Enhancing Distance Learning with Panoramic Video

Jonathan Foote and Don Kimber
FX Palo Alto Laboratory, Inc.
3400 Hillview Avenue
Palo Alto, CA 94304
{foote, kimber}@pal.xerox.com

Abstract

This paper describes a new system for panoramic two-way video communication. Digitally combining images from an array of inexpensive video cameras results in a wide-field panoramic camera, from inexpensive off-the-shelf hardware. This system can aid distance learning in several ways, both by presenting a better view of the instructor and teaching materials to the students, and by enabling better audience feedback to the instructor. Because the camera is fixed with respect to the background, simple motion analysis can be used to track objects and people of interest. Electronically selecting a region of this results in a rapidly steerable "virtual camera." We present system details and a prototype distance learning scenario using multiple panoramic cameras.

1.Introduction

Attempts to replicate the traditional lecture setting in distance learning have been of limited success. A big constraint is the absence of face-to-face communications. However as bandwidth and Moore's Law increase, the number of bits available for teleconferencing will increase as well. Here we present one approach to enhancing distance learning. Using a panoramic video system, we seek to replicate the affordances of "face-to-face" lectures at a distance. We assume that higher picture resolution and a wider field of view can help communication via video [1]. Though there remain significant problems such as gaze direction and audio, we are experimenting with using panoramic video to enhance the distance learning experience. We feel that panoramic video has two major value propositions for distance learning. The first is to replace a human camera operator. Instead of physically moving a steerable camera, the equivalent effect can be achieved by extracting an interesting portion of a larger image. This can automatically create a good video image the lecturer without the need for a human operator.

The second advantage of panoramic video addresses one of the more problematic areas of distance learning, which is audience feedback During a real-time distance learning, it is often difficult for the instructor to gauge the attention state of the audience. In a face-to-face lecture situation., there are a host of non-verbal cues that audience members can communicate their level of attention to the instructor. Every good teacher can differentiate the with the wandering stares and puzzled looks that indicate boredom or incomprehension from the forward posture and rapt gaze of engagement. This feedback is highly critical to the lecturing process, so the instructor knows when to slow down or go into more depth given the audience feedback; however, this is also difficult or impossible to convey to a remote location. Capturing lectures on video requires a human operator to orient, zoom, and focus the video or motion picture camera.

This paper presents FlyCam, a system that generates a seamless panoramic video images from multiple adjacent cameras [2]. The name alludes to the compound eyes of insects that form sophisticated images from an array of cheap sensors. FlyCam component cameras are mounted on a rigid substrate such that each camera's field of view overlaps that of its neighbor. The resulting images are aligned and corrected using digital warping, and combined to form a large composite image. The result is a seamless high-resolution video image that combines the views of all cameras. Because cameras are mounted in fixed positions relative to each other, the same composition function can be used for all frames. Thus the image composition parameters need only be calculated once, and the actual image composition can be done quickly and efficiently, even at video rates.

Because a FlyCam is fixed with respect to the background, straightforward motion analysis can detect the location of people in the image. This can be used to electronically "pan" and "zoom" a "virtual camera" by cropping and scaling the panoramic image. In this

