An Immersive Bidirectional System for Life-size 3D Communication

Claudia Plüss (Kuster)
ETH Zurich
claudia.pluess@inf.ethz.ch

Nicola Ranieri
ETH Zurich
ranierin@inf.ethz.ch

Jean-Charles Bazin
ETH Zurich
jean-charles.bazin@inf.ethz.ch

Tobias Martin
ETH Zurich
tobias.martin@inf.ethz.ch

Pierre-Yves Laffont
ETH Zurich
pierre-yves.laffont@inf.ethz.ch

Tiberiu Popa
Concordia University
tiberiu.popa@concordia.ca

Markus Gross
ETH Zurich
grossm@inf.ethz.ch

ABSTRACT

Telecommunication and video conferencing are an integral part of modern society with implications in many aspects of everyday life. However, compared to a meeting in person, the sense of presence is still limited in electronic communication. In this paper, we present a novel system for life-size 3D telecommunication. It is designed to create an immersive user experience by seamlessly embedding a remote conversation partner into the local environment. To achieve this, users are captured in 3D by hybrid (color+depth) sensors and displayed on a life-size transparent 3D display. We have built two instances of this system in Zurich and Singapore. They form a complete and fully functional prototype enabling bidirectional communication in real-time over a long distance. We further demonstrate alternative hardware setups, which make our system flexible and adaptable to different usage scenarios.

CCS Concepts

•Computing methodologies → 3D imaging; Mixed / augmented reality; Reconstruction; •Hardware → Displays and imagers;

Keywords

Telepresence systems, Telecommunication, 3D displays, 3D video processing

CASA 2016, May 23 - 25, 2016, ,

ISBN 978-1-4503-4745-7/16/05. . . $15.00

DOI: http://dx.doi.org/10.1145/2915926.2915931

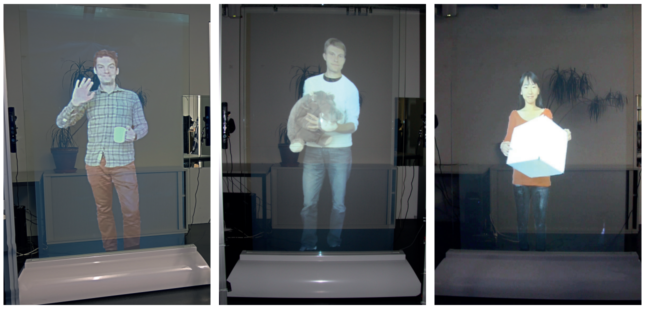

Figure 1: Our telecommunictaion system creates the illusion of a remote person being present in the local environment.

1. INTRODUCTION

In the past decades, globalization has sparked the need for advanced communication technologies and has increasingly made them indispensable. Video conferencing systems such as Skype or Google Hangouts are easily accessible and widely used by companies and individuals.

However, the most popular means of modern communication have a number of shortcomings. Classical video conferencing solutions are typically webcams on a laptop or desktop computer. They narrow the appearance of the remote participant to a cropped and resized “flat” image on a 2D screen, showing only the face or upper body. This limits the sense of immersion and may hinder effective communication, as many non-verbal cues are lost. First, gestures and body language are only partly conveyed, even though they are an essential part of human communication. Second, perceptual realism is limited due to the lack of disparity between the left and right eye’s views (absence of binocular parallax). Third, spatial relationships, such as gaze and gesture directions are not preserved [17]. Finally, the visual discontinuity between the remote scene background displayed on the screen and the local physical environment degrades the feeling of mutual presence.

In this paper, we explore possibilities for next-generation telepresence systems, which address these challenges. Our goal is to provide a fully immersive communication experience, and to increase the sense of mutual presence among participants situated in two geographically distant locations. In contrast to the numerous methods which aim to give users the feeling that they are at a remote or at a virtual location [6, 30, 3], we aim to create the illusion of distant persons "being here" (see Figure 1).

Thus, we propose to *integrate* life-size, full-body, three-dimensional representations of remote persons into the local environment. This enables immersive interaction including body language, binocular parallax, and 3D spatial awareness.

Our main contribution is a bidirectional life-size 3D telecommunication system, which implements this vision and allows communication in real-time over a large distance. We develop a set of hardware and software components, namely for capture, transmission and display, to enable life-size 3D telepresence setups with practical impact. Furthermore, we modify the acquisition method by Kuster et al. for efficient fusion of data from multiple color and depth sensors [11] to satisfy the requirements of real-time teleconferencing. We demonstrate the flexibility and generality of our telepresence framework by adapting it to two relevant telecommunication applications. Depending on the usage scenario, different configurations of the acquisition and display modules can be easily combined.

To achieve the concept of "being here", we acquire a remote person's appearance and geometry using hybrid (color + depth) sensors and then display this person on our life-size transparent 3D display. The display shows a pair of stereo images and can be viewed by one or multiple users wearing passive polarized glasses. We have built two instances of this system in Zurich and Singapore, enabling bidirectional communication across several thousand kilometers.

