Gaze Awareness and Interaction Support in Presentations

Kar-Han Tan, Dan Gelb, Ramin Samadani, Ian Robinson, Bruce Culbertson, John Apostolopoulos

HP Laboratories
HPL-2010-187

Keyword(s):

Distance learning, Gaze awareness, Paralanguage, Presentation recording, Remote presentation, Telepresence

Abstract:

Modern digital presentation systems use rich media to bring highly sophisticated information visualization and highly effective storytelling capabilities to classrooms and corporate boardrooms. In this paper we address a number of issues that arise when the ubiquitous computer-projector setup is used in large venues like the cavernous auditoriums and hotel ballrooms often used in large scale academic meetings and industrial conferences. First, when the presenter is addressing a large audience the slide display needs to be very large and placed high enough so that it is clearly visible from all corners of the room. This makes it impossible for a presenter to walk up to the display and interact with the display with gestures, gaze, and other forms of paralanguage. Second, it is hard for the audience to know which part of the slide the presenter is looking at when he/she has to look the opposite way from the audience while interacting with the slide material. It is also hard for the presenter to see the audience in these cases. Even though there may be video captures of the presenter, slides, and even the audience, the above factors add up to make it very difficult for a user viewing either a live feed or a recording to grasp the interaction between all the components and participants of a presentation. We address these problems with a novel presentation system which creates a live video view that seamlessly combines the presenter and the presented material, capturing all graphical, verbal, and nonverbal channels of communication. The system also allows the local and remote audiences to have highly interactive exchanges with the presenter while creating a comprehensive view for recording or remote streaming.

External Posting Date: November 21, 2010 [Fulltext] Approved for External Publication
Internal Posting Date: November 21, 2010 [Fulltext]
Presented at ACM Multimedia, Firenze, Italy, October 25, 2010

Gaze Awareness and Interaction Support in Presentations

Kar-Han Tan, Dan Gelb, Ramin Samadani,
Ian Robinson, Bruce Culbertson, John Apostolopoulos
Hewlett-Packard Laboratories
1501 Page Mill Road
Palo Alto, CA, USA

ABSTRACT

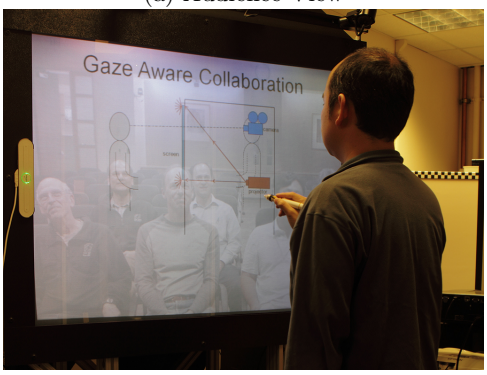
Modern digital presentation systems use rich media to bring highly sophisticated information visualization and highly effective storytelling capabilities to classrooms and corporate boardrooms. In this paper we address a number of issues that arise when the ubiquitous computer-projector setup is used in large venues like the cavernous auditoriums and hotel ballrooms often used in large scale academic meetings and industrial conferences. First, when the presenter is addressing a large audience the slide display needs to be very large and placed high enough so that it is clearly visible from all corners of the room. This makes it impossible for a presenter to walk up to the display and interact with the display with gestures, gaze, and other forms of paralanguage. Second, it is hard for the audience to know which part of the slide the presenter is looking at when he/she has to look the opposite way from the audience while interacting with the slide material. It is also hard for the presenter to see the audience in these cases. Even though there may be video captures of the presenter, slides, and even the audience, the above factors add up to make it very difficult for a user viewing either a live feed or a recording to grasp the interaction between all the components and participants of a presentation. We address these problems with a novel presentation system which creates a live video view that seamlessly combines the presenter and the presented material, capturing all graphical, verbal, and nonverbal channels of communication. The system also allows the local and remote audiences to have highly interactive exchanges with the presenter while creating a comprehensive view for recording or remote streaming.

