Received: 10 March 2023 | Revised: 23 June 2023 | Accepted: 21 July 2023

DOI: 10.1002/jsid.1262

RESEARCH ARTICLE

WILEY

Toward physically realistic vision in teleoperation: A user study with light-field head mounted display and 6-DoF head motion

Nicolai Bechtel[1] | Bernhard Weber[1] | Pascal Severin[1] | Jaime Sancho Aragon[2] | Laurie Van Bogaert[3] | Michael Panzirsch[1]

[1]Institute of Robotics and Mechatronics, German Aerospace Center (DLR), Wessling, Germany

[2]Research Center on Software Technologies and Multimedia Systems, Universidad Politécnica de Madrid, Madrid, Spain

[3]Laboratory of Image Synthesis and Analysis (LISA), Université Libre de Bruxelles (ULB), Brussels, Belgium

Correspondence
Michael Panzirsch, Institute of Robotics and Mechatronics, German Aerospace Center (DLR), Wessling, Germany.
Email: michael.panzirsch@dlr.de

Funding information
Horizon 2020 Framework Programme, Grant/Award Number: 951989

Abstract

Besides haptics, the visual channel provides the most essential feedback to the operator in teleoperation setups. For optimal performance, the view on the remote scene must provide 3D information, be sharp, and of high resolution. Head-mounted displays (HMD) are applied to improve the immersion of the operator into the remote environment. Still, so far, no near-eye display technology was available that provides a natural view on objects within the typical manipulation distance (up to 1.2 m). The main limitation is a mismatch of the 3D distance and the focal distance of the visualized objects (vergence-accommodation conflict) in displays with fixed focal distance. This conflict potentially leads to eye strain after extended use. Here, we apply a light-field HMD providing close-to-continuous depth information to the user, thus avoiding the vergence-accommodation conflict. Furthermore, we apply a time-of-flight sensor to generate a 2.5D environment model. The displayed content is processed with image-based rendering allowing a 6 degree-of-freedom head motion in the visualized scene. The main objective of the presented study is evaluating the effects of view perspective and light-field on performance and workload in a teleoperation setup. The reduction of visual effort for the user is confirmed in an abstract depth-matching task.

KEYWORDS

light-field displays, vergence-accommodation conflict

1 | INTRODUCTION

High-performance teleoperation of robots requires a transparent exchange of kinesthetic, tactile, visual, and auditory information between the human operator and the robotic system since the capabilities depend largely on the quality of immersion into the robot's environment. These environments can reach from industrial[1] to healthcare[2] or space exploration scenarios[3]. Recently, the worldwide ANA Avatar XPRIZE competition (e.g., Schwarz et al.[4] and Vaz et al.[5]) for intuitive and immersive teleoperation was arranged focusing on manipulation tasks involving a variety of objects. One of the most advanced setups for visual feedback based on spherical rendering was presented in Schwarz and Behnke.[6] A flexible 6 degree-of-freedom (DoF) adaptation of the view perspective was enabled for the operator, however at the cost of an additional robotic arm allowing for a moving camera. Another promising concept is the strategy of Aykut et al.[7] based on stereo fish-eye cameras and field-of-view (FoV) adaptation for the sake of delay compensation in the head motion.

In most telerobotic setups, the remote environment is visualized to the human operator via head-mounted displays since they enable intuitive stereo vision at large head motions. Still, recent studies in VR environments[8,9] indicate that operating with conventional head-mounted displays (HMD, with fixed focus) cause high visual effort or even eye strain for the operator, thus heavily limiting the application period in VR as well as telerobotics due to reduced usability. One of the main reasons for increased visual effort is the so-called vergence-accommodation conflict (VAC) resulting from non-matching 3D and focusing distance. In conventional HMDs, the focus plane is fixed at around 1.5- to 2-m distance. That means that eyes cannot focus an object outside this depth area. Regarding the analysis of Banks et al.,[10] the VAC is very pronounced for objects in distances below approximately 1 m for conventional HMDs. These distances are typical human arm reaching distances in VR as well as telerobotic applications.

In literature, for instance, holographic displays[11] and such based on microlens arrays, multiple depth planes,[12,13] or varifocal elements (e.g., focus-tunable lenses[14,15] or deformable membrane mirrors[16]) were proposed to display different depth areas. The light-field head-mounted display of this work (Creal Zorya[17]) was already applied in Panzirsch et al.[9] This work confirmed for the first time that light-field HMDs can reduce the visual effort for the user in VR applications with perfect visual information.

Here, we present the results of a teleoperation user-study using the same light-field HMD with a real-time (10–15 fps, 50- to 150-ms delay) video pipeline based on real-time view synthesis. The video pipeline is inspired by the Reference View Synthesis (RVS),[18-21] Creal's Spatio-Temporally Amortized Light-Field (STALF) renderer, which is able to generate the light-field needed by Creal's Zorya,[20] and using a steady Microsoft Kinect Azure sensor (without pan-tilt unit) plus a depth refinement software: Kinect Refinement Tool (KiRT).[22] Such sensors can be positioned in an array at certain locations of the robotic environments requiring precise visual information from different perspectives. Relevant applications can, for instance, be found at CERN[23] or the International Space Station with the exocentric cameras[24] mounted outside of it. The sensor and software setup captures 2.5D information allowing for 6-DoF adaptation of view perspective with light-field depth information to the operator. In contrast to this, a camera on a pan-tilt unit provides a good overview on the environment around the robot while the perspective change resulting purely from the rotations of the pan-tilt unit (without translations) provides only little additional 3D information on a specific object. From one incremental area of the scene, almost exactly the same light rays enter the sensor lens irrespective of the pan-tilt motion.

Translational perspective changes are usually not available in standard teleoperation systems. Still, views from the side onto an object in front of the robot supports the user in analyzing the inclination of an object and gives further depth information. For instance, in case of the humanoid robot DLR Justin, the stereo camera setup is integrated into the head such that views on a grasped object are often very steep such that the orientation of the object is not obvious to the user. This is one of the major teleoperation challenges in the DLR space project Surface Avatar.[25] Similarly, in the DLR project AI-In-Orbit-Factory,[26] the image sensor is used for artificial intelligence (AI) as well as for teleoperation. For AI, this sensor has to be positioned steady at a certain distance from the scene. The distance of the sensor renders the depth interpretation of distant objects very difficult for teleoperation since the human eye requires more information than AI. Real-time view synthesis promises large benefits in such applications.

