

NimbRo Avatar: Interactive Immersive Telepresence with Force-Feedback Telemanipulation

Max Schwarz*, Christian Lenz*, Andre Rochow, Michael Schreiber, and Sven Behnke

Abstract—Robotic avatars promise immersive teleoperation with human-like manipulation and communication capabilities. We present such an avatar system, based on the key components of immersive 3D visualization and transparent force-feedback telemanipulation. Our avatar robot features an anthropomorphic bimanual arm configuration with dexterous hands. The remote human operator drives the arms and fingers through an exoskeleton-based operator station, which provides force feedback both at the wrist and for each finger. The robot torso is mounted on a holonomic base, providing locomotion capability in typical indoor scenarios, controlled using a 3D rudder device. Finally, the robot features a 6D movable head with stereo cameras, which stream images to a VR HMD worn by the operator. Movement latency is hidden using spherical rendering. The head also carries a telepresence screen displaying a synthesized image of the operator with facial animation, which enables direct interaction with remote persons. We evaluate our system successfully both in a user study with untrained operators as well as a longer and more complex integrated mission. We discuss lessons learned from the trials and possible improvements.

I. INTRODUCTION

Telemanipulation and telepresence are key cornerstones of robotics. On the one hand, they enable robots to perform tasks which are currently beyond the capabilities of autonomous perception, planning, and control methods—the human intellect is still unmatched in its ability to perceive, plan, and react to unforeseen situations. On the other hand, they allow humans to work in remote environments without needing to travel or to expose themselves to potential dangers, such as in disaster response. The COVID-19 pandemic has further highlighted the need of and potentials for teleoperation systems. Telerobotic systems can help reducing contacts and thus lower the infection risk. This not only includes medical work, but also helping persons requiring assistance in their activities of daily life.

The ANA Avatar XPRIZE Challenge¹ fosters development of telerobotic and telepresence systems for these and other use cases. It focuses on immersiveness and intuitive operation—both for the remote operator as well as so-called *recipients* interacting with the robot. We present our telerobotic system designed for the challenge (see Fig. 1), which is capable of locomotion in human environments, full 3D immersion, dexterous bimanual manipulation, and interaction with the recipient. Our system has an approximately humanoid shape, with two arms ending in anthropomorphic



Fig. 1. NimbRo Avatar interacting with a human recipient. Top: Remote environment. Bottom left: Human operator. Bottom right: Operator VR view.

hands. It features a 6D-movable head, carrying cameras and a telepresence screen. Our operator station facilitates full force feedback to the operator’s wrists and fingers. The operator wears a VR head-mounted display to fully immerse into the remote environment. We build the whole system using only off-the-shelf components which allows for easy replication and maintenance. In summary, our contributions include:

- 1) An anthropomorphic telemanipulation robot (avatar) with advanced manipulation and communication capabilities,
- 2) a matching operator station, allowing full telepresence and force-feedback telemanipulation,
- 3) a real-time operator head animation technique, and
- 4) integration into a telerobotic avatar system, including a VR low-latency rendering technique and a force-feedback system optimized for low-latency operation.

We evaluate our system in lab trials, with both trained and untrained operators to assess the intuitive control. Furthermore, we demonstrate a complex integrated mission consisting of six separate tasks—modeled after the ANA Avatar XPRIZE Challenge rules.

*Equal contribution.

All authors are with the Autonomous Intelligent Systems group of University of Bonn, Germany; schwarz@ais.uni-bonn.de

¹<https://www.xprize.org/prizes/avatar>

II. RELATED WORK

Telemanipulation robots are complex systems consisting of many components, which have been the focus of research both individually as well as on a systems level.

a) Telemanipulation Systems: The DARPA Robotics Challenge (DRC) 2015 [1] resulted in the development of several mobile telemanipulation robots, such as DRC-HUBO [2], CHIMP [3], RoboSimian [4], and our own entry Momaro [5]. All these systems demonstrated impressive locomotion and manipulation capabilities under teleoperation, even with severely constrained communication. However, the DRC placed no emphasis on intuitiveness of the teleoperation controls or immersion of the operators. To our knowledge, our team was the only one using a VR HMD and 6D magnetic trackers to perceive the environment in 3D and to control the robot arms—the rest of the teams relying entirely on 2D monitors and traditional input devices to control their robots. All teams, including ours, required highly trained operators familiar with the custom-designed operator interfaces. Furthermore, since the DRC was geared towards disaster response, the robots did not feature any communication capabilities for interacting with remote humans.

In our subsequent work [6], we developed the ideas embodied in the Momaro system further. The resulting Centauro robot is a torque-controlled platform capable of locomotion and dexterous manipulation in rough terrain. It is controlled by a human sitting in a dedicated operator station, equipped with an upper body exoskeleton providing force feedback and a VR HMD. Still, Centauro is focused on disaster response and does not have any communication facilities.

Schmaus *et al.* [7] discuss the results of the METRON SUPVIS Justin space-robotics experiment, where an astronaut on the ISS controlled the Justin robot on Earth, simulating an orbital robotics mission. Instead of opting for full immersion and direct control, the authors relied on a 2D tablet display and higher levels of autonomy, allowing the astronaut to trigger autonomous task skills.

