Disambiguation of Horizontal Direction for Video Conference Systems

Mabel Mengzi Zhang, Seth Rotkin, and Jürgen P. Schulze

University of California San Diego
9500 Gilman Dr, La Jolla, CA 92093
{mabel.m.zhang,sethrotkin}@gmail.com, jschulze@ucsd.edu

Abstract. All existing video conferencing systems which support more than two sites and more than one user at each site suffer from directional ambiguity: not only is it generally impossible for two remote users to look each other in the eyes, but even just horizontal directionality is not preserved. Under horizontal directionality we understand that the direction of the users' gaze or pointing fingers does not match what the other participants perceive. We present a video tele-conferencing concept, which, by combining existing software and hardware technologies, achieves horizontal directionality for multiple sites and participants at each site. Our solution involves multiple cameras, as well as large stereo or multi-view display walls at each site. Because building a physical prototype of our proposed system would have been fiscally impossible for us, we instead built a prototype for our virtual realit CAVE. In this publication we report on our experiences and findings with this prototype.

1 Introduction

Teleconferencing using distributed virtual reality (VR) as opposed to traditional 2D video based tele-conferencing has repeatedly been shown to have the potential to be more realistic because of the more natural interactions 3D environments allow [7]. The reason why VR can work better than 2D video is that it can allow realistic eye contact and directionality, which means that when a person turns to the image of another on the display device, that other person perceives correctly that he or she has been turned to, and everybody else in the tele-conference can see that those two participants are facing each other. In practice, none of these VR based approaches have been commercially successful, we hypothesize that this is because of the high level of software complexity involved, the level of achievable visual accuracy, and the inherent latency such approaches introduce into an already latency-prone application due to long distance network transfers.

Our approach started with the concept of the Cisco TelePresence systems, which are among the most sophisticated commercial tele-conferencing systems. We directly use camera images, which allows for realistic imagery at the remote site, creating a stronger notion of presence to feel that the participants share the same physical space, which is the primary goal of our and many prior tele-conferencing systems.

In this publication, we are going to summarize prior work next, then describe our approach, then we present our implementation, and finally discuss the insight we gained with our VR prototype.

G. Bebis et al. (Eds.): ISVC 2011, Part II, LNCS 6939, pp. 24–32, 2011.

2 Related Work

To our knowledge, this is the first virtual reality simulator of a multi-site videoconferencing system. It is modeled after the Cisco TelePresence 3000 system, which, along with the HP Halo system (now purchased by Polycom), could be considered the state of the art of video teleconferencing systems [14]. Both systems are steps in the direction we are exploring in this paper. They utilize multiple displays and cameras, specifically placing the cameras so that some viewing directionality is maintained. However, neither system can provide correct directionality for all participants because each participant is only captured by one camera, so that all participants see the same view of each participant. For example, if a participant looks directly at his or her dedicated camera, it will appear to all remote participants as if that person looks directly at them.

One of the most state-of-the-art approaches to video tele-conferencing is that of Maimone and Fuchs [8], which utilizes five Microsoft Kinect devices in order to achieve one-on-one video conferencing with a high level of realism, including eye contact. This work shows that re-creating eye contact is still a hot topic for video-conferencing, and that multiple cameras can be employed to achieve this effect. Our approach goes beyond this work in that it outlines a concept which scales to many participating sites with multiple users at each site. Also, we don't require the 3D reconstruction of the participants, which is also at the core of related work by Wu et al. [17] and Chu et al. [2], which adds latency, and by using regular cameras instead of the Kinect the depth range for how far the user can be from the camera is less limited. Of course, Maimone's approach has the benefit of creating a textured 3D model of each participant, which can be used for more than just the simulation of eye contact.

Probably the most related to our proposed system's video camera setup is that of HP's Coliseum system [1]. They propose installing multiple cameras around the user to be able to re-create multiple views of the user from different viewing angles. We achieve this by mounting the cameras in a linear array, which permits us to support more than one user, and our system adds multi-site directional consistency and a concept of how to display the imagery on a screen for multiple viewers.