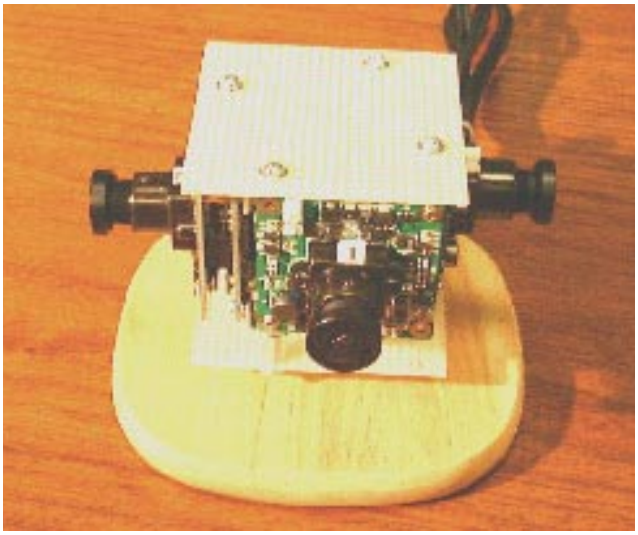

Figure 1. FlyCam videocamera array. Height =4 cm

system, an appropriate camera view can be automatically determined by finding motion of human images. Thus the system can serve as an automatic camera operator, by steering a real or virtual camera at the most likely subjects. For example, in a teleconference, the camera can be automatically steered to capture the person speaking. Also, it is possible for remote viewers to control their own virtual cameras; for example, someone interested in a particular feature or image on a projected slide could zoom in on that feature while others see the entire slide.

2.Technical Details

2.1. FlyCam configurations

The philosophy behind FlyCam was to achieve computationally reasonable panoramic imaging with a minimum of expensive or special-purpose equipment. To this end, a FlyCam is composed of inexpensive (< $150) miniature color video board cameras. Figure

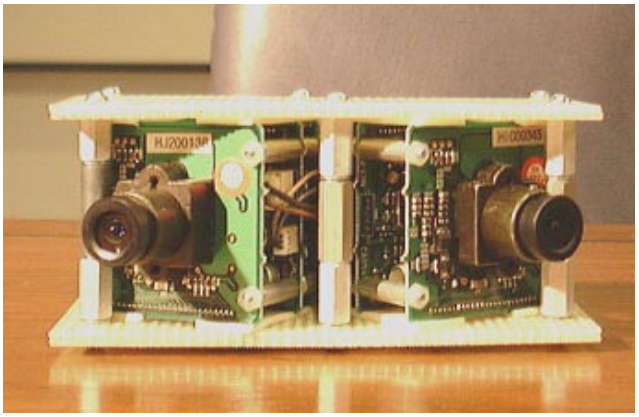

Figure 3. FlyCam videocamera array. Height =4 cm

1 shows a FlyCam prototype constructed from four video cameras. Though cameras are mounted as close together as practical, they do not share a common center of projection (COP). It is not necessary to align or optically calibrate the cameras in any way, as long as their fields of view overlap slightly. Each camera has a 2.5 mm lens offering a an approximately 115-degree field of view, thus the component camera images overlap somewhat. The small focal length yields a large depth-of-field and thus all objects are in focus from a distance of a few centimeters to infinity. Figure 2 shows a still frame from a 360-degree panoramic video.

We have found that a full 360-degree field of view is not ideal for typical distance learning applications. There is practically always a preferred direction where the audience faces the lecturer in the “front” of the room. To this end, we have developed the two-camera system shown in Figure 3. Two-camera systems can be configured with lenses in the rages of 3.5 to 8mm, yielding combined fields of view of.

A particular advantage of using fewer cameras is that images can be combined at full resolution at reason-

Figure 2. 360-degree panoramic video image

able frame rates. Because of the narrower field of view and higher resolution, two-camera systems are more appropriate for distance-learning applications. A less-than-360 field of view is not a disadvantage, as cameras are aligned with a typical viewpoint such as the lecturer as seen from the audience or vice-versa.

2.2. Piecewise image stitching

We use a piecewise bilinear warping of quadrilateral regions to both correct for lens distortion and to map images from adjacent cameras onto a common image plane so they can be merged. First, a number of image registration points are determined by imaging a structured scene and manually identifying the points in the different camera images that correspond to each registration point in the scene. In practice, we image a grid of squares, and use the corners as registration points. The four corners of each square form a quadrilateral "patch" in the image of each camera. Bilinear transformations are used to warp each quadrilateral patch. Each patch is mapped into a square "tile" in the panoramic image. Every image patch is then warped back to a square and tiled with its neighbors to form the panoramic image. Figure 5 shows the raw camera images with the patches while Figure 4 shows the composite panoramic image. Note how the vertical lines warped by lens distortion are straightened in the panorama.

