Capturing and Presenting Shared Multi-Resolution Video

Don Kimber, Qiong Liu, Jonathan Foote, Lynn Wilcox

FX Palo Alto Laboratory, 3400 Hillview Ave. Bldg. 4, Palo Alto, CA 94304

ABSTRACT

We present a framework, motivated by rate-distortion theory and the human visual system, for optimally representing the real world given limited video resolution. To provide users with high fidelity views, we built a hybrid video camera system that combines a fixed wide-field panoramic camera with a controllable pan/tilt/zoom (PTZ) camera. In our framework, a video frame is viewed as a limited-frequency representation of some "true" image function. Our system combines outputs from both cameras to construct the highest fidelity views possible, and controls the PTZ camera to maximize information gain available from higher spatial frequencies. In operation, each remote viewer is presented with a small panoramic view of the entire scene, and a larger close-up view of a selected region. Users may select a region by marking the panoramic view. The system operates the PTZ camera to best satisfy requests from multiple users. When no regions are selected, the system automatically operates the PTZ camera to minimize predicted video distortion. High-resolution images are cached and sent if a previously recorded region has not changed and the PTZ camera is pointed elsewhere. We present experiments demonstrating that the panoramic image can effectively predict where to gain the most information, and also that the system provides better images to multiple users than conventional camera systems.

Keywords: Collaborative and automatic camera control, panoramic video, video capture and presentation, webcams.

1. INTRODUCTION

It is becoming increasingly common to provide remotely viewable video of meetings, presentations, performances, etc. Controlling the video cameras can be problematic, especially if multiple remote viewers have different viewing needs. Researchers have proposed fully automatic camera control systems, including Bell Core's Auto-Auditorium [2], Cornell's lecture capturing system [7], Microsoft's ICAM system [6], and AT&T's Automated Cameramen [5]. However, these systems typically rely on state-of-the-art audio and computer vision techniques that may not be robust enough for real world use. Furthermore, different audience members may have different interests and preferences regarding what they would like to see. For example, one viewer of a presentation may wish to see the speaker, another may want to read written notes on a white board, and yet another may prefer to view a display being used as part of the presentation. Different remote audience members may be presented with individualized views, but there typically are many more viewers than available cameras. One approach is to use a wide field panoramic camera which captures the entire view, and provide individual views using digital panning and zooming [4]. Unfortunately, the resolution of panoramic cameras is limited, and so it may not be possible to provide high enough quality close up views to satisfy users' needs (e.g. to read presentation text.)

We have built FlySPEC, a hybrid camera system shown in Figure 1. (The name is derived from the FlyCam panoramic camera used for the wide-angle view [4] and the acronym SPEC for SPot Enhanced Camera.) FlySPEC combines a wide angle panoramic camera with a pan/tilt/zoom (PTZ) camera which can provide higher resolution views of interesting regions. A web-based interface provides each remote viewer with a small panoramic video image together with a larger close-up, as shown in Figure 3. Users may request close-up views of any region in the panorama simply by marking that region with a box outlining the desired field of view. Although the framework described in this paper can be easily extended to include multiple PTZ cameras, we will confine discussion to the case of a single PTZ camera.

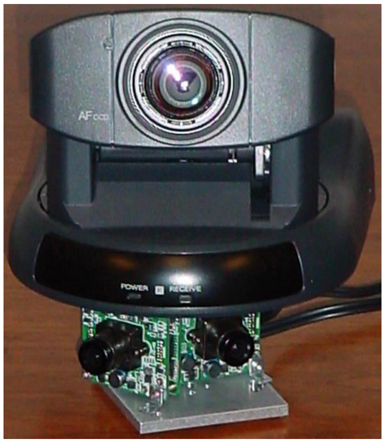


Fig. 1: The FlySPEC Camera: A pan-tilt-zoom camera (top) is combined with a fixed panoramic camera system (below.)

The basic problem discussed here is where to point the PTZ camera, and which close-up view to provide each remote user. This can be difficult because at a given time it may not be possible to simultaneously provide high-resolution views for all requested regions. Our approach to this problem is motivated by the human perception and rate-distortion theory. The human visual system combines a wide area low-resolution retinal visual field with a much smaller but quickly controllable area of foveal vision. Essentially, the PTZ camera in FlySPEC plays a role analogous to the fovea. A remarkable thing about the human visual system is that the low and high resolution areas of our sensory visual field are so well integrated that we have a conscious illusion of a visual field which is much higher resolution over its entirety than is actually the case [13]. Rate-distortion theory offers methods to maximize the fidelity of signals transmitted through band-limited channels. Our approach to FlySPEC camera control combines these ideas to integrate low and high-resolution video to provide the best fidelity images.

Conceptually, the 'true' world view from a camera can be thought of as an ideal image function *w*. This is sampled by cameras which can capture the image signal only up to some maximum spatial frequency determined by camera resolution and level of optical zoom. Furthermore, the finite display resolution available to remote viewers imposes another limited information channel. The goal is to maximize the overall quality of each user's view, determined by image fidelity (e.g. in mean squared error) together with the users level of interest in those views. Conceptually, the system maintains a band limited representation w_S of the ideal image *w*, where the cutoff frequency varies spatially depending on where the PTZ camera is pointed. Therefore, a key task is to determine how to control the PTZ camera to maintain the best quality user views. Obviously, user requests to view specific regions guide the system by indicating regions where improved resolution is most desirable.