2. RELATED WORK

A wealth of telecommunication systems have evolved in the past decades. In the following, we provide an overview focusing on end-to-end systems, which include capture, transmission, and display components. The most widely used mainstream applications for 2D video conferencing include Skype and Google Hangouts, while Cisco's TelePresence TX 9000, HP's HALO or Polycom TPX are examples of advanced commercial systems. Participants are acquired by color cameras and then directly displayed on 2D screens.

The early work of Fuchs et al. [5] demonstrates 3D teleconferencing using a "sea" of cameras. It enables users wearing a head-mounted display to look around a remote environment captured by 11 cameras and reconstructed through stereo techniques. Another pionnering work is the virtualized reality system by Kanade et al. [9], which captures dynamic scenes and renders them from novel viewpoints. One of the first bidirectional platforms for telepresence has been the Blue-C project [6]. Users are captured inside a CAVE-like environment and can interact with geographically distant participants in a shared virtual environment.

In the context of 3D TV, a number of high-quality live systems have been presented [16, 27, 2]. They are based on large camera arrays, arranged in a row or in a grid, and autostereoscopic 3D displays, which do not require glasses. Novel views are generated by image-based rendering techniques. While the image quality for these systems is generally visually attractive due to the high sampling of the capture space, they require rather elaborate hardware setups and generate large amounts of data to be processed. Furthermore, they are sensitive to calibration and synchronization of the large camera and projector arrays. In contrast, our capture setups consist of a small number of off-the-shelf devices.

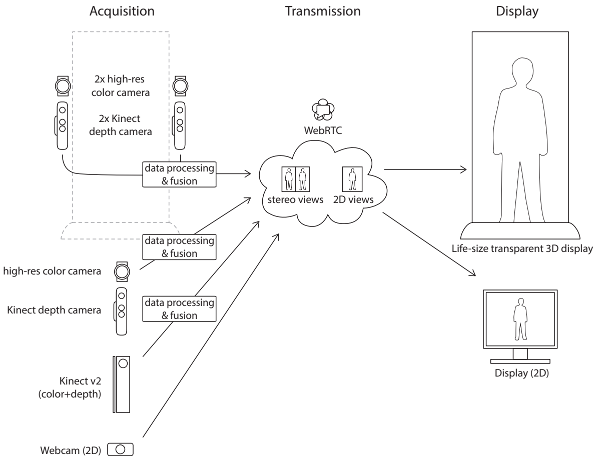

Figure 2: System overview. Schematic illustration of our acquisition, transmission, and display modules, and the data flow between them. The hardware options for the acquisition and display modules are displayed on the left and on the right, respectively. For bidirectional communication, an acquisition and display module are combined at each site.

Focusing on the display component, Jones et al. [8] propose a one-to-many teleconferencing system using structured light for 3D acquisition, and a volumetric 3D display to render a spatially faithful view of the participant's head. With the emergence of inexpensive real-time depth cameras such as the Kinect, the task of acquiring 3D geometry for scene reconstruction has become straightforward. Consequently, teleconferencing systems incorporating one or several such sensors have been proposed [28, 14].

Recent systems for teleimmersion [24, 30, 3, 18] propose advanced 3D acquisition methods including color and depth sensors, 3D sound rendering, and video coding techniques. They allow users to interact and collaborate in a shared virtual environment.

The systems mentioned above propose means to create a window to a remote location, or to insert 3D representations of participants into a virtual space. This is conceptually different from our goal: We aim to provide the illusion that remote participants are present in the *local* physical environment. We embed them into 3D space via a life-size transparent display without showing their remote scene background.

3. SYSTEM OVERVIEW

Our system implements the full teleconferencing pipeline, consisting of *acquisition*, *transmission*, and *display*. An overview is shown in Figure 2. As illustrated, different hardware options exist for the acquisition and display modules depending on the usage scenario. In order to create a 3D representation of the captured scene, we combine color and depth cameras to acquire video footage along with geometric information. To fuse and process the input data streams, we adjust the

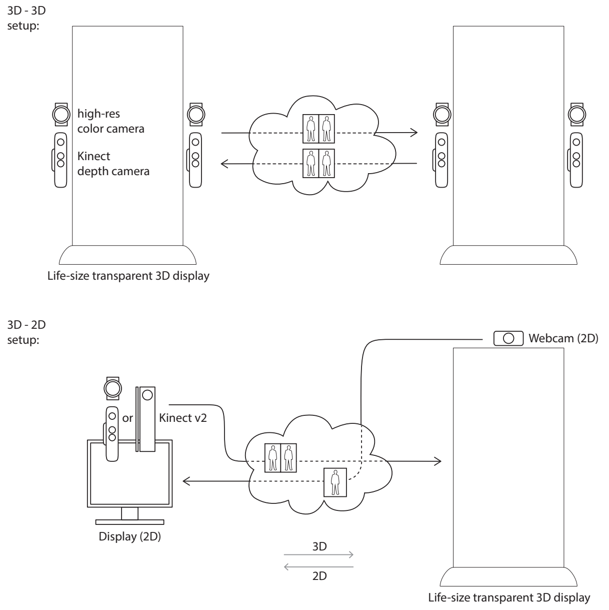

Figure 3: Proposed bidirectional video conferencing setups for different usage scenarios.

method by Kuster et al. [11] for efficient reconstruction of geometric and color data from multiple sensors. As a result, we obtain textured geometry, that can be rendered from novel viewpoints. This allows generating left/right stereo views for the 3D display from a virtual stereo camera. Next, we use WebRTC to transmit the rendered views over the Internet via web-browsers. Audio information is transmitted in the same manner. Finally, the remote participant is displayed on our life-size transparent 3D display.