Categories and Subject Descriptors

H.4.3 [**Information Systems Applications**]: Communications Applications—*Computer conferencing, teleconferencing, and videoconferencing* ; H.5.3 [**Information Interfaces and Presentation (I.7)**]: Group and Organization Interfaces—*Synchronous interaction*

MM'10, October 25–29, 2010, Firenze, Italy.

(a) Audience View



(b) Presenter View

Figure 1: A remote presentation where the audience sees the presenter's interactions with the slide material while the presenter sees the audience.

General Terms

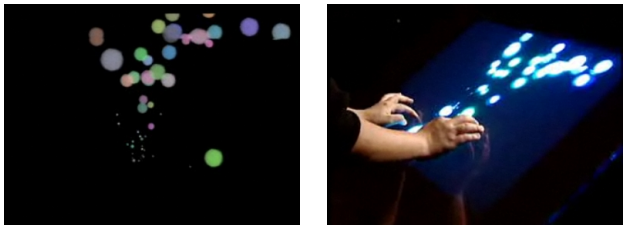
Algorithms, Design

Keywords

Distance learning, Gaze awareness, Paralanguage, Presentation recording, Remote presentation, Telepresence

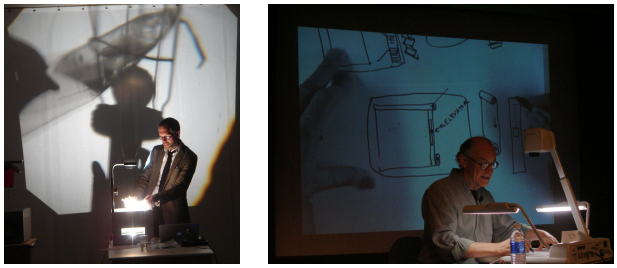
1. INTRODUCTION

One of the most significant applications of video communications is the *oral presentation*, in which a *presenter* disseminates information to an *audience* often aided by audio-visual material like PowerPoint/Keynote slides and video content. From the elementary school classroom to the corporate boardroom, and even in military command centers, presentations are critical tools for education, persuasion, and



(a) Content Only (b) Content + Hands

Figure 2: Two of the views presented to the audience during a popular TED Talk on novel touch interfaces. A recording of the talk is available online [5].



Overhead Projector Document Camera

Figure 3: Physical transparencies and documents allow natural presenter interaction by gesturing and sketching. Photos by (left) monnezza@flickr and (right) Matthias Müller-Prove.

coordination. With the advent of computers, displays, and networked multimedia communications, presenters are given the power to project their ideas and vision to larger and larger groups of audiences sometimes distributed around the globe. Time and again, we have seen great presenters launch their ideas, products, and almost immediately create global awareness of critical issues with highly effective presentations.

Many large scale presentations are highly polished productions that capture every key aspect of the presentation with multiple cameras and a production crew to strategically switch to the right views at the right moments to deliver the right messages. An example from the popular TED Talks is shown in Fig. 2, where the presenter was introducing a new multi-touch user interface. In the recorded video, the audience is shown the presentation from a number of different camera angles in addition to the visual content being synthesized by the multi-touch demos. In this particular example, it is crucial to show the audience what the presenter is doing with his hands on the multi-touch demo, and a camera is dedicated to showing a view of the presenter's hands, as shown in Fig. 2(b). Most presentations, even for some large scale events, are not as well-produced. In a typical presentation, only the slides are shown on the large presentation screen and broadcast to remote sites. One can easily imagine the diminished effectiveness of the same TED Talk if the audience was only shown the demo content, like that shown in Fig. 2(a). Even when a video stream of the presenter is provided, as many remote presentation systems offer today, it is still far from the engaging, visceral experience that presentations can be.

Perhaps not surprisingly, the more one relies on the presentation system to reach larger audiences, the more constraints one has in terms of interactions. Consider the conventional overhead projector or document camera (Fig. 3), where the presenter writes on transparencies or paper. The audience can easily see the real time interaction of the presenter's hands and pen with the slides. Conveying interaction with slides using a mouse pointer is considerably less expressive than using the presenter's hands.

