Immersive VR-based Live Telepresence for Remote Collaboration and Teleoperation

MICHAEL WEINMANN[1], PATRICK STOTKO[1], STEFAN KRUMPEN[1] & REINHARD KLEIN[1]

Abstract: We present a framework for sharing immersive live telepresence experiences to groups of remote users for arbitrary-sized environments. Our framework builds upon RGB-D data capture of the local environment (by a person or robot) and involves real-time 3D reconstruction, scalable data streaming and visualization to a multitude of remote users at modest bandwidth requirements and low latency while preserving the visual quality of current real-time reconstruction approaches. We evaluate our live-telepresence system regarding its performance, its visual quality and respective user experiences, and show its beneficial use in the context of robot teleoperation, where we achieve a higher situation awareness induced by the VR-based scene exploration in comparison to purely video-based teleoperation.

1 Introduction

Sharing immersive live telepresence experiences has received increasing attention in recent years with applications in entertainment, teleconferencing, remote collaboration, site exploration, education, robotics, cultural heritage and medical rehabilitation. The impression of telepresence - defined as the subjective experience of being in an environment that may differ from the user's actual local physical surrounding - heavily relies on the ability of users to interactively explore the respective scene while avoiding any kind of discomfort such as motion sickness. This requires accurate scene visualization at high framerates and low latencies.

Purely video-based solutions, i.e. in terms of 360 degree videos, strongly restrict the scene exploration to views in the vicinity of the camera poses during scene capture and do not meet the strong real-time constraints required for live scenarios (LUO et al. 2018; SERRANO et al. 2019). Therefore, sharing live-captured scenes as needed for teleconferencing (ORTS-ESCOLANO et al. 2016) or remote collaboration (VASUDEVAN et al. 2011; FAIRCHILD et al. 2016; MOSSEL & KRÖTER 2016; STOTKO et al. 2019a; STOTKO et al. 2019b) typically involves data capture, 3D reconstruction, data streaming and visualization. All these steps have to be achieved within strong real-time constraints while preserving the visual quality of the scene, considering typically available network bandwidth and client-side hardware. Whereas capturing a small fixed-sized region of interest as typical for teleconferencing (ORTS-ESCOLANO et al. 2016) and small-scale remote collaboration (VASUDEVAN et al. 2011; FAIRCHILD et al. 2016), i.e. within a few square meters in single rooms, allows the exploitation of expensive well-calibrated setups with statically placed cameras, the efficient capturing, data representation and streaming become significantly more challenging when capturing scenes of arbitrary size with a moving camera.

[1] University of Bonn, Institute of Computer Science II – Visual Computing, Endenicher Allee 19A, D-53115 Bonn, E-Mail: [mw, stotko, krumpen, rk]@cs.uni-bonn.de

This paper addresses the task of sharing immersive live telepresence experiences for arbitrary-sized environments with groups of remote users based on efficient large-scale real-time 3D reconstruction, data streaming and visualization (see Figure 1). By design, the proposed system shifts the hardware requirements from the involved users towards the cloud. In addition, the system runs at low/modest bandwidth requirements with low latency and can handle network interruptions. As a result, this approach allows a re-thinking of well-established exploration processes towards (1) VR-based remote consulting/collaboration, where experts may save the travel time and costs, (2) VR-based remote exploration of contaminated scenes and thereby avoiding the exposure of people to danger, (3) VR-based education scenarios, where expensive excursions may be replaced with immersive group-scale telepresence in the respective environments or (4) VR-based robot teleoperation where remotely-located humans benefit from a higher situation awareness within the respective local environment of the robot. Example applications shown in this paper include remote collaboration as well as robot teleoperation.

