FLYSPEC: A Multi-User Video Camera System with Hybrid Human and Automatic Control

Qiong Liu, Don Kimber, Jonathan Foote, Lynn Wilcox, John Boreczky
{liu, kimber, wilcox, foote, johnb}@fxpal.com
FX Palo Alto Laboratory, 3400 Hillview Ave. Bldg. 4, Palo Alto, CA 94304

ABSTRACT

FlySPEC is a video camera system designed for real-time remote operation. A hybrid design combines the high resolution of an optomechanical video camera with the wide field of view always available from a panoramic camera. The control system integrates requests from multiple users so that each controls a virtual camera. The control system seamlessly integrates manual and fully automatic control. It supports a range of options from untended automatic to full manual control. The system can also learn control strategies from user requests. Additionally, the panoramic view is always available for an intuitive interface, and objects are never out of view regardless of the zoom factor. We present the system architecture, an information-theoretic approach to combining panoramic and zoomed images to optimally satisfy user requests, and experimental results that show the FlySPEC system significantly assists users in a remote inspection tasks.

Keywords

Collaborative and automatic camera control, gesture based camera control, panoramic video, video production, video communication, video conferencing, distance learning, webcams.

1. INTRODUCTION

Remote viewing over the Internet is increasingly popular for many applications, such as distance learning, security, and sporting events. Through remote video access, students can attend classes from their dorms, scientists can participate in seminars held in other countries, executives can discuss critical issues without leaving their offices, and web surfers can view interesting events through webcams.

This paper reports the design and development of FlySPEC, a hybrid video camera system that combines the best aspects of pan-tilt-zoom (PTZ) cameras and a panoramic camera. This allows remote users to see a wide-field view as well as optically zoom in on interesting details. With a conventional PTZ camera it is difficult to locate objects outside the camera's filed of view, especially at high zoom factors. The panoramic view is always available in the FlySPEC system, which allows users to repoint the PTZ camera without changing the zoom or needing to "hunt around."

A problem with conventional PTZ cameras is they do not scale well to multiple users, as there is only one camera and no good way to arbitrate conflicts in control requests. A particular contribution of this paper is a hybrid control system that integrates camera requests from multiple users. We present an information-theoretic framework that combines multiple camera sources so that control requests from multiple users are optimally satisfied given the available camera hardware. Additionally, this system integrates any automatic control algorithms to combine the convenience of an automatic "hands-off" system with the flexibility of manual control. An additional benefit is that the control algorithm may be trained on user requests. By learning the characteristics from many user requests over time, the system can learn to generate control output that better mimics human operation.

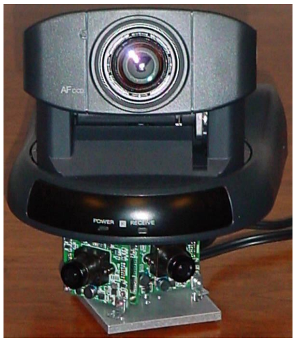


Figure 1. The FlySPEC Camera: A pan-tilt-zoom camera (top) is combined with a fixed panoramic multi-camera system (below)

Many existing camera-control systems are single-operator-controlled systems. To produce good videos with this kind of system, corporations and schools generally need to hire experts to operate them. Hiring video professionals may greatly increase the video production cost. It is also not very convenient for video users if the camera operator's schedule is not flexible. To address these problems, researchers 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 vision techniques that may not be robust enough for real world use. The approach taken by FlySPEC is to introduce remote viewers into the camera control loop to adjust the camera view when interesting shots are not being captured. In contrast to a professionally controlled camera system, the FlySPEC system is more economical and flexible. Compared with a fully automatic camera control system, the audience supervision may reduce unsatisfactory video shots that cannot be detected automatically. Besides these advantages, the FlySPEC system can also provide different views to different remote users to satisfy individual viewing needs.

The FlySPEC system design must address a number of issues. First, how to introduce human supervision into an automatic camera control loop? Second, how to build a system that does not require a camera control professional? Third, how to resolve conflicts when different control supervisors have different requests? Next, how to produce multiple video streams to satisfy different users' viewing preferences? Finally, how can we efficiently use a limited number of cameras to produce multiple high quality video streams for different viewing needs?

Our design provides new solutions for the five problem areas above. An automatic camera control system can save people's time and effort by continuously predicting interesting spots that the camera users want to watch. The FlySPEC system keeps the automatic camera control module intact and puts the human supervision module in parallel with the automatic control module. This design allows the automatic control unit and the human supervision mechanism to make separate decisions for the camera management. These camera management decisions are then fed to a control coordinator where human decision overrides the machine decision. When a human decision is not available, the machine decision is passed to the camera directly.

FlySPEC avoids using a professional camera operator by encouraging video users to supervise the camera management results. Since video users are always available in regular meetings or video conferences, allowing video users to supervise the video production is more convenient than meeting a professional's tight schedule. Besides the scheduling convenience, video users generally have more background and interests in the broadcast events, and have better chances to emphasize important events properly.