Even in the absence of any user requests, the system can automatically guide the PTZ camera based on image statistics. (See [14] for an excellent review article on the statistics of natural images.) For example, at a given time the current estimate w_S and a spectral model of image statistics can be used to predict where to point the PTZ camera to gain the most additional high frequency information. We will describe this approach in more detail using a squared error fidelity criterion together with a $1/f^2$ spectral power model. This framework can also include more perceptually or semantically meaningful measures of fidelity. For example, feature-based models could be used to predict regions containing text and the resolution necessary for legibility. The system can also use a temporal model to estimate the error in regions that do not seem to be changing. The simplest model is that the absence of change in a region of the low-resolution panoramic view means that a recent high resolution PTZ view is likely to still be a good view. This is very valuable when the multiple user requests can't be satisfied by a single PTZ camera.

The remainder of this paper is structured as follows. In Section 2, we give an overview of the FlySPEC system including architecture and description of the user interface. In Section 3, we describe the optimization framework used to control FlySPEC in more detail. In Section 4 we describe experiments for evaluating the system. We conclude with discussion and future directions in Section 5.

2. OVERVIEW OF THE FLYSPEC SYSTEM

The FlySPEC system includes camera hardware, video servers, and web-based clients as shown in Figure 2. The hardware consists of a hybrid camera system including a FlyCam panoramic camera [4], a Sony EVI-D30 PTZ camera, and the servers necessary to access and control the cameras. The FlyCam stitches video from two wide-angle cameras to produce 1200x480 resolution image frames with an approximately 100-degree field of view. The Sony camera and frame grabber hardware can provide 768x492 resolution images and can be pointed to anywhere in the field of view of the panoramic camera, and can be zoomed in to areas as small as about 4.4 degrees field-of-view. The close proximity of the FlyCam and PTZ camera makes it easy to control the PTZ camera to cover a given region, using a simple affine transformation to map a region in the panoramic image to pan/tilt/zoom parameters for the PTZ camera. The zoom value is determined from the width of the desired region.

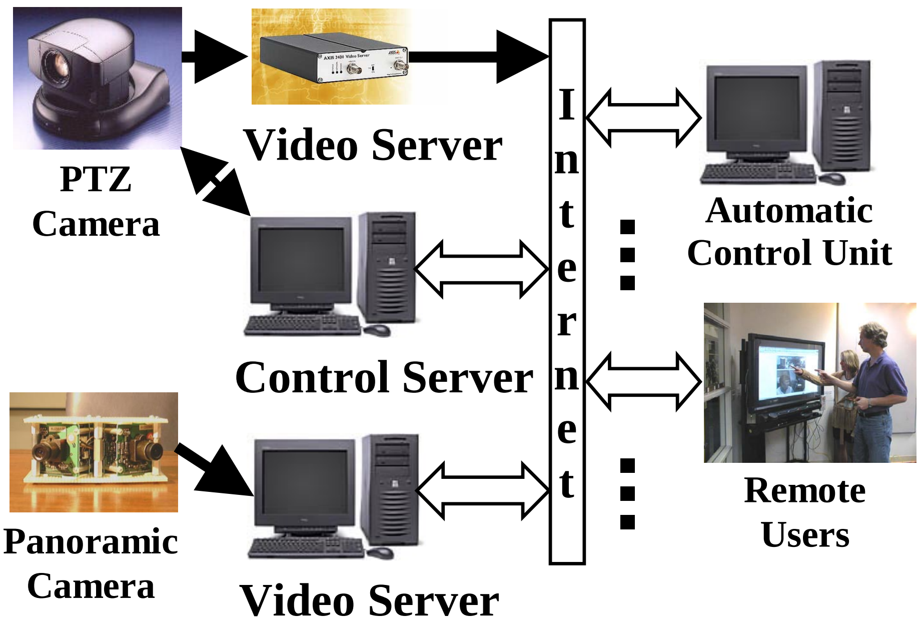
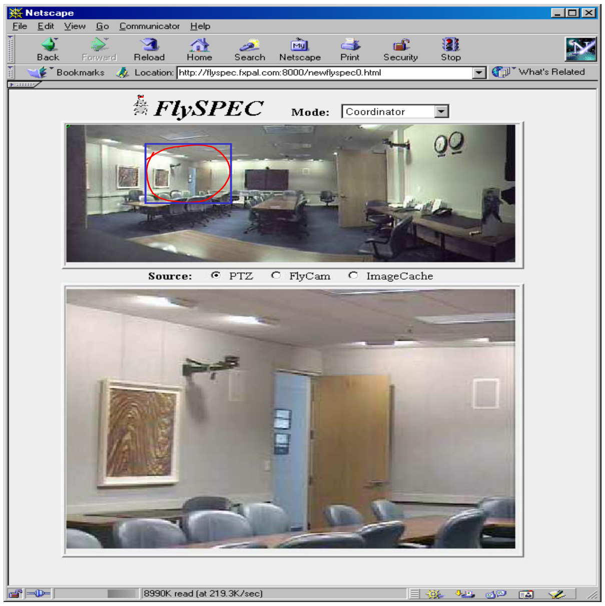


Fig. 2: The FlySPEC system architecture.