2 Methodology

As illustrated in Fig. 1, the presented framework for immersive group-scale telepresence in live-captured scenes involves

- a **local user (or robot)** capturing the local environment based on RGB-D sensors as present in mobile phones or the Microsoft Kinect,
- a cloud-based **real-time reconstruction framework** for on-the-fly/online scene capture and camera localization based on volumetric fusion that reconstructs a dense 3D model in real-time and streams both the camera pose and the reconstructed scene parts to the server component,
- a cloud **server** to manage the global scene model and to control the data transmission according to the requests by remotely connected users, which requires reliable, efficient data representation and communication/management under concurrent memory operations while also handling possibly occurring network instabilities (for robot teleoperation, a further management of the coupling between user interactions and respective robot actions has to be taken into account),
- **visualization components** that update the locally generated meshes for the individual remote users according to the already transmitted data and provide the functionalities of interactive scene exploration with features such as measuring distances, marking objects and the collaboration with other connected experts (e.g. via VoIP), and
- remote users that may independently and immersively explore and interact with the live-captured scene while communicating with the person (or robot) capturing the scene.

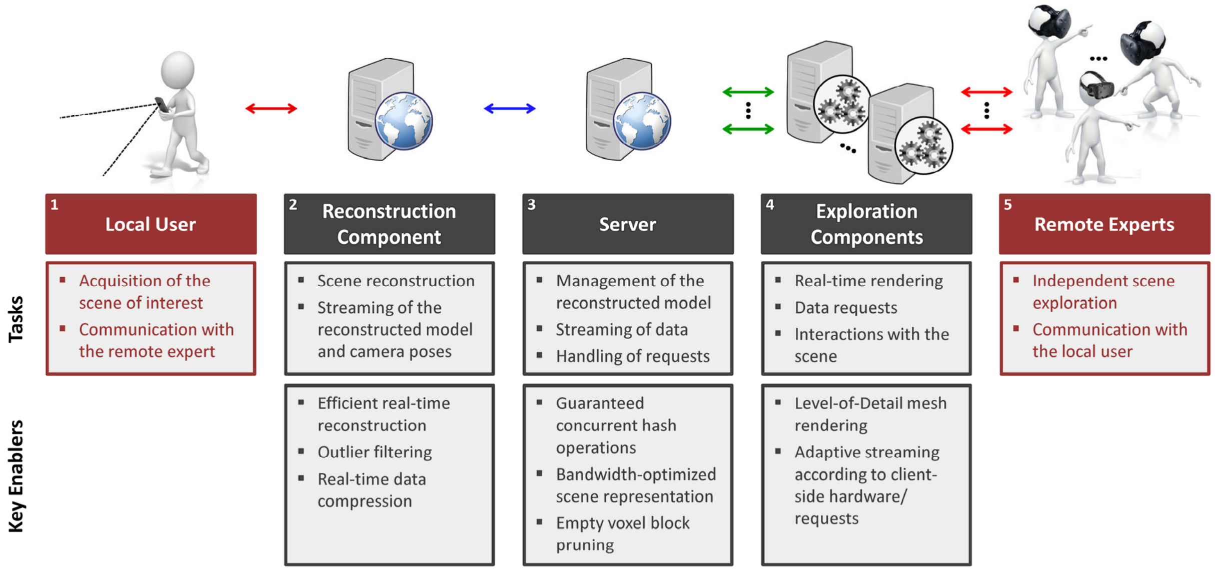
1 Local User
2 Reconstruction Component
3 Server
4 Exploration Components
5 Remote Experts
Tasks
Key Enablers
Acquisition of the scene of interest
Communication with the remote expert
Scene reconstruction
Streaming of the reconstructed model and camera poses
Management of the reconstructed model
Streaming of data
Handling of requests
Real-time rendering
Data requests
Interactions with the scene
Independent scene exploration
Communication with the local user
Efficient real-time reconstruction
Outlier filtering
Real-time data compression
Guaranteed concurrent hash operations
Bandwidth-optimized scene representation
Empty voxel block pruning
Level-of-Detail mesh rendering
Adaptive streaming according to client-side hardware/ requests

Fig. 1: Overview of the framework, its components and the challenges to be solved for sharing immersive live telepresence experiences for arbitrary-sized environments with groups of remote users. Images are partially provided by PresenterMedia.com

In the following, we describe the major steps towards an efficient live-telepresence system and its extension for robot teleoperation.

