Tele-Robotics VR with Holographic Vision in Immersive Video

Gauthier Lafruit
Laurie Van Bogaert
gauthier.lafruit@ulb.be
laurie.van.bogaert@ulb.be
Université Libre de Bruxelles
Laboratory of Image Synthesis and Analysis
Brussels, Belgium

Jaime Sancho Aragon
jaime.sancho@upm.es
Universidad Politécnica de Madrid
Research Center on Software Technologies and Multimedia Systems (CITSEM)
Madrid, Spain

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

Grégoire Hirt
gregoire.hirt@creal.com
CREAL
Lausanne, Switzerland

Klaus H. Strobl
klaus.strobl@dlr.de
Institute of Robotics and Mechatronics
German Aerospace Center (DLR)
Wessling, Germany

Eduardo Juarez Martinez
eduardo.juarez@upm.es
Universidad Politécnica de Madrid
Research Center on Software Technologies and Multimedia Systems (CITSEM)
Madrid, Spain

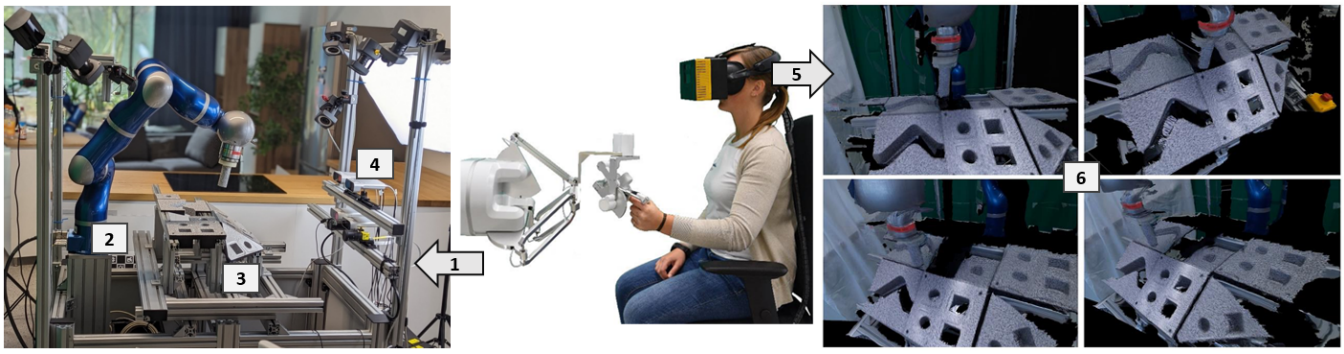

Figure 1: Tele-operating (1) the robot arm (2) in the scene (3). Various viewpoints are captured with RGBD cameras (4), out of which a virtual viewpoint is synthesized (6) for the tele-operator's head pose (5). The synthesized view is projected into a Holographic HMD, providing correct eye accommodation to fore- and background objects [15].

ABSTRACT

We present a first-of-its-kind end-to-end tele-robotic VR system where the user operates a robot arm remotely, while being virtually immersed into the scene through force feedback and holographic vision. In contrast to stereoscopic head mounted displays that only provide depth perception to the user, the holographic vision device projects a light field, additionally allowing the user to correctly accommodate his/her eyes to the perceived depth of the scene's objects. The highly improved immersive user experience results in less fatigue in the tele-operator's daily work, creating safer and/or longer working conditions. The core technology relies on recent advances in immersive video coding for audio-visual transmission developed within the MPEG standardization committee. Virtual viewpoints are synthesized for the tele-operator's viewing direction from a couple of colour and depth fixed video feeds. Besides of the display hardware and its GPU-enabled view synthesis driver, the biggest challenge hides in obtaining high-quality and reliable depth images from low-cost depth sensing devices. Specialized depth refinement tools have been developed for running in real-time at zero delay within the end-to-end tele-robotic immersive video pipeline, which must remain interactive by essence. Various modules work asynchronously and efficiently at their own pace, with the acquisition devices typically being limited to 30 frames per second (fps), while the holographic headset updates its projected light field at up to 240 fps. Such modular approach ensures high genericity over a wide range of free navigation VR/XR applications, also beyond the tele-robotic one presented in this paper.

IXR '22, October 14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9501-4/22/10.
https://doi.org/10.1145/3552483.3556461

KEYWORDS

immersive video, view synthesis, holographic HMD, foveated light field rendering, tele-robotic 6DoF-VR

ACM Reference Format:
Gauthier Lafruit, Laurie Van Bogaert, Jaime Sancho Aragon, Michael Panzirsch, Grégoire Hirt, Klaus H. Strobl, and Eduardo Juarez Martinez. 2022. Tele-Robotics VR with Holographic Vision in Immersive Video. In *Proceedings of the 1st Workshop on Interactive eXtended Reality (IXR '22), Oct. 14, 2022, Lisboa, Portugal*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3552483.3556461

1 INTRODUCTION

When talking about immersive video applications, one often thinks of 360° videos. They are immersive indeed, but only with three Degrees of Freedom (3DoF) where the user's head pose (i.e., its three rotational angles) selects the viewport that is rendered in the VR/XR headset. In applications with tactile interaction like tele-robotics, however, there is a need for 6DoF, where also user translations induce perspective changes in the VR/XR headset. 6DoF therefore often calls for explicit 3D reconstructions by e.g. photogrammetry [21], which is computationally demanding. As a matter of fact, there exists a technology in-between video (2D) and 3D that gives the illusion of 3D immersion, while relying mostly on 2D video processing. It is referred to as 2.5D and holds all the promises of 6DoF.