To enable bidirectional communication, we combine an acquisition and a display module at each site. We propose two setups for immersive 3D video conferencing (Figure 3):

1. The 3D-3D setup enables symmetric 3D to 3D communication with a life-size transparent display at each site, equipped with hybrid (color+depth) cameras at each side of the display panel.

2. The 3D-2D setup enables a combination of 3D and 2D communication, were a 3D capture and 2D display module are connected to a 2D capture and 3D display module. This setup can be used if displaying a remote person in 3D is required at only one of the sites, for instance in a one-to-many teaching, speech, or presentation scenario.

Both systems allow unintrusive yet complete 3D acquisition. Users stand approximately 2m in front of the acquisition/display setup. During the 3D acquisition stage, two frontal stereo views of the captured scene are rendered. They correspond to the approximate eye positions of the remote user, which enables binocular parallax on the transparent display. Concretely, we place a virtual stereo camera at the center of the two depth cameras (3D-3D setup), or at the location of the single depth camera (3D-2D setup). The views sent to the remote location are rendered from the positions of this left/right virtual stereo camera.

In the following, we describe the individual components of our system in more detail.

4. COMPONENTS

4.1 Acquisition

We demonstrate several hardware options for acquisition (see Figure 2, left). All of them consist of off-the-shelf devices.

- The first acquisition option is built around our life-size display to enable a symmetric configuration where users are simultaneously captured and displayed in 3D at both sites. Combining camera views from two angles (at each side of the screen) ensures that we obtain a full frontal 3D reconstruction of the user's body with minimal occlusions. We mount a Kinect depth camera and a Point Grey Grasshopper2 color camera on each side of the display panel. Their respective resolutions are 640×480 and 1280×960 pixels. In principle, we could use the color camera that is built into the Kinect device. However, for our system we prefer the Grasshopper2 cameras for their higher resolution and better image quality. The two camera pairs are about 1.3m apart.

- The second option consists of a single unit from the above configuration, i.e., one Kinect depth camera and one Grasshopper2 camera placed centrally in front of the user. This option may be chosen for the asymmetric scenario (Figure 3, bottom).

- As an alternative frontal setup, the color and depth sensor of a single Kinect v2 device may be used. Their resolutions are 512×424 and 1920×1080 pixels, respectively.

- As a fourth option we enable 2D capture with a standard webcam. No virtual view rendering is performed in this case.

All cameras are intrinsically and extrinsically calibrated using a standard checkerboard-based approach [34] and bundle adjustment.

At the end of the acquisition stage, our goal is to render the captured scene from novel viewpoints corresponding to the eye positions of the remote user. However, especially in the 3D-3D setup, the acquired raw input geometries suffer from spatial as well as temporal noise (see Figure 4a and the accompanying video). Additional artifacts are due to misalignment of the color and depth sensors caused by calibration inaccuracies.

Therefore, to provide a pleasant viewing experience, these hybrid input data streams need to be carefully processed and combined into a common, consistent representation. Several methods have been developed to process and refine hybrid data [23, 12]. While they can provide visually appealing results for single hybrid cameras, it is unclear how they can be extended to multiple devices. Applying these methods on each hybrid camera independently tends to provide corrupted geometry due to inevitable misalignment (see experiments in [11]).

To address this problem, Kuster et al. proposed an approach based on temporal moving least squares (TMLS) surfaces [11]. TMLS allows to efficiently fuse geometric and

