Active surveillance using dynamic background subtraction.

Kenneth M. Dawson-Howe,
Dept. of Computer Science, Trinity College, Dublin. Ireland.
Tel. +353-1-6081220   Fax. +353-1-6772204
Kenneth.Dawson-Howe@cs.tcd.ie

August 13, 1996

Abstract

A prototype active surveillance system is presented. The system has two cameras, one of which is rigidly fixed & has a wide angle lenses while the other is mounted on a pan-tilt head & has a zoomed lenses. The images from the wide angle camera are analyzed using dynamic background subtraction together with normalized cross correlation in order to identify moving people within the scene. An approximate mapping is determined (off-line) between the pan-tilt angles and the position of the zoomed image within the wide-angle view. This mapping is improved (for each frame) through 1-D image based correlation and hence moving people are located in the zoomed view.

The system has been tested on a difficult real-world scene and image sequences from both cameras are presented. The potential of the system as an 'intelligent' security device and the power of the dynamic background subtration & correlation mechanism are clearly demonstrated.

Keywords: Active vision, Surveillance, Motion tracking

Abbreviated Running title: Active Surveillance

Introduction

When installing a surveillance camera a balance must be struck between coverage and magnification [1]. As a result, as most surveillance cameras are intended only as a deterrent, they generally view a large area and provide images of people which are below the resolution which would be required in order to identify them. To overcome this problem more complex camera systems are required and one such system is presented in this paper.

The system presented has two cameras, one of which is rigidly fixed and uses a wide-angle lenses in order to view a large area of the scene. The images from this camera are analyzed in order to detect moving objects/people (See Figure 1 (a)). The second camera in the system is mounted on a pan-tilt head and has a zoomed lenses in order to obtain high resolution views of parts of the scene (See Figure 1 (b)). The pan-tilt head is driven so that the zoomed camera views the moving objects in the scene, on the basis of the information provided by the wide-angle camera.

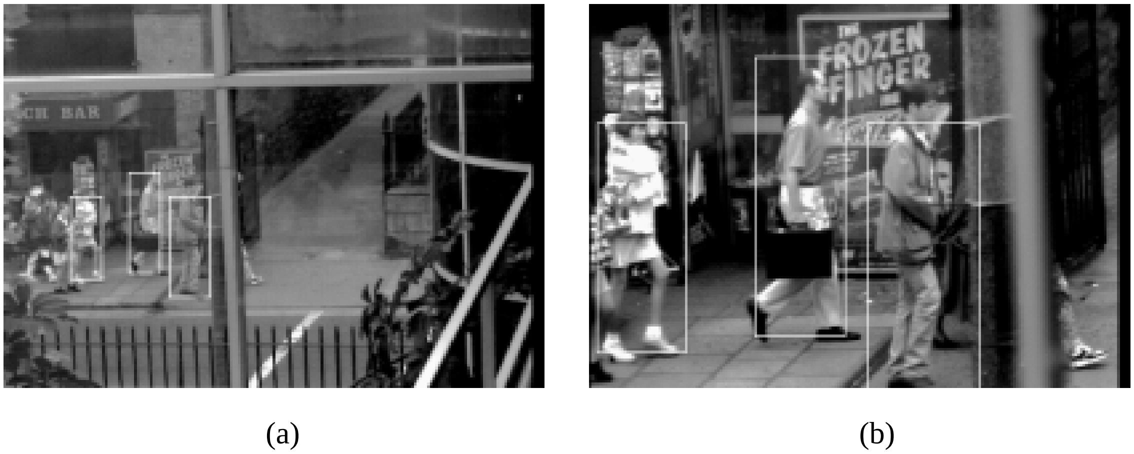


Figure 1: Wide angle (a) and zoomed (b) camera views showing the tracked objects marked with rectangles.

Many systems have been developed for object tracking based on image subtraction [2, 3], edge detection [2, 4], optical flow [5, 6, 7], corner detection [8, 9], etc. These systems vary considerably in their complexity and in the domains which they address.

One of the significant differences in the domain is that some of the systems assume that the camera from which the tracking is done is rigidly fixed [2, 4, 5] while others address the less constrained task of tracking moving objects from a moving platform [3, 6, 7, 9, 10]. The system described in this paper takes an approach which is between these two extremes, as it makes use of a rigidly fixed camera for the basic object tracking, but uses a second, moving camera to provide high resolution images of the moving objects.