In 2.5D, a sparse set of cameras captures information about the 3D scene, afterwards rendering any 6DoF viewpoint to the scene for a VR/XR free navigation and/or 3D rendering experience. For instance, Google Starline [13] describes an end-to-end architecture capturing RGB colour and depth with a couple of RGBD cameras, transmitting the compressed audio-visual information towards the client-side decoder for rendering on an autostereoscopic display. Their target application is a 3D teleconference between individual participants. Taking benefit of applicative constraints like the detection of the area of interest (the participants' faces), some technological finetuning and simplifications are introduced, in order to reach high-quality 3D rendering in real-time. Per-camera (simulcast) compression reduces the bitrate for transmission over a local network, but the compression rate is still insufficient to undercut typical data rates encountered in everyday's large-scale consumer networks.

Similar needs over a wide range of immersive applications prompted the Motion Picture Experts Group (MPEG) standardization committee to develop new compression schemes, one currently known as the MPEG Immersive Video (MIV) coding technology [4] published in October 2022 as international standard. Beyond specifying a compression scheme staying well within the 15-25 Mbps range, MIV provides an architectural framework with a software toolset covering pre-processing (e.g., depth estimation) and post-processing (e.g., view synthesis) tools, which are the main subject of the present paper. In contrast to the application domain that MIV was originally developed for – immersive video for entertainment (e.g., cinematic VR) – we will rather focus on a professional tele-robotic VR application for remote manipulation of objects in hazardous environments: a robot arm is manipulated remotely, while having a view from within the scene thanks to on-site surrounding cameras.

Since hazardous working places do not provide sufficient freedom to set up anything-anywhere, the scene of interest is surrounded by a small number of fixed-position cameras, which views are fused to provide the tele-operator any viewpoint to the scene, so he/she can guide the robotic arm for optimal interaction with the remote scene. The synthesized virtual views should hence reach a quality level that effectively helps the tele-operator in conducting his/her work with as little fatigue as possible. For example, close views to the scene within the tele-operator's personal space (i.e., within arm reach) should be rendered with the highest visual comfort using a Holographic Head Mounted Display (H2MD) [3] where the user's eyes can focus onto any object in space (foreground or background) at will. It is, however, not mandatory to reconstruct the scene in 3D at the highest level of photo-realism; good visual feedback with holographic vision (i.e., foreground/background eye accommodation) and zero-delay rendering within the VR headset are the main targets here. Such approach helps in setting up a relatively simple system at reasonable cost.

The remainder of the paper describes the end-to-end architecture for our immersive tele-robotic application using the MIV framework as the starting point. In section 2, we describe the overall architecture made of an OpenXR front-end for rendering in the XR headset, as well as a couple of Dynamic Link Libraries (DLL) for capturing the scene information, each DLL being dedicated to a particular capturing modality. In section 3 we present the specific choices we made to support our immersive tele-robotic application. Finally, in section 4, we present how all these architectural choices are put together into a tele-robotic application scenario. User tests will soon be conducted to confirm that the immersive video and holographic vision approach bring an added value to the tele-robotic operator.

2 GENERIC END-TO-END ARCHITECTURE

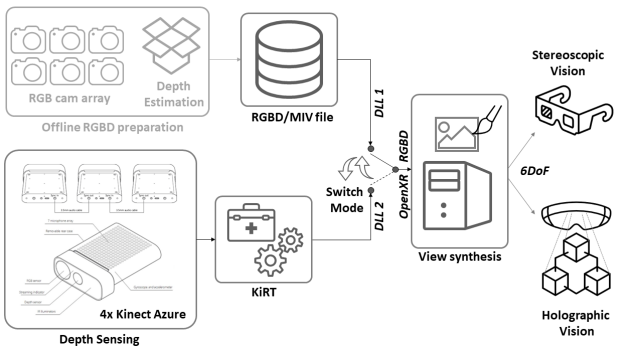

Figure 2: Overall flowgraph

Figure 2 provides a high-level view on the software pipeline developed within the context of our tele-robotic application. The module at the right serves as OpenXR driver for the Holographic HMD developed by CREAL, a company designing VR and AR light field headsets, cf. Figure 3 [3]. As shown in Figure 4 and [15], light fields ensure correct eye accommodation without the need of eye tracking and artificially blurring regions outside the object of interest, actually resolving the vergence-accommodation conflict [8] as confirmed in [16] for synthetic content with perfect depth maps. We call this "Holographic Vision". As a matter of fact, CREAL's

headset synthesizes a light field at a frame rate of up to 240 Hz, also avoiding cyber-sickness when rapidly moving the head.

At the left of Figure 2, various capturing modules (DLL1 and DLL2) can be plugged in for providing the RGBD (colour + Depth) views fed to the view synthesis module that will literally compute - i.e. synthesize - any virtual viewpoint to the scene for each new user's head pose [2, 7]. The pipeline hence supports the transformation of fixed camera feeds into intermediate virtual viewpoints that are delivered to the VR/XR headset, either in stereoscopic or holographic viewing format.

Finally, note that the capturing DLL modules are not required to run at the same speed as the VR/XR headset, since new virtual views can always be synthesized from the last DLL output. Therefore the view synthesis module and its rendering shaders run asynchronously from the DLLs.

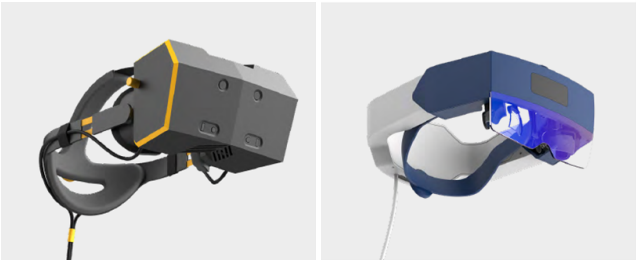

Figure 3: CREAL's Holographic Head Mounted Devices (H2MD) for VR (left) and AR (right)

