VIDEO CROSS-TALK REDUCTION AND SYNCHRONIZATION FOR TWO-WAY COLLABORATION

Ramin Samadani, John G. Apostolopoulos, Ian Robinson and Kar-Han Tan

Multimedia Communications and Networking Lab, HP Labs

ABSTRACT

Recent two-way collaboration prototypes attempt to improve natural interactivity, correct eye contact and gaze direction, and media sharing using novel configurations of projectors, screens, and video cameras. These systems are often afflicted by *video cross-talk* where the content displayed for viewing by the local participant is unintentionally captured by the camera and delivered to the remote participant. Prior attempts to reduce this cross-talk purely in hardware through various forms of multiplexing (e.g., temporal, wavelength (color), polarization) have performance and cost limitations. In this work, careful system characterization and subsequent signal processing algorithms allow us to reduce video cross-talk. The signals themselves are used to detect temporal synchronization offsets which then allow subsequent reduction of the cross-talk signal. Our software-based approach enables the effective use of simpler hardware and optics than prior methods. Results show substantial cross-talk reduction in a system with unsynchronized projector and camera, and with cross-polarizing filters for nominal separation of the video signals.

Index Terms— video cross-talk, visual echo cancellation, immersive collaboration, synchronization, projector-camera systems.

1. INTRODUCTION

Two-way collaboration systems attempt to provide natural interaction among participants at two different locations. Desired attributes of these systems include natural frontal view with correct eye contact and eye gaze, as well as the ability to interact on a shared surface. This is achieved by both displaying information for each local participant and capturing information to deliver to the remote participant through the same surface. These systems are often afflicted by *video cross-talk* which arises when the video signal to be displayed to the local user interferes with the local video signal that one desires to capture with the camera. Prior systems used optics and hardware to separate the cross-talk signal: [1] used synchronized temporal multiplexing of the video signals, and [2] used wavelength multiplexing using precision multiple-passband optical filters. A related prior approach [3] applied to two remote participants working with shared projected content on white boards, and a portion of the view of the camera contained the desired local signal, e.g., text or hand drawings on a white board, and segmentation was applied to classify each pixel as cross-talk or not. Additional emerging applications in portable configurations [4] may also require cross-talk reduction.

Our method reduces video cross-talk via algorithms implemented in software, significantly reducing cost and hardware complexity compared to conventional hardware and/or physics-based approaches. Unique aspects of our work include the use of the signals themselves for synchronization, the careful modeling of the forward path from projector to camera, including space-varying blur, and a temporal mixture model for the reconstruction that works well even when there are large motions or scene transitions in the cross-talk signal. Figure 1 shows a video frame with cross-talk and the corresponding frame with cross-talk reduced using our approach.

Fig. 1. Cross-talk input and output of the algorithm.

2. PROBLEM SETUP

Figure 2 shows the configuration of the system described in [2], but using polarizers instead of multiple-passband optical filters for the signal separation. The figure shows the light paths and the signals used by the software processing pipeline described in this paper. The local participant is shown to the left of a holographic diffusing screen. On the screen is projected the image of the remote participant, as well as any additional information intended to be displayed. The characteristics of the holographic screen allows viewing of the local participant by the camera, but in addition there is visual *cross-talk* that comes from the back-scattered projected light. Even though orthogonal polarizers in front of the projector and the camera minimize cross talk, the residual cross talk is still visible and objection-

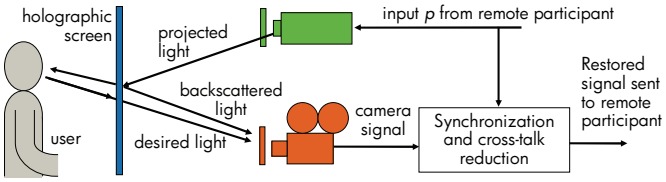

Fig. 2. Light paths and signals for the experimental system.

able, as seen in the top image of Figure 1. The human visual system is highly sensitive to structured patterns, and therefore the cross-talk is more noticeable than, e.g., noise with an equal amount of power.

