AN ONLINE VIDEO COMPOSITION SYSTEM

Qiong Liu, Xiaojin Shi, Don Kimber, Frank Zhao, Frank Raab

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

ABSTRACT

This paper presents an information-driven online video composition system. The composition work handled by the system includes dynamically setting multiple pan/tilt/zoom (PTZ) cameras to proper poses and selecting the best close-up view for passive viewers. The main idea of the composition system is to maximize captured video information with limited cameras. Unlike video composition based on heuristic rules, our video composition is formulated as a process of minimizing distortions between ideal signals (i.e. signals with infinite spatial-temporal resolution) and displayed signals. The formulation is consistent with many well-known empirical approaches widely used in previous systems and may provide analytical explanations to those approaches. Moreover, it provides a novel approach for studying video composition tasks systematically. The composition system allows each user to select a personal close-up view. It manages PTZ cameras and a video switcher based on both signal characteristics and users' view selections. Additionally, it can automate the video composition process based on past users' view-selections when immediate selections are not available. We demonstrate the performance of this system with real meetings.

1. INTRODUCTION

Many existing video capture systems are professional-operator-controlled systems. To increase video capture flexibility and reduce labor cost, researchers proposed fully automatic video capture systems, including Bell Core's Auto-Auditorium [2], Cornell's lecture capturing system [9], Microsoft's ICAM system [8], and AT&T's Automated Cameramen [4]. However, these systems typically rely on state-of-the-art audio and vision techniques that may not be robust enough for real world use.

Our system overcomes problems of these systems by encouraging video viewers to compose video online and seamlessly merging manual composition and automatic composition. Similar to professional-operated systems, our system can be operated by human. Unlike professional-operated systems, our system hands the view selection task to regular viewers who are interested in the topic. Similar to a fully automatic system, our system can automatically compose video when no users want to control the system. Unlike a fully automatic system, our system allows convenient manual correction of imperfect automatic composition.

The system uses a hybrid camera, FlySPEC, [6,7] that combines the high resolution of a PTZ video camera with the wide field of view always available from a panoramic camera (Figure 1). By constructing a high-fidelity video canvas using video from the PTZ camera and the panoramic camera, our system enables each user to check details of a selected region using gestures over the canvas. Based on users' requests distributed on the canvas, we also design an algorithm for maximizing the overall video fidelity with one FlySPEC[7].

In this paper, we extended our approach for single FlySPEC control to online video composition using multiple FlySPEC cameras located at different view points and a video switcher. The online video composition system is named MSPEC which stands for multiple FlySPECs. Figure 1 shows the control interface of an MSPEC and a FlySPEC camera. In this interface, the three panoramic views come from panoramic cameras of three FlySPECs, and the close-up view comes from one selected FlySPEC camera. Similar to the single FlySPEC control, the MSPEC system also needs to move every PTZ camera to the right pose. Unlike the single FlySPEC control, the MSPEC can select the best output video stream from multiple FlySPEC streams. This design gives the system more chances to output better streams when one FlySPEC is not enough to handle the capturing task well.

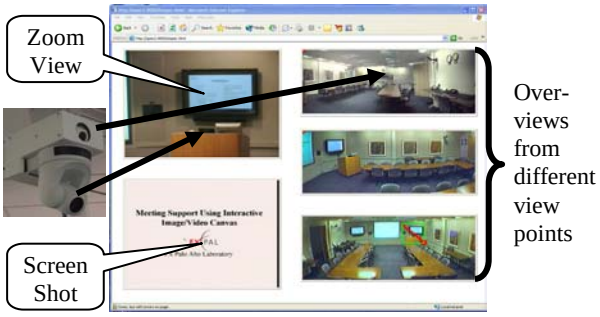


Figure 1. The control interface of a MSPEC and one of FlySPEC cameras.

With the MSPEC system, a user can compose a close-up view video stream by selecting a rectangular region in one of the panoramic views from time to time. When our video server receives the selection information, the server will send the user a close-up video stream according to the request. This close-up video may come from a PTZ camera, a panoramic camera, an image cache, or a mixture of all these sources depending on our video composition strategy and all users' requests. When users don't, won't, or can't provide their region selections, our algorithm can automate the best video stream selection based on users' past selection patterns. The automatic composition and users' manual composition are seamlessly integrated into the system to support a range of options from untended automatic to full manual composition. In the following paper, we present our

online video composition framework and some video composition experiments.

