ConnectBoard: A remote collaboration system that supports gaze-aware interaction and sharing

Kar-Han Tan , Ian Robinson, Ramin Samadani, Bowon Lee,
Dan Gelb, Alex Vorbau, Bruce Culbertson, John Apostolopoulos

Multimedia Communications and Networking Lab
Hewlett-Packard Labs
1501 Page Mill Road
Palo Alto, California, USA.

***Abstract*—We present ConnectBoard, a new system for remote collaboration where users experience natural interaction with one another, seemingly separated only by a vertical, transparent sheet of glass. It overcomes two key shortcomings of conventional video communication systems: the inability to seamlessly capture natural user interactions, like using hands to point and gesture at parts of shared documents, and the inability of users to look into the camera lens without taking their eyes off the display. We solve these problems by placing the camera *behind* the screen, where the remote user is virtually located. The camera sees *through* the display to capture images of the user. As a result, our setup captures natural, frontal views of users as they point and gesture at shared media displayed on the screen between them. Users also never have to take their eyes off their screens to look into the camera lens. Our novel optical solution based on *wavelength multiplexing* can be easily built with off-the-shelf components and does not require custom electronics for projector-camera synchronization.**

I. Introduction

With the advent of modern high-fidelity video communication technologies, remote collaboration systems are increasingly being employed to facilitate meetings between geographically distributed users. While there are clear benefits in reducing the carbon footprint, time, and expenses required to transport users to a common physical location, a real face-to-face meeting is still a qualitatively better experience than a remote meeting. We believe this is due to the fact that many sometimes subtle aspects of co-located meetings are lost in a remote meeting. We are particularly interested in addressing the following issues:

- **Nonverbal communications** In a co-located meeting one user can simply point and gesture at a document with his or her hands and the other user can instantly grasp the nonverbal meaning that is being conveyed. In a remote meeting, for electronic documents users typically have to use pointing devices to communicate through cursor movements, and hand gesturing can only be captured with document cameras and physical media. More subtle, but just as important, is communicating gaze direction accurately. This enables a user to see where the other's attention is focused - are they looking where I'm pointing?
- **Eye Contact** In a co-located meeting, it is easy to have eye contact while users engage in conversation. In a conventional remote meeting, eye contact is virtually impossible because if a user is looking at the image of the other user on screen, then he or she is not looking at the camera and the other user gets a view of the user looking typically downwards. If a user wants to look at the camera, then he or she cannot be looking at the image of the other user. This problem is only exacerbated when users work close to the screen.

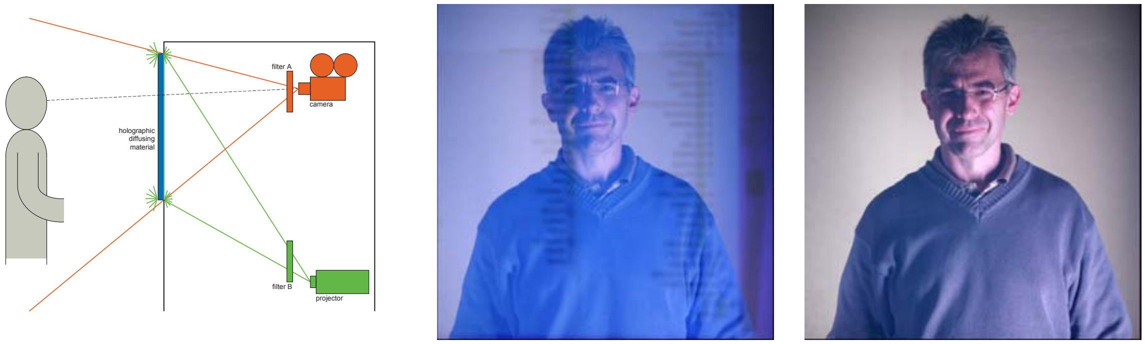
These issues get in the way of natural interactions and users are forced to constantly work *around* them. Our goal is to remove these barriers to natural communications and create a highly intuitive experience so that users can simply have engaging, productive interactions.

