ELSEVIER

Robotics and Autonomous Systems 35 (2001) 221–228

Robotics and Autonomous Systems

www.elsevier.com/locate/robot

Real-time human activity monitoring exploring multiple vision sensors

Paulo Peixoto*, Jorge Batista[1], Helder J. Araujo[2]

Department of Electrical Engineering, Institute of Systems and Robotics (ISR), University of Coimbra, 3030 Coimbra, Portugal

Abstract

In this paper, we describe the monitoring of human activity in an indoor environment through the use of multiple vision sensors. The system described in this paper is made up of three cameras. Two of these cameras are active and are part of a binocular system. They operate either as a set of three static cameras or as a set of one fixed camera and an active binocular vision system. The human activity is monitored by extracting several parameters that are useful for their classification. The system enables the creation of a record based on the type of activity. These logs can be selectively accessed and provide images of the humans in specific areas. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Surveillance; Active vision; Real-time tracking

1. Introduction

Automated monitoring of human activity is important for many applications. The problem of analysing human activity in video has been the focus of several researchers' efforts and several systems have been described in the literature [4,5,8–10,13,15,16]. Many of these systems consist of a computer vision system to detect and segment a moving object and a higher level interpretation module.

In very specialised applications other sensors are used besides vision. Automatic interpretation of the data is very difficult and most systems in use require human attention to interpret the data. These systems are characterised by the storage of large amounts of data that require no specific action. A single and specific event which may require immediate intervention can be difficult to find among a lot of redundant information.

Tracking human motion in an indoor environment is of interest in several applications. A considerable amount of work has been devoted to tracking humans with a single camera. Images of the environment are acquired either with static cameras with wide-angle lenses (to cover all the space), or with cameras mounted on pan and tilt devices (so that all the space is covered by using good resolution images) [6,7,11,12,14,17]. In some cases both types of images are acquired but the selection of the region to be imaged by the pan and tilt devices depends on the action of a human operator. The combination of several modalities of imaging devices enables the achievement of robust performance. In addition, the monitoring of events requires the use of multiple sensing agents. This is an important and essential step towards the full automation of high-security applications in man-made environments.

* Corresponding author. Tel.: +351-239796275; fax: +351-239406672.

E-mail addresses: peixoto@isr.uc.pt (P. Peixoto), batista@isr.uc.pt (J. Batista), helder@isr.uc.rt (H.J. Araujo).

[1] Tel.: +351-239796289; fax: +351-239406672.

[2] Tel.: +351-239796216; fax: +351-239406672.

0921-8890/01/$ – see front matter © 2001 Elsevier Science B.V. All rights reserved.
PII: S0921-8890(01)00117-8

The system described in this paper tries to explore the combination of several vision sensors in order to cope with the proposed goal of autonomously detecting and tracking human intruders in man-made environments. In the current setup the system is made up of three cameras that can operate in two modes: passive mode and active mode. In the passive mode the three cameras remain static and monitor the environment. When specific events occur the system starts operating as a combination of a static camera and a binocular active system (entering in the active mode).

The system is also able to detect and log the presence of targets in some specific areas of interest. This log file can be consulted off-line in order to search for some particular event.

2. Global vision

During the active mode of operation the global vision system (the wide-angle static camera) is responsible for the detection and tracking of all targets visible in the scene. It is also responsible for the selection of the target that is going to be tracked by the binocular active system.

2.1. Ground plane correspondence

In order to redirect the attention to a new target the active vision system should know where to look for it. Since, the position of the target is known in the static camera image, we will need to map that position in terms of rotation angles of the neck (pan and tilt) and vergence (we are assuming a symmetric vergence configuration). The goal would be to fixate the active vision system on the target head.

Assuming that all target points considered in the static camera image lie on the ground plane then any point on this plane can be mapped to a point in the image plane of the static camera using a homography [3].

For each detected target on the image plane $p(x,y)$, we compute the corresponding point on the ground plane. Then the relationship between the point $P(X,Y)$ in the plane and the joint angles can be derived directly from the geometry of the active vision system (see Fig. 1):

$$\theta_p = \arctan\frac{X}{Y},$$

$$\theta_v = \arctan\frac{B/2}{\sqrt{X^2+Y^2}-D},$$

$$\theta_t = \arctan\frac{H-h}{\sqrt{X^2+Y^2}},$$

where B is the baseline distance.