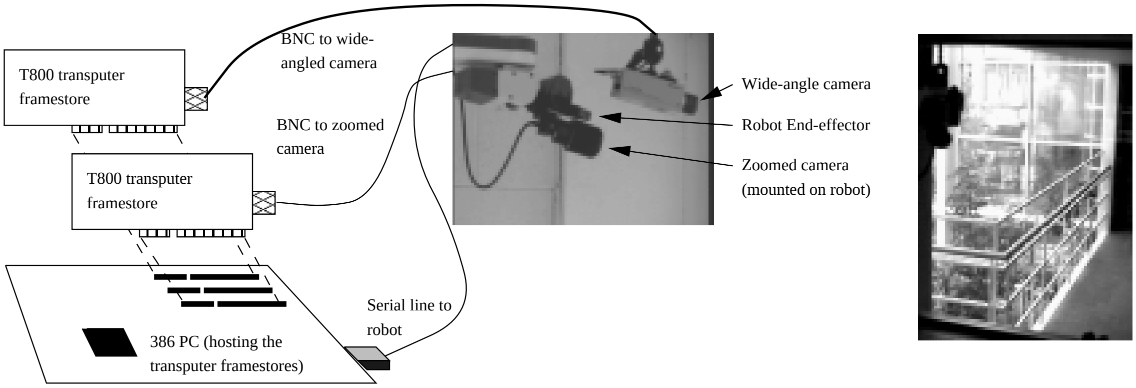
By using a rigidly fixed camera for the motion detection it is possible to make use of a dynamic background subtraction technique, contrary to the typical view that such a simple technique cannot be used in a complex domain [5].

As a result of this approach a number of problems must be dealt with including the registration of $(pan, tilt)$ position to zoomed image position (i, j) and the image based improvement of that registration while performing active tracking.

Architecture

The prototype system presented in this paper was developed on a 386 PC hosting two T800 transputer framestores & controlling a SCARA configuration robot which is employed in this application simply as a pan-tilt mechanism (See Figure 2 (a)).

While this technology is somewhat outdated (as modern processors are orders of magnitude faster & current pan-tilt head technology is much more reliable and


T800 transputer framestore
BNC to wide-angled camera
BNC to zoomed camera
T800 transputer framestore
386 PC (hosting the transputer framestores)
Serial line to robot
Wide-angle camera
Robot End-effector
Zoomed camera (mounted on robot)

(a) (b)

Figure 2: The basic system architecture (a) is shown together with a view taken from behind the two cameras (b) (where the cameras and robot end effector appear in black in the top left).

efficient than the robot arm used), it does provide a reasonable testbed for the development of the required algorithms and, in fact, a number of tracking systems have previously been developed on such transputer platforms [11, 12]. The use of such relatively slow technology for the system prototype provides a good level of confidence that the final system will be able to cope with larger and more complex scenarios.

The test environment selected for the prototype system was a real world scene which was visible directly from our research laboratory (See Figure 2 (b)). The scene was a section of the pavement beside a road outside a local shop which was around 30 meters from the camera system and was viewed through two windows (one in laboratory and the other in our building's foyer).

Object Tracking

In order to allow high-resolution images of the people in the scene to be acquired it is reasonable to assume that such people move about in the scene. To monitor the scene reliably it is essential that the processing time per frame be as low as possible. Hence it is important that the techniques which are employed are as simple and as efficient as possible. For that reason the well known technique of background subtraction [2, 3, 5] was selected for this application.

Background subtraction allows moving objects to be detected by taking the point-by-point absolute difference of the current image and a background image which must be acquired when there are no moving objects in the scene (See equation 1).

$$Moving_t(i,j) = |Image_t(i,j) - Background(i,j)| \tag{1}$$

Such a mechanism is impractical for the surveillance system described in this paper as it may not be possible to obtain a background image with no moving

