



Immersive 3D Telepresence

Henry Fuchs and Andrei State,
University of North Carolina at Chapel Hill
Jean-Charles Bazin, *ETH Zürich*

Cutting-edge work on 3D telepresence at a multinational research center provides insight into the technology's potential, as well as into its remaining challenges.

For more than two decades, individually and with many collaborators, we have actively explored immersive 3D telepresence technology. Since 2011, we have been working within the BeingThere International Research Centre for Tele-Presence and Tele-Collaboration, a joint research effort among Nanyang Technological University (NTU) in Singapore, ETH Zürich in Switzerland, and the University of North Carolina (UNC) at Chapel Hill.

The BeingThere Centre is directed by Nadia Magnenat-Thalmann at NTU, Markus Gross at ETH Zürich, and Henry Fuchs at UNC. We invite readers to visit the Centre's website (<http://imi.ntu.edu.sg/BeingThereCentre>), which provides information about the dozens of faculty, staff, and students researching 3D telepresence as well as mobile avatars, virtual humans, and various 3D scanning and display technologies.

Here we present a brief overview of some of our recent immersive 3D telepresence work, focusing on major issues, recent results, and remaining challenges, mainly with respect to 3D acquisition and reconstruction, and 3D display.

We do not discuss issues related to real-time data transmission, such as networking and compression.

TELEPRESENCE

Researchers have long envisioned “telepresent” communication among groups of people located in two or more geographically separate rooms, such as offices or lounges, by means of virtual joining of the spaces. As Figure 1 shows, shared walls become transparent, enabling participants to perceive the physically remote rooms and their occupants as if they were just beyond the walls—life size, in three dimensions, and with live motion and sound.

An ideal implementation would provide wall-size, multiuser-autostereoscopic (or *multiscopic*, that is, showing individualized 3D views to each user) displays along the shareable walls, allowing encumbrance-free, geometrically correct 3D viewing of the remote sites. Together with directional sound, such a system should create a convincing sense of co-presence within the joint real-virtual space, enabling almost any kind of natural interaction and communication short of actually stepping across the seemingly transparent walls into the other rooms (Figure 2a).

Alternatively, if a display were mounted on a moving platform (Figure 2b), remote participants could move anywhere in the local environment. With the help of a transparent screen, they could be even more effectively integrated with that space and its occupants, at the expense of not showing their own remote environments.

WHY 3D MIGHT BE BETTER THAN 2D

With conventional 2D teleconferencing systems such as Skype, and even with high-end systems such as Cisco TelePresence TX9000, the imagery seen by all participants at one site is exactly what is acquired by the one or more cameras located at the remote site(s). This is fundamentally different from in-person, face-to-face meetings, where each participant sees the surroundings from his or her own point of view, and each point of view is unique because participants are sitting or standing in different locations around the room.

In face-to-face meetings, we each change our location and direction of gaze so naturally that we hardly give it a thought. In addition, when someone is looking at us, not only do we see that person looking at us, but everyone else can observe that person looking at us from his or her own point of view. It has been shown that mutual gaze enhances human communication,¹ and thus we also aim to offer this capability in the systems we design.

Natural movement in 3D space, situational awareness, gaze direction, and eye contact are very difficult to provide in 2D teleconferencing, where all participants see the remote scene from the fixed viewpoint(s) of the remote camera(s). Hence, to achieve most of the benefits of face-to-face interaction, we believe that each local participant should receive personal imagery of the remote environment that matches his or her dynamically changing point of view.

The importance of 3D display is an active research question and appears to depend on the target application and context.² For example, while 2D display might be sufficient for casual one-to-one video conferencing, 3D display could play a key role in telepresence scenarios such as collaborative work, 3D object or data manipulation, and remote space immersion.

3D TELEPRESENCE REQUIREMENTS

To generate the novel views for each participant, two major approaches are being used: image-based methods and 3D reconstruction. Image-based methods³ require deployment of many cameras and are appropriate when the novel viewpoints are close to the cameras' physical locations. In contrast, 3D reconstruction estimates the scene's actual 3D shape (objects, people, background, and so on), resulting in a dynamic geometric model that can then be rendered for these novel viewpoints. Moreover, such a geometric model can be enhanced with synthetic representations of objects of interest.