To compute the tilt angle, we must know the target height which can be easily computed, since we can obtain the projection of the targets head and feet points (detected in the image plane) on the ground plane. Assuming that the static camera height and position, relative to a predefined referential on the

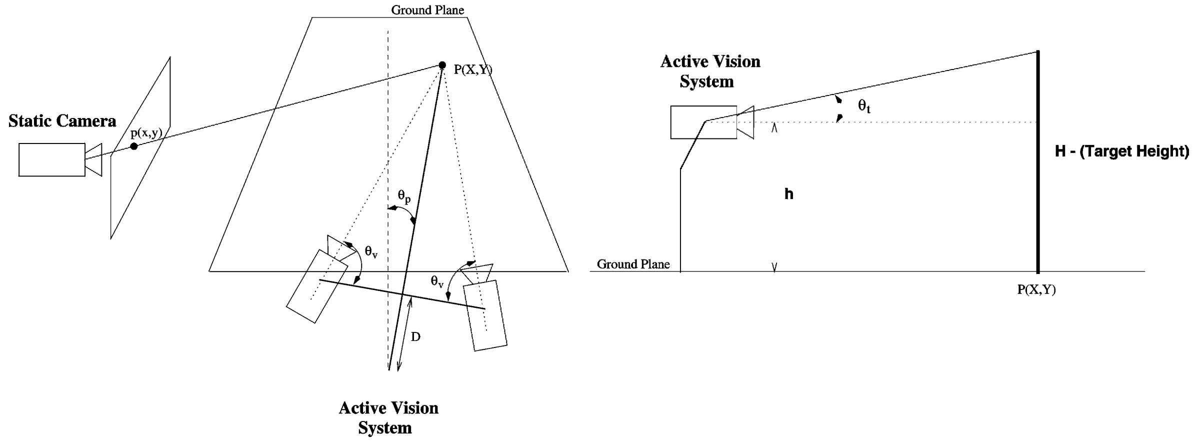


Fig. 1. Correspondence between image points and ground plane points.

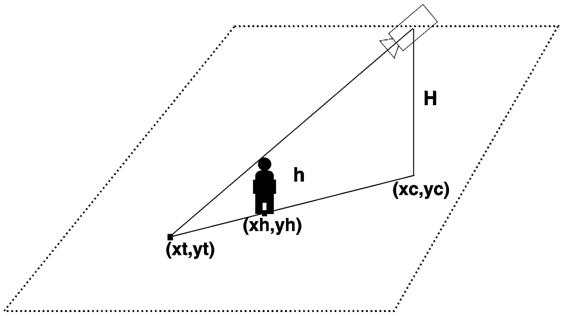


Fig. 2. Target height computation.

ground plane, are known we can compute the target height (see Fig. 2). In fact the camera height and position can be estimated if we know the position and height of two objects in the scene (we used a doorway and a closet).

2.2. Target tracking and initialisation

One of the most important steps in detecting objects in video is to localise, where motion is occurring in a frame. The simplest technique is to use image differencing of consecutive frames of video, to see where motion has occurred. Another, more sophisticated, approach is the use optical flow/image motion algorithms. The information provided by the optical flow algorithms is more detailed than simple change detection.

In each frame a segmentation procedure based on optical flow allows the detection of the possible available targets. A Kalman filter is attached to each detected target and the information returned by the filter is used to predict the location of the target in the next frame. The prediction is used to estimate a bounding box around the expected new position of the target. This bounding box is then compared to the new detected blobs. If a match occurs then the target position is updated. If the uncertainty in position becomes too large over a significant amount of time then the target is considered to be lost and the associated tracking process is terminated. This can occur when the target walks out of the image, is heavily occluded or stops.

When two or more people pass by each other the segmented region could result in one big blob. In this particular case the system tries to recognise each of the individual targets using the predicted bounding boxes. Since we also compute the height of each target, we can use this measurement to certify that the correct match was made.

This problem of the overlapping of the targets on the image also determines the maximum number of targets that the system is able to track. If the number of targets is such that the segmented region is large and the confidence on the trajectory of each previously detected target is low the system is unable, without any kind of additional information, to detect the targets and their trajectories.

