TOWARDS NEXT GENERATION 3D TELECONFERENCING SYSTEMS

C. Kuster[1], *N. Ranieri*[1], *Agustina*[2], *H. Zimmer*[1], *J.C. Bazin*[1], *C. Sun*[3], *T. Popa*[1], *M. Gross*[1]

[1]IVC, ETHZ, Switzerland, [2]BeingThere Centre, IMI, NTU, Singapore, [3]SCE, NTU, Singapore

ABSTRACT

Teleconferencing is becoming more and more important and popular in today's society and is mostly accomplished using 2D video conferencing systems. However, we believe there is a lot of room for improving the communication experience: one crucial aspect is to add 3D information, but also freeing the user from sitting in front of a computer. With these improvements, we aim at eventually creating a fully immersive 3D telepresence system that might improve the way we communicate over long distances. In this paper we review and analyze existing technology to achieve this goal and present a proof-of-concept, but fully functional prototype.

Index Terms — 3D teleconferencing, 3D display, 3D acquisition

1. INTRODUCTION

In our globalized world, people want to communicate with persons far away. Consequently, advanced communication and remote collaboration are a central pillar of our modern society, affecting both our private and work life. Classical means of remote communication are telephones and more recently video conferencing systems like Skype. While video conferencing enhanced the communication experience by adding visual information, it still suffers from a number of shortcomings. First, users are required to sit in front of a computer or at least carry a smartphone or similar device. So it is tedious to roam around freely and using gestures is almost impossible although they are a vital part of human communication. Second, users do not make eye contact as they look into the screens instead of the cameras capturing them. Third, today's video conferencing is typically restricted to capture only the upper part of the human body and also does not provide any 3D information, but just 2D video.

To solve above shortcomings, we aim to develop a communication system that seamlessly integrates the remote person in the environment of the other participants, resulting in a fully immersive 3D telepresence experience. To achieve this, we envision a mobile robot platform with a transparent auto-stereoscopic display and a 3D capture device; see Figure 1b. The latter captures full-body 3D information of the users. After applying a gaze correction algorithm to ensure eye contact, the 3D information is transmitted to the other platforms where it is eventually visualized in 3D on the autostereoscopic displays. This system would solve all mentioned problems as it is mobile, provides full body 3D information and also ensures eye contact. We believe that such a system has the potential to change the way people communicate. However, there are several perceptual and technical issues that need to be tackled. In this paper, we present a short survey and analysis of the available technology to meet the requirements for our 3D telepresence system (Section 2) and also present a proof-of-concept, but fully functional prototype implementation (Section 3).

2. COMPONENTS

The core components of teleconferencing systems are *acquisition*, *transmission* and *display*. However, the traditional 2D display devices that are currently deployed in commercial teleconferencing systems limit the perceptual realism of a scene. In contrast, a 3D display device would yield a more immersive and realistic experience. An important challenge is that 3D displays generally require rendering the scene from multiple viewpoints. Therefore, in tandem with a 3D display, an appropriate acquisition system is required to capture more complex information such as the scene geometry, in addition to video footage. Finally, the acquired data has to be transmitted across long distances to the remote location which also poses challenges as the acquired 3D data is typically larger than standard 2D video streams and also more sensitive to compression artifacts. In the following subsections we provide an in-depth analysis of each component by providing a short summary of existing works and a description of our approach.

2.1. Acquisiton

Our goal is to capture a complete three dimensional representation of a dynamic scene. This allows rendering it from arbitrary viewpoints, for example to create imagery for a 3D display. Where for classic (binocluar) stereo two images are necessary, a typical automultiscopic display shows several views from slightly different angles. One way to generate these views is to use a large number of up to 64 video cameras [1, 2]. However, these camera systems need to be carefully calibrated and synchronized and are of course also very expensive. Furthermore, considerable computing power and memory is necessary to process the large amount of data. An alternative approach that requires fewer cameras but more sophisticated processing is to use captured 3D geometry to render a scene from novel viewpoints, which requires to fill in the areas not captured by the cameras in an appropriate way. In the Blue-C project [3], a proxy 3D geometry from the visual hull is used to render a person from different viewpoints. Waschbüsch et al. [4] obtain 3D information in form of depth maps from structured light-assisted stereo matching, whereas Knoblauch et al. [5] use a fast GPU stereo algorithm. The main challenge in these systems is to compute scene geometry from images in real-time. Recently, depth cameras such as the Microsoft Kinect have made this task straightforward. The systems by Bogomjakov et al. [6], Tola et al. [7] and Maimone et al. [8] use one or more depth cameras to capture 3D representations of a person. However, the raw geometry from any depth camera-based system will be noisy and incomplete (see Figure 2). This can create visually disturbing artifacts if the data is displayed on a 3D screen. Therefore, it is necessary to post-process the depth maps in order to obtain high quality results.

