LEARNING AUTOMATIC VIDEO CAPTURE FROM HUMAN'S CAMERA OPERATIONS

Qiong Liu, Don Kimber

FX Palo Alto Laboratory, 3400 Hillview Avenue, Palo Alto, CA94304

ABSTRACT

This paper presents a video acquisition system that can learn automatic video capture from human's camera operations. Unlike a predefined camera control system, this system can easily adapt to its environment changes with users' help. By collecting users' camera-control operations under various environments, the control system can learn video capture from human, and use these learned skills to operate its cameras when remote viewers don't, won't, or can't operate the system. Moreover, this system allows remote viewers to control their own virtual cameras instead of watching the same video produced by a human operator or a fully automatic system. The online learning algorithm and the camera management algorithm are demonstrated using field data.

1. INTRODUCTION

Videoconferencing has been gradually adopted by general public for watching meetings, presentations, and performances remotely. To reduce the cost of capturing meetings and presentations, engineers have constructed various video acquisition systems. For instance, the AT&T's Automated Cameramen [3] tracked moving objects based on the radial-profile difference between the acquired image and the background image. The Bell Core's Auto-Auditorium [1] provided audiences a "combination shot", with the speaker placed in a picture-in-picture box in the lower corner of the slide image when the system could not determine automatically whether the most important image should be of the speaker or of the screen. In Cornell's lecture capturing system [5], a video alternating between two cameras helped the system to produce more engaging presentations. The Microsoft's ICAM system [4] mimicked the structure of a video production team via specific rules. There are three major drawbacks of these predefined fully automatic camera control systems.

First, these systems do not have sufficient bad-shot-correction mechanisms. These fully automatic systems can barely avoid bad shots based on state-of-the-art audio/vision techniques. If a camera control system does not allow remote viewers to correct the problem when a bad shot happens, remote users may miss important information during a meeting or conference.

Second, these systems force all remote users to watch the same video stream without considering users' various preferences. This prevents remote users from following different events in a meeting. For example, when the camera for broadcasting focuses on the speaker of a lecture, a remote user may want to check the white board. If the system cannot let the remote user check the white board, this user may get lost.

Third, system installers have to set a large number of environment-specific parameters for these systems. If this kind of predefined camera control system is set in a time varying environment, such as a multifunctional room, it will be difficult for the system to have good performance.

In this paper, we present a camera control system that tries to attack these major problems, and our effort will mainly focus on solving problem 3. In section 2, we briefly introduce the camera hardware and the control interface. In section 3, we describe the control strategy and the learning approach. Experimental results on field data are given in section 4. Conclusions are presented in section 5.

2. CAMERA HARDWARE AND CONTROL INTERFACE

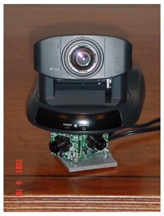


Figure 1. The Camera Hardware Construction

To provide a high quality virtual camera to every remote user, we construct the camera hardware by installing a pan/tilt/zoom (PTZ) camera on top of a panoramic camera. The hardware construction is shown in Figure 1. With this construction, remote users may monitor the entire camera operation environment with the panoramic camera while they may also obtain details of an event with the PTZ camera.

To obtain remote users' opinions on controlling their virtual cameras, we provide the interface shown in Figure 2 to remote users. With this interface, a remote user may request close-up video by selecting his/her region of interest in the overview window. When our video server receives the selection information, the server will send the user a close-up video of that region. This close-up video may come from the PTZ camera or the high-resolution panoramic camera depending on our camera management strategy and all users' requests.