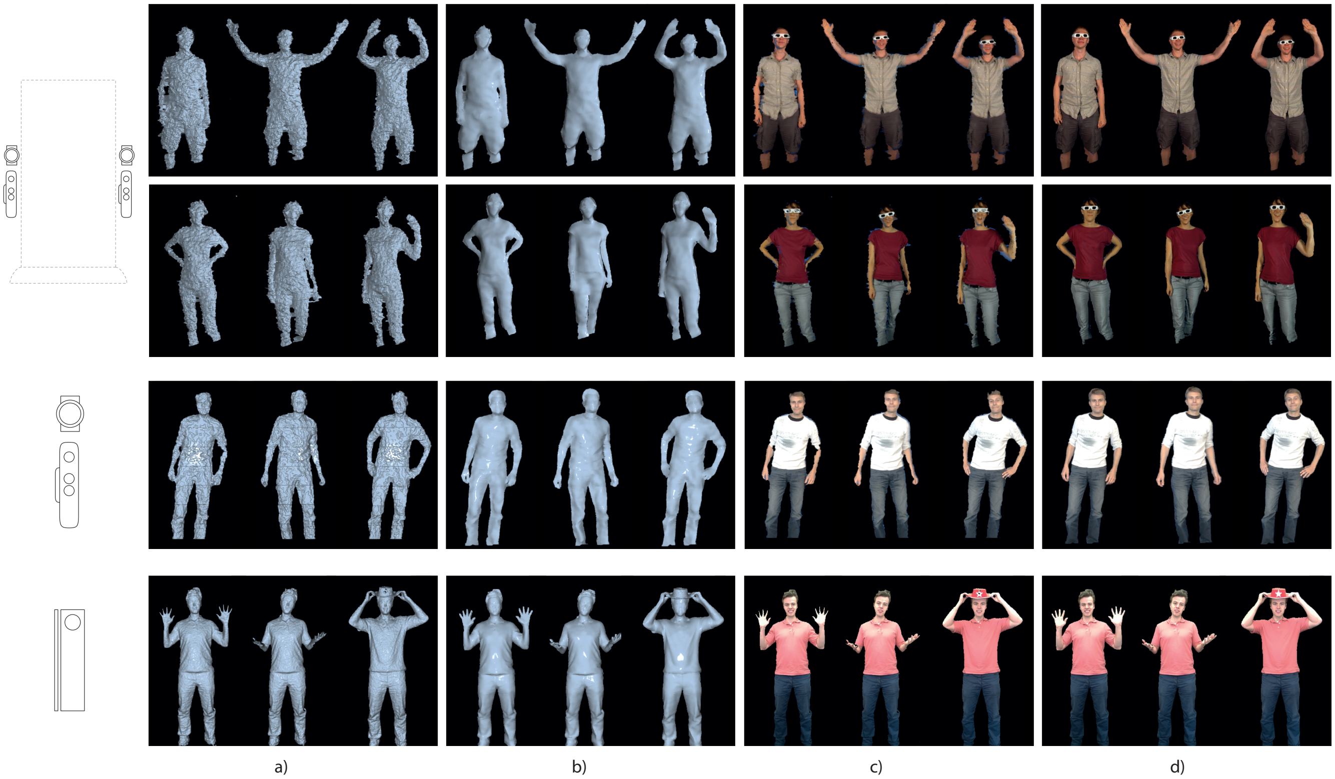

Figure 4: Processing and fusion of color and depth data with three capture setups. Note that the black background appears transparent on our lifesize display. First and second row: Two high-res color cameras and two Kinect depth cameras placed at the sides of the transparent display (virtual view rendered from the center). Third row: One high-res color camera and one Kinect depth camera placed centrally in front of the user. Fourth row: One central Kinect v2 sensor (color and depth). a) Raw geometry. b) Processed geometry. c) Raw geometry with texture. d) Processed geometry with texture. The processing is able to preserve details in the geometry, remove noise, provide more accurate silhouettes, and reduce texture bleeding. Please zoom in for better appreciation of the details.

color information from multiple hybrid sensors into a spatially and temporally consistent 3D representation. Furthermore, it is tolerant to noisy and incomplete input data. 3D reconstruction is performed by computing an implicit surface based on neighboring input 3D points, taking both normal and color information into account. The implicit surface can then be sampled by projecting points onto it. In our case, the projected points correspond to the input points from the depth maps. In this work, we use TMLS as a common framework for all our 3D capture setups, including the ones consisting of a single hybrid camera, for which we obtain comparable results to [23, 12].

Even though TMLS can be fully executed on the GPU, the original version proposed in [11] operates at 2-3fps for two hybrid cameras. Therefore, in order to meet the performance requirements of live teleconferencing, we adjust the original method and propose a truly interactive capture system (20-30fps on a standard desktop computer). The modifications are listed in the following.

- **Neighbor lookup.** The implicit TMLS surface at a given point is defined by the neighboring sample points in space and time. In the GPU implementation, this corresponds to a large number of texture lookups, constituting one of the main bottlenecks of the original TMLS system. For this reason, we select a subset (around 25%) of the neighbors to perform the lookups. As a consequence, some of the small-scale surface details may be smoothed out. However, this has no noticeable effect on the visual quality of the results, since in our target application the surface is always textured.

- **Rendering.** Instead of employing point-based rendering, which entails some computational overhead, we render the surface as an individual triangle mesh for each depth camera. After projecting the input points onto the computed implicit surface, [11] proposes to render them as splats, i.e., small disks that represent the new surface [35]. In contrast, we maintain the grid-based alignment of the 3D points from the original depth maps. This allows to convert the points from each depth camera into a mesh after processing, which makes rendering more straightforward, as the standard graphics pipeline is targeted and optimized mainly for meshes. As the fields of view of the depth cameras overlap, the surface may be rendered multiple times in these regions. However, the meshes are well aligned, since all vertices are samples of the same reconstructed MLS surface.

- **Segmentation.** In contrast to the boundary refinement approach described in [11], we simplify segmentation by assuming a known background. We perform frame differencing or chromakey-based segmentation, resulting in comparable segmentation quality at higher performance.