2. ONLINE VIDEO COMPOSITION

The goal of video composition is to acquire as much required information as possible for viewers with limited video channels. This goal can be formulated as a video quality maximization problem. In this formulation, we use the video reaching each FlySPEC camera as the best quality video which has infinite spatial and temporal resolution. Let $f(x,y,t)$ be the ideal video, where x and y are panoramic image canvas coordinates and t denotes time. Due to limited resolution of imaging sensors, a FlySPEC camera may only obtain an approximation $\hat{f}(x,y,t)$ of the ideal signal $f(x,y,t)$. Various regions of $\hat{f}(x,y,t)$ are transmitted to viewers according to their requests. To improve the video quality for each viewer, we have to improve $\hat{f}(x,y,t)$ estimation to reduce the difference between the displayed videos and the ideal video.

With the current MSPEC system, there are two ways to improve video quality for viewers. First, the system can change the PTZ camera pose to improve $\hat{f}(x,y,t)$ estimation. Second, the system can use a buffered high quality image, $\hat{f}_{t-T}$, to substitute for $\hat{f}(x,y,t)$ when some image regions do not change over a short time period T between consecutive video frames.

Denote $\{R_i\}$ as a set of non-overlapping small regions, N as the total number of requests, and $p(R_i,t \mid O)$ as the probability of viewing region-R_i details conditioned on environmental observation O at time t (e.g. the probability of viewing region-R_i when skin-color, body shape etc. appear in that region.) The total weighted distortion $D[\hat{f}_{t-T}, f_t]$ between users' requested images and the real image can be estimated with:

$$D[\hat{f}_{t-T}, f_t] \approx \sum_i N \cdot p(R_i,t \mid O) \cdot \int_{R_i} |\hat{f}(x,y,t-T) - f(x,y,t)|^2 \, dxdy, \quad (1)$$

Since all cameras have limited resolutions, $\hat{f}(x,y,t)$ is typically modeled as a band limited representation of $f(x,y,t)$ with cutoff frequency determined by the resolution of a camera. 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. The band limited model assumes $\hat{F}(\omega_{xy},t) = F(\omega_{xy},t)$ below certain spatial-frequency *a(t)* and $\hat{F}(\omega_{xy},t) = 0$ above the frequency. Let $F_{M,t}$ be $F(\omega_{xy},t) - F(\omega_{xy},t-T)$ and $F_{S,t}$ be $F(\omega_{xy},t) - \hat{F}(\omega_{xy},t)$, the above integration may be estimated with:

$$\int_{R_i} \left| \hat{f}_{t-T} - f_t \right| dxdy = \int_{R_i} |\hat{F}(\omega_{xy},t-T) - F(\omega_{xy},t)|^2 \, d\omega_{xy} = \int_{R_i,\omega_{xy} \le a_i(t-T)} |F_{M,R_i,t}|^2 \, d\omega_{xy} + \int_{R_i,\omega_{xy} > a_i(t-T)} |F_{S,R_i,t}|^2 \, d\omega_{xy} \quad (2)$$

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 missed because of the limited resolution of the cached image. By sampling region R_i at frequency $a_i(t)$ and updating the cached image, the expected distortion reduction is:

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

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

In our system, the sampling frequency of a region is directly related to the camera zoom level at that region. Therefore, the above distortion can be adjusted by changing the camera zoom level associated with region R_i. With equation 1-3, the total distortion reduction (information gain) over all requested images is proportional to:

$$\Delta D \approx \sum_i p(R_i,t \mid O) \cdot \Delta D_{R_i}. \quad (4)$$

To improve 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 *(P,T,Z)*, corresponding to pan/tilt/zoom, as the best pose for the PTZ camera. *(P,T,Z)* can be obtained with

$$(P,T,Z) = \arg\max_{(p,t,z)} (\Delta D), \quad (5)$$

where *(P,T)* decides the location of the updated regions and *Z* decides the sampling frequency of those updated regions.

With above control equations, the system can move each PTZ camera to form a very high-resolution image for future requests when the environment is static. In a dynamic environment, the algorithm will guide the PTZ camera to follow moving objects that interest most viewers.

Denote q as a PTZ camera id number, Q as the best PTZ camera for the zoom view in the above interface, and $\Delta DMAX_q$ as the maximum distortion reduction of camera q. Q can be obtained with

$$Q = \arg\max_q (\Delta DMAX_q). \quad (6)$$

Our system sends Q to the video switcher for selecting the best output video stream.

2.1 Estimating the Distortion Reduction from an Image Cache Update

Since the system cannot try all PTZ camera poses in practice, it has to seek the optimal camera pose via simulation before moving each PTZ camera. More specifically, the system has to try the distortion reduction equations (3) and (4) with sampling regions and cutoff frequencies corresponding to various camera poses, and select the optimal camera pose based on equation (5).