II. Previous Work

Eye contact is one of the oldest problems in telepresence. The first attempts used half silvered mirrors like those in the Teleprompter [1], which are still widely used today by television newscasters and public speakers. While Teleprompters are one-way communication devices, similar devices like Gazecam [2] and the EuroPARC Reciprocal Video Tunnel [3] were used in teleconferencing systems. These systems allow users to look at the remote user's image while looking into the camera at the same time. Using a half silvered mirror, which is typically angled at 45 degrees from the display, results in a large footprint. Stray reflections off the mirror can also create distracting views say of the ceiling or floor. Creating eye contact using this method thus typically results in deep enclosures that limit the range of usable viewing angles (which fits the 'tunnel' metaphor).

Another way to implement eye contact is to use switchable liquid crystal diffusers, a technique demonstrated by Shiwa at NTT [4]. Such a diffuser can switch quickly between two states: transparent and diffusing. In the transparent state, synchronized cameras can capture images of the user; In the diffusing state synchronized projectors can render images

MMSP'09, October 5-7, 2009, Rio de Janeiro, Brazil.
978-1-4244-4464-9/09/$25.00 ©*2009 IEEE.*



(a) (b) (c)

Fig. 1. Optical backscatter removal. (a) Display-camera setup. (b) Backscatter clutters user image. (c) Backscatter removed.

of the remote user. This technique was also used in *blue-c* [5]. More recently SecondLight [6] used a switching diffuser to allow projection onto objects above the screen, enabling tangible viewing of overlay visualizations. The switching diffuser technique allows smaller footprints and wider viewing angles. The key technical limitation [7], [8] is that currently available diffusers may not switch fast enough, especially in larger screens, resulting in flickering images. It is possible to overdrive the screens to achieve higher switching frequencies, but transition times between the two states are still significant enough to reduce the actual duty cycles of both the synchronized projector and camera, resulting in dim or noisy images.

Instead of half mirrors or switching diffusers, TouchLight [9] uses a screen which diffuses only light incident from pre-specified angles, and allows light to pass through otherwise. This gives a transparent screen which can display images if the projector is placed at the right location. It does not require special synchronized cameras and projectors, thus offering greater freedom for designers. Eye contact is, however, not supported by TouchLight as the diffuser bounces light from the projector back into the camera as well. This *backscattered* portion of the displayed content gets superimposed on the image of the user captured by the camera, as shown in Fig. 1(b). HoloPort [8] uses a similar transparent screen, and uses synchronized projector and camera to counter the backscatter problem.

A creative variation of the half silvered mirror technique was used in ClearBoard [7], where a polarizing film was sandwiched between a projection screen and a half mirror. The system has a ‘drafting table’ design where the work and display surface is placed at a 45 degree angle, and a camera captures the mirror image of the user from above. Images from the display are blocked by another polarizing filter on the camera to ensure that only the mirror image is captured. However the drafting table design produces an unnatural view of the remote user, who would appear to be leaning backwards while working on the shared surface, even though he/she is sitting upright.

ClearBoard allows eye contact and simultaneously supports nonverbal communication as both users can see one another’s hands and they can gesture and point at objects on the shared surface. C-Slate [10] renders hands overlaid on a shared surface that vary in transparency based on distance between the user’s hands and the shared surface. When the user’s hands are far from the surface they appear highly transparent but by the time they touch the surface and start to manipulate digital objects, the hands appear opaque. C-Slate, however, does not support eye contact. Highly stylized silhouettes of hands can also be used for nonverbal communications [11].

III. Our Solution

As we have seen, there have been several attempts at designing systems which simultaneously support eye contact and nonverbal communications. Ideally, we would like to create a ClearBoard-like experience, except with a vertical surface that capture a frontal view of the user. We would also like to be able to build our system from off the shelf components, without requiring synchronization between projectors and cameras, so that they can operate at their respective optimal frame rates and exposure settings.

To meet these requirements, we employ a directionally-selective diffuser similar to that used in HoloPort. Unlike the HoloPort system, we choose not to use active synchronized projector cameras. Instead, we use *wavelength multiplexing* in the visible light spectrum. The idea is that if the projector outputs light in spectral ranges that do not overlap those observable by the camera, the camera will not sense the backscattered light emitted from the projector. A wavelength multiplexed projector-camera pair can be built with the use of interference filters originally designed for viewing stereoscopic 3D movies [12]. As can be seen in Fig. 1(c), we are able to optically remove all backscatter.

IV. Color Processing

Interesting engineering choices exist for the specific optical filter passbands used in our system, and the necessary color correction methods differ depending on these choices. In this section, we discuss the color correction for the prototype system we built using off-the-shelf filters from *Infitec* [12] with ordinary projectors and cameras. Two color characterizations [13] are needed: 1) for the camera with the first Infitec filter; and 2) for the projector with the second, complementary Infitec filter.