In our prototype, we use a hybrid setup with one Microsoft Kinect depth camera and one Point Grey Grasshopper camera that

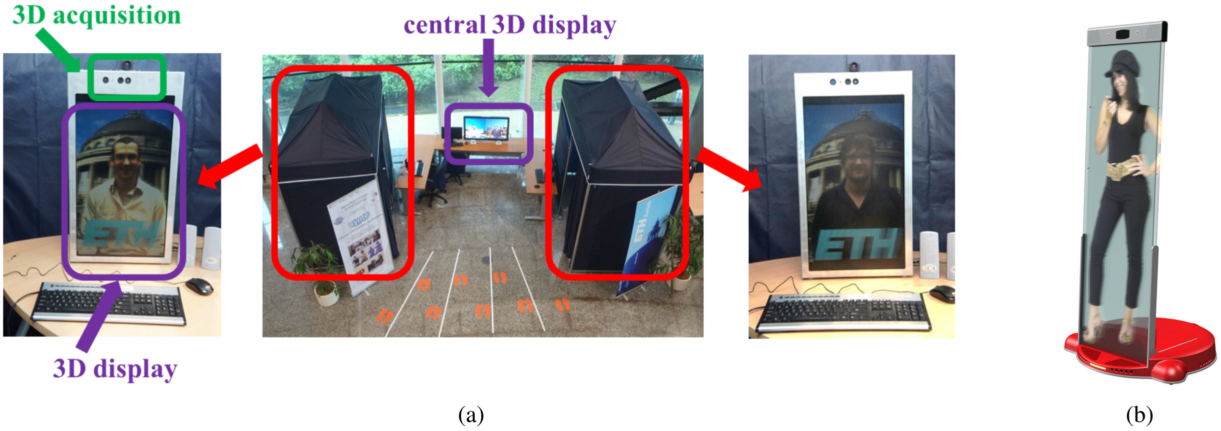


Figure 1: (a) A prototype implementation of a static, room-based 3D telepresence system. (b) Sketch of a mobile platform.

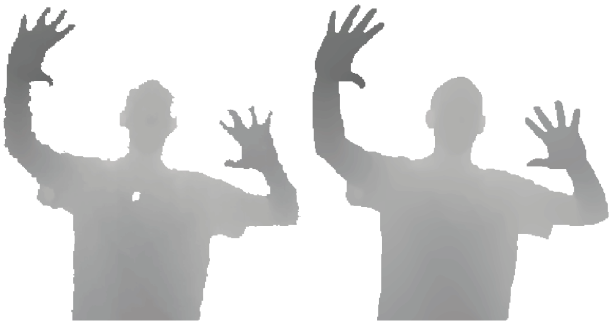


Figure 2: Depth map before and after processing

replaces the low quality RGB camera of the Kinect. Following [9], we perform denoising, outlier removal and foreground/background segmentation on the depth data in real-time, where the segmentation is used to mask out the captured person for further processing. Our algorithms all incorporate both color and depth information to align depth discontinuities of the output geometry with color edges in the video footage. An example of a depth map before and after processing can be seen in Figure 2.

