Semantic Saliency Driven Camera Control for Personal Remote Collaboration

Cha Zhang[†], Zicheng Liu[†], Zhengyou Zhang[†] and Qi Zhao[‡]

[†] *Communication and Collaboration Group, Microsoft Research*
One Microsoft Way, Redmond, WA 98052 USA
{chazhang,zliu,zhang}@microsoft.com

[‡] *Department of Computer Engineering, UC Santa Cruz*
1156 High Street, Santa Cruz, CA 95064 USA
zhaoqi@soe.ucsc.edu

***Abstract*—This paper presents a camera combo system for personal remote collaboration applications. The system consists of two different cameras. One camera has a wide field of view, and the other can pan/tilt/zoom (PTZ) based on analysis of the images captured by the wide angle camera. Unlike traditional approaches which usually drive the PTZ camera to follow the person or his/her head, our system is capable of capturing general objects of interest in remote collaboration. For instance, when the user raises something trying to show it to the remote person, our system will automatically position the PTZ camera to zoom in at the object. At the core of our system is a semantic saliency map that overcomes many limitations of low-level saliency maps computed from preliminary image features. We demonstrate how such a semantic saliency map can be computed through contextual analysis, sign analysis and transitional analysis, and how it can be used for PTZ camera control with a novel information loss optimization based virtual director. The effectiveness of the proposed method is demonstrated with real-world sequences.**

I. Introduction

As globalization continues to spread throughout the world economy, it is increasingly common to find product teams where team members reside in different time zones. Many companies are looking for video-conferencing solutions to improve collaborations between their remotely located team members. Driven by this demand and thanks to the rapidly improving network bandwidth and computer performances, video-conferencing has become increasingly popular. One of the most critical issues in immersive video-conferencing is video quality. Expensive high-definition video cameras are often used in modern telepresence systems [1], and webcams are widely used as the de-facto device in personal remote collaboration. Although some high-end cameras can produce very decent video images, there is an inherent tradeoff between resolution and field of view in standard static cameras. For instance, with a regular camera that has 60-70 degrees field of view pointing at a whiteboard 3-4 meters away, even at 2 megapixel resolution, it would not be possible for the remote meeting attendees to read the texts on the whiteboard. On the other hand, to provide the user as much flexibility as possible, a wide angle camera is necessary to keep the person in the view when he/she moves around.

Fig. 1. The camera combo hardware.

In this paper, we explore the usage of a pair of cameras, namely, a wide angle camera and a pan-tilt-zoom (PTZ) camera (as shown in Fig. 1), to provide high quality video for video-conferencing. The primary target of application is personal remote collaboration in offices, though most of the developed techniques can be applied in meeting rooms as well. The wide angle camera monitors the room, detects and tracks people in the room, and analyzes user activities in order to intelligently drive the PTZ camera and generate videos to be sent to the remote collaborator. The idea of such a camera combo has indeed been studied in many other applications, such as surveillance and monitoring [2], lecture recording [3], smart meeting rooms [4], etc. However, most of these projects were satisfied with an algorithm accurately tracking people in the field of view, and generating close-up shots of the persons that are being tracked. We argue that in personal remote collaboration, human is not the only subject of interest. For instance, one may hold an object or a paper document for the remote participant to have a look. If some diagrams are necessary to explain things, one may want to draw on the physical whiteboard behind him/her. Ideally an intelligent camera shall understand the user's attention and frame the PTZ camera accordingly.

For this purpose, we propose to compute a *semantic saliency* map based on the input wide angle video and use the semantic saliency map to control the PTZ camera (Note the same technique can be used to control a high resolution camera for digital pan, tilt and zoom). In contrast to the low level saliency computation algorithms that has attracted a lot of attention recently [5], [6], a semantic saliency map integrates knowledge

from high-level semantic analysis, thus is more suitable for high-level camera control in many applications. Our semantic saliency map is computed based on three analysis components: contextual analysis, sign analysis and transitional analysis. We demonstrate the effectiveness of the proposed method with various scenarios during a personal remote collaboration session.