The characterization of the camera subsystem is complex and nonlinear because the filters are not within a linear transformation of colorimetric XYZ. Nevertheless, we found approximate but satisfactory results by determining a transformation composed of 1D transfer functions for each color channel followed by a matrix for converting to CIE XYZ. To determine the parameters of the transfer functions and the conversion matrix, we measured the CIE XYZ of a MacBeth color chart using a Photo Research PR650 spectrophotometer as well as obtaining RGB values from the video camera that was used in the system. Our characterization software extracted the MacBeth patches, and our numerical fitting software generated both the 1D monotonic curves for the gray scales, as well as the matrix coefficients. The measurements were conducted with the same lights that are used during system use.

The characterization of the projector subsystem was more straightforward. Even though the Infitec filters have unconventional multiple spectral passbands, when they are placed in front of an LCD projector, additive color theory [13] applies. Matrix multiplication converts XYZ to display primaries (with the second Infitec filter) and per channel 1D functions convert to nonlinear output projector color space. To determine the parameters for the projector characterization, we used a diffuse white standard target and measured the CIE XYZ of different input digital RGB values displayed through the projector subsystem. Subsequently, we determined the parameters of the matrix and 1D functions using our numerical fitting software.

The overall transformation from camera to projector then consists of 1D input transfer functions, matrix operation (the concatenation of the camera and projector matrices) and output transfer function per color channel. Traditional linear matrix primary correction, however, was observed to cause a large amount of clipping. This is because of a large power imbalance between the three primaries due to the alignment of the projector spectral power with the Infitec filter passbands (verified using spectrophotometer measurements). For this reason, after we determined the primary correction parameters and the 1D output functions using spectrophotometric measurements, we developed a non-linear primary correction algorithm, shown in Fig. 2.

The input 1D lookup functions are defined by the transformation from camera to input color space. Similarly, the output 1D lookup functions are defined by the transformations to the projector output space. We use traditional methods of measuring color ramps followed by numerical function fitting or function fitting and inversion to provide for the input and output 1D lookup functions, but we will not discuss its details here. The rest of this discussion focuses on the primary correction of Fig. 2.

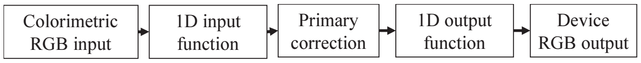


Fig. 2. Flowchart of the non-linear color correction used.

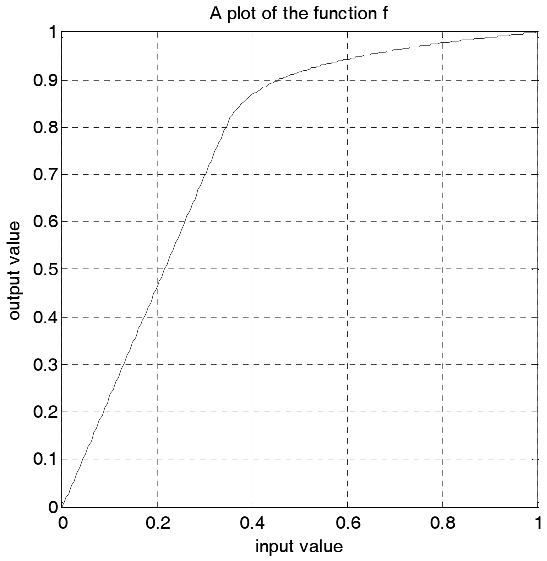


Fig. 3. A plot of the function $f(x)$

One approximate color correction that we implemented and tested involved a simple modification to matrix multiplication. Consider standard matrix multiplication

$$\begin{pmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \\ m_{31} & m_{32} & m_{33} \end{pmatrix} \begin{pmatrix} r \\ g \\ b \end{pmatrix} = \qquad (1)$$

$$\begin{pmatrix} m_{11}r & m_{12}g & m_{13}b \\ m_{21}r & m_{22}g & m_{23}b \\ m_{31}r & m_{32}g & m_{33}b \end{pmatrix} = \begin{pmatrix} m_1^T \boldsymbol{c} \\ m_2^T \boldsymbol{c} \\ m_3^T \boldsymbol{c} \end{pmatrix},$$