2.1 Efficient Immersive Live-Telepresence System for Remote Collaboration

The key to an immersive live-telepresence experience is given by efficient data capture, data management and visualization as discussed in detail in our respective investigations (STOTKO et al. 2019a; STOTKO et al. 2019b; STOTKO et al. 2019c). In the following, we briefly summarize respective optimizations separately for the respective components:

- **Reconstruction component:** We use an RGB-D-based real-time reconstruction framework (NIEßNER et al. 2013). In a pre-processing step, we filter potentially unreliable data from the RGB-D frames provided by the camera to reduce the data amount and increase the robustness of the RGB-D data used during reconstruction. For this purpose, we discard samples at depth discontinuities and noisy regions within the depth data. Furthermore, we apply a filter to reduce the number of unnecessary allocations during the volumetric data fusion which leads to a significant speed-up of the reconstruction and also reduces the amount of data that are later queued for the streaming to the server. Finally, we also discard voxels that contain only very few and possibly unreliable observations.
- **Server:** To efficiently maintain the global 3D scene model and control the streaming to connected remote users' exploration components, we assign each exploration component a stream set based on GPU hash data structures. Here, our data structure is designed to support guaranteed thread-leveled concurrent insertion, retrieval and removal of entries that are the prerequisite of efficient and reliable data updates. Furthermore, we convert the scene model to a highly bandwidth-efficient representation based on Marching Cubes (MC) indices (LORENSEN et al. 1987) to support pruning empty voxel blocks and enable efficient streaming to a large number of connected exploration components.

- **Visualization component:** To achieve efficient rendering, we group the received scene data into larger mesh blocks for which we asynchronously generate triangle mesh data as well as three levels of detail.

2.2 VR-based Immersive Teleoperation and Live Exploration with a Mobile Robot

The aforementioned live-telepresence system can also be used for robot teleoperation as shown by Stotko et al. (2019d). We equipped a remotely operated ground robot (Schwarz et al. 2018) with a Microsoft Kinect v2 RGB-D camera that can be moved via the robot's arm. As the poses of the robot's individual components cannot be computed using only the camera pose, the robot client sends them to the live-telepresence framework where the current robot state is directly visualized (see Fig. 2). To allow an adequate immersive teleoperation experience, we additionally remove potential jittering effects due to imperfect camera poses by applying a temporal low-pass filter on the robot's base pose.

In comparison to other VR-based teleoperation approaches, our technique overcomes the limitations regarding a sparse scene reconstruction in terms of a sparse point cloud (Bruder et al. 2014) and the high latency induced by video-based VR approaches (Kurup & Liu 2016).