The second contribution of this paper is a minimum information loss framework for PTZ camera control or virtual director. The field of view of the PTZ camera can be considered as a cropping window from the wide angle camera. The goal of the virtual director is to find the optimal location and scale of the cropping window. Traditionally this is fulfilled by defining a set of ad hoc rules. In this paper, we observe that both cropping and scaling may lead to information loss. Such loss cannot be reduced simultaneously if the output video resolution is fixed. For instance, increasing the cropping window size will include more salient regions but cause more resolution loss due to scaling. We propose a novel framework for virtual director by minimizing a cost function that seeks the best tradeoff between these two information loss factors.

The rest of the paper is organized as follows. An overview of our system is presented in Section II. In Section III we describe a few basic techniques for semantic analysis. The semantic saliency map is introduced and computed in Section IV. The virtual director that controls the camera based on the semantic saliency map is given in Section V. Experimental results and conclusions are presented in Section VI and VII, respectively.

II. System Overview

A. The Camera Combo

We construct the camera combo by using two Axis network cameras, as shown in Fig. 1. The fisheye camera is an Axis 212 PTZ network camera, though we zoom out to the maximum and use it as a fixed zoom fisheye camera. The field of view of this wide angle camera is around 140 degree, which is sufficient to cover a typical office or meeting room environment. The PTZ camera is an Axis 213 PTZ network camera, which has built-in 26$\times$ optical zoom and auto focus. Both cameras are operated with 640$\times$480 pixels resolution at 25 frames per second (fps).

The cameras are mounted on a custom-made base. The camera centers are roughly aligned at the same height. The distance between the two cameras is around 15cm, which is practically negligible when the observed object is a few meters away from the camera combo.

Both cameras need to be calibrated in order to compute corresponding pan/tilt and zoom level of the PTZ camera from regions specified in the fisheye camera. This task is non-trivial because the PTZ camera will be constantly moving during the application. We adopted the two step procedure proposed by Sinha and Pollefey [7] to calibrate the PTZ camera. In the first step, the intrinsic parameters of the PTZ camera are determined by capturing a set of images for a static scene at different pan/tilt angles at the camera's lowest zoom setting, following an algorithm originally proposed by Hartley [8]. In the second step, we fix the pan and tilt of the camera, and monotonically increase the zoom level in order to compute the intrinsic parameters across discrete steps of zoom levels. Intrinsic parameters at arbitrary zoom levels are then interpolated from the discrete instances.

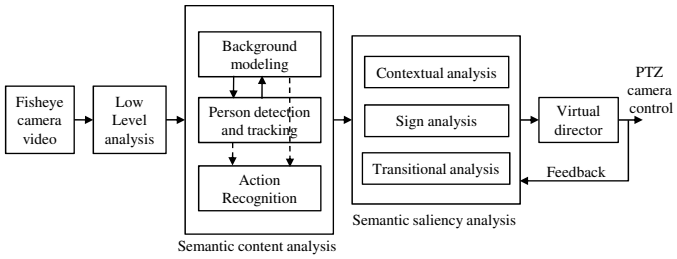

Fig. 2. The system diagram of the proposed approach.

Fisheye lens differ from an ordinary rectilinear lens in that the projection from a 3D ray to 2D image position in the fisheye lens is non-perspective. Dedicated calibration schemes are thus necessary for fisheye lens [9]. In our work, we found a simple equiangular model [10] widely used in computer graphics rendering is sufficient for computing corresponding pan/tilt angle given a target pixel location in the fisheye image.

B. System Diagram

As mentioned in the introduction, we propose to compute a semantic saliency map in order to determine where the user's attention should be for tele-collaboration. Fig. 2 shows the basic diagram of our proposed system. Given the input video from the wide-angle camera, we first extract simple low level features such as color, texture, motion, etc. These features are then fed to the semantic analysis module to perform high-level video analysis, such as person detection/tracking, background modeling and action recognition. These high-level analysis results are then used to compute the semantic saliency map through various components such as contextual analysis, sign analysis, transitional analysis, etc. A virtual director will determine where to point the PTZ camera given the semantic saliency map. Note there is also feedback from the PTZ's camera control to the semantic saliency analysis, which will be detailed later.