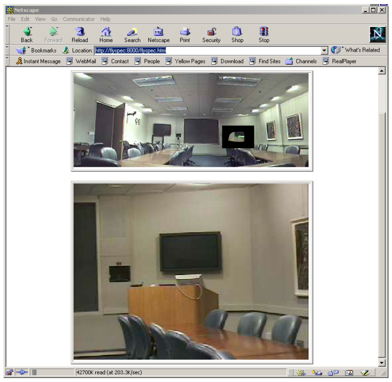

Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://flyspec:8000/flyspec.html
What's Related
Instant Message WebMail Contact People Yellow Pages Download Find Sites Channels RealPlayer

Figure 2. Web-based Graphical Interface for Remote Users to Control Their Virtual Camera

3. LEARNING CAMERA CONTROL FROM USERS' VIDEO REQUESTS

Conceptually, the ideal image received by the camera system may be represented with $f(x,y,t)$, where *x* and *y* are coordinates used by the panoramic image. Due to the limitations of sensors, a practical system may only obtain an approximation $\hat{f}(x,y,t)$ of the ideal signal $f(x,y,t)$. To efficiently use available sensors, the system moves the PTZ camera to maintain high quality of acquired signals. Moreover, the system also includes video buffer to cache the past image, $\hat{f}_{t-T}$, for future references when some image regions do not change over a short period of time. Let $F(\omega_{xy},t)$ and $\hat{F}(\omega_{xy},t)$ be the spectrum representation of $f(x,y,t)$ and $\hat{f}(x,y,t)$ respectively, where ω_{xy} is the rotational spatial frequency. Denote $\{R_i\}$ as a set of non-overlapping small regions, $p(R_i,t\,|\,O)$ as the probability of viewing region-*R_i* details conditioned on environmental observation *O*, and *T* as a short period of time between video frames. The total distortion $D[\hat{f}_{t-T}, f_t]$ between users' requested images and the real image might be modeled with:

$$D[\hat{f}_{t-T}, f_t] \approx \sum_i p(R_i,t\,|\,O)\cdot\int_{R_i} |\,\hat{F}(\omega_{xy},t-T)-F(\omega_{xy},t)\,|^2 \; d\omega_{xy} \quad .(1)$$

Assume $\hat{F}_{R_i}(\omega_{xy},t)$ is a band limited representation of $F_{R_i}(\omega_{xy},t)$ and the cutoff frequency of $\hat{F}_{R_i}(\omega_{xy},t)$ is $a_i(t)$. Let $F_{M,R_i,t}$ be $F_{R_i}(\omega_{xy},t)-F_{R_i}(\omega_{xy},t-T)$ and $F_{S,R_i,t}$ be $F_{R_i}(\omega_{xy},t)-\hat{F}_{R_i}(\omega_{xy},t)$, the integration may be rewritten as:

$$\begin{aligned}&\int_{R_i} |\,\hat{F}(\omega_{xy},t-T)-F(\omega_{xy},t)\,|^2 \; d\omega_{xy} \\ &= \int_{R_i,\omega_{xy}\le a_i(t-T)} |\,F_{M,R_i,t}\,|^2 \; d\omega_{xy} + \int_{R_i,\omega_{xy}>a_i(t-T)} |\,F_{S,R_i,t}\,|^2 \; d\omega_{xy}\end{aligned} \quad . \qquad (2)$$

This integration reflects the distortion between the real image and the cached image, where the first term on the right side reflects the distortion caused by environmental changes, and the second term reflects the distortion caused by environmental details. By sampling region R_i at frequency $a_i(t)$ and updating the cached image, the expected distortion reduction is:

$$\Delta D_{c,R_i} = \begin{cases} \displaystyle\int_{R_i,\omega_{xy}\le a_i(t-T)} |\,F_{M,R_i,t}\,|^2 \; d\omega_{xy} + \int_{R_i,a_i(t)\ge\omega_{xy}>a_i(t-T)} |\,F_{S,R_i,t}\,|^2 \; d\omega_{xy} & * \\ \displaystyle\int_{R_i,\omega_{xy}\le a_i(t-T)} |\,F_{M,R_i,t}\,|^2 \; d\omega_{xy} - \int_{R_i,a_i(t)<\omega_{xy}\le a_i(t-T)} |\,F_{S,R_i,t}\,|^2 \; d\omega_{xy} & ** \end{cases}$$

$$where \quad \begin{cases} * & a_i(t-T)\le a_i(t) \\ ** & a_i(t-T)> a_i(t) \end{cases} . \qquad (3)$$

