TEMPORAL MODULATION FOR COMPUTATIONAL VIDEO CROSS-TALK REDUCTION

[1]*Dookun Park*, [2]*Ramin Samadani*, [2]*Kar-Han Tan, and* [2]*Dan Gelb*

[1]Stanford University, [2]Hewlett-Packard Labs

ABSTRACT

Many recent prototypes for video collaboration, digital media sharing and gesture interfaces provide a video signal for display on a screen or surface and capture another video signal through the same screen or surface. The media captured in such systems, for transmission or for gesture user interfaces, needs to be separated from the displayed video. Otherwise, *video cross-talk* occurs. The prior, widely used temporal multiplexing avoids cross-talk by synchronizing camera capture with screen display so that the camera only captures when the screen does not display signal. This approach suffers from light loss (both displayed and captured) and increased display flicker due to the lower duty cycle of the displayed signal. This paper describes a new method, computational temporal modulation, that temporally modulates the displayed signal. The intentionally mixed signals captured by the camera are subsequently separated using computations. Our approach results in brighter display with less flicker and more signal captured by the camera. Experiments using a prototype collaboration system show good quality cross-talk reduction with light-weight real-time computation.

***Index Terms*—** video cross-talk, visual echo cancellation, temporal multiplexing, projector-camera systems.

1. INTRODUCTION

Recent prototype systems that combine collaboration with media sharing [1, 2, 3] aim to 1) extend the productivity of remote collaborations with natural gaze awareness, and media sharing as though separated by a vertical, transparent sheet of glass [1]; 2) use a table-based interaction surface [2]; 3) allow natural use of shared digital whiteboard [3]. In addition, gesture-based interfaces [4] capture lightfields through a display to extract gestures to serve as a user interface. All these prototypes need to separate a video signal meant for display on a screen or surface from a desired captured video signal through the same screen or surface. Otherwise, the display video causes cross-talk or interference to the desired captured video.

Previous approaches for cross-talk reduction include wavelength multiplexing [1] and temporal multiplexing (TM) [2, 4], but these techniques lose half the light available, and temporal multiplexing additionally suffers from flicker due to the low duty cycle of the displayed signal. The binary classification approach to signal separation in [3] applies best to spatially limited extent cross-talk such as text and drawings. Closest to our work is the model-based approach of [5] which preserves light, but which requires heavy computation to implement the required colorimetric, geometric, and optical transformations of a full projector-camera model.

Figure 1 shows on the left, best seen by zooming in the pdf file, two consecutive video frames containing cross-talk with temporal multiplexing, for example spurious cars are seen in the skyline to the left of the person. The right shows the corresponding output frame with cross-talk reduced using our *Pass-Mask* algorithm, described in Section 3.2. Our target system is illustrated in Figure 2. It consists of a camera, a projector, and a screen. This configuration is similar to the system described in [1], but without any optical filters for the signal separation. Light paths and signals used by the software processing pipeline described in this paper are shown in the figure.

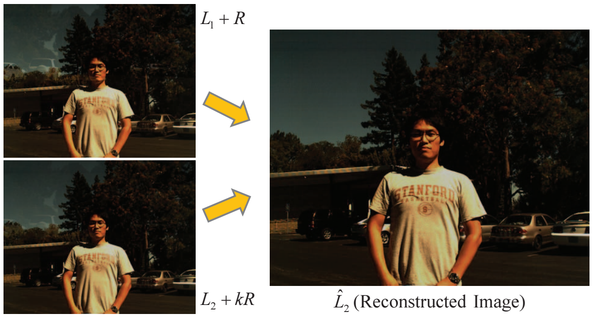

Fig. 1. Two inputs(left) with cross-talk and output(right) with reduced cross-talk of our *Pass-Mask* algorithm

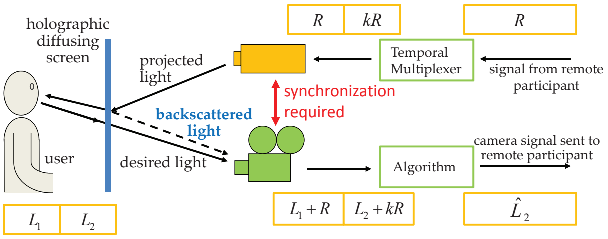

Fig. 2. Light paths and signals for the experimental system.