Compared with the various saliency extraction approaches in the literature [5], [6], a significant difference of our approach is the semantic analysis layer embedded between low level analysis and saliency analysis. For different applications, the components in this middle layer may differ, but they all provide critical information to the saliency analysis module to make the saliency map more meaningful. To give a concrete example, assume during conferencing the user stands up and walks to the whiteboard to write something. With the traditional low-level saliency computation methods, both the person and the chair are moving, and both have high saliency. However, if high-level semantic information is provided, the system may easily distinguish the motion between the user and the chair, which can accordingly reduce the saliency score of the chair region.

Our saliency computation scheme also differs significantly from traditional approaches. In particular, we fuse the results from three types of analysis, namely, contextual analysis, sign analysis and transitional analysis. Contextual analysis computes the saliency map in the context of the application being concerned. For instance, in personal tele-collaboration, the upper body region usually has a much higher priority to be shown on the remote side such that the remote user may see the facial expression and gesture. Sign analysis impact the saliency map by recognizing special activity signs conducted by the user. Usually that requires an agreement made between the user and the system beforehand. Transitional analysis studies the impact of focus transition had on the saliency map itself. It takes the feedback from the camera control, which may impact the saliency map. For instance, if the PTZ camera has been given a close shot of the person for a long time, the saliency value around the shot region may reduce gradually to encourage the display of other interesting regions.

The virtual director component in Fig. 2 differs from traditional rule-based virtual directors such as those in [3]. We propose a novel optimization based virtual director that intends to minimize the joint information loss caused by the pan, tilt and zoom of the PTZ camera. The tradeoff is between zooming into the scene for more scene details, and zooming out for covering a larger field of view. With the help of the various semantic saliency analysis methods in the previous stage, we show such an optimization scheme can produce as good as, if not better than, rule based camera controls.

III. Semantic Content Analysis

The semantic saliency map is based on results from a few semantic content analysis modules operated on the wide-angle image, such as background modeling, person detection and tracking, action recognition, etc.

A. Background Modeling

Since in an office or meeting room environment most of the background objects are static, we construct a background model given the video sequence from the fisheye camera. There have been many background modeling schemes proposed in the literature, such as those based on Gaussian distributions [11], mixture of Gaussians [12], non-parametric kernel density estimators [13], etc. We implemented an algorithm based on per-pixel Gaussian distribution modeling, with "high-level" guidance from the person detector and tracker that will be briefly described in the next subsection, as was suggested in [14]. That is, regions identified by the detector and tracker will not be considered during background model updating, even if they have been static for a long time.

Fig. 3 shows some results of our background modeling algorithm. In the top row, the person has been walking around and just sat down. This is a relatively simple scenario and we obtained a full mask of the human body. In the bottom row, the person has been sitting there without motion for a long time. Due to the lack of motion, the body gradually merged with the background model. However, the head region is still very clear in the foreground mask because it is not used for background updating according to the "high-level" guidance mechanism.

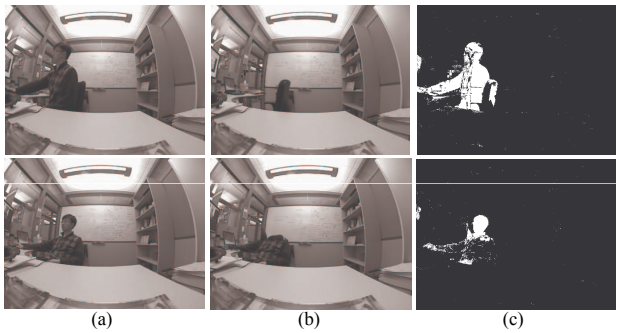

Fig. 3. Background modeling result. (a) Current video frame; (b) current background image; (c) foreground mask.

