Real-time Egocentric Superimposition of Operator's Own Body on Telexistence Avatar in Virtual Environment

MHD Yamen Saraiji[1] Charith Lasantha Fernando[2]

Masahiro Furukawa[3] Kouta Minamizawa[4] Susumu Tachi[5]

Graduate School of Media Design

Keio University, Japan

ABSTRACT

During teleoperation manipulation, the synchronization of the user behavior and a remote avatar is important to deliver the sensation of being in that remote place. Current telexistence technologies allow full upper body posture synchronization through multi-DOF humanoid robot structures and allow the operator to control the remote body as his own. However, it does not preserve a consistent feedback, such as the human like skin tones, operator's hand shape and the current outfit he is wearing during the operation. Thus in this paper we propose a new method that provides the operator's body shape, complexion, and light correction using real-time visuals taken from a see-through camera placed in the HMD and superimposed over robot vision. By using hand and arm trajectory from a kinematics solver, a virtual representation is used to generate masking images that isolate his local body appearance and superimpose it into the virtual environment. Local body appearance is captured via a see-through HMD. This paper describes the design and implementation of the above technique and obtained basic results.

Keywords: Telexistence, Video See-through HMD, Self-superimposing.

Index Terms: I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Virtual Reality; I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism— Color, shading, shadowing, and texture

1 INTRODUCTION

In telexistence, real-time mapping between user body and the controlled avatar robot is important to achieve a self-projection in the remote environment. We have already revealed that the behavior can achieve it despite of mapping user's appearance [1]. On the other hand, in first-person view operation reflecting body visuals and appearance on the controlled avatar would create a stronger impression to the user of being actually in the remote space rather than a surrogate robot.

[1]e-mail: yamen@tachilab.org
[2]e-mail: charith @tachilab.org
[3]e-mail: m.furukawa @tachilab.org
[4]e-mail: kouta @tachilab.org
[5]e-mail: tachi @tachilab.org

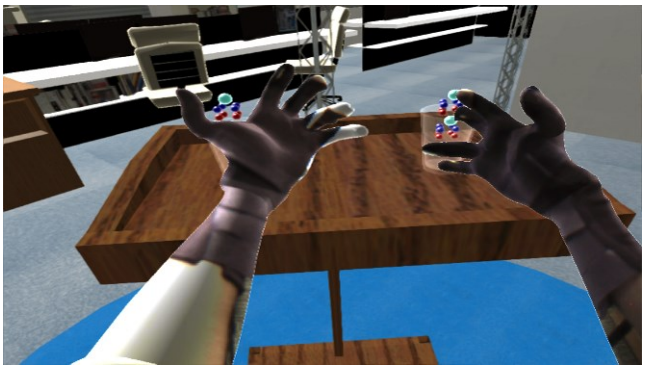

Figure 1: User's physical arms interacting in the virtual world.

In Substitutional Reality [2], user is able to experience a recorded past events in a first-person view. The immersive feedback and responsive movement of user's head would manipulate user's consciousness and makes the user believe of being in that recorded past. However, this method does not allow real-time interaction with one's own body. Furthermore, the moment when the user tries to look at his body and failing to perceive it as expected, the immersive experience of belonging to some other space would be reduced.

Regarding user's visual feedback, it was previously addressed in Telexistence for mutual interaction with other persons. Previous work in Telesar 4 [3] aims to deliver the surrounding environment to the user from the first point of view of the user via an immersive display. Also in [4][5], real-time captured images of operator's body were projected in the remote environment to provide knowledge to audience of the person who is operating.

So considering first-person view environments in which the user would have interaction using his body, visual consistency would become an important point to address to maintain the experience of being in that environment. Visual consistency requires the expected trajectory of the hands to be synchronized with the user's hand in real-time, as well as matching the relative position between the user's eyes and hands correctly. In other words, bodily kinematics and visual consistency are key points to be maintained during the operation. In "TELESAR V" [6], a 53 DOF dexterous slave robot was designed to maintain the bodily and kinematics consistency between master and slave. In regards to visual consistency, the aim is to provide a seamless operating visual experience to the user to see his own body rather than a different body – robot. Thus the problem is about how to introduce a physical object, which is operator's body, into another environment such as in teleoperation environment or into a virtual environment.