The user interface runs in a web page as shown in Figure 3. At the top is a panoramic video image, above a close-up view. If the user watches passively, the system automatically determines the close-up view. However if the user wants a more detailed view of a particular region, he/she can circle it with the mouse. The bounding box of the indicated region is then enlarged and presented in the close-up view, adjusted to match the available aspect ratio. At any given time, the close-up window may display either the current PTZ video, the digitally zoomed view from the panoramic camera, or a cached still image of the selected region. In the absence of user input, the system chooses the appropriate source using the methods described in the next section. However, the user is free to override the system's choice at any time and select a different view from the panorama.

Because normally many remote users are watching an event, the demands on any one user for controlling the camera are light. Typically certain shots, such as a close-up of the speaker giving a presentation, will be of interest to a number of users. Selections of interesting regions made by some viewers can benefit other users who have made no specific requests. However, this assumes no 'civic responsibility' on the part of any viewers – viewers simply mark the regions that they want to see. In the case of multiple simultaneous requests, the framework described in the next section is used to arbitrate. Similar (i.e. overlapping) requests can usually be well handled by choosing a PTZ view that covers the requested regions. Requests for close-ups of non-neighboring regions can be handled sequentially.


Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Stop
Bookmarks
What's Related
FlySPEC

Location: http://flyspec.fxpal.com:8000/newflyspec0.html

Mode: Coordinator

Source: ● PTZ ○ FlyCam ○ ImageCache

8990K read (at 219.3K/sec)

Fig. 3: The FlySPEC Web interface for remote viewing and control. The user has selected a region of interest in the panoramic view at top, which is displayed as a close-up below.

3. A FRAMEWORK FOR CAPTURING AND PRESENTING VIDEO

Here we describe the framework for controlling the PTZ camera and determining which views to present to users. We first describe how the system maintains its own representation of the world in the absence of users. This is useful for applications such as constructing variable resolution mosaics of a static scene. Because the primary consideration is to provide optimal views for multiple users, we will next show how the framework solves that problem.

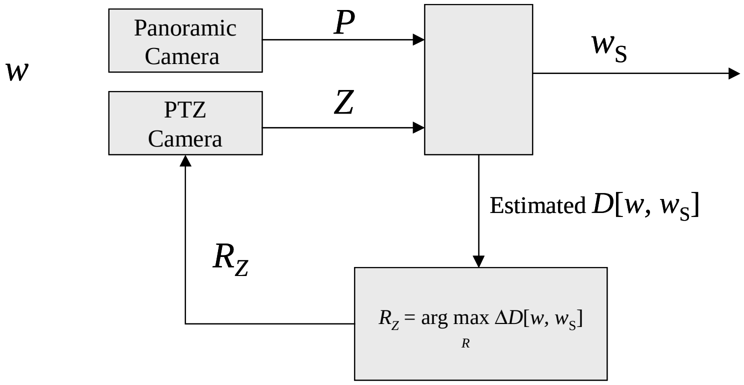


Fig. 4: The FlySPEC control system structure. Here w is the 'true' image function, w_S is the system representation of w, D is distortion, and R_Z is the region where the PTZ camera is pointed.

In our setup, we assume a "real world" view available from the FlySPEC, modeled as an image function $w(x,y,t)$ which gives the intensity along any direction x,y at any time t. Different coordinate systems could be used, but for convenience we will use a cylindrical coordinate system matching the image cylinder of the panoramic camera. The output from the panoramic camera P can be understood as a spatial frequency band limited sampled representation of $w(x,y,t)$. Likewise, when the PTZ camera *Z* is pointed at a rectangle R_Z in the *x,y* plane, its output can be viewed as a sampling of *w(x,y,t)* in the region R_Z, but with higher resolution corresponding to a higher upper band limit. The output from *P* and *Z* can conceptually be combined into a representation $w_S(x,y,t)$ of *w(x,y,t)*. Note that $w_S(x,y,t)$ is variable resolution, and will contain higher frequency information in regions where a zoomed view is available.

We consider a measure $D[w_S, w]$ which describes the difference or “distortion” between w_S and *w*. Conceptually, this is the information lost by representing *w* with w_S. Many loss or distortion functions may be used, but we will focus on the weighted squared error

$$D[w_s, w] = \int c(x,y,t)[w_s(x,y,t) - w(x,y,t)]^2 dxdydt = \int c(x,y,t)e(x,y,t)dxdydt ,$$

where *c(x,y,t)* is a weighting function which can emphasize that certain regions are considered more important or interesting than others.

A complete analysis of the optimal continuous-time control of the PTZ camera would be highly complex, computationally intractable, and is beyond the scope of this paper. We simplify the analysis by assuming camera motions are made only at discrete times, which may occur every few seconds or when user requests occur, and also by making a greedy decision at each moment of where to point the PTZ camera.