Results of our processing pipeline are available in Figure 4, as well as in the accompanying video. We show input and output frames from the three described options for 3D acquisition. The raw input geometries exhibit different levels of noise. The setup with two camera pairs mounted at the sides of the transparent display (Figure 4, first and second row) is the most challenging for processing and reconstruction. In this case, data streams from four sensors are combined. Additionally, the rendered virtual views are far away (approximately 0.65m) from the physical camera locations, and the data suffers from interference between the two Kinect depth sensors. When the user is captured frontally by a high-res color camera and Kinect depth sensor (Figure 4, third row), we can achieve higher quality results and preserve more details in the geometry. Here, a single image can be used to texture the entire geometry and the virtual views are close to the physical camera locations. Therefore, there are no visible seams in the texture, less distortion, and smaller occluded areas, which have to be filled in. Finally, the quality of the input geometry from the Kinect v2 has improved significantly compared the original Kinect sensor (Figure 4, fourth row). This further benefits our results. Overall, the rendered output views are dependent on the quality of the input data. However, in all cases our accelerated TMLS processing is able to improve both, the geometry and the texture of our reconstructions. In Figure 4, note the reduced noise in the geometry while small-scale details are preserved (Figure 4a vs. Figure 4b), the improved silhouettes, for example the hand regions, and the reduced texture bleeding (Figure 4c vs. Figure 4d, first and second row). In addition, the supplementary video shows how TMLS is able to alleviate temporal flickering.

4.2 Transmission

A wide number of transmission approaches exist, depending on the 3D data representation: stereo images, images and depth maps, images and 3D point clouds, or meshes. Specialized coding and compression techniques have been proposed, such as [7, 33] for dynamic meshes, the popular H.264 for video encoding [32], and its extensions to stereo and multiview videos plus depth [31, 4].

We choose to transmit stereo pairs, since this approach has low bandwidth requirements and lets us take advantage of existing efficient compression techniques and transmission protocols for images. In addition, we can make use of established frameworks for communication over the Internet, which support synchronized transmission of 2D video and audio data.

For transmission, the rendered stereo views from the acquisition module are combined side-by-side into a single image. Each teleconferencing site is equipped with a microphone and speakers. We use WebRTC [1] to transmit video and audio streams over the Internet. Besides a web-browser on each end, our transmission module consists of a server, which acts as the common connection point for the handshake between the two parties. Each acquisition or display module registers with the server via a website, specifying if they support stereo or single image data. After the appropriate modules have been connected by a human operator, the data streams are sent peer-to-peer. If desired, one acquisition module can be connected with several displays.

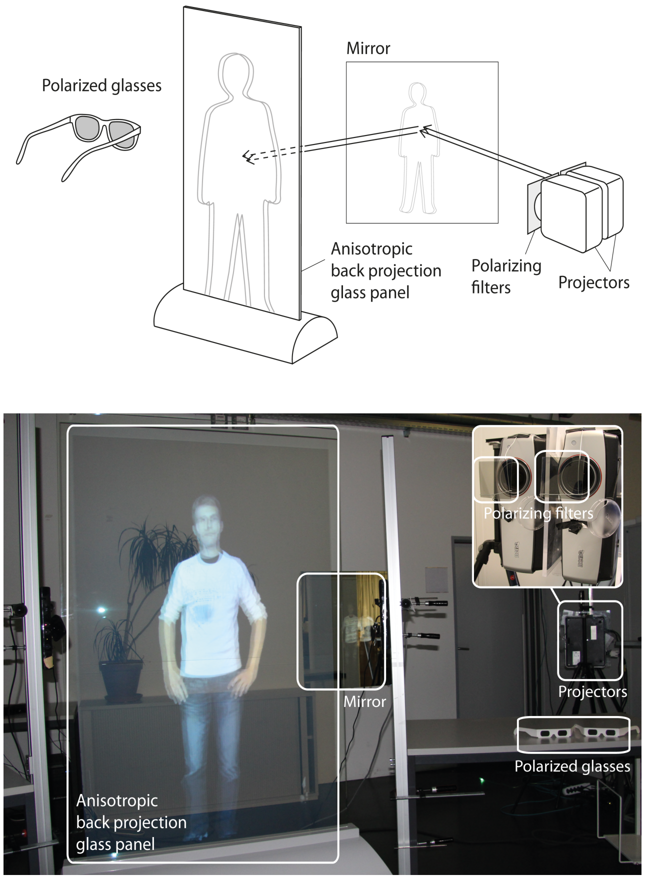

Figure 5: Our life size transparent display system. A pair of projectors with mounted linear polarizing filters project stereoscopic images onto an anisotropic back projection glass panel. A mirror is used for optical path folding to reduce the required space for the setup. Polarized glasses are used to perceive 3D.

Interfacing the transmission module is simple. On the sender side, image data is provided to the web-browser through the Microsoft DirectShow Framework. Concretely, the acquisition module is implemented to mimic a standard webcam, which can be selected as a video source in the browser when starting a WebRTC session. On the receiver end, images are displayed directly in the browser window.

4.3 Display

Few truly transparent display technologies are available. Transparent LCD and organic LED as deployed by companies such as Samsung and LG have a high percentage of light loss and would require controllable backlight. Other approaches as the one proposed by Lee and Hong [13] make use of a half mirror to overlay an opaque 3D display with the reality. However, the light loss of half mirrors is also relatively high, which dampers the immersive experience. A truly transparent display has been presented by Sun et al.