2.2. Transmission

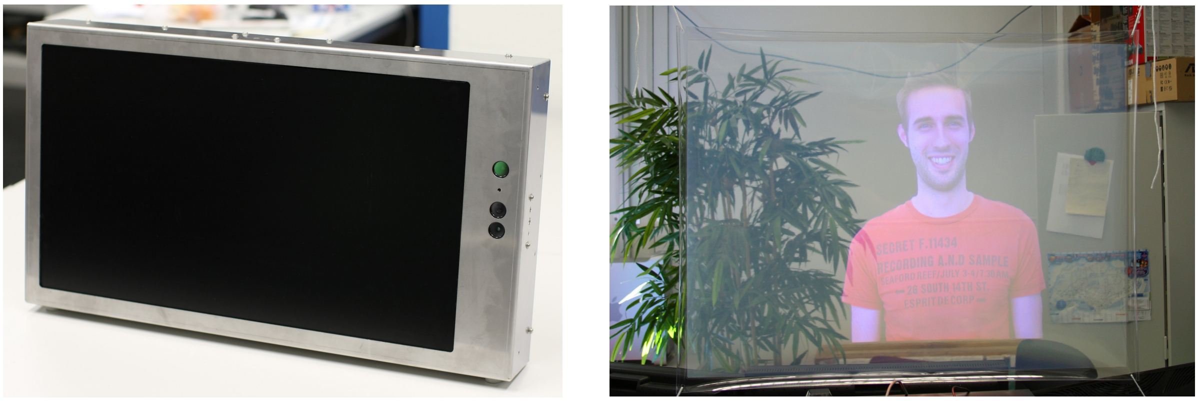
Different teleconferencing systems adopt different approaches in transmitting their data over the network, depending on the requirements of the acquisition and display components, as well as on the capabilities of the network infrastructure. In 2D teleconferencing systems, the 2D video is compressed into a standardized media format (MPEG-4, H.264, etc.) for transmission over low/medium bandwidth networks. In 3D teleconferencing systems, a high bandwidth network is commonly assumed and 3D video data may be transmitted in various formats such as 3D-point clouds [3, 10] or depth maps [11]. In case of high bandwidth availability, systems like [11] transmit the data uncompressed, thus requiring hundreds to thousands Mbps bandwidth. Other systems attempted to reduce the bandwidth requirements in various ways, e.g. by performing colour reduction, background pixel removal, and z-lib compression to the data before transmission [10], or transmitting only the different point-cloud information [3]. However, the resulting bandwidth requirement is still typically higher than the capabilities of commercial Internet environments.

We follow an approach similar to the NTII, which transmits the 3D video data as RGB colour images plus depth maps. But, we apply a standard H.264 2D video compression using the x264 library [12] before transmission. The colour image is in 24-bit RGB format and the associated depth map is stored in the R channel of a 24-bit RGB data. These data, which are in 1024x768 resolution, are combined into a single 2D image of resolution 1024x1536 for compression and then transmitted over RTP/UDP. This approach of transmitting the colour and depth data combined in a single 2D video stream frees us from the issue of synchronizing these two data, but at the expense of a higher bandwidth requirement due to the two additional channels for the depth map. The entire data transmission is performed following the standard H.323 protocol [13], which is the most popular real-time communication protocol, and making use of the open-source implementation [14]. Overall, we managed to significantly reduce the bandwidth requirement from 586 Mbps to less than 1 Mbps for each stream. This means our system is safely compatible with today's limited Internet bandwidth.

Finally, it is important to note that compression methods such as H.264 are designed to keep visual artifacts to a minimum in standard video footage. In our context of free-viewpoint 3D telepresence systems, it means that even small compression artifacts in the depth map may cause significant and visually disturbing changes in the geometry, especially at depth discontinuities. We suggest to overcome these artifacts by rendering the geometry from a novel viewpoint directly on the acquisition side and then transmitting the modified depth map to the remote rendering sites.

2.3. Display

A variety of 3D display technologies are available, each offering different advantages and drawbacks. Parallax barrier based systems use a barrier layer to selectively block colored pixels of a second layer into certain directions [15, 16]. Each subpixel is therefore only visible from a specific angle within the field of view. Correct addressing of these subpixels allows to multiplex different images into different directions. Different images can also be multiplexed using small lenticular lenses as shown by Lippmann [17]. This basic idea has been extended by various authors, an extensive overview on recent advances can be found in [18]. Volumetric displays do not multiplex different images into different directions but create a volume of controllable lightsources approximating the plenoptic function of a scene. There are numerous implementations, e.g. using a highspeed projector in combination with shutter screens [19] or a turning mirror [20]. A comprehen-



(a) A dual layer parallax barrier based automultiscopic display. Each layer consists of a LCD panel. The front layer is used as barrier, the back layer as image generator.

(b) Illustration of a stereoscopic projection on an anisotropic transparent backprojection foil. Content can be shown in 3D using nVidia 3D Vision and shutter glasses.

Figure 3: Two different implementations of 3D capable displays for teleconferencing systems.

sive survey on these techniques is available in [21]. In our work, two display prototypes have been designed and built, both targeting our final goal, a transparent autostereoscopic display.

Parallax barrier display. Our first display is a parallax barrier based automultiscopic display (Figure 3a) using two Acer HN274H liquid crystal layers with a spacing of 15mm in between. The front LCD with removed diffusers is used as black and white / transparent parallax barrier while the back layer provides the color images. Sixteen different views are rendered interleaved to the backlayer, and in each frame both barrier and images are shifted to regain spatial resolution similar to Kim et al. [22]. The layers work at 120Hz divided by three time-multiplexing steps and provide a perceived depth of approximately ±10cm.