where the color vector $\boldsymbol{c}$ is composed of the color components (r, g, b) and m_i^T refers to the ith row of the matrix M. Empirically, we have found that for our system, one of the matrix diagonal components is much larger than the others. In this case, one can modify the operation to be:

$$\begin{pmatrix} m_{11}r & m_{12}g & m_{13}b \\ m_{21}r & f(m_{22}g) & m_{23}b \\ m_{31}r & m_{32}g & m_{33}b \end{pmatrix} \qquad (2)$$

where the function f is given by

$$f(g) = \begin{cases} m_{22}g & \text{if } g < g_0 \\ \frac{(g+k)}{(1+k)}^{\gamma} & \text{if } g \geq g_0 \end{cases} \qquad (3)$$

Figure 3 shows a plot of the function in Eq. (3). This function is of the same form as the nonlinear gamma functions used in colorimetric spaces like sRGB, and it simply matches a power function to a linear function with the same function value and slope at the threshold value g_0. The parameters k and γ may be determined once g_0 is specified. We set $g_0 = 0.8$ and performed the operation in Eq. (2), followed by a simple component clip (still needed because of the off-diagonal terms). Even though this approximation performs well in this example, it will not work for the case where two of the diagonal elements are large. A better approach is to apply the function f to the output of the matrix multiplication. When the first and third diagonal components are large, we apply the following transformation

$$\begin{pmatrix} f(m_1^T \boldsymbol{c}) \\ m_2^T \boldsymbol{c} \\ f(m_3^T \boldsymbol{c}) \end{pmatrix}, \tag{4}$$

In the more general multidimensional case, the idea of the nonlinear correction is that a companding non-linearity applied to each output component of a traditional 3×3 matrix multiplication reduces the clipping and results in smooth color transitions. The compander is similar to the one dimensional gamma functions of colorimetric spaces, but extended to multiple dimensions,

$$f(m_i^T \boldsymbol{c}) = \begin{cases} D_c\|\boldsymbol{c}\| = m_i^T \boldsymbol{c} & \text{if } D_c\|\boldsymbol{c}\| < y_0 \\ \Big((\frac{(\|\boldsymbol{c}\|+k)}{(\|\boldsymbol{c}_{max}\|+k)} \Big)^{\gamma} & \text{otherwise} \end{cases} \tag{5}$$

where m_i^T is the i^{th} row of the traditional 3×3 matrix correction, D_c is the directional derivative of the matrix multiplication in the direction of input color $\boldsymbol{c}$, and $\boldsymbol{c}_{max}$ is the maximum norm color (the color that just clips) in the direction of $\boldsymbol{c}$. This multidimensional correction softly compands the colors near the gamut boundary. An observation of the actual correction matrices also shows that the computation may be optimized by observing that certain off-diagonal elements in the matrices are near zero. This allows the expensive computations determining γ and $\boldsymbol{c}_{max}$ (which are color direction dependent) and the color norm $\|\boldsymbol{c}\|$ to be precomputed and stored in small (one dimensional) lookup tables so that full 3D color correction is not necessary.

During the operation, one has the choice of conducting the companding operation described independently per color component, or to adjust the colors towards black by applying the same minimum gain factor to all the color components. The tradeoffs are darker but more accurate color tones when adjusting towards back, and somewhat less accurate color tones but brighter ones when adjusting the colors independently per component. An example of color correction is shown in Fig. 4.

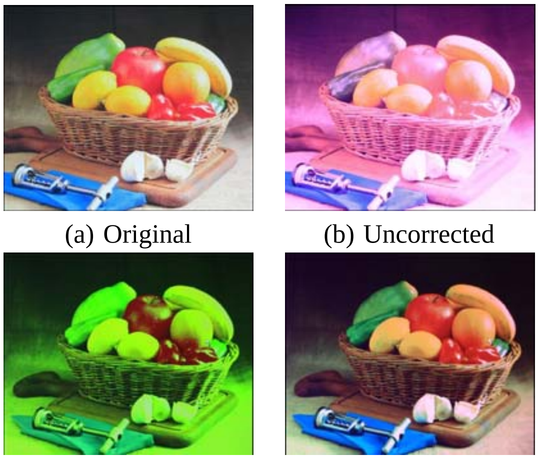

(a) Original
(b) Uncorrected
(c) Pre-distorted
(d) Corrected

Fig. 4. Color correction results. (a) Original image (b) Original image displayed through a wavelength multiplexing filter, clearly showing an unnatural color shift. (c) Our algorithm pre-distorts the original image (d) After being displayed through the filter, the resulting image is close to the original. Note that all images were projected onto a white surface and photographed and so the camera imaging process had introduced additional color transformation and is only an approximation of what is actually observed by the human eye.