The reference view synthesis software is a depth image-based rendering (DIBR) method able to synthesize virtual views. Such DIBR is able to produce realistic results by using color images and their depth without calculating the 3D mesh of the scene. It is one of the reference tools of the MIV standard.[27] It has already been used with stereoscopic headsets in Bonatto et al.[18] achieving 2×90 Hz. The version of RVS (RVSVulkan) used in this study differs from the publicly available code[1] in the Graphics API used (Vulkan in place of OpenGL) and presents changes related to the telerobotic setup that is described in Lafruit et al.[22]

Obtaining high-quality depth maps of a scene for DIBR is a challenging process. Depth estimation methods like MPEG Depth Estimation Reference Software (DERS)[28] or Immersive Video Depth Estimation (IVDE)[29] take in the order of tens or hundreds of seconds to generate a depth frame and hence are not fast enough for teleoperation applications. Other works have explored the acceleration of such methods employing graphic processing units (GPUs).[30] Nonetheless, the acceleration is not enough for real-time applications, leading researchers to investigate solutions where depth information is generated by employing time-of-flight (ToF) sensors. These sensors are able to produce real-time depth information

[1]https://gitlab.com/mpeg-i-visual/rvs.

at the expense of lower-quality depth maps[31] that can be refined after capture.[32,33] KiRT, the software employed to refine Microsoft Azure Kinect depth maps in this work, follows that trail to improve the quality of the real-time depth information captured.

The main objective of this study is the evaluation of the visual effort reduction through 6-DoF view synthesis and light-field visualization of nonperfect visual information captured in real-time. For this sake, an experimental telerobotic setup was built, consisting of a DLR lightweight robot equipped with force-torque sensor and a haptic interface providing force feedback to the user. The task focuses on 3D perception and depth matching of abstract objects. Two main comparisons are investigated: comparison *C-I* of light-field (*LF on*, *C-I-1*) and conventional stereo (*LF off*, *C-I-2*), and comparison *C-II* of 6-DoF head motion from different initial poses *C-II-B* and *C-II-C* and a reference condition with steady view *C-II-A*.

The following hypotheses were formulated:

H1 The light-field technology eliminates the vergence-accommodation conflict and thus visual effort (the load on the human vision system) and workload are reduced during teleoperation.

H2 Real-time view synthesis improves depth perception such that accuracy in depth positioning is improved.

H3 Large rotational and translational perspective changes improve depth perception for more distant objects such that depth positioning is improved.

The paper is structured as follows: Section 2 introduces the materials and methods including the description of the experimental setup, study procedure, and metrics. The results are presented in Section 3 and discussed together with the limitations of the study in Section 4. Finally, Section 5 concludes the work.

2 | MATERIALS AND METHODS

2.1 | Sample

$N = 28$ (14 females, 14 males) subjects with an average age of $M = 27.1$ years ($SD = 4.2$ years; range: 22–42 years) participated in the study. Fourteen participants required visual acuity correction and thus wore their glasses or contact lenses during the experiment. None of the subjects indicated to have problems with depth perception.

2.2 | Apparatus

The light-field head-mounted display (HMD) Zorya of the company Creal is equipped with a background display (standard LCD screen, resolution 1600 × 1440 px, 100° FoV) and two light-field displays in the center covering a FoV of 30° with approximately 40 px/° angular resolution. The HMD allows the adjustment of the interpupillary distance (IPD), that is, the individual distance between the two eyes, between 58 and 72 mm. The reader is referred to Panzirsch et al.[9] for detailed information on the display technology.

In the *LF off* conditions *C-I-2*, the inner light-field displays were flattened to a fixed focal distance via software, corresponding to a conventional high-resolution stereoscopic display. The peripheral background display alone (without deactivated light-field display) could not be used for this condition due to its clearly lower resolution.

The telerobotic setup of this study consisted of a DLR lightweight robot (depicted in Figure 1) and the haptic interface lambda.7 by Force Dimension (see Figure 2). The wrench measured by the force-torque sensor installed at the wrist of the robot could be displayed to the operator on the lambda.7—though not needed for the

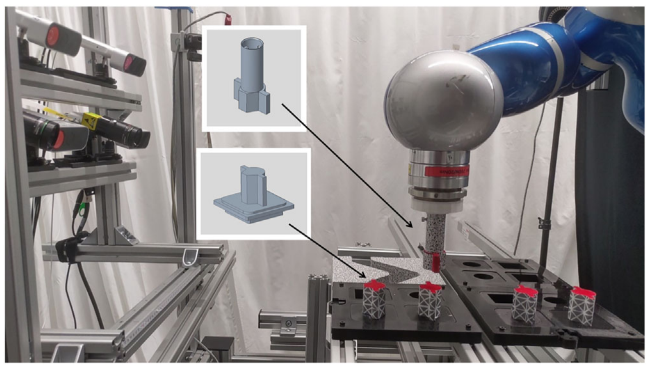


FIGURE 1 Light-weight robot with camera array and experimental scene.

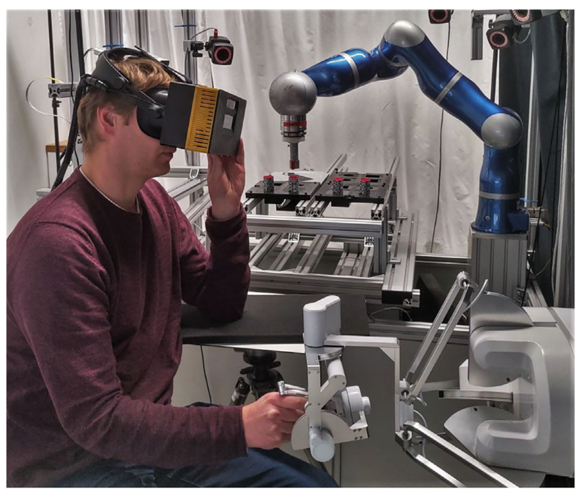


FIGURE 2 User interface: HMD, haptic interface and arm rest.

task of this study. The gripper DoF of the haptic interface had to be closed to activate the motion of the device. The lambda.7 provides a large translational workspace (∅240 mm × 170 mm), which was further extended through a scaling of 2:1 in order to allow accurate positioning of the robot. The device haptically guided the user to an initial position after having activated the gripper.

The pipeline software can be subdivided into three submodules. The first step is to acquire RGBD (RGB + depth) images of the scene using Kinect cameras and then refine them using Kinect Refinement Tool (KiRT). The second step is to synthesize virtual views and their associated depth map at the position of the eyes of the user. The last part is the Creal software, which is responsible to manage the headset and create the light field using STALF. The three submodules of the video pipeline ran asynchronously, working as independent actors that produce and/or consume video streaming of other submodules. They are implemented in a Windows computer equipped with an NVIDIA RTX3090 GPU and connected as follows. KiRT is written in C++ and CUDA and compiled as a dynamic loading library that communicates with the view synthesis module using a C interface. The reimplementation of RVS uses C++ and Vulkan and communicates with the software that manages the headset using the OpenXR standard.

The robot control software coupling robotic manipulator and lambda.7 was implemented in Matlab/Simulink and executed on a Linux computer at 1 kHz.