Part of our work is based on ideas in a patent by Fields [3], which proposes that for multi-site video conferencing, the sites should virtually be arranged as the edges of an equilateral polygon with n edges, an n-gon. Fields also proposes using an array of cameras, an array of displays, and using camera interpolation (view synthesis, for instance Seitz and Dyer [15]) to create views of the participants from arbitrary directions. Our approach differs from Fields' in that we support multiple viewers at each site with correct viewing angles by utilizing multi-view displays such as 3D stereo or autostereoscopic displays. Also, Fields did not simulate or implement his approach but only described the idea.

Another related approach which uses VR technology is MASSIVE [5]. It uses a spatial model of interactions which extracts the "focus" and "nimbus" of the conferees, making the perception of conferees at the remote site relative to the local site's conferee positions and orientations. It also allows users to use audio, graphics, and text media through the network. Our system focuses on directionally-correct viewing, which MASSIVE does not address.

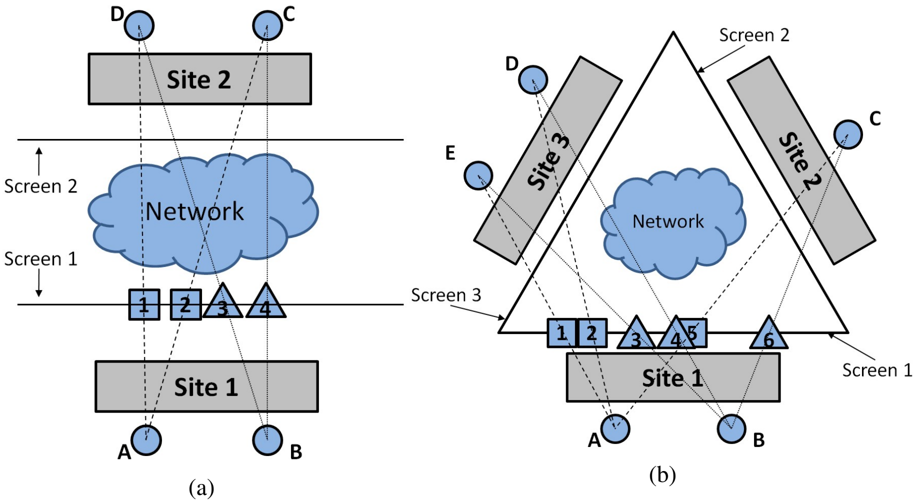

Fig. 1. Teleconferencing scenarios with (a) two and (b) three participating sites. The circles indicate users, the rectangles indicate the tables the users sit at. In front of each table is a screen. The dashed lines are the lines of sight for User A, the dotted lines are those for User B. The numbered squares are the intersections of the lines of sight for User A, the numbered triangles are those for User B. Those squares and triangles indicate where on the screen the users at Site 1 have to see the other participants for consistent lines of sight between any two users; these are also the locations where cameras would have to be installed.

A different family of approaches to achieve correct eye contact directions uses the real-time generation of 3D human models. Ohya [9,10] proposes the Virtual Space TELeconferencing (VISTEL) system. By generating a 3D model of the participants and modeling their motion on a screen, this approach is able to achieve motion parallax and correct eye contacts. Even though the interaction may be smooth with real-time generation, the 3D images look artificial compared to images captured by cameras.

Similarly, Kishino's three-site virtual teleconference system reconstructs an entire human body in 3D and uses two large screens to create the sensation of a common space [6]. Notably, it discusses virtual object manipulation, which is simulated in our system as 3D datasets that float in the middle of the conference space and may be manipulated with a navigation pointer. Kishino's and Yoshida's work uses gesture and speech recognition to detect the user's intent in building virtual objects, whereas we propose a system where the participants interact with datasets using a pointing device [18].