V. Image Processing

Besides color processing, we also apply geometric corrections to the image. Due to the use of wide angle lenses in our cameras, they exhibit significant radial distortion [14]. We estimated the parameters using the MATLAB Camera Calibration Toolbox [15] and applied the warp using Intel Performance Primitives library. As we used consumer camcorders in our system, which have limited low light performance, under some indoor lighting conditions there can be visible image noise. When necessary, we turn on a real time O(1) GPU bilateral filter-based denoising operator [16].

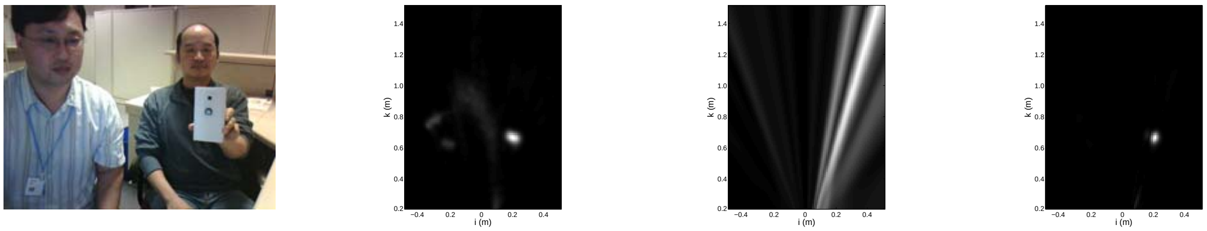
VI. Audio

To enable a natural and immersive audio experience, we use spatial audio acquisition and rendering. We placed an array of two microphones along the top edge of the screen for audio acquisition, and a loudspeaker on each side of the screen for stereo audio rendering (See Fig. 6). In order to suppress acoustic feedback from the loudspeakers to the microphones, we use a multichannel acoustic echo cancellation algorithm on four (2 × 2) signal paths. Using the microphone array, we localize and track an active talker in the local acoustic scene. Localization allows us to enhance the captured speech signal by beamforming. From the monaural speech signal we generate two audio streams by applying the inter-aural time and intensity differences. The synthesized stereo audio stream is encoded and transmitted to the remote location.

In uncontrolled environments, background noise and reverberation may degrade the performance of acoustic source localization algorithms. We plan to use multimodal sensor arrays consisting of cameras, microphones, and active depth sensors [17]. Fig. 5 illustrates that multimodal fusion [18] can potentially provide accurate 3D sound source localization that is impossible with the individual modalities.

VII. Streaming

The video and audio streams in ConnectBoard are processed using a flexible software dataflow framework called Nizza [19] that we have developed. This enables rapid prototyping and experimentation by combining existing modules such as camera capture and compression with ConnectBoard specific components such as the color processing (See section IV). Our system currently uses MPEG2 video compression and MPEG1 layer 2 audio compression. MPEG2 video coding provides a good trade-off of computational complexity and compression efficiency. If sufficient CPU is available then more advanced codecs such as H.264 are available in our system. On the capture side, audio and video streams are acquired and echo cancellation is performed on the input



(a) Frontal RGB View (b) Depth PDF Top View (c) Acoustic PDF Top View (d) Joint PDF Top View

Fig. 5. Multimodal joint 3D PDF in the spatial domain. (a) Frontal view of the scene. (b) Top view of the 3D PDF from the depth sensor. (c) Top view of the 3D PDF from the pair of microphones. (d) Joint 3D PDF. Clearly object localization is much easier in the joint PDF than the separate depth and acoustic PDFs which are fairly noisy.

audio signals. The streams are compressed and packaged into MPEG transport streams for network transmission. Audio and video streams from the remote participant are received over the network, reassembled and decompressed. After decompression the necessary color correction is applied. Currently the color correction is performed on the CPU, while geometric correction is performed on the GPU. For efficiency we plan on moving the color correction to the GPU since the GPU is very well suited to such processing.

