Reach-Through-the-Screen: A New Metaphor for Remote Collaboration

Jonathan Foote, Qiong Liu, Don Kimber, Patrick Chiu, and Frank Zhao

FX Palo Alto Laboratory Inc., 3400 Hillview Ave., Palo Alto, CA 94304 USA

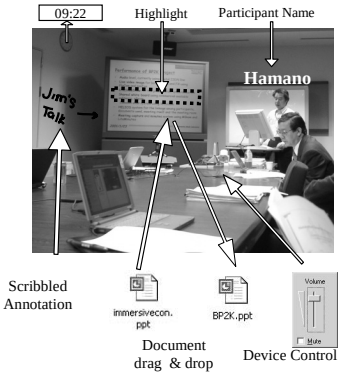
Abstract. For some years, our group at FX Palo Alto Laboratory has been developing technologies to support meeting recording, collaboration, and videoconferencing. This paper presents several systems that use video as an active interface, allowing remote devices and information to be accessed "through the screen." For example, SPEC enables collaborative and automatic camera control through an active video window. The NoteLook system allows a user to grab an image from a computer display, annotate it with digital ink, then drag it to that or a different display. The ePIC system facilitates natural control of multi-display and multi-device presentation spaces, while the iLight system allows remote users to "draw" with light on a local object. All our systems serve as platforms for researching more sophisticated algorithms to support additional functionality and ease of use.

1 Introduction

The Immersive Conferencing group at FX Palo Alto Laboratory investigates novel technologies for communication and collaboration. At FXPAL, our approach is to extend commodity hardware and software with novel approaches and interfaces that add functionality and simplify ease of use. Our work depends on a long tradition of research in this area by workers at Xerox PARC, MIT, Sony, UNC, Stanford, and elsewhere, and we apologize that space does not permit more than a cursory review of related work.

1.1 "Reach-Through-the-Screen" Interaction

Many of our systems are unified by a common interaction metaphor we call "RTS," for "reach-through-the-screen." In RTS, a video image of a particular location becomes an active user interface. We use the intuition that "things on the network are closer than they appear:" that is, devices visible in the video are available for control over the network, even if they are in a remote location. Devices and functions visible in the video image become available for the user to control via the video. For example, the video image of a display device becomes an active interface to that device: via RTS, a user may drag and drop a presentation file onto the video image, where it is then displayed. Similarly, dragging a file onto a printer image will case the printer to print that file. By



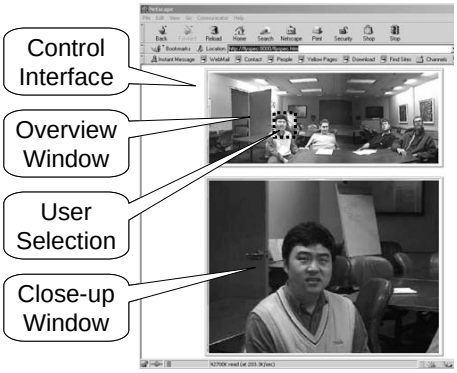


Fig. 1. (Left) "Reach through the screen" interface allows device access through an active video window.

Fig. 2. (Right) FlySPEC user interface. Circling a region in the panoramic video points the PTZ camera for a close-up view

interacting with an image, users are not encumbered with headgear or special eyeglasses. This is a non-immersive "window-on-the-world" (WoW) style of augmented reality [1]. Figure 1 shows some of the interaction possibilities available with RTS (though we have not fully implemented all of them). For example, right-clicking on an audio device could call up the control panel for that device. A hallmark of RTS interaction that in many cases it gives remote participants richer control than local users. In this paper, we review four related systems that use the RTS metaphor. The FlySPEC system allows a camera to be controlled by pointing or circling regions of interest in the video. The NoteLook system allows presentation graphics to be "dragged" from a video image, annotated, and then "dropped" onto a display device to display the annotations. The ePIC system allows simple control of multi-screen presentation displays and devices using the RTS metaphor, while the iLight camera-projector system allows real-world objects to be annotated from a video image. Drawing on the video image causes the drawn annotations to be projected onto the scene. With all these systems, we have implemented a basic robust version as a platform for both use and research. In many cases, we plan to augment the systems with automatic functionality from computer vision, sensors, and/or machine learning techniques.