Figure 2 shows the local user to the left of a holographic diffusing screen. The *remote signal* projected on the screen is the video of the remote user together with any additional shared media content. At the same time, the camera needs to capture the video of the local user. The remote signal is intended for the local user only. In practice, however, the remote signal is reflected into the camera by the screen, and this results in visual *cross-talk* that is transmitted to the remote user. The human visual system is very sensitive to the structured (not random noise) cross-talk and it must be reduced.

In temporal multiplexing [2, 4], the camera is turned off for a half cycle, during which time the projector displays the remote signal for the local user, and the projector is turned off for the complementary half cycle, during which time the camera captures the local signal. With this simple approach, a synchronized camera projector system can remove cross-talk without any additional processing. However, the averaged on-off projected signal brightness is 50% as bright as a projector always on. Moreover, display flicker may be a problem because of the on-off cycling of projector display. The camera also captures 50% signal due to the multiplexing. The next section describes our approach to overcome the shortcomings of TM.

978-1-4673-2533-2/12/$26.00 ©2012 IEEE

2. PROBLEM FORMULATION

In this work, we intentionally modulate in a controlled manner the remote signal input to the projector, capture the mixed signals with the camera, and computationally reduce the cross-talk. Because the limitations of the traditional TM solution are due to the projector being fully off for half cycle, we instead only reduce the projector brightness for half cycle. At the same time, the camera continuously captures frames without being turned off. This results in increased average projector brightness, reduced display flicker, and increased camera signal. This method requires a signal reconstruction algorithm, however, since the projector and the camera share the screen at the same time and cross-talk occurs. Thus, we call this technique *Computational Temporal Modulation* (CTM).

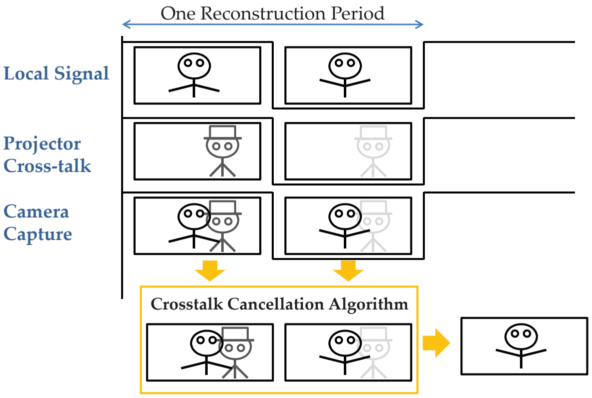

Fig. 3. Timing diagram of our system. We repeat the projector signal with different gains, removing the remote motion issue.

Figure 3 shows graphically the flow of CTM cross-talk reduction. The projector display and the camera capture are temporally oversampled by a factor of two. During one reconstruction period, we display the same projector frame, $\mathbf{R}(m,n)$, with two different intensity gains, 1 and k, with $0 < k < 1$. Repeating $\mathbf{R}(m,n)$ removes remote signal motion, allowing simpler and more effective signal reconstruction. During one reconstruction period, the modulated remote frames are captured by the camera, mixed with the local signals, $\mathbf{L_1}(m,n)$ and $\mathbf{L_2}(m,n)$. The two consecutive frames, $\mathbf{I_1}(m,n)$ and $I_2(m,n)$, captured by the camera are given by,

$$\begin{aligned} \mathbf{I_1}(m,n) &= \mathbf{L_1}(m,n) + \mathbf{R}(m,n), \\ \mathbf{I_2}(m,n) &= \mathbf{L_2}(m,n) + k\mathbf{R}(m,n), \end{aligned} \tag{1}$$

with (m,n) row and column pixel indices. Bold letters convey that the pixel values are vectors with RGB components. In general, the two consecutive local frames, $\mathbf{L_1}$ and $\mathbf{L_2}$, may differ due to motion or noise. Our goal is to estimate the local signal $\mathbf{L_2}$ based on the two captured images, $\mathbf{I_1}$ and $\mathbf{I_2}$. The signals $\mathbf{L_1}$, $\mathbf{L_2}$, $\mathbf{R}$ are not available, only the mixed signals $\mathbf{I_1}$ and $\mathbf{I_2}$ and the known constant k are inputs to the reconstruction algorithm.

3. RECONSTRUCTION ALGORITHMS

We describe two reconstruction algorithms starting from (1). The first, the *Naive algorithm*, assumes no changes between the local frames, with $\mathbf{L_1}(m,n) = \mathbf{L_2}(m,n)$. The second, the *Pass-Mask algorithm*, performs well even with local motion, benefiting from visual masking of the remote cross-talk signal, but it does not require expensive motion estimation. In principle, changes between local frames may be due to a variety of causes, including motion in the local scene or capture noise. Experiments find that motion is the important primary cause and noise has been negligible.