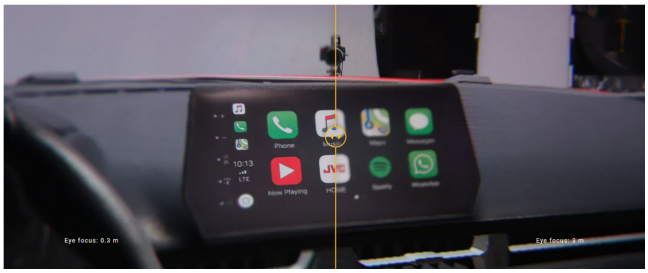

Figure 4: Light field rendering supports correct eye accommodation on foreground (left) and background objects (right)

2.1 Holographic Vision

The CREAL's headset driver follows the Varjo quad-view foveated rendering approach through OpenXR, cf. XR _VARJO _quad _views in [9]. From an input of four reference views from fixed camera positions, any virtual viewpoint for the user's head pose is synthesized with the MIV-inspired Depth Image-Based Rendering (DIBR)[1] [2, 7, 24]. Additionally, a smart optical system projects a dense light field into the user's eyes [18, 19]. This is what creates the truly Holographic Vision effect, as will be explained in Section 2.1.2.

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

2.1.1 OpenXR driver. As shown in Figure 5 for the ToysTable sequence [1] used in MIV [4], CREAL's OpenXR driver first synthesizes stereoscopic views (they can be used into any conventional, stereoscopic HMD), on top of which additional micro-parallax views are added for providing a dense light field to each eye in the holographic HMD. The driver hence supports both stereoscopic and holographic vision HMDs. Its software is partly written in Vulkan (close-to-metal OpenGL) with parametrizable view synthesis shaders [24], supporting various rendering options.

Indeed, while VR for entertainment targets photo-realism (e.g., 6DoF extension of 360° video), tele-robotic applications do not necessarily need to provide photo-realistic renderings. It may even be beneficial to avoid arbitrary inpainting options that may confuse the tele-operator's work. For instance, in the center image of Figure 6 a thin region behind the crest of the horse has not been captured by the cameras surrounding the scene, and gets hence sometimes disoccluded under a couple of virtual viewpoints. The synthesis engine will typically inpaint such regions, filling them with sometimes dis-gracious elongated triangles that flicker in time. This creates visual discomfort to the tele-operator, who might be better served by keeping the disocclusion areas black. This may be regarded as a beneficial shadow effect that increases the visual depth cue to the tele-operator. A parameter in the rendering shader of the virtual view synthesizer allows to switch between inpainting and black shadow operational modes; one being more suited for immersive video entertainment, the other for tele-operation applications as in the present paper.

2.1.2 Fovea dense light field. The Holographic Vision is obtained by projecting a dense light field at high frame rate (up to 240 Hz) into the user's eyes through the semi-transparent mirror of Figure 5 (bottom-centre). The peripheral regions in the retina being less responsive in human vision, they are refreshed at a lower frame rate (up to 90 Hz), keeping more data bandwidth for the critical visual information of the fovea. Optically merging the fovea and peripheral vision images remains a challenge, leading to the circular delineation in Figure 6 (also observe the bluish colour boundary north-west from the central wooden horse), especially in the absence of adequate colour calibration.

As shown in Figure 7, the dense light field is obtained by a time-multiplexed image projection through a light modulator and a pinlight array. These light sources are successively switched on and off, virtually projecting the light modulator's image objects at the depth corresponding to their respective disparity shifts (which includes the shift when jumping from one light source to the next). All the (time-multiplexed) light rays entering the eye pupil and hitting the retina will then trigger an eye lens response that will adjust its focal length to the perceived depth of the object of interest. For instance, for the yellow butterfly at the top-left of Figure 7, each image and pin light source shift will create a specific incidence angle of the light rays (two extreme light rays entering the pupil are shown in Figure 7) such that all light rays will focus on one point on the retina. Each point of the yellow butterfly will undergo the same phenomenon and will eventually be observed as sharp by the user. The blue tree further away in the scene will have other disparity shifts, hence the eye's focal length set for the butterfly will cause a different visual effect for the blue tree that will be projected

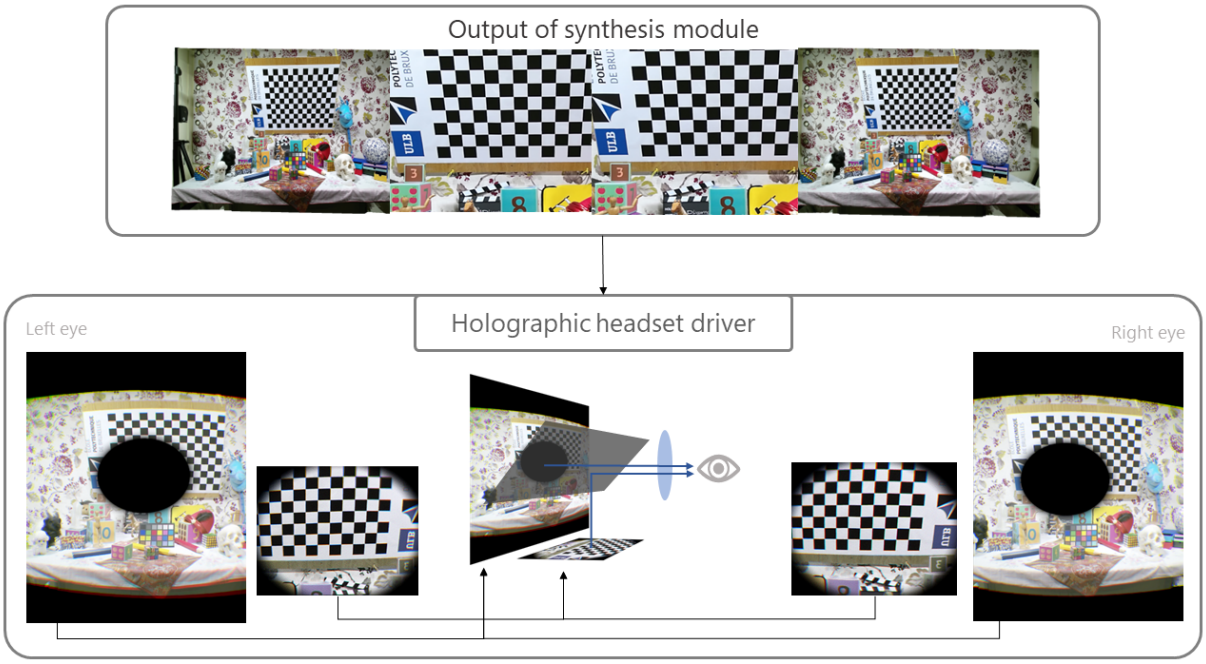

Figure 5: DIBR synthesized output for conventional stereoscopic headsets (top-two outer images) and the holographic headset (bottom) using the four images for the peripheral stereoscopic and central foveated holographic vision regions.