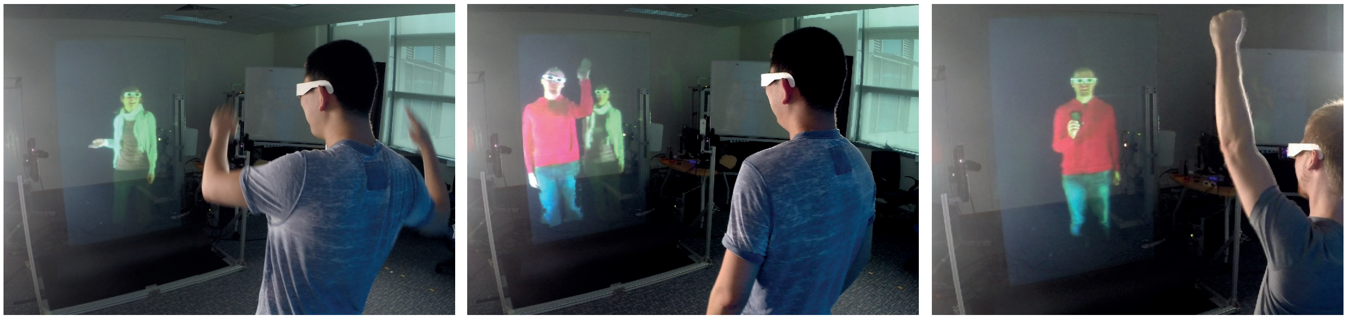

Figure 6: Bidirectional 3D video conferencing session between Zurich and Singapore (3D-3D setup). The system supports one or multiple users at each site.

[26]. However, it is only available in 2D. They use phosphor embedded in a glass, which is activated by an UV projector creating a self emissive transparent 2D display.

Another direction of displays, based on thin plastic films, are based on the Pepper's ghost illusion technique [25], a century old trick used in theatres to fade in and fade out pictures onto a stage. Several companies such as Hologram USA and MusionCanada offer high-end display solutions, which are based on this illusion technique. They mostly aim for live performances on large stages. While these state-of-the-art systems are able to create high fidelity results, they do not provide a sense of 3D in their current configuration.

As shown in Figure 5, in our work we deploy an alternative approach using an anisotropic back projection foil in combination with a polarized stereo projection setup [22]. The panel with embedded foil was obtained from Vision Optics GmbH. The life-size back projection screen measures 1.2m (width) × 2.1m (height). It selectively diffuses light incident from a very specific angle (in our configuration 38° to the surface normal) while preserving its polarization. This makes the screen robust against environment light as it can pass the screen unhampered when coming from a direction other than the selection angle, which results in a high transparency.

We use a pair of BenQ SH910 projectors with mounted linear polarizing filters. This creates two bright FullHD images on the transparent screen, each with different polarization. Thus, when viewed with corresponding polarized glasses, the image separation can be used to provide stereoscopic content. Transparency can be achieved by projecting black (no light), allowing the silhouette of the remote particpant to be seamlessly integrated into the local environment.

We use a mirror for optical path folding to reduce the required space of the setup. Simple homographies are used for calibration, to rectify and align the images of the left and the right eye on the projection surface [21]. The system is programmed to display either top/down or side-by-side stereo content received over a web browser using a HTML5 enabled webGL web page.

5. LIVE 3D VIDEO CONFERENCING

We have built two instances of the combined 3D acquisition and 3D life-size display setups (Figure 3, top): one in Zurich, and one in Singapore. Figure 6 and the supplementary video show live 3D video conferencing sessions between the two sites. Both, the acquisition and display modules run on standard desktop computers (Intel Core i7-2600K 3.40 GHz, 16 GB RAM, GeForce GTX 680). The transmitted stereo pairs have a resolution of 1080×960 pixels in total. We measured a network latency of 1.4s (round-trip) with a network bandwidth of 0.35Mbps and a framerate of 20fps. Note that these measurements were taken over an encrypted (VPN) connection.

In our second setup (Figure 3, bottom), we combine 2D and 3D video conferencing. Views of the transparent display in this scenario can be seen in Figure 1 and the supplementary video. This asymmetric system can for instance be used in a one-to-many teaching or speech scenario.

Note that our system is easy to use: Once the appropriate acquisition/display modules are connected, users can directly walk up to the setup and start interacting with remote participants. We do not require any initial calibration, user-dependent parameters, or learning. Furthermore, we do not assume human body silhouettes a priori. Therefore, one or multiple persons, but also any other objects can be captured and displayed.

We have presented our system to various groups and individuals, including 150 participants at a demo day. The informal feedback we received was positive. Users enjoyed the life-size 3D representation, especially when the remote participant virtually walks in and appears on the display. Previously, we have tested the same technology on a smaller scale display, showing only the upper body of the participant. The life-size version better conveyed body language and allowed for more natural user interaction.

