Loosely-Coupled Telepresence Through the Panoramic Image Server

Michael Jenkin[1] James Elder[2] Greg Pintilie[1]
jenkin@cs.yorku.ca jelder@yorku.ca gregp@cs.yorku.ca
Departments of Computer Science[1] and Psychology[2]
York University
4700 Keele St., North York, Ontario, Canada, M3J 1P3

Abstract

While computer vision systems can clearly assist in surveillance tasks, taking the human out of the loop entirely proves to be difficult or undesirable in many applications. Human operators are needed to detect events missed by automatic methods, to identify false alarms, and to interpret and react appropriately to complex situations. A key challenge in partially-automated systems is how best to combine machine algorithms for event detection, analysis and tracking with telepresence systems controlled by one or more human operators. Given the disparity in performance between the human visual system and typical robotic cameras, we argue that direct coupling of human and machine is inappropriate. We propose instead to couple human and machine components indirectly, through a database called the Panoramic Image Server. We show how this loose coupling allows machine and operator surveillance priorities to be resolved while providing a fast and natural telepresence environment for one or more human operators.

1 Introduction

Developing completely autonomous surveillance systems is difficult for a number of reasons. To be practical, systems must function over a broad class of potentially dynamic topographies, and thus cannot be calibrated and optimized to the extent possible for inspection or CAD/CAM systems. A surveillance system must be highly reliable in that the probability of failing to detect an important event is very low: autonomous computer vision techniques will be hard-pressed to meet this demand. The false negative rate must also be low, since reacting to false alarms can be costly and disruptive.

On the other hand, there is a clear role for computer vision in surveillance. Human surveillance is costly, and fatigue and boredom can reduce the thoroughness and effectiveness of the visual sweep. While vision algorithms are never perfect, at least they do not get tired, and they can potentially notice events that escape human attention.

In the medium term, systems which combine computer and human surveillance symbiotically may be most effective. In this paper, we address the question of how to couple human and machine to optimize the effectiveness of surveillance and to avoid disorienting the human operator. We focus here on practical, low-cost solutions.

In order to obtain an image with reasonable levels of distortion at reasonable cost, sensors must necessarily have a limited field of view. One can employ multiple sensors, but this increases system cost and installation difficulty. Moreover, fusing sensors to provide a seamless telepresence and isotropic event detection is a very difficult task. The most effective means of obtaining low-cost panoramic surveillance capability is through a single camera with two-dimensional rotational servoing.

Unfortunately, combining human and computer surveillance effectively through a single robotic camera is difficult due to the disparate properties of the three components (human, computer, camera). While human analysis of a visual scene may be highly complex, image analysis algorithms for surveillance are at this stage relatively primitive. While the binocular field of view of the human visual system is relatively large (roughly 200×160 deg), and resolution is a strong function of eccentricity (M-scaling), typical cameras have a relatively small field of view, on the order of 30×30 deg, at a constant resolution.

A further difference is speed: low-cost off the shelf hardware simply cannot keep up with human gaze shifts. Human head and eye saccadic movements reach a maximum velocity of up to 500 deg/second [3]. While there are a number of robotic camera systems under development which approach this per-

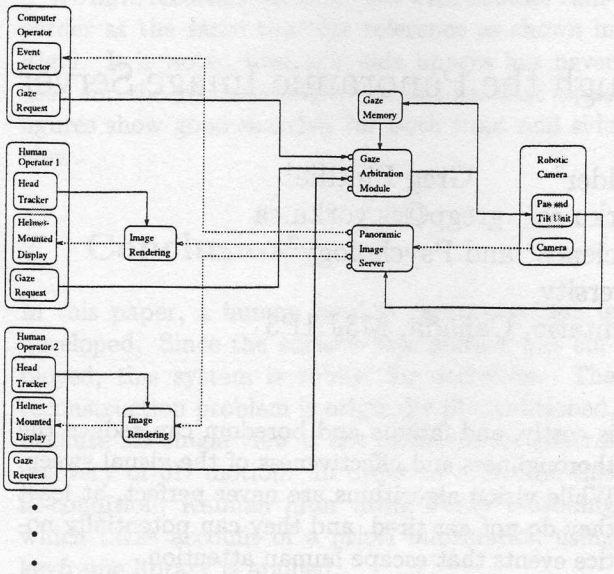

Figure 1: System design. Gaze information flow is indicated by solid lines, image information flow by dotted lines.

(a) Physical PTU

(b) Operator

Figure 2: System hardware

formance [5, 4, 2, 1], they are not presently available off the shelf, and when they are their cost may be prohibitive for widespread use.

Obtaining a slew rate comparable to the human head also does not mean that human gaze can be accurately tracked. Tracking requires a continuous feed of updated head or eye position, velocity or acceleration. Small delays can lead to large errors: in the space of a 20 ms delay in sensing head position, the head may have moved up to 10 deg. This demand means that the human operator and camera must be linked by a dedicated line. Using an existing switched network for communication is out of the question: any short interruption in the tracking feed will result in unacceptable disorientation for the human operator.