Figure 1. An artist's depiction of the BeingThere Centre's multiroom telepresence concept shows three geographically remote rooms, virtually joined as if co-located and separated by seemingly transparent walls.

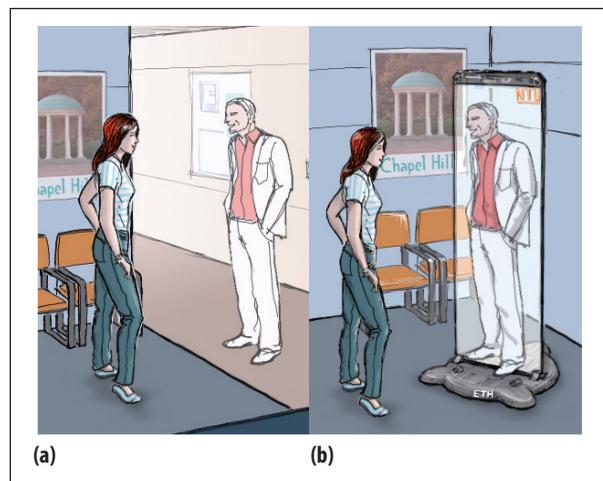


Figure 2. Telepresence implementations. (a) Natural face-to-face interaction between two participants in a room-based scenario. (b) Remote participant displayed on a mobile, life-size, transparent stereoscopic display.

To provide dynamic personalized views to each telepresence participant, we reconstruct the 3D environment of each site and display it in 3D at each of the other sites. This requires three distinct but closely coupled processes:

- continuously scan each environment to build and maintain an up-to-date 3D model of it, including all people and objects;
- transmit that 3D model to the other sites; and
- generate and display to each participant the appropriate 3D view of each distant room and its contents.

Our recent work has emphasized the acquisition and display challenges, both active research areas in the 3D vision and graphics communities.

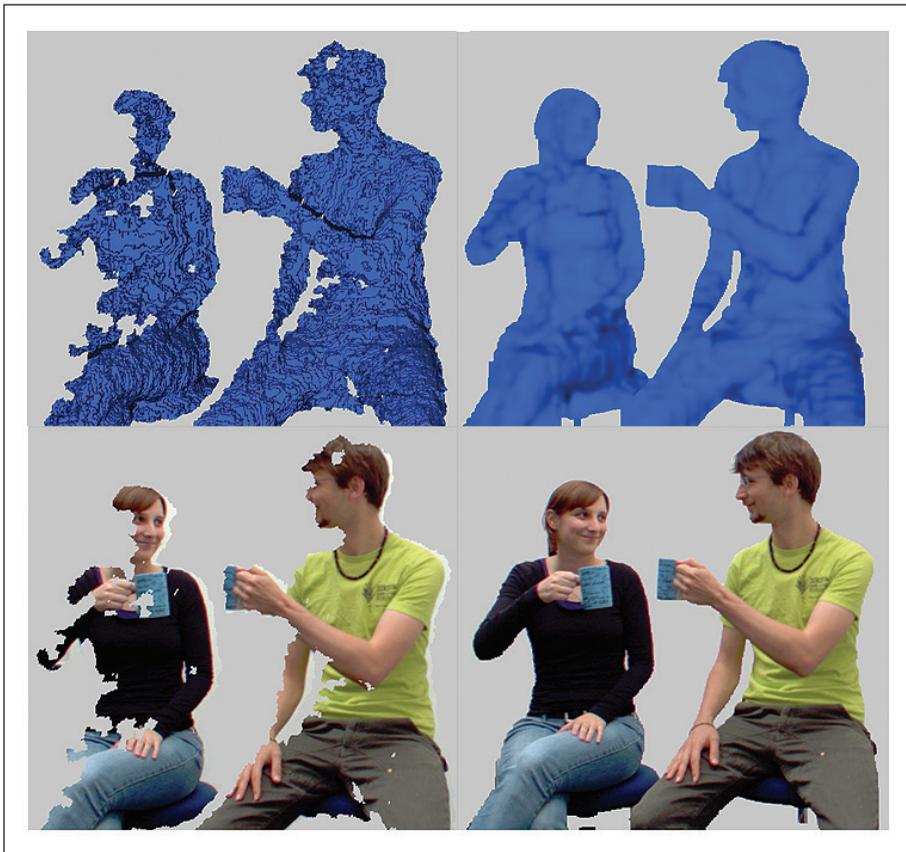


Figure 3. Real-time automatic 3D reconstruction of participants from a single color-plus-depth (RGB-D) camera. The top row shows the geometry, and the bottom row shows its textured version. The images on the left show raw RGB-D data; the images on the right show RGB-D data after filtering and the application of occlusion and photometric-consistency operations.