6. LIMITATIONS AND FUTURE WORK

Our current display requires users to wear passive 3D glasses. Compared to active systems, it does not need synchronization devices and bulky, more expensive active glasses. However, passive 3D glasses, like their active counterparts, still obscure the eyes. This may impede natural communication and affect the sense of immersion for some applications. Therefore, we are actively investigating solutions, which do not require wearing glasses.

Our system allows multiple users at each site, and in the current version, they all see the same 3D view. A direction for future work would be to extend our display component to stereoscopic multi-user transparent displays in order to provide individual views to each user based on their position and eye gaze [20][10].

Our entire system runs at full HD resolution, and while our results are visually attractive, the capabilities of the human eye are much higher. To catch up with human eye resolution and thus bring realism to a higher level, we are looking forward to next-generation cameras and displays with 8K or even higher resolution. Transmission of such high-resolution data will require wider bandwidth capabilities, such as Internet2 [29], in combination with novel video codecs.

Another development worth investigating are recent wearable see-through displays with augmented reality features. This includes commercial products such as Google Glass, or Microsoft's Holographic Goggles, or devices that are still in research stage [15][19]. Since these glasses are transparent, the user's eyes are clearly visible, which could achieve better eye contact, and provide a brighter view of the local environment.

7. CONCLUSION

In this paper we propose a full-fledged bidirectional communication system, consisting of acquisition, transmission, and display modules. The core of our system is a life-size transparent 3D display and the corresponding software components for real-time 3D data acquisition, processing, and rendering. This is the key to realizing the concept of "being here", i.e., to give the user the illusion that the remote communication partner is present in the local environment. We built two prototypes of our system in Zurich and Singapore, enabling a novel and immersive video conferencing experience with real-time performance across a large geographic distance.

8. ACKNOWLEDGMENTS

This research, which is carried out in part at BeingThere Centre, is supported in part by the Singapore National Research Foundation under its International Research Centre@Singapore Funding Initiative and administered by the IDM Programme Office.

9. REFERENCES

[1] Web real-time communication (WebRTC), 2011.

[2] T. Balogh and P. T. Kovács. Real-time 3D light field transmission. In *SPIE Photonics Europe*, pages 772406–772406. Int. Soc. for Optics and Photonics, 2010.

[3] S. Beck, A. Kunert, A. Kulik, and B. Froehlich. Immersive group-to-group telepresence. *IEEE TVCG*, 19(4):616–625, 2013.

[4] Y. Chen, M. M. Hannuksela, T. Suzuki, and S. Hattori. Overview of the MVC + D 3D video coding standard. *J. Visual Communication and Image Representation*, 25(4):679–688, 2014.

[5] H. Fuchs, G. Bishop, K. Arthur, L. McMillan, R. Bajcsy, S. Lee, H. Farid, and T. Kanade. Virtual space teleconferencing using a sea of cameras. In *Proc. First Int. Conf. on Medical Robotics and Computer Assisted Surgery*, volume 26, pages 161–167, 1994.

[6] M. Gross, S. Würmlin, M. Näf, E. Lamboray, C. P. Spagno, A. M. Kunz, E. Koller-Meier, T. Svoboda, L. Van Gool, S. Lang, K. Strehlke, A. V. Moere, and O. G. Staadt. blue-c: a spatially immersive display and 3D video portal for telepresence. *ACM Trans. Graphics (Proc. SIGGRAPH)*, 22(3):819–827, 2003.

[7] S.-R. Han, T. Yamasaki, and K. Aizawa. Time-varying mesh compression using an extended block matching algorithm. *IEEE Trans. Circuits Syst. Video Techn.*, 17(11):1506–1518, 2007.

[8] A. Jones, M. Lang, G. Fyffe, X. Yu, J. Busch, I. McDowall, M. Bolas, and P. Debevec. Achieving eye contact in a one-to-many 3D video teleconferencing system. In *ACM Trans. Graphics (Proc. SIGGRAPH)*, volume 28, page 64, 2009.

[9] T. Kanade, P. Rander, and P. Narayanan. Virtualized reality: Constructing virtual worlds from real scenes. *IEEE Multimedia*, 4(1):34–47, 1997.

[10] A. Kulik, A. Kunert, S. Beck, R. Reichel, R. Blach, A. Zink, and B. Froehlich. C1x6: a stereoscopic six-user display for co-located collaboration in shared virtual environments. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 30(6):188, 2011.

[11] C. Kuster, J.-C. Bazin, A. C. Öztireli, T. Deng, T. Martin, T. Popa, and M. Gross. Spatio-temporal geometry fusion for multiple hybrid cameras using moving least squares surfaces. *CGF (Eurographics)*, 33(2):1–10, 2014.

[12] M. Lang, O. Wang, T. Aydin, A. Smolic, and M. Gross. Practical temporal consistency for image-based graphics applications. *ACM Trans. Graphics (Proc. SIGGRAPH)*, 31(4):34, 2012.

[13] B. Lee and J. Hong. Transparent 3d display for augmented reality. In *Photonics Asia*, pages 855602–855602. Int. Soc. for Optics and Photonics, 2012.

[14] A. Maimone and H. Fuchs. Encumbrance-free telepresence system with real-time 3D capture and display using commodity depth cameras. In *IEEE/ACM ISMAR*, pages 137–146, 2011.