In contrast to the discussed prior works, our avatar system is specifically designed to operate in human workspaces and to interact with humans. While individual aspects of this problem setting have been addressed before (and will be discussed below), there exists, to our knowledge, no integrated system designed for this purpose.

b) 3D VR Televisualization: Live capture and visualization of the remote scene is typically done using data from RGB or RGB-D cameras. There are many examples of static and movable stereo cameras on robots, which are directly visualized in a head-mounted display [8]–[10]. However, these approaches are limited either by a fixed viewpoint, or considerable camera movement latency, potentially creating motion sickness. In contrast, our system hides latencies by correcting for viewpoint changes through spherical rendering [11].

RGB-D sensors allow rendering from free viewpoints [12], [13], removing head movement latency. However, these sensors produce sparse point clouds, which can be difficult

to visualize in a convincing way. Reconstruction-based approaches [6], [14], [15] address this issue by aggregating point clouds over time and building dense representations, which can be viewed without movement latency. They still, however, struggle with many reflective and transparent materials, because the depth sensors cannot measure them. An additional drawback is that reconstruction-based approaches usually cannot deal with dynamic scenes—which is an issue when interacting with the environment and human recipients. In contrast, our method always displays a live stereo RGB stream, which has no difficulties with materials or dynamic scenes.

c) Force Feedback: Teleoperation systems use typically stationary devices to display any force feedback captured by the remote robot to the human operator [6], [16], [17]. In contrast, wearable haptic devices [18] are usually more lightweight and do not limit the operator’s workspace. However, they cannot display absolute forces to the operator.

Much recent and ongoing research focuses on stable teleoperation systems in time-delayed scenarios [19], [20]. Large time delays for teleoperation in earth-space scenarios are investigated in [21], [22]. In our application, we assume smaller distances between the operator station and the avatar robot. Thus, our force feedback controller does not need to handle such high latencies.

d) Face Animation: Visualizing facial expressions of persons wearing VR HMDs is a well-known task, also in other contexts. Usually, IR eye tracking cameras capture eye poses and expressions such as frowns, while a standard camera captures the unobscured lower part of the face. A special requirement in the Avatar challenge is that the method needs to be quickly adaptable to a new operator, as less than one hour of setup time is allotted.

A first category of HMD facial animation methods is based on explicit 3D representations. Olszewski *et al.* [23] train a neural regressor to output blend shape weights, which deform a face mesh. On the other hand, Codec Avatars [24], [25] are an implicit model, trained on multiple (usually many) images of the operator.

All the mentioned methods require either extensive manual work (3D modeling), complicated capture setups (3D reconstruction), or long training times, all of which are infeasible in the Avatar challenge. In contrast, our 2D approach is based on taking a single image of the operator and does not require any on-site training. However, the resulting quality will be lower than models especially trained or adapted to the operator at hand.

III. OPERATOR STATION

The operator controls the avatar through a dedicated operator station (Fig. 2a, Fig. 3). Two Franka Emika Panda 7DoF robotic arms are used for haptic teleoperation. The arms serve dual purposes: They measure the operator wrist pose precisely and allow direct inducement of forces at the wrist, thus facilitating force feedback. The operator hands are connected to the Panda arms through two SenseGlove hand exoskeleton haptic devices and OnRobot HEX-E 6-axis

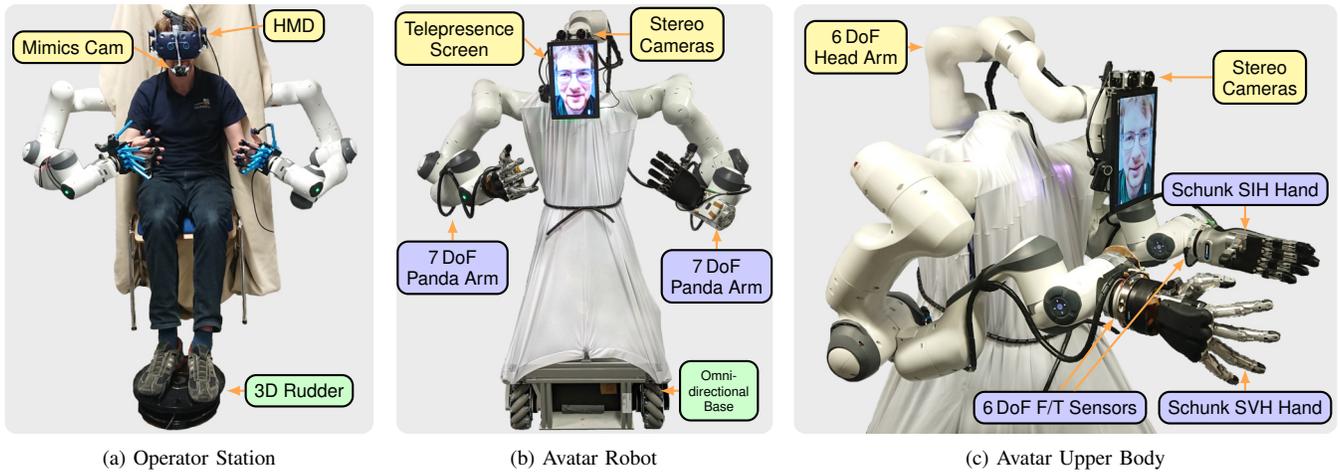


Fig. 2. NimbRo operator station and avatar robot. All components involved in the VR/telepresence system are colored yellow. The manipulation components are shown in blue. Finally, the locomotion system is colored green.