2 The FlySPEC Camera System

A traditional pan/tilt/zoom (PTZ) camera system cannot allow multiple remote users to point the camera to different positions at the same time. To serve different viewing requests, a straightforward approach is to capture the event with a panoramic camera that covers every possible view and serve different

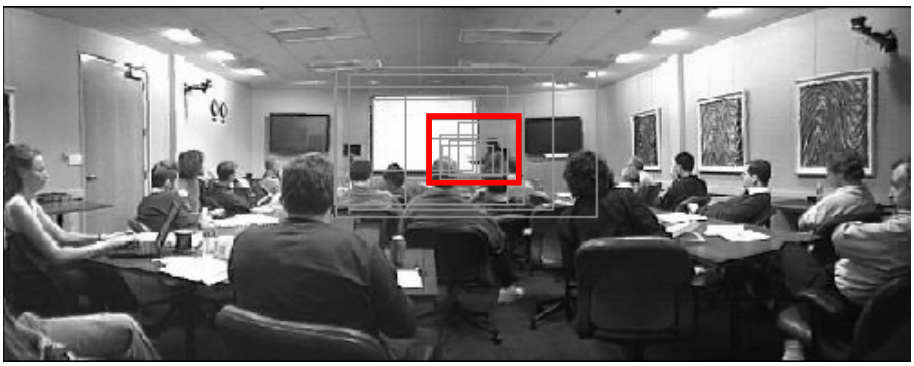


Fig. 3. Multiple user requests for SPEC camera zoom

viewing requests through electronic pan/tilt/zoom. However, a panoramic camera generally lacks the required resolution. FlySPEC is the name of our SPot Enhanced Camera system constructed by combining motorized PTZ cameras with a panoramic camera [3]. Using this combined system, users view a low-resolution panoramic view of a scene simultaneously with a customized close-up video. The panoramic video is the same for all users. The close-up video for a user is selected by marking a region in panoramic video with a gesture. Thus the panoramic video serves as an interface: clicking or circling objects in the panorama yields a zoomed-in video close-up.

2.1 Camera management and automatic control

A FlySPEC system has a limited number of physical PTZ cameras. A basic function of the FlySPEC system is to minimize conflicting PTZ requests and maximize the view quality for all remote users. This is done by serving view requests with a combination of cropped and zoomed images from both cameras. Each user therefore controls a "virtual camera" that behaves like a personal PTZ camera, (but may have reduced resolution). This is a problem for remote applications like classes, seminars, or sports games. To tackle the problem caused by multiple users, we optimize the view returned to each user by maximizing a cost function. Our FlySPEC system currently uses the overall electronic zoom factor as the cost function. Specifically, our management software tries to move PTZ cameras to positions that minimize the average electronic zoom for all users, thus resulting in the best images for all [2]. For example, Figure 2 shows multiple human zoom requests as rectangles; this is satisfied by zooming the camera to the bounding box of all requests. Our current research is to use machine learning to associate sensor data with human control actions to automatically develop good camera control strategies. For example, given motion and audio sensors, we could learn that motion near the podium results in camera movement from human controllers, while motion near the door does not.

