Free Viewpoint Teleconferencing Using Cameras Behind Screen

Sehoon Lim, Luming Liang, Yatao Zhong, Neil Emerton, Tim Large, Steven Bathiche
Microsoft Applied Sciences, Redmond WA

Abstract

Natural interaction in videoconferences can be made by correcting gaze/perspective, scale, and position by using cameras placed behind a partially transparent front emitting OLED panel and frame interpolation of deep neural networks. The diffraction artifacts from through-screen imaging are removed by a deconvolution method.

Author Keywords

Viewpoint; teleconferencing; camera behind screen; through-screen imaging; organic light-emitting diode (OLED); diffraction; point-spread function (PSF); signal-to-noise ratio (SNR); modulation transfer function (MTF); deconvolution; convolutional neural network (CNN); deep neural network (DNN); frame interpolation; frame synthesis.

1. Introduction

Teleconferencing has always been powerful tool to connect people. Even though the quality of compression, cameras, and displays have steadily been improving for decades, telepresence is still challenged in reproducing the physical ques and dynamics that we are used to in physical person-to-person's meetings. Some of these physical dynamics include front-parallel pose, human scale, 3D perceptive, view-dependent rendered image, and eye contact. Even with holographic 3D displays [1], market's feedback has not been successful. The hardware system is complex and expensive. The data transfer of 3D experience is also limited by commercial network speeds for standard image quality. For example, video conferencing software such as Teams and Zoom takes 2D video image in one end and shows 2D display image in another end. Gaze correction is a method to improve eye contact in 2D image data by eye repainting which corrects for the angular offset of the camera and the center of the screen [2]. However, this experience is limited to small screens where the angular offset is limited.

Gaze and perspective awareness are a long-standing problem in person-person teleconferencing for productive inter-personal communication. Associated with gaze correction are several other issues which add to the sense of abnormality in video conferences, including incorrect space perspective, inappropriate scaling of the remote participants, socially awkward spatial location of conference participants (Figure 1). The angular offset between the camera and the displayed image of the remote participant mean that the participants will not experience a sense of eye contact (Figure 2). Looking directly into either the eyes on the screen or the camera breaks the eye contact and may miss subtle non-verbal feedback cues. Spatial factors also affect conversational dynamics but are not taken into accounts by current videoconferencing systems. Both the arrangement of participants relative to each other and the distance between them are meaningful aspects of non-verbal communication, as illustrated in Edward Hall's work on proxemics [3], and Adam Kendon's work on F-formations [4]. Such factors could be applied to the virtual environment of a remote conversation by adjusting the speaker's position and size on the display.

Figure 1. Spatial awareness (left), interaction spaces (middle), and proxemics (right)

Figure 2. Camera behind screen (left) vs. Camera at the top bezel (right)

We propose a solution using transparent OLED display with hidden cameras behind the panel. The users are segmented, scaled, viewpoint corrected to achieve a more natural conversational situation. Modern image processors are employed both to detect people in the scene and make scene adjustments, and to correct the diffraction resulting from through-screen imaging. The concept of embedding image sensors into displays is not new, but dates back to the early days of computing when it was first recognized that lack of gaze awareness in video conferencing. The problem is created by the mismatch between the view of the camera and the position of the remote participant on the screen. Modern video conferencing systems still do not solve this problem.

Using the camera behind the display reduces the offset between the point of view of the remote person, and the image of the remote person on the screen. In the literature, a user typically detects true eye contact if the discrepancy is less than ~3 degrees. It is important to note this is about allowing the user to detect the remote participant's intention or attention, it is not about fixing gaze direction. Industrial design also demands to minimize the display's bezel widths by eliminating camera notch and hole. It effectively increases screen utilization resulting in benefits of more information as well as aesthetics in user's experience.

We demonstrate an end-to-end teleconferencing system that delivers critical cues of gaze and perspective awareness using cameras behind screen. Cameras are placed behind a transparent OLED (t-OLED) which is specially designed to improve light transmission with minimal impact on image blur and color shift. To deal with image degradation through display a deconvolution method is used. A learning-based method of deep neural network (DNN) detects/segments main presenter which is rescaled/overlaid

 ISSN 0097-996X/21/5201-0218-$1.00 © 2021 SID

in the background. Cameras behind screen support the communicative ques/dynamics in the way that the correct viewpoint is synthesized by DNN using synchronized images and the correct position/scale are preserved. In this paper characterizations of t-OLED sample are presented for camera behind screen. Image processing for deconvolution and inter-frame synthesis are discussed. The concept of free viewpoint teleconferencing is demonstrated by experimental results.