Therefore the total distortion reduction (information gain) over all requested images is proportional to:

$$\Delta D \approx \sum_i p(R_i,t\,|\,O)\cdot \Delta D_{c,R_i} \; . \qquad (4)$$

To ensure video quality, the control strategy of our system is to maximize the distortion reduction ΔD by using proper cameras (i.e. the PTZ camera, the panoramic camera, or no-updating) to update the cached image. Denote *(X,Y,Z)*, corresponding to pan/tilt/zoom, as the best pose for the PTZ camera. *(X,Y,Z)* can be obtained with

$$(X,Y,Z) = \underset{(x,y,z)}{\arg\max}\left(\Delta D\right). \qquad (5)$$

When $\Delta D_{c,R_i}$ has a negative value, the system may choose not to update region R_i for increasing ΔD.

3.1. Estimating the distortion reduction between the real image and the cached image

Since the system cannot try all PTZ camera poses in practice, it has to seek the optimal camera pose via simulation before moving the PTZ camera. More specifically, the system has to try the distortion reduction equation with cutoff frequencies corresponding to various camera poses, and select the optimal camera pose.

During computer simulation, accurate estimation of equation (3) is difficult without sufficient camera resolution. To compensate this problem, we have to use image/video power spectrum models to assist the evaluation of information gains corresponding to various poses. According to Dong and Atick [2], if a system captures object movements from distance zero to infinity, $|\,F_{S,R_i}\,|^2$ and $|\,F_{M,R_i}\,|^2$ statistically fall with spatial frequency, ω_{xy}, according to $1/\omega_{xy}^{m}$ and $1/\omega_{xy}^{m-1}$ respectively, where *m* is around *2.3*.

Based on these simple models and various camera poses, the estimation of each distortion term may vary. Due to space limit, we only give the estimation procedure of the most general case. More specifically, we assume that only the panoramic video is available for the estimation. Let *b* be the spatial cutoff frequency of the panoramic video. Since the panoramic video is available for cache update at any time, we have $b \le a_i(t)$, and $b \le a_i(t-T)$. Let $E_{s,i,t}$ be the R_i-region AC-power between spatial frequency *1* and *b*, $E_{m,i,t}$ be the R_i-region frame-difference AC-power between spatial frequency *1* and *b*, $J_{m,i,t}$ be the R_i-region frame-difference power up to spatial frequency *b*, and $\hat{f}_b(x,y,t)$ acquired by the panoramic camera be the band-limited representation of $f(x,y,t)$. The estimation of $E_{s,i,t}$, $E_{m,i,t}$, and $J_{m,i,t}$ are very straightforward. For example,

$$J_{m,i,t} = \int_{R_i} \left| \hat{f}_b(x,y,t) - \hat{f}_b(x,y,t-1) \right|^2 dxdy\ . \tag{6}$$

$E_{s,i,t}$, $E_{m,i,t}$ can be estimated in a similar way. With these values, terms for $\Delta D_{c,R_i}$ may be obtained with:

$$\begin{aligned} \int_{R_i, a_i(t) \ge \omega_{xy} > a_i(t-T)} |F_{S,R_i,t}|^2\ d\omega_{xy} &= \frac{\left[a_i(t) - a_i(t-T)\right] \cdot b}{a_i(t) \cdot a_i(t-T) \cdot (b-1)} \cdot E_{s,i,t} \\ \int_{R_i, a_i(t-T) \ge \omega_{xy} > a_i(t)} |F_{S,R_i,t}|^2\ d\omega_{xy} &= \frac{\left[a_i(t-T) - a_i(t)\right] \cdot b}{a_i(t) \cdot a_i(t-T) \cdot (b-1)} \cdot E_{s,i,t}\ \ . \\ \int_{R_i, \omega_{xy} \le a_i(t-T)} |F_{M,R_i,t}|^2\ d\omega_{xy} &= J_{m,i,t} + \frac{1 - \left[b/a_i(t-T)\right]^{0.3}}{b^{0.3} - 1} \cdot E_{m,i,t} \end{aligned} \tag{7}$$