Therefore, in this paper we propose a novel method to filter out the operator's real visuals in a real-time manner and superimpose in the virtual environment as shown in Figure 1. User's own body visuals are captured using See-Through Head Mounted Display

(STHMD). Then the operator will be able to see his hands and arms in a real-time within a virtual environment. Maintain a correct perception of the body was also addressed in this paper. Furthermore, light and color correction of the two spaces will affect the body visuals, so a lighting and color correction step is also discussed in this paper.

2 Related Works

The visual consistency can be categorized into two sections depending on the point of view: third-person and first-person view. In telexistence visual consistency from audience perspective is important for mutual interaction, which is referred to as third person view. This type of visual consistency is out of our study in this paper, but was covered in [3][4][5].

In regards of first-person view, Noyes and Sheridan [7] proposed a predictive graphic displays deployed to augment visual feedback and provide a superimposed virtual representation of the slave robot. That virtual slave robot is visible even though some occluders existed between slave's cameras and hands.

In [8], a study about illusory ownership of a virtual body used a kinematic based approach to project virtual hands and arms of user's body into his virtual environment. However, this study didn't aim to include egocentric images of user's own body.

Image based approaches, such as "Chroma-Keying", allow one to isolate subjects from the surrounding background in real-time. However, using it for teleoperation applications enforces the design of the cockpit to follow its constraints, such as no additional objects that block the keying. A related method in virtual reality applications was proposed [9] to isolate the user's hands and arms from the background by detecting skin colors and tones and apply a similar approach to Chroma-Keying to define the subject's area. However, because it uses the color space only, it would lack the information of body's geometry and kinematics.

In this paper, we focus on addressing visual consistency for the operator by proposing a new method to superimpose user's body visuals on the slave's vision, and to provide an estimated 3D geometry of user's body which can be used for depth test in the virtual environments. The proposed self-projection technique can be integrated with many teleoperation applications as well as in Augmented Reality applications by using mounted camera(s) and body tracking techniques. We describe how the system is organized, and implementation details of this method.

3 System Design

To use this method of self-superimposing, we use TELESAR V platform that has a special See-Through HMD, a kinematic generator, and a 53 DOF virtual slave robot that can mimic operator movements.

The proposed method works by obtaining real-time egocentric images of operator's own body to be superimposed on the virtual environment replacing the virtual robot. An active mask is created based on the kinematic trajectory of the user arms and hands, and this was used to mask out the unwanted scenes from the remote background. In addition, the difference between lighting of the two environments was corrected.

Figure 2 shows the basic system overview that explains the flow. In this diagram, the "operator" is one who operates the system via TELESAR V cockpit and controls the robot. A real-time kinematic solver was used to map operator's posture into the slave robot as joint angles values.

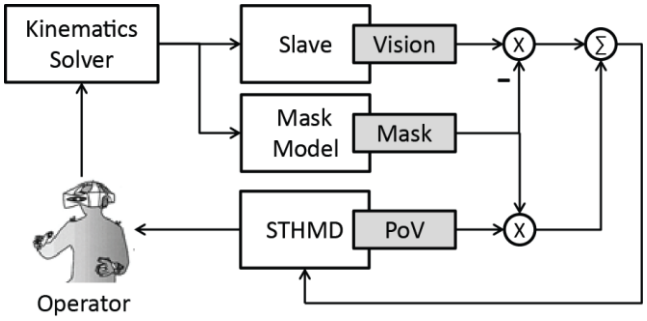

Figure 2: Superimposing flow diagram.

These values will be sent to control the virtual robot and the simulated masking model simultaneously. These are illustrated as "Slave" and "Mask Model" respectively. The masking model is a simulated model of the slave's body that follows the trajectory of the user. The masking model resembles a human model not a robot model, thus the external edges and surfaces would be soft and with no corners as in robot structures. The video feed is captured from the operator's STHMD which has a stereo camera system, and outputs Point of View (PoV) images.

For masking, we define three basic operations: " $\otimes$ " masks out parts of an image by using reference mask layer, "-" Inverts the reference mask, and "$\sum$" which combines several images together into a single one. The resulting images are provided to the user in his HMD. Further details are described in the following sections.

Ideally, in STHMD design user eyes' and cameras' point of view should be the same to obtain the same observation of the eyes from the cameras see-through, further explanation about this point will be discussed in System Implementation section.