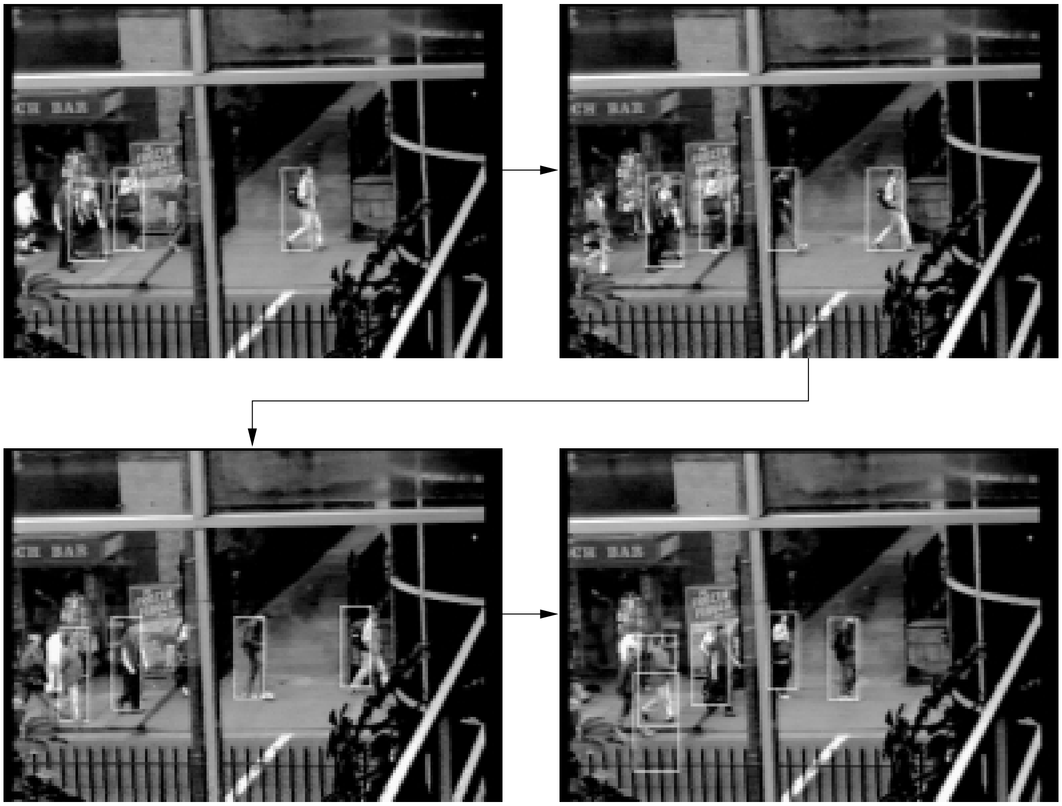


Figure 3: A sequence of images from the wide angle camera showing moving objects as they are tracked. Note that the people to the very far left of the scene are not tracked as they are not within the defined region of interest.

objects, and more importantly the background of the scene may change due to lighting conditions or 'stationary' objects being moved (e.g. a gate being opened and then left open). As Bartolini *et al.* [5] point out, "algorithms based on direct grey-level comparison are not robust enough against sudden lighting changes". It is possible though to overcome these problems by using a dynamic background together with normalized cross correlation to evaluate any changes.

The dynamic background is initialized with the first image acquired (whether or not that image contains any moving objects), and is updated if a point changes and remains changed for a number of frames (See equation 2).

$$Moving_t(i,j) = |Image_t(i,j) - Background_t(i,j)| \tag{2}$$

$$where$$

$$Background_t(i,j) = \begin{cases} Image_0(i,j) & t \leq 2 \\ Image_t(i,j) & Image_t(i,j) \neq Background_t(i,j) \quad and \\ & Image_t(i,j) \approx Image_{t-1}(i,j) \approx Image_{t-2}(i,j) \\ Background_{t-1}(i,j) & otherwise \end{cases} \tag{3}$$

Lighting changes are still a problem as the background subtraction mechanism initially responds to the lighting changes in the same way that it would respond to a moving object as it takes a number of frames for the background to be updated. To overcome this the image was divided into 64 by 64 regions and the normalized cross correlation N (See equation 4) was computed for any region in which points changed (as determined by the background subtraction). The normalized cross correlation is independent of average intensity and hence moving objects can be discriminated from lighting changes.

$$N = \frac{\sum_i \sum_j Image_t(i,j).Background_t(i,j)}{\sqrt{\sum_i \sum_j (Image_t(i,j))^2}.\sqrt{\sum_i \sum_j (Background_t(i,j))^2}} \tag{4}$$

This method proved too slow for the prototype system being developed and hence a region of interest (around half the size of the image) was defined and the image was subsampled (one in every four pixels along both the I & J axes).