PREVIOUS WORK

The dream of 3D telepresence has inspired researchers for many decades, but, due to technological difficulties, prototypes emerged slowly in the 1980s and 1990s. The basic approach for creating 3D reconstructions of room-size environments involves deploying multiple cameras around the space and using their imagery with various stereo reconstruction techniques to continually update the 3D model of that space, including, of course, the moving people. A UNC-led team conducted an early experiment in 3D teleconferencing using such a “sea” of cameras.⁴ Carnegie Mellon University researchers created one of the first systems to capture dynamic scenes and render them from new viewpoints using a set of 51 cameras fixed on a five-meter dome.⁵ ETH Zürich’s blue-c was perhaps the first bidirectional 3D telepresence system—it scanned as well as displayed in 3D a participant at each of two locations.⁶ Later work at the Electronic Visualization Laboratory of the University of Illinois at Chicago introduced simultaneous 3D display for two or three local users.⁷

Early prototypes used bulky head-mounted displays (HMDs).⁴ While providing a strong 3D illusion of a distant person or environment, these early HMDs were inadequate if for no other reason than they required each participant to view other distant or local partners by means of a helmet or goggles—hardly a satisfactory illusion. The display in blue-c was a considerable improvement: a CAVE (computer-assisted virtual experience) environment)-like experience with head-tracked stereo display using active shutter stereo glasses for the single local user.⁶

Although today’s stereo shutter glasses are still so dark that they preclude effective eye contact and thus impede some forms of natural interaction, they are the only currently available technology supporting fully individualized stereo views for multiple local users. An excellent example of a multiuser stereo display (and 3D acquisition) system was developed by a team at Bauhaus

University.^{8,9} It supports up to six local users through six stereo projectors rear-projecting onto the same screen area. The ingenious design permanently assigns each projector the task of showing a primary color (red, green, or blue) and a single eye’s view (left or right) to all six users, with each user’s view displayed at one of six time slots during each video frame.

3D ACQUISITION AND RECONSTRUCTION

3D acquisition and reconstruction are the technologies that feed the 3D telepresence pipeline. They have to meet critical requirements of accuracy, completeness, and speed. A room-size environment can be viewed by remote partners from a multitude of viewpoints located anywhere in the remote site’s “telepresence room.” For example, consider Figure 1: one of the seated participants at the UNC site (foreground) might get up and walk up very close to the wall display to conduct a semiconfidential, low-volume conversation with one of the NTU participants, perhaps as illustrated in Figure 2a. Supporting such natural behavior

requires that the 3D telepresence system acquire and reconstruct minute details of each environment with high accuracy.

People's continuous movement—walking, gesturing, changing facial expressions, and so on—makes this task exceptionally difficult. Yet such details must be captured and properly reconstructed at remote sites without annoying or misleading visual artifacts. Today's teleconferencing users are accustomed to high-definition 2D video and are unlikely to accept jarring image-quality degradation in exchange for true viewpoint-specific dynamic stereoscopy.

The most popular traditional 3D reconstruction strategy has been to use numerous conventional color cameras. The recent emergence of inexpensive color-plus-depth (RGB-D) cameras, such as Microsoft's Kinect, has revolutionized 3D reconstruction, and we have been using them in many of our telepresence projects. For small-scale scenarios with few participants (typically up to two at each site), one or two Kinects are sufficient. While Kinects can extract textured geometry in real time, their data contains spatial and temporal noise as well as missing values, especially along depth discontinuities; therefore, raw Kinect data must be processed to eliminate or reduce those.