In the domain of head-mounted displays, there are also approaches that construct the virtual environment relative to the head position and orientation [4]. However, the interaction in such systems suffers from the disadvantages of head-mounted displays, such as limited resolution and field of view, noticeable lag on head motion, and the inability to directly see local conference participants.

3 The VR Tele-conferencing System Mock-up

The purpose of our VR prototype application is to verify the feasibility and effectiveness of our proposed tele-conferencing system. Our prototype is built as a C++ plug-in for the COllaborative VIsualization and Simulation Environment (COVISE) [13], which is based on OpenSceneGraph [11] as the underlying graphics library. It runs at interactive frame rates (20-30 frames per second) in our StarCAVE, a 5-sided, 15 HD screen, rear-projected CAVE (Cave Automatic Virtual Environment)-like system with an optical tracking system. All parameters and user interactions can be controlled from within the immersive environment with a wireless 3D wand. Our application allows the user to interactively study the impact of the various camera and display parameters and accurately try out different vantage points and viewing, as well as pointing angles.

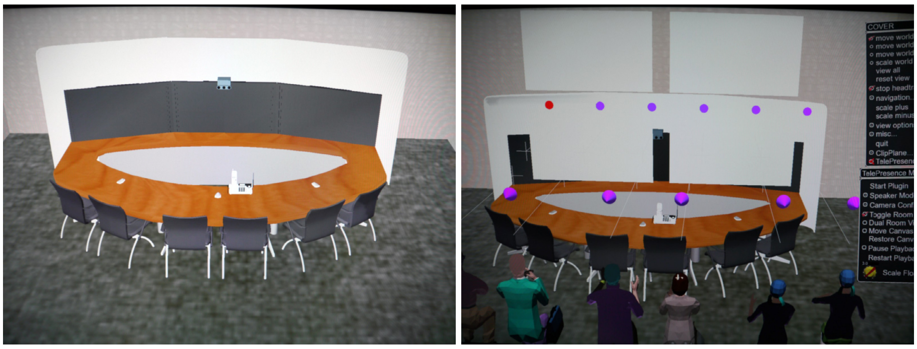

Fig. 2. Left: Our virtual conference room with only the static elements. Right: The full system with the dynamic objects, which includes movable screens, cameras, and participants.

Our virtual teleconferencing room draws ideas from the Cisco TelePresence 3000 system. That system consists of a room with three screens at the front and a half of an oval table, with room for up to six participants [16]. On top of the middle screen are three cameras with fixed focus. They each point to one pair of users at the table, and when viewed side-by-side the create a continuous image of the participants. In our virtual model, we replaced the three screens with a larger high resolution screen, which could in practice be constructed out of an array of narrow bezel LCD panels. And we replaced the cameras with an array of six cameras along a line in front of the wall, each focusing on one participant (instead of two).

We modeled the static components of our conferencing room and the users in Autodesk 3ds Max and exported them to an OSG file for COVISE to read. The dynamic elements, such as the cameras and the screens, are created by our plug-in on the fly using OpenSceneGraph primitives. Figure 2 shows the static and dynamic elements of our virtual conference room.

To mock-up the concept of a multi-site conferencing system, our prototype simulates a two-site conference by displaying two conference rooms back-to-back, which helped us in debugging what each conference participant should see on the screen. In such a setup, a straight line drawn from one participant to another in the other conference room illustrates the line of sight between those two participants, which is indicated as the dotted and dashed lines in Figure 1. This line intersects the display wall in front of the users in each room. These intersections are the locations in each of the two rooms where cameras should be placed and where the other participant should show up on the screen, in order to simulate correct viewing directionality. If we draw a line for every pair of participants, the number of intersection points with the screen equals the number of cameras and images needed. Now one can introduce a threshold distance, below which two neighboring cameras and images are to be merged into one, in order to reduce cost and increase available screen real-estate.