One of the problems that sometimes arise with these kind of methods is the presence of shadows on the floor that could lead to an incorrect segmentation of the targets pixels (see Fig. 3). One way to overcome this problem is to rely on the top portion of the blob where segmentation is much more robust to light vari-

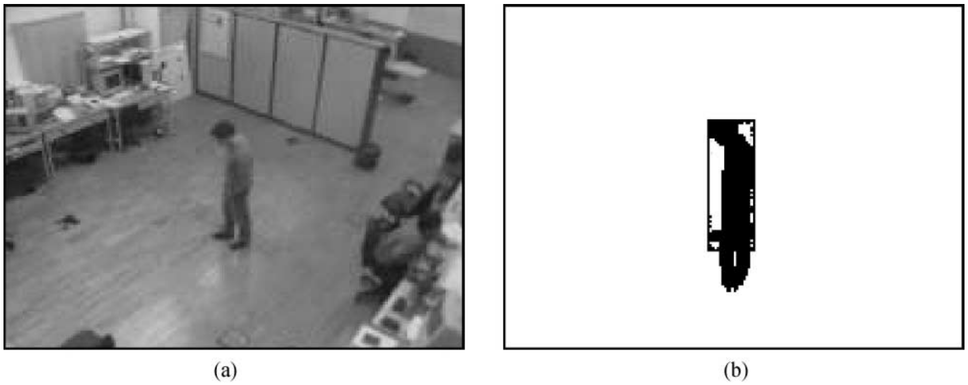


Fig. 3. Dealing with shadows: (a) real image captured by the static camera; (b) result of the segmentation process.

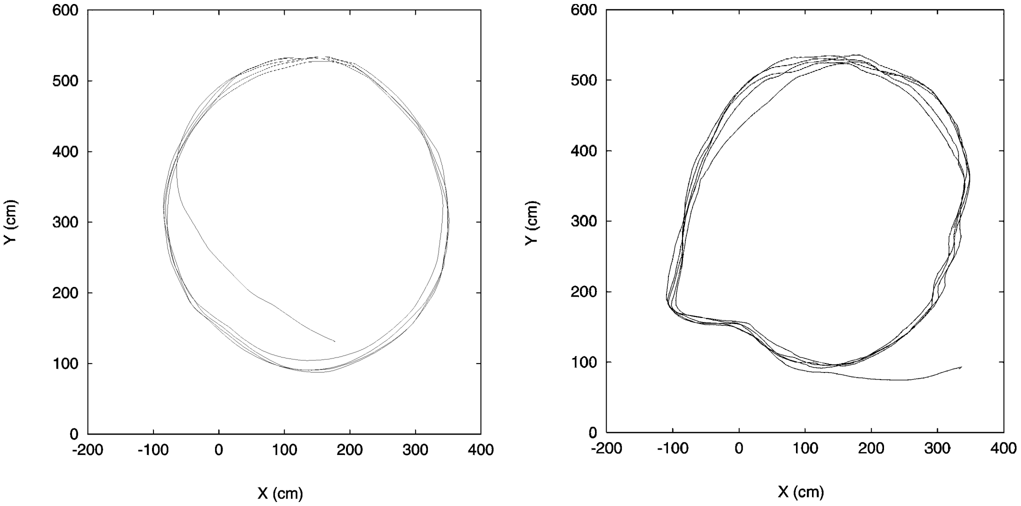


Fig. 4. Example of the influence of the shadows on the 2D mapping of the targets on the ground plane. In these two experiments the subject moved along a circular path. On the right, the result obtained where shadows are included as part of the segmented blob. On the left, the result obtained by ignoring the shadows using the height computation.

ations. If the height of the target is known with a level of confidence above a certain threshold, then the projection of the target on the ground plane can be established using the estimated height of the target. We are then able to compute the number of pixels that the blob should have on the image. Using this kind of approach, we can improve the quality of the target mapping on the ground plane. The example shown in Fig. 4 clearly shows that the presence of a shadow can induce a false mapping on the ground plane. In both examples the subject moved along a circular path. On the right we can see the result obtained by assuming that the shadows are part of the segmented blob. The result path has a clear bias near the spot where the shadow appears (see Fig. 3). The image on the left shows the result obtained by ignoring the shadows using the height computation.

An important aspect of the system is its ability to keep track of the 2D trajectory of the targets on the ground plane. This feature is interesting specially for the posterior reconstruction of the trajectory of each target.

2.3. Evaluation

Some experiments were made to determine the performance of the system regarding both the 2D trajectory estimation and the height estimation.