To detect moving people in the scene it was necessary to aggregate the information obtained from the region correlations. This was done by applying a 2-D averaging filter (actually implemented as two 1-D filters) to the correlations where the filter shape was defined by the expected average height and width of people in the scene. Local maxima which are bordered by regions with relatively small changes then represent possible moving people within the scene. The size/shape of the moving region can be refined somewhat at this stage by checking the original correlation values.

An example sequence of images from the tracking system is shown in Figure 3.

Registration

In order to drive the pan-tilt head to direct the zoomed camera to a particular part of the scene it is necessary to determine a mapping between (i, j) values in the wide-angle scene and $(pan, tilt)$ values.

The mapping is determined by driving the pan-tilt head to a number (25 in this case) of different positions and determining through template matching the (i, j) position in the wide-angle view which corresponds to each of the zoomed views. Note that the magnification between the two views must be supplied in order to allow the zoomed view to be scaled down so that it corresponds to the wide angle view. In addition when template matching it was determined that it was important to consider only those sections of the wide-angle view which are part of the scene being viewed (e.g. it was important not to base the registration on the window frame which appears as part of the test scene, as it will be shifted somewhat relative to the rest of the scene due to the slightly different camera angles). See the upper left corner of Figure 4 for an example fo the regions of interest used for template matching.

In order to speed up the process of template matching the search for the best fit was done at a number of different resolutions. At the lowest resolution the templates are compared in all possible positions (subject to having sufficient points within the regions of interest in the wide angle view). At higher levels only the positions corresponding to the highest (16) correlations from the previous level are evaluated; See Figure 4.

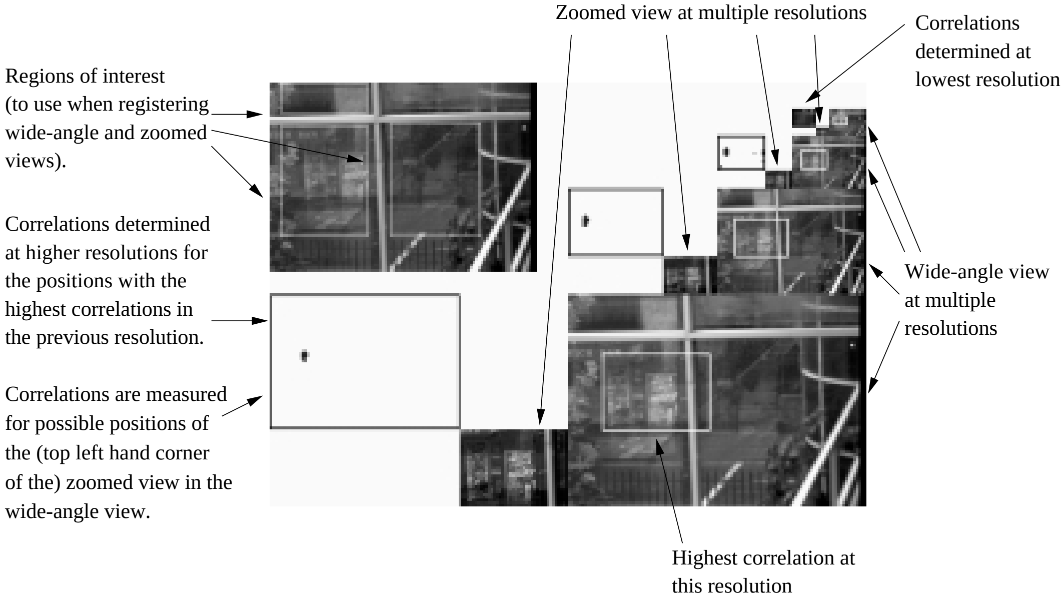

Zoomed view at multiple resolutions
Correlations determined at lowest resolution
Regions of interest (to use when registering wide-angle and zoomed views).
Correlations determined at higher resolutions for the positions with the highest correlations in the previous resolution.
Correlations are measured for possible positions of the (top left hand corner of the) zoomed view in the wide-angle view.
Wide-angle view at multiple resolutions
Highest correlation at this resolution

Figure 4: Multi-resolution image correlation. This correlation is done for 25 different zoomed views in order to allow a mapping from $(pan, tilt)$ to image position (i, j) (i.e. how the zoomed view relates to the wide-angle view).