2.3 | Experimental setup

Figures 1 and 2 present the experimental setup including the robotic manipulator with experimental scene, camera array, and the user interface consisting of the HMD, haptic interface, and foot pedals. Participants were asked to rest their elbows on the tabletop to avoid too much physical effort during the experiment. Holding the HMD served reducing the physical load on the head and neck and enabled the participant to keep the HMD in the right pose. This is important to ensure optimal visibility on the central light-field displays. A foot pedal was used to couple to and decouple from the robot.

2.4 | Experimental task

The scene with the pegs of the depth-matching task is visualized as a 3D model in Figure 3. The robot was equipped with the end-effector depicted in Figure 1 that had the same profile as the four pegs.

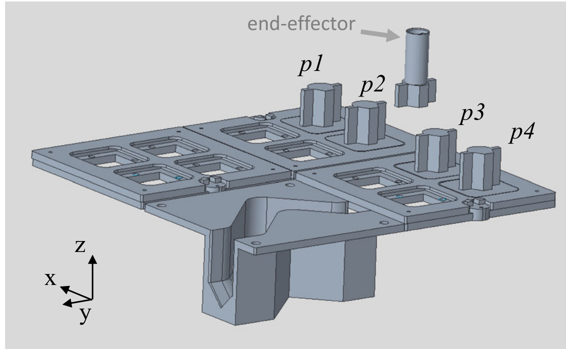


FIGURE 3 The scene seen from initial perspective for setting *C-II-C*.

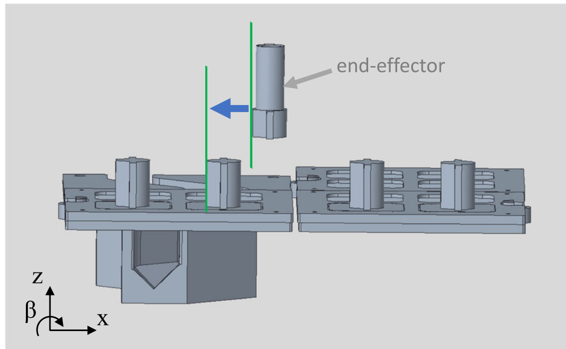


FIGURE 4 Exemplary robot starting pose at column 3.

Figures 4 and 5 present a view onto the scene from the right-hand side. In the beginning of each subtask (different column positions), the robot's end-effector was automatically positioned at a distance behind the column that had to be matched. This distance was identical for all pegs. The task was the depth-matching of the front plane of end-effector and column as marked in green in Figure 4. The robot was fixed in orientation and in z- and y-direction such that the user only had to vary the x-position of the robot via the haptic interface. The task was initialized by the experimenter. When the subject reached the subjectively perceived correct depth position (compare Figure 5), she/he informed the experimenter who logged the final pose. Since the end-effector was positioned 1.7 cm above the column tip (see Figure 4), force-feedback was irrelevant and the visual effort was intensified since the user had to refocus between end-effector and column depth.

For comparison *C-II*, the perspectives on the scene were varied. Two initial perspectives were set. Figure 6 presents the initial view of perspective setting *C-II-A* and *C-II-B* and Figure 3 the one of *C-II-C*. These initial view perspectives are additionally marked with light dashed arrows. In setting *C-II-A*, the head motion was disabled

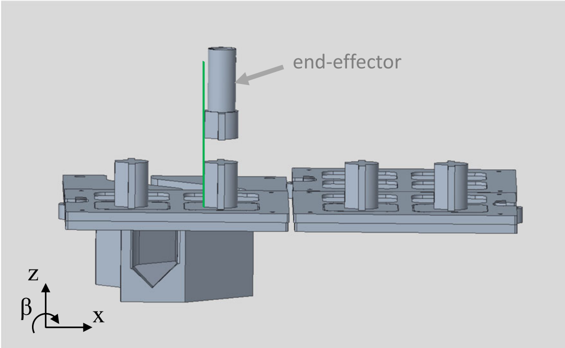


FIGURE 5 Exemplary robot goal pose at column 3.

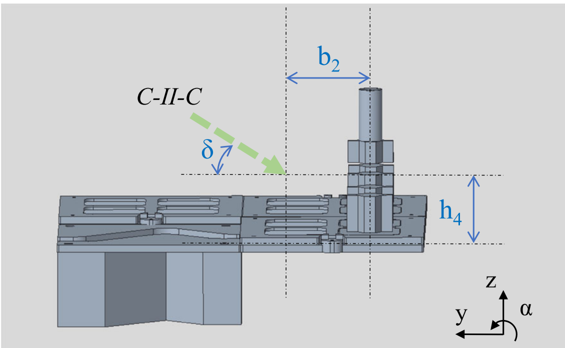


FIGURE 6 Initial perspective for setting *C-II-A* and *C-II-B*.

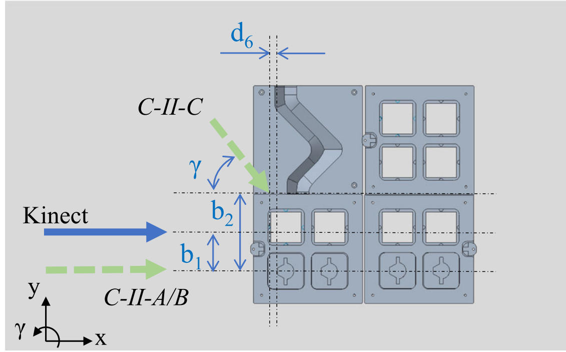


FIGURE 7 Distances in top view.

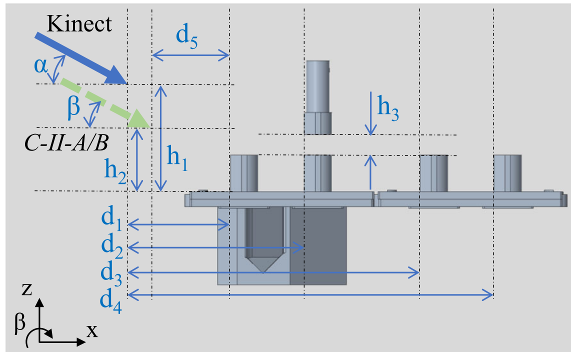


FIGURE 8 Distances in side view.

to achieve a fixed straight view, while a full 6-DoF head motion with unlimited motion range was enabled in *C-II-B* and *C-II-C*. This setting design of *C-II-A* was chosen as a reference setting for multiple reasons. First, the setting represents the choice of different standard solutions for visual feedback in teleoperation. The fixed camera position comparable with the project Space Factory[26] as described above and a stereo camera on a pan-tilt unit are represented in *C-II-A*. The rotations of a pan-tilt unit give no additional cues on a central scene in front of the user (since the same light rays enter the lens from the respective parts of the scene) but extends the overview on the surroundings. That means, the pan-tilt camera is in the center of a scene while in applications requiring high precision, the manipulated object is in the center and should be viewable from different sides. Second, the light-field displays provide a 30° FOV, which required that—to ensure that the scene is viewed through the central light-field displays and not the background display—regarding the aimed workspace/scene, the rotations of the pan-tilt had to be restricted.

