46266.2020.00084

[42] Tong Wu, Liang Pan, Junzhe Zhang, Tai WANG, Ziwei Liu, and Dahua Lin. 2021. Balanced Chamfer Distance as a Comprehensive Metric for Point Cloud Completion. *Advances in Neural Information Processing Systems* 34 (2021), 29088–29100. https://proceedings.neurips.cc/paper_files/paper/2021/file/f3bd5ad57c8389a8a1a541a76be463bf-Paper.pdf

[43] Brian Yamauchi. 1997. A frontier-based approach for autonomous exploration. In *Proceedings 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation CIRA'97.'Towards New Computational Principles for Robotics and Automation'*. IEEE, IEEE, Monterey, CA, USA, 146–151.

[44] Enes Yigitbas, Ivan Jovanovikj, and Gregor Engels. 2021. Simplifying Robot Programming Using Augmented Reality and End-User Development. In *Human-Computer Interaction – INTERACT 2021*, Carmelo Ardito, Rosa Lanzilotti, Alessio Malizia, Helen Petrie, Antonio Piccinno, Giuseppe Desolda, and Kori Inkpen (Eds.). Lecture Notes in Computer Science, Vol. 12932. Springer International Publishing, Cham, 631–651. https://doi.org/10.1007/978-3-030-85623-6_36

[45] Rui Zeng, Yuhui Wen, Wang Zhao, and Yong-Jin Liu. 2020. View planning in robot active vision: A survey of systems, algorithms, and applications. *Computational Visual Media* 6, 3 (2020), 225–245. https://doi.org/10.1007/s41095-020-0179-3