The (i, j) to $(pan, tilt)$ mapping is modeled using a bi-linear transformation (See equations 5 & 6).

$$pan = a_1 + a_2.i + a_3.j + a_4.i.j \tag{5}$$
$$tilt = b_1 + b_2.i + b_3.j + b_4.i.j \tag{6}$$

For any given observation of (i, j) and the corresponding $(pan, tilt)$ values there will probably be a small discrepancy between the $(pan, tilt)$ values observed and those predicted by the mapping functions. These discrepancies may be expressed as

$$\sigma_i = a_1 + a_2.i + a_3.j + a_4.i.j - pan \tag{7}$$
$$\gamma_i = b_1 + b_2.i + b_3.j + b_4.i.j - tilt \tag{8}$$

In order to minimize the discrepancies σ_i & γ_i, values for $a_1 ... a_4$ and $b_1 ... b_4$ must be determined which make

$$E_\sigma = \sum_{i=1}^{n} \sigma_i^2 = \sum_{i=1}^{n} (a_1 + a_2.i + a_3.j + a_4.i.j - pan)^2 \tag{9}$$
$$E_\gamma = \sum_{i=1}^{n} \gamma_i^2 = \sum_{i=1}^{n} (b_1 + b_2.i + b_3.j + b_4.i.j - tilt)^2 \tag{10}$$

as small as possible (where n is the number of observations). To do this each of the first partial derivatives must be equated to zero and by rearranging a linear equation for each partial derivative (w.r.t. the various a_x and b_x coefficients) may be obtained. Having two sets of 4 equations each with 4 unknowns it is a routine matter to solve for the 8 co-efficients.

The inverse mapping (from $(pan, tilt)$ to image position (i, j)) may be determined simply by rearranging the equations (although the solution is quadratic giving the possibility of two roots).

Active Tracking

Having identified potential moving people and determined the mapping between $(pan, tilt)$ and image position (i, j) it is possible to track objects as they move about within the wide-angle scene and hence direct the zoomed camera to view the relevant section of the scene. An example of such tracking by the zoomed camera is shown in Figure 5.

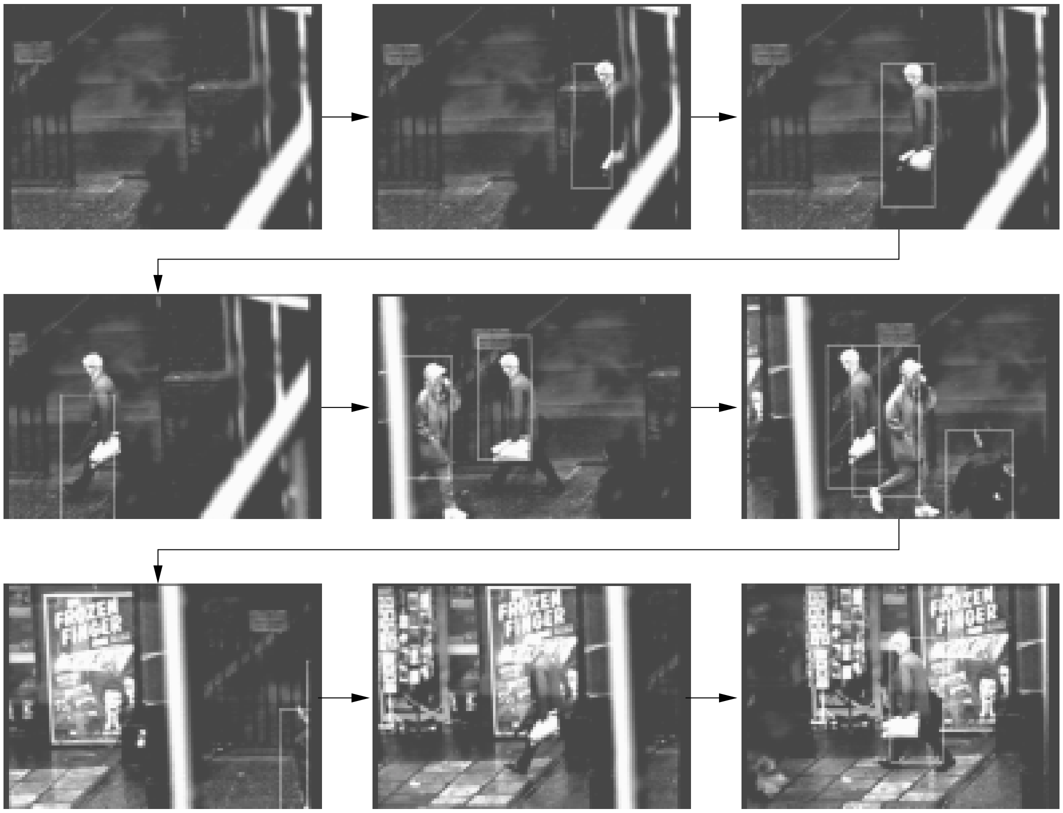


Figure 5: A sequence of images from the zoomed camera showing the moving objects as they are tracked (by the wide-angle camera). Note that in the fifth frame the process for refining the estimate of position failed in one direction, in the seventh frame the individual being tracked has disappeared behind part of the window frame, and in the eighth frame the person was not located by the moving-object detection software as he blended quite well into the background.