4 System Implementation

4.1 Operator's visual acquisition

To acquire operator's appearance in real-time, we used a STHMD as shown in Figure 3, which captures stereo video images of user's body and background environment. The HMD, which we designed, provides wide-angle vision of H×V (61◦×40◦), a HD LCD 1280×800 px, and two USB cameras, Point Grey Research, Inc. Firefly (FFMV-03M2M), with almost identical field of view lens of the HMD (horizontal 62◦), and a resolution of 640x480 px @ 60 FPS. These cameras were installed on the HMD with a horizontal distance of 65mm.

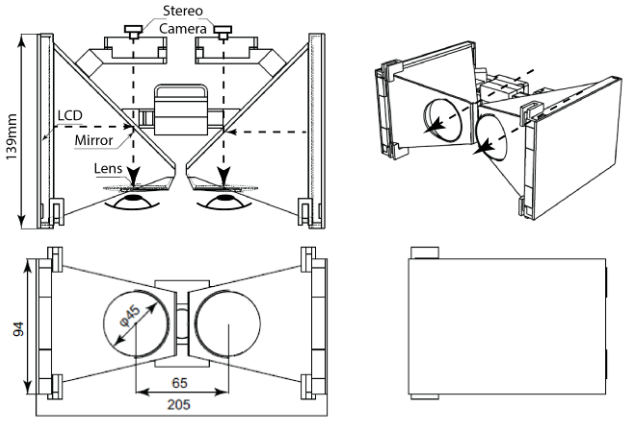

Figure 3: See-through Head Mounted Display (STHMD).

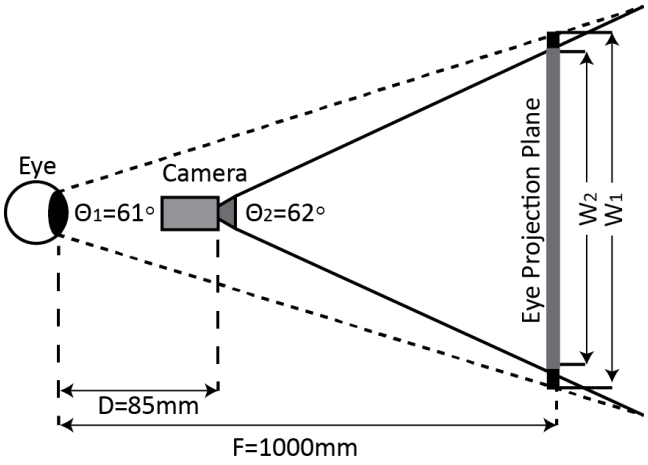

Figure 4 Camera Displacement from operator's eyes.

Because camera's aspect ratio (1:1.333) is different from LCD's aspect ratio (1:1.6), a cropping is required to match the scale difference and to avoid any scale distortion when mapping the images to the screens. Thus, by maintaining the width of the cameras at 640 px, the height can be easily calculated as in (1):

$$H = \frac{W}{R} = \frac{640}{1.6} = 400\,px \tag{1}$$

As a result, the effective cameras resolution is 640×400 px.

Regarding STHMD camera setup, Figure 4 describes the location and the parameters for eyes and cameras field of view, with the corresponding focal distance. Due to the displacement between operator's eyes and the see-through camera system (D=85mm), the optical path length between them is not the same, resulting a perspective distortion in which the user will perceive his hands closer than if seen by naked eyes. This can be solved by either changing the design of the STHMD cameras, in which optical distance of the cameras matches eyes focal length (F=1000mm), or by correcting the perspective in the software.

In this implementation, a software based correction method was used. To change camera projection plane to match eyes projection plane, a virtual viewport is created and the rendered contents are mapped into it. To calculate the cropping viewport width W_2, first we determine the eye projection plane projected width W_1 in respect of F and Θ_1, and W_2 in respect of camera distance to projection plane $(F\text{-}D)$ and its field of view angle Θ_2 as in (2):

$$\begin{aligned} W_1 &= 2F\tan\left(\frac{\Theta_1}{2}\right) = 1178mm \\ W_2 &= 2(F-D)\tan\left(\frac{\Theta_2}{2}\right) = 1099mm \end{aligned} \tag{2}$$