Choosing the same visual sensor and image processing pipeline allowed for a comparison irrespective of the remaining artifacts of the prototype pipeline status and sensor limitations.

The distances and angles of Kinect Azure and view perspectives *C-II-A* to *C-II-C* marked in Figures 6–8 are as follows: $d_1 = 0.43$ m, $d_2 = 0.51$ m, $d_3 = 0.65$ m, $d_4 = 0.73$ m, $d_5 = 0.23$ m, $d_6 = 0.08$ m, $h_1 = 0.25$ m, $h_2 = 0.12$ m, $h_3 = 0.017$ m, $h_4 = 0.2$ m, $b_1 = 0.095$ m, $b_2 = 0.315$ m, $\alpha = 20^\circ$, $\beta = 20^\circ$, $\gamma = 45^\circ$, and $\delta = 35^\circ$.

This task design required a comparably high level of accuracy, therefore leading to a high operator workload. Thus, although requiring low user experience, the visual effort was increased. The Azure position was chosen such that all front edges of the peg were visible. The sensor was shifted slightly to the left to ensure that the nonfrontal planes did not show artifacts as holes. Figures 9 and 10 present the initial views for all conditions resulting from this choice of sensor position. Note that by using additional cameras, accounting for correct camera calibration and correcting depth map errors, such artifacts can be prevented.

2.5 | Experimental design and procedure

A within-subject design with *LIGHTFIELD* (*LF on*,*C-I-A* vs. *LF off*, *C-I-B*) and *PERSPECTIVE* (*C-II-A*, *C-II-B*, and *C-II-C*) as experimental factors was implemented, while

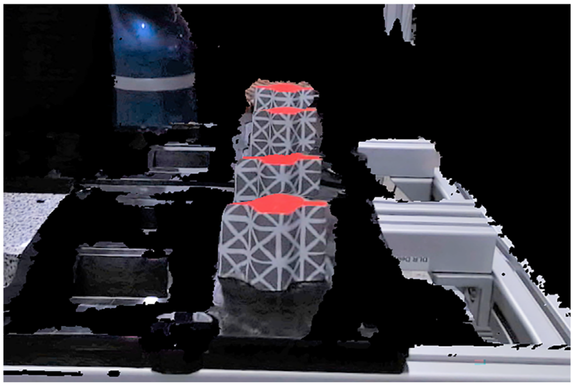


FIGURE 9 View from *C-II-A* and *C-II-B* initial perspective.

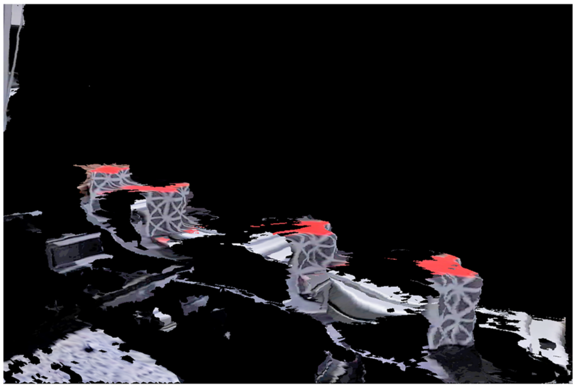


FIGURE 10 View from *C-II-C* initial perspective.

the order of all conditions was counterbalanced across subjects. The order of pegs (*p*1, *p*2, *p*3, and *p*4; compare Figure 3) was systematically varied since it potentially interacted with the light-field and perspective conditions.

After presentation of the study background by the experimenter, the participants had to fill the informed consent and the demographic questionnaire. The participants were told that the main performance criteria is accuracy in depth direction. It was recommended to use the arm rest for left and right arm and to hold the HMD with the left hand. When attaching the HMD, the participants had to adjust the eye distance and to vary the vertical and horizontal HMD pose with respect to the head to find the optimal pose and match the eye-box of the light-field displays.

The following experimental blocks with three *PERSPECTIVE* conditions were performed twice, with *LIGHTFIELD* condition (*LF on*) and non-light-field condition (*LF off*). In the C-I block, the depth-matching task had to be performed with three different *PERSPECTIVE* settings (*C-II*). For each *PERSPECTIVE* setting, all pegs had to be matched in depth in systematically varied order. Also, the orders of *LIGHTFIELD* (*C-I*) and *PERSPECTIVE* conditions (*C-II*) were systematically varied. During each main block *C-I*, the items of the post-trial questionnaires were rated verbally since the HMD should not be layed down during the block. Both post-block questionnaires were filled out by the participant. The training was done only in the very first block. Before the first run, in both main *C-I* blocks, the participant had to set up the HMD correctly. After each change of the *PERSPECTIVE* setting (in the *C-II* block), the HMD position was re-initialized, such that the user had to look straight for a pose reset.

2.6 | Measures and statistical analysis

2.6.1 | Objective measures

The motion path lengths (measured at input device), the completion times, as well as the position error, the number of changes in motion direction, and sign of the error were recorded as objective measures.

2.6.2 | Subjective measures

In an interim questionnaire after each trial, participants subjectively rated the overall workload[34] (scale ranging from 1 = *very low* to 20 = *very high*), the degree to which they felt present in the VR (in %, 0% = *no experience of presence*; 100% = *like in the real world*), how confident they felt in performing the tasks (from 1 = *not at all* to 7 = *very confident*), the quality of depth perception (from 1 = *no depth perception at all* to 7 = *like in reality*), and finally how well they could recognize objects (from 1 = *not at all* to 7 = *perfectly*), referring to the sharpness (exposure) of an object in front of the background or the objects in the close environment respectively as an indicator for the quality of the depth of field effect and the aid it provides to the user. The postcondition questionnaire included the Simulator Sickness Questionnaire (SSQ,[35] with nausea, disorientation, and oculomotor symptom clusters), the Visual Strain Questionnaire (VSQ[36]), and the NASA Task Load Index (NASA-TLX[37]) questionnaire.

2.6.3 | Statistical analysis

For the objective measures, 2 (LIGHTFIELD: off vs. on) × 3 (PERSPECTIVE) × 4 (PEG) repeated measures ANOVAs (rmANOVA) were performed. In case of nonsphericity, Greenhouse-Geisser (GG.) corrections were made.

Post hoc comparisons were performed with Bonferroni corrections.