Transparent stereoscopic screen. The second display uses a transparent anisotropic backprojection foil that diffuses incident light only from a certain angle. Such foils are optimized for a specific center of projection by gradually adapting the diffusing angle over the whole foil and thus are especially robust against disturbing environment light. Silhouette carved images can be realized by projecting black background color, as black regions in the projected image will be perceived as transparent. We use an Acer H5360 projector that supports nVidia 3D Vision [23] together with a Holographic-Optical Projection Screen (HOPS ®) foil in a 38° configuration [24]. Synchronized with shutter-glasses, the projector displays time-multiplexed stereo image pairs, providing classic stereoscopic imaging. With enabled nVidia 3D Vision, an arbitrary mesh can be rendered to display silhouette carved objects, overlaid with the real scene through the transparent projection foil (Figure 3b). Due to the fact that black regions in the projected image will be perceived as transparent, this approach is limited to bright scenes. However, if the dark patch is small enough, it will still be perceived as black, as human perception focuses on the surrounding brighter colors. Also, the HOPS foil shows a strong fall-off in brightness outside the field of view of ± 35° as it does not diffuse but redirects the incident light.

As further steps towards the final goal of transparent autostereoscopy, eye-tracking [25] will replace both the shutter-glasses as well as the need for rendering more than two views as done in the automultiscopic approach. Furthermore, a transparent optical element will be used for multiplexing the stereoscopic image pair.

3. PROTOTYPE

In this section, we present a proof-of-concept prototype for 3D telepresence/teleconferencing. We built three instances of this prototype and deployed them in the BeingThere Centre laboratories at ETH Zurich and NTU Singapore. This way we could test and experiment with communication over very long distances. Figure 1a shows a set-up with two instances (inside the tents) placed side by side and a central screen (Philips automultiscopic 3D screen) that displays the 3D views of the users/participants of each tent to the audience. Each system is equipped with a 3D acquisition unit (Section 2.1), a 3D display (Section 2.3), microphone and speakers. The data are transmitted between the sites using the technology described in Section 2.2. The 3D screen displays the counterpart participant located at the other site and also the participant at the current site (mirror effect). We also provide a feature that detects the static background behind the participant and replaces it by any other 3D background images.

Our current prototype is composed of the following processing systems: one computer for each acquisition unit to process the 3D data (segmentation, filtering, etc.), one computer for each display unit to render the views for the 3D display, plus an optional computer to drive the central screen, if any. We believe that with an optimized implementation and adapted hardware (especially graphics card) one single computer would suffice per site.

Our tests over Internet from Zurich to Singapore which are more than 10,000km apart, revealed a net performance of $10 - 15$ frames per second, given our acquisition system running at 15Hz. Even with four concurrent 3D video streams (two streams in each direction), we obtained satisfying results in terms of data transmission, end-to-end latency, frame rate and visual appearance. Due to the low bandwidth requirement, our system does not assume any high-bandwidth networking infrastructure, contrary to many 3D telepresence systems, giving us the advantage to appeal a wider set of potential users.

We have also presented our prototype to a large audience and received a very positive feedback. One or several persons could enter each tent to test our prototype and they could see the participants situated inside the other tent, and also themselves, in 3D. The participants especially appreciated that several persons can use the system simultaneously (i.e. our system does not assume single users, contrary to standard head-tracked 3D display), our prototype is based on off-the-shelf hardware (and thus is close to the consumer level) and is easy to use (no user setting nor specific calibration is needed).

4. CONCLUSION AND FUTURE WORK

We presented a short technological survey of the main components of next generation 3D teleconferencing systems as well as a proof-of-concept prototype that is currently deployed in the BeingThere Center laboratories at ETH Zurich and NTU Singapore.

In future, we plan to mount our 3D teleconferencing system on a mobile robotic platform. This alleviates the problem of having to sit in front of a computer or the tedious carrying of a smartphone-like device which prevents gesturing. Developing this mobile platform poses several technical challenges like reducing power consumption, foreground segmentation with a mobile camera and cloud processing to reduce computations on the platform.

5. 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.

6. REFERENCES

[1] W. Matusik and H. Pfister, "3D TV: a scalable system for real-time acquisition, transmission, and autostereoscopic display of dynamic scenes," *ACM Transactions on Graphics*, vol. 23, no. 3, pp. 814–824, 2004.