With that it is possible to calculate the cropping ratio as in (3) and apply it to LCD resolution to calculate the final image size (W_{VP}, H_{VP}) as in (4):

$$Ratio = \frac{W_2}{W_1} = \frac{1099}{1178} = 0.933 \tag{3}$$

$$\begin{aligned} W_{VP} &= W_{LCD}\,Ratio = 1280 \times 0.933 = 1194\,px \\ H_{VP} &= H_{LCD}\,Ratio = 800 \times 0.933 = 746\,px \end{aligned} \tag{4}$$

In Figure 5 stereo raw output images from both cameras show operator's hands after being rescaled and corrected according to the previous equations.

4.2 Kinematics solver

The virtual model and the slave representation in the remote environment need to be matched. Thus we utilize the same kinematic solver [6] to control the slave robot "TELESAR V" in our virtual environment's slave representation. The solver in this system receives five tracking points from the operator's body: shoulders, hands, and the head, and maps them into 23 DOF body. The operator wears a set of data gloves to track fingers bending, and map them into 15 DOF per hand. Kinematic solver generates joint angle values to be used in Virtual Telesar to simulate the slave avatar that represents operator's body posture, and assist in the masking algorithm.

The next step describes how this avatar will be used to generate the masking layer of operator's body.

4.3 Masking

To isolate operator's hands and arms from the background pictures, a masking image was generated using the virtual model which is simulated in Virtual Telesar Platform [10]. In the current implementation, a fixed length size 3D model representing user's arms and hands was used as a reference for masking. This model receives the same joint angles which were generated by the kinematics solver for the slave robot. Due to the fixed length size model, it will not be appropriate to be used for different users.

In the simulated environment, the virtual cameras match the position and the field of view of STHMD cameras. This would match both the projected size and position between user and virtual model.

In Figure 6, a generated black and white image representing the mask of arms and hands. White areas show the parts which remain, and black areas are the parts to erase. The process of generating the mask was done in the GPU as a post-processing phase, and was rendered into a separated GPU Render Target to be used in the fusion step.

4.4 Image Fusion

After acquiring the required information of the estimated representation of the user's body via the masking image, a masking operation is applied over PoV images. An inverted mask cuts the outer area of the slave's Vision. Figure 7 shows the resultant masked images of operator's STHMD vision.

Masking procedure was done using a post-processing shader on the GPU, the following masking function (5) was used:

$$R = M \otimes P + \overline{M} \otimes V \tag{5}$$

Here R is the result image, M is the masking image, P is the PoV image, and V is the virtual environment Vision image. The operation $\otimes$ is a masking operator between mask image and colored image.

As a result of direct image fusion, the superimposed hands have flat look and hard edges when combined with the remote environment, which is a result of different lighting conditions between the two environments. To solve this, the next step describes a proposed lighting and color correction approximation for the resultant images.

4.5 Lighting and color correction

In order to match lighting conditions of slave's environment, an estimated lighting model is applied on PoV images. Because operator's body was constructed as a 3D model, the required parameters which represent the surface of the object exist in this representation. For lighting, surface normals were mainly used to calculate shading values in respect of the source lights in the virtual environment.

Figure 5: Images obtained by See-Through HMD (Left/Right).

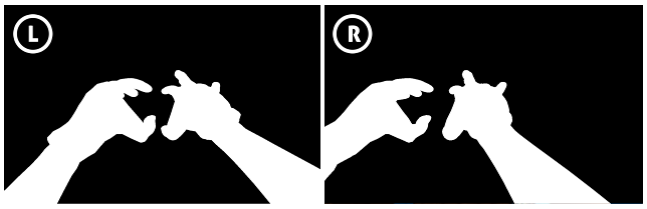

Figure 6: Mask images generated by the virtual model.

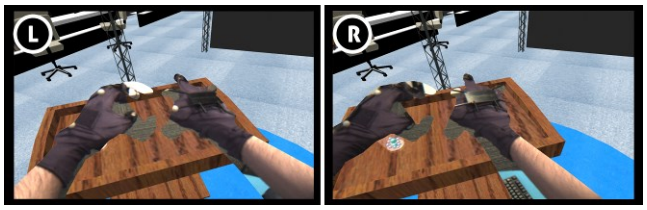

Figure 7: Result after fusing the pictures with the mask.