The entire tracking system is based on a Pentium II 300 MHz computer equipped with an RGB Matrox Meteor frame grabber that captures simultaneously images from the three cameras, avoiding the problems of image synchronisation. The full tracking system (both static camera and the active vision system) run at approximately 25 Hz on 384 × 288 images, including the vision routines, control routines and logging routines.

In the first example three persons, with different heights, walked into the scene. Fig. 5 shows the computed height for each target in each frame. The targets real heights are presented in Table 1.

The height is computed using the process described in Section 2.1. Since we are interested in having an incremental algorithm in which the height value is updated in every frame, we assume that the height estimations follow a normal distribution with parameters μ and σ^2. If h_t is the computed height on time t, we

Table 1
True heights of subjects represented in Fig. 5

Target	Subject A	Subject B	Subject C
True height (cm)	191	176	184

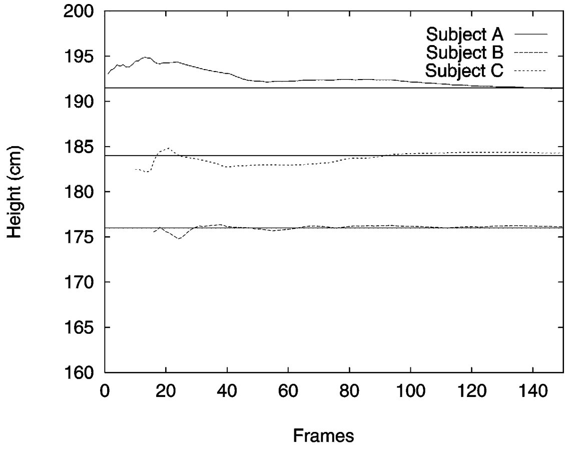


Fig. 5. Results of target height computation.

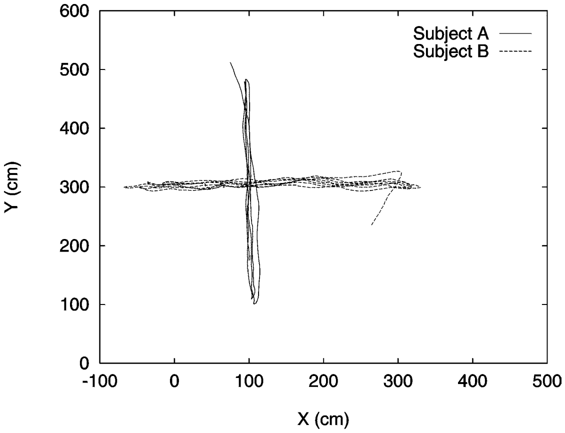


Fig. 6. Results of the 2D mapping of the targets onto the ground plane: subject A moves from left to right and subject B moves back and forth.

update the statistical parameters using

$$\mu_t = \alpha\mu_{t-1} + (1-\alpha)h_t,$$
$$\sigma_t = \alpha\sigma_{t-1}^2 + (1-\alpha)(h_t - \mu_t)^2.$$

The constant α controls the update rate of the statistical information: if α is low then the adaptation is quicker but the learned value can deviate from the true statistics. If α is set high then a sufficient amount of data needs to be gathered for the solution to converge to the correct distribution. We use the statistics to refine future estimations of the height. We placed a *gate* on acceptable values for the new estimations (in practice one can use a 3σ *gate*). Any value larger/lower than $\mu \pm 3\sigma$ is rejected in subsequent updates. The gate has the effect of excluding the outlying height computations (caused for instance by an incorrect segmentation of the target), gradually reducing the estimated value of the height to its true value. As long as the target is visible for a sufficient amount of time this gated adaptation is guaranteed to converge to the correct height value. The system is able to determine the height of the targets after several frames with an average error of 2 cm.

The trajectory of each target on the ground plane is computed and stored for posterior off-line analysis. We performed some experiments in order to evaluate the performance of the system in terms of mapping errors. We show here two examples of those tests. In both cases two persons walked on the scene trying to maintain two straight line trajectories, marked with a tape, on the ground plane. In the first example, one of the persons moved from left to right while the other one moved back and forth (Fig. 6). In the second example both persons moved along a diagonal path (Fig. 7).