For the subjective measures, nonparametric tests were performed. For the post-trial questionnaires, the effect of LIGHTFIELD (off vs. on) was tested performing Wilcoxon tests on the average ranks across perspectives. Then, the effects of PERSPECTIVE was investigated using Friedman tests on the ranks averaged across both LIGHTFIELD conditions. Again, Wilcoxon tests were performed for the pairwise comparisons of the three perspectives. Also for the post-block questionnaires, Wilcoxon tests were conducted.

3 | RESULTS

In the following analysis of the objective data, three subjects had to be excluded because the data were partially damaged.

3.1 | Objective data

Final position (translational error). RmANOVA on final positions revealed significant main effects of PERSPECTIVE [$F(1.54,36.85) = 4.55$ (GG.); $p < 0.05$] and PEG [$F(2.34,56.04) = 18.35$ (GG.); $p < 0.001$] and a significant interaction effect between both factors [$F(4.22,101.30) = 3.23$ (GG.); $p < 0.05$]. First, the errors were significantly smaller with Perspective C-II-C compared to Perspective C-II-A and C-II-B (both $ps < 0.05$). Second, the smallest error occurred for Peg 4 (all $ps < 0.01$), followed by Peg 3 ($p_{3vs1} < 0.01$), Peg 2, and Peg 1. The overall pattern revealed that only for Perspective *C-II-B* the errors for Peg 1 and 2 were significantly higher compared to Peg 3 and 4 (all $ps < 0.01$ in post-hoc comparisons); see Figure 11.

Path length. RmANOVA showed a significant main effect of PERSPECTIVE [$F(2,48) = 13.47$; $p < 0.001$]. Post hoc comparisons indicated that path lengths were significantly longer with Perspective *C-II-B* compared to *C-II-A* and *C-II-C* (both $ps < 0.01$). Moreover, a significant LIGHTFIELD × PEG interaction was found [$F(3,72) = 3.10$; $p < 0.05$], indicating that for Peg 1, a significant effect of LIGHTFIELD was evident ($p < 0.05$); that is, path lengths were longer in the *LF on* (*C-I-A*) compared to the *LF off* condition.

Motion time. A significant PERSPECTIVE main effect [$F(2,48) = 10.82$; $p < 0.001$] occurred; times were significantly longer with Perspective *C-II-B* compared to *C-II-A* and *C-II-C* (both $ps < 0.05$).

Direction changes in direction and error. Similar to motion times, a significant PERSPECTIVE main effect was evident [$F(1.62,38.79) = 10.79$ (GG.); $p < 0.001$] and significantly more direction changes in velocity were found with Perspective *C-II-B* compared to *C-II-A* and *C-II-C* (both $ps < 0.05$). The findings for changes in error showed the very same effect [$F(1.60,38.42) = 11.18$ (GG.); $p < 0.001$], post hoc comparisons for Perspective *C-II-B* versus *C-II-A* and *C-II-C* also reached significance (both $ps < 0.05$).

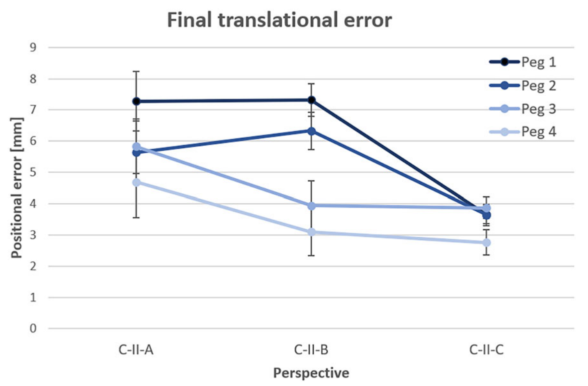


FIGURE 11 Final translational error.

Table 1 summarizes the results in the performance data.

3.2 | Subjective data

Subsequent to each experimental trial, subjects rated their experiences/impressions in a post-trial questionnaire. At the end of the two experimental blocks (*LF on* vs. *LF off*), participants also completed a post-block questionnaire.

3.2.1 | Post-trial questionnaire

Overall workload. Comparing the perceived overall workload with Wilcoxon test indicated a significant effect of LIGHTFIELD [$Z = 1.98$; $p < 0.05$]; that is, the workload was rated lower in the *LF on* compared to the *LF off* condition ($M_{LF} = 6.73$; $M_{noLF} = 7.27$). Subsequent explorative analyses also revealed that the above main effect was more evident for participants with VR experience LIGHTFIELD [$Z = 2.14$; $p < 0.05$] compared to participants without VR experience. This positive effect of *LF on* versus *LF off* only occurred with Perspective *C-II-B* and *C-II-C* (both $ps < 0.05$), but not with Perspective *C-II-A*.

Sense of Presence. However, Friedman test revealed a significant effect of PERSPECTIVE [$Chi^2 = 11.6$; $p < 0.01$];

TABLE 1 Results—objective measures.

	Perspective C-II-A	Perspective C-II-B	Perspective C-II-C	Statistical sign
Final position error [mm]				
LF ON (C-I-A)	5.83 (6.18)	5.16 (3.01)	3.39 (0.94)	Main Effects Perspective and Peg
LF OFF (C-I-B)	5.89 (1.27)	5.19 (3.39)	3.57 (1.62)	Interaction Effect Persp. × Peg
Path length [mm]				
LF ON (C-I-A)	69.49 (32.98)	85.55 (34.69)	61.70 (21.00)	Main Effect Perspective
LF OFF (C-I-B)	59.42 (22.47)	82.32 (21.65)	59.21 (13.99)	Interaction Effect LF × Peg
Completion time [s]				
LF ON (C-I-A)	11.91 (6.52)	16.92 (8.78)	12.91 (5.68)	Main Effect Perspective
LF OFF (C-I-B)	11.12 (4.82)	16.02 (5.72)	12.78 (3.99)	
Changes in vel. [#]				
LF ON (C-I-A)	248.15 (142.69)	368.67 (207.43)	275.91 (126.63)	Main Effect Perspective
LF OFF (C-I-B)	235.82 (105.60)	346.07 (135.05)	276.68 (98.31)	
Changes in err. sign [#]				
LF ON (C-I-A)	239.86 (130.27)	342.93 (185.50)	260.52 (113.94)	Main Effect Perspective
LF OFF (C-I-B)	225.92 (93.46)	329.68 (120.60)	260.67 (81.51)	

subsequent Wilcoxon tests indicated that the sense of presence was significantly higher for Perspective *C-II-B* compared to Perspective *C-II-A* ($p = 0.001$).

Confidence. Again, a significant effect of PERSPECTIVE was found [$\text{Chi}^2 = 7.3$; $p < 0.05$]. Subjects were most confident with Perspective *C-II-C* compared to Perspective *C-II-A* and *C-II-B* ($p = 0.001$ and $p < 0.01$) and more confident with Perspective *C-II-B* than with *C-II-A* ($p < 0.05$).