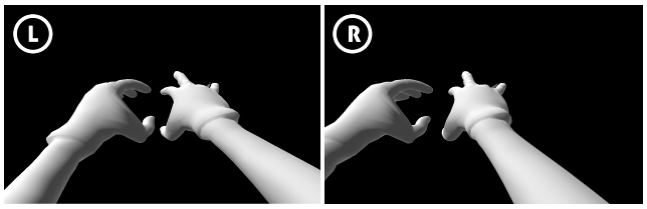

Figure 8: Estimated light correction layer.

Figure 9: Color correction and light matching.

Lighting shaders, which are a combination of Blinn-Phong and Fresnel shading equation (6) was used to generate the correction layer:

$$L_F = \max(0, 1 - (\vec{N} \cdot \vec{V}))^2 \quad (6)$$

Where L_F is the Fresnel value, N is Normal vector of the model's pixel, and V is the normalized view vector to the pixel.

These lighting equations are calculated per-pixel bases and combined into a "Light correction layer" as shown in Figure 8. And in Figure 9, the color corrected image shows a consistent lighting appearance between both environments.

Figure 10 is taken from a different frame, and shows a side by side comparison of the same scene with and without lighting correction. Noticeable flat and sharp edges appeared in the first image on the left.

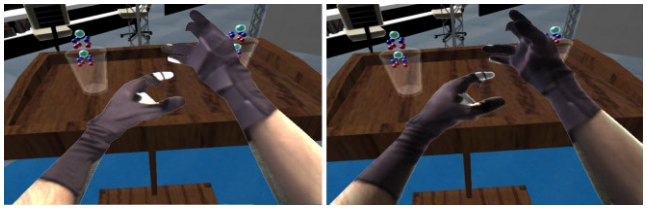

Figure 10 (Left) No light correction (Right) With light correction.

However, in the right image, the body color was affected by the same light used in the virtual environment; also the unwanted extra images that were not cropped with the mask became less visible because of the shading along edges.

Light correction model parameters (color tone, light sources) can be changed depending on the remote environment setup. Also light placement is possible in the virtual environment to replicate the actual situation of the remote environment.

4.6 System and Environment Setup

This method was carried out in real-time with a frame rate exceeding 60FPS using CPU core i7 at 2.30GHz/8GB of memory and NVidia GeForce GT 650M. Image processing and fusion were done on the GPU. User's active movement was captured by 14 Natural Point OptiTrack V100:R2 cameras. HMD has 5 reflective markers and another 4 tracking points for shoulders, and hands have 3 markers each. Finger movements are measured by 5DT Data Glove 5 Ultra which provides only 5 DOF for each hand, this yielded the lack of mapping to 15 fingers joints per hand.

The environment was our laboratory room as shown in Figure 5 and it was replicated in Virtual Telesar as shown in Figure 7 for prototyping and testing. The room matched the physical one in terms of scale and object localization in the scene.

In this experiment, the pupillary distance for the HMD was set to 65mm which is considered as an average distance for humans, although it can be changed via HMD slider.

5 Discussion

5.1 User Experience

In the experiment, the user was asked to wear a jacket for tracking the shoulder movements, and pair of gloves to track hands and fingers movements. Also head movement was tracked using the STHMD. User's posture is first calibrated for IK tracking algorithm, then the user was asked to experience the surrounding environment and the virtual objects around him. The user was able to see his superimposed hands and arms at the same relative position of his real hands to his eyes. Also in the experiment, several lights were used for testing the estimated light correction layer; user arms and hands were affected by the lights and matched the surrounding environment appearance.

5.2 Experimental Results

In the experiment, it was possible to preserve the hands where we expected it to be. Also the color tone of the imposed images was improved after applying the light correction model to the user's body images. The appropriate experimental distance for the hands in the current conditions and design of the STHMD is approximately 15 to 40 cm from operator's eyes. The camera position, which is not conjugated with the eyeballs, produced a slight distortion of the captured hands. The issue can be improved using the conjugated optical system with some external mirror.