3.1. Naive Algorithm : No motion in the local signal

Assuming no motion, the following reconstruction formula, the *naive algorithm* may be derived from (1):

$$\begin{aligned} \hat{\mathbf{L}}_\mathbf{2}(m,n) &= \left(\frac{k}{1-k}\right)[\mathbf{I_2}(m,n) - \mathbf{I_1}(m,n)] + \mathbf{I_2}(m,n) \\ &= G\Delta\mathbf{I}(m,n) + \mathbf{I_2}(m,n), \end{aligned} \tag{2}$$

where $\Delta\mathbf{I}(m,n) = \mathbf{I_2}(m,n) - \mathbf{I_1}(m,n)$ and the constant $G = k/(1-k)$. For further interpretation, substitute for $\mathbf{I_1}$ and $\mathbf{I_2}$ of (2), the right hand sides of (1):

$$\begin{aligned} \hat{\mathbf{L}}_\mathbf{2}(m,n) &= G\left[\Delta\mathbf{L}(m,n) - (1-k)\cdot\mathbf{R}(m,n)\right] + \mathbf{I_2}(m,n) \\ &= G\Delta\mathbf{L}(m,n) + \mathbf{L_2}(m,n) \end{aligned} \tag{3}$$

where $\Delta\mathbf{L}(m,n) = \mathbf{L_2}(m,n) - \mathbf{L_1}(m,n)$.

If there is no local motion, $\Delta\mathbf{L}(m,n) = 0$, and $\hat{\mathbf{L}}_\mathbf{2}(m,n) = \mathbf{L_2}(m,n)$ perfectly reconstructs the local signal. When there is motion, the reconstructed signal $\hat{\mathbf{L}}_\mathbf{2}(m,n)$ is a sum of the desired local signal $\mathbf{L_2}(m,n)$ and the additional term $G\Delta\mathbf{L}(m,n)$. Experiments show objectionable *motion artifacts* typically occurring at the boundaries of moving objects, confirming that the added term is often due to motion induced pixel differences, amplified by gain factor G.

Higher sampling rates reduce $\Delta\mathbf{L}(m,n)$, and the motion artifacts are somewhat reduced. This partial solution, however, even with our 120 frames per second projector-camera system, shows objectionable motion artifacts due to the amplification. A more fundamental treatment is needed to reduce the motion artifacts.

3.2. Pass-Mask Algorithm

To understand the pixel adaptive *Pass-Mask algorithm*, the source of pixel value variations, $\Delta\mathbf{I}(m,n)$ is reviewed. The two main causes for the image differences (2) are 1) pixel differences due to motion in the local signal; and 2) modulation of the remote projector signal:

$$\Delta\mathbf{I}(m,n) = \Delta\mathbf{L}(m,n) - (1-k)\mathbf{R}(m,n). \tag{4}$$

In (3), the $-(1-k)\mathbf{R}(m,n)$ term is required to reduce the cross-talk, but the undesired $\Delta\mathbf{L}(m,n)$ term generates visual artifacts. We would like to remove only the $\Delta\mathbf{L}(m,n)$ component from $\Delta\mathbf{I}(m,n)$, but this quantity is not simple to detect and remove since we have no prior information about the local and remote frames that compose the mixed captured frames. The blind source separation approach [6] to this problem is very difficult. We instead develop an effective solution based on video enhancement that reduces visible motion artifacts. Instead of full separation of the two sources of pixel value variation in $\Delta\mathbf{I}(m,n)$, we classify $\Delta\mathbf{I}(m,n)$ into two categories, *definitely affected by motion* and *possibly not affected by motion*, and subsequently process the pixels adaptively.

The classification is based on the characteristics of the $-(1-k)\mathbf{R}(m,n)$ term. For the $\mathbf{R}(m,n)$ pixel values, $\mathbf{R}(m,n) \geq 0$. The modulation value k satisfies $0 < k < 1$. Additionally, there is a maximum $\mathbf{R}(m,n)$ value captured by the camera that depends on the experimental setup. This maximum, $\mathbf{R_{max}}$, is empirically determined by displaying a constant white image on the projector and capturing the reflected cross-talk with the camera in a dark room.