Consider a region *R* where $w_S(x,y,t)$ is band-limited to spatial frequencies *k* less than some cutoff *a*, and assume that *c(x,y,t)* is constant *c* in that region. The contribution to $D[w_S, w]$ from that region at time *t*, denoted *D(R,a,t)* is

$$D(R,a,t) = c\int_R [w_s(x,y,t) - w(x,y,t)]^2 dxdy .$$

In the spatial frequency domain, this can be computed as:

$$D(R,a,t) = c\int [W_s(k_x,k_y,t) - W(k_x,k_y,t)]^2 dk_x dk_y .$$

Because the estimate $w_S(x,y,t)$ is assumed to be a band limited representation of *w(x,y,t)* sampled with spatial frequency cutoff $k_x,k_y < a$, then we find:

$$D(R,a,t) = c \int_{k_x,k_y>a} W(k_x,k_y,t)^2 dk_x dk_y dt .$$

Now this quantity is not in fact known, because the system only has the estimate $w_S(x,y,t)$ of the true (and unknown) image function *w(x,y,t)*. However what is relevant to choosing the PTZ camera direction is the expected error for region *R*, denoted *E(R,a,t)* :

$$E(R,a,t) = \mathrm{E}\left[\int_{k_x,k_y>a} W(k_x,k_y,t)^2 dk_x dk_y dt \right] .$$

This can be used to select where to point and zoom the camera. Suppose that for a rectangle *R*, the current spatial frequency cutoff is *a*, and that zooming into that region would provide a new higher frequency cutoff *b*. (The smaller the rectangle *R*, the greater *b* is.) Then the expected distortion reduction for rectangle *R* is:

$$\Delta D(R,a,t) = c(E(R,a,t) - E(R,b,t)) .$$

Therefore, a simple strategy for comparing candidate regions for where to point the PTZ camera is to pick the region *R* which maximizes ΔD.

For practical implementation, it is convenient to tile the panoramic image into small fixed-size regions, indexed by row and column *i,j*. During operation, the system maintains matrices of key quantities: a_{ij} is the current spatial frequency cutoff of the estimate of in region *i,j* , E_{ij} is an estimate of the expected error for the region, and c_{ij} is the weight for the region. Then the contribution to overall distortion for a rectangle *R* is given by

$$\Delta D(R,a,t) = \sum_{i,j} c_{ij} \Delta E_{ij} \ ,$$

where the sum is taken over tiles in the region *R*. The error estimate E_{ij} depends on the cutoff frequency a_{ij} for the region, as well as the current estimate $w_S(x,y,t)$ around the region. In particular, spectral characteristics of $w_S(x,y,t)$ can be used to estimate the error at higher unobserved frequencies. The spectral energy distribution for many images is well modeled statistically as falling with frequency as $1/f^{\,2}$. So, if the spectral magnitude at $f=1$ is M_{ij} for region *i,j*, the energy in the frequency range $[a_{ij}, b_{ij}]$ (and thus the projected error reduction by obtaining frequencies up to b_{ij}), is given by

$$\Delta E_{ij} = \int_{a_{ij}}^{b_{ij}} \frac{M_{ij}}{f^2} df = M_{ij} \left(\frac{1}{a_{ij}} - \frac{1}{b_{ij}}\right) .$$

The value of b_{ij} is determined by the spatial sampling frequency of the PTZ image *R*, given by $b_{ij} = N / width(R)$ where *N* is the number of image pixels per row.

3.1. Increasing video fidelity for multiple users

The primary task of FlySPEC is to provide high quality views for multiple remote users. The framework described above is easily extended to that situation, as shown in Figure 5. The FlySPEC user interface provides each user with a low-resolution panoramic view P_u, and a close-up view C_u of a region. Due to network bandwidth and display size limitations, P_u and C_u are of limited resolution. For example, in our usual FlySPEC configuration the panoramic display P_u is 500x200 pixels, whereas the panoramic camera *P* has about 1200x480 resolution. Likewise, the close-up image is 320x240 even though the PTZ camera can provide 768x492 pixel images.

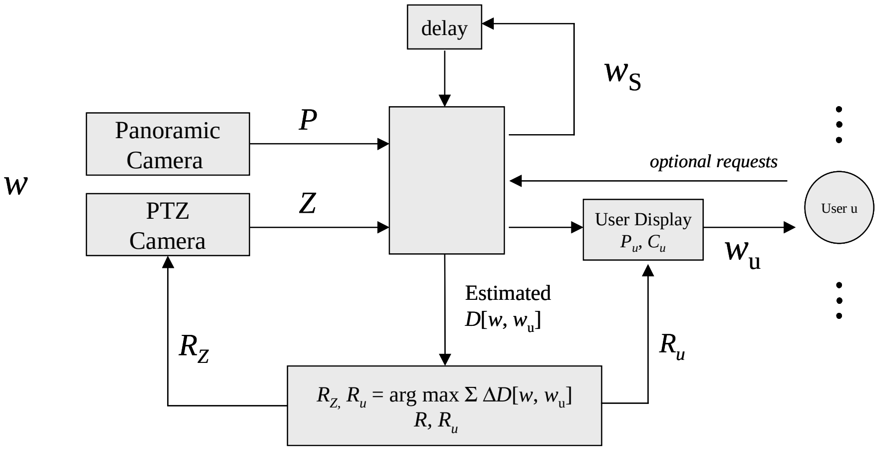


Fig. 5: The FlySPEC control system structure. Here *w* is the ‘true’ image function, w_S is the system representation, *D* is distortion, R_Z is where the PTZ camera is pointed, R_u is the close-up region shown to user *u*, and w_u is the representation of *w* based on the user display.