2.3. Border patch cross-fading

The luminance across cameras will not be even, primarily because the component cameras have "auto-iris" functions that adapt their gain to match the available light. Component cameras imaging a scene with variable lighting will tend to have different gains, hence patches imaged by adjacent cameras will have

Figure 5. Raw camera images, showing "patches"

different luminances. Thus even when the panoramic image is geometrically correct, seams will be apparent from the brightness differences across cameras [2]. We minimize this problem by the simple measure of cross-fading edge patches. Redundant patches at are used at the edge of each camera -- that is, at camera borders, the same patch is imaged from each neighbor camera. Because these patches are then corrected to a square of known geometry, they can be combined by cross-fading them. The pixel value in a patch is given by a linear combination of the component patches, such that pixels on the left come from the left camera, pixels on the right come from the right camera, and pixels in the middle are a linear mixture of the two. This proves quite effective for hiding the camera seams, to the extent that they can be difficult to detect even when the observer knows where to look.

Figure 4. Composite panoramic video frame

Figure 6. FlyCam webcam application

2.4. Optical and stereo issues

In keeping with our "better, faster, cheaper" philosophy, no attempt is made to align component cameras to a common center of projection. In any case, it is not practical to achieve a common COP without elaborately aligned mirrors or other optical apparatus. Thus the panoramic image will have imperfections due to disparity between the cameras. We minimize this in several ways. First of all, there is no disparity for objects near the calibration distance. Because the baseline distance between component cameras is quite small, an object can move far from the optimal distance without noticeable disparity. Blending the border patches reduces the disparity artifacts even further. For typical distance learning scenarios, subjects rarely get close enough to the FlyCam that disparity is noticeable. More distant objects will also have a slight disparity, but typically these are smooth walls so such artifacts are not visible. Current work is to use the disparity "bug" as a feature: by calculating a measure of disparity it is possible to segment foreground (the lecturer or audience) from the background [6].

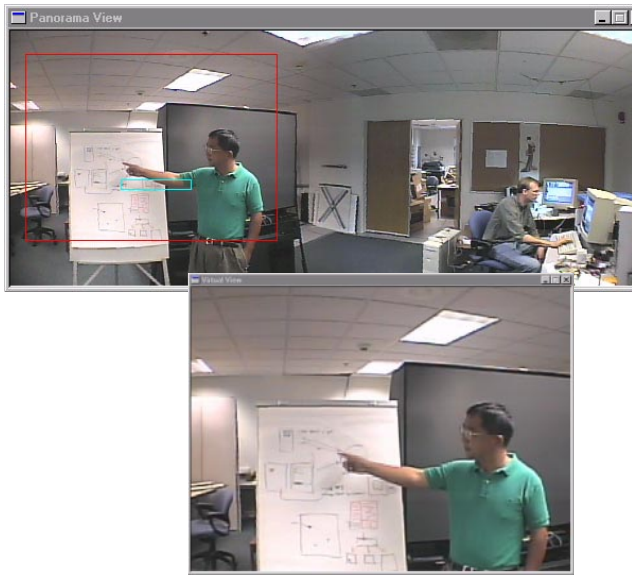

Figure 7. Automatic camera steering from motion.

3."Virtual" video cameras from panoramic images

An immediate application of panoramic images is to select a normal-sized "virtual" view by cropping the larger panorama. Virtual cameras can be panned/zoomed virtually instantaneously, with none of the limitations due to moving a physical camera. In addition, an unlimited number of different views are available at any one time, unlike a physical camera. We have built a FlyCam server application that functions as a virtual steerable video camera, allowing each client to request an individual view from the panoramic image. Figure 6 shows the client; the virtual camera is steered by clicking on the panoramic image or the left/right arrows, while the "+" and "-" zoom controls have the obvious functionality. Unlike other webcams that use a steerable camera, every client can choose their own unique combination of pan, tilt, and zoom.