Low latency is required for any remote collaboration system to be usable. We have measured the end-to-end latency of ConnectBoard using a photographic technique. We take a single photograph that contains both a live view of a running timer, and a view of the same timer as captured and processed by the system. The difference between the two timers yields the total end-to-end latency. We are currently using off-the-shelf consumer HD camcorders with HDMI output. While being inexpensive, these cameras unfortunately have internal buffering which can add close to a hundred milliseconds of delay compared to other professional cameras. Even with this delay, the end-to-end latency was measured at 260 to 270 milliseconds.

VIII. Prototype Design

The design of the prototype system had to take a number of factors into account. To accommodate the ability to walk up and use the system, the screen had to be sized and positioned to allow users with a range of heights to see each other given typical "social distances", and to comfortably use the board as an interaction surface. Our latest prototype, shown in Fig. 6 below, uses a 50" holographic rear projection screen from Sax3D, mounted 42" off the floor, that is designed for a close-mounted, short-throw projector. This has the advantage of a simpler and more compact layout than systems with a longer throw, which require folding the optical path with mirrors.

With the partially diffusing screens used, a significant hot spot occurs if the user can see the projector through the screen. To prevent this, screens are designed to accept an image projected from a glancing angle, so that the projector can be placed outside of the usual sight lines. However, the close interaction distance, the fixed location of the projector with respect to the screen, and the desire to accommodate a range of user heights, made it necessary to mount the projector above the screen.

The camera should be mounted so as to capture a view similar to what the remote participant would see if they were actually standing on the other side of the screen. The view had to encompass the whole of the display (so that the remote party can see gestures directed at any part of the screen surface). To achieve this from a typical user distance, however, requires a wide-angle lens that can lead to significant distortions in the image. We are investigating the impact of moving the camera further back and using a longer focal length lens.

IX. Conclusions and Future Work

ConnectBoard allows users to be closer to the screen, which inspires new user interaction scenarios and also new interaction design challenges. Ideally, when users stand closer to the screen, and perceptually closer to one another, the experience can be more immersive. This fills the user's field of view with the image of another person and allows them to speak at a more intimate volume.

One of the more common and productive workplace interaction scenarios is the standing hallway meeting, sometimes called a "scrum" or "watercooler" meeting. These meetings among two or three people are usually brief, informal, and often spontaneous. Interactions like these are well-suited for ConnectBoard, however they will require a low-friction interactive design that encourages spontaneous meetings.

When users stand closer to the screen they can also touch the screen, which supports the experience vision of users interacting as if separated by a transparent sheet of glass. This brings to mind another common and productive scenario of two co-workers collaborating at a whiteboard. In real life, people stand shoulder to shoulder at a whiteboard, taking turns writing and turning to one another to discuss. With ConnectBoard they would stand on opposite sides of a transparent "greaseboard", writing on the board and discussing "through" the glass. One challenge this scenario presents is that as one user writes on the board, the other user would see a mirror image of the writing. To correct this, we flip each user's relative image horizontally, which could detract from the perceived reality of the experience.

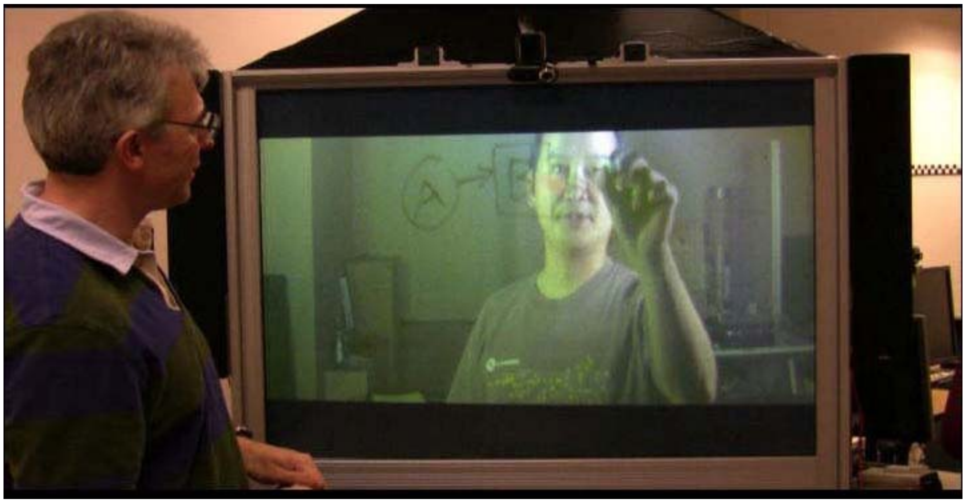


Fig. 6. ConnectBoard in action, linked to a second prototype in a remote location. Note that the image is horizontally flipped so that drawings made on the screen appear correct to both users.