Our aim is to find ways to identify some of the rich interaction tools available to presenters in more intimate, colocated settings, and make it possible for large scale presentations to have the same degree of interaction richness and attempt to go beyond the capabilities of current systems by capturing and conveying gaze awareness, and other forms of paralanguage [11], [12]. In this paper, we focus on the following issues:

Nonverbal communications When the presenter is addressing a large audience the slide display often needs to be very large and placed high enough so that they are clearly visible from all corners of the room. This makes it impossible for a presenter to walk up to the display and interact with the display with gestures, gaze, and other forms of paralanguage.

Gaze awareness It is hard for the audience to know which part of the slide the presenter is looking at when he/she has to look the opposite way from the audience while interacting with the presented material. It is also hard for the presenter to see the audience in these cases.

Our work is related to presentation and meeting capture systems and smart rooms or spaces [3], [4], [2], [8], [14], and [6]. As far as we know, all of the above systems employ conventional displays and camera systems and do not attempt to capture gaze awareness. The closest related works in spirit are [17], [9], and [10], which attempt to create composites of the presenter with shared media or whiteboard. While [17] successfully captures a presenter's interaction with a whiteboard, gaze awareness is sometimes not captured as the presenter's back would be facing the camera when working on the whiteboard.

2. OUR SOLUTION

Our solution is a presentation device based on a see-through display, as shown in Fig. 1(b). As the presenter interacts with slides shown on the display, a camera captures a video stream of the presenter and the system digitally combine the slides and presenter video stream to create a coherent view for the audience, as shown in Fig. 1(a). This allows gaze awareness and rich nonverbal communications to be captured and delivered.

2.1 See-through displays

A number of researchers have attempted to create see-through displays with various techniques. These include half-silvered mirrors [1], mirrors with polarizers [7], time division [15], and wavelength division [16]. We chose to use the wavelength division design of ConnectBoard [16] as the

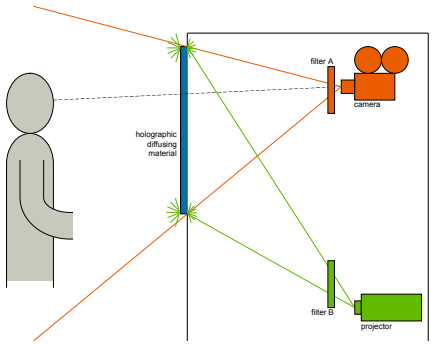


Figure 4: A see-through display.

system uses passive optical elements that are available off-the-shelf, and does not require custom electronics. The basic design is shown in Fig. 4. We have also built a touch sensing film and an electronic marker system into our presentation system so that one can naturally manipulate content and freely create sketches on the vertical surface.

2.2 Combining the visual signals

By separating the *presenter* and *shared media* light sources, our system enables compelling video presentations by allowing novel designs and special effects by *compositing* [13] the video information. In addition, the video is improved by avoiding quality degrading camera capture of the shared media.

For the compositing we use the well known compositing rules [13] in an RGBA representation where A represents an alpha channel with space and time varying alpha values $\alpha(x, y, t)$ with x and y spatial pixel coordinates, and time t. The novelty is in the different mechanisms for generating the alpha values for the compositing operation, but even the simplest approach with a global value $\alpha(x, y, t) = \alpha = \frac{1}{2}$ already provides good results because of the good quality of the digital transmission of the shared media.

Careful compositing of the two signals provides improved clarity through increased contrast and reduced visual masking. Figure 5 shows the processing block diagram. The input video media frames are content-analyzed (optionally jointly) in the *content analysis* block, which generates *per pixel* alpha values that are fed into the compositor to combine the media frames. We currently use alpha blending but any of the Porter-Duff operations [13] may be used. We tailor the $\alpha(x, y, t)$ values to preserve the contrast of the presentation information, by analyzing the shared media colors for slide presentations and preserving the lighter colors by setting $\alpha = max(R, G, B)$ where α represents the weight of the shared media.