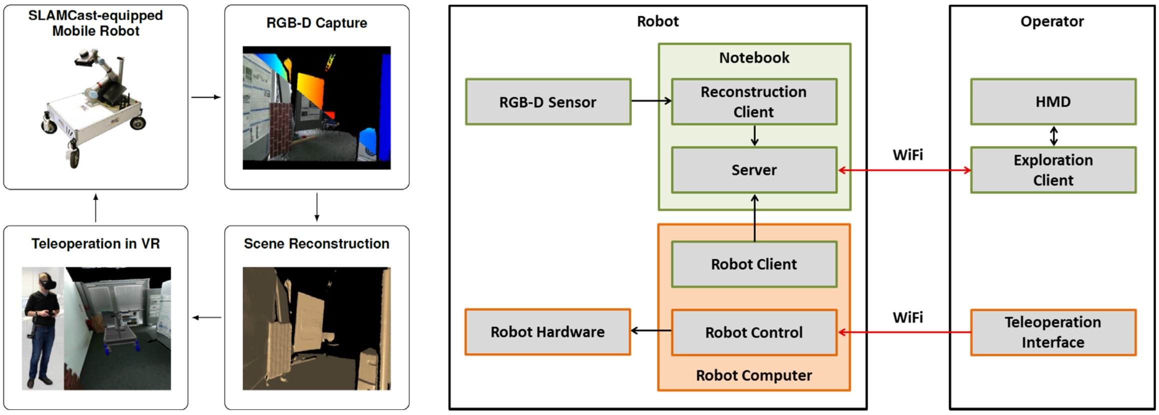

Fig. 2: (left) High-level overview of our VR-based robot teleoperation and scene exploration system: An operator controls a robot using a live-captured and reconstructed 3D model of the robot's local environment. (right) Detailed implementation of the teleoperation system: The system allows the operator to enter the reconstructed scene without being limited to the specific view of the camera on the robot. The teleoperation of the robot is based on standard teleoperation devices (e.g. a gamepad). Components of the live-telepresence system are marked in green and orange components belong to the robotic system

3 Results

In this section, we provide an evaluation of the basic live-telepresence system as well as an evaluation of its use for robot teleoperation.

3.1 Evaluation of Live-Telepresence System

In the following, we evaluate our live-telepresence system regarding its performance, its visual quality and a study of respective user experiences.

Performance Analysis We measured the bandwidth for streaming the data to the exploration components as well as the streaming latency and component scalability using several datasets captured with off-the-shelf RGB-D cameras. Even with low streaming rates of 512 blocks/request at a request rate of 12Hz, a low latency was observed. Furthermore, the required mean (and maximum) bandwidth was around 13MBit/s (and 25 MBit/s respectively) while the server was able to handle 24 components simultaneously using standard consumer hardware without introducing further latency (see Fig. 3).

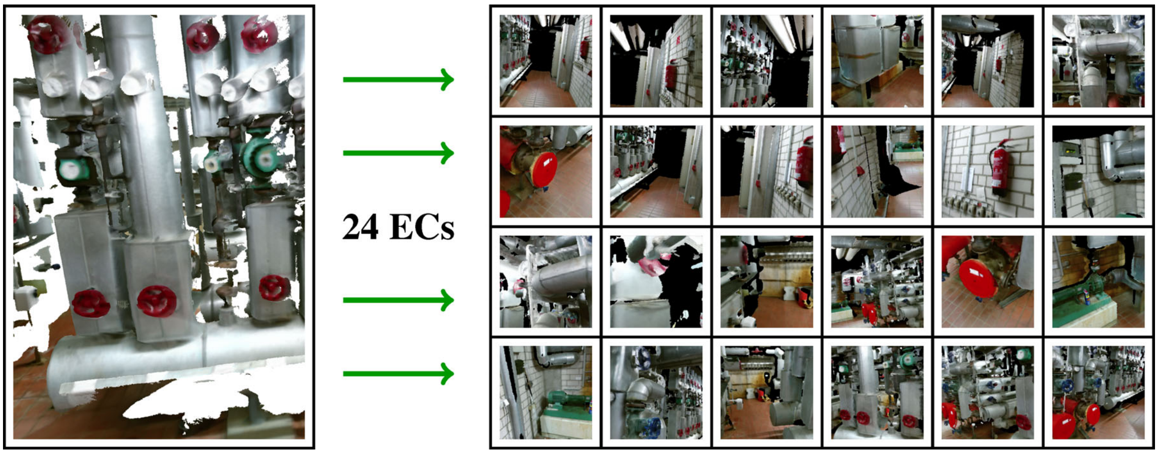

Fig. 3: Illustration of our scalable telepresence system which enables sharing live telepresence experiences of high-quality scene reconstructions to more than 24 exploration components