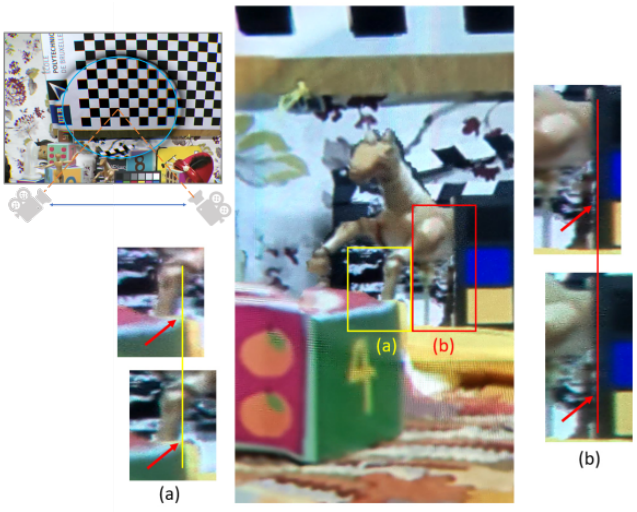

Figure 6: Micro-parallax images captured through a single headset lens (centre) for two extreme capturing positions through the H2MD fovea (top-left) with zoom-in on the front (a) and rear leg (b) of the wooden horse in the scene.

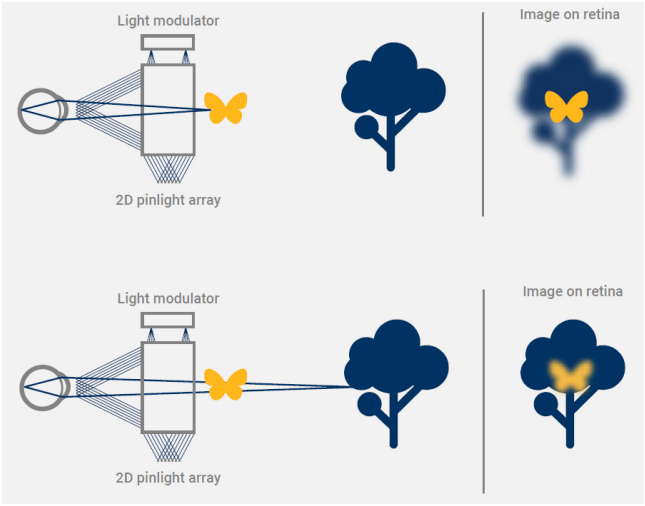

Figure 7: The pinlight array and light modulator system (left) creates a dense light field, supporting eye accommodation to front- and background objects (right) [19].

over several positions on the retina. The user then perceives an averaging effect and hence a blurred blue tree, as shown at the top-right in Figure 7. If, however, the user would like to see the blue tree as sharp, he/she will unconsciously change his/her eye's focal length to counteract the doubled/ghosted tree images that cause the blur, hence realigning the tree shifts on his/her retina to obtain a sharp blue tree, as shown at the bottom-right of Figure 7. At the very same moment, however, a discrepancy appears between the yellow butterfly disparity shifts imposed by the light modulator and pinlight array pair, on one hand, and the ones captured on the

retina, on the other hand, this time perceiving a blurring effect on the yellow butterfly.

It goes without saying that the Holographic Vision optical system explained above has been designed for the anatomy (retina curvature, light receptors density, etc) and physiology of the human eye (focal length, eyes' vergence, etc). It's therefore very difficult to capture the corresponding eye accommodation effect with a conventional or machine vision camera; they have a very different layout and focal length compared to our eyes. We could, however, capture the micro-parallax images projected from the fovea with a relatively simple camera, based on motion parallax, i.e. moving the camera over two extreme positions while looking through one lens of the headset, as shown at the top-left of Figure 6. Observing the wooden horse's front and back legs, cf. regions (a) and (b) in Figure 6, we could note different disparity shifts between the horse's legs and surrounding objects. For example, the bottom-left and bottom-right pair of images in Figure 6 show that after aligning a front object (cube or colour chart) over its two extreme viewing positions (shown by the vertical yellow and red lines), a different disparity shift with the horse's legs can be observed, cf. the difference in gaps pointed by the red arrows in the figure. This indeed suggests that various images are projected from different directions into the fovea with disparities (and hence corresponding depths) that change from one object to another in the scene. This is exactly what a light field provides and what triggers the Holographic Vision, as explained in the very beginning of this section. [23] shows that such motion parallax caused by slight head and/or saccadic eye movements re-enforces the depth cues (especially at occlusion boundaries), even in conventional stereoscopic headsets. Our experience suggests that this is even more pronounced with a holographic headset, which we hope to confirm soon with user tests in a real tele-robotic setup, cf. section 4.

2.2 DLL multi-modal capture

From the capturing point of view, we have chosen to foresee various DLLs (cf. left to center of Figure 2), each targeting a specific data acquisition modality, instead of developing a monolithic acquisition software module that would be difficult to manage. The two main acquisition DLLs foreseen so far are a multi-RGBD file reader (inspired by the MIV format, but not yet including any compression) and a multi-cam capturing pipeline delivering multi-RGBD data at its output, as further explained in the following.