These are just a few of the user interaction opportunities created by the ConnectBoard concept and we look forward to reporting on these and others in future publications.

References

[1] J. Oppenheimer, "Prompting apparatus," US Patent 2883902, 1959, filed Oct 1954.

[2] S. R. Acker and S. R. Levitt, "Designing videoconference facilities for improved eye contact," *Journal of Broadcasting and Electronic Media*, vol. 31, no. 2, pp. 181–191, 1987.

[3] B. Buxton and T. Moran, "Europarc's integrated interactive intermedia facility (iiif): early experiences," in *Proceedings of the IFIP WG 8.4 confernece on Multi-user interfaces and applications*. Amsterdam, The Netherlands, The Netherlands: Elsevier North-Holland, Inc., 1990, pp. 11–34.

[4] S. Shiwa and M. Ishibashi, "A large-screen visual telecommunication device enabling eye contact," *SID Digest*, vol. 22, pp. 327–328, 1991.

[5] 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. V. Moere, and O. Staadt, "blue-c: a spatially immersive display and 3d video portal for telepresence," in *Proceedings ACM SIGGRAPH*, 2003, pp. 819–827.

[6] S. Izadi, S. Hodges, S. Taylor, D. Rosenfeld, N. Villar, A. Butler, and J. Westhues, "Going beyond the display: a surface technology with an electronically switchable diffuser," in *Proceedings 21st annual ACM symposium on User interface software and technology (***UIST***)*, 2008, pp. 269–278.

[7] H. Ishii and M. Kobayashi, "Clearboard: a seamless medium for shared drawing and conversation with eye contact," in *Proceedings of the ACM SIGCHI conference on Human factors in computing systems (***CHI***)*, 1992, pp. 525–532.

[8] M. Kuechler and A. Kunz, "Holoport - a device for simultaneous video and data conferencing featuring gaze awareness," in *Proceedings of the IEEE conference on Virtual Reality*. IEEE Computer Society, 2006, pp. 81–88.

[9] A. Wilson, "Touchlight: An imaging touch screen and display for gesture-based interaction," in *Proceedings International Conference on Multimodal Interfaces (***ICMI***)*, 2004.

[10] S. Izadi, A. Agarwal, A. Criminisi, J. Winn, A. Blake, and A. Fitzgibbon, "C-slate: A multi-touch and object recognition system for remote collaboration using horizontal surfaces (**tabletop**)," *International Workshop on Horizontal Interactive Human-Computer Systems*, vol. 0, pp. 3–10, 2007.

[11] P. Tuddenham and P. Robinson, "Territorial coordination and workspace awareness in remote tabletop collaboration," in *Proceedings of the 27th international conference on Human factors in computing systems (***CHI***)*. ACM, 2009, pp. 2139–2148.

[12] H. Jorke and M. Fritz, "A new stereoscopic visualization tool by wavelength multiplex imaging," in *Proceedings Electronic Displays*, Sep 2003.

[13] H. Lee, *Introduction to color imaging science*. Cambridge University Press, 2005.

[14] D. C. Brown, "Close-range camera calibration," in *Photogrammetric Engineering*, vol. 37, no. 8, 1971, pp. 855–866.

[15] J.-Y. Bouguet, "Camera calibration toolbox for matlab," http://www.vision.caltech.edu/bouguetj/calib_ doc/index.html.

[16] Q. Yang, K.-H. Tan, and N. Ahuja, "Realtime ø(1) bilateral filtering," in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (***CVPR***)*, 2009.

[17] S. B. Gokturk, H. Yalcin, and C. Bamji, "A time-of-flight depth sensor - system description, issues and solutions," in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition Workshop (***CVPRW***)*, vol. 3, 2004.

[18] B. Lee and K.-H. Tan, "Maximum a posteriori multimodal object localization with a depth sensor and stereo microphones," in *Proc. IMMERSCOM*, May. 2009.

[19] D. Tanguay, D. Gelb, and H. H. Baker, "Nizza: A framework for developing real-time streaming multimedia applications," HP Labs Tech Report HPL-2004-132 http://library.hp.com/techpubs/2004/HPL-2004-132.html, 2004.