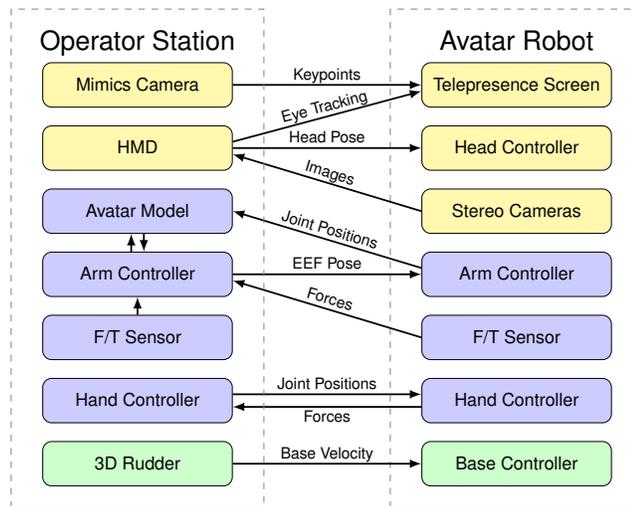


Fig. 3. Information flow between the system components. Same coloring as in Fig. 2.

force-torque sensors between Panda arms and SenseGloves. The force/torque sensors are used to measure the slightest hand movement in order to assist the operator in moving the arm, reducing the felt mass and friction to a minimum. The SenseGlove haptic interaction device features 20 DoF finger joint position measurements (four per finger) and a 1 DoF haptic feedback channel per finger (i.e. when activated the human feels resistance, which prevents further finger closing movement).

The operator wears an HTC Vive Pro Eye head-mounted display, which offers 1440×1600 pixels per eye with an update rate of 90 Hz and 110° diagonal field of view. While other HMDs with higher resolution and/or FoV exist, this device offers eye tracking, which is important for visualizing the operator's face on the avatar robot. The HMD features headphones for audio communication. We mounted an additional camera (Logitech StreamCam) in front of the operator's lower face to capture their mimics. The camera is also used for audio capture.

A 3D Rudder foot paddle device is used to capture 3-axis

locomotion control commands. The operator can command translation velocity by pitching and/or rolling their feet, as well as command yaw velocities by corresponding foot yaw movements.

To ensure safety, the operator station robotic arms are limited in the force they can exert, causing an immediate shut-down of the system when limits are exceeded. In addition, a second, supervising person can shut down the system through an E-Stop device.

A. Force-Feedback Control

Our force feedback controller [26] commands joint torques to each Panda arm and reads the current hand pose to generate the commanded hand pose sent to the avatar robot. The controller runs with an update rate of 1 kHz. To keep the operator and avatar kinematic chains independent, a common control frame is defined in the middle of the palm of both the human and robotic hands, i.e. all necessary command and feedback data are transformed such that they refer to this frame before being transmitted. As long as no force feedback is displayed to the operator, we generate a weightless feeling for moving the arm. Even though the Panda arm has a quite convenient teach-mode using the internal gravity compensation when zero torques are commanded, the weightless feeling can be further improved by using precise external force-torque sensors. The force-torque measurements are captured with 500 Hz and smoothed using a sensor-side low-pass filter with a cutoff frequency of 15 Hz.

In order to prevent the operator from exceeding joint position or velocity limits of the Panda arm, we command joint torques pushing the arm away from those limits. In addition, the avatar arm limits are displayed in a similar way with low latency, using an operator-side avatar model to predict the avatar arms' movement.

The hand controllers map measured operator finger joint angles to the avatar hands. For the right Schunk SVH hand, torque feedback in the form of motor currents is available, which is used to provide per-finger force feedback to the operator.



Fig. 4. Immersive 3D Visualization: Operator view (cropped) from different head perspectives.

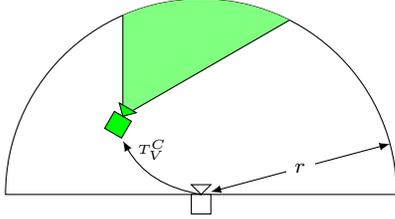


Fig. 5. Spherical Rendering compensates pose differences between VR eye and robot camera poses (transform T_V^C). The real camera FoV (green) is projected onto a sphere around the VR camera with radius r .

Any error state of the arms (see Section IV-A) is displayed to the operator as a colored overlay in VR.

B. 3D Visualization

We took special care to develop an immersive 3D visualization approach, which displays the environment around the avatar in VR to the operator. Our method achieves both low-latency response to head movements as well as real-time streaming of dynamic scenes. State-of-the-art systems can usually only achieve one of these goals. Our approach uses a 6D movable head on the robot, carrying a stereo camera with human baseline (see Section IV). The head mimics operator movements exactly. To compensate for movement latency, which introduces a pose difference T_V^C , we use spherical rendering [11]. In short, the wide-angle camera images are rendered on the operator side as spheres with radius $r = 1$ m, centered on the camera pose at time of capture. The operator can move their head in VR freely with low latency. For translations, this causes a small amount of distortion (see Fig. 5), but this is unnoticeable in most situations and is quickly corrected when the robot head arrives at its target pose.

IV. AVATAR ROBOT

The avatar robot is designed to interact with humans and made-for-humans everyday objects and indoor environments and thus features an anthropomorphic upper body (Fig. 2b, Fig. 3). Two 7DoF Franka Emika Panda arms are mounted in slightly V-shaped angle to mimic the human arm configuration. The shoulder height of 125 cm above the floor allows convenient manipulation with objects on a table, as well interaction with both, sitting and standing persons. The shoulder width of under 90 cm enables easy navigation through standard doors.