These differences, in algorithm complexity, field of view, M-scaling, and speed, suggest that optimal saccadic strategies for the camera/computer system and the human operator are likely to be quite different. Given these differences, it does not make sense to couple these systems too closely. Rather, we must find a way in which both systems can act with relative autonomy, but can coordinate where coordination is useful or necessary.

In this paper we describe a surveillance system which couples human and machine components indirectly, through a dynamic database called the Panoramic Image Server. We show how this loose coupling allows machine and operator surveillance priorities to be resolved while providing a fast and natural telepresence environment for the human operator. Human and camera are free to follow saccadic strategies which are optimized for their respective visual systems.

Loose coupling through the Panoramic Image Server allows the system to tolerate significant delays between head tracking signals and camera systems, permitting telepresence operation over large distances through low-cost switched connections. This loose coupling also permits multiple human operators to engage in surveillance through a common camera system, facilitating consultation and verification of complex events.

2 System Design

The overall design of the system is schematized in Figure 1. Flow of gaze information is depicted by solid lines, flow of image information by dotted lines. On the right is shown the robotic camera unit, which captures limited field of view images of the scene from a sequence of gaze directions, and conveys these (image, gaze direction) pairs to a software module called the Panoramic Image Server. The Panoramic Image Server may be either local or remote to the camera unit: in our prototype system it resides on the same SGI Indy machine that controls the camera unit shown in Figure 2a.

On the left side of Figure 1 is shown the *computer operator* and several *human operators*. The computer operator is a software module which receives panoramic image updates from the Panoramic Image Server, and uses these to detect potentially important events. When such events are detected, the computer operator sends a request to the gaze arbitration module to shift gaze to the location of interest.

Each human operator is tightly coupled to an image rendering module, which is local to the operator. In our system, image rendering is done on the SGI Indigo^2 machine that tracks the head movements of the operator. The image rendering module obtains periodic updates of the visual panorama from the panoramic image server and samples the head tracker at fixed intervals, obtaining updates of the relative position of the operator's head. The image rendering module uses these two signals to synthesize appropriate image frames, which are sent to a helmet mounted display system (see Figure 2b).

Human operators may signal an interesting or suspicious event through the keyboard. This generates a gaze request to the gaze arbitration module to move the camera in the direction the operator is looking.

The gaze arbitration software module arbitrates gaze requests from the computer operator and multiple human operators. It also incorporates a memory of the history of gaze directions, which is used to give priority to "stale" views which have not been updated for a relatively long time.

The key modules of the proposed surveillance system are described in more detail below.

2.1 The Panoramic Image Server

The job of the Panoramic Image Server is to receive (image, gaze direction) pairs from the camera unit, and to use this information to update a database of the visual panorama viewable from the camera at all possible gaze angles. Camera motions are modelled as perfect rotations. The offset of the tilt axis of our camera system from the optical centre of our camera leads to errors in constructing the panorama for surfaces near the camera and camera motions in the tilt direction. However, we find empirically that human operators are able to cope with these errors surprisingly well, and we have developed methods for preventing these errors from generating false alarms in the computer operator (see below).

We parameterize both the image frames I generated by the camera and the panorama P by discrete pan and tilt angles (ϕ, θ) as well as by discrete time t. Each image frame is initially 640×480 pixels, with a resolution of roughly 2.6×3.5 minutes of visual arc. For our prototype implementation, the panoramic image resolution is a factor of 4 lower in linear dimension, with a resolution of 10.4×14.0 minutes of arc, so that each image maps to a 160×120 pixel patch of the panoramic image. Since the total visual field over all gaze directions of our camera system is roughly 346×106 deg, this results in a panoramic image of size $1,996 \times 454$ pixels.

To reduce blocking effects in the overlapping of multiple frames at neighbouring gaze directions, updating of the panorama follows an exponential decay rule over time. Letting (ϕ_G, θ_G) represent the gaze angle of the camera, (ϕ_I, θ_I) represent discrete directions relative to the centre of image frame I, and (ϕ, θ) represent discrete directions relative to the centre of the panorama P, the updating rule is:

$$P(\phi, \theta, t) = (1-\gamma)P(\phi, \theta, t-1) + \gamma I(\phi_G + \phi_I, \theta_G + \theta_I)$$

where γ is the half life of an image snapshot. Whenever a gaze direction is updated, a residual is computed as $r(\phi, \theta) = P(\phi, \theta, t) - P(\phi, \theta, t-1)$.