When multiple users are included in the camera management loop, it is unavoidable that the system will receive conflicting requests for camera management. The FlySPEC system resolves this problem through considering various requests together and generating camera control decisions that can benefit many requests at the same time. This approach crucially reduces superfluous movement of mechanical cameras in a FlySPEC system. It also helps to compensate for imprecise user requests by considering many requests together.

For users who have viewing preferences that are very different from most others, a single-stream camera system may not provide them with what they want to watch. The FlySPEC system handles this problem by generating multiple video streams for requests that are quite different. When all users have the same or similar viewing preferences, the FlySPEC system will not use more bandwidth than a single-stream camera system.

Because the FlySPEC system must occasionally generate multiple video streams, another important research issue is how to generate multiple high-resolution video streams with a limited number of cameras. The FlySPEC camera is constructed with a set of PTZ cameras and a panoramic camera. When the number of requested video streams is less than the number of PTZ cameras, each stream gets a dedicated camera. When the number of requested video streams exceeds the number of PTZ cameras, FlySPEC shares the PTZ cameras among multiple streams. For requests that cannot be served by a PTZ camera, the system generates the requested video from a panoramic camera that covers the entire field of view [4]. There are different criteria for choosing PTZ camera views. The criterion used by FlySPEC is to maximize the real world information acquired by all users.

For more detailed system explanation, the rest of the paper is organized as follows. In Section 2, we present the FlySPEC system including the camera hardware, the FlySPEC control system, the graphical user interface for remote users to manage the camera system, and the algorithm for managing camera resources. We present FlySPEC viewing experiments in Section 3. Concluding remarks and future work are given in Section 4.

2. THE FLYSPEC SYSTEM

The FlySPEC system can be described in four parts: the hybrid camera hardware, the control structure, the camera management algorithm, and the graphical interface for user interaction with the system.

2.1 The FlySPEC Camera

Figure 1 shows a FlySPEC camera. This is a hybrid camera constructed by combining a PTZ camera and a panoramic camera. The panoramic camera covers every possible view, and serves different video requests through electronic pan/tilt/zoom. The PTZ camera can be used to capture details of small objects. The close proximity of the panoramic and PTZ cameras makes it easy to find the correspondence between the PTZ view and a region in the panoramic view. This correspondence is useful for controlling the PTZ camera based on low-resolution panoramic video. It also allows seamless switching of the video source between the panoramic and PTZ cameras.

2.2 The Control Structure

Figure 2 (a) shows a typical control structure of an automatic camera management system. Here, the control unit acquires environmental information with sensors, such as microphones and other video cameras. Based on the sensory information, the control unit can localize the region of interest, and point the camera at the interesting spot. For example, the system can control cameras based on audio [1, 3, 6, 9], vision [2, 5, 6, 11], and other sensory information [8, 10]. Since systems with this control structure do not analyze the output video, the sensors and the controlled camera must be aligned exactly to achieve quality video output. This system also requires the control algorithm to accurately predict the spot in which audience members are interested. However, computer prediction of the region of interest is still an open problem.

To avoid the sensor alignment problem, some scientists prefer to use the video output from the controlled camera directly. This type of system is diagrammed in Figure 2 (b). With this type of control system, the computer will track an interesting spot continuously. Although this overcomes the alignment problem, the region of interest must still be located. Compared to the system of Figure 2 (a), the system of 2 (b) is limited by the view of the controlled camera. For example, if the system loses the tracked object, it is hard for the system to recover. The system also has difficulty finding interesting regions outside the camera view. In summary, state-of-the-art automatic camera control systems are still unable to produce video as well as humans do. To improve video quality, the camera control system still requires human involvement during its video capture process.

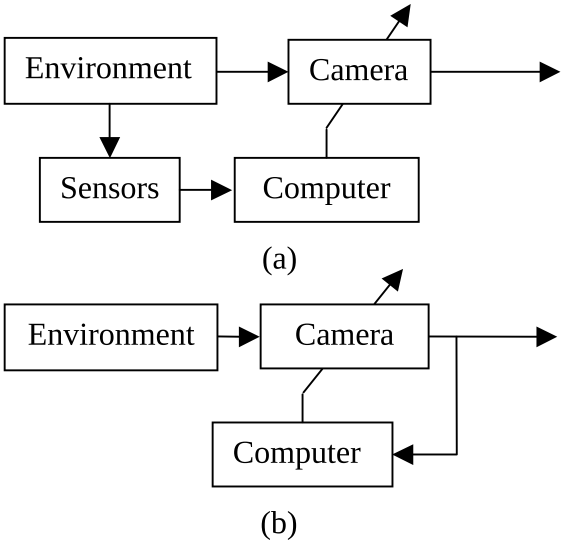


Figure 2. An automatic camera control system. (a) A camera control system guided by extra environment sensors. (b) A camera control system which uses the view of the controlled camera.