Simple or sophisticated content analysis, including computer vision and image analysis techniques, may provide a variety of different effects. Here are some examples: 1) Slide transitions are detected. Dissolves that start by fully presenting the slides and fading into an adaptive alpha blend may highlight the information in the slides at each transition, focusing the viewer's attention; 2) Audio and video activity detection of the presenter may be used to modify the alpha values to emphasize more the speaker or emphasize more the shared media, depending on presenter gestures or motions; 3) Activity detection and spatial analysis of the shared media may be used to classify the shared media as slide presentation or video and different compositing treatments could be used for each class of shared material; 4) The size of the shared media may be automatically adjusted based on content analysis of the spatial frequencies in the shared media frames, for example small font size media may be enlarged for better visibility; 5) Depth based cameras may be used to only blend presenter information only when gestures are near the screen, emphasizing the gestures instead of the full presenter video.

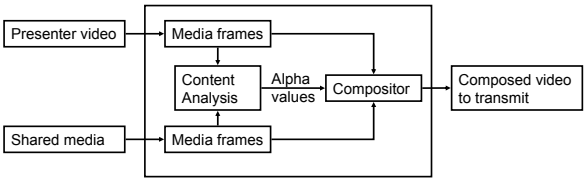


Figure 5: Video signal compositor processing block.

In addition to the automatic methods just described, the compositing operation may be overridden or be fully under interactive control of the presenter, the receiving audience or a professional A/V technician. No system modification is needed to allow control by the presenter or a local A/V technician but providing control to the receiving audience requires modifying the system to transmit separately the two video streams and conduct the compositing at the receiving processor.

2.3 System Architecture

We have created various processing components for audio and video capture, compression/decompression, networking, and rendering. Our system typically uses DirectShow compatible cameras for capturing the presenter and audience as they are widely supported on the Windows platform. We do support multiple camera interfaces beyond DirectShow that can be used as needed. For audio capture we rely on the ASIO standard as it enables synchronized multi-stream audio capture with well controlled input delay. In order to capture the presenter's materials in a general manner independent of what application is being used we rely on operating system calls to capture an image of the application as rendered on the GPU. While this typically requires a read-back of the application window from the GPU to main memory this still results in lower latency than is observed if the presenter's application is aquired using a video capture card or a similar method.

For video compression we rely on either H.264/AVC or Mpeg-2 codecs. The H.264 codec provides better compression for comparable signal-to-noise levels at the expense of additional computation requirements. Our audio signals are compressed using Mpeg-1 Layer 2 or AAC codecs. Compressed streams are sent using RTP. We do not retransmit lost frames in order to minimize latency. For video rendering we have created a flexible GPU-based compositor component. We can support multiple video streams with either per-pixel alpha values, or a single alpha value for the entire

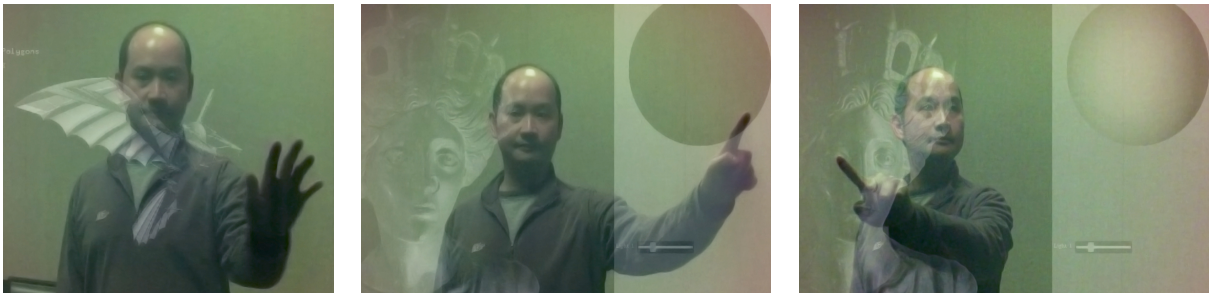


Figure 6: Examples of interactive applications.

stream. The compositor can perform dynamic repositioning and blending of the various streams as required.