However, a noticeable shift of the masked area (especially at the fingers) caused a slight outline of the local environment to

appear, which is mainly regarded as sensor tracking precision problem. Also, a noticeable shift of hands position was due to the fixed sized masking model which is used, a preprocessing is required to acquire operator body's dimensions and reflect them to the model to match the size and thickness of user's arms. Also the used tracking system of body, arms and hands affects masking quality. Depending on the number of DOF that can be generated from the system, it will affect the generated trajectory of the masking body. Using inverse kinematics method to generate the trajectory using hands position only showed a mismatch posture of elbows between operator and virtual model. Other tracking methods such as direct forward kinematics by tracking individual joints of the arm would provide better resolution of the tracked subject.

Regarding the performance, frame rate was limited to 60 FPS for rendering the previous scene. A slight delay of tracking the movement of the user was noticeable when he performs sudden moves. This issue is related to the tracking delay of the operator's body.

Finally, the design of the STHMD plays an important role on the final resultant image in which the camera system's virtual position must match the operator's eyes position. This is essential to remove the produced perspective distortion that results by moving the camera's projection plane from operator eyes one.

6 Conclusion

This paper describes a self-superimposing system where the operator's body appearance and behavior are superimposed into virtual environments from the first point of view. This system aims to preserve visual and appearance consistency between the user and his avatar. For visual capturing a See-Through Head Mounted Display (STHMD) was used to acquire first-person view 3D stereo images of the operator's body, and the perspective was corrected by redefining the projected area of the captured images. For the body construction, visualization and simulation, Virtual Telesar was used in the process. A 3D virtual model was used as a reference to construct a masking layer, and then applied on STHMD's output to produce isolated images of user's body only; a calibration step is necessary to be done to match that 3D model into the different users in order to generate a correct matching mask. When these images were composed over remote environment's feedback images, a difference of the color was noticeable. To solve this issue, a proposed lighting and color correction method was also presented.

Acknowledgement

This work was supported by JSPS Grant-in-Aid for Scientific Research A (23240017).

References

[1] S. Tachi, H. Arai and T. Maeda. Tele-existence master-slave system for remote manipulation. II. *Proceedings of the 29th IEEE Conference on Decision and Control*, pages 85-90 vol.1, 5-7, 1990.

[2] K. Suzuki, S. Wakisaka, and N. Fujii. Substitutional reality system: a novel experimental platform for experiencing alternative reality. *Scientific reports*, 2, 2012.

[3] S. Tachi, K. Watanabe, K. Takeshita, K. Minamizawa, T. Yoshida, and K. Sato. Mutual telexistence surrogate system: Telesar4-telexistence in real environments using autostereoscopic immersive display. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 157–162. IEEE, 2011.

[4] S. Tachi, N. Kawakami, H. Nii, K. Watanabe, and K. Minamizawa. TELEsarPHONE: Mutual telexistence master-slave communication system based on retroreflective projection technology. *SICE Journal of Control, Measurement, and System Integration*, 1(5):335–344, 2008.

[5] S. Beck, A. Kunert, A. Kulik, and B. Froehlich. Immersive Group-to-Group Telepresence. *IEEE Transactions on Visualization and Computer Graphics*, 19(4), pages 616-625, 2013.

[6] C. L. Fernando, M. Furukawa, T. Kurogi, S. Kamuro, K. Minamizawa, S. Tachi, et al. Design of TELESAR V for transferring bodily consciousness in telexistence. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5112–5118. IEEE, 2012.

[7] M. Noyes and T. Sheridan. A novel predictor for telemanipulation through a time delay. In *Proceedings of Annual Conference on Manual Control*. NASA Ames Research Center, 1984.

[8] D. Banakou, R. Groten, and M. Slater. Illusory ownership of a virtual child body causes overestimation of object sizes and implicit attitude changes. *Proceedings of the National Academy of Sciences*, 110(31), pages 12846-12851, 2013.

[9] F. Steinicke, G. Bruder, K. Rothaus, and K. Hinrichs. Poster: A virtual body for augmented virtuality by chroma-keying of egocentric videos. In *IEEE Symposium on 3D User Interfaces. 3DUI 2009*, pages 125-126. IEEE, 2009.

[10] M. Y. Saraiji, C. L. Fernando, M. Furukawa, K. Minamizawa, and S. Tachi. Virtual Telesar-designing and implementation of a modular based immersive virtual telexistence platform. In *IEEE/SICE International Symposium on System Integration (SII)*, pages 595–598. IEEE, 2012.