Figure 3 (a) shows the control structure of a traditional human operated camera system. Here, the human operator continuously monitors the video output of the camera, and adjusts the camera pan/tilt/zoom based on the video output and past experience. If the operator is provided with additional sensors, such as a panoramic camera or a stereo sound system, the system can be described with Figure 3 (b), and the operator can generally act more responsively to environment changes. Since the system described in Figure 3 (b) generally allows the operator to produce better video output, the FlySPEC system uses this approach.

Compared to a state-of-the-art automatic camera control system, a human controlled camera is almost always better at capturing regions of interest. On the other hand, a carefully designed automatic camera control system can predict interesting events most of the time, and can relieve a human operator from having to continuously control the camera. To preserve the benefits of both methods, we combine the human-operated system and the automatic system as shown in Figure 4.

In this system, both the human operator and the automatic control unit can access data from sensors such as the panoramic camera and microphone arrays. In addition, the human operator can monitor the output video. With this system, the human operator and the automatic control unit can make separate decisions based on environmental information. These decisions may be different. To resolve conflicts, the human decision and the control unit decision are sent to a coordinator unit before the decision is sent to the camera. Since we believe the human choice is generally better than the automatic control, the human's decision overrides the automatic unit decision inside the coordinator. When no human decision is available, the decision made by the automatic control unit is used instead.

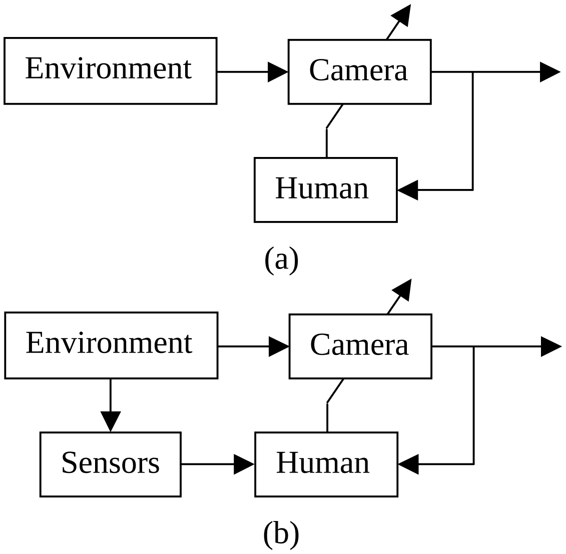


Figure 3. Human-operator controlled camera system. (a) The operator can only monitor the environment with the controlled camera. (b) The operator can monitor the environment with other sensors.

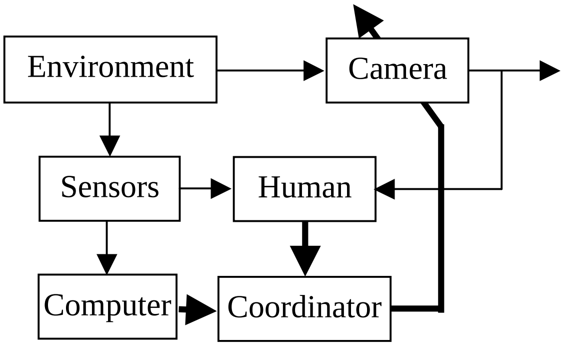


Figure 4. A single-user controlled FlySPEC system. The thick lines in this figure indicate the flow of camera control signals.

With this system, the human operator's job is to monitor the video production process instead of operating the camera continuously. To ensure the quality, the human operator only needs to adjust the camera when the automatic system misses the region of interest. Thus the system is fully automatic when no human operator controls it. For an automatic system, which may miss a large number of interesting events, human operator can drastically decrease the miss rate. Compared with a manual camera control system, this system can substantially reduce the human operator's effort. In summary, this system allows users to make the tradeoff between operator effort and video quality.

2.3 The Graphical User Interface For Supervising The Video Capture

To supervise the video production process, we designed the graphical user interface shown in figure 5. In the web browser window, the upper window shows a resolution-reduced video from the panoramic camera, and the lower window shows the close-up video produced by the FlySPEC system. In other words, the panoramic camera view provides sensory information about the environment to the human operator, and the lower window provides output video to the operator for feedback. Using this interface, the human operator adjusts the video output by selecting an interesting region in the panoramic view with a simple mouse-based gesture. After the interesting region is marked with a line or a circle, the region inside the bounding box of the mark will be shown in the close-up view window.

