MULTICHANNEL VIDEO/AUDIO ACQUISITION FOR IMMERSIVE CONFERENCING

Qiong Liu, Don Kimber, Jonathan Foote, Chunyuan Liao

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

ABSTRACT

This paper presents an information-driven audiovisual signal acquisition approach. This approach has several advantages: users are encouraged to assist in signal acquisition; available sensors are managed based on both signal characteristics and users' suggestions. The problem formulation is consistent with many well-known empirical approaches widely used in previous systems and may provide analytical explanations to these approaches. We demonstrate the use of this approach to pan/tilt/zoom (PTZ) camera management with field data.

1. INTRODUCTION

To reduce the signal acquisition cost for videoconferencing, researchers have developed various systems, for example, see [1,3,4,6,7]. Many of these use various heuristics to improve signal quality, yet a proper method to measure the quality of acquired audiovisual signals is still missing in the literature. This paper presents: a unified method for measuring the audiovisual signal quality, a signal capture device, and a device control strategy based on the proposed method. Guided by the proposed method, the paper also seeks analytical explanations of some widely used approaches.

In a meeting room environment, audiovisual signals may be viewed as space-time varying signals at any place. If perfect signals can be captured at a signal acquisition site and rendered at a remote site, a remote meeting participant may experience these signals just as if he/she were at the signal acquisition site. Ideally, the participant should be able to look around or freely turn her/his head to follow sound in a specific direction. Additionally, we also expect remote participants to share data acquisition devices to reduce system cost. As the number of sensors and available computational power restrict the signal acquisition channel, sensor management becomes a task of selecting proper content for the channel to reduce the distortion of acquired signals.

A hybrid system shown in Figure 1 can be used to capture audiovisual signals in time and space. It combines a wide-angle panoramic camera, a pan/tilt/zoom (PTZ) camera, and a microphone array. While the panoramic camera and a single microphone in the microphone array may provide overview signals of a conference room at low computational cost, the PTZ camera and the highly directional steerable microphone array with all available microphones may effectively enhance acquired signals in specific directions according to the guidance of the overview signal. With limited number of PTZ cameras and limited computational power for steering a microphone array to specific directions, our approach focuses on controlling PTZ cameras and highly directional audio beams to maximize users' information gain.

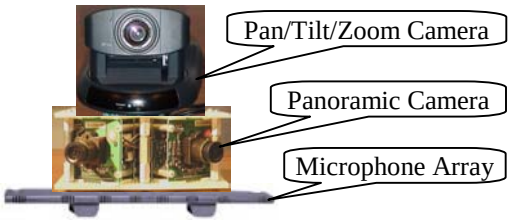


Figure 1. Signal acquisition sensor arrangement

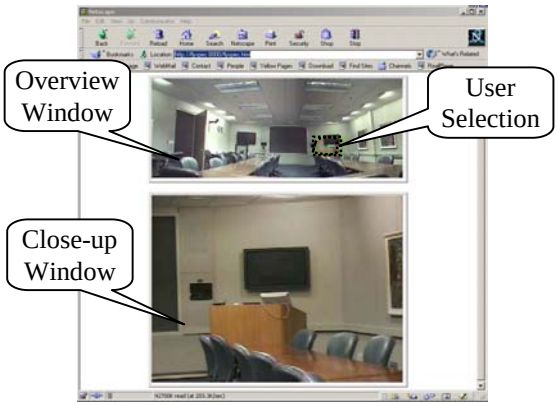


Figure 2. Web-based Graphical Interface for Remote Users to Control Their Virtual Cameras

For remote users to submit directional information of their interested events, our system provides an interface shown in Figure 2 to every remote user. With this interface, a remote user provides directional information by selecting a region in the overview window. When the system server gets parameters of a user selection, the server will send the user close-up video and directional enhanced audio corresponding to that region.

The remainder of this paper is structured as follows. In Section 2, we formulate our approach for audiovisual signal acquisition. In Section 3, we discuss the relations between our formulation and some well-known signal acquisition strategies. Experiments on video data acquisition are presented in Section 4. Concluding remarks are given in Section 5.