3. CONCLUSION

We have presented a solution that addresses the issues of capturing rich nonverbal communications and creating a coherent view for the audience by combining a video stream of the presenter with the slides. Fig. 6 shows a presenter interacting with a 3D model and a relighting application. In all cases, it is easy for the local and remote audience to see where the presenter is looking and what they are pointing at. We believe that this novel capability enables an entirely new class of presentation systems that creates an enhanced experience for local and remote audiences, and makes it easier for the presenter to receive feedback from the audience.

4. REFERENCES

[1] S. R. Acker and S. R. Levitt. Designing videoconference facilities for improved eye contact. *Journal of Broadcasting and Electronic Media*, 31(2):181–191, 1987.

[2] P. Chiu, A. Kapuskar, S. Reitmeier, and L. Wilcox. Room with a rear view: Meeting capture in a multimedia conference room. *IEEE Multimedia*, pages 48–54, October–December 2000.

[3] R. Cutler, Y. Rui, A. Gupta, J. Cadiz, I. Tashev, L.-w. He, A. Colburn, Z. Zhang, Z. Liu, and S. Silverberg. Distributed meetings: a meeting capture and broadcasting system. In *ACM MULTIMEDIA '02: Proceedings of the tenth ACM international conference on Multimedia*, pages 503–512, 2002.

[4] G. Golovchinsky, P. Qvarfordt, B. van Melle, S. Carter, and T. Dunnigan. Dice: designing conference rooms for usability. In *CHI '09: Proceedings of the 27th international conference on Human factors in computing systems*, pages 1015–1024, 2009.

[5] J. Han. Jeff han demos his breakthrough touchscreen. TED Talks. http://www.ted.com/talks/jeff_han_demos_his_breakthrough_touchscreen.html, February 2006.

[6] E. A. Isaacs, T. Morris, T. K. Rodriguez, and J. C. Tang. A comparison of face-to-face and distributed presentations. In *CHI '95: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 354–361, 1995.

[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***)*, pages 525–532, 1992.

[8] B. Johanson, A. Fox, and T. Winograd. The interactive workspaces project: Experiences with ubiquitous computing rooms. *IEEE Pervasive Computing*, 1(2):67–74, 2002.

[9] I.-J. Lin. Active shadows: Real-time video object segmentation in a camera-display space. In *Proceedings International Workshop on Image Analysis for Multimedia Interactive Services (***WIAMIS***)*, 2004.

[10] I.-J. Lin and H. Chao. Integrating contextual video annotation into media authoring for video podcasting and digital medical records. *HP Labs Technical Report*, HPL-2007-9, 2007.

[11] A. Mehrabian. *Silent Messages*. Wadsworth, Belmont, CA, 1971.

[12] A. Pennycook. Actions speak louder than words: Paralanguage, communication, and education. *TESOL Quarterly*, 19(2):259–282, June 1985.

[13] T. Porter and T. Duff. Compositing digital images. *ACM SIGGRAPH Computer Graphics*, 18(3):253–259, 1984.

[14] Y. Shi, W. Xie, G. Xu, R. Shi, E. Chen, Y. Mao, and F. Liu. The smart classroom: Merging technologies for seamless tele-education. *IEEE Pervasive Computing*, 2(2):47–55, 2003.

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

[16] 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 *Proceedings IEEE Workshop on Multimedia Signal Processing (***MMSP***)*, 2009.

[17] Z. Zhang. Computer vision technologies for remote collaboration using physical whiteboards, projectors and cameras. In *CVIIE '05: Proceedings of the Computer Vision for Interactive and Intelligent Environment*, pages 109–122. IEEE Computer Society, 2005.