B. Person Detection and Tracking

Person detection is one of the most important components for video analysis. It has been very widely studied in literature [15]. Consequently we will only briefly describe the techniques used in our system.

We employ the face detector developed in [16] for detecting frontal faces in the environment. Afterwards, the face is tracked via a modified kernel based tracking algorithm that integrates with the result from background modeling. In the original kernel based tracking method proposed in [17], the goal was to search for a nearby region which has the smallest distance between the target histogram $\mathbf{q}(u)$ and the region's histogram $\mathbf{p}(u)$, where u is the color bin index. Both histograms were computed by imposing a kernel with Epanechnikov profile. In our modified tracking algorithm, we enforce a mask derived by the background modeling procedure on the histogram computation. That is, only pixels that are classified as foreground pixel are used for computing the histogram. We found that such a masking scheme works very well in preventing tracked objects from being attracted by background regions with similar colors.

C. Action Recognition

Given the output from person tracking and background substraction, it is not difficult to design simple action recognition algorithms to detect usual actions. In the context of our particular application for personal tele-collaboration, we built recognition modules for detecting two simple actions, hand waving and finger pointing based on skin color classification over the foreground mask provided by the background modeling module. Such an approach is similar to the previous work in [18] and has been shown to work well in practice. The main difference is that instead of using a generic skin color model [18], we construct the model from the tracked face region, which is more accurate.

Fig. 4 shows two example scenes where both the face and hands are detected and tracked. For action recognition, more sophisticated approaches could certainly be applied in

Fig. 4. Results of face and hand tracking.

our framework, such as the method based on 3D volumetric features [19].

IV. Semantic Saliency Analysis

The results of semantic content analysis are used to compute where the user's visual attention should be, namely, the saliency map. We formulate the saliency analysis problem as follows. Assume at any instance t, the content analysis module provides information as $\Omega_t = \{M_t(\mathbf{x}), R_t, \mathcal{A}_t, \cdots\}$, where $\mathbf{x}$ is the pixel index, $M_t(\mathbf{x})$ is the foreground mask as shown in Fig. 3 (c), $R_t = \{\mathbf{x}_0^t, \mathbf{x}_1^t\}$ is a rectangle region that represents the tracked person's head ($\mathbf{x}_0$ is the top left corner and $\mathbf{x}_1$ is the bottom right corner), $\mathcal{A}_t$ is the recognized human action. Note if more content analysis modules are available, they can all be integrated into Ω_t. A saliency map, defined as $S_t(\mathbf{x})$, can be computed based on the history of Ω_t, namely:

$$S_t(\mathbf{x}) = \Psi(\Omega_t, \Omega_{t-1}, \cdots, \Omega_{t-N}), \tag{1}$$

where N is the length of history.

In the simplest form, the foreground mask $M_t(\mathbf{x})$, or a certain form of motion segmentation results, can be directly used as the saliency map [20], [21]. That is:

$$S_t(\mathbf{x}) = \Psi(\Omega_t) = M_t(\mathbf{x}). \tag{2}$$

Such a simple saliency map contains some semantic information and may work well for certain applications. However, it does not work well for camera control during remote collaboration because it places equal emphasis on upper and lower bodies, any moving objects such as a chair, etc. Furthermore, it cannot respond to human actions such as body gestures that may intend to guide/control the camera attention. The three semantic analysis methods below intend to overcome these shortcomings. Nevertheless, the foreground mask $M_t(\mathbf{x})$ will serve as the base map that shall be enhanced by the following analysis modules.