A. Manipulation

The Panda arms have a sufficient payload of 3 kg, a maximal workspace of 855 mm and the extra degree of

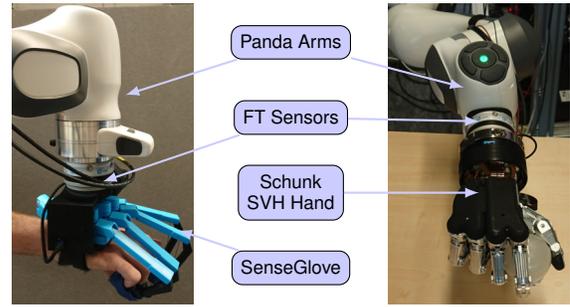


Fig. 6. Haptic interaction system. Operator interface (left) and avatar robotic hand (right).

freedom gives some flexibility in the elbow position while moving the arm. Despite existing torque sensors in each arm joint, we mounted additional OnRobot HEX-E 6-axis force/torque sensors at the wrists for more accurate force and torque measurements close to the robotic hands, since this is the usual location of contact with the robot's environment (see Fig. 6). The avatar is equipped with two Schunk hands. A 20DoF Schunk SVH hand is mounted on the right side. The nine actuated degrees of freedom provide very dexterous manipulation capabilities. The left arm features a 5DoF Schunk SIH hand for simpler but more force-requiring manipulation tasks. Both hand types thus complement each other.

Software-wise, the arms are driven with a Cartesian Impedance controller towards the operator's hand pose. The arms feature a safety mechanism which prevents the arms from exceeding certain joint torque, position, and velocity limits. In case a software stop occurs (for example when exceeding torque limits while touching a table) the arm can recover automatically by smoothly fading between the current and target arm pose.

Any force/torque measured by the sensor on the wrist is transmitted to the operator side (see Section III). Similarly, the hands receive position commands and transmit current measurements back to the operator station.

B. 3D Perception

Our robot's head is mounted on a UFACTORY xArm 6, providing full 6D control of the head pose. The robotic arm is capable of moving a 5 kg payload, which is more than enough for a pair of cameras and a small color display for telepresence. Furthermore, the arm is very slim, which results in a large workspace while being unobtrusive. Finally, it is capable of fairly high speeds (180°/s per joint, 1 m/s at the end-effector), thus being able to match dynamic human head movements.

Two Basler a2A3840-45ucBAS cameras are mounted on the head in a stereo configuration. The cameras offer 4K video streaming at 45 Hz and are paired with C-Mount wide-angle lenses, which provide more than 180° field of view in horizontal direction. We also experimented with Logitech BRIO webcams with wide-angle converters, which offer auto-focus but can only provide 30Hz at 4K, resulting in visible stutters with moving objects. The Basler cameras

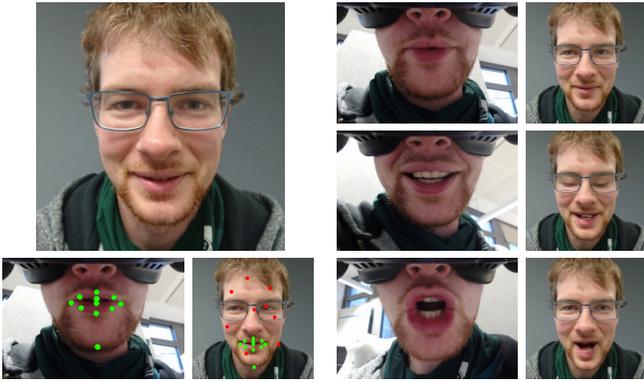


Fig. 7. Facial animation for the telepresence screen. The photograph of the operator (top left) is warped and inpainted using a learned network. The network operates on keypoints (bottom left) extracted from the lower face camera (green), and head keypoints (red), modified by eye tracking data (not shown). The right column shows the generated result. Note that the generated image shows a variety of mouth and chin poses, as well as eye movements and blinking.

are configured with a fixed exposure time (8 ms) to reduce motion blur to a minimum. For transmission, the raw camera images are MJPEG-compressed on the onboard GPU [27]. The entire pipeline achieves 30–40 ms latency from camera exposure start to outputting the image to the VR HMD, as measured by the camera-provided timestamps.

C. Telepresence Screen

For life-like conversations, it is necessary to display the operator’s face with all its expressiveness, which is difficult to achieve since the operator is wearing an HMD. To address this, we extract facial keypoints from the lower face camera and the eye tracking data. We train a keypoint detector network based on an hour glass architecture² for lower face parts. For training, we crop images from a VoxCeleb2 [28] sequence and use extracted keypoints [29] as ground truth.

We also learn unsupervised keypoints for the entire face following the method of Siarohin *et al.* [30]. The unsupervised keypoints include an eye keypoint controlling gaze direction and blinking and are used to animate a given face in a different facial expression. During training, we sample two images of the same person, denoted *source* and *target*. We extract both keypoint types from the target image and ask the generator network to produce the target image, given source image and target keypoints.

During inference, we extract the unsupervised keypoints in the operator photograph and apply the eye tracking data captured by the HMD on the eye keypoint in pixel space. We extract the lower face keypoints in the images obtained with the operator cam. Unlike in training, we merge the keypoints in the source image operator pose and feed them into a generator network, following Siarohin *et al.* [30]. The source image is then adjusted to the target keypoints by the generator network. We removed keypoint Jacobians from the pipeline, since our keypoint detector does not produce them

²Network architecture: 128×128 input size, five encoder blocks Conv2d-BN-ReLU-AvgPool, five decoder blocks Upsample-Conv2d-BN-ReLU. Output is one heat map per keypoint.

in the upper face, due to occlusions caused by the VR glasses. The generated image is then displayed on the avatar’s head in real time. Figure 7 shows example outputs of the entire pipeline.