The projector digital video signal, $p(m_1, m_2, l)$, is also sent to software to modify the camera video signal, $c(n_1, n_2, k)$, by first detecting synchronization offset and subsequently generating cross-talk reduced signal $\hat{c}_d(n_1, n_2, k)$. The signals $p(m_1, m_2, l)$ and $c(n_1, n_2, k)$ are unsynchronized, or in other words, the unknown time difference between l and k is not fixed nor is it an integer. For example, the camera and projector may operate at different capture and display frequencies, and even if they are at the same frequency there can be a time-varying phase (time offset) between them. Therefore, our solution also needs to address the synchronization problem.

3. PROPOSED SOLUTION

Our proposed solution starts conceptually at the back of the screen (the side towards the camera in Figure 2). The linearity of light means that the radiance emerging from the back of the screen is,

$$s(x, y, t) = s_p(x, y, t) + s_d(x, y, t) \quad (1)$$

where continuous signal $s(x, y, t)$ represents radiance, composed of two terms: (1) $s_p(x, y, t)$, from the video of the remote participant displayed by the projector, and resulting in the cross-talk signal in the camera, and (2) from the desired light $s_d(x, y, t)$ coming from the room, containing the local participant to the left of the screen. At the camera, because of the polarizer, the two signals are attenuated differently, but linearity continues to hold because we control the camera and enforce its linearity by setting the camera gamma to the identity. The resulting video frames at the camera are given by

$$c(n_1, n_2, k) = c_p(n_1, n_2, k) + c_d(n_1, n_2, k). \quad (2)$$

Where the functions $c()$, $c_p()$ and $c_d()$ are 3-color functions of discrete spatial indices n_1 and n_2 and discrete temporal index k (color index not indicated for simplicity).

The algorithm inputs are the corrupted signal $c(n_1, n_2, k)$ and a short sequence of frames $p(m_1, m_2, l)$ for $l \in [l_{min}(k), l_{max}(k)]$. The output is an estimate of the desired $c_d(n_1, n_2, k)$. Linearity allows us to solve the signal subtraction problem for any arbitrary interfering cross-talk signal. In the related prior approach [3] participants worked with shared content on white boards, and the desired local signal, e.g., text or hand drawings on a white board covered only a portion of the board. The authors used image segmentation to classify pixels as cross-talk or not, and tuning parameters reduced the misclassification errors. In our problem scenario the entire view of the camera contains desired signal as well as cross-talk and it is not possible to segment the cross-talk artifacts for removal.

3.1. Projector-Camera Static Forward Model

We estimate the transformation (forward model $f()$) from projector signal $p(m_1, m_2, l)$ to camera cross-talk signal $c_p(n_1, n_2, k)$ to subtract the estimated signal $\hat{c}_p(n_1, n_2, k) = f(p(m_1, m_2, l))$ from Equation 2. This section covers the transformation from a single projector frame to the camera signal, and Sections 3.2 and 3.3 cover the required temporal aspects needed for unsynchronized projectors and cameras. To obtain good results, all of the photometric, geometric and optical factors that comprise $f()$ need characterization. Related work in projector-camera modeling [5, 6, 7, 8] developed inverse models in order to modify multiple projector input signals to result in uniform and well-blended signals on a screen. The camera is used incidentally in these applications to characterize the inverse model but it is not used during operation. In our case, in addition to the forward model, we need signal subtraction at the camera to provide the cross-talk reduced signals to the remote participants. The techniques used in the related works are similar to ours at a high level, but the technical details differ for each component. In addition, we are not aware of prior work on space-varying blur modeling or temporal mixing, as is required to obtain good results for our application.

Fig. 3. The static portions of the forward model transforming projector input signals $p(m_1, m_2, l)$ to camera cross-talk signal $\hat{c}_p(n_1, n_2, k)$.