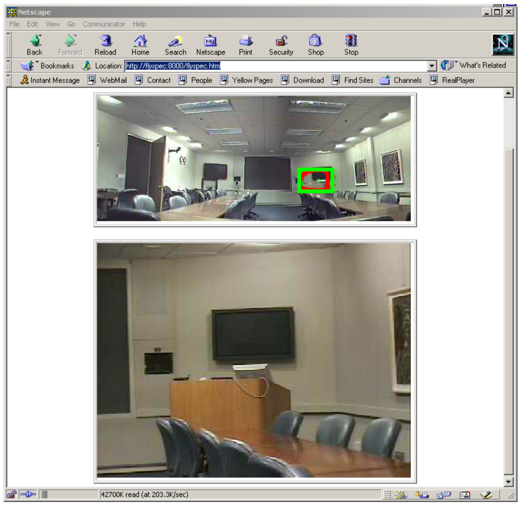

Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://flyspec:8000/flyspec.html What's Related
Instant Message WebMail Contact People Yellow Pages Download Find Sites Channels RealPlayer
42700K read (at 203.3K/sec)

Figure 5. Web-based Graphical User Interface for Supervising the Video Production

2.4 Organize the Audience Members for Camera Control

In previous sections, we described a flexible tradeoff between the human operator's effort and the video quality. However, it still needs a specific operator to monitor the process for quality video production. To further decrease operator effort, we propose a system that uses multiple remote viewers in the camera control process. Audience participation in the video production has several advantages:

- Audience members are available for most video production events, and are a cheap source of human input to the production control process.
- Audience members are normally very interested in important broadcasting events.
- Audience members are humans. Therefore, they are more knowledgeable than a state-of-the-art automatic system.
- Audience members normally have more background on the broadcasting event than the professional camera operator.
- Distributing the video monitor and camera adjustment to many audience members makes the system robust, more accurate regarding general opinion about regions of interest, and reduces the effort involved by any one person.

Due to these advantages of involving audience members in video capture, the system described in Figure 4 can be further improved as shown in Figure 6.

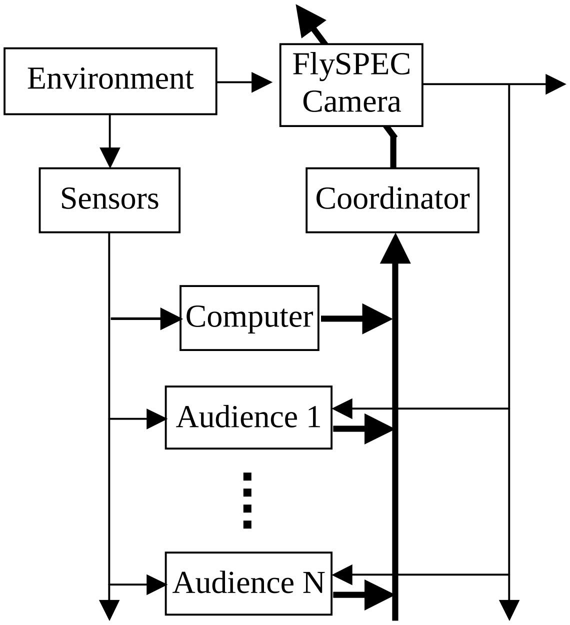


Figure 6. System that allows audience members to supervise the video production process.

This control structure can be implemented as shown in Figure 7. Although using audience members to supervise the video capture process has many advantages, it also generates the problem of how to resolve the conflicts among different control commands.

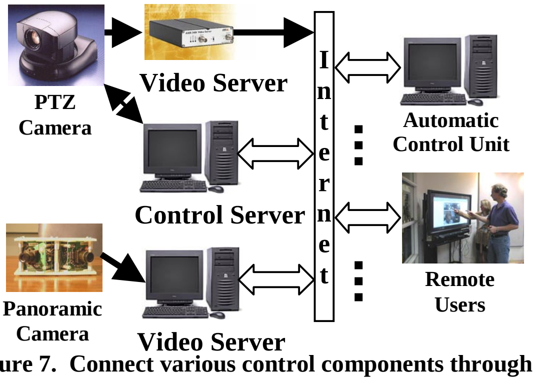


Figure 7. Connect various control components through the Internet.

A management algorithm coordinates conflicting commands. Commands from audience members are separated into groups, and different cameras are used to serve different audience groups. In our FlySPEC system, the PTZ camera can perform optical zoom and allows audience members to see small objects clearly. On the other hand, the panoramic camera can provide different video streams to multiple users simultaneously, and can provide reasonable video for a wide field of view. The FlySPEC coordinator uses the PTZ camera to serve the group with the most demanding image requests, and uses the panoramic camera to serve general audience members who do not need as much resolution.

Separating audience members into groups can be treated as an optimization problem. Our approach is motivated by an analogy with rate-distortion theory. Displays are limited information channels, and the goal is to optimize those channels to provide each user with the highest fidelity view possible.

The FlySPEC cameras produce a panoramic video image *P* and a close-up image *Z* of a region R_Z. Because the cameras are approximately collocated, *P* and *Z* can be thought of as limited resolution representations of an 'ideal' image *W*. Because *P* is relatively low resolution and *Z* covers only a limited region R_Z, the ideal image *W* is unknown to the system. However conceptually *P* and *Z* can be used to estimate *W*.