Figure 3 shows the quality enhancement by one of our 3D reconstruction techniques (<http://beingthere.ethz.ch/videos/VMV2011.mp4>).¹⁰ For larger environments or more participants, Figure 4 shows a typical real-time 3D reconstruction of a room scene using 10 Kinect cameras, which represents the approximate limit of the amount of data that can be processed today within a single common PC in real time (www.cs.unc.edu/TelepresenceVideos/RealTimeVolumetricTelepresence.mp4).¹¹ We have also used RGB-D cameras for real-time gaze correction, and developed a Skype plug-in to provide convincing eye contact between videoconferencing participants (<http://beingthere.ethz.ch/videos/SA2012.mp4>).¹²

3D DISPLAY

When the 3D telepresence display only needs to show a single distant individual, we might choose to project that person without his or her background environment on a human-size transparent display to give a strong illusion of that distant individual's presence in the local environment (Figure 2b). Figure 5 shows one of our implementations of this concept, with a transparent screen displaying rear-projected stereoscopic imagery.^{10,15}



Figure 4. Virtual views of real-time 3D room reconstructions from 10 Kinect RGB-D cameras.

If there is only one local participant and the reduced eye contact of stereo glasses is acceptable, then a simple head-tracked stereo display might be adequate. We have built numerous such systems, including the transparent display¹³ of Figure 5 and others consisting of several large stereoscopic TVs forming a wall-size personal display window into the remote site.

A more significant challenge arises when there are multiple local participants, in which case we seek to present to each participant the correct stereoscopic view from each of their positions, preferably without any encumbering stereo glasses. To achieve this multiscope display, we have been exploring techniques developed for compressive light-field displays at the MIT Media Lab.¹⁴ Together with its team, we have recently improved such displays by optimizing the light-field views only for the current spatial locations of all viewers.¹⁵ Figure 6 shows two photos of our optimized display, simultaneously taken from two

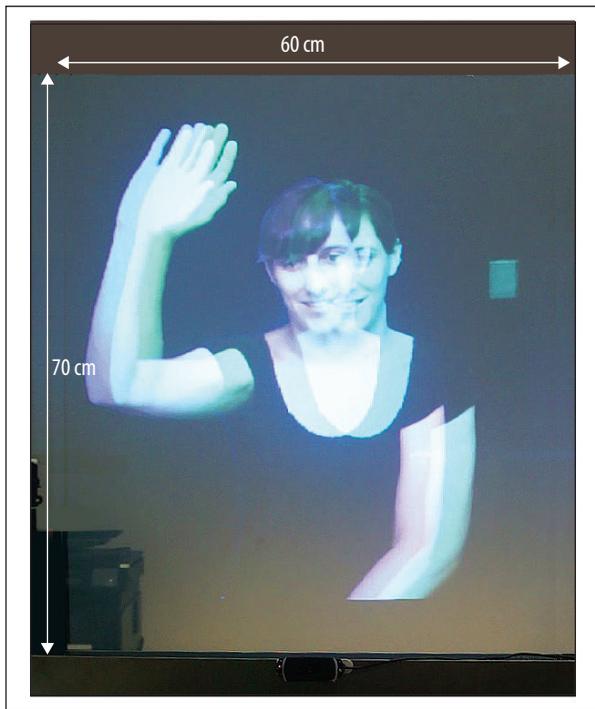


Figure 5. Photo of near-life-size stereoscopic transparent rear-projection display, showing both left and right eye stereo images (the display is viewed with passive stereo glasses). Note that furniture in the background is visible through the transparent display.

different vantage points without any filters or specialized glasses. We plan to build larger displays from tiled copies of this 27-inch prototype.

REMAINING CHALLENGES

Despite recent progress in 3D telepresence, challenges remain both in 3D acquisition and reconstruction and in 3D display.

3D acquisition and reconstruction

Accurate and rapid 3D reconstruction of an entire meeting room will require dozens of RGB-D cameras, more than can be operated by a single PC today. To that end, we must design a distributed real-time acquisition system; such a system's work becomes increasingly challenging as the acquisition volume increases. System complexity is indeed a significant issue: today, even high-end teleconferencing systems use only a small number of displays and a few high-quality cameras. In contrast, immersive 3D telepresence systems will likely require many more cameras, advanced unconventional displays, and considerably more processing.

Another challenge is that the quality of images obtained through real-time 3D capture and reconstruction is not on par with that of 2D images directly acquired by conventional cameras. Reconstruction artifacts and missing 3D

data are intolerable to users accustomed to high-definition image quality even from inexpensive webcams. However, there have been rapid advances in consumer-grade RGB-D cameras, both in existing offerings (Microsoft Kinect, PrimeSense Capri) and in new models (Google's Project Tango, Intel's RealSense 3D camera).

