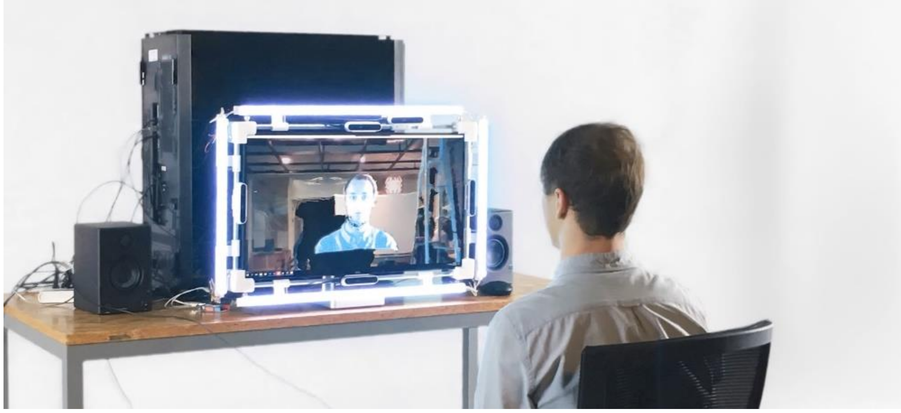
for the International Symposium on Electronic Art (ISEA),
June 2022, Barcelona, Spain



TeleWindow: A Flexible System for Exploring 3D Immersive Telepresence Using Commodity Depth Cameras

Cameron Ballard*, David Santiano*, Michael Naimark*
New York University Shanghai
Shanghai, China
clb468@nyu.edu dss441@nyu.edu michael@naimark.net
*all authors contributed equally

Abstract

Video conferencing has become an essential part of everyday life for many people. However, traditional 2D video calls leave much to be desired. Eye-contact, multiple viewpoints, and 3D spatial awareness all make video conferencing much more immersive. We present TeleWindow: a frame with mountable volumetric cameras that attaches to a display for immersive 3D video conferencing. The full system consists of a 3D display and a frame with up to four volumetric cameras. Our system was flexible conceptually as well as physically. We intended our research to be "unfettered" rather than focus and directed, e.g., for anything directly entrepreneurial and commercial: "art as unsupervised research" [1]. In contrast to similar work, our focus was on making an accessible system for exploring immersive teleconferencing so cost of materials and required technical knowledge is kept to a minimum. We present a technical baseline for immersive, easily replicable 3D teleconferencing.

Keywords

display technologies, telepresence / tele-existence, human computer interaction, mixed/augmented reality

Introduction

If the COVID-19 pandemic has shown us anything, it's the need for effective video communication tools in the contemporary world. Remote telepresence systems provide benefits across many use cases, from education to healthcare to everyday conversation. However, traditional video conferencing still has many barriers to an immersive experience, including a lack of eye contact, a static viewpoint, and no three-dimensional spatial awareness. Aside from the highest-end solutions, current commercial systems are sufficient for communication, but fail to provide the immersive experience of a face-to-face conversation. Accurate 3D spatial representation may be necessary in industry and healthcare use-cases, and immersive communication makes a huge difference in educational and everyday instances.

Much work has already been done implementing software and hardware methods for real-time 3D teleconferencing. However, most of these systems require significant resources, technical experience, and equipment to function. To achieve widespread immersive teleconferencing, these systems must be accessible to individuals and organizations without access to cutting edge expensive technology.

While existing hardware and software for 3D remains prohibitively complex or expensive, much will rapidly become more accessible as computing systems continue to develop. Volumetric capture cameras and workflows are also becoming increasingly more ubiquitous as they can be found on a multitude of mid and high-end mobile devices.

As such, we focus on creating an easily replicable physical system to demonstrate the feasibility of widely available immersive video communication. We present a modular system built with commodity depth cameras that can be replicated and installed by an end-user on a home system. Our work establishes a baseline for easily replicable 3D teleconferencing implementations.

Motivation and Related Work

Thomas Sheridan identified three main operational components of virtual "presence:" accurate sensory information, the viewer's ability to change their viewpoint, and their ability to manipulate objects remotely [2]. Since we focus on video communication, we do not consider remote manipulation a crucial function of our system and instead focus on the first two requirements identified by Sheridan. To tackle these problems, we require accurate, real-time capture and display of three-dimensional data with a dynamically updating rendering perspective.

Many systems have since been developed that meet some or all of these requirements for virtual telepresence, as described in "Immersive 3D Telepresence" [3]. An early effort from UNC reconstructed a 3D view using a "sea" of two-dimensional cameras [4]. Carnegie Mellon used a dome of 51 cameras to reconstruct a three-dimensional screen [5]. Schreer et.al. presented a system concept for a real-time 3D teleconferencing system [6]. Finally, blue-c was perhaps the first actual implementation that represented real telepresence; it reconstructed two 3D scenes and shared the data between two systems [7].