A. Contextual Analysis

Contextual analysis refers to enhancements made to the base saliency map based on the context of the particular application being studied. Take personal tele-conferencing as an example. It is a common knowledge that in tele-conferencing, the face and upper body of the user is much more important than his/her lower body, or other moving objects such as a chair. As a result, a soft-masking operation may be imposed on the base saliency map to emphasize the head and shoulder region. Mathematically, we compute a contextual score as:

$$s_t^C(\mathbf{x}) = \exp\{-\frac{1}{2}(\mathbf{x} - \mathbf{x}_R)^T \Sigma_R^{-1}(\mathbf{x} - \mathbf{x}_R)\} \tag{3}$$

where $\mathbf{x}_R$ is the mask center determined by the current head tracking region R_t, Σ_R is the covariance matrix of the soft contextual mask. The semantic saliency map can be computed as:

$$S_t^C(\mathbf{x}) = s_t^C(\mathbf{x}) M_t(\mathbf{x}). \tag{4}$$

When more than one persons are in the room, we may soft mask each person as above. Alternatively, if a speaker detection algorithm such as [22] is available in the previous stage, the contextual mask can be place on the speaker's head and shoulder region only, while other people's saliency regions will all be attenuated.

B. Sign Analysis

When people communicate with each other, they use speech, expression and gestures extensively to deliver their messages. In this paper, we generally call them *signs*. Signs are audio/visual signals that are agreed between the users or the users and the computers in order to communicate their intension or status. Consequently, signs will have a strong impact on the visual attention. For instance, if the user points to a few equations on the whiteboard, it shall be the content on the whiteboard that receives the full attention. From camera control point of view, the computer should be able to recognize signs made by the user and move the PTZ camera to focus on the user's intended regions of interest.

Unfortunately, low level saliency analysis will not be able to recognize such semantic intentions. We rely on the action recognition module in the previous stage to perform sign-based saliency analysis. In our application, we assume that when the user wave his/her hand, or use his/her hand to point to a certain region, that region nearby the hand will be the focus of attention. That is, when $\mathcal{A}_t$ is hand-waving or finger pointing, define a sign score as:

$$s_t^S(\mathbf{x}) = \exp\{-\frac{1}{2}(\mathbf{x} - \mathbf{x}_A)^T \Sigma_A^{-1}(\mathbf{x} - \mathbf{x}_A)\}, \tag{5}$$

where $\mathbf{x}_A$ is the hand region center given by the action recognition module, Σ_A is the covariance matrix of the soft contextual mask. The semantic saliency map can be computed as:

$$S_t^S(\mathbf{x}) = s_t^S(\mathbf{x}) + M_t(\mathbf{x}). \tag{6}$$

Note we use summation instead of multiplication in order to raise the saliency values around the hand region.

C. Transitional Analysis

Transitional analysis studies the transition or change in saliency regions as time pass by. For example, if a person has been paying exclusive attention to a particular object for a long time, it is very likely that his/her attention will be distracted to some nearby interesting object. In the study conducted in [3], it has been shown that professional videographers

often add some randomness to the scene switching to improve aestheticity. Such distraction from the most salient object can be modeled by a saliency fading procedure as below.

Let the PTZ focused region in the past N time instances be $\{F_{t-1}, \cdots, F_{t-N}\}$. Given any pixel $\mathbf{x}$, compute its past attention score as:

$$s_t^T(\mathbf{x}) = \exp\{-\sum_{\tau=t-N}^{t-1} \frac{\delta_{F_\tau}(\mathbf{x})}{A_{F_\tau}} \alpha^{\tau-t}\}, \tag{7}$$

where $\delta_{F_\tau}(\mathbf{x})$ is an index function which takes value 1 if $\mathbf{x}$ is inside region F_τ and 0 otherwise. A_{F_τ} is the area of region F_τ. α, larger than 1, is a parameter controlling the fading speed of saliency. It can be seen that if a pixel has never been observed by the PTZ camera, its score is 1. Otherwise, the score is smaller than 1 but greater than 0.

The semantic saliency map after transitional analysis is:

$$S_t^T(\mathbf{x}) = s_t^T(\mathbf{x}) M_t(\mathbf{x}), \tag{8}$$

The transitional analysis is particularly useful if there are multiple salient objects in the scene. The fading procedure provides a natural mechanism to allow the virtual director to switch between multiple salient objects.

