# Turning Your Video Monitor into a Virtual Window

Jeremy R. Cooperstock
Department of Electrical & Computer Engineering
University of Toronto
Toronto, Ontario M5S 1A4
jer@dgp.toronto.edu

Koichiro Tanikoshi*, William Buxton†
Computer Systems Research Institute
University of Toronto
Toronto, Ontario M5S 1A4
(tanikosi | buxton)@dgp.toronto.edu

**Abstract - A video conference system that allows a video attendee to look around an entire conference room simply by moving his or her head is described. In order to locate the attendee's head, a differential image is produced by removing a reference view from the current video image. The orientation of a motorized camera is then determined directly by the head position of the video attendee. The increased affordances of this mechanism are discussed.**

## I. INTRODUCTION

In conventional video-conference settings, video attendees often feel a lack of *presence* in meetings because they are only provided with the view from a stationary camera. These disengaged visitors feel as if they are watching the meeting through a peep-hole rather than attending as full participants.

To address this problem, we have built a camera control system that uses the video image of a person's head to control the pan, tilt, and zoom of a camera in a remote location. With no equipment besides a video camera and monitor at their site, users can peer through a virtual window [4] into another location, choosing their view as desired.

The attendee can control the orientation and zoom factor of the camera through natural head movements, as if peering through a window. To provide a position lock on a desired view, we have also introduced a freeze mechanism that permits the attendee to continue moving while maintaining a fixed orientation of the remote camera. These abilities greatly enhance the user's sense of presence in video-conference meetings [2].

The remainder of this paper discusses the system architecture, head-detection techniques, camera control algorithms and ongoing work.

## II. SYSTEM ARCHITECTURE

We use a SUN SPARC2 workstation with a frame grabber as the controller, and a motorized video camera to provide the attendee with a dynamic view of the conference room. The attendee's image is provided to the frame grabber, as shown in Fig. 1. An important point to recognize is that no special equipment is required at the site of the attendee. The attendee need only send his or her video image to the conference room, where all of the processing is performed.

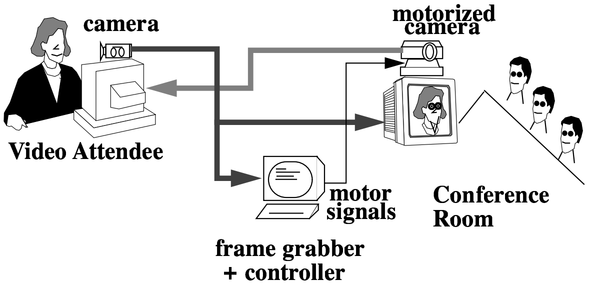

Figure 1. Configuration of equipment at user location and remote site. The video image of the attendee is provided to the conference room in addition to the frame grabber that processes the image to provide camera control motor signals.

The system software operates in two stages. Stage 1 computes the position and size of the attendee's head, while stage 2 provides camera control. The video image of the attendee is captured by the frame grabber and provided to stage 1. Since the current bottleneck in our system is the transfer of image data from the frame grabber to the computer, we use a quarter-size gray-scale image of 180 x 120 pixels to minimize the cost of this operation.

* Author is visiting from Hitachi Research Laboratory, Hitachi Ltd., Japan.

† Author is Principal Scientist, User Interface Research, Alias Research Inc., Toronto, Ontario.

Stage 1 produces a differential image from the current video image and a reference view, typically the first frame captured before the person entered the scene. This results in removal of the background, leaving only the attendee in the image. The top of the attendee's head is then located by scanning this image for a coherent set of non-zero pixels.

In stage 2, the head position is translated to orientation of the motorized camera. In order to reduce susceptibility to noise and potentially irritating constant small-scale camera motions, we apply a low-pass filter to the head parameters and further require that any changes to the camera orientation be of some minimum magnitude. The ratio of head width to frame width is also calculated and used to determine the zoom factor. These parameters are then sent to the camera through a serial interface. Stage 1 and 2 are repeated indefinitely, approximately twice per second. The operation of this system is illustrated in Fig. 2.

## III. HEAD DETECTION

With current technology, the grabbed image is unstable. Pixel intensities of the background image may vary due to noise in the video signal or changes in lighting conditions. It should be noted that our intention was to implement a basic version of the virtual window concept for videoconference environments rather than research state-of-the-art head-tracking techniques. With this in mind, we investigated three simple methods of head-detection.

Our initial attempt involved a comparison of the current video image with the previous frame to produce a differential image. Any pixel whose intensity had changed significantly was considered relevant. All other pixels were discarded. The problem with this approach is that little or no head movement between successive frames results in a sparse differential image of the head, which may be overwhelmed by background noise.

A second attempt involved pre-processing of each image by convolution with an edge-detection filter, and using the resulting frames to produce the differential image. Although edge information is stable under variable lighting conditions, background noise remained a problem. Furthermore, the loss of detail of the face made it difficult to locate the user's head.

Our third approach was to produce the differential image using an initial reference frame, taken without the attendee in the scene. While the need to begin the process with the user out of the camera view may be inconvenient, the results are far more stable than previous methods. Surprisingly, we found that this method also works fairly well in a large number of cases in which the reference image contains the attendee. However, this approach suffers under lighting variations or camera perturbations.