Fig. 4. (Left) NoteLook pen annotation client

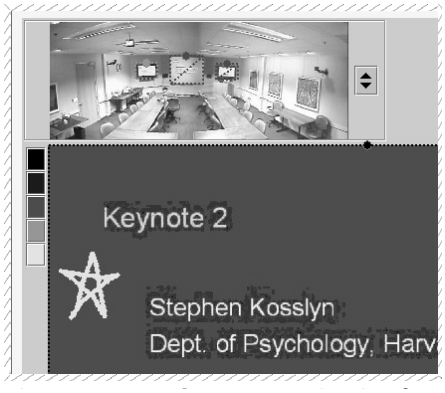


Fig. 5. (Right) NoteLook pen annotation interface showing active panoramic image with hotspots

2.2 Camera registration and calibration

Panoramic video is supplied by the FlyCam system developed at our lab, which employs multiple cameras and stitches the video images together [4]. A two-camera system provides a panoramic video with a horizontal field of view of approximately 110 degrees. (Recently, we have been constructing cheaper but low-resolution panoramic cameras by equipping inexpensive webcams with wide-angle lenses.) A chief asset of the panoramic camera is that it is fixed with respect to the scene. This allows sophisticated functions without having to compensate for camera motion. For example, registering devices with the areas they appear in the panoramic image allows them to be manipulated through that area without having to locate them with expensive and fragile computer vision techniques.

2.3 Devices and Hotspots

Availability and accessibility of devices are graphically depicted by the hotspots overlaid on the active video window. In practice, we used the fixed panoramic video image, and define hotspots to be particular regions of this image. The possible device operations are indicated by colors and styles of the highlight borders. We are currently working on automatically registering the PTZ camera image with the panoramic image, so that hotspots become active in the zoomed-in video as well.

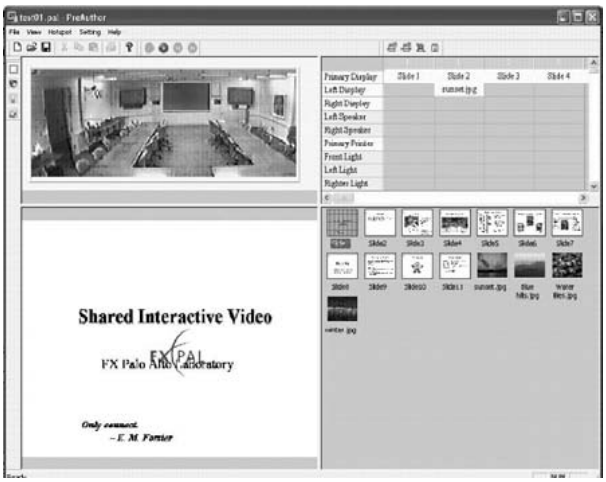


Fig. 6. ePIC presentation authoring and control system

3 The NoteLook Collaboration System

The NoteLook annotation system is designed to be run on a wireless pen tablet [5]. NoteLook allows collaboration and annotation around digital presentation graphics, using digital ink. In operation, the NoteLook system displays a thumbnail view of the conference room in the top left corner. Panoramic video shows a view of the environment with hotspots overlaid on the displays appearing in the video image. "Hotspots" that correspond to active devices are outlined with colored rectangle overlays on the live video image. Besides the two video channels from the front and back meeting room cameras, additional channels may be hooked up during a teleconference to interact with displays at a remote location. A user can grab an image of a slide showing on one of the wall displays by dragging from that display's hotspot into the note page area. The image appears in full resolution, and can be drawn upon with digital ink. After annotating the slide, the user can send the annotated image up to a selected wall display with a drag-and-drop operation. Users may also scribble notes on blank note pages with the pen. The image of any note page, whether it contains an annotated slide or only ink strokes, can be "beamed up" to any wall display. At the end of a session, the user can save the note pages for publishing on the Web. Note pages are time-stamped and automatically associated with any recorded meeting video in a Web interface, such that clicking on a particular annotation starts video playback at the time the annotation was made. More details and features of NoteLook are described in [6].