Our prototype uses two types of cameras: one simulates the cameras on top of the screens that look at the participants, the other simulates what the participants see. In each room, there are six cameras of each type, all of which can be moved, panned, and tilted to adjust the view. These camera views are projected onto our virtual screen at the front of the conferencing room. The camera images from the cameras pointed at the remote participants are displayed with a render-to-texture approach just above the table, at the same height the Cisco system displays them, to display the users' heads at their natural height.

Above those images we display two of the six images from the cameras of the local viewers, to show what the six participants see. The latter images would not be displayed in a physical setup, they are only used to verify that the virtual users see the correct images. The operator of our teleconference simulator can use the 3D wand to adjust the position of the screens (with alpha blending, so that overlapping images can be blended together to form a continuous image), the position of the cameras, and the pan and tilt of the cameras. The user can also select which user's views to display.

3.1 Automatic Camera Selection

In a typical usage scenario, from the array of six in each room, the user chooses a set of active speakers, one speaker in one room, and two speakers in the other room. Thus, there are two pairs of active participants, where the single active participant in the first room can look at either or both of the active participants in the second room. In presentation mode, two cameras in each room are automatically chosen, each camera looking at one speaker, such that this camera is the closest to the line of sight between the pair of speakers 3. The images from these four chosen cameras are then projected onto the screens in each room.

3.2 Viewing 3D Models

Viewing datasets as a group is a task that often helps to make discussions more clear. Our system demonstrates how participants of the conference can view datasets in the middle of the conference space together as a group. Our concept of the n-gon naturally lends itself to displaying the data model that is being discussed by the group in the center of the virtual n-gon. Our prototype can load a 3D dataset into the virtual space between the conference participants and the display wall, and it can be manipulated by moving and scaling it with the 3D wand.

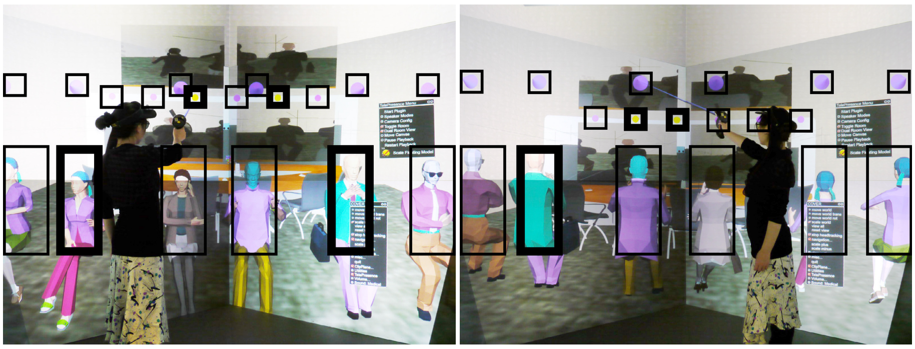

Fig. 3. Local (left) and remote (right) rooms with conference in session. For clarity, boxes are superimposed onto the photos: non-bold and bold small boxes denote inactive and active cameras, respectively; non-bold and bold large boxes denote inactive and active participants, respectively. After three active (bold) participants are chosen, one local (left) and two remote (right), four mounted cameras (bold) are automatically chosen, such that they are each the camera with the smallest possible distance to the line of sight between the pair of active participants.

When viewing a data set in the middle of the virtual conference space, there are two ways in which the data is displayed: one can either display the data set in the way it would be seen if it was a physical object, viewed by people around it, so that everybody sees a different side of the object. Or, alternatively, the object can be displayed so that every user sees the same side of it, which is similar to showing every participant the view of the same camera pointed at the object. Our prototype supports both concepts.

4 Discussion

In this section we are going to discuss various topics we came across while developing and using the VR teleconferencing simulator. We strongly believe that the simulator acted as a catalyst to more rapidly gain insight into the complicated setup of cameras and displays our prototype proposes, but this claim is hard to quantify.