Depth Perception. Here, the significant effect of PERSPECTIVE [$\text{Chi}^2 = 17.9$; $p < 0.001$] showed that depth perception with Perspective *C-II-A* was rated significantly lower than for the other two perspectives (both *p*s < 0.001), and ratings were also lower for *C-II-B* compared to *C-II-C* ($p < 0.05$).

Object Recognition. As before, a significant effect of PERSPECTIVE [$\text{Chi}^2 = 23.5$; $p < 0.001$] was found with Friedman's test, indicating significantly lower values for Perspective *C-II-C* compared to *C-II-A* ($p = 0.001$) and *C-II-B* ($p < 0.001$).

3.2.2 | Postcondition questionnaire

Here, the sum scores for the simulator sickness questionnaire (SSQ), the visual strain questionnaire (VSQ), the NASA-TLX, and the SUS usability questionnaire were analyzed using Wilcoxon tests comparing *LF on* versus *LF off* conditions. Yet, no significant difference was found for any of the measures.

4 | DISCUSSION

In the present study, subjects were instructed to position the pegs in the telerobotic setup as accurately as possible. Consequently, the main performance measure is the final positional error. Besides, path lengths, motion times, and direction changes were analyzed as secondary measures, which mainly provide information about the extent of positional corrections. Together with the subjective ratings, subsequent to each trial and each block, the above stated hypotheses were tested. First, it was hypothesized that light-field technology leads to reduced visual effort:

4.1 | H1

Comparing the results from the experimental blocks with (*LF on*) versus without light-field (*LF off*) showed no significant effects on the positional accuracy, which is also in line with prior experiments of the authors.[9] However, path lengths were significantly longer when positioning the most distant peg (Peg 1) with light-field activated. This effect can best be explained by the fact that the depth of field resolution decreases the further away an object is. The most distant object (where the positioning accuracy was also the lowest) apparently led to the greatest uncertainty in positioning, leading to longer trajectories (and motion times, although significance was not reached for this metric) with light-field activated. Obviously, the subjects tried to use the additional information of the light-field display (potentially increased confidence) and to correct the position accordingly, even if they could not achieve a more accurate result compared to the condition without light-field. Contrary to the prior study[9] (where the depth map was perfect and 3D instead of 2.5D), no evidence was found that light-field technology directly reduces the visual strain of the human operator and no effects in the SSQ and VSQ queries were found. Yet, the overall workload ratings indicated lower values when light-field was available compared to the condition without. This effect was particularly evident for subjects who had at least basic experience with VR technology and when 6-DoF head motion was enabled (*C-II-B* and *C-II-C*). This finding provides at least initial evidence on favor of H1.

4.2 | H2

The findings for positional accuracy showed that the best results were obtained with *C-II-C* compared to the other perspectives. For perspective *C-II-B*, the level of accuracy was moderated by peg distance (see Figure 11); that is, for the more near pegs 3 and 4, better results were achieved as compared to the more distant pegs 1 and 2.

Analyzing the secondary performance measures (path lengths, motion times, direction changes) indicated the longest paths and times as well as highest number of direction changes for *C-II-B*, also indicating that with the additional visual information subjects performed more corrective motions compared to perspective *C-II-A*. With perspective *C-II-C* the benefits of view synthesis are more evident, with regard to the secondary measures when compared to *C-II-B*. This is discussed in more detail for hypothesis H3. The subjective ratings showed (1) a higher sense of presence with *C-II-B* compared to *C-II-A*, (2) a higher level of confidence with view synthesis than without, and overall the highest confidence level with *C-II-C*, (3) improved depth perception with view synthesis, and (4) the lowest ratings for *C-II-C* in terms of object recognition. The sense of presence increases with 6-DoF head motion but is limited by the 2.5D nature of the visualization. The confidence seems to increase with the quality of depth information (optimal view from the side in *C-II-C*) which corresponds to the results on accuracy. Still, due to the 2.5D visualization, the object recognition is obviously reduced when looking from the side. In summary, H2 is at least partially confirmed.

4.3 | H3

Specifically large rotational and translational perspective (*C-II-C*) changes improve depth perception for more distant objects such that depth positioning is improved. Indeed, the positional accuracy and confidence was highest in *C-II-C* compared to *C-II-A* and *C-II-B*, which confirms H3. Additionally, motion paths and times were shorter and fewer direction changes occurred with perspective *C-II-C* when compared to *C-II-B*. This indicates that the subjects did not move their head in condition *C-II-B* up to the perspective *C-II-C*, potentially because of limited comfort (such that the depth information was better for perspective *C-II-C* despite 6-DoF head motion in *C-II-B*). Or, this indicates that moving the head to this perspective costs additional time (which would not explain the reduced results on depth accuracy in *C-II-B* when compared to *C-II-C*). Object recognition was rated worse in *C-II-C* compared to the other perspective, which can be explained by the fact that only 2.5D was available in this perspective, as discussed above. Note that test subjects may have referred object recognition to the correct interpretation of an object's shape rather than to the visibility of the object. Overall, H3 was confirmed. It has to be mentioned that the benefits regarding depth accuracy of C-II-C were not limited to distant objects.

4.4 | Summary

Providing translational motions for change of view perspective to the operator increases confidence and depth matching accuracy when compared to state-of-the-art solutions such as pan-tilt units[25] or steady cameras.[26] Furthermore, large changes of view perspective (*C-II-C*) are especially helpful for more distant objects (compare Figure 11). Although the 2.5D visualization, which was especially emphasized in the perspective looking from the side onto the scene, limited the object recognition and sense of presence, the depth accuracy was not negatively affected. Such large variations in view perspective are reachable when different initial perspectives are provided to the user or especially translational head motions are scaled up. Still, in future work, the maximal scaling with regard to video delay and frames per second among others needs to be evaluated. Robotic solutions for 6-DoF camera motions may reduce the 2.5D limitations, but are by far costlier and have presumably a much more limited workspace than the digital HoviTron solution due to the robot workspace itself or safety aspects such as collision avoidance. Subjectively rated, the light-field visualization *C-I-A* could not be differentiated from standard visualization *C-I-B*. Still, the workload was significantly reduced through light-field providing a natural, physically correct depth visualization.

4.5 | Limitations

The depth sensor of the Kinect Azure is based on the time-of-flight principle. Such sensors typically have inaccuracies at edges when only one plane of this edge is visible to the sensor resulting in artifacts as wrong edge positions with errors of up to 0.5 cm at about 50-cm distance from the sensor. Therefore, in this study, we designed the task scene according to this sensor weakness: Insertion tasks were not feasible since the reference edge of the insertion was distorted through the described artifact. The pegs of the *depth-matching task* were designed to have mainly edges with two visible planes by adding a wall separating the cylindrical and cornering sides of the peg. Note that within the HoviTron project, other sensor types such as light-field cameras were investigated for which this artifact did not appear.