There also has been progress in 3D reconstruction algorithms. Recent work demonstrates improvements in reconstruction quality from processing of shadows from the infrared projector in an RGB-D camera,¹⁶ accumulation of temporal data for fixed as well as deformable objects such as people (www.cs.unc.edu/TelepresenceVideos/VR2014.mp4),¹⁷ and spatiotemporally consistent reconstruction from several hybrid cameras (<http://beingthere.ethz.ch/videos/EG2014.mp4>).¹⁸ However, most of these techniques are not yet capable of real-time performance.

3D display

Today, even state-of-the-art wall-size stereo displays that provide personalized views to each freely moving user require specialized stereo glasses.⁹ For many, these dark glasses would be uncomfortable and visually unacceptable, as they impede eye contact. The more attractive alternatives—high-quality, large-format multiscopic displays—are still in a basic research phase and remain to be built but would have an enormous impact on the field.

Augmented reality (AR) eyeglass-style displays could enable more flexible interaction among participants than even multiscopic displays. Using 3D models of the remote and local environments, these HMDs could achieve the most-powerful-yet sense of combined presence,¹⁹ as Figure 7 shows (www.cs.unc.edu/TelepresenceVideos/AugmentedRealityTelepresence.mp4). The newest designs promise significant improvements over older-style goggles: more transparency for better eye contact and a brighter view of the local environment, as well as a wider field of view and an eyeglass form factor suitable for long-term wear.²⁰ We hope for a convenient see-through AR display like Google Glass and the Lumus DK-32, but with a wider field of view such as the Oculus Rift.

The most encouraging aspect for the future of 3D telepresence is that relevant technologies are rapidly advancing because of consumer interest in the various components: ever-higher-quality large-format video displays, ever-higher-quality RGB-D cameras, and ever-faster GPUs for gaming and entertainment. Most of these technologies can be easily repurposed for the demanding scenarios of 3D telepresence. In the past five years, the advances have been greater than in the previous 15; we expect continuing and exciting improvements in the next few years.