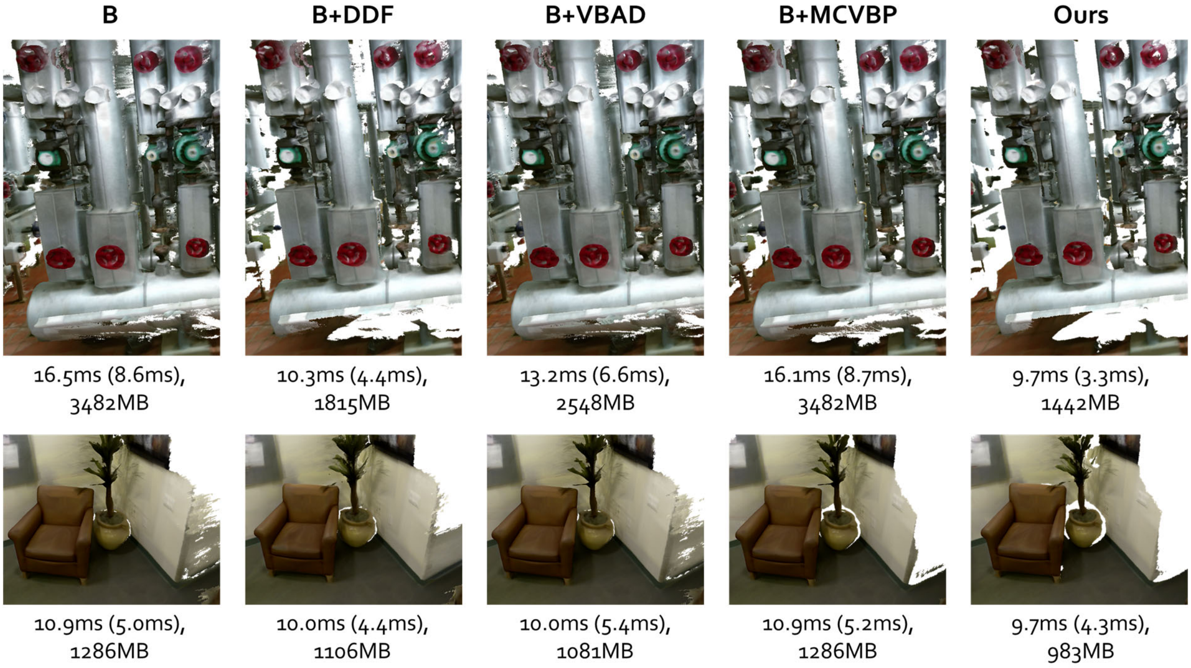

Fig. 4: Comparison of visual quality, mean runtime (and standard deviation) as well as memory requirements for different system variants. For more detailed descriptions, we refer to the discussion by STOTKO et al. (2019b)

Visual Quality To ensure a high degree of immersion, we also evaluated the visual quality of the reconstructed 3D models (see Fig. 4). As a result of filtering outliers in the input and model data, our reconstruction component generates smaller models with higher quality in comparison to standard 3D reconstruction which also benefits streaming bandwidth and scalability (see Figure 2). This may be explored for guiding the user to perform a more thorough scene acquisition resulting in more complete 3D models with higher accuracy.

Evaluation of User Experience To evaluate the practicality of our framework for telepresence in live-captured scenes, we immersed 18 subjects (mean age of 28.0 years) into an on-the-fly captured scene based on standard VR devices, where the current local scene model was visualized according to the participant's current pose. This way, the users were able to interactively inspect the scene independent from the camera's current pose. The user ratings (see Fig. 5) indicate that the users experienced a high degree of situation awareness and self-localization in the simultaneously captured scene and could easily assess the terrain for navigation purposes. Furthermore, they reported the controls for scene interaction (i.e. teleporting, performing distance measurements, etc.) to be intuitive. The ratings regarding the resolution of the reconstructed model and the speed of movements were slightly lower. Future improvements regarding texture resolution and regarding the overall model quality as well as the increasing availability of affordable VR devices, and with it the higher familiarity to the control mechanisms, may address these aspects.