3.1. Virtual camera control using motion analysis

Though a user can easily select a desired image, we can eliminate human input entirely for a truly automatic lecture recording system. To this end, we have implemented automatic camera control algorithms to select an appropriate virtual camera view. Because the FlyCam is fixed with respect to the background, motion analysis does an excellent job of detecting interesting foreground objects, such as the lecturer.

Figure 9. FlyCam view of remote audience.

Motion is determined by computing the absolute value of the frame-to-frame pixel differences. This is thresholded to remove small motions and quantization noise. The total number of non-zero pixels in the thresholded image is a good measure of the motion in the FlyCam's field of view. The first and second spatial moments of the thresholded difference image are calculated. The first spatial moment is a good estimate of the centroid of a moving object, while the second spatial moment estimates the spatial distribution of motion in the panoramic image. The resulting location estimate is smoothed over time using a Kalman filter. Straightforward heuristics control the virtual camera based on the motion analysis. The virtual camera is set to follow the center of any moving object. The new image coordinates x at frame t are set to a function of the coordinates form the previous frame and the new estimate of the motion centroid $\hat{x}$ as follows:

$$x_t = \alpha x_{t-1} + (1-\alpha)\hat{x}$$

Figure 8. Distance learning scenario using multiple video panoramas

The parameter $\alpha, 0 \le \alpha < 1$ serves as "inertia;" if it is large the virtual camera will move only slowly towards the motion. A moderate value of α serves both to mimic the dynamics of a physical camera and to smooth jitters due to noise inherent in the motion analysis.

For a distance learning environment, several ad-hoc methods enhance the lecturer tracking. If there is a projection screen or video monitor, it is desirable to make sure that moving images on the display are not tracked. To this end, our tracking algorithms can be set to include a "dead zone," or regions where motion is ignored. Additionally, our lecture room has a tall podium that obscures much of the lecturer. To ensure proper tracking, this region is replaced with a copy of the motion above it. Thus small movements above the podium are magnified so the tracking will continue despite occlusion of the speaker's torso. Again, these enhancements are straightforward because the FlyCam does not move with respect to the scene, which is not the case with most conventional video cameras. In the aggregate, these tracking functions add to the realism and interest of the video, and mimic the performance of a human operator. Figure 7 shows our automatic camera system in action.

4.FlyCam for Distance Learning

Figure 8 shows a prototype distance learning scenario in use at FX Palo Alto Laboratory. The instructor stands at the front of the room, and the local audience in the middle facing forward. In the front left of the room is a flat panel display that shows a view of the

System	Resolution	Bandwidth	Stitching/Warping Artifacts	motion images
Film-based panoramas (IPIX, QTVR)	Excellent	Low	Some	No
Wide-angle systems (Columbia, BeHere)	Poor	Moderate	Few	Yes
Polycameras (FlyCam, Columbia, USC)	Good	Moderate to high	Some	Yes

Table 1. Taxonomy of panoramic imaging systems

local audience. This is also the view that a remote instructor would have of the local audience. The front right display shows the remote audience to both the instructor and the local audience. Presentation graphics are displayed on the large central display; the remote audience sees these as well via a remote monitor connection. In addition, the remote audience sees a zoomed view of the instructor via an automatic tracker that follows the instructor as she or he walks about the front of the room.

5.Related work

There has been considerable prior work on combining multiple images into a panoramic scenes; enough that limited space precludes a fuller set of references. Many approaches have been to compose existing still images into a panorama that can be dynamically viewed [7,8], or by compositing successive video frames into a still panorama [9]. Because all these techniques involve computationally expensive image registration (that is, aligning images with unknown displacements) none of these techniques can be done practically at video rates. In contrast, the system presented here uses cameras with fixed, known alignments, so displacements need not be calculated at all.