2.2.1 DLL MIV file reader. The depth maps of ToysTable [1] have been estimated with the Depth Estimation Reference Software (DERS) [17] developed in the MPEG community and have therefore the same optical axis as the RGB colour camera views, which creates favourable conditions for applying MIV. This is an important difference compared to depth sensing devices where the depth map is taken from a slightly different position than the RGB image, as will be explained in subsection 3.1. On the other hand, [6] shows that with as little as four RGBD images it is possible to synthesize the dense light field of a hologram, hence likewise, it should also be possible to inject the multiview RGBD images of ToysTable into the OpenXR renderer of section 2.1 for holographic visualization in the H2MD. As already explained in section 2.1.2, Figure 6 shows the pictures taken from the ToysTable holographic light field through the H2MD lenses, clearly suggesting that a user wearing the H2MD would be able to correctly accommodate his/her eyes to the objects of interest, irrespective of their perceived position in space. Compared to Figure 4, the depth maps used for synthesizing the light field are estimated, thus imperfect, but this hardly causes any visual artefacts in the captured images of Figure 6.

For obtaining the same effect on a live event like a tele-robotic application, all the data acquisition and depth estimation/sensing actions have to be performed in real-time with virtually zero delay. This application scenario is covered in the next sections.

2.2.2 DLL multi-RGBD capture. The RGBD file-from-disk reader DLL described above implicitly assumes that the depth and colour information from each captured viewpoint have the same optical axis. This comes for free when performing depth estimation (e.g., DERS) on the RGB input views, but this requires heavy compute power which makes it practically impossible to reach real-time, zero-delay performances. For end-to-end real-time capture and rendering, it is highly recommended to use depth sensing devices, but their RGB & D images are rarely aligned by design. This often causes incorrect virtual view synthesis results in disoccluded areas (i.e., around object silhouettes) and thus must be properly handled upfront. Since there exists a variety of depth sensing devices flooding the market, it is unthinkable to have a single generic DLL that can handle all of them in the same way; it is more likely that each manufacturer comes with its own driver to overcome this RGB & D misalignment issue [26]. For reasons which will become clear in the next section, we have chosen to use four Azure Kinect devices to capture the scene from different fixed viewpoints, and we developed a DLL that outputs filtered RGBD images with perfect RGB & D alignment. The refinement tool used for this purpose is named the Kinect Refinement Tool (KiRT), shown at the bottom-left in Figure 2 and further described in more detail in subsection 3.3.

3 TELE-ROBOTIC PRACTICAL CONSIDERATIONS

3.1 Azure Kinect RGBD

The best-known low-cost depth sensing devices that are available on the market are Intel's Realsense devices (e.g., L515, D415) [10, 11] and the Azure Kinect [14]. Comparisons between these depth sensor devices in the context of view synthesis applications [25] suggest that the L515 and Azure are good candidates for further studies. The Azure Kinect has a depth sensor resolution of only 640x576 pixels (in Narrow Field-of-View mode – NFoV; in Wide Field-of-View – WFoV - it is even less), while D415 and L515 have a higher resolution of 1280x720 and 1024x768, respectively. The depth precision and angular resolution are however similar for Azure and L515 (worse for D415) over a depth range up to 1.5m, and for a depth range above 1.5m they are better for the Azure Kinect [20]. Clearly, it is misleading to only consider the spatial resolution to decide which depth sensor is most appropriate. For example, it is very convenient for depth computation that the cameras are synchronously triggered without any performance drawback, which is not feasible using Realsense depth sensing devices. After in-house tests with aforementioned devices, we decided to use the Azure Kinect in our

tele-robotic application (incidentally, Intel decided to discontinue its L515 Realsense product family [12]).

Azure Kinect has different operating modes, cf. Table 1: NFoV vs. WFoV, and binned vs. unbinned. To take benefit of the highest possible depth resolution, it was decided to use the unbinned mode, especially for sustaining high-precision DIBR-based Holographic Vision in the tele-operator's personal space, i.e., everything within arm reach in the user's visual vicinity, where the eye accommodation is continuously active. Even though the depth range of the WFoV operating mode starts at 25 cm, while it is twice as far for the NFoV, we could better overcome a depth wave-beating effect in the latter mode. The two depth modes are depicted in Figure 8 (b) and (c).

Mode	Resolution	Fov	Range
NFOV unbinned	640x576	75°x65°	0.5-3.86
NFOV binned	320x288	75°x65°	0.5-5.46
WFOV binned	512x512	120°x120	0.25-2.88
WFOV unbinned	1024x1024	120°x120°	0.25-2.21
Color(16:9)	720p to 4K	90°x59°	-
Color(4:3)	4096x3072	90°x74.3°	-
Color(4:3)	2048x1536	90°x74.3°	-

Table 1: Operating mode of kinect's depth sensors [14].

The depth gathered from the four Azure Kinects is aligned to its RGB counterpart using Microsoft API and their intrinsic and extrinsic calibration (see section 3.2). An example of this alignment is shown in Figure 8 (d). Then, the four RGBD captures are transferred to the GPU, where the rest of the process takes place. Not only the refinement tool (see section 3.3) benefits from the GPU processing of the information; the OpenXR driver (see Section 2.1.1) needed for the rendering also operates in the GPU through Vulkan.

3.2 Intrinsic and extrinsic calibration

The Kinect Azure cameras excel in the accuracy of their factory calibration w.r.t. the Realsense depth sensing devices. Nevertheless it is judicious to validate and virtually always intrinsically recalibrate the RGB cameras in order to allow sub-pixel matching accuracy of the downstream tasks and to better extrinsically calibrate the spatial layout of the depth cameras array. In addition, a more accurate model of the cameras facilitates a more efficient and accurate operation of the KiRT, see subsection 3.3.