[15] A. Maimone, X. Yang, N. Dierk, A. State, M. Dou, and H. Fuchs. General-purpose telepresence with head-worn optical see-through displays and projector-based lighting. In *IEEE VR*, pages 23–26, 2013.

[16] W. Matusik and H. Pfister. 3D TV: a scalable system for real-time acquisition, transmission, and autostereoscopic display of dynamic scenes. In *ACM Trans. Graphics (Proc. SIGGRAPH)*, volume 23, pages 814–824, 2004.

[17] D. Nguyen and J. Canny. Multiview: improving trust in group video conferencing through spatial faithfulness. In *Proceedings of the SIGCHI conf. on Human factors in computing systems*, pages 1465–1474. ACM, 2007.

[18] V. Nguyen, J. Lu, S. Zhao, D. Jones, and M. Do. Teleimmersive audio-visual communication using commodity hardware [applications corner]. *IEEE Signal Processing Magazine*, 31(6):118–136, 2014.

[19] T. Oskiper, M. Sizintsev, V. Branzoi, S. Samarasekera, and R. Kumar. Augmented reality binoculars. In *IEEE/ACM ISMAR*, pages 219–228, 2013.

[20] T. Peterka, R. Kooima, D. Sandin, A. Johnson, J. Leigh, and T. DeFanti. Advances in the dynallax solid-state dynamic parallax barrier autostereoscopic visualization display system. *IEEE TVCG*, 14(3):487–499, 2008.

[21] N. Ranieri and M. Gross. Vision-based calibration of parallax barrier displays. In *IS&T/SPIE Electronic Imaging*, pages 90111D–90111D. Int. Society for

Optics and Photonics, 2014.

[22] N. Ranieri, H. Seifert, and M. Gross. Transparent stereoscopic display and application. In *IS&T/SPIE Electronic Imaging*, pages 90110P–90110P. Int. Society for Optics and Photonics, 2014.

[23] C. Richardt, C. Stoll, N. A. Dodgson, H.-P. Seidel, and C. Theobalt. Coherent spatiotemporal filtering, upsampling and rendering of RGBZ videos. *CGF (Eurographics)*, 31(2):247–256, 2012.

[24] O. Schreer, I. Feldmann, N. Atzpadin, P. Eisert, P. Kauff, and H. Belt. 3dpresence -a system concept for multi-user and multi-party immersive 3d videoconferencing. In *5th Europ. Conf. on Visual Media Production (CVMP)*, pages 1–8, 2008.

[25] J. Steinmeyer. *Hiding the Elephant: How Magicians Invented the Impossible and Learned to Disappear*. Carroll & Graf Publishers, 2003.

[26] T. Sun, S. Wu, and B. Cheng. 54.4: Novel transparent emissive display on optic-clear phosphor screen. In *SID Symposium Digest of Technical Papers*, volume 44, pages 755–758. Wiley Online Library, 2013.

[27] Y. Taguchi, T. Koike, K. Takahashi, and T. Naemura. TransCAIP: a live 3D TV system using a camera array and an integral photography display with interactive control of viewing parameters. *IEEE TVCG*, 15(5):841–852, 2009.

[28] E. Tola, C. Zhang, Q. Cai, and Z. Zhang. Virtual view generation with a hybrid camera array. *CVLAB-Report-2009-001 (EPFL)*, 2009.

[29] H. Towles, W. Chen, R. Yang, S. Kum, H. Fuchs, N. Kelshikar, J. Mulligan, K. Daniilidis, L. Holden, B. Zeleznik, A. Sadagic, and J. Lanier. 3D tele-collaboration over Internet2. In *Int. Workshop on Immersive Telepresence*, 2002.

[30] R. Vasudevan, G. Kurillo, E. Lobaton, T. Bernardin, O. Kreylos, R. Bajcsy, and K. Nahrstedt. High-quality visualization for geographically distributed 3-D teleimmersive applications. *IEEE Trans. on Multimedia*, 13(3):573–584, 2011.

[31] A. Vetro, T. Wiegand, and G. J. Sullivan. Overview of the stereo and multiview video coding extensions of the H.264/MPEG-4 AVC standard. *Proc. of the IEEE*, 99(4):626–642, 2011.

[32] T. Wiegand, G. J. Sullivan, G. Bjntegaard, and A. Luthra. Overview of the H.264/AVC video coding standard. *IEEE Trans. Circuits Syst. Video Techn.*, 13(7):560–576, 2003.

[33] J. Yang, C. Kim, and S. Lee. Semi-regular representation and progressive compression of 3-d dynamic mesh sequences. *IEEE TIP*, 15(9):2531–2544, 2006.

[34] Z. Zhang. Flexible camera calibration by viewing a plane from unknown orientations. In *ICCV*, volume 1, pages 666–673, 1999.

[35] M. Zwicker, H.-P. Pfister, J. Van Baar, and M. Gross. Surface splatting. In *Proceedings of the 28th annual conf. on Computer graphics and interactive techniques*, pages 371–378. ACM, 2001.