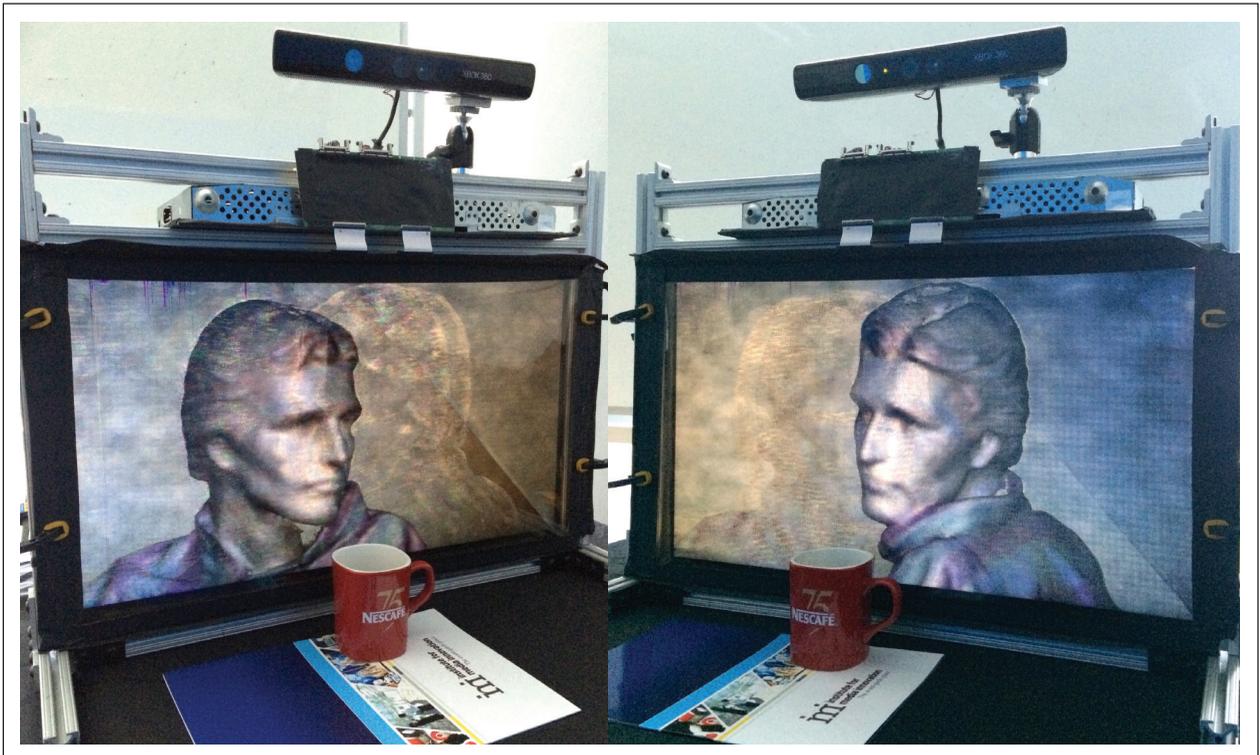


Figure 6. Simultaneously taken photos of a multiscope display from two different viewpoints. Note that the virtual head faces perpendicularly out of the screen; its spatial relationship to the mug in the foreground is preserved, and each viewer sees a different side of the head.

We see no obvious roadblocks to the realization of immersive 3D telepresence. As with other dramatic changes, such as the move from analog to digital television, the older technology can remain dominant during decades of incremental development. However, as cost and effectiveness of new 3D telepresence technologies continue to improve, the advantages of 3D telepresence over 2D teleconferencing will become increasingly attractive. 

Acknowledgments

We gratefully acknowledge our colleagues within and outside of the BeingThere Centre: codirectors Markus Gross and Nadia Magnenat-Thalmann; project leaders Tat Jen Cham, I-Ming Chen, Marc Pollefeys, Gerald Seet, and Greg Welch; assistant director Frank Guan; professors Jan-Michael Frahm, Anselmo Lastra, Tiberiu Popa, Miriam Reiner, and Turner Whitted; and research assistants Nate Dierk, Mingsong Dou, Iskandarsyah, Claudia Kuster, Andrew Maimone, Tobias Martin, and Nicola Ranieri. Special thanks to Renjie Chen for the photos of our multiscope display and to Mingsong Dou for the 3D head-scan data. We



Figure 7. View through an experimental augmented reality head-mounted display showing a distant participant across the local table and a virtual couch model on the table.

are also grateful to Mark Bolas of the University of Southern California and Tracy McSheery of PhaseSpace for the experimental optical see-through HMD through which the image in Figure 7 was acquired.

This research was supported in part by the BeingThere Centre, a collaboration among ETH Zürich, NTU Singapore, and



UNC Chapel Hill, supported by ETH, NTU, UNC, and the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the Interactive Digital Media Programme Office. Part of this research was also supported by Cisco Systems, and by the US National Science Foundation, Award IIS-1319567 “HCC: CGV: Small: Eyeglass-Style Multi-Layer Optical See-Through Displays for Augmented Reality.”