During computer simulation, accurate estimation of equation (3) is difficult without sufficient camera resolution. To compensate for this problem, we use Dong and Atick's image/video power spectrum models [3] to assist the evaluation

of dstortion reduction corresponding to various poses. According to these models, 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 available images, the estimation of each distortion term may vary. Due to space limit, we only give the estimation procedure of a typical case. More specifically, we assume that only the panoramic videos are available for the estimation. Let *b* be the spatial cutoff frequency of a panoramic video. Since the panoramic video is available for cache update at any time, *b* cannot be larger than the spatial cutoff frequencies of cached images. In other words, 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 a band-limited representation of $f(x,y,t)$. $J_{m,i,t}$ can be estimated with:

$$J_{m,i,t} = \int_{R_i} \left| \hat{f}_b(x,y,t) - \hat{f}_b(x,y,t-T) \right|^2 dxdy \cdot \tag{7}$$

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

$$\int_{R_i, a_i(t) \ge \omega_{xy} > a_i(t-T)} |F_{S,R_i,t}|^2 \, d\omega_{xy} = \frac{[a_i(t) - a_i(t-T)] \cdot b}{a_i(t) \cdot a_i(t-T) \cdot (b-1)} \cdot E_{s,i,t}$$

$$\int_{R_i, a_i(t-T) \ge \omega_{xy} > a_i(t)} |F_{S,R_i,t}|^2 \, d\omega_{xy} = \frac{[a_i(t-T) - a_i(t)] \cdot b}{a_i(t) \cdot a_i(t-T) \cdot (b-1)} \cdot E_{s,i,t} \tag{8}$$

$$\int_{R_i, \omega_{xy} \le a_i(t-T)} |F_{M,R_i,t}|^2 \, d\omega_{xy} = J_{m,i,t} + \frac{1 - [b/a_i(t-T)]^{0.3}}{b^{0.3} - 1} \cdot E_{m,i,t}$$

2.2 Weighting Distortions According to Users' Requests

To compute the distortion of all requests, 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. Assume *N* is the total number of requests and n_i users request the view of 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:

$$p_t(R_i \mid O) = \frac{n_i}{N}. \tag{9}$$

2.3 Automate Video Composition without Users' Requests

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 history 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

$$p_t(R_i \mid O) = p(R_i \mid O) = \frac{p(O \mid R_i) \cdot p(R_i)}{p(O)}. \tag{10}$$

In a tele-conferencing environment, it is reasonable to assume that signals from different sources (i.e. objects), such as a presenter or an audience member, are independent. 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 can further assume that R_i and O_{other} are independent. With these assumptions, $p(R_i \mid O)$ may be estimated with

$$\begin{aligned} p(R_i \mid O) &= p(R_i \mid O_i, O_{other}) \\ &= \frac{p(O_i \mid R_i, O_{other}) \cdot p(R_i, O_{other})}{p(O_i, O_{other})} \\ &= \frac{p(O_i \mid R_i) \cdot p(R_i)}{p(O_i)} \end{aligned}. \tag{11}$$

The observation O_i may be further separated into 'independent' features $O_i = \{o_1, o_2, \ldots, o_n\}$ as [1, 10] suggested. With these independent features, $p(R_i \mid O)$ may be estimated with

$$\begin{aligned} & p(R_i \mid O) \\ &= \frac{p(o_1 \mid R_i) \cdot p(o_2 \mid R_i) \cdots p(o_n \mid R_i) \cdot p(R_i)}{p(o_1) \cdot p(o_2) \cdots p(o_n)} \end{aligned}, \tag{12}$$

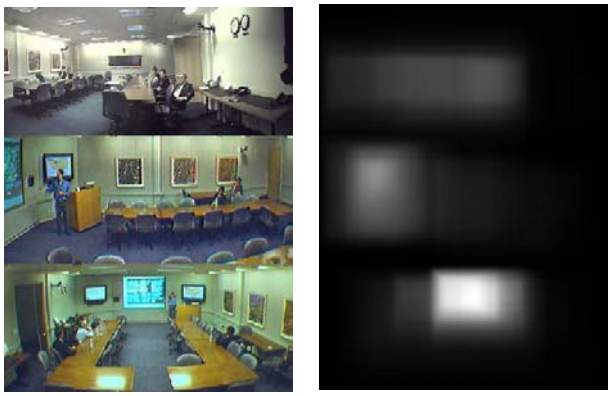
where $p(R_i)$ is the prior probability of selecting region R_i, and $p(o_j \mid R_i)$ is the probability of observing o_j in R_i when R_i is selected. Probabilities on the right side of this equation may be 'learned' online. With the $p(R_i \mid O)$ estimate available, it is straightforward to compute equation (5) for the optimal PTZ camera pose. This enables the system to automate video composition based on users' past selection patterns.