4.1 View Interpolation

The more local and remote users there are, the more cameras are needed to capture correct lines of sight for the users. This relationship for the number of cameras C required for L local and R remote users is $C = R \times L$, resulting in rapidly growing numbers as the number of users increases ($O(n^2)$). For example, in Figure 1(a), this number would come out to four for each of the sites, in Figure 1(b) it is six for each of Sites 1 and 3, and four for Site 2.

In order to reduce the number of cameras required by the system for correct lines of sight, view synthesis approaches could be used. Seitz and Dyer [15] have shown that this can be done by using just two cameras images; we hypothesize that more cameras are needed for higher quality images the more remote participants there are.

4.2 User View Separation

In our simulator we experimented with displaying the different views for the different users in separate windows side by side. Each user gets to see a correct camera view of all remote participants, but their locations can overlap with the other local participants'. Hence, it is straightforward to consider display technology which can separate the views for the different users and for each user hide those views generated for the other local user(s).

Stereo displays would allow for two separate views if the stereo glasses were modified to show either two left-eye images, or two right-eye images. This is easier to accomplish with polarizing glasses than active stereo shutter glasses. This approach is intriguing, but would require for the users to wear glasses, which would defeat the purpose of the system of allowing direct eye contact and correctness of gaze direction. Auto-stereoscopic displays can remedy this issue [12], and their quality has increased significantly throughout the past few years. Many of these displays can generate eight or more views. Most of them, however, would require for the users to sit in very specific locations so that they see their dedicated views. This constraint might not be too hard to satisfy, given that in the current Cisco system, the users also need to sit in relatively specific locations in order to show up in the right location on the screen.

4.3 Directional Correctness

In our approach, we only consider horizontal directional correctness, but not vertical. Our assumption is that every user is located at the same "height", so that only horizontal directionality matters. Vertical correctness could be achieved if the cameras were installed at eye level, behind or in front of the displays. Or, view interpolation could help solve this problem without obstructing the screen by doubling the number of cameras and installing the additional ones below the screen, and then interpolate between every vertical pair of cameras.

4.4 View Sharing for 3D Models

Our video conferencing simulator implements the visualization of 3D models in the middle of the virtual conference space. Since for 3D models no physical cameras are needed, it is very easy to superimpose the rendering of this 3D model onto the video streams from the cameras. In this way it is possible to allow directionally correct viewing of 3D models in the midst of the participants, if the 3D model is virtually placed inside of the n-gon, which is where in Figure 1 the Network cloud is located. The limitation of this approach is that it only works well if the object is smaller than the virtual space between the conference participants. Ideally, the object is displayed below the sight lines between the participants, similar to the projection approach in Cisco's teleconferencing system. This could be accomplished by extending the display walls down to floor level.

Another strategy for view sharing could be that the system could automatically detect which participant is speaking and then show every participant the 3D object's view of that participant, so that everybody has the same view as the speaker.

5 Conclusion and Future Work

We presented a virtual reality simulation tool for the development of future video conferencing systems, and discussed some of the unique features of our system and its limitations. Based on these simulations, we were able to confirm our hypothesis about directional disambiguation of multi-user, multi-site video conferencing systems, which was that by using a virtual site arrangement as an equilateral polygon we are able to convey gaze and pointing direction correctly between all conference participants. The simulator itself proved to be very capable of giving the user the impression of a real system, which gave us a much better basis for discussions and insight than sketches and descriptions would have.

The next steps towards a future video conferencing system are to implement some of the proposed technology in the simulator: for instance, the algorithmic interpolation between camera views, and a simulation of multi-viewer display systems. In addition, we would like to verify our findings with live video and human participants to see if the human perception matches our findings. Eventually, we would like to build a physical prototype system to confirm the findings from our simulator.

Acknowledgements. This research was made possible in part by the support of a Cisco Research Center Grant.