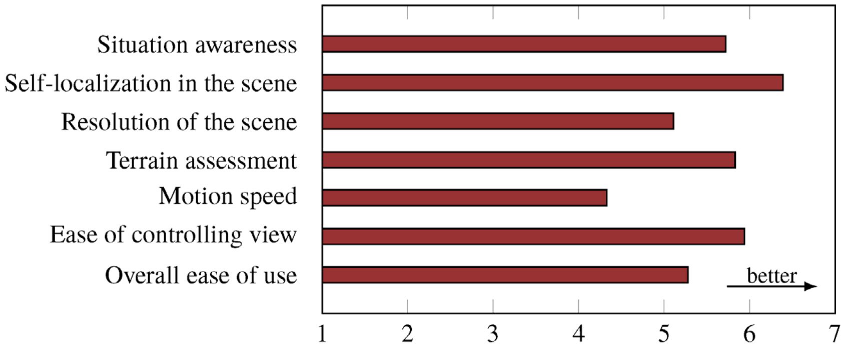

Fig. 5: Assessment of user experience based on mean ratings on a 7-point Likert scale

3.2 Evaluation of VR-based Robot Teleoperation

In the following, we evaluate our VR-based robot teleoperation system regarding its usability in comparison to video-based robot teleoperation as well as the quality of remote users' interactions with the scene.

Evaluation of User Experience For the assessment of the benefits of our immersive VR-based teleoperation system in comparison to video-based approaches, we conducted a user study. The participants (20 participants with a mean age of 29.25 years) were asked to maneuver a robot through a course with challenges of different difficulties based on two different stimuli. During the VR mode, the participants were immersed into the local environment of the robot based on standard VR devices and were able to follow the robot through the course. Here, the HMD-based scene inspection corresponding to the already reconstructed scene parts is not limited to the current view of the camera. Thus, we expect a higher degree of immersion and situation awareness as users obtain a better impression regarding 3D distances in the scene as well as occurring obstacles. During the video mode, the robot teleoperation was purely based on video data corresponding to the current view of the camera moved by the robot. As a consequence, there is no possibility of observing information outside the current camera view which we expect to result in a lower degree of situation awareness due to the reduced perception quality of distances between objects in the scene as well as occurring obstacles. To validate our expectations, we collected ratings of a variety of relevant aspects based on a 7-point Likert (see Figure 6), the number of collisions with the environment made in the different modes and the total execution time required to navigate the robot from the starting point to the target location. The ratings provided by the users indicate that our VR system is beneficial for self-localization in the scene, maneuvering around narrow corners, avoiding obstacles, assessing terrain for navigability as well as regarding the ease of controlling the view. For these aspects, the boxes defined by the medians and interquartile ranges do not overlap indicating a significant difference in favor of the VR-based teleoperation. In addition, the VR mode was judged to be well-suited for teleoperation and to allow moving the robot to target positions in an easier manner. These statistical results support the higher degree of situation awareness for the VR-based teleoperation.

However, it seems as if the higher degree of immersion for VR-based teleoperation results in a more time-consuming inspection that limits the speed of robot motion. The video-based mode allowed a faster completion of the course due to the lacking possibility to inspect the situation, e.g. by walking around the robot, at the cost of a higher number of collisions. Furthermore, the video-based mode was rated slightly better regarding the perceived latency as well as the resolution. The latter aspect may be addressed by future work on enhancing the texture resolution.

Interaction of Remote Users with the Scene In the context of disaster management, typical user interactions with the scene for exploring contaminated sites and establishing evacuation plans include the measurement of distances such as door widths to determine whether a different robot or the required equipment would fit through narrow spaces. For this reason, we implemented a tool for measuring 3D distances based on the controllers of the involved HMD device, where the resulting measurement accuracy is determined by the voxel resolution specified for the scene representation (see Fig. 6). We observed errors of up to 1 cm which we expect to be sufficient for disaster management scenarios. Furthermore, our system allows the user to label areas as interesting, suspicious or incomplete. This information is integrated into the overall map and may be used by the capturing robot to return to a particular position as well as to complete or to refine the scan.