Camera parameters like exposure time, aperture and gain are fixed during the $\mathbf{R_{max}}$ measurement and during system operation. Since the cross-talk is only caused by the projector, $\mathbf{R_{max}}$ is only dependent on the system configuration, and not on external conditions like room light, background, etc. Based on this analysis, the quantity $-(1-k)\mathbf{R}(m,n)$ is bounded by,

$$-(1-k)\mathbf{R_{max}} \leq -(1-k)\mathbf{R}(m,n) \leq \mathbf{0}. \tag{5}$$

The notation $\mathbf{A} \leq \mathbf{B}$ means the color components $A_i \leq B_i$, for all i. If there is no motion, i.e. $\Delta\mathbf{L}(m,n) = 0$, then $\Delta\mathbf{I}(m,n)$ always falls within the bounds of (5). The quantity $\Delta\mathbf{L}(m,n)$ in (4) is not used for the classification, since with motion its bound is difficult to determine, since it is set by the colors of objects and backgrounds, illumination, amount of motion, etc.

The binary valued classification mask image P is defined by

$$P(m,n) = \begin{cases} 1, & \text{if } -(1-k)\mathbf{R_{max}} \leq \Delta\mathbf{I}(m,n) \leq \mathbf{0} \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

Based on mask P, the *Pass-Mask algorithm* provides the estimate,

$$\hat{\mathbf{L}}_{\mathbf{2}}(m,n) = GP(m,n)\Delta\mathbf{I}(m,n) + \mathbf{I_2}(m,n). \tag{7}$$

When $P(m,n) = 0$, the pixel is definitely affected by motion and the estimate $\hat{\mathbf{L}}_{\mathbf{2}}(m,n) = \mathbf{I_2}(m,n)$ does not reduce cross-talk in favor of reducing motion artifacts. Local motion is likely to have the beneficial effect of visually masking the remote cross-talk. When $P(m,n) = 1$, the pixel is possibly not affected by motion, and the cross-talk reduction algorithm (2) is applied. This reconstruction is called *Pass-Mask* because binary mask P allows only the selected pixels to pass to the cross-talk reduction step.

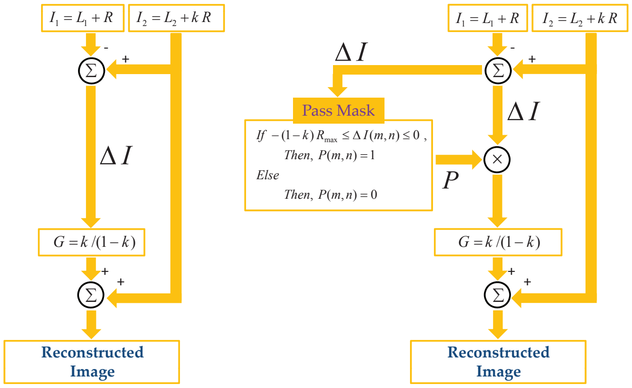

Fig. 4. Flowchart of the *Naive algorithm* on the left, and the *Pass-Mask* algorithm on the right. The Pass-Mask algorithm substantially reduces motion artifacts seen in the Naive algorithm.

Figure 4 shows the flowcharts of the naive algorithm on the left, and the pass-mask algorithm on the right. The diagrams shown are meant to clarify the differences between the two algorithms, but the actual implementations combine some of the steps.

4. EXPERIMENTAL RESULTS

4.1. Hardware System and Real-Time Software

The experimental setup of Figure 2 uses an NEC NP-U300X projector, a HOPS Glass holographic diffusing screen produced by Sax3D, an nVidia FX4800 graphics board, a Matrox Helios XCL frame grabber, and a Basler a504kc high speed camera. We used the 3D shutter glasses signal (60Hz) from the graphics board to provide synchronization between camera and projector. Hardware and a TI microprocessor provided a frequency doubler that allowed operation at 120hz, the maximum display rate of the projector. The camera-projector timing relationship is shown in Figure 3.

The value k in (1) affects different aspects of the CTM reconstruction algorithm. Higher k is desired for higher average brightness and less flicker in the displayed video. The average brightness, with respect to full projection, is $B = 0.5(1+k)$ so that for $k = 0.5$, $B = 75\%$. Higher k, on the other hand, causes undesirable amplification of noise power (variance), given by $H = [(\frac{1}{1-k})^2 + (\frac{k}{1-k})^2]$ which for $k = 0.5$ is 5. Higher k is also undesired because $G = \frac{k}{1-k}$ in (3) also amplifies motion related pixel differences. In practice, the motion artifacts are much more objectionable than the noise amplification (which was not visible in our experimental setting). We empirically chose $k = 0.5$ in light of these tradeoffs.