For audio communication, the avatar is equipped with a speaker in its torso and a microphone attached to the head.

D. Locomotion & Tether

The robot upper body is mounted on a holonomic platform with four Mecanum wheels, capable of up to 2.5 m/s movement, although we usually cap the operator command at 1.5 m/s for safety reasons.

Our system currently operates with a communication & power tether, providing a 1 Gbit/s Ethernet connection through which all communication between operator station and avatar robot takes place (see Fig. 3). In its present configuration, the system consumes around 200 Mbit/s of bandwidth. Since the stereo 4K camera feed is responsible for nearly all of the bandwidth, the bandwidth can be freely adapted by changing video feed resolution and/or frame rate. The system can be operated with moderate amounts of communication latencies. When exceeding 50 ms, the force feedback mechanism will start to become unstable.

E. Safety

Both subsystems, the operator station and avatar robot, are designed to operate in contact with humans. Thus, safety mechanisms are highly important. The Franka Emika Panda arms employed in both systems feature considerable measures that detect abnormal situations, such as excessive force or speed, and immediately switch into a safe shutdown state, which holds the current position.

On any communication loss or if the operator station is paused, the avatar arms remain in their current pose and will smoothly fade to the new target pose upon receiving new commands. During operation, the avatar arms are compliant using our impedance controller.

The avatar robot also features a 6DoF head arm, which has protective measures similar to the Panda arms. If the E-Stop is activated, it will stop immediately. Furthermore, the holonomic base with its four wheels becomes de-energized immediately when the E-Stop is pressed, allowing the robot to be easily moved by hand.

V. EVALUATION

Our system has been evaluated in two larger experiments, focusing on intuitiveness and immersion in a user study and system capability in a more complex longer integrated mission.

A. User Study

In the user study experiment, participants were asked to perform several remote manipulation tasks. The avatar robot was stationed in a kitchen, out of sight and hearing range of the operator (other end of the building). The participants did not see the objects they were supposed to manipulate before the task, although most of them were familiar with the kitchen.

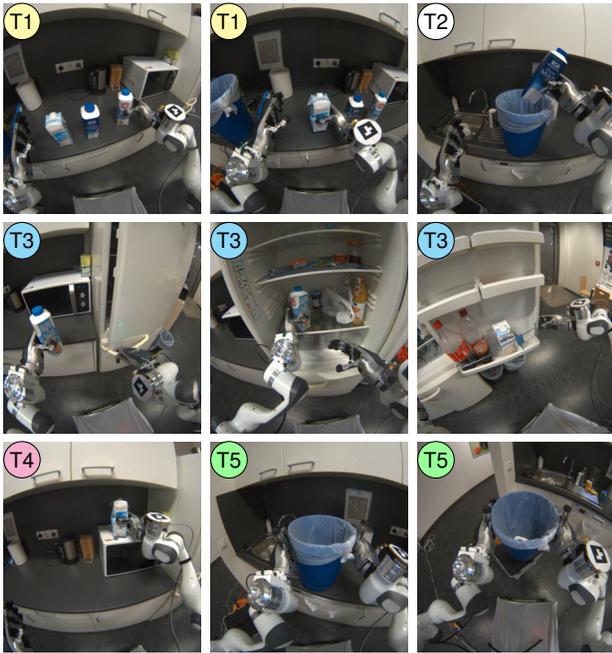


Fig. 8. User study tasks. We show cropped versions of the operator left eye VR view. Tasks: ① Sorting the cartons by weight. ② Throwing away the empty carton. ③ Opening the fridge, putting the half-full carton inside, and closing the fridge. ④ Storing the full carton on the shelf. ⑤ Grasping the waste basket with both hands and putting it down.

Due to a motor failure of the right SVH hand, the robot was equipped with two SIH hands for this test—slightly lowering the manipulation capabilities, as the SVH hand is better suited for small-scale, dexterous manipulation.

See Fig. 8 for an overview of the tasks to be performed. For the first task, the avatar should navigate to three milk cartons. The operators were told that one was full, one half full, and one empty. They should use any means available to find out which was which.

The next three tasks were to put the full carton on a shelf, store the half-full carton in the fridge (opening³ and closing it), and throw the empty one into a waste basket. These tasks could be performed in any order. The fifth and final task was to move the waste basket from the kitchen counter to its proper place, requiring bimanual manipulation.

The ongoing COVID-19 pandemic limited us to our immediate colleagues for the user study, severely constraining the scope of the study. Out of ten total participants, two were trained operators with deep knowledge of the system, five more were members of the robotics group but were unfamiliar with the system, and the final three were other staff members with no relation to robotics. All untrained operators were given a few minutes to familiarize themselves with the system, with the avatar being given a few unrelated example objects to manipulate.

Table I shows immediately quantifiable results. As a first observation, it is apparent that the trained operators could perform the task three to four times quicker, which is ex-

³During four runs, we experienced hardware failures in the right hand and had to open the fridge with human assistance.