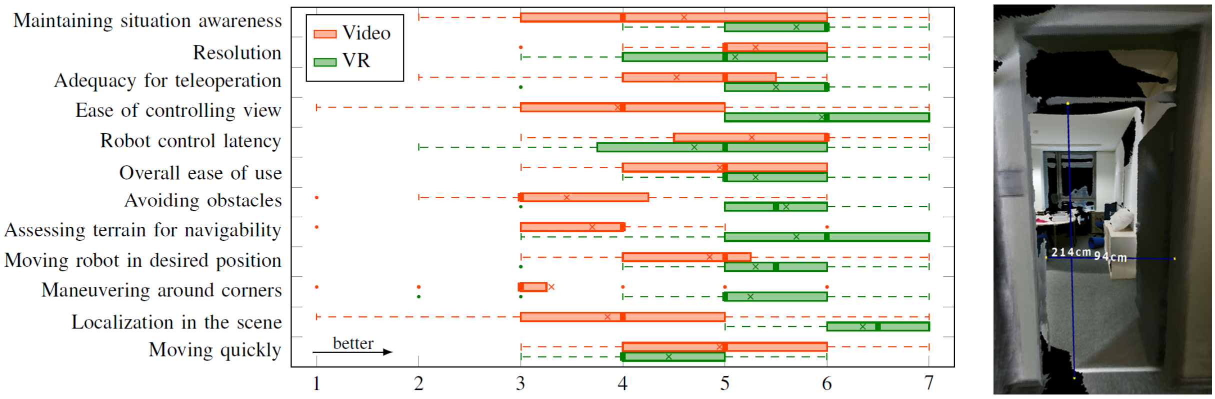

Fig. 6: (left) Statistical results of a user study, i.e. median, lower and upper quartile (includes interquartile range), lower and upper fence, outliers (marked with a dot) as well as the average value (marked with an x), for different aspects. Our VR-based system achieved higher ratings on the 7-point Likert scale than a video-based approach for most of the aspects. (right) Example of interactively taken measurements (94 cm and 214 cm) of the width and height of a door taken to guide the further management process. The ground truth values are 95 cm and 215 cm

4 Conclusions

We summarized our investigations towards scalable sharing of immersive live-telepresence experiences beyond room-scale based on efficient real-time 3D reconstruction and streaming. For this purpose, we (1) integrated several filters to discard unreliable data and thereby improve model compactness and robustness, (2) developed a bandwidth-optimized scene representation for the global scene model on the server, (3) introduced a novel hashing approach to allow concurrent thread-leveled insertion, removal and retrieval while preserving key uniqueness, and (4) use level-of-detail rendering of the mesh data for the exploration clients. While this system is suitable for remote collaboration purposes, we also described its extension towards VR-based robot teleoperation where we showed the benefits over purely 2D video-based teleoperation. Future challenges include the further improvement regarding model quality, streaming efficiency and collaborative capturing.

5 Acknowledgements

We thank Max Schwarz, Christian Lenz and Sven Behnke for their support in integrating our telepresence system on a robot platform for teleoperation.

6 References

Bruder, G., Steinicke, F. & Nüchter, A., 2014: Poster: Immersive point cloud virtual environments. IEEE Symposium on 3D User Interfaces, 161-162.

Kurup, P. & Liu, K., 2016: Telepresence Robot with Autonomous Navigation and Virtual Reality: Demo Abstract, ACM Conference on Embedded Network Sensor Systems, 316-317.

FAIRCHILD, A.J., CHAMPION, S.P., GARCÍA, A.S., WOLFF, R., FERNANDO, T. & ROBERTS, D.J., 2016: A Mixed Reality Telepresence System for Collaborative Space Operation. IEEE Transactions on Circuits and Systems for Video Technology **27**(4), 814-827.