To allow repeatable and quantitative performance testing, we simulated a video conference by replacing the remote video with a pre-recorded video of a remote participant, which we projected. A local user viewed the pre-recorded video and the camera captured the local user with cross-talk. We implemented software running in real-time which we can switch between three settings: *No processing*, *Naive*, and *Pass-mask*. Informal visual inspections during system operation provided the following findings: 1) Cross-talk with *No processing* is noticeable and cross-talk reduction is required, 2)*Naive algorithm* works well with little motion in the local signal but annoying artifacts are very visible when there is motion in the local signals. 3)*Pass-mask algorithm* works well with or without local motion. The next section provides quantitative simulation results.

4.2. Results and Comparison

Since the performance of cross-talk reduction methods is dependent on the remote and local signals, we need to use the same remote and local signals for each reconstruction method to get objective quantitative results and comparisons. Repeating remote signal is straightforward since we just need to play the same video on the projector. However, duplicating the exact same local scene, when there is motion, is impossible. Thus, we first recorded cross-talk of the remote signal in a dark room and mixed the recorded cross-talk with another pre-recorded local video for simulation. Call the local frame $\mathbf{L}$, the cross-talk frame $\mathbf{C}$, and the reconstructed frame $\hat{\mathbf{L}}$. To obtain quantitative cross-talk reduction rates, we define the following quantities: Power of frame $\mathbf{F}$ is $\mathcal{P}(\mathbf{F}) = \frac{1}{MN}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} \mid \mathbf{I}(m,n) \mid^2$, where, M and N are the height and width of an image respectively. The cross-talk residual is $\mathbf{C}_{residual} = \hat{\mathbf{L}} - \mathbf{L}$, and the cross-talk reduction rate is $\eta = \frac{\mathcal{P}(\mathbf{C}) - \mathcal{P}(\mathbf{C}_{residual})}{\mathcal{P}(\mathbf{C})}$.

First, we compare the *Pass-Mask algorithm* with the *Naive algorithm*. The $\mathbf{R_{max}}$ vector measured and used for our experiments is $[68 \;\; 78 \;\; 102]^T$, where 8 bit pixel values are between 0 and 255. Experimental results are best seen in video, but Figure 5 shows the results of the *Naive* algorithm on the left, and the *Pass-Mask* algorithm on the right. On the bottom of the arm, and below the shadow of the hand, the naive algorithm results in very visible artifacts, whereas the pass-mask algorithm substantially reduces these artifacts.

We operated the real-time cross reduction system for a long time, with good results. We also stressed the system by using a video with high motion, an arm waving very fast at about $2 \sim 3$ Hz. Figure 6 shows the cross-talk reduction rate, η, for 155 frames. The observed

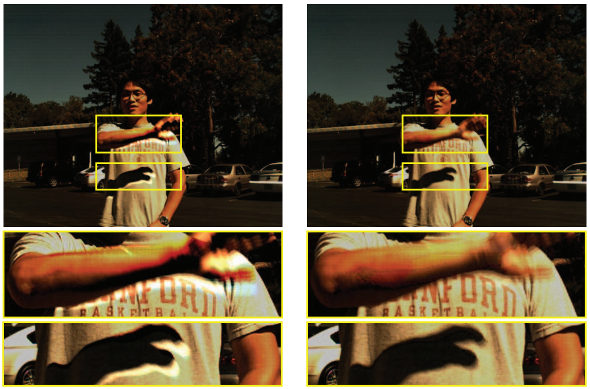

Fig. 5. Results of the *Naive algorithm* on the left, and the *Pass-Mask algorithm* on the right. Note the disturbing white or light haloes near the arm and the hand's shadow in the *Naive*. These artifacts are not visible in the *Pass-Mask* results, and although we do not reduce cross-talk in these regions, it is not noticeable.

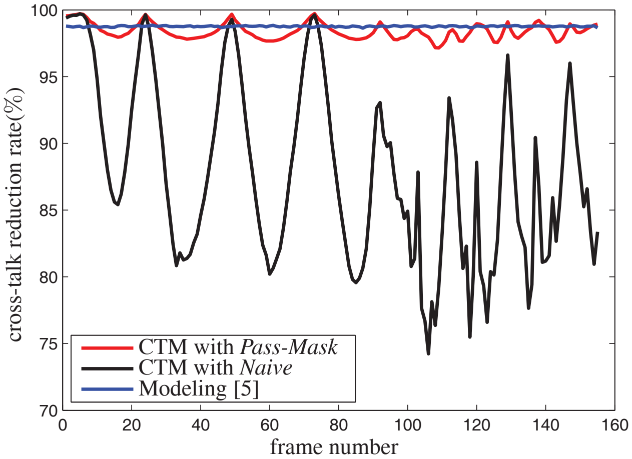

Fig. 6. The red, black, and blue curves are the cross-talk reduction rates of the *Pass-Mask*, the *Naive*, and the method of [5] respectively