With the advent of commercially available depth cameras, contemporary systems have been able to capture and display 3D scenes in real-time. UNC developed one such system using multiple Kinect cameras placed around a room [8]. Most recently, Google announced "Project Starline", a massive and expensive system for 3D teleconferencing [9]. Other similar attempts have been made to also provide physical presence in a remote space, such as the Telehuman [10]. While these systems clearly demonstrate the possibility of remote telepresence systems, they rely on significant setup, including multiple cameras spread throughout the room, expensive equipment, and significant technical experience. Our system implements similar methods with out-of-the box technology and a simple frame around the display for a more flexible and modular system.

Most current depth cameras use a combination of point cloud data for depth representation and 2D video for texture information to achieve accurate 3D RGB video data. Many commercial systems are available for volumetric data capture, including the ZED stereo cameras, Kinect from Microsoft [11], and the RealSense cameras from Intel [12]. These cameras use some combination of infrared, RGB, and lidar data to generate 3D visual point clouds. We settled on the RealSense camera for our system, as discussed in the System Description section.

To process volumetric data and register point clouds we turned to the Point Cloud Library [13]. PCL is a comprehensive C++ library for the intake and manipulation of point cloud data, and capable of interfacing with most current commercial depth cameras.

System Description

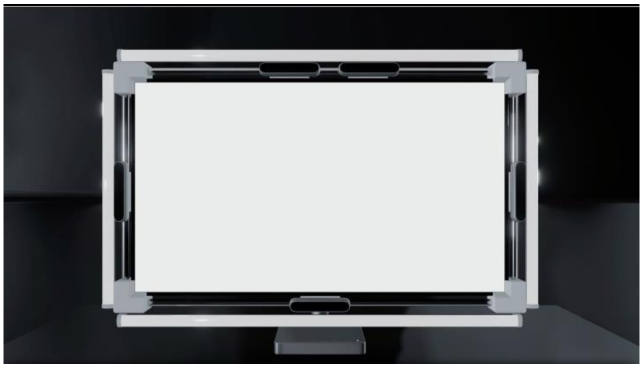


Figure 1. A CGI rendering of the frame around a display. The final system included four cameras.

Hardware

The physical implementation of our system consists of an eye-tracked stereo lenticular display from SeeFront 3D Technology with a simple frame built around the display to hold four Intel RealSense D415 volumetric capture cameras and attached lighting. The frame and cameras can be seen in Figure 1. We decided to use RealSense D415s as they proved to be a good balance of form factor, cost, performance, and accuracy; a good fit for our goals of accessibility, especially when compared to other popular depth cameras such as the Kinect. Four cameras were used in the initial build of our system with configurations available for simpler setups using fewer volumetric cameras. Lights were added to provide even illumination of the scene and increase depth camera accuracy. The physical system itself was designed to eliminate the need for required wearables such as headsets or tracking modules.

Capture Pipeline

The capture pipeline begins with a non-real-time point cloud registration sequence to roughly align the RealSense cameras using an ICP algorithm [14]. A quick and rough initial alignment is all that is needed after the volumetric cameras are locked into the frame. The multitude of cameras are used to overlay point clouds on top of each other to minimize typical occlusions on a human figure (i.e., beneath the chin or behind the arms) from lack of depth data. Point cloud data is culled so that depth data is constrained to the subject within the capture zone. In contrast, previous work focused on room-sized teleconferencing and required cameras distributed around the room. For a one-to-one use case like ours, the view is restricted to the movement typical of a face-to-face conversation and does not require angles of view that extend past the frame. An example of the point cloud merge can be seen below in Figure 3.

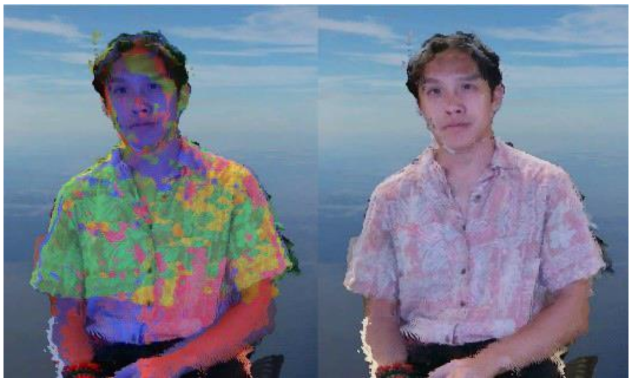


Figure 2. An example of the merged point clouds. On the left, the different colors each represent a view from one of the cameras.