2. AUDIOVISUAL SIGNAL ACQUISITION AND SIGNAL QUALITY ESTIMATION

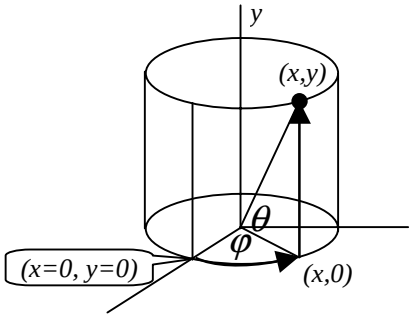


Figure 3. The coordinate system

A major goal of teleconferencing is to let remote participants experience a conference as they would in the real conference room. If a system could acquire a high quality signal at a specific location and present it signal to remote users, they may experience signal changes just as they would at the data acquisition site. To achieve high quality signal acquisition, we must both define the signal quality and move active sensors (such as the PTZ camera) to improve the signal quality averaged over all users.

Conceptually, the ideal signal received at a given point may be represented with $f(\varphi,\theta,t)$, where φ and θ are spatial angles used to identify the direction of a coming signal and t is the time. For derivations in later applications, we use a cylindrical coordinate system of Figure 3 to describe the signal. In Figure 3, a line passing the origin and a point on a cylindrical surface is used to define the signal direction. The point on the cylindrical surface has coordinates *(x,y)*, where *x* is the arc length between *(x=0, y=0)* and the point's projection on *y=0*, and *y* is the height of the point from the plane *y=0*. With this coordinate system, the ideal signal is represented with $f(x,y,t)$. In practice, a signal acquisition system may only get an approximation $\hat{f}(x,y,t)$ of the ideal signal $f(x,y,t)$ due to the limitation of sensors. Our sensor control strategy is to maximize the quality of the acquired signal $\hat{f}(x,y,t)$.

The information loss of representing f with $\hat{f}$ may be defined with

$$D[\hat{f}, f] = \sum_i p(R_i, t \mid O) \iiint_{R_i,T} \left| \hat{f}(x,y,t) - f(x,y,t) \right|^2 dxdydt, \quad (1)$$

where $\{R_i\}$ is a set of non-overlapping small regions, T is a short time period, $p(R_i, t \mid O)$ is the probability of viewing region-R_i details (conditioned on environmental observation O).

This probability may be obtained directly based on users' requests. Suppose there are $n_i(t)$ requests 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 period, then $p(R_i,t|O)$ may be estimated as

$$p(R_i, t \mid O) = \frac{n_i(t)}{\sum_i n_i(t)}. \quad (2)$$

$\iiint_{R_i,T} \left| \hat{f}(x,y,t) - f(x,y,t) \right|^2 dxdydt$ is easier to estimate in the frequency domain. If ω_x and ω_y represent spatial frequencies corresponding to x and y respectively, and ω_t is the temporal frequency, the distortion may be estimated with

$$\begin{aligned} &\iiint_{R_i,T} \left| \hat{f}(x,y,t) - f(x,y,t) \right|^2 dxdydt \\ &= \iiint_{R_i,T} \left| \hat{F}(\omega_x,\omega_y,\omega_t) - F(\omega_x,\omega_y,\omega_t) \right|^2 d\omega_x d\omega_y d\omega_t \end{aligned}. \quad (3)$$

The problem of acquiring a high quality signal is equivalent to reducing $D[\hat{f}, f]$. Assume $\hat{f}(x,y,t)$ is a band limited representation of $f(x,y,t)$. Reducing $D[\hat{f}, f]$ may be achieved by moving steerable sensors to adjust cutoff frequencies of $\hat{f}(x,y,t)$ in various regions $\{R_i\}$. Assume the region i of $\hat{f}(x,y,t)$ has spatial cutoff frequencies $a_{x,i}(t)$, $a_{y,i}(t)$, and temporal cutoff frequency $a_{t,i}(t)$. The estimation of $\iiint_{R_i,T} \left| \hat{f}(x,y,t) - f(x,y,t) \right|^2 dxdydt$ may then be simplified to

$$\begin{aligned} &\iiint_{R_i,T} \left| \hat{f}(x,y,t) - f(x,y,t) \right|^2 dxdydt \\ &= \iiint_{\substack{R_i,T \\ \omega_x > a_{x,i}(t) \\ \omega_y > a_{y,i}(t) \\ \omega_t > a_{t,i}(t)}} \left| F(\omega_x,\omega_y,\omega_t) \right|^2 d\omega_x d\omega_y d\omega_t \end{aligned}. \quad (4)$$

So the optimal sensor control strategy is to move high-resolution (i.e. in space and time) sensors to certain locations at certain time periods so that the overall distortion $D[\hat{f}, f]$ is minimized.