Overall, if all the three analysis components are available, we obtain the final semantic saliency map as:

$$\begin{aligned} S_t(\mathbf{x}) &= \Psi(\Omega_t, \Omega_{t-1}, \cdots, \Omega_{t-N}) \\ &= [s_t^C(\mathbf{x}) M_t(\mathbf{x}) + s_t^S(\mathbf{x})] s_t^T(\mathbf{x}). \end{aligned} \tag{9}$$

Fig. 5 shows the procedure of computing the semantic saliency map for a typical scene. Note the intensity of the semantic maps in Fig. 5(c)(d)(e) are re-scaled to make them visible.

V. Virtual Director

Once the saliency map has been computed, it is the virtual director's responsibility to determine where to focus the PTZ camera. In the field of view of the wide-angle camera, it is equivalent to finding a cropping rectangular region for the PTZ camera to show. During this process, however, tradeoff has to be made. If the focused region is too large, the PTZ camera will be able to see most of the salient object, resulting in a small spatial information loss due to cropping. However, since the resolution of the PTZ video is limited, there will be resolution information loss because the camera cannot zoom in too much to reveal details of the object. On the other hand, if the cropping region is small and the PTZ camera zooms in closely to show the details, there will be spatial information loss due to the cropping of the field of view of the camera.

Our formulation takes both information loss into consideration and seeks for a trade-off between the two. In particular, we represent the video information loss function with two terms, i.e.,

$$\mathbf{L}(\mathcal{V}, \hat{\mathcal{V}}) = \mathbf{L}_s(\mathcal{V}, \hat{\mathcal{V}}) + \lambda \mathbf{L}_r(\mathcal{V}, \hat{\mathcal{V}}), \tag{10}$$

where $\mathbf{L}(\mathcal{V}, \hat{\mathcal{V}})$ is the information loss function between the observed world and the video captured by the PTZ camera. $\mathbf{L}_s(\mathcal{V}, \hat{\mathcal{V}})$ is the information loss due to the limited field of view (cropping), and $\mathbf{L}_r(\mathcal{V}, \hat{\mathcal{V}})$ is the information loss due to resolution. λ is a weighting parameter that balances the two types of information loss.

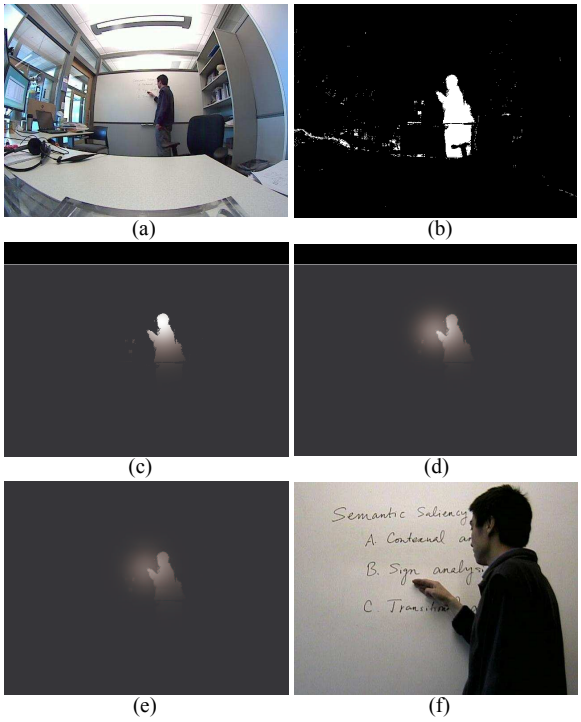

Fig. 5. Computation of the semantic saliency map. (a) The image captured by the wide angle camera. (b) The foreground mask or base map $M_t(\mathbf{x})$ obtained from background modeling and person detection/tracking. (c) Apply contextual analysis to the base map $s_t^C(\mathbf{x})M_t(\mathbf{x})$. (d) Apply sign analysis to the previous result $s_t^C(\mathbf{x})M_t(\mathbf{x}) + s_t^S(\mathbf{x})$. (e) Apply transitional analysis and obtain the final semantic saliency map $S_t(\mathbf{x})$. (f) The image captured by the PTZ camera after camera control.