4 The ePIC Presentation Control System

Today, more and more meeting environments are equipped with a richer variety of displays, often including secondary or multiple projectors or large-screen displays. This gives a presenter richer options for conveying information to others. For example, a presenter can use the primary display for a text presentation, while using another display to show a supporting figure or video. Existing authoring tools, such as PowerPoint, are excellent for creating presentations for a single display, but provide no support for multiple devices, or remote presentations. To enable better use of media-rich environments, we designed ePIC, a tool for authoring and replaying presentations on arbitrary device configurations, that also supports real-time control [7] (as in FlySPEC and Notelook). ePIC complements, but does not replace tools used to author specific media. It can organize media prepared for simple devices and synchronously present them in one or more multimedia venues. For example, the EPIC system can import a conventional PowerPoint file and re-author it for effective presentation on multiple displays. Our prototype supports arbitrary configurations of displays, printers, speakers and room lights. EPIC, which stands for Environment Picking Image Canvas, uses live or static images of the presentation environment as a graphical user interface (GUI). This allows users to visually select and control presentation devices, even in remote locations. Users may drag slides or media files onto a visual representation of the intended device. For example, a user can drag a slide thumbnail on top of any of the displays visible in Figure 8 to show a slide on that display. The live panoramic video is especially useful for remote presentations, as it allows the presenter to see exactly what the audience is seeing at any time. Figure 8 shows the ePIC user interface. At top left is the active video canvas, showing room control hotspots (besides displays, ePIC can control other room features such as loudspeakers, printers, cameras, and even room lighting.) At top right is the slide timeline. Users can arrange slides and room control actions by time in the presentation and device. The bottom left panel shows a detailed slide view, or zoomed video image from the PTZ camera. At bottom right the slides in a PowerPoint presentation are available either for immediate drag-and-drop operations or for arranging into a multiscreen presentation. EPIC provides both a preview mode and a playback mode. In the preview mode, the actual room devices are not controlled. Rather, a preview of the presentation is rendered in the video image canvas. By displaying slide thumbnails in the hotspot bounding boxes (see Figure 4), the user can navigate the presentation using the usual controls, and see a visual indication of what slides are rendered on each display device. The image canvas can contain a rendering of a VRML model of the venue, if available. In this case, slide thumbnails are rendered in 3-D on the display surfaces. Because the rendering viewpoint can be selected, the user can preview the presentation from any point in the venue. Figure 9 shows an example of a VRML output, including a pop-up menu for selecting the viewpoint. Users may also zoom in and out of the rendering to see details or overviews of the presentation. Currently, we are investigating ways to

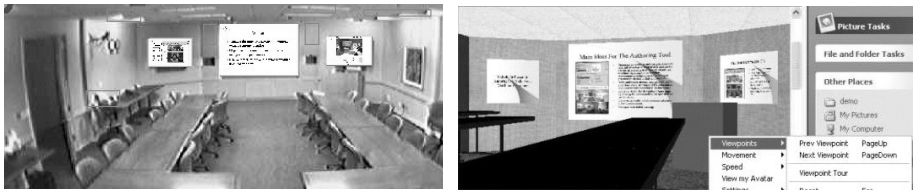


Fig. 7. Previewing ePIC presentations on video (left) and VRML (right) representations

incorporate machine learning to assist multi-screen presentations and optimal slide placement [9].

5 The iLight Projection Drawing System

The iLight system allows remote users to draw using light projected on a real-life object or scene [8]. Using an aligned camera/projector system, live video from the scene is presented as a drawable video image to one or more remote users. Graphics drawn on the video canvas are projected onto the scene visible in the video. Thus remote users may annotate particular scene objects or features for local viewers, and can also display text and project arbitrary digital images. Unlike related systems, there are no video feedback problems, nor are computer vision techniques required to detect objects or actions. Imagine a shared whiteboard as an example scenario. The iLight system's video projector "draws" remote images on a local whiteboard, while sending live whiteboard video to the remote drawer. In operation, remote users draw on the camera image with familiar graphical tools such as rectangles and (digital) pens, while local users draw directly on the whiteboard using (physical) dry-erase markers. The local projector, fixed with respect to the video camera, projects the remotely drawn images onto the whiteboard. Because the iLight camera and projector are exactly aligned, the projected image is drawn exactly where the remote user intends. Local users see the ink and projected light intermixed on the whiteboard, while remote users see a camera image of the same. Thus local and remote users can freely draw and annotate each other's work as if they shared a local whiteboard. Though neither can erase the other's marks, the remote user has functionality not available with physical ink. For example, the remote user may project any arbitrary image, as well as copy-and-paste regions of the camera image. We are currently investigating using camera feedback to enhance the projection onto surfaces with nonuniform reflectivity. By monitoring the reflection visible in the camera image, it is possible to modify the projected image to compensate for color and brightness variations in the scene. For example, when projecting on a checkerboard surface, the projection can be brightened in the dark squares and dimmed in the light squares to result in a more uniform perceived illumination.