TABLE I

USER STUDY: COMPLETION TIMES AND SUCCESS RATES

Operator group	Completion time [min:sec]		Correct
	mean	stddev	
Untrained (8)	8:05	1:15	75%
Trained (2)	2:51	1:01	100%

pected. Still, even untrained operators could perform the task in very reasonable time, given the task complexity. To put these times into perspective, our untrained operators solved the entire mission with five tasks in 8 minutes, while the average task completion time of the top five teams during the DRC was around 5 minutes [5], excluding any locomotion—and with fully trained operators. We feel that these results show that our operator interfaces are intuitive and that the system can be immediately applied to challenging tasks by novices. We did not notice any difference in performance between the robotics group members and non-roboticists.

While all participants managed to solve the manipulation tasks, two operators did not sort the milk cartons correctly, confusing the half-full and full cartons. Nearly all operators relied on force feedback to judge the carton weight, with some operators using visual feedback to confirm the empty carton, which responds differently to contacts with the robot end-effector.

After performing the tasks, participants were asked to answer a questionnaire. We again took inspiration from the ANA Avatar XPRIZE challenge rules, which specify a series of questions for the human operator. Figure 9 shows an aggregated view of the responses. We can immediately see that the answers were mostly positive. Although there was no baseline system available, we can gain insight by comparing the question responses to each other.

For example, nearly all operators felt completely present in the remote environment (Q4). Furthermore, the operators felt safe while controlling the system (Q6), and said it felt intuitive to operate the arms (Q8). On the downside, control of the fingers was less intuitive, probably due to the inexact mapping between human and robot fingers and the missing finger force feedback (as stated above, the more sensitive SVH hand was not available for the study). Furthermore, the operators reported that they felt less sure about the safety on the avatar side (Q7). When questioned, the participants indicated that this had mostly to do with situational awareness during locomotion. While it is easy to see the space in front of the robot, it is harder to see to the sides and impossible to see the space behind the robot. We aim to improve this by adding separate rear cameras for locomotion in the future.

B. Integrated Mission

In the integrated mission, a trained operator performed a sequence of six tasks involving locomotion, precise manipulation, and communication and interaction with a recipient. The tasks are inspired by the ANA Avatar XPRIZE challenge rules and designed to evaluate different system components (see Fig. 10). In the first task, the operator meets the recipient

- 1) Were you able to clearly see and hear what was happening in the remote space?
- 2) Did you get the necessary haptic feedback to complete the tasks?
- 3) Were you able to sense your own position and movement?
- 4) Did you feel present in the remote environment?
- 5) Was it easy and comfortable to use the Avatar System?
- 6) Did you feel safe using the Avatar System?
- 7) Did you feel the Avatar System was safe for the remote environment?
- 8) Was it intuitive to control the arms?
- 9) Was it intuitive to control the fingers?
- 10) Could you judge depth correctly?
- 11) Was the VR experience comfortable for your eyes?

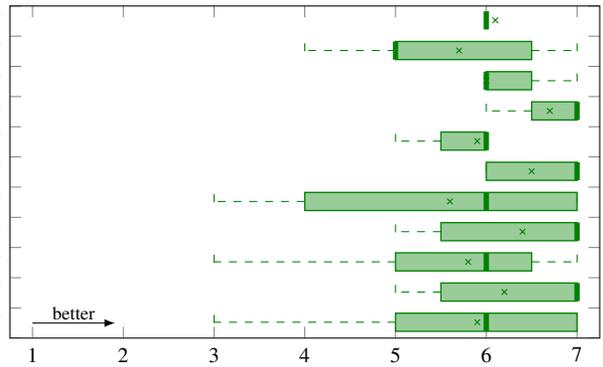


Fig. 9. Statistical results of our user questionnaire. We show the median, lower and upper quartile (includes interquartile range), lower and upper fence, outliers (marked with \bullet) as well as the average value (marked with \times), for each aspect as recorded in our questionnaire.

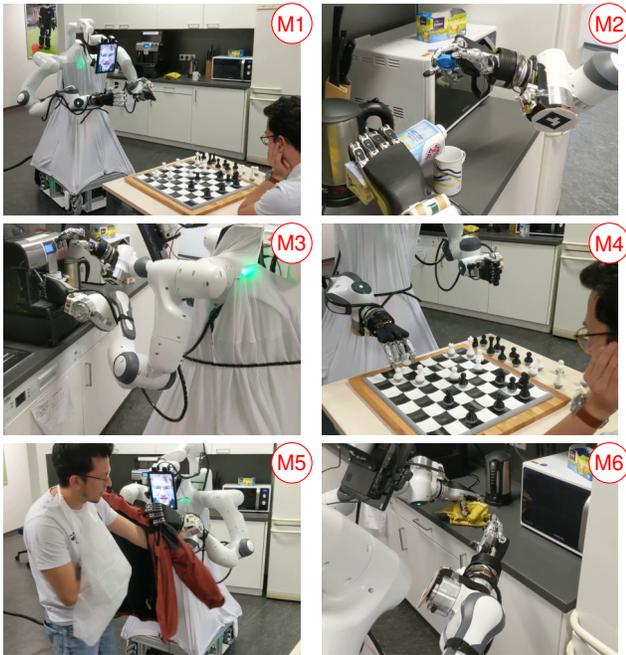


Fig. 10. Integrated mission. The operator used the avatar to: 1) Meet the recipient. 2) Get some milk from the fridge. 3) Make a coffee. 4) Play some chess with the recipient. 5) Help the recipient to put on his jacket. 6) Clean up the kitchen.