2. Transparent OLED

Placing a camera behind a display conflicts with the needs of high-quality camera imaging which conventionally requires a clear aperture to receive enough uninterrupted light from the scene. The display panel in front of the camera prevents fulfillment of the imaging requirements by modulating the incident light due to the display's 3D structure. A large OLED display panel is typically a combination of stacked optical layers such as color filters, pixel structure, and a substrate. The pixel structure consists of anode, various bandgap management layers like hole injection layer, OLED, and cathode.

LG Display redesigned the OLED structure placing all the addressing lines, transistors, anode and color filters into opaque areas and the OLED material and cathode, which are substantially transparent, uniform everywhere. The resulting OLED structure is shown in Figure 3. Aspect ratios of open area are vertically 94% and horizontally 50%. The display resolution is FHD 1920x1080.

Figure 3. Microscope image

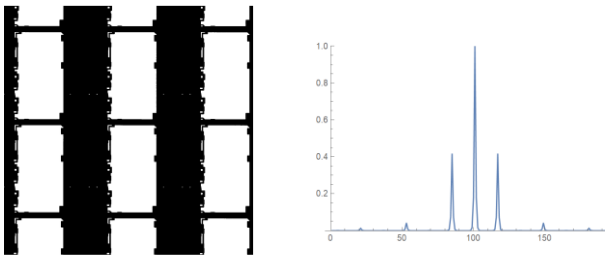

Figure 4. Aperture function and resulting horizontal diffraction pattern

The diffraction pattern produced by the display is calculated by Fourier transformation of the aperture function, shown in Figure 4. The display aperture pattern produces diffraction primarily in the horizontal direction where the open aperture duty cycle is 50%.

The display transmission varies from 35% in blue (460nm) to 45% in red (650nm) as shown in Figure 5. Although the t-OLED sample absorbs more light in blue than in red, this color error is readily corrected in camera image white balance.

The camera system and display combined MTF was calculated using Zemax OpticStudio and is shown in Figure 6. The sample's MTFs are different in horizontal and vertical directions. The

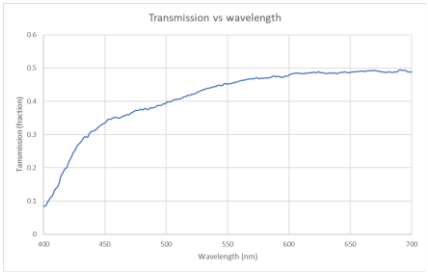

Figure 5. Spectral transmission

horizontal MTF has zeros after the main lobe however the vertical MTF monotonically attenuates in the spatial frequency. This one directional information loss results from the opening ratio in the panel design. The information loss causes one dimensional image blur through the display.

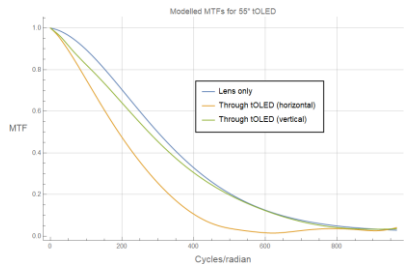

Figure 6: Calculated MTF

The measured system MTF values are shown in Figure 7. These are measured using slant edge technique [5,6].

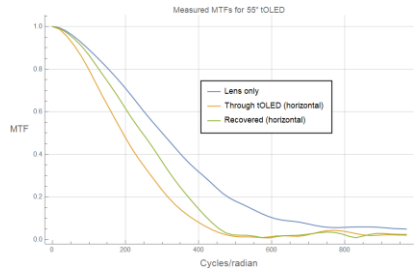

Figure 7: Horizontally measured MTF

The display has a wide viewing angle, good color gamut and adequate luminance for office applications even though it has a very high aperture ratio, as shown in Figure 8.