References

1. M. Argyle and M. Cook, *Gaze and Mutual Gaze*, Cambridge Univ. Press, 1976.
2. J.P. McIntire, P.R. Havig, and E.E. Geiselman, “Stereoscopic 3D Displays and Human Performance: A Comprehensive Review,” *Displays*, vol. 35, no. 1, 2014, pp. 18–26.
3. H.-Y. Shum, S.-C. Chan, and S.B. Kang, *Image-Based Rendering*, Springer, 2008.
4. H. Fuchs et al., “Virtual Space Teleconferencing Using a Sea of Cameras,” *Proc. 1st Int’l Conf. Medical Robotics and Computer Assisted Surgery (MRCAS 94)*, 1994, pp. 161–167.
5. T. Kanade, P. Rander, and P.J. Narayanan, “Virtualized Reality: Constructing Virtual Worlds from Real Scenes,” *IEEE MultiMedia*, vol. 4, no. 1, 1997, pp. 34–47.
6. M. Gross et al., “blue-c: A Spatially Immersive Display and 3D Video Portal for Telepresence,” *ACM Trans. Graphics*, vol. 22, no. 3, 2003, pp. 819–827.
7. T. Peterka et al., “Advances in the Dynallax Solid-State Dynamic Parallax Barrier Autostereoscopic Visualization Display System,” *IEEE Trans. Visualization and Computer Graphics*, vol. 14, no. 3, 2008, pp. 487–499.
8. A. Kulik et al., “C1x6: A Stereoscopic Six-User Display for Co-located Collaboration in Shared Virtual Environments,” *ACM Trans. Graphics*, vol. 30, no. 6, 2011; doi:10.1145/2070781.2024222.
9. S. Beck et al., “Immersive Group-to-Group Telepresence,” *IEEE Trans. Visualization and Computer Graphics*, vol. 19, no. 4, 2013, pp. 616–625.
10. C. Kuster et al., “Towards Next Generation 3D Teleconferencing Systems,” *Proc. 3DTV-Conf.: The True Vision—Capture, Transmission and Display of 3D Video (3DTV-CON 12)*, 2012; doi:10.1109/3DTV.2012.6365454.
11. A. Maimone and H. Fuchs, “Real-Time Volumetric 3D Capture of Room-Sized Scenes for Telepresence,” *Proc. 3DTV-Conf.: The True Vision—Capture, Transmission and Display of 3D Video (3DTV-CON 12)*, 2012; doi:10.1109/3DTV.2012.6365430.
12. C. Kuster et al., “Gaze Correction for Home Video Conferencing,” *ACM Trans. Graphics*, vol. 31, no. 6, 2012; doi:10.1145/2366145.2366193
13. N. Ranieri, H. Seifert, and M. Gross, “Transparent Stereoscopic Display and Application,” *Proc. SPIE*, vol. 9011, 2014; doi:10.1117/12.2037308.
14. G. Wetzstein et al., “Tensor Displays: Compressive Light Field Synthesis Using Multilayer Displays with Directional Backlighting,” *ACM Trans. Graphics*, vol. 31, no. 4, 2012; doi:10.1145/2185520.2185576.
15. A. Maimone et al., “Wide Field of View Compressive Light Field Display Using a Multilayer Architecture and Tracked Viewers,” to appear in *Proc. SID Display Week*, 2014.
16. T. Deng et al., “Kinect Shadow Detection and Classification,” *Proc. IEEE Int’l Conf. Computer Vision Workshops (ICCVW 13)*, 2013, pp. 708–713.
17. M. Dou and H. Fuchs, “Temporally Enhanced 3D Capture of Room-Sized Dynamic Scenes with Commodity Depth Cameras,” *Proc. IEEE Conf. Virtual Reality (VR 14)*, 2014, pp. 39–44.
18. C. Kuster et al., “Spatio-Temporal Geometry Fusion for Multiple Hybrid Cameras Using Moving Least Squares Surfaces,” *Computer Graphics Forum*, vol. 33, no. 2, 2014; doi:10.1111/cgf.12285.
19. A. Maimone et al., “General-Purpose Telepresence with Head-Worn Optical See-through Displays and Projector-Based Lighting,” *Proc. IEEE Conf. Virtual Reality (VR 13)*, 2013; doi:10.1109/VR.2013.6549352.
20. A. Maimone and H. Fuchs, “Computational Augmented Reality Eyeglasses,” *Proc. IEEE Int’l Symp. Mixed and Augmented Reality (ISMAR 13)*, 2013, pp. 29–38.

Henry Fuchs is the Federico Gil Distinguished Professor of Computer Science at the University of North Carolina at Chapel Hill, and codirector of the BeingThere Centre. His research interests include telepresence, augmented reality, and graphics hardware and algorithms. Fuchs received a PhD in computer science from the University of Utah. He is a member of the National Academy of Engineering. Contact him at fuchs@cs.unc.edu.

Andrei State is a senior research scientist in the Department of Computer Science at the University of North Carolina at Chapel Hill, and cofounder of InnerOptic Technology, which creates virtual reality guidance for surgeons. His research interests include telepresence and virtual and augmented reality. State received a Dipl.-Ing. (aer) from the University of Stuttgart, Germany, and an MS in computer science from UNC Chapel Hill. Contact him at andrei@cs.unc.edu.

Jean-Charles Bazin is a senior researcher in the ETH Zürich Computer Graphics Laboratory, and conducts research at the BeingThere Centre. His research interests include various topics in computer vision and graphics, such as image/video editing and 3D data processing. Bazin received a PhD in electrical engineering from KAIST, South Korea, and an MS in computer science from Université de Technologie de Compiègne, France. Contact him at jebazin@inf.ethz.ch.



Selected CS articles and columns are available for free at <http://ComputingNow.computer.org>.