A group at Columbia has created an omnidirectional digital camera using curved mirrors [10]. In this system, a conventional camera captures the image reflected from a parabolic mirror, resulting in a hemispherical field of view. Digitally processing the reflected image allows the construction of distortion-free images for any user selected portion of the acquired omnidirectional image, albeit at limited resolution. The drawback of this approach is that subimages extracted from the hemispherical image will be limited in resolution to a small fraction of the single camera, and the necessary image warping will be extreme to regenerate unwarped images. In contrast, the system presented here has virtually unlimited resolution at all viewing angles. If more resolution is desired, the system can be configured with additional cameras. A group at UNC uses 12 video cameras arranged in two hexagons, along with a mirror apparatus to form a common COP. The UNC group devised a similar approach to panoramic image composition using the texture mapping hardware. of a SGI O2 [4]. Another group at Columbia has taken a similar approach using an array of board cameras. Instead of piecewise image warping, a table lookup system directly warps each image pixel into the composite panorama [5]. A group at USC has created a system using 5 cameras arranged in a “+” configuration for automatic panning and tilting [11].

There is no shortage of prior research in person tracking systems. Systems based on steerable cameras must compensate for camera motion as well as the event when a face goes out of view of the camera. This is much less of a problem for a panoramic camera with a motionless and much wider field of view. At least one system uses a panoramic image from a hemispherical mirror to point a steerable camera [12].

5.1. Future work

Besides investigating other camera configurations and resolutions, we are also investigating the use stereo disparity to improve person tracking. We are also investigating audio source location from a microphone array to augment the automatic tracking [13]. By automatically pointing a virtual camera towards an audio source, the FlyCam system could better capture audience questions.

6.References

[1] Moore, G., “Sharing Faces, Places, and Spaces: the Ontario Telepresence Project Field Studies,” Chap 14 in *Video-Mediated Communication*, ed. Finn, K., Sellen, A., and Wilbur, S. Lawrence Erlbaum Asscoiates, Mahwah, New Jersey, 1997.

[2] Foote, J., and Kimber, D., “FlyCam: Practical Panoramic Video,” in *Proceedings of IEEE International Conference on Multimedia and Expo*, vol. III, pp. 1419-1422, 2000.

[3] G. Wolberg, *Digital Image Warping*, IEEE Computer Society, Press, 1992

[4] A. Majumder, et al., “Immersive teleconferencing: a new algorithm to generate seamless panoramic video imagery,” in *Proc. ACM Multimedia 99*, Orlando, FL, pp. 169-178, 1999.

[5] R. Swaminathan and S. Nayar, “Non-metric calibration of wide-angle lenses and polycameras,” in *Proc. Computer Vision and Pattern Recognition*, June 1999

[6] Darrell, T., Gordon, G., Woodfill, W., Baker, H., “A magic morphin mirror,” in *SIGGRAPH '97 Visual Proceedings*, ACM Press. 1997.

[7] S. Chen and L. Williams, “View interpolation for image synthesis,” in *Computer Graphics (SIGGRAPH'93)*, pp.279-288, August 1993.

[8] IPIX, the Interactive Pictures Corporation, http://www.ipix.com

[9] Teodosio, L., and Bender, W., “Salient Video Stills: content and context preserved,” in *Proc. ACM Multimedia 93*, Anaheim, CA, pp.39-46, 1993.

[10] Nayar, S., “Catadioptric omnidirectional camera.” In Proc. of *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Peurto Rico, June 1997

[11] Nicolescu, M., and Medioni, G., “Electronic Pan-Tilt-Zoom: A Solution for Intelligent Room Systems,” in *Proceedings of IEEE International Conference on Multimedia and Expo*, vol. III, pp. 1581-1584, 2000

[12] Huang, Q., Cui., Y., and Samarasekera, S., “Content based active video data acquisition via automated cameramen,” in *Proc. IEEE International Conference on Image Processing (ICIP) '98*

[13] Wang, C., and Brandstein, M., “A hybrid real-time face tracking system,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) '98*, IEEE