At initialization, the Panoramic Image Server scans through the entire (ϕ, θ) space through a sequence of overlapping views. Each pixel of $P(\phi, \theta, 0)$ is then initialized to the mean of the grey level values recorded in the direction (ϕ, θ) over all overlapping views which include (ϕ, θ). A second panoramic map $\sigma(\phi, \theta)$ is used to store the standard deviation of $P(\phi, \theta, 0)$, for use in the Gaze Arbitration Module. Figure 3 shows the output of this initialization phase. A panoramic view of the environment is obtained together with an estimate of the grey level uncertainty in each viewing direction.

The display side of the panoramic image server is straightforward. Upon receiving each new image frame from the camera unit, it updates the panorama and stores it in a local file. The updated panorama is obtained by the operator units over the local area network as requested.

2.2 Gaze Arbitration Module

The gaze arbitration software module receives gaze requests from the computer and human operators and considers these requests together with a memory of history of gaze directions to generate a gaze command, which is sent to the camera servoing unit to shift gaze direction. The gaze memory is a database isomorphic to the visual panorama which specifies for each discrete visual direction the previous time at which the grey level value in the

Initial panoramic view

Baseline noise

Figure 3: Initial panoram

panorama was updated. Denoting this panoramic gaze database as $G(\phi,\theta,t)$, and the halfwidth of each image in pan and tilt directions as (ϕ_I^M,θ_I^M), the update rule is

$$G(\phi,\theta,t) = \begin{cases} t & \text{if } |\phi_G-\phi| < \phi_I \text{ and} \\ & \quad |\theta_G-\theta| < \theta_I \\ G(\phi,\theta,t-1) & \text{otherwise} \end{cases}$$

The major task of the Gaze Arbitration Module is the prioritization of physical gaze directions of the camera. Two separate issues are involved

1. The camera should view what the human operators want to have updated.
   As multiple operators may want access to the same panorama, it is not desirable to simply track the operator's head motion.

2. The Panoramic view should be as fresh as possible.
   This involves re-viewing the entire panorama but it is not desirable to simply scan through space. Dynamic regions should be viewed more frequently than static regions, and it is desirable to view locations indicated by the operator more frequently than other regions.

In order to accomplish these tasks, the Gaze Arbitration Module selects the gaze direction which maximizes the interest function Ω

$$\Omega(\phi,\theta,t) = \alpha\frac{t-G(\phi,\theta,t)}{t} + (1-\alpha) * \frac{|r(\phi,\theta)|}{\sigma(\phi,\theta)}.$$

Here $r(\phi,\theta)$ is the update residual at (ϕ,θ) and $\sigma(\phi,\theta)$ is obtained from the offline initialization phase. Ω essentially seeks out pixels which have not been seen recently (first term) and which have a residual larger than the expected residual for this update.

In order to accommodate a user's request that a certain gaze direction should be prioritized for viewing, the value $G(\phi,\theta,t)$ in the panoramic gaze database is updated using

$$G'(\phi,\theta,t) = G(\phi,\theta,t)/2$$

for each request to view in direction (ϕ,θ). This essentially "ages" this view direction and makes it more desirable for the attention mechanism.

3 Hardware

For our prototype system we used a Watec camera with a field of view of 28 × 28 deg and a resolution of 640 × 480 pixels, providing an angular resolution of approximately 2.6 × 3.5 minutes of visual arc. The camera was mounted on a pan and tilt

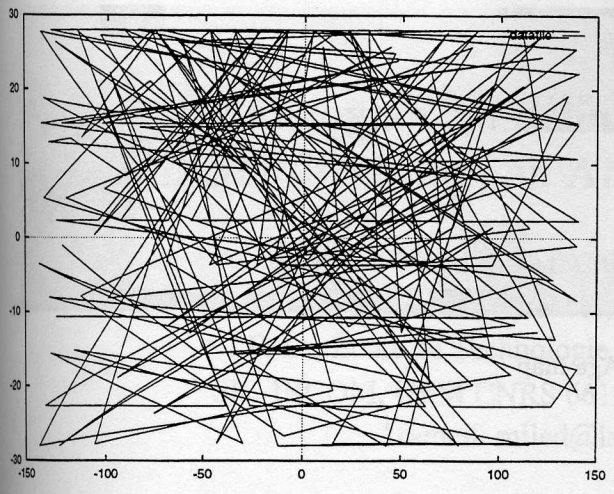

Figure 4: Gaze directions of the camera

unit from Directed Perception (model PTU). The range of motion of the unit was roughly ± 140 deg in pan and ± 50 deg in tilt. The nominal resolution of the unit was 3.1 arc min. Although the pan axis was near the optical centre of the camera, the tilt axis was approximately 2 inches below, resulting in significant forward and backward translation of the camera. Modeling camera movements as pure rotations leads to a maximum error of roughly 0.3 deg (6 pixels) for a surface at a distance of 1m and a field of view of 28×28 deg.