It would have been optimal to integrate additional nondistorted separate stereo camera systems with fps, delay, resolution, and so forth matching with the video pipeline. Still, the video pipeline is in a prototype status such that certain artifacts (as described above for the Azure) could not be removed yet. Therefore, for the sake of comparability, the same video information was used for all conditions.

The sensor was shifted slightly ($b_1 = 9.5$ cm) to the left such that the view from the right might have lead to different results since the right side of the object was not perfectly visible. Note that by using additional cameras, accounting for correct camera calibration, and correcting depth map errors, a large portion of the scene should be possible to cover in future.

The study lasted 1 h for the participants on average. During the first block with training phases, the HMD was worn for 30 min and for another 20 min during the second and third block each. The weight of the HMD and the limited usability due to its prototype state might have influenced the results toward the end of the study.

5 | CONCLUSION

This work presented the first statistical analysis of benefits of light-field visualization in near-eye displays involving a real-time video pipeline. While performing a depth-matching task, the participants perceived significantly lower overall workload with light-field display when compared to a display configuration with fixed focal plane. Besides the light-field display, the effects of real-time view synthesis were evaluated. When compared to a state-of-the-art teleoperation visualization, the benefits of view perspective changes in 6-DoF were evident especially for more distant objects. For distant objects, it was found that a separate initial perspective close to these objects eased the depth-matching for the user. In future work, the HoviTron pipeline equipped with more advanced depth sensors should be compared to a real stereo camera stream.

ACKNOWLEDGEMENTS

Open Access funding enabled and organized by Projekt DEAL.

ORCID

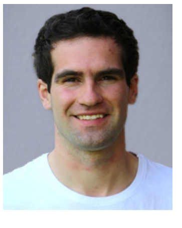
Bernhard Weber https://orcid.org/0000-0002-7857-0201
Michael Panzirsch https://orcid.org/0000-0002-0647-7147

REFERENCES

1. Lee J, Balachandran R, Sarkisov YS, De Stefano M, Coelho A, Shinde K, et al. Visual-inertial telepresence for aerial manipulation. *2020 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE; 2020. p. 1222–9.
2. Vogel J, Leidner D, Hagengruber A, Panzirsch M, Bauml B, Denninger M, et al. An ecosystem for heterogeneous robotic assistants in caregiving: core functionalities and use cases. IEEE Robot Autom Mag. 2020;28(3):12–28.
3. Panzirsch M, Pereira A, Singh H, Weber B, Ferreira E, Gherghescu A, et al. Exploring planet geology through force-feedback telemanipulation from orbit. Sci Robot. 2022;7(65): eabl6307.
4. Schwarz M, Lenz C, Rochow A, Schreiber M, Behnke S. Nimbro avatar: interactive immersive telepresence with force-feedback telemanipulation. *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE; 2021. p. 5312–9.
5. Vaz JC, Dave A, Kassai N, Kosanovic N, Oh PY. Immersive auditory-visual real-time avatar system of ANA Avatar XPRIZE finalist Avatar-Hubo. *2022 IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO)*, IEEE; 2022. p. 1–6.
6. Schwarz M, Behnke S. Low-latency immersive 6D televisualization with spherical rendering. *2020 IEEE-RAS 20th International Conference on Humanoid Robots (HUMANOIDS)*, IEEE; 2021. p. 320–5.
7. Aykut T, Karimi M, Burgmair C, Finkenzeller A, Bachhuber C, Steinbach E. Delay compensation for a telepresence system with 3D 360 degree vision based on deep head motion prediction and dynamic FoV adaptation. IEEE Robot Autom Lett. 2018;3(4):4343–50.
8. Biener V, Kalamkar S, Nouri N, Ofek E, Pahud M, Dudley JJ, et al. Quantifying the effects of working in VR for one week. arXiv e-prints, arXiv–2206;2022.
9. Panzirsch M, Weber B, Bechtel N, Grabner N, Lingenauber M. Light-field head-mounted displays reduce the visual effort: a user study. J Soc Inf Display. 2022;30(4):319–34.
10. Banks MS, Kim J, Shibata T. Insight into vergence/accommodation mismatch. *Head-and Helmet-mounted Displays XVIII: Design and Applications*,