Our model uses static (time-invariant) characterizations of: 1) color transformation; 2) geometric transformation; 3) space-varying color gain and offset, and 4) space-varying blur. Figure 3 shows a block diagram of the transformation. An example of the predicted cross talk from an input projector frame is shown in Figure 4.

Fig. 4. Digital input sent to projector $p(m_1, m_2, l)$ shown on the left and the transformed cross-talk signal $\hat{c}_p(n_1, n_2, k)$ by applying the transformation shown in Figure 3 is shown on the right.

Video test patterns including color patches, grid patterns, horizontal and vertical stripes, and uniform white, black and gray level signals are sent to the projector while the room is dark to estimate the different parameters of $f()$. We begin at step *space-varying offset and gain* in Figure 3. By averaging captured uniform white video frames and black video frames, we determine the (spatially-varying) white response, $W(n_1, n_2)$, and the black response, $B(n_1, n_2)$, of the projector-camera system. For input $c_I(n_1, n_2, k)$, the output is given by

$$c_O(n_1, n_2, k) = c_I(n_1, n_2, k)[W(n_1, n_2) - B(n_1, n_2)] + B(n_1, n_2) \quad (3)$$

Given this gain offset transformation, the global color transformation is determined next by fitting between measured colors and color values $c_I()$ computed using the inverse of Equation 3. Measured average color values for gray input patches are used to determine 1D

lookup tables applied to the input color components, and measured average color values for primary R,G,B inputs are used to determine a color mixing matrix using the known digital input color values. Computing the fits using the spatially renormalized colors allows our color model to fit the data with a small number of parameters.

The geometric transformation is determined using a traditional polynomial mesh transformation model [5].

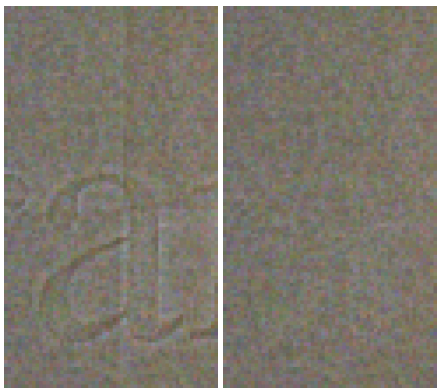

Fig. 5. Blur uncompensated on the left, but compensated on the right. The images are enhanced for print, but the unenhanced structured artifacts are very visible and objectionable in the videos.

The final space-varying blur step shown in Figure 3 is required to obtain good results at edges in the cross-talk signal. If the blur is not applied, objectionable halo artifacts remain visible in the restored signal, as seen in Figure 5, where the left shows results when the space-varying blur is not modeled (halos of the letter "a" and vertical step edge are visible), and the right shows improved results that incorporate the space-varying blur model.

The parameters of the space-varying blur were determined by estimating separable blur kernels in the horizontal and vertical directions. Captured horizontal and vertical step edges at different locations in the frames were fit using scaled *erf* functions. The standard deviations σ of best fit are also the parameters for the space-varying Gaussian blurs we apply. The range of values found are $\sigma \in [1, 4]$. The sparsely sampled blur estimates, 50 points each for horizontal and vertical estimates, were interpolated to a dense set of horizontal and vertical blur parameters,$\sigma_h(n_1, n_2)$ and $\sigma_v(n_1, n_2)$.

Direct implementation of space-varying blur,

$$c_b(n_1, n_2) = \sum_{n_1', n_2'} G(n_1, n_2, n_1', n_2') c_u(n_1', n_2') \tag{4}$$

is expensive so we modified (to provide unity gain filters) the efficient method described in [9] where Gaussian filters of arbitrary width are approximated by a linear combination of space invariant Gaussian filters of predetermined width. The linear (but shift variant) operation of Equation 4 is now approximated by

$$c_b(n_1, n_2) \approx \sum_i \alpha_i(n_1, n_2) \sum_{n_1', n_2'} G_i(n_1 - n_1', n_2 - n_2') c_u(n_1', n_2'). \tag{5}$$

For our implementation, $i = 4$, so that four separable convolutions, are followed by pixel-wise linear combination with weights $\alpha_i(n_1, n_2)$ that are precomputed for efficiency.