(who suffers from an arm injury) and offers help, specifically to make a coffee for the recipient. This demonstrates situational awareness (Where is the recipient?) and good verbal and non-verbal communication. The dexterous and precise manipulation capabilities are tested in the second and third tasks: The avatar gets milk from the fridge, opens the carton, pours milk in the cup and makes a coffee using the coffee machine. Especially opening and closing the fridge requires a good haptic feedback for the operator. The fourth and fifth tasks evaluate the human-robot interaction. First, the avatar and recipient play a few moves of chess. This mainly tests vision capabilities (recognizing the different chess pieces by shape) and precise manipulation capabilities (handling the rather small chess pieces). Afterwards, the operator helps the recipient to put on his jacket, which involves close contact between the human and avatar. Good haptic feedback and low impedance are key to make this experience non-threatening and comfortable for the recipient. Lastly, in the

TABLE II
INTEGRATED MISSION: TASK COMPLETION TIMES

Trial	Task completion time [s]						Total
	Intro	Milk	Coffee	Chess	Jacket	Clean	
1	27	156	80	105	49	85	502
2	44	38	-	-	-	-	-
3	26	106	88	102	58	106	486
4	45	140	133	178	69	129	694

final task the avatar cleans up the kitchen which involves putting back the milk into the fridge and wiping the kitchen counter. This task was designed to evaluate the locomotion capabilities since a lot of different locations need to be accessed. A video of this integrated mission is available⁴.

Although the operator was familiar with the system, most tasks were not trained beforehand. We did test if the robot was able to open a milk carton before the trial. Table II shows the timing results for the four attempted trials. Considering that some tasks cannot be compared by time (i.e. conversations were not predefined and thus vary in length), the results show quite similar task durations for all trials. Except for task two in Trial 2 (the avatar knocked over the milk inside the fridge), all tasks were executed successfully. The operator was able to correct smaller mistakes such as dropping the lid of the milk carton onto the counter, or losing a chess piece by an imperfectly executed grasp, without any external help. Overall, this integrated mission experiment shows not only the raw capabilities of the robotic platform, but also the reliability and intuitive control of the avatar which allow adaptation to unforeseen circumstances and tasks.

C. Lessons Learned

During both the user study and the integrated mission, we identified several strong and weak points of our system, which may not be directly reflected in the quantitative results already presented. First of all, many user study participants expressed surprise at being back in their actual environment, which they had forgotten about while controlling the Avatar. This indicates that our system has a high level of immersion.

Second, some participants were significantly smaller than we had planned for, resulting in difficulty reaching the 3D

⁴A video demonstrating the integrated mission can be found here: http://ais.uni-bonn.de/videos/IROS_2021_NimbRo_Avatar

Rudder and, more significantly, resulting in robot wrist poses closer to the body and thus more potential self-collisions. These persons complained of jerky behavior as the system displayed forces in response to position limit violations. In a future iteration, we will address this by reducing the size of the avatar wrists and by adapting the initial head pose to the operator size.

Overall, the system proved to operate reliably during all tests. We noticed that the avatar arms stopped due to exceeded torque limits some times, but the automatic recovery allowed safe continuation of operation. We will improve the control loop further to avoid triggering the stop in any case.

VI. CONCLUSION

We developed a system capable of executing complex dexterous and interaction tasks remotely. It is especially suited for typical everyday indoor environments and interactions with both remote made-for-human objects and environments as well as remote persons. The system has proven itself in a small user study with untrained operators, as well as in a longer and more complex integrated mission. The mission demonstrates a chain of both highly difficult and useful tasks as well as natural human-human interaction through the avatar. We also identified some points of possible improvement, which we will address in future work.