Conceptually, the user has an estimate $w_u(x,y,t)$ of the true image, and an individual weighting function $c_u(x,y,t)$. We seek to maximize the fidelity of this estimate

$$D[w_u, w] = \int c_u(x,y,t)[w_u(x,y,t) - w(x,y,t)]^2 dxdydt = \int c_u(x,y,t) e_u(x,y,t) dxdydt \, .$$

Given multiple users, we seek to optimize the fidelity of all views provided to all users. So the goal is to minimize

$$D_{TOTAL} = \sum_{u} D[w_u, w],$$

where it is understood that w_u depends on R_Z and R_u. The FlySPEC system must choose which regions R_u to present close-up views C_u for each user u, and what region R_Z to point the PTZ camera towards. Thus, the goal is to minimize D_{TOTAL} over the choices for R_Z and R_u for each u.

It is important to realize that different users have different needs, as reflected by their individual weighting functions $c_u(x,y,t)$. The weighting function could be modeled in various ways, for example

$$c_u(x, y, t) = c_0(x, y, t) + s_u(x, y, t),$$

where $c_0(x,y,t)$ is a "prior" for the interest function, and $s_u(x,y,t)$ is an indicator function equal to 1 if x,y is in the user-selected view region and zero otherwise (or if the user has made no requests). In the simplest case $c_0(x,y,t)$ is uniform, but it could also be learned from previous user actions or adjusted for *a priori* knowledge. For example, a region around a lectern or display could be set to higher values. Also, selections from many users could be averaged over time to estimate preferred regions in the scene.

3.2. Caching images to anticipate user requests

An interesting problem is how to proceed when no user requests are available. In this situation, the system cannot directly increase the fidelity of any user's view. However, this period can be used to cache image data to improve the response to anticipated user requests. Consider a static environment with image function $w(x,y)$. There still exists a time-varying estimate $w_S(x,y,t)$ of $w(x,y)$ which can be thought of as a variable-resolution mosaic. As the PTZ camera is focused on a smaller region R, the spatial frequency cutoff a increases, and so the error $E(R,a,t)$ decreases. Since the users' viewing events can happen at any time, decreasing the error between $w_S(x,y,t)$ and $w(x,y)$ can significantly increase the fidelity of future images selected by the users. This can be done by constructing a mosaic of the static scene by caching multiple high-resolution views. Since this must be done as rapidly as possible, the order in which regions are selected for higher resolution views can be chosen to reduce $D[w_S, w, t]$ the fastest. When the mosaic is produced by tiling the overall view into a set of slightly overlapping close-up regions imaged by the PTZ camera, it is interesting to consider the order in which those regions should be captured. As the number of captured close-ups increases, $D[w_S, w, t]$ will monotonically decrease. However the amount of decrease after any given number of images is captured depends on order in which images are captured. Since the improvement from each close-up is independent from any other, the greedy algorithm is actually optimal.

Even in dynamic scenes that change over time, there are often large areas that are static, at least for extended periods of time. This can be exploited to provide high fidelity views of a region even when the PTZ camera is pointed elsewhere. Each time the PTZ camera is moved away from a region, FlySPEC preserves the last high-resolution image of that region. Later when a view of that region (or a sub-region within it) is requested, the recorded image can be used. However, representing a region with an earlier image will result in distortion, which can be characterized to match our framework. A simple model is to assume that if the low-resolution view of that region has not changed, it is a static region, and that a high-resolution view would not have changed either. For example in a presentation, a speaker may use a large display for presentation graphics and a whiteboard for occasional notes. FlySPEC can capture close-up views of the notes or slides, then point the PTZ camera at the speaker. The recorded images of the notes or slides are available as close-up views without requiring the PTZ camera. However when a slide is changed, this is apparent from the panoramic camera, and FlySPEC will be able to determine that the previous recorded image of the slide is no longer a valid (i.e. high fidelity) representation of the slide area.

Image caching is thus a powerful way to better leverage the PTZ camera. However, it can be misleading in situations where important information is available only in highly zoomed-in views. Figure 6 shows a mostly static scene that has changed as a person entered a room. The system correctly determined where to point the PTZ camera to update w_S for the portion of the scene where the person was now standing. However, it failed to identify the changes in clocks visible in the upper right portion of the scene. If a close-up view of the clocks were recorded, it would be highly misleading to present that view later. To avoid user confusion, the web interface (shown in Figure 3) clearly indicates using radio buttons when cached images are being shown, and allows the user to select live video.