We model the spatial information loss by the total amount of saliency falling outside the field of view of the PTZ camera, computed on the image observed by the wide-angle camera. Given the semantic saliency map as $S_t(\mathbf{x})$, we first normalize it such that the summation over the whole image is 1. That is,

$$\sum_{\mathbf{x}} S_t(\mathbf{x}) = 1. \tag{11}$$

Let the field of view of the PTZ camera correspond to a rectangular region $\mathcal{W}_t$, the spatial information loss is defined as:

$$\mathbf{L}_s(\mathcal{W}_t) = \sum_{\mathbf{x} \notin \mathcal{W}_t} S_t(\mathbf{x}) = 1 - \sum_{\mathbf{x} \in \mathcal{W}_t} S_t(\mathbf{x}) \tag{12}$$

The resolution information loss is computed based on the energy difference between the PTZ camera view and its down-sampled version which has the same resolution as the image of the wide angle view. Let the resolution information loss for a particular cropping window $\mathcal{W}_t$ be modeled as:

$$\mathbf{L}_r(\mathcal{W}_t) = \sum_{\mathbf{x} \in \mathcal{W}_t} l_r(\mathbf{x}, scale(\mathcal{W}_t))., \tag{13}$$

where $l_r(\mathbf{x}, scale(\mathcal{W}_t))$ is a loss value computed at pixel location $\mathbf{x}$ for the scale change from cropping window $\mathcal{W}_t$ to the resolution of the PTZ camera.

The loss function $l_r(\mathbf{x}, scale(\mathcal{W}_t))$ has to be estimated in order to perform the minimization of the combined information loss function in Eq. (10). We adopt a data driven approach to solve this problem. Before the camera control process starts, a number of cropping windows are selected for the PTZ camera to zoom in (at different zoom levels) and capture the corresponding images. For each image $I(\mathbf{x})$ captured by the PTZ camera, we perform a low-pass filter and obtain a smoothed image $I_l(\mathbf{x})$. The low-pass filter is to assure that when $I_l(\mathbf{x})$ is down-sampled to $I_l^d(\mathbf{x})$, which has the same resolution as the wide angle camera, there is no over-smoothing or aliasing. The image $I_l^d(\mathbf{x})$ is then divided into small patches $p_j, j = 1, \cdots$ with size 5×5 pixels. The energy of each patch's edge map is computed and quantized into 8 bins. For each patch and its corresponding region P_j in $I(\mathbf{x})$ or $I_l(\mathbf{x})$, a loss value can be computed as:

$$l_j = \sum_{\mathbf{x} \in P_j} |I(\mathbf{x}) - I_l(\mathbf{x})|^2. \tag{14}$$

For each scale of zooming, the loss values are averaged across all the corresponding PTZ images for each bin of patch edge energy. The end result is a look-up table which provides a loss value for each patch edge energy bin and each scale of zooming. Such a look-up table is used to compute $l_r(\mathbf{x}, scale(\mathcal{W}_t))$ in Eq. (13). That is, for each pixel $\mathbf{x}$ in the wide angle image, we take its 5×5 pixels neighborhood and compute its edge energy. The corresponding loss value is then obtained through a simple table look-up.

Given the overall cost function of Eq. (10), an exhaustive search scheme is used to find the cropping region inside the wide-angle view that minimize the combined information loss in order for the PTZ camera to move to. Such an exhaustive search is affordable with the integral image approach, populated by the face detector developed in [23].

In practice we found that the cropping windows obtained by the above algorithm tend to be too tight around the foreground object. To increase the aesthetics of the scene, we expand the computed cropping window by a certain predetermined percentage, e.g., 25% in both width and height.

VI. Experimental Results

We have built the camera combo system for camera control. The performance of our system is best shown with a short video demonstrating its usage during a personal communication session. Such a video is available at:

http://research.microsoft.com/∼chazhang/mmsp08video.wmv

VII. Conclusions and Future Work

We presented a framework for camera control based on the computation of semantic saliency map and an information loss optimized virtual director. Compared with low level saliency maps, the proposed semantic saliency map can better describe the user's attention through contextual analysis, sign analysis and transitional analysis. The information loss optimization framework is a more principled approach for camera control compared with traditional ad hoc approaches.