2.4. Active vision system visual routines

The active vision system is responsible for pursuit of a specific target in the scene. There should exist some kind of priority scheme in order to choose what

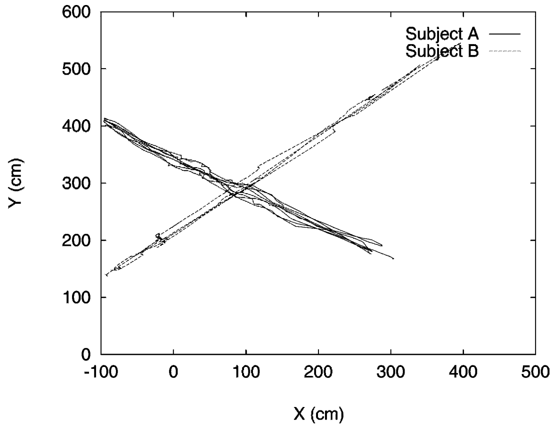


Fig. 7. Results of the 2D mapping of the targets onto the ground plane: both subjects move diagonally.

target to pursue. Of course this priority scheme is dependent on the type of activity in which the target is involved and the relevance of that activity in terms of the application. A priority level is dynamically assigned to each newly detected target. This allows the system to sort the several targets available in the scene according to their priority level. The highest priority target will be the one that will get the attention of the active vision system.

During the tracking process the motion of the active vision system must satisfy two basic requirements:

1. stabilise the images of the selected target on both retinas;
2. maintain fixation on the target.

The tracking task is achieved using two different steps: fixation and smooth pursuit. In the first one the attention of the active vision system is directed to the target with the fastest velocity possible and in the second one the target is tracked [2]. For a more detailed description of the active vision system and for some performance characterisation please refer to [1].

3. Human activity logging

A fundamental problem to be addressed in any surveillance or human activity monitoring scenario is that of information filtering: how to decide whether a scene contains an activity or behaviour worth analysing. Our approach to detection and monitoring of such situations is based on the fact that typically actions are somehow conditioned by the context in which they are produced. For instance the action of opening a closet only makes sense in the near vicinity of a closet.

We assumed the concept of “context cells” to discriminate portions of the scene where any behaviour can be important in terms of the application. It is assumed that a set of rules is known “a priori” and that these rules have the adequate relevance for the purpose of the monitoring task. It is also assumed that these context cells have enough information to trigger a logging event in order to register the actions of the human subjects.

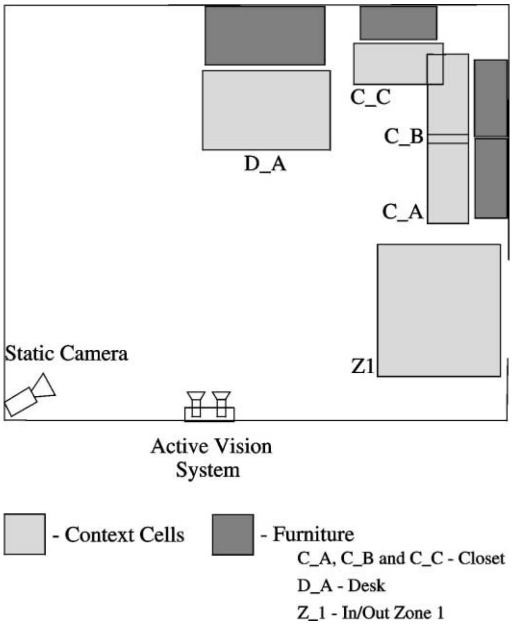


Fig. 8. Context cells definition.

Since in this first approach, we only have as an input the position of the target in the plane and its height we defined a very simple set of context cells in our lab (see Fig. 8). Three different context cells were defined: closets, desks and in/out zones that correspond to areas where targets can enter/exit the scene. The rule used to describe the possible actions in the desk context cell is shown in Fig. 9. Two actions are logged in this case: “near desk” and “seated at desk”.

To establish if a certain target is in a certain context cell, we take into account the time that the target spends on that particular cell. In each cell we define

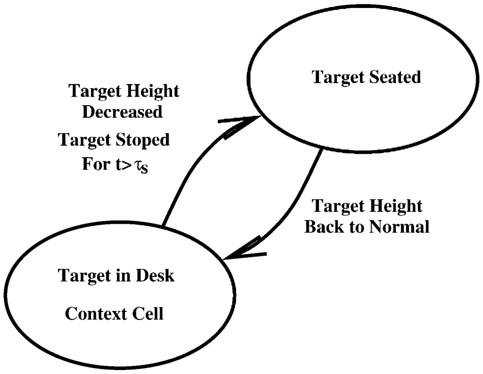


Fig. 9. An example of a rule used to describe an action, in this case the use of a desk.