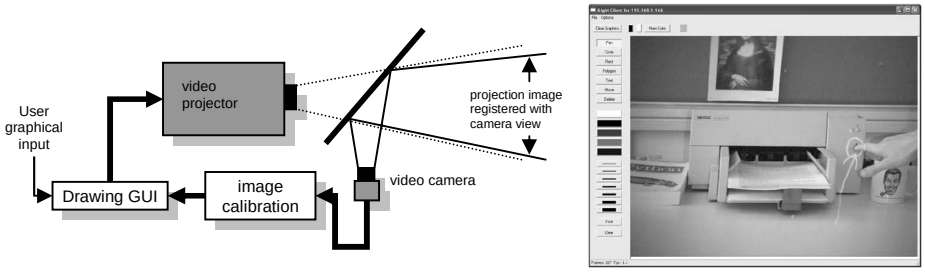


Fig. 8. (Left) iLight camera/projector system

Fig. 9. (Right) iLight user interface. User drawings on the video window are projected onto the actual scene.

6 Conclusion

We have presented four systems that support remote communication and collaboration, using a powerful and intuitive "reach-through-the-screen" interface. Perhaps most importantly, these systems serve as platforms for investigating algorithms, such as computer vision and machine learning, that will hopefully support additional advanced functions and ease of use.

References

1. M. Tani, K. Yamaashi, K. Tanikoshi, M. Futakawa, and S. Tanifuji, "Object-oriented video: Interaction with real-world objects through live video." In *Proc of CHI '92,* ACM Press, pp. 593–598
2. Q. Liu and D. Kimber, "Learning Automatic Video Capture from Human Camera Operations." In *Proc. IEEE Intl. Conf. on Image Processing*
3. Q. Liu, D. Kimber, J. Foote, L. Wilcox, and J. Boreczky. "FLYSPEC: A Multi-User Video Camera System with Hybrid Human and Automatic Control." In *Proc. ACM Multimedia 2002,* pp. 484–492, Juan-les-Pins, France, December 1-6, 2002.
4. J. Foote and D. Kimber, "FlyCam: Practical Panoramic Video." In *Proc. IEEE Intl. Conf. on Multimedia and Expo*, vol. III, pp. 1419–1422, 2000.
5. P. Chiu, Q. Liu, J. Boreczky, J. Foote, T. Fuse, D. Kimber, S. Lertsithichai, and C. Liao, "Manipulating and Annotating Slides in a Multi-Display Environment." In *Proc. of INTERACT '03*, pp. 583–590
6. P. Chiu, A. Kapuskar, S. Reitmeier, and L. Wilcox. "NoteLook: Taking Notes in Meetings with Digital Video and Ink." In *Proc. ACM Multimedia '99*, Orlando, Florida, November 1999.
7. C. Liao, Q. Liu, D. Kimber, P. Chiu, J. Foote, and L. Wilcox, "Shared Interactive Video for Teleconferencing. In *Proc. ACM Multimedia 2003*, pp. 546–554
8. J. Foote and D. Kimber, "Annotating Remote Reality with iLight," in preparation.
9. Q. Liu, F. Zhao, and D. Kimber, "Computer Assisted Presentation Authoring for Enhanced Multimedia Venues," in preparation.