Intrinsic camera calibration without additional lens distortion adaptation has been used to validate the factory calibration of the four Kinect Azure cameras using the camera calibration toolbox DLR CalDe and DLR CalLab [22]. For this, the only intrinsic parameters released are the focal length (common in both main directions) and the 2D position of their principal point; skew was not allowed, nor necessary. The typical residual errors in this configuration amount to 6 pixels RMS (root mean square). After additionally releasing the typical 2-parameters (3rd and 5th degrees) undistorted-to-distorted Brown-Conrady radial lens distortion model [5] to intrinsically fully calibrate the camera, the RMS of the residual errors reduces to typically 0.8 pixels. Note that this lower RMS errors (as well as the higher factory errors) are due to the fact that the calibration software allows to fill the whole image with measurement corners (partial pattern projections are possible), hence de-calibration becomes more apparent and a better calibration of the radial distortion model is achieved.

The camera calibration toolbox DLR CalDe and DLR CalLab further allows for the simultaneous intrinsic and extrinsic calibration of stereo cameras consisting of more than two cameras during the same optimization step, as opposed to the non-optimal, inconsistent approach of repeated 2-cameras stereo extrinsic optimizations. To this end, one or more images of a calibration pattern common to all cameras are taken. In the case of a set of cameras that hardly share a field of view, the calibration pattern can include several reference frames for all images to refer to a common reference frame, which will be used for the extrinsic calibration of the cameras by registration of the calibration plate model among all camera reference frames. It is worth mentioning that the intrinsic camera parameters are here fixed to the optimally estimated ones explained above. Eventually, the 6DoF rigid-body transformations between all cameras are consistently estimated and the final estimated accuracy amounts to 1.0 pixel RMS.

3.3 Kinect Refinement tool (KiRT)

Azure Kinect depth maps present artifacts in slanted borders and reflective surfaces (see Figure 8, black pixels) degrading the global quality of the depth maps and producing holes in the virtual camera synthesis. These artifacts are even more pronounced when the depth image is aligned with its RGB counterpart, as the alignment produces disoccluded areas given their different points of view.

KiRT is a GPU-accelerated tool intended for refining Azure Kinect depth maps in real-time before they are used in the rendering. It employs the information of the four Azure Kinect devices and their intrinsic/extrinsic parameters to inpaint the depth map artifacts. The algorithm is inspired on MPEG-DERS [17], focusing on its GPU real-time acceleration, and hence simplifying its most time-consuming stages. In order to achieve the real-time acceleration, the process only addresses the incorrect pixels of the depth map, extending the real depth information. An example of the depth map refinement is shown in Figure 8 (e).

4 TELE-ROBOTIC WORKLOAD EVALUATION

In the next steps of the project, the added value of using the technology of Figure 2 with the H2MD described in section 2.1 will be thoroughly assessed.

The evaluation study within the tele-robotic setup of Figure 1 is designed around two hypotheses: (a) the light-field visualisation reduces the vergence-accommodation conflict, and (b) the 6-DoF head motions enabled through view synthesis (software pan-tilt) improve the system usability. These two criteria will be assessed in real robotic working conditions with the tele-operator answering a questionnaire around the ease of use and his/her workload and associated fatigue.

We will compare the proposed technology to a reference condition in which the operator can switch his view between two fixed cameras looking onto the scene from various angles. Here, we decided not to consider a standard 2-DoF pan-tilt unit since the

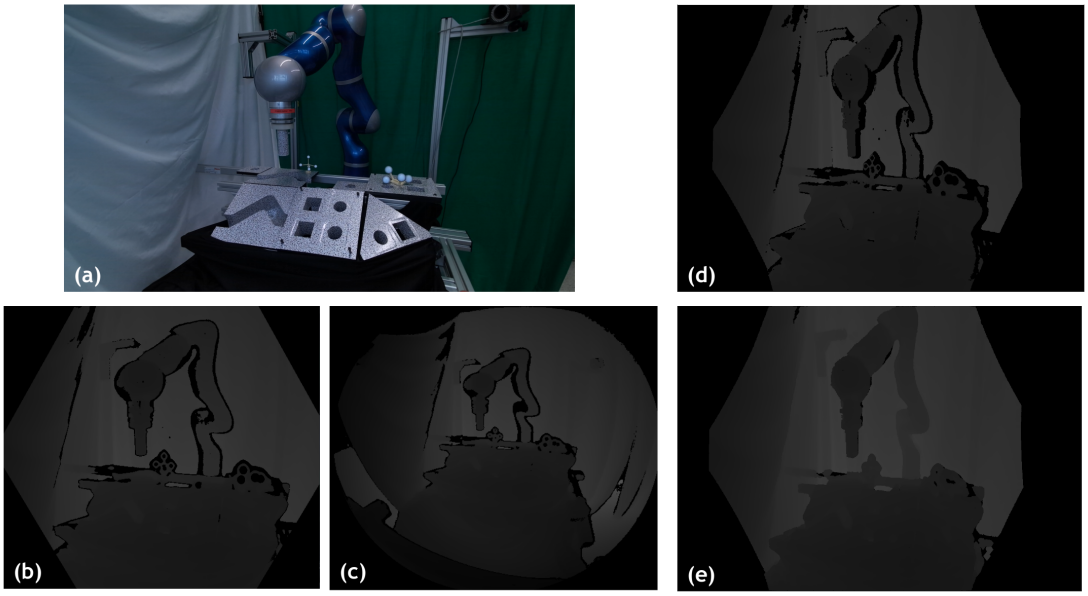

Figure 8: Example of Azure Kinect Capture: (a) color image, (b) NFOV unbinned depth, (c) WFOV binned depth, (d) NFOV depth aligned to RGB and (e) NFOV depth aligned to RGB refined using KiRT.

respective orientation changes would not improve the view onto the central scene.