Unfortunately the $(pan, tilt)$ information returned by the robot was not to be accurate enough when converted into (i, j). This was readily apparent as the boxes drawn around the objects on the zoomed view were offset by a significant amount (from the objects that they represented). This problem was solved by refining the estimate of (i, j) through correlation of the zoomed view and the wide-angled view. To do this in an efficient fashion, the correlation was done for one row and one column only. The row and column were selected carefully from the zoomed view, so that they provided the greatest likelihood that the (i, j) values would be tuned successfully. A technique similar to those described in [13] and [14] in which the i and j values are tuned separately was employed.

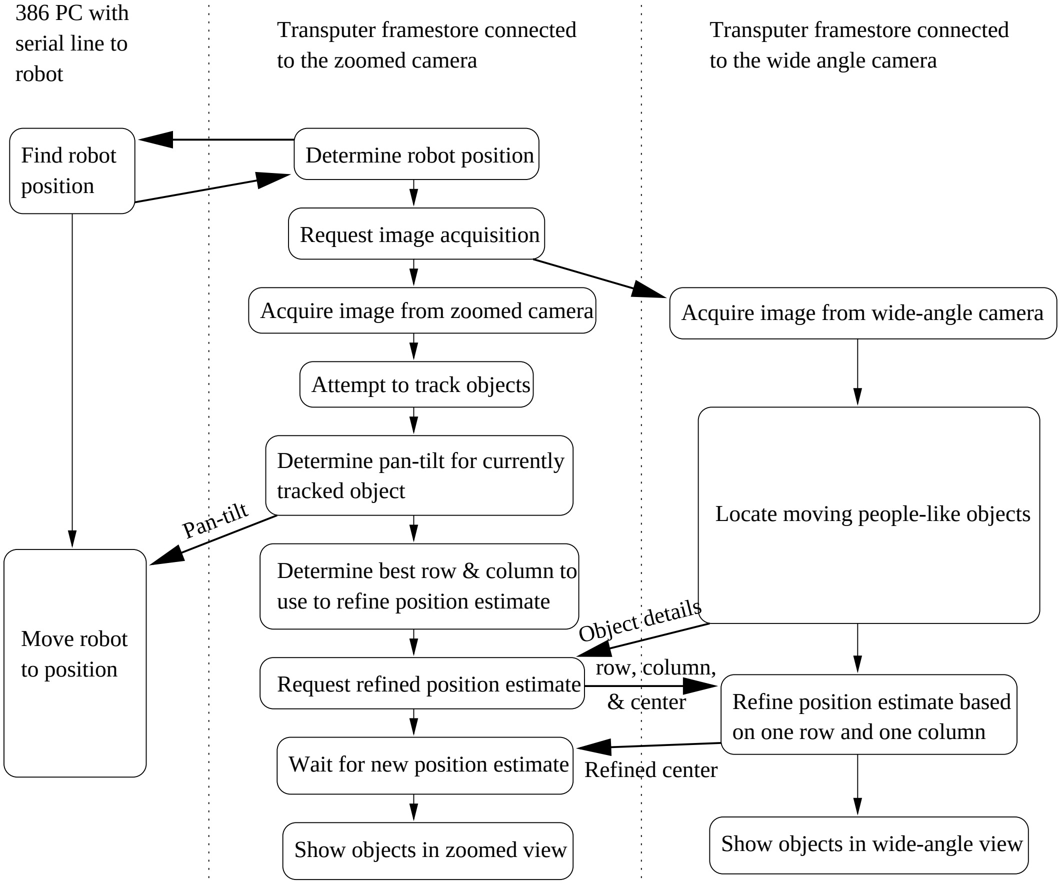

386 PC with serial line to robot
Transputer framestore connected to the zoomed camera
Transputer framestore connected to the wide angle camera
Find robot position
Determine robot position
Request image acquisition
Acquire image from zoomed camera
Acquire image from wide-angle camera
Attempt to track objects
Determine pan-tilt for currently tracked object
Locate moving people-like objects
Pan-tilt
Move robot to position
Determine best row & column to use to refine position estimate
Object details
Request refined position estimate
row, column, & center
Refine position estimate based on one row and one column
Wait for new position estimate
Refined center
Show objects in zoomed view
Show objects in wide-angle view