vol. 8735, International Society for Optics and Photonics; 2013. p. 873509.
11. Park J-H, Kim S-B. Optical see-through holographic near-eye-display with eyebox steering and depth of field control. Opt Express. 2018;26(21):27076–88.
12. Tan G, Zhan T, Lee Y-H, Xiong J, Wu S-T. Polarization-multiplexed multiplane display. Opt Lett. 2018;43(22):5651–4.
13. Zhan T, Zou J, Lu M, Chen E, Wu S-T. Wavelength-multiplexed multi-focal-plane seethrough near-eye displays. Opt Express. 2019;27(20):27507–13.
14. Konrad R, Padmanaban N, Molner K, Cooper EA, Wetzstein G. Accommodation-invariant computational near-eye displays. ACM Trans Graphics (TOG). 2017;36(4):1–12.
15. Rathinavel K, Wang H, Blate A, Fuchs H. An extended depth-at-field volumetric near-eye augmented reality display. IEEE Trans Visual Comput Graph. 2018;24(11):2857–66.
16. Dunn D, Tippets C, Torell K, Kellnhofer P, Akşit K, Didyk P, et al. Wide field of view varifocal near-eye display using see-through deformable membrane mirrors. IEEE Trans Visual Comput Graph. 2017;23(4):1322–31.
17. Sluka T. Near-eye sequential light-field projector with correct monocular depth cues. *European Patent EP3542206B1*; 2017.
18. Bonatto D, Fachada S, Lafruit G. RaViS: real-time accelerated view synthesizer for immersive video 6DoF VR. *Society for Imaging Science and Technology (IS&T) - Electronic Imaging*; Burlingame, USA; 2020. Place: San Francisco, California, USA.
19. Bonatto D, Fachada S, Rogge S, Munteanu A, Lafruit G. Real-time depth video-based rendering for 6-DoF HMD navigation and light field displays. IEEE Access. 2021;9:146868–87.
20. Bonatto D, Hirt G, Kvasov A, Fachada S, Lafruit G. MPEG immersive video tools for light field head mounted displays. *2021 International Conference on Visual Communications and Image Processing (VCIP)*, IEEE; 2021. p. 1–2.
21. Fachada S, Bonatto D, Teratani M, Lafruit G. View synthesis tool for VR immersive video.. In: Sobota DB, editor.. 3D Computer Graphics. Rijeka: IntechOpen, 2022. https://doi.org/10.5772/intechopen.102382
22. Lafruit G, Van Bogaert L, Aragon JS, Panzirsch M, Hirt G, Strobl KH, Martinez EJ. Tele-robotics VR with holographic vision in immersive video. *Proceedings of the 1st Workshop on Interactive Extended Reality*; 2022. p. 61–8.
23. Di Castro M, Ferre M, Masi A. CERNTAURO: a modular architecture for robotic inspection and telemanipulation in harsh and semi-structured environments. IEEE Access. 2018;6:37506–22.
24. Fong T, Rochlis Zumbado J, Currie N, Mishkin A, Akin DL. Space telerobotics: unique challenges to human–robot collaboration in space. Rev Human Fact Ergon. 2013;9(1):6–56.
25. Lii NY-S, Schmaus P, Leidner D, Krueger T, Grenouilleau J, Pereira A, et al. Introduction to surface avatar: the first heterogeneous robotic team to be commanded with scalable autonomy from the ISS. *Proceedings of the International Astronautical Congress, IAC*, International Astronautical Federation, IAF; 2022.
26. Kempf F, Mühlbauer MS, Dasbach T, Leutert F, Hulin T, Radhakrishna Balachandran R, et al. AI-In-Orbit-Factory-AI approaches for adaptive robotic in-orbit manufacturing of modular satellites. *Proceedings of the International Astronautical Congress, IAC*; 2021.
27. Boyce JM, Doré R, Dziembowski A, Fleureau J, Jung J, Kroon B, Salahieh B, Vadakital VKM, Yu L. MPEG immersive video coding standard. Proc IEEE. 2021;109(9):1521–36.
28. Rogge S, Bonatto D, Sancho J, Salvador R, Juarez E, Munteanu A, Lafruit G. MPEG-I depth estimation reference software. *2019 International Conference on 3D Immersion (IC3D)*, IEEE; 2019. p. 1–6.
29. Mieloch D, Stankiewicz O, Domaski M. Depth map estimation for free-viewpoint television and virtual navigation. IEEE Access. 2020;8:5760–76.
30. Sancho J, Sutradhar P, Rosa G, Chavarrias M, Perez-Nunez A, Salvador R, et al. GoRG: towards a GPU-accelerated multiview hyperspectral depth estimation tool for medical applications. Sensors. 2021;21(12):4091.
31. Xie Y, Fachada S, Bonatto D, Teratani M, Lafruit G. View synthesis: lidar camera versus depth estimation. 2021.
32. Gong X, Liu J, Zhou W, Liu J. Guided depth enhancement via a fast marching method. Image Vis Comput. 2013;31(10): 695–703. https://www.sciencedirect.com/science/article/pii/S0262885613001108
33. Jiang L, Xiao S, He C. Kinect depth map inpainting using a multi-scale deep convolutional neural network. *Proceedings of the 2018 International Conference on Image and Graphics Processing*, ICIGP 2018. Association for Computing Machinery; New York, NY, USA; 2018. p. 91–5. https://doi.org/10.1145/3191442.3191464
34. Vidulich MA, Tsang PS. Absolute magnitude estimation and relative judgement approaches to subjective workload assessment. *Proceedings of the Human Factors Society Annual Meeting*, vol. 31, SAGE Publications Sage CA: Los Angeles, CA; 1987. p. 1057–61.
35. Kennedy RS, Lane NE, Berbaum KS, Lilienthal MG. Simulator sickness questionnaire: an enhanced method for quantifying simulator sickness. The Int J Aviat Psychol. 1993;3(3): 203–20.
36. Howarth PA, Istance HO. The association between visual discomfort and the use of visual display units. Behav Inf Technol. 1985;4(2):131–49.
37. Hart SG, Staveland LE. Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. Advances in Psychology, vol. 52: Elsevier, 1988. p. 139–83.

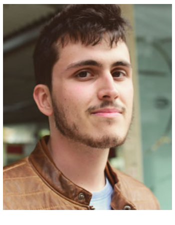
AUTHOR BIOGRAPHIES



Nicolai Bechtel received his Master of Science in Computational Engineering from the University of Applied Sciences Munich in 2018. Since then, he has been conducting research in the field of haptics and virtual reality as a research assistant at the Center for Robotics and Mechatronics of the German Aerospace Center (DLR) in Oberpfaffenhofen. His research focuses on haptics, multidynamic simulations, and development of virtual reality environments. He is currently working on topics such as Multi-Contact Haptics and Model-Based Teleoperation.

Bernhard Weber received his Dipl.-Psych. and PhD degree at the University of Würzburg, Germany, in 2004 and 2008, respectively. From 2008–2010, he was with the German Aerospace Center (DLR), Institute of Flight Guidance, Brunswick, Germany and since 2010 at the DLR Institute of Robotics and Mechatronics, Wessling, Germany, as a human factors expert. His main research interests are human factors in telerobotic systems, evaluation of haptic interaction technology, and sensorimotor performance under conditions of microgravity.

Pascal Severin is currently enrolled in the Master's program of Human Factors Engineering at TUM and received his Bachelor of Science in Psychology from the University of Vienna in 2020. As a working student at the German Arospace Center (DLR) in Oberpfaffenhofen, he has worked on research of Mixed Reality HMIs in teleoperation, hybrid robotic surgery, and human position sense under conditions of altered gravity.



Jaime Sancho Aragón received his MSc degree in Systems and Services Engineering for the Information Society from Universidad Politécnica de Madrid (UPM), Spain, in 2018. He is currently a PhD student at the Electronic and Microelectronic Design Group (GDEM) in the Software Technologies and Multimedia Systems for Sustainability (CITSEM) Research Center, UPM. His research interests include biomedical real-time systems and immersive computer vision technologies.

Laurie Van Bogaert obtained her master's degree of science in Computer Science and Engineering in 2021 from the Université Libre de Bruxelles, Belgium. She is currently pursuing her PhD degree on Multi-Plane Image with Transparency and Specularity for 6 degrees of Freedom Virtual Reality. She received a FRIA funding for her PhD from the Walloon Region, Belgium. Her research interests are novel view synthesis, real-time rendering, Vulkan, Cuda and OpenXR.

Michael Panzirsch received his diploma in mechanical engineering from the Technische Universität München in 2010. Since then he is with the Department for Analysis and Control of Advanced Robotic Systems of the German Aerospace Center (DLR) in Oberpfaffenhofen as a research associate. He finished his PhD thesis on passivity-based multilateral control for delayed teleoperation at the Polytechnical University of Madrid (UPM) in October 2018. His main areas of research interests are teleoperation of stationary and mobile robots, haptics, and healthcare robotics. Currently, he is working on topics such as haptic augmentation, model-augmented teleoperation, and shared control.

How to cite this article: Bechtel N, Weber B, Severin P, Sancho Aragon J, Van Bogaert L, Panzirsch M. Toward physically realistic vision in teleoperation: A user study with light-field head mounted display and 6-DoF head motion. J Soc Inf Display. 2023;31(12):663–74. https://doi.org/10.1002/jsid.1262