Previous equations described a way to compute the distortion when participants' requests were available. When participants' requests are not available, the estimation of $p(R_i, t \mid O)$ may become a problem. This may be overcome by using the system's past experience of users' requests. Specifically, if we assume that the probability of selecting a region does not depend on time *t*, the probability may be estimated as

$$p(R_i, t \mid O) = p(R_i \mid O) = \frac{p(O \mid R_i) \cdot p(R_i)}{p(O)} \quad (5)$$

We may consider O as an observation space of $\hat{f}$. By using a low dimensional observation space, it is easier to estimate $p(R_i, t \mid O)$ with limited data. With this probability estimation, the system may automate the signal acquisition process when remote users don't, won't, or cannot control the system.

3. RELATED WORK

This sensor management approach is consistent with many common empirical approaches for teleconferencing video/audio acquisition. The systems described in [1][3][6] all track moving objects based on image frame differences, and use high-resolution cameras to capture tracked objects. This camera control strategy corresponds to accumulating AC-components along ω_t for distortion measurement, and increasing a_x and a_y to reduce video distortion of the tracked object. This approach did not consider the distortion accumulation along the spatial domain. Therefore, even when a moving object has no detailed texture, this system may still move its high-resolution camera to shoot that object. For example, when a white screen is turned off to black, the system will move the high-resolution camera to that position even though the black screen has no useful content on it.

In Cornell's lecture capturing system [7], a video stream randomly alternating its source between two cameras helped the system to produce more engaging presentations. This strategy corresponds to balancing the temporal cutoff frequencies between two video channels. Since normal video spectrum has high magnitude at low temporal frequencies, balancing the temporal cutoff frequencies between two channels may statistically reduce the overall distortion and provide more information to remote users. With more targeted analysis of users' requests and video characteristics, our approach is expected to reduce this distortion further, and provide better information to remote users.

Different from the Cornell system, the Bell-Core 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. From our point of view, this camera management practice tried to balance spatial cutoff frequencies of two video channels for overall distortion reduction. In theory, this control strategy is only slightly different from the Cornell system. Therefore, its pros and cons are similar to the Cornell system. Customized rules, such as showing the slide when it first appears may further reduce video distortion statistically. It also fits well with our formulation.

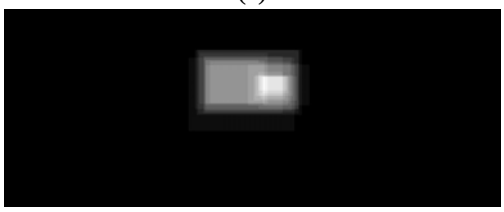
Systems described in [1][6] define regions, such as the stage area, to restrict interesting shot search. This strategy corresponds to the probability term based on users' selections. Since these regions were defined during system installations, they lack adaptability to environmental changes. Simple (e.g. all regions are defined with rectangles) and uniform (i.e. every pixel in the region is weighted equally) definition of these regions also makes it difficult to avoid shooting unwanted events.

For teleconferencing audio sensor management, researchers used beam-forming techniques to improve SNR of the audio signal when one or multiple speakers are presented [4]. From a different point of view, this strategy may be considered as improving spatial resolution of the audio signal. It also aligns well with our distortion reduction formulation.

In [5], we published our early work on video camera management based on single image frame contents. In this paper, our formulation is improved to handle time-varying signals. This improvement makes the formulation more generalizable to various other signal (e.g. audio) acquisitions.

4. EXPERIMENTS

(a)



(b)

Figure 4. Users' selections and corresponding probability estimation

The equations in section 2 can be directly used for active sensor management. We use a conference room PTZ camera control example to demonstrate our sensor management method. A panoramic camera (described in the introduction) was used to record 10 presentations in our corporate conference room and 14 users were asked to select interesting regions on a few uniformly distributed video frames, using the interface shown in Figure 2. Figure 4(a) shows a typical video frame and corresponding selections highlighted with boxes. Figure 4(b) shows the probability estimation based on these selections. In Figure 4(b), lighter color corresponds to higher probability value and darker color corresponds to lower value.