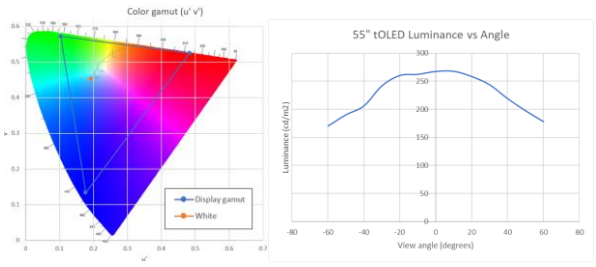

Figure 8. Display luminance, view angle and color gamut

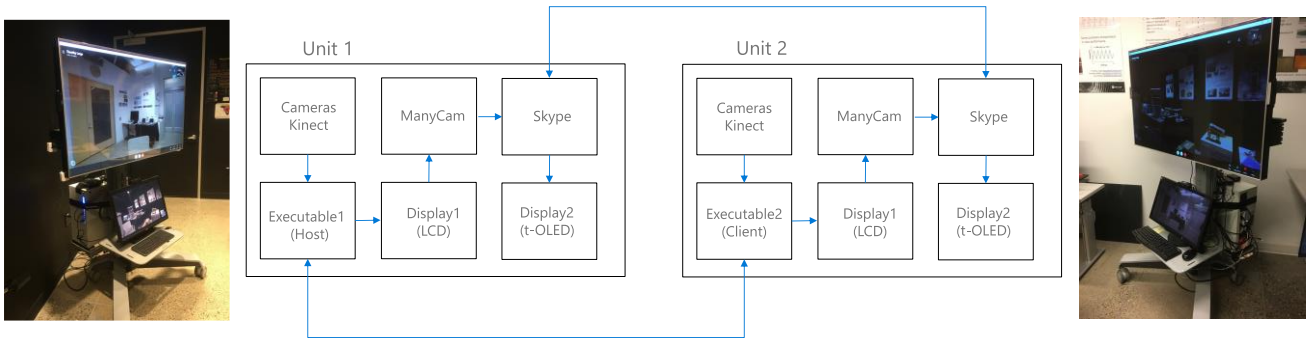

Figure 9. Bidirectional demo architecture

3. Cameras Behind Screen

Through-screen cameras allow free viewpoint experience. An end-to-end system is designed for bidirectional demo and it is digitally connected by commercial conferencing software of Skype over internet in Figure 9. Python code (Executable) controls cameras and a Kinect for measuring the position of user, and it communicates the positional information of user between the host and client. The code also controls the position and scale of foreground in display image to realize the communication dynamics and cues. As a result, it plays a role of functional wrapper for user experience in whole system. Kinect is a time-of-flight depth sensor with 512x424 pixels that enables body tracking and 3D positioning. ManyCam software virtualizes the conferencing

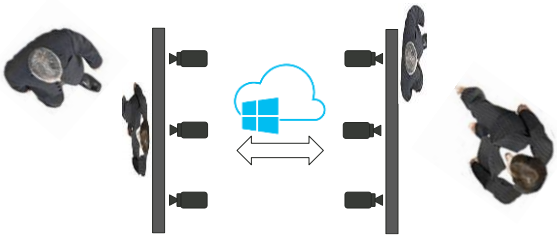

Figure 10. Foreground adjustment for correct viewpoint in F-formation

image that is fed into the Skype video channel to show in the other end. Three cameras are horizontally placed right behind the t-OLED covering the whole width of display in Figure 9. The user's position in one end determines correct viewpoint of the other user and the positional dynamics make F-formation (Figure 10). The foreground figure of user is segmented/scaled/synthesized/overlaid in the display with correct viewpoint. The gaze correction of bidirectional demo is demonstrated in Figure 11.

Audio was provided by Samson microphone, and speakers either side of t-OLED. Though existing telepresence systems like Skype only transmit mono audio, the system uses the position of the participant to synthesize their remote position in audio space by using left-right phase delays of the mono audio. Inter-aural time difference of the two channels is the main cue for Azimuthal localization. The binaural sound rendering augments realistic experience along with the video conferencing.

4. Image Sharpening and Interpolation

To improve the image quality, it is necessary to use image recovery

Figure 11. Gaze correction in demo operation

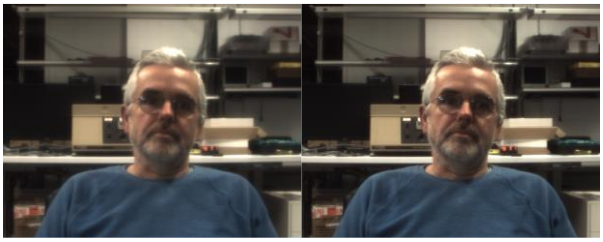

Figure 12. Before deconvolution (left) vs. After deconvolution (right)

to remove the blur caused by the display structure. It is also necessary to remove annoying camera switching as the participants move around, by interpolation.

The Wiener deconvolution method was used for deblurring. With this method the approximate spectral noise characteristics of the image are considered, and the filter weights the degree of deblurring according to the SNR ratio at a given spatial frequency. An example of deconvolved frames is shown in Figure 12.

In order to interpolate between camera images, we devised a convolutional neural network (CNN) structure to find the speaker within the image.