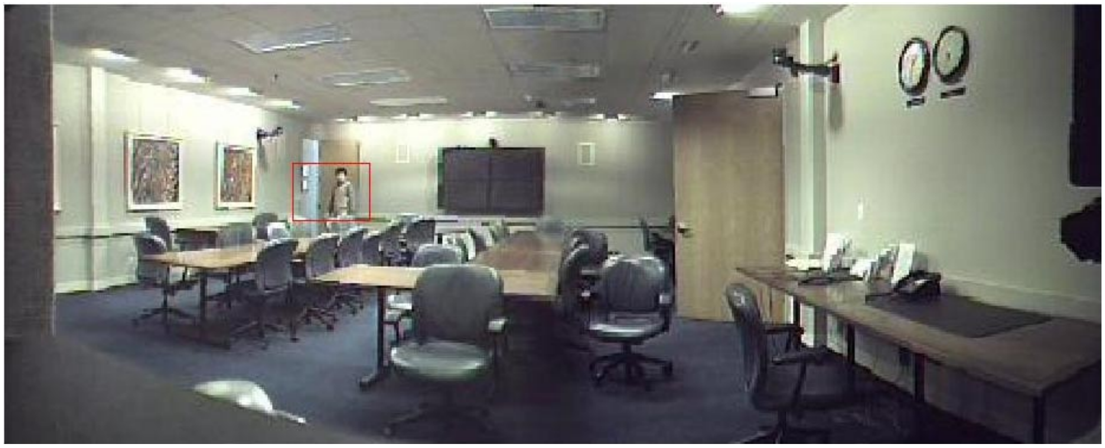


Fig. 6: Using image change to select a close-up view. The red region was automatically selected when the subject entered the room. Change of the clocks, however, was not detected.

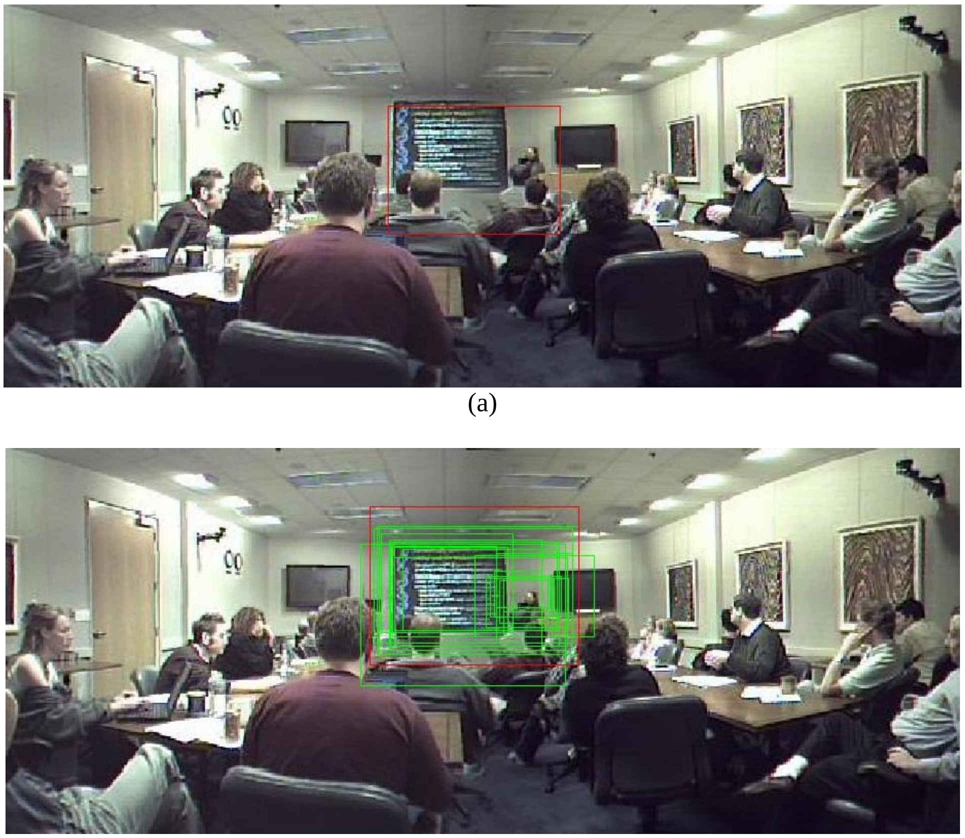
4. EXPERIMENTS

4.1. Providing low distortion user views

In this study, we deployed our system in a conference room, and grabbed 22 images with the panoramic camera during more than 10 presentations. We asked 14 subjects to mark each image with regions that they wanted to watch in the close-up view. We then used this data to compare the overall fidelity of views presented to users for three possible camera management schemes. The first is a panoramic system with no PTZ camera, in which only digital zooming is used for close-ups. The second is a hybrid system including the panoramic camera as well as a PTZ camera, but where the PTZ camera is used only to serve the most demanding (i.e. request that is expected to have the highest distortion) request. The third scheme was to use the method described in Section 3.1 to determine where the PTZ camera should be pointed, and which view should be presented to each user. In this method, each user is shown a close-up view of his/her requested region, taken from the PTZ camera if it covers the requested region and otherwise taken from the panoramic camera. Given that rule, the expected distortion can be computed for any placement of the PTZ view. The system chooses the region for the PTZ view which leads to minimum overall expected distortion. Figure 7(b) shows the set of requested views for an image, together with the PTZ positioning chosen to minimize overall distortion. Table 1 shows the result of this study. Clearly the camera management algorithm described in this paper results in significant overall improvement in image quality.