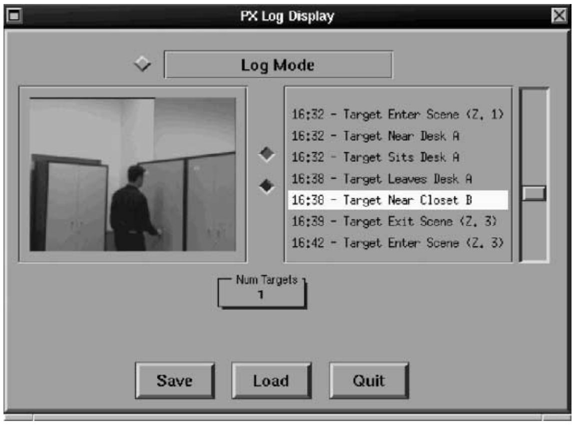
a certain threshold time τ_s that determines the minimum amount of time that the target should stay in the cell in order for the action to be logged. In some cells, such as the case of the in/out cells this time τ_s is equal to zero since in this particular case we are only interested in the event of entering/leaving the scene.

An advantage of this concept based on the context is that it can be used to predict the expected appearance and location of the target in the image, to predict occlusion and initiate processes to deal with that occlusion.

The system creates log files that describe the actions that occurred during a certain period of time. A picture is taken by the active vision system for each recorded action. These images can then be used for posterior analysis and processing, for instance for identification purposes. Fig. 10 shows an example of a portion of a log file recorded by the system.

Different actions require different views in order to understand what is going on. For instance if the target is near a closet, then his hands, not his head, should have the attention of the system. The advantage of the use of the active vision system is that if the "best" view needed for a particular action understanding has not been captured then the active vision system can be instructed to redirect its attention to the right place. Once again the definition of "best" view is context dependent and if this context is known then the search space for the best view can be substantially reduced.

Another aspect of this work (not yet in real-time) is the modelling of more complex behaviours using the same underlying principle. The logging ability can be extended to the detection of more elaborate actions like the detection of opening or closing of closets, and other typical indoor actions.



```
16:32 - Target Enter Scene (Z. 1)
16:32 - Target Near Desk A
16:32 - Target Sits Desk A
16:38 - Target Leaves Desk A
16:38 - Target Near Closet B
16:39 - Target Exit Scene (Z. 3)
16:42 - Target Enter Scene (Z. 3)
```

Fig. 10. An example of a typical human activity log output.

Table 2
Classification results of context cells

Context cell Context cell	Closet A	Closet B	Closet C	Desk A	I/O Zone 1
Classification (%)	91.4	88.6	85.7	94.3	100

Some experiments were made in order to evaluate the performance of the logging module of the system. Several persons walked around in the scenes and their activity in terms of the defined cells were registered by a human operator. At the same time the system was trying to track the same actions. At the end of the experiments the human report was compared with the one produced by the system. A correct match was considered to be one with an exact correspondence between the system and the human operator. Table 2 represents the result obtained in terms of percentage of correct classifications.

The results shown here state that there is in general a good classification of the context cell visited by the targets. The worst results were obtained on the Closets B and C possibly because this two areas overlap slightly and any small error in the 2D mapping of the target could result in a misclassification of the cell.

4. Conclusions

In this paper, we described a real-time system aimed at detecting and tracking targets in man-made environments. This system is based on a global view of the scene and a binocular tracking system. The specific features of the active system enable the tracking of humans while handling some degree of occlusion. The degree of occlusion that can be tolerated depends upon the target distance. Behaviour modelling can advantageously use the 3D trajectories reconstructed both with the data from the global view camera and the data from the active system. Logging of human activity is performed in real time and by analysing changes in that data (changes in height and position) some limited interpretation of action is performed. In addition, the redundancy of the system enables cross-checking of some types of information, enabling greater robustness.

References

[1] J. Batista, P. Peixoto, H.J. Araujo, Visual behaviors for real-time control of a binocular active vision system, Control Engineering Practice 5 (10) (1997) 1451–1461.