Video frame interpolation is popular to make high frame-rate video given data and hardware. The intermediate frame is interpolated by inferring motion estimation, occlusion reasoning and image synthesis between consecutive frames. This frame interpolation in temporal domain could be applied to frame interpolation for spatially separated cameras. In our system the correct viewpoint is

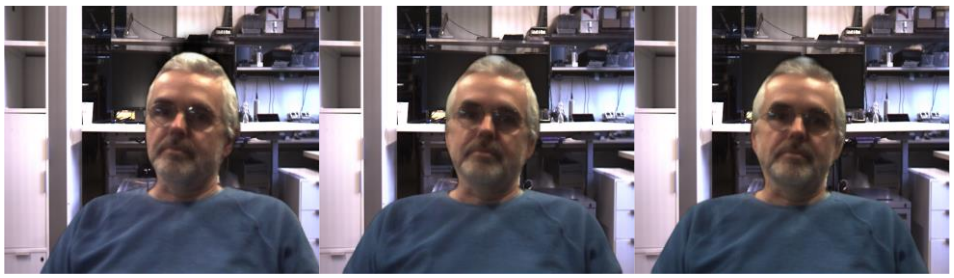

Figure 13. Horizontally synthesized viewpoints from dual camera feed

synthesized between two synchronized cameras behind the display.

Given two synchronized frames of I(0) and I(1), but spatially displaced, an interframe of I(t) is synthesized where $t \in (0, 1)$. We adopted a compression-driven network design for frame interpolation which prevents over-parameterization in DNN models. Deformable separable convolution (DSC) is the key component of frame interpolation. The deformable convolution adds pixel offset vectors and the separable means different pixels have different kernel weights that enables flexibility in handling large and complex motion. Dilation of the starting point of the offsets helps to explore wider area.

We compressed the baseline model for frame synthesis [7] by re-training the weights with sparsity constraint. The sparsity regularization increases the number of zeros in the weights and reduces the model complexity. Feature pyramid warped the frames at multiscale using five feature levels inherent from Unet's encoder [8]. This multiscale method is efficient and stable to capture the motion in the feature space. We used the Vimeo-90K dataset [9] for training, which contains 51312/3782 video triplets of size 256×448 for training/validation. We further augment the data by randomly flipping them horizontally and vertically as well as perturbing the temporal order.

Two of the three 2 MP FLIR cameras (Blackfly) produced 2 MP RGB images and they were linked by GPIO pins: one was set as master and the other as slave for synchronization. Without synchronization, the viewpoint of synthetic frame could be deteriorated by temporal motion of cameras. The master camera

Figure 14. Foreground overlaid on background

triggers the slave camera at every active exposure. The camera lens of aperture F/2 was focused on a test chart in 2-meter distance. The camera operates at 20 FPS, 20 ms shutter speed, and data is output using raw 16-bit image format. Furthermore, the camera gain was set to 15 dB to compensate light attenuation from the display. The same camera conditions were used for both the cameras.

A horizontally synthesized viewpoint from two cameras were also shown in Figure 13. The left and right-hand images showed two extreme cases of perspective and the middle one showed the intermediate perspective between them. The background occlusion from foreground was filled by previous frames in the video stream. There was no severe artifact due to the frame synthesis. Finally, the adjusted foreground was overlaid in a background with correct viewpoint and gaze in Figure 14.

5. Conclusion and Discussion

Machine learning and embedded cameras make possible a new class of more natural videoconferencing devices. This paper demonstrated critical communication cues could be retained by using cameras behind display and DNN computation in a compact hardware form.

6. Acknowledgements

The authors are grateful to LG Display for providing samples of transparent OLED and for helpful discussions on this project.

7. References

1. SeeReal Technologies. Core enabling technologies for holographic 3D displays (H3D) - SeeReal Technologies
2. The Verge. Microsoft's AI-powered 'Eye Contact' feature is finally coming to the Surface Pro X - The Verge; online magazine.
3. Hall, Edward T: The Hidden Dimension ISBN 0-385-08476-5; 1966, reprinted 1990.
4. Kendon, Adam: Conducting Interaction: patterns of behavior in focused encounters, ISBN 0-521-38036-7; 1990
5. CPIQ P. Standard for camera phone image quality. Institute of Electrical and Electronics Engineers (IEEE); 2015.
6. ISO/TC42/WG18. Resolution and spatial frequency response. International Organization for Standardization (ISO); 2014.
7. Lee H, Kim T, Chung T, Pak D, Ban Y, and Lee S. AdaCoF: Adaptive Collaboration of Flows for Video Frame Interpolation; CVPR; 2020
8. Ronneberger O, Fischer P, and Brox T. U_Net: Convolutional Networks for Biomedical Image Segmentation; Computer Vision and Pattern Recognition; May 2015.
9. Xue T, Chen B, Wu J, Wei D, and Freeman W. Video enhancement with taskoriented flow; International Journal of Computer Vision; 127(8):1106–1125; 2019.