One future work is to improve the semantic saliency map and virtual director so that it can drive the camera based on predictions of motion in the future. This is often done by professional videographers, which can greatly improve the stableness of the PTZ camera.

References

[1] A. W. Davis and I. M. Weinstein, *Telepresence 2007 – Taking Videoconferencing to the Next Frontier*. Wainhouse Research Segment Report, 2007.

[2] A. Hampapur, S. Pankanti, A. Senior, Y.-L. Tian, L. Brown, and R. Bolle, "Face cataloger: Multi-scale imaging for relating identity to location," in *IEEE conference on Advanced Video and Signal Based Surveillance*, 2003.

[3] Y. Rui, A. Gupta, J. Grudin, and L. He, "Automating lecture capture and broadcast: technology and videography," *ACM Multimedia Systems Journal*, vol. 10, no. 1, pp. 3–15, 2004.

[4] Q. Liu, D. Kimber, J. Foote, L. Wilcox, and J. Boreczky, "FLYSPEC: A multi-user video camera system with hybrid human and automatic control," *Proc. ACM Multimedia*, 2002.

[5] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11.

[6] T. Liu, J. Sun, N.-N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2007.

[7] S. N. Sinha and M. Pollefeys, "Towards calibrating a pan-tilt-zoom camera network," in *OMNIVIS 2004, workshop on Omnidirectional Vision and Camera Networks held in conjunction with ECCV*, 2004.

[8] R. I. Hartley, "Self-calibration of stationary cameras," *International Jounral of Computer Vision*, vol. 22, no. 1, pp. 5–23, Feb/March 1997.

[9] Y. Xiong and K. Turkowski, "Creating image-based vr using a self-calibrating fisheye lens," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 1997.

[10] R. Kingslake, *A History of the Photographic Lens*. Academic Press, 1989.

[11] C. Wren, A. Azarbayejani, T. Darrel, and A. Pentland, "Pfinder: Real time tracking of the human body," *In IEEE Trans. on Pattern Analysis and Machine Intelligence*, 1997.

[12] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, 1999, pp. 246–252.

[13] A. Elgammal, D. Harwood, and L. S. Davis, "Non-parametric model for background subtraction," in *Proc. European Conf. on Computer Vision (ECCV)*, Dublin, Ireland, June 2000.

[14] M. Harville, "A framework for high-level feedback to adaptive, per-pixel, mixture-of-gaussian background models," in *Proc. European Conf. on Computer Vision (ECCV)*, 2002.

[15] E. Hjelmas and B. K. Low, "Face detection: A survey," *Computer Vision and Image Understanding*, vol. 83, pp. 236–274, 2001.

[16] C. Zhang and P. Viola, "Multiple-instance pruning for learning efficient cascade detectors," in *Neural Information Processing Systems (NIPS)*, 2007.

[17] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 25, no. 5, pp. 564–577, May 2003.

[18] F. Wang, C.-W. Ngo, and T.-C. Pong, "Gesture tracking and recognition for lecture video editing," in *Proc. of Int. Conf. on Pattern Recognition (ICPR)*, 2004.

[19] Y. Ke, R. Sukthankar, and M. Hebert, "Efficient visual event detection using volumetric features," in *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, 2005.

[20] R. P. Wildes, "A measure of motion salience for surveillance applications," in *Proc. Of IEEE Int. Conf. on Image Processing (ICIP)*, 1998.

[21] Y.-L. Tian and A. Hampapur, "Robust salient motion detection with complex background for real-time video surveillance," in *IEEE Computer Society Workshop on Motion and Video Computing*, 2005.

[22] C. Zhang, P. Yin, Y. Rui, R. Cutler, and P. Viola, "Boosting-based multimodal speaker detection for distributed meetings," in *IEEE Int. Workshop on Multimedia Signal Processing (MMSP)*, 2006.

[23] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. of CVPR*, 2001.