3. VIDEO COMPOSITION EXPERIMENTS

In our corporate conference room, we captured 56 meeting segments with three synchronized panoramic video cameras during 14 presentations. Then we asked 19 subjects to mark each meeting segment, which includes 3 synchronized video segments captured by different panoramic video cameras, with a rectangular region that s/he wants to watch in a close-up view. After getting inputs from these subjects, we used data corresponding to 30 meeting segments as training data to estimate $p(O \mid R_i)$ and $p(R_i)$. The estimate of $p(R_i)$ is shown in Figure 2, where whiter points correspond to higher $p(R_i)$ values.

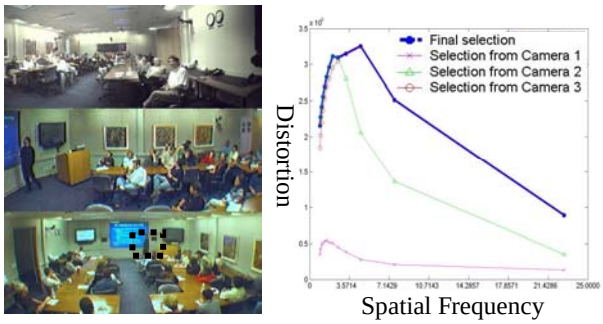
We tested our camera control algorithm with 26 other meeting segments. Figure 3 (a) shows a snapshot of a meeting segment using three panoramic views. If remote viewers can only watch one close-up stream and they do not send their requests for this meeting segment, our system will automatically choose the dotted black box shown in Figure 3 (a) as the PTZ

camera view. The optimal PTZ camera view selection is illustrated in Figure 3 (b), which shows the maximum distortion reductions corresponding to various PTZ cameras at various zoom levels. It also shows the maximum distortion reduction of using all three cameras. The horizontal axis of Figure 3 (b) reflects the spatial frequency associated with various zoom levels. The unit of this axis is based on the spatial frequency of a panoramic image. Since we cut the image canvas into small regions for fast optimization, the zoom level corresponds to a set of discrete values, and the best camera pose for that zoom level is computed. The optimal PTZ camera view, which is marked with the dotted black box in Figure 3 (a), corresponds to the highest distortion reduction point in Figure 3 (b).



(a) (b)

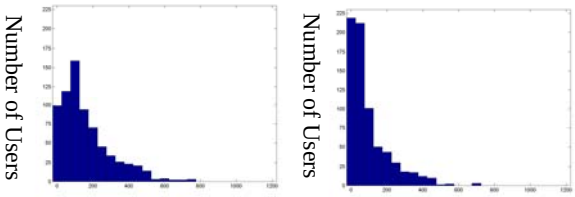
Figure 2. Estimation of $p(R_i)$ (a) A typical meeting shot that reveals the conference room arrangements. (b) Users' preferences to various regions $p(R_i)$.



(a) (b)

Figure 3. The maximum distortion reductions corresponding to various zoom levels and the best PTZ-camera pose selection (dotted black box).

Figure 4 shows the distortion statistics based on different PTZ camera control strategies. Figure 4 (a) reflects the visual distortion distribution when PTZ cameras are used for smallest field-of-view requests. Figure 4 (b) reflects the visual distortion distribution when PTZ cameras are controlled using our algorithm. Compared with Figure 4 (a), the peak shift in Figure 4 (b) reveals obvious user's view improvement come from using our control strategy. In this experiment, a system using our control strategy have 38% less distortion than a system using PTZ cameras for smallest field-of-view requests, and 51% less distortion than a system using no PTZ camera.



(a) (b)

Figure 4. The distortion statistics based on different PTZ camera usages. (a) Use PTZ cameras for smallest field-of-view requests. (b) Control using our algorithm.

4. CONCLUSIONS

We investigated the video composition problem within a signal distortion optimization framework. The composition strategy developed in this paper aligns well with many well-known composition rules. It also helped us to understand some problems overlooked by empirical approaches. Online video composition experiments based on our formulation further convinced us of the usefulness of this framework. Our experiments also challenged us with the problems of better probability estimation and user satisfaction evaluation.

5. REFERENCES

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

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

[4] 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.*

[5] C. Kyriakakis, P. Tsakalides, and T. Holman, "Acquisition and Rendering Methods for Immersive Audio," IEEE Signal Processing Magazine, pp. 55 – 66, January 1999.

[6] Q. Liu, D. Kimber, J. Foote, L. Wilcox, and J. Boreczky, "FLYSPEC: A Multi-User Video Camera System with Hybrid Human and Automatic Control," *Proceedings of ACM Multimedia 2002*, pp. 484 – 492, Juan-les-Pins, France.

[7] Q. Liu, and D. Kimber, 2003. LEARNING AUTOMATIC VIDEO CAPTURE FROM HUMAN'S CAMERA OPERATIONS. In *Proc. of IEEE International Conference on Image* Processing *2003*.

[8] 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.

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

[10] B.A. Olshausen, and D.J. Field, 1996. Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vis. Res. 37*: 3311-25.