3.2. Signal-based synchronization

Prior methods have used hardware synchronization [1] or have added decodable bar-code symbols into the video stream [3]. Here, we use a signal processing method that has the advantage that it does not require additional hardware or additional information placed in the video signal. We find that a method that processes single frames at a time is sufficient to detect the synchronization offset between the digital projector stream $p(m_1, m_2, l)$ and the captured camera video $c(n_1, n_2, k)$. The method consists of applying the forward model to the projector signal in a buffer of frames to generate the estimate for the cross talk signal $\hat{c}_p(n_1, n_2, l)$ and computing the projection

$$\hat{l} = \arg\max_l \frac{\sum_{n_1, n_2} d(\hat{c}_p(n_1, n_2, l)) d(c(n_1, n_2, k))}{\sum_{n_1, n_2} d(\hat{c}_p(n_1, n_2, l))^2} \tag{6}$$

Where $d()$ is a bandpass filter that detrends and mean-subtracts its input signals, without which spurious matches may occur. Equation 6 is similar to a bank of matched filters, where the filters are the estimates $d(\hat{c}_p(n_1, n_2, l))$ for different l values. The value $\hat{l}$ identifies one of the interfering frames in our unsynchronized system.

3.3. Temporal mixture estimation

Lack of synchronization means that the cross-talk signal at the camera may arise from a number of frames projected on the screen during the camera frame integration time. In the following we assume that the nominal projected video and camera frame rates are the same, however the approach generalizes to the case of different frame rates. When the nominal display and camera frame rates are the same, we assume the cross-talk originates from two consecutive projected frames. The signal-based synchronization identifies one of the interfering frames, $\hat{l}$, as having the maximum effect. For simplicity assume that the two frames $p(m_1, m_2, \hat{l})$ and $p(m_1, m_2, \hat{l} + 1)$ produce the cross-talk (we discuss shortly how to identify if it is frames $\hat{l}$ and $\hat{l} + 1$ or $\hat{l}$ and $\hat{l} - 1$). The forward model for the cross-talk,

$$\hat{c}_p(n_1, n_2, k, \alpha) = \alpha f(p(m_1, m_2, \hat{l})) + (1-\alpha) f(p(m_1, m_2, \hat{l}+1)), \tag{7}$$

corresponds to the physical assumption that the projector displays frame $\hat{l}$ for a proportion α of the time, and the remaining camera capture time, $1 - \alpha$, the projector displays frame $\hat{l} + 1$.

To estimate α, we use a total variation measure, elsewhere applied for image denoising and restoration. The total variations of a differentiable image $I(x, y)$ is defined [10] by

$$tv(I(x, y)) \equiv \int_\Omega \mid \nabla I(x, y) \mid d\Omega. \tag{8}$$

Approximating tv using the sum of absolute values of horizontal and vertical differences of a frame, we determine α by minimizing

$$\hat{\alpha} = \arg\min_\alpha tv[c(n_1, n_2, k) - \hat{c}_p(n_1, n_2, k, \alpha)]. \tag{9}$$

Informal justification of Equation 9 is that signal $c(n_1, n_2, k)$ in Equation 2 has edges from desired signal $c_d(n_1, n_2, k)$ and spatially uncorrelated edges from cross-talk signal $c_p(n_1, n_2, k)$. Minimizing the total variation finds the signal that leaves only the edges of the desired signal. Finding $\hat{\alpha}$ uses a line search of a function computed using simple image differencing operations. The actual implementation involves two line searches since we do not know whether the required frames correspond to times $\hat{l}$ and $\hat{l} + 1$, or $\hat{l}$ and $\hat{l} - 1$.

The total variation method has been very reliable in our tests. The center image of Figure 6 shows that estimating the cross-talk from single frames is not effective (checkerboard and text cross-talk remain) but the temporal model used in the right image successfully reduces the cross-talk.