periodicity in the plots is caused by first repetitive waving of the hands and then the body swaying in the test video used for the local signal. The average reduction rates of *Pass-Mask* and *Naive* are **98.4%** and **87.9%** respectively. The peak points in the plots correspond to frames with very little motion in the local signal. For those peak points, the performance of *Pass-Mask* and *Naive* are very similar, as expected, since *Pass-Mask* reconstruction is identical to *Naive* reconstruction when there is no motion in the local signal.

Next, we compare the cross-talk reduction performance of the *Pass-Mask algorithm* with the modeling based approach of [5]. From Figure 6, the *Pass-Mask* algorithm has higher fluctuation in performance than the modeling based method of [5]. This is because the CTM technique compares two consecutive frames, thus is has lower performance with *motion* in the local signal. On the other hand, the modeling method focuses on the cross-talk itself, thus its performance is fairly independent of the local signal. When there is little motion in the local signal, the performance of CTM is better than that of the modeling based method. The average reduction rates for the *Pass-Mask* and the modeling based method of [5] are very similar, **98.4%** and **98.8%**, respectively.

Compared to prior work, *Pass-Mask* achieves a high cross-talk reduction rate without requiring complex and computationally expensive geometric and photometric modeling of [5], but it does require camera-projector synchronization. Although the cross-talk reduction rate, η, decreases with high motion in the local signal, we have informally observed that the video quality is high, and the cross-talk is not seen. The reduction rate is always higher than 97% and the faint artifacts are masked near the motion boundaries. Thus, in terms of implementation, computational load, and quantitative and visual results, the *Pass-Mask algorithm* using the CTM technique offers an attractive solution for the visual cross-talk reduction problem.

5. CONCLUSIONS

We described a solution to video cross-talk reduction by using temporal modulation of one of the video signals, and the adaptive reconstruction algorithm called *Pass-Mask*. This algorithm uses simple, adaptive operations easily implemented in real-time. The approach performs well in the presence of motion in the local signal, and the frequency doubling of the input to the reconstruction system altogether removes the effect of motion in the remote signal. The approach has been tested using the collaboration testbed described, but the general approach of intentionally modulating the projector signal, and subsequently separating the video signals, applies to diverse system configurations and architectures that require display of one video signal while capturing another video signal through the same surface. Even though our experiments used fixed camera settings, as long as camera automatic gain settings change slower than the 120 Hz, the *Pass-Mask* algorithm is still effective since it only compares two consecutive frames. $\mathbf{R_{max}}$ is affine(i.e. linear with constant bias) to exposure, so knowing it for two gain settings should allow gain changes to be tracked without requiring new calibration.

6. REFERENCES

[1] K.-H. Tan, I. Robinson, R. Samadani, B. Lee, D. Gelb, A. Vorbau, B. Culbertson, and J. Apostolopoulos, "Connectboard: A remote collaboration system that supports gaze-aware interaction and sharing.," in *IEEE MMSP 2009*, Rio de Janeiro, Brazil, Oct. 2009.

[2] 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 of the 21st annual ACM symposium on User interface software and technology*. ACM New York, NY, USA, 2008, pp. 269–278.

[3] M. Liao, M. Sun, R. Yang, and Z. Zhang, "Robust and Accurate Visual Echo Cancelation in a Full-duplex Projector-camera System," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 10, pp. 1831–1840, 2008.

[4] M. Hirsch, D. Lanman, H. Holtzman, and R. Raskar, "Bidi screen: a thin, depth-sensing lcd for 3d interaction using light fields," *ACM Transactions on Graphics (TOG)*, vol. 28, no. 5, pp. 159, 2009.

[5] R. Samadani, J. Apostolopoulos, I. Robinson, and K.-H. Tan, "Video cross-talk reduction and sychronization for two-way collaboration," in *IEEE ICIP 2010*, Hong Kong, Sept 2010.

[6] A.M. Bronstein, M.M. Bronstein, M. Zibulevsky, and Y.Y. Zeevi, "Sparse ICA for blind separation of transmitted and reflected images," *International Journal of Imaging Systems and Technology*, vol. 15, no. 1, pp. 84–91, 2005.