FlySPEC presents each user *u* with a resolution-reduced version of *P* and a close-up image C_u of a region R_u. Since the panoramic image *P* is the same for all users, the control task is choosing the C_u for user *u* to provide the best estimation of the real world. We formalize this by assuming there is a value Q_u that describes the quality of image C_u. The quality value Q_u can be defined with

$$Q_u = \int w_u(x,y) I_u(x,y,R_Z) \qquad (1)$$

where w_u is a weighting window and $I_u(x,y,R_Z)$ is a measure of how much more information would be provided to the user around (*x*,*y*) compared with what they already know. Roughly speaking, $I_u(x,y,R_Z)$ is the 'objective' aspect of Q_u, describing how much 'raw information' is gained by the user, whereas $w_u(x,y)$ can be thought of as the 'subjective' aspect of Q_u describing how important that particular information is.

$I_u(x,y,R_Z)$ can be modeled with varying levels of sophistication. We will take $I_u(x,y,R_Z)$ as the actual 'true resolution' presented to the user around *x*,*y*. This is the measure of actual information provided by camera (in comparison with electronic zoom, which provides no real additional information). The true resolution of the user's view C_u is limited by the resolution actually available to the FlySPEC system, as well as the resolution of the user's display. Assuming a small region around *(x,y)* has M_{PTZ} pixels in the PTZ view, $M_{panoramic}$ pixels in the panoramic view, and $M_{display}$ pixels in C_u, $I_u(x,y,R_Z)$ can be modeled as resolution gain as follows:

$$I_u(x,y,R_Z) = \min\left(\frac{M_{PTZ}}{M_{panoramic}}, \frac{M_{display}}{M_{panoramic}}\right) \qquad (2)$$

If position (*x*,*y*) does not have a PTZ camera view, M_{PTZ} is set to $M_{panoramic}$ at (*x*,*y*). $w_u(x,y)$ can also be modeled with varying levels of sophistication. We use an indicator function based on the user's selection:

$$w_u(x,y) = \begin{cases} 1 & if\,(x,y) \in Selection \\ 0 & otherwise \end{cases} \qquad (3)$$

With the definition of $I_u(x,y,R_Z)$ and $w_u(x,y)$, if audience member *u* marks region R_u (region inside the bounding box of the mark) as the interesting region, the quality of image C_u is defined as:

$$Q_u(R_Z, R_u) = \int_{R_u} I_u(x,y,R_Z) \qquad (4)$$

If *K* is the number of users, the overall system quality *Q* is the sum of each user's image quality Q_u.

$$Q = \sum_{u=1}^{K} Q_u(R_Z, R_u) \qquad (5)$$

Searching for the best PTZ camera view R_Z directly with eq. (5) is a very time consuming task. From equations (1)-(5), we realize that the best PTZ camera view R_Z should overlap with users' selections, and this view should not include large regions not covered by users' selections. To make the algorithm feasible, the FlySPEC system generates candidates for R_Z based on this prior knowledge, and compares the overall quality *Q* over those candidates. Bounding boxes of all possible combinations of users' selections are considered as candidates. Let *A* be that set of bounding boxes. Then the PTZ camera view R_Z can be optimized as follows.

$$R_Z = \arg\max_{R \in A} \sum_{u=1}^{K} Q_u(R, R_u) \qquad (6)$$

Eq. (1-6) works for a system regardless of whether it has image memory. For a one-PTZ-camera FlySPEC system without memory, resolution gain $I_u(x,y,R_Z)$ is the same for all time instants. For a system with image memory, $I_u(x,y,R_Z,t)$ will depend on what data the system saves, the history of users' selections, and the time-varying 'ideal' image *W*(*t*). For example, if more pixels at a position (*x*,*y*) are captured by the PTZ camera and there is no intensity change at that position in the panoramic view, $I_u(x,y,R_Z,t)$ can be updated with a higher value based on how many pixels are captured and saved for that position. On the other hand, when intensity change at that position is detected by the panoramic camera, $I_u(x,y,R_Z,t)$ will be set to 1 and all PTZ-camera captured pixels corresponding to that position will be deleted. With image memory, the PTZ camera can be used more efficiently to provide high quality video to different audience members if their selections are significantly different.

This optimization algorithm can also be extended beyond one FlySPEC system. For example, if we put a FlySPEC system at the back of a conference room and a video capture card can capture high-resolution video of a screen in front of the conference room, then the system can perform the optimization based on both the camera and the screen capture. In this optimization, $I_u(x,y,R_Z,t)$ values at positions corresponding to the screen can be set to a high constant depending on the resolution of the screen capture. If the $I_u(x,y,R_Z,t)$ values at some positions are set to a high constant, the PTZ camera will not normally move to those positions during the control process. This optimization process can be naturally extended to systems with multiple FlySPECs and other high-resolution image sources.