It's worth noting that much of the research with 3D video capture and display uses the Microsoft Kinect rather than RealSense cameras. We found that images and point clouds were generally more accurate with a Kinect, but multiple cameras caused interference between infrared screens used for depth-mapping, and their bulky nature made mounting difficult. The accuracy of RealSense cameras is more than sufficient to exhibit an accessible 3D teleconferencing setup, and has improved considerably since we acquired them for this project.

Rendering and Display

First, the RealSense cameras return RGB texture data and a grayscale depth map. Next, the point cloud is extracted from the depth map and colored with the RGB texture using RealSense software. The transformation matrix from the registration step is used to align the point clouds within virtual space. Eye-tracking data is received from the SeeFront display's eye-tracker, which is used in an off-axis projection algorithm to present a head-tracked multiscopic view of the captured volume [15]. Eye tracking can also be performed on the RGB texture data with existing open-source computer vision software such as OpenCV.

Discussion and Future Work

The system itself runs in real-time with a resolution of 3840 x 2160 at 30 frames per second with all four cameras running. Point cloud registration can also be manually fine-tuned to more closely align point clouds given errors during ICP alignment. To save cost and avoid volumetric streaming issues, we worked with a single standalone system, to explore both real-time properties (as a "Telemirror") and record/playback opportunities (as a "Telerecorder"). This allowed us to focus on issues of immersion rather than bandwidth.

Our results may be described as "noisy and inaccurate" yet "cinematic and authentic." We wanted to minimize the cartoonishness of smoothed, highly interpolated avatars in favor of a "raw" look. We often used Star Wars' Princess Leia hologram as exemplar: even though it was completely fictive and made for a sci-fi film, it consciously included vertical raster lines and noise spikes. This intentional representation of noise and unfiltered data minimizes any "uncanny valley" effect, and our test users appreciated this alternative-to-avatar approach.

Test users also recognized glasses-free 3D as a necessary component. There is a certain magic, sitting in front of a 2D display, when the eye tracker locks on the users' eyes and the image instantly becomes stereoscopic. It is a unique media experience.

Given a single system, we could only explore multiscopic, "multi-view" perspective via pre-recording. For example, a single volumetric frame can be grabbed which allows a user to look around their own face, something impossible to do with a mirror. We find a multiscopic perspective is a key feature for maintaining immersion in future TeleWindows.

Future work may involve exploring the middle ground between a "low-tech" solution, our solution presented with the TeleWindow system, and high-end teleconferencing setups like Google's Project Starline. General improvements can be made to our system such as refinement and streamlining of the current capture and rendering pipeline. Since the initial development, new cameras have become available, and volumetric capture technology has already become more accurate with cleaner captures. Implementation of real-time deep learning based rendering helpers may also minimize occlusions and clean up depth data for better representations of human beings. But our exploration also has a twist.

As summer 2020 approached, the COVID pandemic had engulfed the world. We were in our third year of TeleWindow exploration, with the bulk of our work taking place during the summer with recent Media Arts and Computer Science graduates. Our flexible, "unsupervised" approach gave us a well-informed view of both the big picture of immersive teleconferencing and the in-the-trenches experience of what works, what doesn't, challenges, and opportunities. We decided to spend summer 2020 applying what we learned to something focused and practical.

We focused on how we could help college students as millions went online for the first time. Students were the perfect subject for our research because a) they were regular users of teleconferencing software and b) we could safely assume that nearly all of them had a laptop and smartphone.

The result was a “cheap simple hack” of separating the speaker view from the presentation or seminar view by moving it to a smartphone above the laptop’s camera and screen [16]. It’s not where we expected to go. We began by considering predictable solutions like computational view interpolation, relighting, and more spatial sound. But once we tried our hack, it worked much better than we expected. We produced several hundred and freely distributed them around our campus. Other universities participated as well. We don’t think we would have discovered this solution without the “unfettered” “unsupervised” TeleWindow exploration behind us.

We anticipate moving forward in both directions: continued exploration of the original TeleWindow and further development of “cheap hacks,” along with various versions in between.

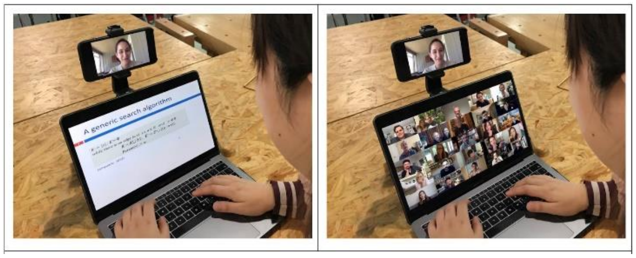


Figure 3. Our cheap simple hack for telepresence.