[2] J. Batista, P. Peixoto, H.J. Araujo, Real-time visual surveillance by integrating peripheral motion detection with foveated tracking, in: Proceedings of the IEEE Workshop on Visual Surveillance, IEEE Computer Society, January 1998, pp. 18–25.

[3] K.J. Bradshaw, I. Reid, D. Murray, The active recovery of 3D motion trajectories and their use in prediction, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (3) (1997) 219–234.

[4] C. Bregler, Learning and recognizing human dynamics in video sequences, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'97), 1997, pp. 568–574.

[5] H. Buxton, S.G. Gong, Visual surveillance in a dynamic and uncertain world, Artificial Intelligence 78 (1–2) (1995) 431–459.

[6] J.L. Crowley, Coordination of action and perception in a surveillance robot, IEEE Expert 2 (4) (1987) 32–43.

[7] Y. Cui, S. Samarasekera, Q. Huang, M. Greiffenhagen, Indoor monitoring via the collaboration between a peripheral sensor and a foveal sensor, in: Proceedings of the IEEE Workshop on Visual Surveillance, IEEE Computer Society, 1998, pp. 2–9.

[8] J.W. Davis, A.F. Bobick, The representation and recognition of action using temporal templates, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'97), 1997, pp. 928–934.

[9] L. Davis, R. Chellapa, Y. Yacoob, Q. Zheng, Visual surveillance and monitoring of human and vehicular activity, in: Proceedings of DARPA'97, 1997, pp. 19–27.

[10] J. Fernyhough, A.G. Cohn, D.C. Hogg, Building qualitative event models automatically from visual input, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV'98), 1998, pp. 350–355.

[11] D.M. Gavrila, L.S. Davis, 3D model-based tracking of humans in action: A multi-view approach, in: Proceedings of CVPR'96, 1996, pp. 73–79.

[12] E. Grimson, P. Viola, O. Faugeras, T. Lozano-Perez, T. Poggio, S. Teller, A forest of sensors, in: Proceedings of the DARPA Image Understanding Workshop, Vol. 1, New Orleans, LA, 1997, pp. 45–50.

[13] Y. Guo, G. Xu, S. Tsuji, Understanding human motion patterns, in: Proceedings of the IEEE International Conference on Pattern Recognition (ICPR'94), 1994, pp. B325–B329.

[14] T. Kanade, R.T. Collins, A.J. Lipton, P. Anandan, P. Burt, L. Wixson, Cooperative multisensor video surveillance, in: Proceedings of the DARPA Image Understanding Workshop, 1997, pp. 3–10.

[15] D. Murray, K. Bradshaw, P. MacLauchlan, I. Reid, P. Sharkey, Driving saccade to pursuit using image motion, International Journal of Computer Vision 16 (3) (1995) 205–228.

[16] N. Oliver, B. Rosario, A. Pentland, A Bayesian computer vision system for modeling human interactions, in: Proceedings of the First International Conference on Computer Vision Systems (ICVS'99), 1999, pp. 255–272.

[17] M. Yedanapudi, Y. Bar-Shalom, K.R. Pattipati, Imm estimation for multitarget–multisensor air-traffic surveillance, Proceedings of the IEEE 1 (1997) 80–94.

Paulo Peixoto received his B.Sc. degree in Electrical Engineering and M.S. degree in Systems and Automation from the University of Coimbra in 1989 and 1995, respectively. He is currently a Ph.D. candidate in the Department of Electrical Engineering at the University of Coimbra. He is also a member of the Portuguese Institute for Systems and Robotics (ISR), where he is a researcher. His research interests include computer vision, active vision and visual surveillance.

Jorge Batista received the B.Sc. degree in Electrical Engineering from the University of Coimbra in 1986. In 1992 he received the M.S. degree in Systems and Automation and in 1999 he received the Ph.D in Electrical Engineering both from the University of Coimbra. Jorge Batista is currently an Assistant Professor in the Department of Electrical Engineering, University of Coimbra, and a researcher at the Institute of Systems and Robotics. His research interests include camera calibration, computer vision, active vision and visual surveillance.

Helder J. Araujo is currently an Associate Professor in the Department of Electrical Engineering, University of Coimbra. He is a co-founder of the Portuguese Institute for Systems and Robotics (ISR), where he is now a researcher. His primary research interests are in computer vision and mobile robotics.