Statistical Results of Total Estimated Error	Without Using PTZ Camera	Using PTZ Camera for Most Demanding Request	Managing the PTZ Camera with Our Algorithm
Mean	1805.8	1584.8	761.8
Std.	693.9	599.5	362.3
Median	1908.1	1690.1	673.1

Table 1: Statistical results on estimated error under different situations.



(a)

(b)

Fig. 7: Choices for PTZ view (a) with no user requests, (b) with user requests (shown in green.)

When users selections are not available, the camera system will choose the PTZ view by itself to minimize the distortion between its representation w_S and the 'true' image function. Figure 7(a) shows a typical image, and the red rectangle indicates the region the system chose for the PTZ camera based on this image. Figure 7(b) shows the region chosen given many user requests. When the system builds its internal representation of the 'true' image, the difference image between the 'true' image and the system internal representation is also valuable for guiding the PTZ camera. For example as the system collects a background image for Figure 6 and computes the information gain based on the difference image, the system automatically chooses the red box in Figure 6 for minimizing the distortion.

4.2. Systematically improving image representation for a static scene

One experiment examined the decrease in squared error distortion as close-up views were taken to obtain higher resolution representations of the scene. As 'ground truth' we used 2304x648 image, which was generated by stitching 12 close-up images from a 2x6 grid taken by the PTZ camera. To test the FlySPEC algorithms we down-sampled this image into a 576x162 image which was taken as the output of the panoramic camera. This was then processed using a 48x18 grid of 12x9 pixel tiles for which DCT coefficients were computed for the spectral analysis. Using a $1/f^2$ spectral model, these DCT coefficients were used to predict the distortion in higher spatial frequency for any given choice of where to point the PTZ camera. The image used for this experiment is shown in Figure 8. The red square shows the choice of the single best region for a PTZ close-up.

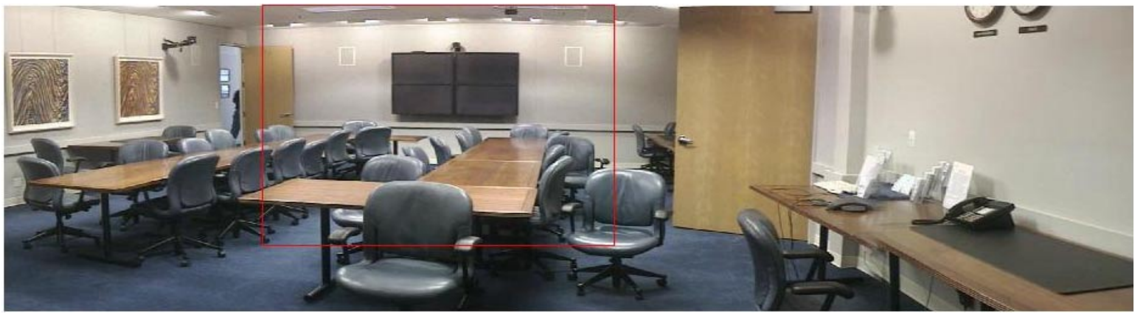


Fig. 8: Panoramic image of conference room used for mosaic experiments.

The ground truth image was used to simulate the improvement in $w_S(x,y,t)$ resulting from collecting a number of close-ups of various regions. For this experiment we considered the results of tiling the full image with N close-up images, for various values of N. A collection of N such close-ups could be used to generate a high-resolution mosaic view. After k of the N images are recorded, $w_S(x,y,t)$ can be thought of as a variable resolution mosaic in which some regions have higher available resolution then others. We compared the improvement in fidelity of $w_S(x,y,t)$ between the case where the k images were simply taken in a sequential left-to-right, top-to-bottom order, and the case when they were collected in an optimal order for minimizing distortion after k pictures.

The results of the experiments are shown in Figure 9. The graphs show the decrease in squared error distortion between $w_S(x,y,t)$ and $w(x,y)$ (the ground truth image of Figure 8.) with increasing numbers of tiles, for three different tile sizes. The dashed blue curves indicate the distortion decrease as mosaics are collected in a left-to-right, top-to-bottom order; the solid red curves indicate the distortion decrease as mosaics are collected in a greedy order based on our high-frequency distortion prediction. If all predictions about high-frequency distortions are correct, the solid curves should be perfect concave curves. The imperfections of these solid curves are caused by distortion mis-predictions. Although the predictions are not always correct, the solid curves are still well below the dashed curves. This fact strongly supports our approach of using spectrally based distortion predictions to choose an order for tiling a scene with images of small regions.