Figure 6: Flowcharts showing the sequence of operations performed (continually) by each of the three processors in the system. Note that the objects which are tracked (from frame to frame) are those identified from the previous frame. The $(pan, tilt)$ which is determined is not achieved by the robot until the next frame, and hence it is necessary to predict the $(pan, tilt)$ position two frames in advance.

The overall flow of operations is shown for each of the three processors in Figure 6. Note that it is necessary to predict the position of the moving target two cycles in advance due to the time taken to locate moving objects & to drive the pan- tilt head to the required position.

Results & Conclusions

The system operates at an average rate of 2.1Hz. Example sequences from both the wide angle and the zoomed camera are shown in Figures 3 & 5 respectively. The system has to balance the need for reliable tracking against the need for a fast frame rate, and hence while the frame rate is extremely good for such out-dated technology moving object identification is not always successfully achieved (e.g. See the eighth frame of Figure 5). This situation is not unusual though, as Bradshaw *et al.* point out, "to obtain the required performance from finite hardware has required compromise, and one must expect the output to contain not only statistical but also occasional gross error" [10].

Considering the application the system has successfully demonstrated the ability to actively track objects in a scene, and hence it shows potential as an 'intelligent' security camera.

References

[1] **Capel, V** *Security Systems and Intruder Alarms*, Heinemann Newnes, Oxford (1989)

[2] **Hogg, D** 'Model-based vision: a program to see a walking person', *Image and Vision Computing* Vol 1 No 1 (February 1983) pp 5-20

[3] **Murray, D and Basu, A** 'Motion Tracking with an Active Camera', *IEEE Trans. PAMI* Vol 16 No 5 (May 1994) pp 449-459

[4] **Sullivan, G** '*A Priori* Knowledge in Vision', in **D Vernon (ed)** *Computer Vision: Craft, Engineering, and Science*, Springer-Verlag, Berlin (1994) pp 58-79

[5] **Bartolini, F, Cappelini, V and Mecocci, A** 'Counting people getting in and out of a bus by real-time image sequence processing', *Image and Vision Computing* Vol 12 No 1 (January/February 1994) pp 36-41

[6] **Reddi, S and Loizou, G** 'Actively keeping a moving target at the center of the field of view', *IEEE Trans. PAMI* Vol 17 No 8 (August 1995) pp 765-776

[7] **Smith, S and Brady, J** 'Optical flow based moving object detection in a moving scene', *IEEE Trans. PAMI* Vol 17 No 8 (August 1995) pp 814-820

[8] **Wang, H and Brady, M** 'Real-time corner detection algorithm for motion estimation', *Image and Vision Computing* Vol 13 No 9 (November 1995) pp 695-703

[9] **Cox, I and Hingorani, S** 'An Efficient Implementation of Reid's Multiple Hypothesis Tracking Algorithm and Its Evaluation for the Purpose of Visual tracking', *IEEE Trans. PAMI* Vol 18 No 2 (February 1996) pp 138-150

[10] **Bradshaw, K, McLauchlan, P, Reid, I and Murray, D** 'Saccade and pursuit on an active head/eye platform', *Image and Vision Computing* Vol 12 No 3 (April 1994) pp 155-163

[11] **Stephens, R** 'Real time 3D object tracking', *Image and Vision Computing* Vol 8 No 1 (February 1990) pp 91-96

[12] **Welch, P and Wood, D** 'Image Tracking in real-time: a transputer emulation of some early mammalian vision processes', *Image and Vision Computing* Vol 11 No 4 (May 1993) pp 221-228

[13] **Horn, B and Bachmann, B** 'Registering Real Images using Synthetic Images', in *Artificial Intelligence - An MIT Perspective*, MIT Press, Boston (1979) pp 129-160

[14] **Dawson, K** '3-D Object Recognition', in **D Vernon and G Sandini** (eds) *Parallel Computer Vision - the VIS a VIS system*, Ellis-Horwood, New York (1992) pp 185-213