To compute the distortion defined with eq.(1), the system needs the result from eq.(4). Since it is impossible to get complete information of $F(\omega_x,\omega_y,\omega_t)$, the system needs proper mathematical models to estimate the result. According to Dong and Atick [2], if a system captures object movements from distance zero to infinity, $F(\omega_x,\omega_y,\omega_t)$ statistically falls with temporal frequency, ω_t, and rotational spatial frequency, ω_{xy}, according to

$$\left|F(\omega_{xy},\omega_t)\right|^2 = \frac{A}{\omega_{xy}^{1.3}\cdot\omega_t^2}, \qquad (6)$$

where *A* is a positive value related to the image energy.

Denote b_{xy} and b_t as the spatial and temporal cutoff frequencies of the panoramic camera, a_{xy} and a_t as the spatial and temporal cutoff frequencies of a PTZ camera. Let

$$E_{xyt} = \int_1^{b_t}\int_1^{b_{xy}} \left|F(\omega_{xy},\omega_t)\right|^2 d\omega_{xy} d\omega_t$$
$$E_{xy} = \int_1^{b_{xy}} \left|F(\omega_{xy},0)\right|^2 d\omega_{xy} \quad . \qquad (7)$$
$$E_t = \int_1^{b_t} \left|F(0,\omega_t)\right|^2 d\omega_t$$

If the system uses the PTZ camera instead of the panoramic camera to capture region R_i, the video distortion reduction achieved by this may be estimated with

$$D_{G,i} = \left[\frac{(a_{xy}^{0.3}-1)\cdot(a_t-1)\cdot b_{xy}^{0.3}\cdot b_t}{a_{xy}^{0.3}\cdot a_t\cdot(b_{xy}^{0.3}-1)(b_t-1)} - 1\right]\cdot E_{xyt,i} + \left[\frac{(a_{xy}^{1.3}-1)\cdot b_{xy}^{1.3}}{a_{xy}^{1.3}\cdot(b_{xy}^{1.3}-1)} - 1\right]\cdot E_{xy,i} + \left[\frac{(a_t-1)\cdot b_t}{a_t\cdot(b_t-1)} - 1\right]\cdot E_{t,i} \quad . \qquad (8)$$

Denote *(X,Y,Z)*, corresponding to pan/tilt/zoom, as the best pose of the PTZ camera. With eq.(1) and eq.(8), *(X,Y,Z)* can be estimated with

$$(X,Y,Z) = \arg\max_{(x,y,z)}[\,p(R_i,t\,|\,O)\cdot D_{G,i}\,]\,. \qquad (9)$$

In our experiment, the panoramic camera has 1200x480 resolution, and the PTZ camera has 640x480 resolution. Compared with the panoramic camera, the PTZ camera can achieve up to 10 times higher spatial sampling rate by performing optical zoom in practice. The camera frame rate varies over time depending on the number of users and the network traffic. We assume the frame rate of the panoramic camera is 1 frame/sec and the frame rate of the PTZ camera is 5 frames/sec. With the above optimization procedure and users' suggestions shown in Figure 4(a), the system selects the rectangular box in Figure 5(a) as the view of the PTZ camera.

(a)

(b)

Figure 5. PTZ camera view selection based on maximizing information gain

When users' selections are not available to the system, the system has to estimate the probability term (i.e. predicts users' selections) according to eq.(5). Due to the imperfection of the probability estimation, the distortion estimation without users' inputs is a little bit different from the distortion estimation with users' inputs. This estimation difference leads the system to a different PTZ camera view suggestion shown in Figure 5(b). By visually inspecting automatic selections over a long video sequence, we find that these automatic PTZ view selections are very close to those PTZ view selections estimated with users' suggestions.

5. CONCLUSIONS

We investigated the immersive conferencing video/audio acquisition problem within a signal distortion optimization framework. The sensor management strategy developed in this paper aligns very well with many well-known sensor management strategies. It also helped us to understand some problems overlooked by empirical approaches. Video acquisition experiments based on our formulation further convinced us of the usefulness of this framework. Our experimental results also challenged us with the problem of better probability estimation.

6. ACKNOWLEDGEMENTS

The authors would like to thank Dr. Juan Liu from PARC for her helpful comments.

7. 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, and S. Samarasekera, "Content based active video data acquisition via automated cameramen," *Proc. IEEE International Conference on Image Processing (ICIP) '98*.

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

[5] 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 2001*, pp. 484 – 492, Juan-les-Pins, France.

[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.

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