References

1. Baker, H.H., Tanguay, D., Sobel, I., Gelb, D., Goss, M.E., Culbertson, W.B., Malzbender, T.: The Coliseum Immersive Teleconferencing System. Technical Report by Hewlett-Packard Laboratories (2002) 25
2. Chu, R., Tenedorio, D., Schulze, J., Date, S., Kuwabara, S., Nakazawa, A., Takemura, H., Lin, F.-P.: Optimized Rendering for a Three-Dimensional Videoconferencing System. In: Proceedings of PRAGMA Workshop on e-Science Highlights, IEEE International Conference on e-Science, Indianapolis, IN, December 8-12 (2008) 25
3. Fields, C.I.: Virtual space teleconference system. US Patent 4,400,724 (August 1983) 25
4. Fuchs, H., Bishop, G., Arthur, K., McMillan, L., Fuchs, H., Bishop, G., Bajcsy, R., Lee, S.W., Farid, H., Kanade, T.: Virtual space teleconferencing using a sea of cameras. In: Proc. First International Conference on Medical Robotics and Computer Assisted Surgery (1994) 26
5. Greenhalgh, C., Benford, S.: Massive: a collaborative virtual environment for teleconferencing. ACM Transactions on Computer-Human Interaction, TOCHI 2 (September 1995) 25
6. Kishino, F., Miyasato, T., Terashima, N.: Virtual space teleconferencing communication with realistic sensations. In: Proc. 4th IEEE International Workshop on Robot and Human Communication (1995) 26
7. Loeffler, C.E.: Distributed virtual reality: Applications for education, entertainment, and industry. Telektronikk (1993) 24
8. Maimone, A., Fuchs, H.: Encumbrance-free Telepresence System with Real-time 3D Capture and Display using Commodity Depth Cameras (2011), `http://www.cs.unc.edu/maimone/KinectPaper/kinect.html` 25
9. Ohya, J., Kitamura, Y., Takemura, H., Kishino, F., Terashima, N.: Real-time reproduction of 3d human images in virtual space teleconferencing. In: Proc. Virtual Reality Annual International Symposium, VRAIS 1993, pp. 408–414 (1993) 26

10. Ohya, J., Kitamura, Y., Takemura, H., Kishino, F., Terashima, N.: Virtual space teleconferencing: Real-time reproduction of 3d human images. Journal of Visual Communication and Image Representation 6, 1–25 (1995) 26
11. OpenSceneGraph. Scenegraph based graphics library (2004), http://www.openscenegraph.org 27
12. Peterka, T., Sandin, D.J., Ge, J., Girado, J., Kooima, R., Leigh, J., Johnson, A., Thiebaux, M., DeFanti, T.A.: Personal varrier: Autostereoscopic virtual reality display for distributed scientific visualization. Future Generation Computer Systems 22(8), 976–983 (2006) 30
13. Rantzau, D., Frank, K., Lang, U., Rainer, D., Wössner, U.: COVISE in the CUBE: An Environment for Analyzing Large and Complex Simulation Data. In: Proceedings of 2nd Workshop on Immersive Projection Technology, IPTW 1998, Ames, Iowa (1998) 27
14. Sandow, D., Allen, A.M.: The Nature of Social Collaboration: How work really gets done. Reflections 6(2/3) (2005) 25
15. Seitz, S., Dyer, C.: Physically-valid view synthesis by image interpolation. In: Proceedings IEEE Workshop on Representation of Visual Scenes (In Conjuction with ICCV 1995), pp. 18–25 (June 1995) 25, 29
16. Szigeti, T., McMenamy, K., Saville, R., Glowacki, A.: Cisco TelePresence Fundamentals. Cisco Press, Indianapolis (2009) 27
17. Wu, W., Yang, Z., Nahrstedt, K., Kurillo, G., Bajcsy, R.: Towards Multi-Site Collaboration in Tele-Immersive Environments. In: Proceedings of the 15th International Conference on Multimedia (2007) 25
18. Yoshida, M., Tijerino, Y.A., Abe, S., Kishino, F.: A virtual space teleconferencing system that supports intuitive interaction for creative and cooperative work. In: Proceedings of the 1995 Symposium on Interactive 3D graphics, SI3D (1995) 26