3.2. Weighting distortions according to users' requests

When multiple users request the cached image, the above distortion should be weighted according to users' requests. In this paper, users' requests to different portions of an image are modeled with a probability function $p_t(R_i \mid O)$. This gives rise to the form of a Bayes estimator. $p_t(R_i \mid O)$ may be estimated directly based on users' requests. Suppose there are n_i users requesting to view region R_i during the time period from *t* to *t+T* when the observation *O* is presented, and *p* and *O* do not change much during this short period, $p_t(R_i \mid O)$ may be estimated with

$$p_t(R_i \mid O) = \frac{n_i}{\sum_i n_i}\ . \tag{8}$$

When users' requests are not available, the estimation of $p_t(R_i \mid O)$ may become a problem. This problem may be tackled by using the system's past experience of users' requests. More specifically, if we assume that the probability of selecting a region does not depend on time *t*, the probability may be estimated with

$$p_t(R_i \mid O) = p(R_i \mid O) = \frac{p(O \mid R_i) \cdot p(R_i)}{p(O)}\ . \tag{9}$$

Signals received by various regions of an image generally come from various sources (i.e. objects), such as a presenter or an audience member. In a tele-conferencing environment, it is reasonable to assume that signals from different sources are independent. On the other hand, it is also reasonable to assume that a human's view selection separates various sources well into two categories (i.e. proper segmentation). Based on these assumptions, the feature vector *O* may be separated into independent feature vectors O_i and O_{other}, where O_i is the feature vector based on the data in R_i and O_{other} is the feature vector based on the data outside of R_i. Moreover, we may further assume that R_i and O_{other} are independent. With these assumptions, $p(R_i \mid O)$ may be estimated with

$$\begin{aligned} & p(R_i \mid O) = p(R_i \mid O_i, O_{other}) \\ & = \frac{p(O_i \mid R_i, O_{other}) \cdot p(R_i, O_{other})}{p(O_i, O_{other})} = \frac{p(O_i \mid R_i) \cdot p(R_i)}{p(O_i)}\ . \end{aligned} \tag{10}$$

The observation O_i may be further separated into "independent" features $O_i = \{o_1, o_2, \ldots, o_n\}$ as [6,7] suggested. With these independent features, $p(R_i \mid O)$ may be estimated with

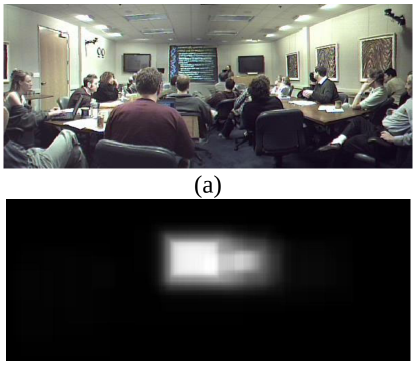
$$p(R_i \mid O) = \frac{p(o_1 \mid R_i) \cdot p(o_2 \mid R_i) \cdots p(o_n \mid R_i) \cdot p(R_i)}{p(o_1) \cdot p(o_2) \cdots p(o_n)}, \tag{11}$$

where $p(R_i)$ is the probability of selecting region R_i, and $p(o_j \mid R_i)$ is the probability of observing o_j in R_i when R_i is selected. Probabilities in this equation may be estimated online. With $p(R_i \mid O)$ available, it is straightforward to compute equation (5) for the optimal PTZ camera pose.

4. EXPERIMENTS

In this study, we deployed our system in a conference room, and grabbed one image per minute (*T=1*) with the panoramic camera during more than 10 short presentations. The total number of images is 120. In these 120 images, we picked 22 images with a uniform 5-minute period and asked 14 subjects to mark each image with regions that they want to watch in the close-up view. These data are then used to estimate $p(R_i \mid O)$ around these selected time instances. With these data, we can also estimated $p(R_i)$ as that shown in Figure 3, where whiter points correspond to higher $p(R_i)$ values.