Fig. 6. Left shows cross-talk from two consecutive frames, middle is the estimate using only one frame, and the right is the estimate using the proposed temporal mixture. The images are enhanced for print.

4. EXPERIMENTAL RESULTS

The experiments conducted using the setup in Figure 2 used a short-throw NEC WT610 DMD projector, HOPS Glass holographic screen produced by Sax3D and a Point Grey Grasshopper 20S4C camera. To determine the parameters for the forward model we generated the test patterns described in Section 3.1, projected them in a dark room, and conducted the characterization described in Section 3.1.

To test the performance of the system we pre-recorded a remote scene of a collaboration participant and played it on the projector, with a local participant viewing the video. The camera captured this video with cross talk. Then we applied the synchronization of sections 3.2 and 3.3, and conducted the cross-talk reduction described in sections 3.1 and 3.3. Figure 1 shows one frame from the resulting video. Viewing the reconstructed video showed significant reduction in cross-talk throughout the video. The videos used for testing were not used to train any of the model parameters.

The digital frames causing the cross-talk are, in most cases, hard to visually identify in the corrupted frames, and ground truth is not available since the projector and camera are not synchronized. But we find that if incorrect frames are used, visible halos occur at moving objects. Since we do not observe such artifacts, the synchronization detection serves its purpose and we presume it is working correctly. Further confirming correctness, for some easily identifiable cross-talk digital frames we have visually confirmed that the synchronization identified the correct frame.

Quantitative results were obtained with a pre-recorded video played on the projector and captured by the camera in an otherwise dark room. The power of captured (cross-talk) signal $c_p(n_1, n_2, k)$ was compared with the power of the residual after cross-talk reduction. The reconstructed video appeared black, but there remained a small amount of residual power. The residual power was on average -18.8 dB below cross-talk power (for 155 frames; 10 seconds at 15 frames per second). Mean and variance were incorporated in the power calculation since a constant mean cross-talk signal with no variance still interferes with the desired signal.

5. CONCLUSIONS

We have shown an effective algorithmic method for projector-camera synchronization and subsequent reduction in video cross-talk by using system modeling and signal processing. This methodology enables flexible use of cheap optical and system components. We used closed-loop projector-camera characterization but the methodology also applies to separate characterization of projector-screen and screen-camera subsystems.

6. REFERENCES

[1] 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.

[2] 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.

[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] P. Mistry, P. Maes, and L. Chang, "WUW-wear Ur world: a wearable gestural interface," in *Proceedings of the 27th international conference extended abstracts on Human factors in computing systems*. ACM, 2009, pp. 4111–4116.

[5] M. Harville, B. Culbertson, I. Sobel, D. Gelb, A. Fitzhugh, D. Tanguay, et al., "Practical methods for geometric and photometric correction of tiled projector displays on curved surfaces," in *Proc. IEEE International Workshop on Projector-Camera Systems (ProCams)*, pp. 52–59.

[6] N. Damera-Venkata, N. Chang, and J. Dicarlo, "A Unified Paradigm For Scalable Multi-Projector Displays," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1360, 2007.

[7] M. Brown, A. Majumder, and R. Yang, "Camera-based calibration techniques for seamless multiprojector displays," *IEEE Transactions on Visualization and Computer Graphics*, pp. 193–206, 2005.

[8] W. Sun, I. Sobel, B. Culbertson, D. Gelb, and I. Robinson, "Calibrating multi-projector cylindrically curved displays for wallpaper projection," in *Proceedings of the 5th ACM/IEEE International Workshop on Projector camera systems*. ACM, 2008, p. 1.

[9] Javier Portilla and Rafael Navarro, "Efficient method for space-variant low-pass filtering," in *VII National Symposium on Pattern Recognition and Image Analysis*, Barcelona, Spain, 1997, vol. 1, pp. 287–292.

[10] P.L. Combettes and J.C. Pesquet, "Image restoration subject to a total variation constraint," *IEEE Transactions on Image Processing*, vol. 13, no. 9, pp. 1213–1222, 2004.