The evaluation scene is presented in Figure 9: trajectories are installed at different angles (1), insertion holes have different shapes and sizes (2), and are installed at different angles (3), which the end-effector of the robot (4) has to reach and/or follow. Figure 10 shows an exemplary tele-operated trajectory (here still with direct visual feedback, which will be replaced by the H2MD) following the evaluation procedure. In this task, the contact to the contours should be avoided as well as possible.

These insertion and trajectory following tasks are classical abstract tasks allowing a detailed analysis of the performance in tele-robotic setups. The main focus in the design of the scene was the creation of different degrees of occlusion through the robotic manipulator, well mimicking a tele-robotic application.

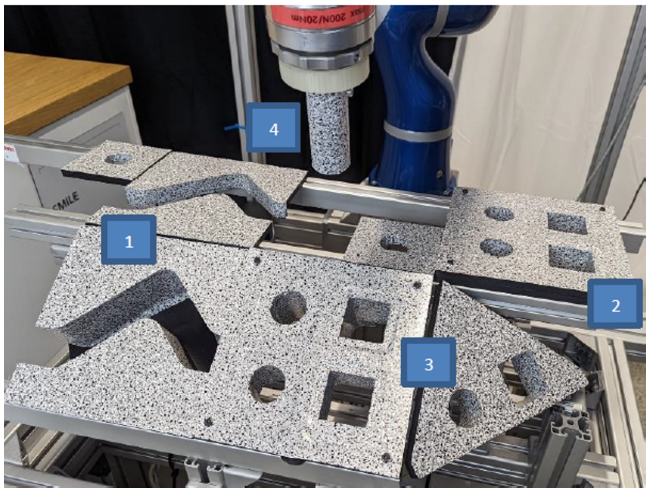

Figure 9: Tele-robotic scene of interest.

5 CONCLUSION

We have presented a first-of-its-kind tele-robotics system architecture where the scene is surrounded with a couple of RGBD depth sensing devices at fixed positions, allowing the real-time synthesis of any virtual viewpoint to the scene, following the tele-operator's head pose. Conventional stereoscopic VR/XR headsets will immerse the tele-operator into the scene with some sense of depth in a 6DoF free navigation experience. With the tele-operator wearing a holographic headset, as proposed in the paper, an additional depth cue is added, i.e. eye accommodation that provides Holographic Vision and a fully immersive experience. User tests will be conducted to confirm that the tele-operator's visual experience is truly enhanced, resulting in a decreased workload and safer working conditions.

ACKNOWLEDGMENTS

This work was supported by the HoviTron project that received funding from the European Union's Horizon 2020 research and innovation program under grant agreement N° 951989.

REFERENCES

[1] Daniele Bonatto, Sarah Fachada, and Gauthier Lafruit. 2021. ULB ToysTable. https://zenodo.org/record/5055543

[2] Daniele Bonatto, Sarah Fachada, Ségolène Rogge, Adrian Munteanu, and Gauthier Lafruit. 2021. Real-Time Depth Video-Based Rendering for 6-DoF HMD Navigation and Light Field Displays. *IEEE Access* 9 (2021), 146868–146887. https://doi.org/10.1109/ACCESS.2021.3123529

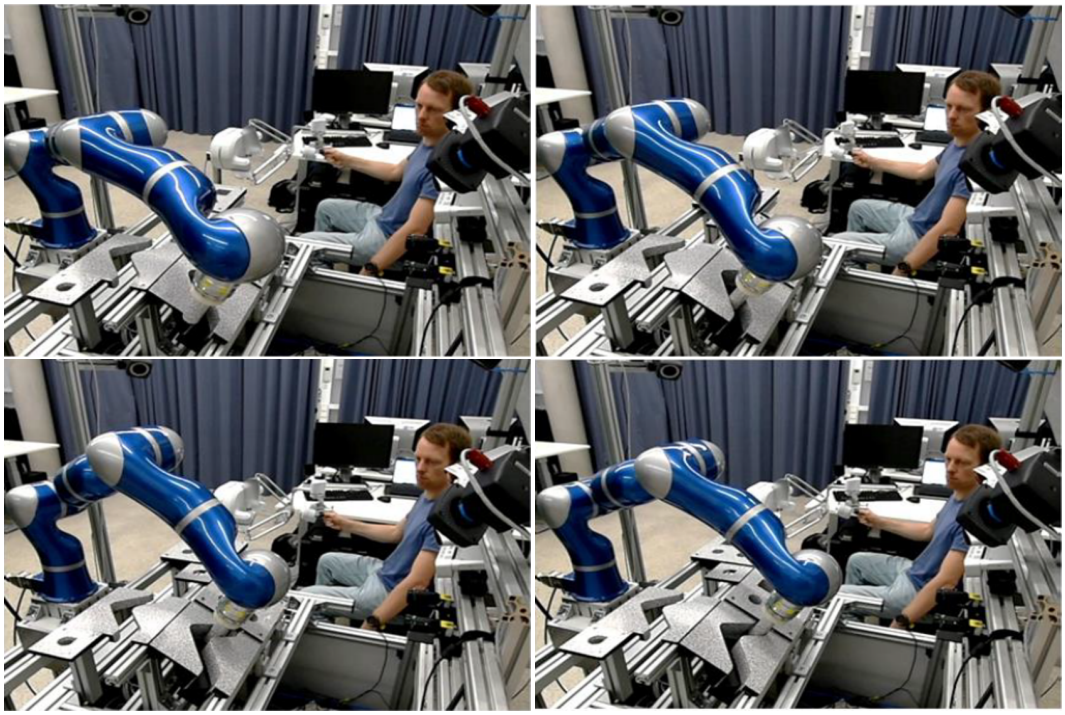

Figure 10: Test trajectory in the tele-robotic setup.