2.5 Learning Camera Control Based On Human Camera Operation

It is difficult to describe human camera operation with simple rules, or to collect sufficient data to adequately train an automatic control system. To increase system flexibility, a learning system can be run in parallel with the automatic control unit, so it can learn camera operation from human inputs. Let a_1, a_2, …, a_R be measurements from environmental sensors, and (x,y) on the panoramic image correspond to a position of the PTZ camera. Then, the destination position (X,Y) for the PTZ camera can be estimated as:

$$\begin{aligned}(X,Y) &= \arg\max_{(x,y)}\left\{p\left[(x,y)\,|\,(a_1,a_2,\cdots,a_R)\right]\right\}\\ &= \arg\max_{(x,y)}\left\{\frac{p[(a_1,a_2,\cdots,a_R\,|\,(x,y)]\cdot p(x,y)}{p(a_1,a_2,\cdots,a_R)}\right\}\\ &= \arg\max_{(x,y)}\left\{p\left[(a_1,a_2,\cdots,a_R)\,|\,(x,y)\right]\cdot p(x,y)\right\}\end{aligned} \tag{7}$$

The zoom level can also be estimated in a similar way. Assuming a_1, a_2, …, a_R are conditionally independent, the camera position can be estimated as:

$$\begin{aligned}(X,Y) &= \arg\max_{(x,y)}\left\{p\left[(x,y)\,|\,(a_1,a_2,\cdots,a_R)\right]\right\}\\ &= \arg\max_{(x,y)}\left\{p[a_1\,|\,(x,y)]\cdot p[a_2\,|\,(x,y)]\cdots p[a_R\,|\,(x,y)]\cdot p(x,y)\right\}\end{aligned} \tag{8}$$

The probabilities in eq. (8) can be estimated online. For example, Figure 8 shows the users' selections during a period of a meeting, and it is straightforward to estimate p(x,y) on these selections.

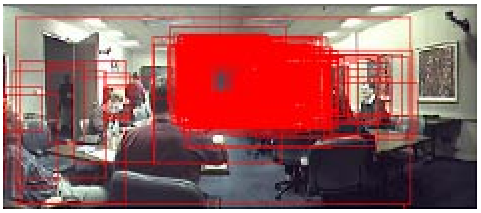


Figure 8. Users' selections during a meeting period in a corporate conference room.

Using progressive learning enables our system to better adapt to environmental changes. Some sensors may become less reliable. For example, if desks are moved so as to block the sound path of a microphone array. To adapt to these changes, a mechanism can learn how informative each sensor is. Assume (U,V) is the interesting position estimated by a sensor (e.g. a camera or a microphone array), and (X,Y) is the camera position decided by the users. How informative the sensor is can be evaluated through online estimation as follows:

$$\begin{aligned}&I\left[(U,V),(X,Y)\right]\\ &= \int_{(U,V),(X,Y)} p\left[(U,V),(X,Y)\right]\cdot\log\frac{p\left[(U,V),(X,Y)\right]}{p(U,V)\cdot p(X,Y)}\end{aligned} \tag{9}$$

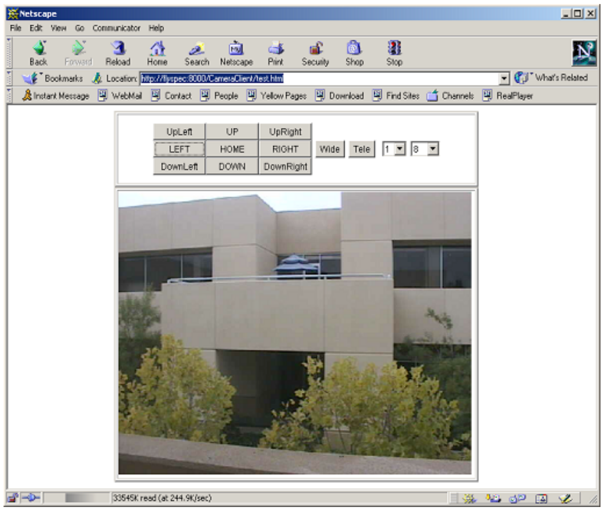
Evaluation of eq. (9) gives us the mutual information between (U,V) and (X,Y). The higher the value, the more important the sensor is to the camera control. When a sensor is broken, disabled, or yields poor information for any reason, the mutual information between the sensor and the human selection will decrease to a very small value, and the sensor will be ignored by the control software. This is helpful in allocating computational power to useful sensors. With similar techniques, the system can disable the rule-based automatic control system when the learning system can operate the camera better.

2.6 Control Client for Seamless Camera Switching

When multiple audience members request different views from the FlySPEC, it is possible that some requests cannot be satisfied by the PTZ camera. When audience members select a region where another source can provide higher quality video (e.g. a screen in a conference room), it is better to provide that. To satisfy users' personal selections, the FlySPEC client selects a proper video source based on the user's request and available video sources. With this functionality, the control client can extract video from the panoramic camera when the PTZ camera cannot reach the requested position or the region is too large for the PTZ camera's field-of-view. The control client can also use video from a source outside of the FlySPEC system if it has better quality video of the desired region.

3. FLYSPEC SYSTEM EXPERIMENTS

We performed three experiments with the FlySPEC system. In the first, we tested whether the FlySPEC control interface was convenient for camera operation. In the second experiment, we tested whether it was reasonable to deliver video from the combination of a PTZ and a panoramic camera. In the third experiment, we tested our optimization algorithm (i.e. equation (1)-(6)) with data collected during more than 10 talks in a corporate conference room.