Many image features may be considered as features from o_1 to o_n. Figure 4. shows the pdf ratio, $p(o_j \mid R_i)/p(o_j)$, variation when we select the frame difference as a feature. In this figure, the horizontal axis corresponds to the absolute value of frame difference, and the vertical axis corresponds to the pdf ratio. If the selected feature is closely related to users' view selections, the PDF function will reveal one or multiple peaks. Otherwise, the function will be flat.



(a)

(b)

Figure 3. Estimation of $p(R_i)$ (a) A typical panoramic shot that reveals the conference room arrangements. (b) Users' preferences to various regions $p(R_i)$.

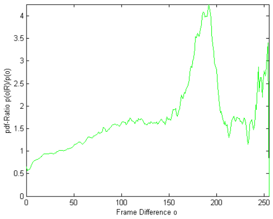
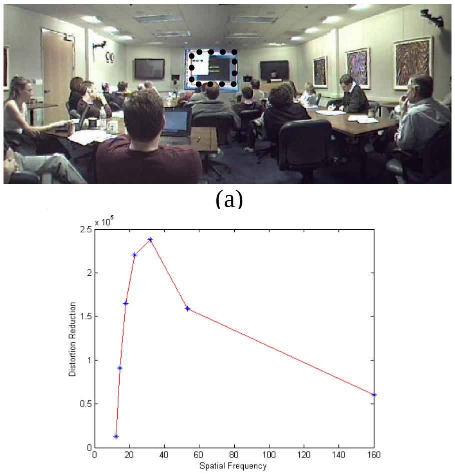


Figure 4. The pdf ratio variation when the absolute value of frame difference is chosen as a feature.



(a)

(b)

Figure 5. PTZ-camera pose selection (dotted black box) and the maximum distortion reductions corresponding to various zoom levels.

Due to space limit, we only demonstrate the PTZ camera control with a constraint experiment. According to given equations, the system can move the PTZ camera to form a very high-resolution image when the environment is static. Assume the cached image reached its highest resolution at time *t-T* and the panoramic image at *t* is shown in Figure 5 (a), the system will choose the dotted black box shown in Figure 5 (a) as the PTZ camera view to maximize the overall distortion reduction. Figure 5 (b) shows the maximum distortion reductions corresponding to various zoom levels (stars in the figure). Since we cut the image into small regions for fast optimization, the zoom level corresponds to a set of discrete values. The horizontal axis of Figure 5 (b) reflects the spatial frequency associated with various zoom levels, where the number tells us the number of PTZ camera sampling points corresponding to 12 pixels in the panoramic view.

5. CONCLUSIONS

By investigating a multi-resolution and multi-user camera management system within a video-distortion optimization framework, we are convinced that users' camera-control inputs are useful for teaching a computer system to take reasonable shots in a videoconferencing environment. We also show that the best PTZ camera pose can be found by using natural image/video statistical models proposed by some neural scientists.

6. REFERENCES

[1] M. Bianchi, "AutoAuditorium: a fully automatic, multi-camera system to televise auditorium presentations," *Proc. of Joint DARPA/NIST Smart Spaces Technology Workshop*, July 1998.

[2] D. Dong and J.J. Atick, "Statistics of Natural Time-Varying Images," *Network: Computation in Neural Systems*, vol 6(3), pp 345-358, 1995.

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

[4] 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.

[5] S. Mukhopadhyay and B. Smith, "Passive Capture and Structuring of Lectures," *Proc. of ACM Multimedia'99*, Orlando, 1999.

[6] A.J. Bell, T.J. Sejnowski, The "independent components" of natural scenes are edge filters, *Vis. Res.* 37(23): 3327-38, 1997.

[7] B.A. Olshausen and D.J. Field, "Sparse coding with an overcomplete basis set: a strategy employed by V1?," Vis. Res. 37: 3311-25.