Conclusion

The system outlined here was designed as a jumping-off point to explore what was possible in the field of immersive teleconferencing given increasingly accessible tools and technology for volumetric capture and display. The system and variations can be easily replicated. Our “simple hack” was one variation, an extreme scaled-down version. An immersive middle ground could contain a 2D, but still head-tracked multiscopic version, using a single clip-on volumetric camera. Our own exploration into telepresence demonstrated the feasibility of a widespread, low-cost immersive teleconferencing experience.

As is often the case exploring emerging media, unanticipated results occur. While some are not so good (but always increase knowledge), others can be magical. We encourage further exploration of more such uncharted territories.

Acknowledgements

We gratefully acknowledge support from Clay Shirky, Xiaojing Zu, and Jeffrey Lehman for this project. Students and recent graduates who have worked with us include Ada Zhao, Barak Chamo, Danxiaomeng Ivy Huang, Deyin Mia Zhang, Diana Yaming Xu, Grace Huang, Jingtian Zong, Mateo Juvera Molina, Nathalia Lin, Runxin Bruce Luo, Tianyu Michael Zhang, Xincheng Huang, Xinran Azena Fan, Yufeng Mars Zhao, Yunru Casey Pan, Zhanghao Chen, and designers Ignazio Moresco and Paolo Salvagione. Thank you all!

References

[1] Bratton, B. 2018, “The New Normal: Planetary-Scale Computation, AI Urbanism and the Expanded Field of Art and Design.” Benjamin Bratton Talk, NYU Shanghai, December 12, 2018.

[2] Sheridan, T. Musings on Telepresence and Virtual Presence. *Presence Teleoperators & Virtual Environments*, 1992.

[3] Fuchs, H., State, A., and Bazin, J.C. “Immersive 3D Telepresence.” *Computer*, 2014.

[4] Fuchs, H. et al. "Virtual Space Teleconferencing Using a Sea of Cameras", *Proc. 1st Int'l Conf. Medical Robotics and Computer Assisted Surgery*, 1994.

[5] T. Kanade, P. Rander and P. J. Narayanan, "Virtualized reality: constructing virtual worlds from real scenes," in *IEEE MultiMedia*, vol. 4, no. 1, pp. 34-47, Jan.-March 1997

[6] Schreer, O., Feldmann, I., Atzpadin, N., Eisert, P., Kauff, P., and Belt, H. J. W., "3DPresence -A System Concept for Multi-User and Multi-Party Immersive 3D Videoconferencing," *5th European Conference on Visual Media Production*, 2008

[7] Gross, M., Würmlin, S., Naef, M., Lamboray, E., Spagno, C., Kunz, A., Koller-Meier, E., Svoboda, T., Van Gool, L., Lang, S., Strehlke, K., Moere, A.V., and Staadt, O. Blue-c: a spatially immersive display and 3D video portal for telepresence. *ACM Trans. Graph*, 2003.

[8] Maimone, A., Bidwell, J., Peng, K., and Fuchs, H. “Enhanced personal autostereoscopic telepresence system using commodity depth cameras.” *Comput. Graph.*, 2012.

[9] Google. Project Starline: Feel like you’re there, together. Accessed in August 2021. https://blog.google/technology/research/project-starline/

[10] Kim, K., Bolton, J., Girouard, A., Cooperstock, J., and Vertegaal, R. TeleHuman: effects of 3d perspective on gaze and pose estimation with a life-size cylindrical telepresence pod. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2012.

[11] Microsoft. Kinect for Windoisws. https://developer.microsoft.com/en-us/windows/kinect/

[12] Keselman, L., Woodfill, J., Grunnet-Jepsen, A., and Bhowmik, A. “Intel RealSense Stereoscopic Depth Cameras.” *ArXiv*, 2017.

[13] Rusu, R. B., and Cousins, S., "3D is here: Point Cloud Library (PCL)," *2011 IEEE International Conference on Robotics and Automation*, 2011.

[14] Rusinkiewicz, S. and Levoy, M. “Efficient variants of the ICP algorithm.” *Proceedings Third International Conference on 3-D Digital Imaging and Modeling*, 2001.

[15] Kooima, Robert. “Generalized Perspective Projection.” 2011.

[16] Naimark, Michael. “A Cheap Simple Hack for Improving Your Online Classtime Experiences” *Medium* (blog). January 8, 2021. https://michaelnaimark.medium.com/a-cheap-simple-hack-for-improving-your-online-class-time-experiences-802071cd34c1