REFERENCES

- [1] E. Krotkov, D. Hackett, L. Jackel, M. Perschbacher, J. Pippine, J. Strauss, G. Pratt, and C. Orlowski, "The DARPA Robotics Challenge finals: Results and perspectives," *Journal of Field Robotics*, vol. 34, no. 2, pp. 229–240, 2017.
- [2] P. Oh, K. Sohn, G. Jang, Y. Jun, and B.-K. Cho, "Technical overview of team DRC-Hubo@UNLV's approach to the 2015 DARPA Robotics Challenge Finals," *Journal of Field Robotics*, vol. 34, no. 5, 2017.
- [3] A. Stentz, H. Herman, A. Kelly, E. Meyhofer, G. C. Haynes, D. Stager, B. Zajac, J. A. Bagnell, J. Brindza, C. Dellin, et al., "CHIMP, the CMU highly intelligent mobile platform," *Journal of Field Robotics*, vol. 32, no. 2, pp. 209–228, 2015.
- [4] S. Karumanchi, K. Edelberg, I. Baldwin, J. Nash, J. Reid, C. Bergh, J. Leichty, K. Carpenter, M. Shekels, M. Gildner, et al., "Team RoboSimian: Semi-autonomous mobile manipulation at the 2015 DARPA Robotics Challenge finals," *Journal of Field Robotics*, vol. 34, no. 2, pp. 305–332, 2017.
- [5] M. Schwarz, T. Rodehutsors, D. Droschel, M. Beul, M. Schreiber, N. Araslanov, I. Ivanov, C. Lenz, J. Razlaw, S. Schüller, et al., "NimbRo Rescue: Solving disaster-response tasks with the mobile manipulation robot Momaro," *Journal of Field Robotics*, vol. 34, no. 2, pp. 400–425, 2017.
- [6] T. Klamt, M. Schwarz, C. Lenz, L. Baccelliere, D. Buongiorno, T. Cichon, A. DiGuardo, D. Droschel, M. Gabardi, M. Kamedula, N. Kashiri, A. Laurenzi, D. Leonardis, L. Muratore, D. Pavlichenko, A. S. Periyasamy, D. Rodriguez, M. Solazzi, A. Frisoli, M. Gustmann, J. Roßmann, U. Süß, N. G. Tsagarakis, and S. Behnke, "Remote mobile manipulation with the Centauro robot: Full-body telepresence and autonomous operator assistance," *Journal of Field Robotics*, vol. 37, no. 5, pp. 889–919, 2020.
- [7] P. Schmaus, D. Leidner, T. Krüger, A. Schiele, B. Pleintinger, R. Bayer, and N. Y. Lii, "Preliminary insights from the METERON SUPVIS Justin space-robotics experiment," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3836–3843, 2018.
- [8] H. Martins, I. Oakley, and R. Ventura, "Design and evaluation of a head-mounted display for immersive 3D teleoperation of field robots," *Robotica*, vol. 33, no. 10, p. 2166, 2015.
- [9] D. Zhu, T. Gedeon, and K. Taylor, "Head or gaze? Controlling remote camera for hands-busy tasks in teleoperation: A comparison," in *Conference of the Computer-Human Interaction Special Interest Group of Australia*, 2010, pp. 300–303.
- [10] P. Agarwal, S. Al Moubayed, A. Alspach, J. Kim, E. J. Carter, J. F. Lehman, and K. Yamane, "Imitating human movement with teleoperated robotic head," in *International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2016, pp. 630–637.
- [11] M. Schwarz and S. Behnke, *Low-latency immersive 6D televisualization with spherical rendering*, International Conference on Humanoid Robots (Humanoids), 2021.
- [12] D. Whitney, E. Rosen, E. Phillips, G. Konidaris, and S. Tellex, "Comparing robot grasping teleoperation across desktop and virtual reality with ROS reality," in *Robotics Research*, 2020.
- [13] D. Sun, A. Kiselev, Q. Liao, T. Stoyanov, and A. Loutfi, "A new mixed-reality-based teleoperation system for telepresence and maneuverability enhancement," *IEEE Transactions on Human-Machine Systems*, vol. 50, no. 1, pp. 55–67, 2020.
- [14] T. Rodehutsors, M. Schwarz, and S. Behnke, "Intuitive bimanual telemanipulation under communication restrictions by immersive 3D visualization and motion tracking," in *International Conference on Humanoid Robots (Humanoids)*, 2015, pp. 276–283.
- [15] P. Stotko, S. Krumpfen, M. Schwarz, C. Lenz, S. Behnke, R. Klein, and M. Weinmann, "A VR system for immersive teleoperation and live exploration with a mobile robot," in *International Conference on Intelligent Robots and Systems (IROS)*, 2019.
- [16] S. Hirche and M. Buss, "Human-oriented control for haptic teleoperation," *Proceedings of the IEEE*, vol. 100, no. 3, 2012.
- [17] F. Abi-Farraj, B. Henze, A. Werner, M. Panzirsch, C. Ott, and M. A. Roa, "Humanoid teleoperation using task-relevant haptic feedback," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 5010–5017.
- [18] J. Bimbo, C. Pacchierotti, M. Agravi, N. Tsagarakis, and D. Praticchizzo, "Teleoperation in cluttered environments using wearable haptic feedback," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017, pp. 3401–3408.
- [19] R. Balachandran, J.-H. Ryu, M. Jorda, C. Ott, and A. Albu-Schaeffer, "Closing the force loop to enhance transparency in time-delayed teleoperation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 10 198–10204.
- [20] H. Wang, P. X. Liu, and S. Liu, "Adaptive neural synchronization control for bilateral teleoperation systems with time delay and backlash-like hysteresis," *IEEE Transactions on Cybernetics*, vol. 47, no. 10, pp. 3018–3026, 2017.
- [21] M. Panzirsch, H. Singh, T. Krüger, C. Ott, and A. Albu-Schäffer, "Safe interactions and kinesthetic feedback in high performance earth-to-moon teleoperation," in *IEEE Aerospace Conference*, 2020.
- [22] L. Guanyang, G. Xuda, L. Lingzhi, and W. Yan, "Haptic based teleoperation with master-slave motion mapping and haptic rendering for space exploration," *Chinese J. of Aeronautics*, vol. 32, no. 3, 2019.
- [23] K. Olszewski, J. J. Lim, S. Saito, and H. Li, "High-fidelity facial and speech animation for VR HMDs," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 6, pp. 1–14, 2016.
- [24] H. Chu, S. Ma, F. De la Torre, S. Fidler, and Y. Sheikh, "Expressive telepresence via modular codec avatars," in *European Conference on Computer Vision (ECCV)*, Springer, 2020, pp. 330–345.
- [25] S. Lombardi, J. Saragih, T. Simon, and Y. Sheikh, "Deep appearance models for face rendering," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, pp. 1–13, 2018.
- [26] C. Lenz and S. Behnke, *Bimanual haptic telemanipulation with predictive limit avoidance using off-the-shelf components*, European Conference on Mobile Robots (ECMR), 2021.
- [27] P. Holub, J. Matela, M. Pulec, and M. Šrom, "UltraGrid: Low-latency high-quality video transmissions on commodity hardware," in *20th ACM International Conference on Multimedia*, 2012, pp. 1457–1460.
- [28] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Conference of the International Speech Communication Association (INTERSPEECH)*, 2018.
- [29] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks)," in *Int. Conf. on Computer Vision (ICCV)*, 2017.
- [30] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," in *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.