We are presently investigating the improved capabilities offered by faster frame grabber hardware and software, both from a technical and user perspective. Promising software approaches include snake-based trackers and connectionist architectures using pre-processed, colour-normalized input images (cf. [3]).

## IV. CAMERA CONTROL

Camera control techniques generally fall into two categories: position and velocity. Position control requires that the desired pan and tilt angles are provided directly to the camera. While this method is simple and accurate, current camera technology results in slow movements, and does not permit the interruption of a movement.

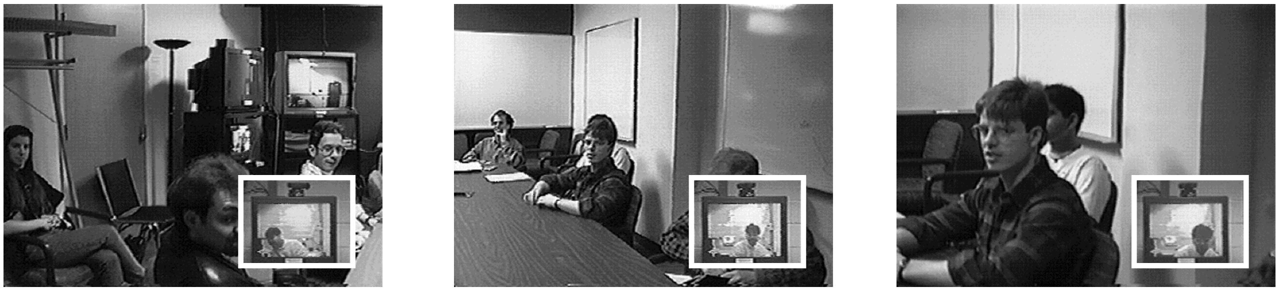

Figure 2. The head-tracking camera control system in operation. The large images represent the view received by the video attendee, while the small inset images represent the appearance of the attendee in the conference room. The motorized camera appears at the top of the video monitor.

With velocity control, the controller provides a velocity vector to the camera and instructs it to stop motion when the desired orientation is reached. Through feedback, corrections to the movement may be provided. This is quick and interruptible, although more complex than position control. For simplicity, we are presently using position control.

Due to the half second delay in image processing, neither method operates in true real-time. There may be a lag of several seconds for the camera to reach the correct orientation, especially for large head movements. However, our users found this delay to be quite tolerable in practice, provided that the camera begins moving within a short time of the head motion. Otherwise, users become confused. With faster video processing capabilities now available, these camera control methods need to be reconsidered.

A potential problem of a head-tracking system for camera control is that the video attendee must remain still in order to continue viewing the same scene. To address this issue, we have introduced a freeze mechanism that locks the camera orientation if the attendee's head remains relatively stable for a certain period (currently, ten seconds). This permits the attendee to select a desired view and once it is locked, freely move about. A simple gesture, such as covering the attendee's own camera lens, can be used to unlock the camera movement and resume head-following.

## V. OBSERVATIONS

In our preliminary testing, we have found that the capability of controlling the orientation and zoom factor of a camera in a remote location adds significantly to the user's sense of engagement in meetings. Because its use only requires individuals to perform the everyday action of looking through a window, anyone can use our system effectively with no special skills or training.

Control of camera orientation via head translation seems to pose no problems on the horizontal (yaw) axis. However, dependency on translation for vertical (pitch) control can be unnatural, since people normally use nodding-like motions, not head translations, to gaze up or down. With an improved head-tracking mechanism, it should be possible to track facial features and use these to provide pitch control.

The virtual window camera control has been incorporated into our video conference environment [1] and we are now analyzing the many interesting consequences of the technology in this setting, particularly the change in social protocols arising from its application.

We see this approach as being equally beneficial to other scenarios besides video-conferencing. Camera control via headtracking would be particularly advantageous in tasks such as teleoperation or surgery, in which the user's hands are required for other tasks.

## VI. FUTURE WORK

An additional feature that we are presently implementing allows the user to name several views, and then select one of these at any time by verbal instruction, supplied to a simple speech recognition system. The speech interface can also be used to conveniently lock the remote camera in a desired orientation and later, unlock it on demand.

Our present system has several limitations, notably the need to remove the user from the view before beginning the head-tracking process, and its inability to function correctly when there is more than one person in the view. To solve these problems, we are investigating other head-detection techniques, such as locating distinct facial features.

## ACKNOWLEDGEMENTS

We wish to thank Bill Gaver of the Royal College of Art, UK, for providing us with the idea of using the virtual windows concept in media space as a way to increase awareness and promote presence. The virtual windows concept originated at the Delft University of Technology, The Netherlands. In addition, we would like to thank Craig McQueen of the University of Toronto, for his idea of using a speech interface to identify different views.

## REFERENCES

1. Cooperstock, J., Tanikoshi, K., Beirne, G., Narine, T., Buxton, W. Evolution of a Reactive Environment. Proceedings of CHI'95, Denver, Colorado, May 1995.

2. Gaver, W., Smets, G., and Overbeeke, C. A virtual window on media space. Proceedings of CHI'95, Denver, Colorado, May 1995.

3. Hunke, M. and Waibel, A. Face Locating and Tracking for Human-Computer Interaction. 28th Asilomar Conference on Signals, Systems and Computers. Monterey, California, November 1994.

4. Overbeeke, C., and Stratmann, M. (1988). Space through movement. Unpublished doctoral thesis, TU Delft, The Netherlands.