LORENSEN, W.E. & CLINE, H.E., 1987: Marching Cubes: A High Resolution 3D Surface Construction Algorithm. 14th Annual Conference on Computer Graphics and Interactive Techniques, 163-169.

LUO, B., XU, F., RICHARDT, C. & YONG, J., 2018: Parallax360: Stereoscopic 360 Scene Representation for Head-Motion Parallax. IEEE Trans. on Visualization and Computer Graphics **24**(4), 1545-1553.

MOSSEL, A. & KRÖTER, M., 2016: Streaming and Exploration of Dynamically Changing Dense 3D Reconstructions in Immersive Virtual Reality. IEEE International Symposium on Mixed and Augmented Reality, 43-48.

NIEßNER, M., ZOLLHÖFER, M., IZADI, S. & STAMMINGER, M., 2013: Real-time 3D Reconstruction at Scale Using Voxel Hashing. ACM Transactions on Graphics **32**(6), 169:1-169:11.

ORTS-ESCOLANO, S., RHEMANN, C., FANELLO, S., CHANG, W., KOWDLE, A., DEGTYAREV, Y., KIM, D., DAVIDSON, P., KHAMIS, S., DOU, M., TANKOVICH, V., LOOP, C., CAI, Q., CHOU, P., MENNICKEN, S., VALENTIN, J. PRADEEP, V., WANG, S., KANG, S.B., KOHLI, P., LUTCHYN, Y., KESKIN, C. & IZADI, S., 2016: Holoportation: Virtual 3D Teleportation in Real-time. Annual Symposium on User Interface Software and Technology, 741-754.

SCHWARZ, M., DROESCHEL, D., LENZ, C., PERIYASAMY, A.S., PUANG, E.Y., RAZLAW, J., RODRIGUEZ, D., SCHÜLLER, S., SCHREIBER, M. & BEHNKE, S., 2018: Team NimbRo at MBZIRC 2017: Autonomous valve stem turning using a wrench. Journal of Field Robotics **36**(1), 170-182.

SERRANO, A., KIM, I., CHEN, Z., DIVERDI, S., GUTIERREZ, D., HERTZMANN, A. & MASIA, B., 2019: Motion parallax for 360 RGBD video. IEEE Transactions on Visualization and Computer Graphics **25**(5), 1817-1827.

STOTKO, P., KRUMPEN, S., HULLIN, M.B., WEINMANN, M. & KLEIN, R., 2019a: SLAMCast: Large-Scale, Real-Time 3D Reconstruction and Streaming for Immersive Multi-Client Live Telepresence. IEEE Transactions on Visualization and Computer Graphics **25**(5), 2102-2112.

STOTKO, P., KRUMPEN, S., WEINMANN, M. & KLEIN, R., 2019b: Efficient 3D Reconstruction and Streaming for Group-Scale Multi-Client Live Telepresence. In Proceedings of IEEE International Symposium on Mixed and Augmented Reality, 19-25.

STOTKO, P., KRUMPEN, S., KLEIN, R., & WEINMANN, M., 2019c: Towards Scalable Sharing of Immersive Live Telepresence Experiences Beyond Room-scale based on Efficient Realtime 3D Reconstruction and Streaming. CVPR Workshop on Computer Vision for Augmented and Virtual Reality.

STOTKO, P., KRUMPEN, S. SCHWARZ, M., LENZ, C., BEHNKE, S., KLEIN, R., & WEINMANN, M., 2019d: A VR System for Immersive Teleoperation and Live Exploration with a Mobile Robot. IEEE/RSJ International Conference on Intelligent Robots and Systems, arXiv preprint arXiv:1908.02949.

VASUDEVAN, R., KURILLO, G., LOBATON, E., BERNARDIN, T., KREYLOS, O., BAJCSY, R. & NAHRSTEDT, K., 2011: High-Quality Visualization for Geographically Distributed 3-D Teleimmersive Applications. IEEE Transactions on Multimedia **13**(3), 573.584.