Human operators wore helmet-mounted displays with integrated head-tracking from Virtual i-O. The NTSC video signal was split over the 100% overlapping displays for the two eyes, providing a binocular resolution of roughly 640×240 pixels over a field of view of approximately 30×11 deg, producing a resolution of roughly 2.81 minutes of arc. Thus a 1:1 mapping of pixels of the panoramic image to the display results in a magnification factor of roughly 4, however operators were provided with zoom control which allowed them to vary magnification over a large range.

4 Results

When left to its own devices, the panoramic image server will continually view the room choosing different gaze directions. Figure 4, for example, shows the gaze directions (in pan/tilt) for the pan and tilt unit for 274 different gazes. Initially the panoramic image server tends to scan the entire image in order to establish a recent reviewing of each pixel. After this, pixels with high variability are chosen while pixels are continually re-scanned.

Figure 5 show the age map associated with viewing the environment shown in Figure 3. The intensity in the image corresponds to the last sample time for this pixel, with lighter coloured pixels being viewed more recently.

5 Discussion and Future Work

The ability to divorce the head-slaved nature of pan and tilt units in security work has many advantages. It allows multiple users to utilize the same hardware. It replaces limitations in the dynamic response of the image acquisition hardware with limitations on the display hardware. It allows the attention of the image capture hardware to be directed by both machine and human operators. And it allows the operator to change his or her view direction without signalling their intentions through the overt motion of the pan and tilt unit.

The major disadvantage associated with using a virtual pan and tilt unit as described here is the need to construct in an on-line manner a panoramic view of the world from individual snapshots. This paper presents a straightforward approach to this problem. For the prototype system discussed here, we have taken the error due to camera translation into account by precomputing a baseline uncertainty map for the scene under surveillance. While this has been shown to be surprisingly effective in practice, it does raise the threshold of events which can be automatically detected, leading to a potentially unacceptable false negative rate. A better alternative may be to use standard grey-level correlation techniques for precomputing an affine remapping of (image frame, gaze direction) pairs to the panoramic image for the scene under surveillance. This coarse remapping can be periodically updated to take into account large changes in the geometry of the scene.

Various attention models can be considered. The model here attends to directions which change - relative to the noise associated with this location - and to locations which have not been viewed recently. Information about the environment could easily be integrated into the view selection function Ω.

The current operator interface to the panoramic image server is based on a head mounted display. Additional input technologies, including the development of a web-based front end are the subject of future work.

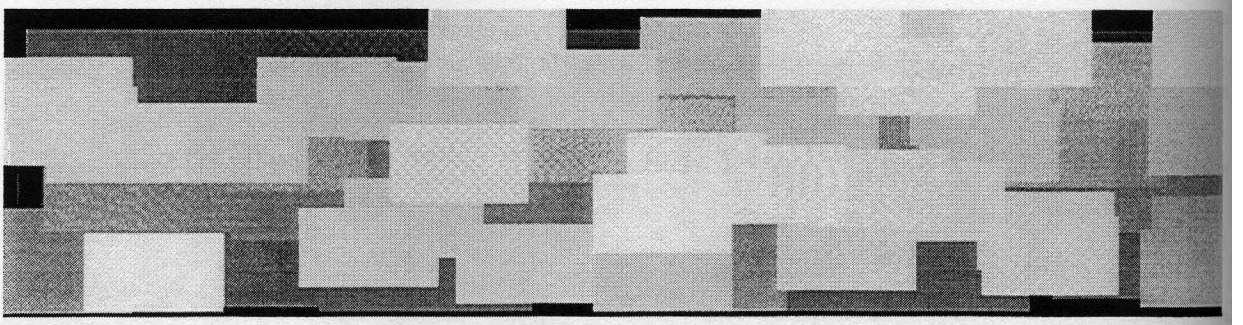

Figure 5: Age map

References

[1] C.M. Brown, D. Coombs, and J. Soong. Real-time smooth pursuit tracking. In A. Blake and A. Yuille, editors, *Active Vision*, chapter 8. MIT Press, Cambridge, MA, 1992.

[2] J.J. Clark and N.J. Ferrier. Attentive visual servoing. In A. Blake and A. Yuille, editors, *Active Vision*, chapter 9. MIT Press, Cambridge, MA, 1992.

[3] D. Guitton and M. Volle. Gaze control in humans: eye-head coordination during orienting movements to targets within and beyond the oculomotor range. *J. Neurophys.*, 58(3):427–459, Sept 1987.

[4] K. Pahlavan and J.-O. Eklundh. A head-eye system - analysis and design. *Computer Vision, Graphics and Image Processing: Image Understanding*, 56(1):41–56, 1992.

[5] P.M. Sharkey, D.W. Murray, S. Vandevelde, I.D. Reid, and P.F. McLauchlan. A modular head/eye platform for real-time reactive vision. *Mechatronics*, 3, 1993.