Location: http://flyspec:8000/CameraClient/test.htm

33545K read (at 244.9K/sec)

Figure 9. Control Interface for a PTZ Camera

3.1 Control Interface Experiment

Our control interface experiment has two goals. First, we want to know if the FlySPEC system can speed up PTZ camera control compared to a conventional control system. Second, we want to compare the users' opinions on the FlySPEC control interface and the classical camera control interface.

The FlySPEC system used in this experiment faces a yellow office building about 15 meters away. To study how users can locate a particular object, we printed out four letter-size blue posters and randomly attached them to visible walls of the adjacent office building. Each poster has a small green character at the center that cannot be recognized without optical zoom available only on the PTZ camera. The characters on the posters are 'S', 'P', 'E', 'C' respectively. The task of a remote user in this study is to visit the posters in the sequence 'S'   'P'   'E'   'C' as fast as possible.

The interface we designed is shown in Figure 5. With this interface, a remote user can control the PTZ camera by marking an interesting region with the mouse in the panoramic video window. The conventional PTZ control interface is shown in Figure 9. This allows a remote user to pan/tilt/zoom the PTZ camera with the control buttons shown in the upper window. If the user continuously pushes a button, the camera will continuously move in the requested direction. If the user releases the button, the camera will stop immediately. In the control window, the two numerical controls are used to set the zoom speed and the pan/tilt speed. A control interface similar to the one shown Figure 9 can be found in many commercial products and research papers. That is the reason that we compare this interface with the FlySPEC interface.

Table 1. Time used for showing characters in a predefined sequence with the interfaces shown in Figure 3 and Figure 4.

Label for every user and statistic items	Time used for showing characters in a predefined sequence with the interface shown in Figure 5. (Second)	Time used for showing characters in a predefined sequence with the interface shown in Figure 9. (Second)
1	25	215
2	10	95
3	15	130
4	19	150
5	20	298
Mean	17.8	177.6
Standard deviation	5.04	71.76
Median	19	150

The speed comparison experiment was performed in two steps. In the first step, five users were instructed to operate the camera individually with the interface shown in Figure 9. The time for finding the proper sequence was recorded for every user, and shown in Table 1. Next, all posters on the building were rearranged. To make a fair comparison, we attempted to keep the camera travel distance similar in all experiments. With rearranged posters, the subjects were instructed to use the improved interface shown in Figure 5. The time for finding the proper sequence is also recorded for every user, and shown in Table 1.

A Wilcoxon Signed Ranks Test was performed to test the significance of the time difference between the FlySPEC system and the conventional PTZ camera control. According to the Wilcoxon test, times spent on these two systems are significantly different at the p=0.05 level. Based on this result, it is reasonable to believe that the FlySPEC system allows remote users to locate interesting regions faster than a conventional PTZ control system. In this study, we did not test the situation where interesting regions dynamically emerge in the environment. Since the PTZ-only system cannot usually view the whole environment, it is reasonable to believe that the FlySPEC will outperform the PTZ-only system in a dynamic environment.

Table 2. Survey results of control interfaces shown in Figure 3 and Figure 4

Label for every user and statistic items	How convenient to operate the FlySPEC system? (Very easy = 1, Very difficult = 5)	How convenient to operate the classical PTZ camera system? (Very easy = 1, Very difficult = 5)
1	2	4
2	2	3
3	1.5	4
4	1	4
5	1	4
Mean	1.5	3.8
Standard deviation	0.4472	0.4
Median	1.5	4

We also interviewed the subjects about the ease of use for the interfaces shown in Figure 5 and Figure 9. The subjects were asked score the two interfaces for ease of use. The scores collected from the subjects are listed in Table 2. After performing a Wilcoxon Signed Ranks Test on data in Table 2, we conclude that the control interface shown in Figure 5 is easier to use than the interface shown in Figure 9 at the p = 0.05 significance level. Besides this experiment, we note the work in [12] which also supports the interface design for FlySPEC.

3.2 Experiment on Serving the Audience with a PTZ Camera and a Panoramic Camera

In this paper, we argued for benefits of a system combining both a PTZ camera and a panoramic camera. Is it a good solution for a video server? The experiment in this section tests if it is desirable to provide customized views to remote users, and if it is practical to let multiple users share one or several high-resolution PTZ cameras.

In this study, we deployed our system in a conference room, and grabbed images with the panoramic camera during more than 10 presentations. Then we asked 14 subjects to mark regions they wanted to watch in the close-up view. Based on requested regions marked on every panoramic image, we selected a region and measured the horizontal distances between this region center and all other region centers. These measured distances are then normalized by ¼ width of the selected region. After we got all normalized distances based on one region, we select a different region in the same image as the reference and get another set of normalized distances. This procedure is performed until all regions are used as the reference region once.