[2] 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 Transactions on Visualization and Computer Graphics*, pp. 841–852, 2009.

[3] M. Gross, S. Würmlin, M. Naef, E. Lamboray, C. Spagno, A. Kunz, E. Koller-Meier, T. Svoboda, L. Van Gool, S. Lang, K. Strehlke, A. Vande Moere, and O. Staadt, "Blue-C: a spatially immersive display and 3D video portal for telepresence," in *SIGGRAPH*, 2003, pp. 819–827.

[4] M. Waschbüsch, S. Würmlin, D. Cotting, and M. Gross, "Point-sampled 3D video of real-world scenes," *Signal Processing: Image Communication*, vol. 22, no. 2, pp. 203–216, 2007.

[5] D. Knoblauch and F. Kuester, "VirtualizeMe: interactive model reconstruction from stereo video streams," in *Proceedings of ACM Symposium on Virtual Reality Software and Technology (VRST)*, 2008, pp. 193–196.

[6] A. Bogomjakov and C. Gotsman, "Reduced depth and visual hulls of complex 3D scenes," *Computer Graphics Forum*, vol. 27, no. 2, pp. 175–182, 2008.

[7] E. Tola, C. Zhang, Q. Cai, and Z. Zhang, "Virtual View Generation with a Hybrid Camera Array," *Tech. report CVLABREPORT*, 2009.

[8] A. Maimone and H. Fuchs, "Encumbrance-free telepresence system with real-time 3D capture and display using commodity depth cameras," in *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2011, pp. 137–146.

[9] C. Kuster, T. Popa, C. Zach, C. Gotsman, and M. Gross, "Freecam: A hybrid camera system for interactive free-viewpoint video," in *Proceedings of Vision, Modeling, and Visualization (VMV)*, 2011.

[10] Z. Yang, K. Nahrstedt, Y. Cui, B. Yu, J. Liang, S. Jung, and R. Bajscy, "TEEVE: The Next Generation Architecture for Tele-immersive Environment," in *IEEE International Symposium on Multimedia (ISM)*, 2005, pp. 112–119.

[11] 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 *International Workshop on Immersive Telepresence*, 2002.

[12] "x264 library," http://www.videolan.org/developers/x264.html.

[13] International Telecommunication Union, "H.323 : Packet-based multimedia communications systems," http://www.itu.int/rec/T-REC-H.323/e.

[14] "H.323 Plus: the Standard in Open Source H.323," http://www.h323plus.org/.

[15] H. Isono, M. Yasuda, and H. Sasazawa, "Autostereoscopic 3-D display using LCD-generated parallax barrier," *Electronics and Communications in Japan (Part II: Electronics)*, vol. 76, no. 7, pp. 77–84, 1993.

[16] G. Ye, A. State, and H. Fuchs, "A practical multi-viewer tabletop autostereoscopic display," in *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2010, pp. 147–156.

[17] G. M. Lippmann, "La photographie integrale," *Comptes-Rendus*, vol. 146, pp. 446–451, 1908.

[18] Y. Kim, K. Hong, and B. Lee, "Recent researches based on integral imaging display method," *3D Research*, vol. 1, pp. 17–27, 2010.

[19] A. Sullivan, "DepthCube solid-state 3D volumetric display," *SPIE Stereoscopic Displays and Virtual Reality Systems*, vol. 5291, no. 1, pp. 279–284, 2004.

[20] A. Jones, I. McDowall, H. Yamada, M. Bolas, and P. Debevec, "Rendering for an interactive 360 light field display," *ACM Transactions on Graphics*, vol. 26, no. 3, pp. 40:1–40:10, 2007.

[21] G. Favalora, "Volumetric 3D displays and application infrastructure," *IEEE Computer*, vol. 38, no. 8, pp. 37–44, 2005.

[22] Y. Kim, J. Kim, J.-M. Kang, J.-H. Jung, H. Choi, and B. Lee, "Point light source integral imaging with improved resolution and viewing angle by the use of electrically movable pinhole array," *Optics Express*, vol. 15, no. 26, pp. 18253–18267, 2007.

[23] "nVidia 3D Vision," http://www.nvidia.com/object/3d-vision-main.html.

[24] "Vision Optics HOPS," http://www.visionoptics.de/index.php?id=18.

[25] K. Perlin, S. Paxia, and J. S. Kollin, "An autostereoscopic display," in *SIGGRAPH*, 2000, pp. 319–326.