[3] Daniele Bonatto, Grégoire Hirt, Alexander Kvasov, Sarah Fachada, and Gauthier Lafruit. 2021. MPEG Immersive Video tools for Light Field Head Mounted Displays. In *2021 International Conference on Visual Communications and Image Processing (VCIP)*. 1–2. https://doi.org/10.1109/VCIP53242.2021.9675317

[4] Jill Boyce, Renaud Dore, Adrian Dziembowski, Julien Fleureau, Joel Jung, Bart Kroon, Basel Salahieh, Vinod Malamal Vadakital, and Lu Yu. 2021. MPEG Immersive Video Coding Standard. *Proc. IEEE* PP (03 2021), 1–16. https://doi.org/10.1109/JPROC.2021.3062590

[5] Duane C. Brown. 1966. Decentering Distortion of Lenses. *Photogrammetric Engineering and Remote Sensing* 32, 3 (May 1966), 444–462.

[6] Sarah Fachada, Daniele Bonatto, and Gauthier Lafruit. 2020. High Quality Holographic Stereograms Generation using four RGBD Images. In *Applied Optics*. OSA Publishing. https://doi.org/10.1364/AO.403787 ISSN: 1559-128X, 2155-3165.

[7] Sarah Fachada, Daniele Bonatto, Mehrdad Teratani, and Gauthier Lafruit. 2022. View Synthesis Tool for VR Immersive Video. In *3D Computer Graphics*, Dr. Branislav Sobota (Ed.). IntechOpen, Rijeka, Chapter 2. https://doi.org/10.5772/intechopen.102382

[8] David M. Hoffman, Ahna R. Girshick, Kurt Akeley, Banks, and Martin S. 2008. Vergence–accommodation conflicts hinder visual performance and cause visual fatigue. *Journal of vision* 8, 3 (2008), 33–33.

[9] The Khronos Group Inc. 2017-2022. The OpenXR Specification, Version 1.0.24. https://www.khronos.org/registry/OpenXR/specs/1.0/pdf/xrspec.pdf

[10] Intel. 2018. Intel Realsense D415. https://www.intelrealsense.com/depth-camera-d415/

[11] Intel. 2019. Intel Realsense L515. https://www.intelrealsense.com/lidar-camera-l515/. [Online; accessed 7-July-2022].

[12] Intel. 2021. Product Change Notification. https://qdms.intel.com/dm/i.aspx/185538FE-9AFC-4245-AD92-F9CEF3FE849F/PCN118463-00.pdf

[13] Jason Lawrence, Dan B. Goldman, Supreeth Achar, Gregory Major Blascovich, Joseph G. Desloge, Tommy Fortes, Eric M. Gomez, Sascha Häberling, Hugues Hoppe, Andy Huibers, et al. 2021. Project Starline: A high-fidelity telepresence system. (2021).

[14] Microsoft. 2021. Azure Kinect DK hardware documentation. https://docs.microsoft.com/en-us/azure/opbuildpdf/kinect-dk/toc.pdf?branch=live

[15] Michael Panzirsch and Alex Kvasov. 2020. EU-H2020 HoviTron project no. 951989, deliverable D5.1 - Proof of Concept 1. https://www.hovitron.eu/public-resources

[16] Michael Panzirsch, Bernhard Weber, Nicolai Bechtel, Nicole Grabner, and Martin Lingenauber. 2022. Light-field head-mounted displays reduce the visual effort: A user study. *Journal of the Society for Information Display* 30, 4 (2022), 319–334.

[17] Segolene Rogge, Daniele Bonatto, Jaime Sancho, Ruben Salvador, Eduardo Juarez, Adrian Munteanu, and Gauthier Lafruit. 2019. MPEG-I Depth Estimation Reference Software. In *2019 International Conference on 3D Immersion (IC3D)*. IEEE, Brussels, Belgium, 1–6. https://doi.org/10.1109/IC3D48390.2019.8975995

[18] Creal SA. 2021. WO2021090107 - Light-field virtual and mixed reality system having foveated projection. https://patentscope.wipo.int/search/en/detail.jsf?docId=WO2021090107

[19] Tomas Sluka, Alexander Kvasov, and Tomas Kubes. 2022. White Paper: Digital Light-Field. https://creal.com/app/uploads/2022/06/CREAL-White-Paper-2022.pdf

[20] Qiang Song. 2021. Performance Evaluation of High-Resolution Time-of-Flight Cameras. Master thesis School of Science and Technology, Universitat Siegen.

[21] E.K. Stathopoulou, M. Welponer, and F. Remondino. 2019. Open-source image-based 3D reconstruction pipelines: review, comparison and evaluation, Vol. Bd. XLII-2/W17. ISPRS – International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 331–338.

[22] K. H. Strobl, W. Sepp, S. Fuchs, C. Paredes, M. Smíšek, and K. Arbter. 2005. DLR CalDe and DLR CalLab. http://www.robotic.dlr.de/callab/

[23] Dmitri Tiron and Michael Langer. July 2018. Microparallax is preferred over blur as a cue to depth order at occlusion boundaries. 18th Vision Sciences Society Annual Meeting. https://doi.org/10.7490/f1000research.1115730.1

[24] Laurie Van Bogaert, Bonatto Daniele, Sarah Fachada, and Gauthier Lafruit. 2022. Novel view synthesis in embedded virtual reality devices. In *Symposium on Electronic Imaging: Engineering Reality of Virtual Reality*. Society for Imaging Science and Technology.

[25] Yupeng Xie, Sarah Fachada, André Souto, Mehrdad Teratani, and Gauthier Lafruit. 2020. EU-H2020 HoviTron project no. 951989, deliverable D1.1 - Camera Calibration and Light Field Processing. https://www.hovitron.eu/public-resources

[26] Yupeng Xie, André L. Souto, Sarah Fachada, Daniele Bonatto, Mehrdad Teratani, and Gauthier Lafruit. 2021. Performance analysis of DIBR-based view synthesis with kinect azure. In *2021 International Conference on 3D Immersion (IC3D)*. 1–6. https://doi.org/10.1109/IC3D53758.2021.9687195