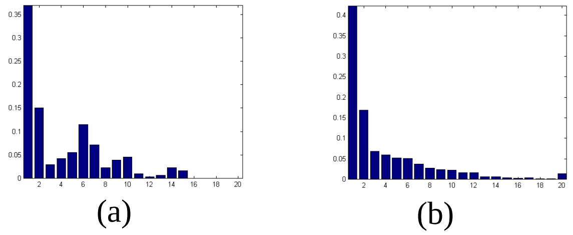


Figure 10. Histogram of the Distances among Various Marked Regions: (a) Distance Histogram Based on One Image. (b) Distance Histogram Based on All Images.

The histograms of these normalized distances are shown in Figure 10. If all users select very similar regions on an image, we should notice a high single peak near bin 1 (within ¼ region width). In Figure 10, high peaks near bin 1 suggest that many users select similar regions. This result strongly supports sharing one or several high-resolution PTZ cameras among audience members. The peaks far from bin 1 in (a) and the heavy tail in (b) suggest that users do not consistently request the same regions. This result supports our idea of customizing videos for remote audience members, and sharing one or several high-resolution PTZ cameras among them.

3.3 Experiment on Selecting the PTZ Camera View for Electronic Zoom Factor Reduction

Here, we used the same system as in the previous experiment. We collected image shots, corresponding user requests, and PTZ camera views selected by our camera management algorithm during more than 10 presentations. Figure 11 is a typical image collected during our experiment. In this figure, users' requests are represented by green boxes. The PTZ camera view selected by our algorithm is represented by a red box.

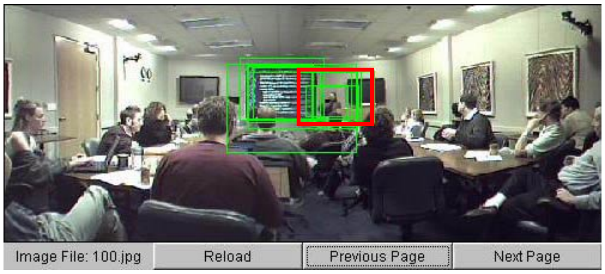


Figure 11. Users' Video Requests (Green) and the Calculated PTZ Camera Position (Red)

From eq. (2-4), it is not difficult to realize that the image quality value Q_u is the number of real pixels that are helpful to the C_u construction. Since the number of pixels is not a direct measure of system performance, we transform Q_u into an electronic zoom-in factor for discussion here. The electronic zoom-in factor is the ratio between the number of displayed pixels and number of received pixels. Because Q_u reflects the number of pixels transmitted to u, the transformation from Q_u to an electronic zoom-in factor is straightforward. Let K be the number of users who request video from the server, and r be the electronic zoom-in factor. Considering all users have 320x240 view windows, the relation between the electronic zoom-in factor and Q_u can be described with eq. (10).

$$r = \frac{1}{K} \cdot \sum_{u=1}^{K} \frac{320 \cdot 240}{Q_u} \qquad (10)$$

The electronic zoom-in factor of the shot shown in Fig. 11. will be 4.84 without a PTZ camera. This zoom-in factor indicates that the system must generate 4.84 times as many pixels for the user image as are actually available from the FlySPEC camera. If the system uses the PTZ camera to serve the most demanding zoom-in request, the zoom-in factor reduces to 3.83. Through managing the PTZ camera with our algorithm, the zoom-in factor is further reduced to 2.57.

Results in Table 3 show considerable zoom factor reduction achieved by our camera-management algorithm. If we install two or more PTZ cameras in the system or activate image-caching mechanism, and perform optimization over time and multiple cameras, this ratio can be further decreased.

Table 3. Statistical results of demanding-pixels/available-pixels under different situations

Statistical Results of Zoom-in Factors	Without Using PTZ Camera	Using PTZ Camera for Highest-Zoom-in Request	Managing the PTZ Camera with Our Algorithm
Mean	5.484	4.235	2.305
Std.	1.599	1.21	0.774
Median	5.29	3.78	2.13

4. CONCLUDING REMARKS AND FUTURE WORK

We have presented the design, implementation, and experimental results of a camera system called FlySPEC that combines the advantages of panoramic and PTZ camera systems. This FlySPEC system has potential in many applications, such as video surveillance and teleconferencing, etc. Experimental results strongly support our approach of combining a panoramic camera and a PTZ camera, and our camera management algorithms. The experiments also revealed that the FlySPEC system facilitates a remote user in locating details of an interesting region much faster than a conventional PTZ camera system. This project can be extended in many aspects as we develop better camera management algorithms and online learning algorithms.

5. ACKNOWLEDGMENTS

The authors would like to thank Hai Tao for useful discussions.

6. REFERENCES

[1] Benesty, J., 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] Brandstein, M., 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,” in Proc. IEEE International Conference on Image Processing (ICIP) ’98.

[6] Q. Liu, Y. Rui, A. Gupta, 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, 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, W. Buxton. Beating the Limitations of Camera-Monitor Mediated Telepresence with Extra Eyes, in Proc. of ACM CHI96, 1996.