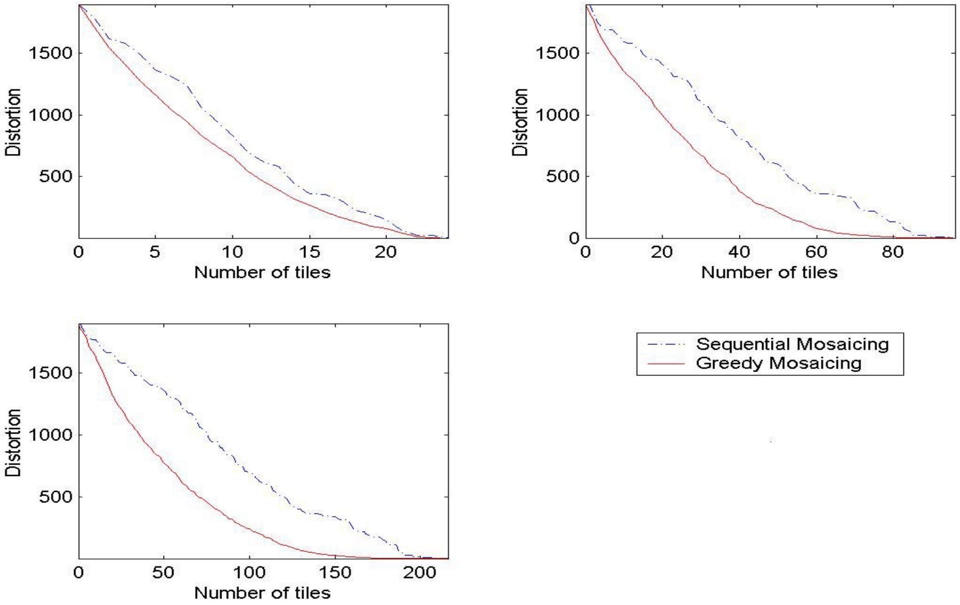


Fig. 9: Comparison of distortion decrease curves.

5. CONCLUSIONS AND FUTURE WORK

We have described a system for video capture and presentation that optimizes the image views for multiple remote users. The system combines a low-resolution wide-angle view with a controllable high-resolution view provided by a PTZ camera. A framework is described for using the low-resolution view, together with user requests when available, to determine how to control the PTZ camera. Experiments suggest that the framework can provide better overall quality to users than conventional camera management schemes.

The FlySPEC system is still in the early stages of development. In the future, we expect to improve the performance of the system by improving the models that underlie the framework. For example, probabilistic models that predict user interest in regions can allow the system to better satisfy users with less user intervention. We are also working on better integrating the views from panoramic and PTZ cameras, so that the representation $w_S(x,y,t)$ is not only an estimate used to determine camera control, but is available as a seamless variable resolution video image.

ACKNOWLEDGEMENTS

The authors would like to thank Hai Tao and Eric Saund for many useful discussions.

REFERENCES

1. J. Benesty, “Adaptive eigenvalue decomposition algorithm for passive acoustic source localization”, *Journal of Acoustics of America*, vol. 107, January 2000, 384-391.
2. M. Bianchi, “AutoAuditorium: a fully automatic, multi-camera system to televise auditorium presentations,” *Proc. of Joint DARPA/NIST Smart Spaces Technology Workshop*, July 1998.
3. M. Brandstein, “A pitch-based approach to time delay estimation of reverberant speech,” *Proc. IEEE ASSP Workshop Appls. Signal Processing Audio Acoustics*, 1997.
4. J. Foote and D. Kimber, “FlyCam: Practical Panoramic Video,” *Proceedings of IEEE International Conference on Multimedia and Expo,* vol. III, pp. 1419-1422, 2000.
5. Q. Huang, Y. Cui, and S. Samarasekera. “Content based active video data acquisition via automated cameramen,” *Proc. IEEE International Conference on Image Processing (ICIP) ’98*.
6. Q. Liu, Y. Rui, A. Gupta, and J. Cadiz. "Automating Camera Management in a Lecture Room," *Proceedings of ACM CHI2001*, vol. 3, pp. 442 – 449, Seattle, Washington, USA, March 31 - April 5, 2001.
7. S. Mukhopadhyay and B. Smith, “Passive Capture and Structuring of Lectures,” *Proc. of ACM Multimedia’99,* Orlando.
8. ParkerVision, http://www.parkervision.com/
9. PictureTel, http://www.picturetel.com/
10. PolyCom, http://www.polycom.com/
11. C. Wang and M. Brandstein, “A hybrid real-time face tracking system,” *Proc. of ICASSP98*, pp. 3737-3740, Seattle, May 1998.
12. K. Yamaashi, J.R. Cooperstock, T. Narine, and W. Buxton. “Beating the Limitations of Camera-Monitor Mediated Telepresence with Extra Eyes,” *Proc. of ACM CHI96*, 1996.
13. D. Dennett, *Consciousness Explained*, Little Brown & Company, 1991.
14. E. Simoncelli, B. Olshausen, “Natural Image Statistics and Neural Representation,” *Annual Revue of Neuroscience* 24:1193-216, 2001.
15. E. Saund, “Image Mosaicing and a Diagrammatic User Interface for an Office Whiteboard Scanner,” *Technical Report*, Xerox Palo Alto Research Center, 1999.