ELSEVIER

Robotics and Autonomous Systems 31 (2000) 99–108

Robotics and Autonomous Systems

www.elsevier.com/locate/robot

Integration of information from several vision systems for a common task of surveillance

Paulo Peixoto*, Jorge Batista, Helder Araújo

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

Abstract

In this paper we describe the integration of vision information obtained by a fixed camera and the information obtained by a binocular active vision system. Both systems acquire images from the same environment. The information redundancy is exploited to achieve a higher degree of robustness in detecting and tracking intruders. Global information about intruders is obtained by the fixed vision system whereas information about a specific intruder (that invades a critical area, for example) is obtained by the binocular system. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Surveillance; Active vision; Real-time tracking

1. Introduction

Several theoretical and technological developments have enabled the development of vision systems that integrate multiple subsystems. The use of several integrated vision systems is specially relevant from the standpoint of the applications. By using several subsystems with different characteristics different types of information can be extracted from the same environment. The different types of information can be advantageously combined to achieve better performance, a higher degree of robustness and to extract more complex forms of data/information. In visual surveillance, and specially for the monitoring of activities, this type of systems (integrating several vision systems/modalities) is very useful. Several of these systems have been recently described [3,4,9,11,14]. This type of setups use heterogeneous sensing systems that are integrated to provide several advantages.

Surveillance applications have specific features that depend upon the type of environment they are designed to operate on. The scenarios of surveillance applications are also extremely varied [6,8,13,14]. Even though there are some surveillance systems using a single camera [12] most of them use multiple cameras [15]. The use of several cameras has several objectives namely to increase the reliability. 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) [7,10]. In some cases both types of images are acquired but the selection of the region to be imaged by the pan and tilt device depends on the action of a human operator. Wide-angle images have the advantage that each single image is usually enough to cover the entire environment. Therefore any potential intrusion is more

* Corresponding author. Tel.: 351-3979-6275; fax: 351-3940-6672.

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

PII: S0921-8890(99)00080-9

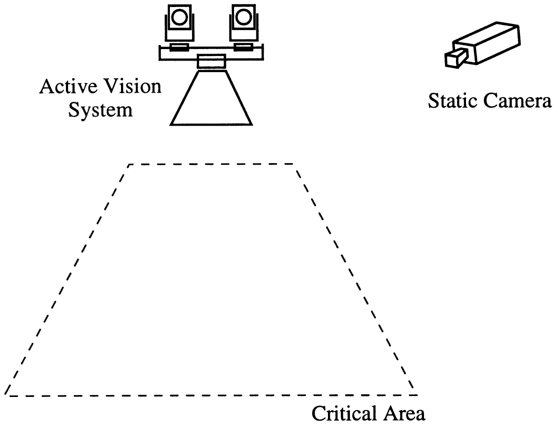


Fig. 1. Integrated visual surveillance system.

easily detected since no scanning is required. Systems based on active cameras usually employ longer focal length cameras and therefore provide better resolution images. Active binocular systems enable the recovery of 3D trajectories by tracking stereoscopically. Proprioceptive data from camera platform can be used to recover depth by triangulation. Trajectories in 3D can also be recovered monocularly by imposing the scene constraint that motion occurs on a plane, typically the ground plane [5].

The system described in this paper is aimed at an application in man-made environments and the detection and analysis of intrusion is one of its main characteristics. It is made up of a static camera and a binocular active system (see Fig. 1). The static camera acquires wide-angle images of the environment enabling the early detection of intrusion. The system associated to the fixed camera can track several moving objects simultaneously using very simple techniques (image differencing and Kalman filters). If one of the intruders approaches a critical area the active system starts its tracking. For that purpose several visual routines were implemented [3]. The active vision system uses optical flow to track binocularly the target in real time. Simultaneously motion segmentation is performed, which enables the extraction of high quality target images. These are useful for further analysis and action understanding. The 3D trajectory of the target is also recovered by using the proprioceptive data.

2. The peripheral vision system

The goal of the peripheral vision system is to command the focus-of-attention of the active vision system in order to control how it will switch its attention between the different targets present in the scene. We define focus-of-attention as the process of using some higher level of knowledge to identify targets of interest and “zoom” in on them iteratively. Any approach to this problem requires (a) the computation of the location of the several subjects present in the scene, and (b) a priori information about the surveillance task required in order to determine who will have the focus-of-attention. In this context the peripheral vision system (using a 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 will have the focus-of-attention of the binocular active vision system.

2.1. Target tracking and initialization

Target segmentation is achieved using a simple differencing scheme. A background image is kept as a reference image and used to describe the stationary portion of the scene. In each frame the acquired image is subtracted from the background, then thresholded. A segmentation procedure allows for the detection of the possible available targets. The assumption is that any differences are presumed to be due to targets. Newly detected targets, not associated with any existing track are initialized and tracked through subsequent frames.

The maximum number of targets that the system is able to track depends on the focal length used and on the static camera position relative to the ground plane. If the number of targets is such that their images overlap simultaneously and if the confidence on the trajectories of the previously detected targets is low, then the system is unable to discriminate the targets. In our indoor experiments, we used a camera positioned 3 m above the ground plane with a 6 mm focal length, and observed that the system performance decreased below 50% of its maximum (in terms of the number of targets consistently tracked).

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. This can occur when the target walks out of the image, is heavily occluded or stops. Currently the system waits for five frames for the target to show up near the predicted position. If that does not happen the track is considered to be invalid and it is terminated.

As previously stated, some kind of higher level information must be assumed in order to assign a priority level to each detected target. The definition of this priority level should be application-dependent and should restrict the location of the active vision system. This priority scheme allows the system to sort the several targets available in the image according to their priority level. The topmost priority target will be the one that will have the focus-of-attention of the active vision system. In the experimental setup used, a very simple scenario was assumed. The priority was defined according to the attitude of the target towards a pre-defined critical zone: a virtual rectangle defined on the ground near an interest zone. Targets leaving this critical area would have a lower priority than those entering the critical area. Targets outside the critical area would have the lowest priority of all.

The image sequence depicted in Fig. 2 exemplifies the tracking process. In this example, three intruders move in front of the active vision system. The peripheral vision system segments and sets up a track process for each intruder. These images of the sequence are separated by time intervals of approximately 500 ms. The trajectories of each intruder along time on the static camera image plane are represented in Fig. 3.

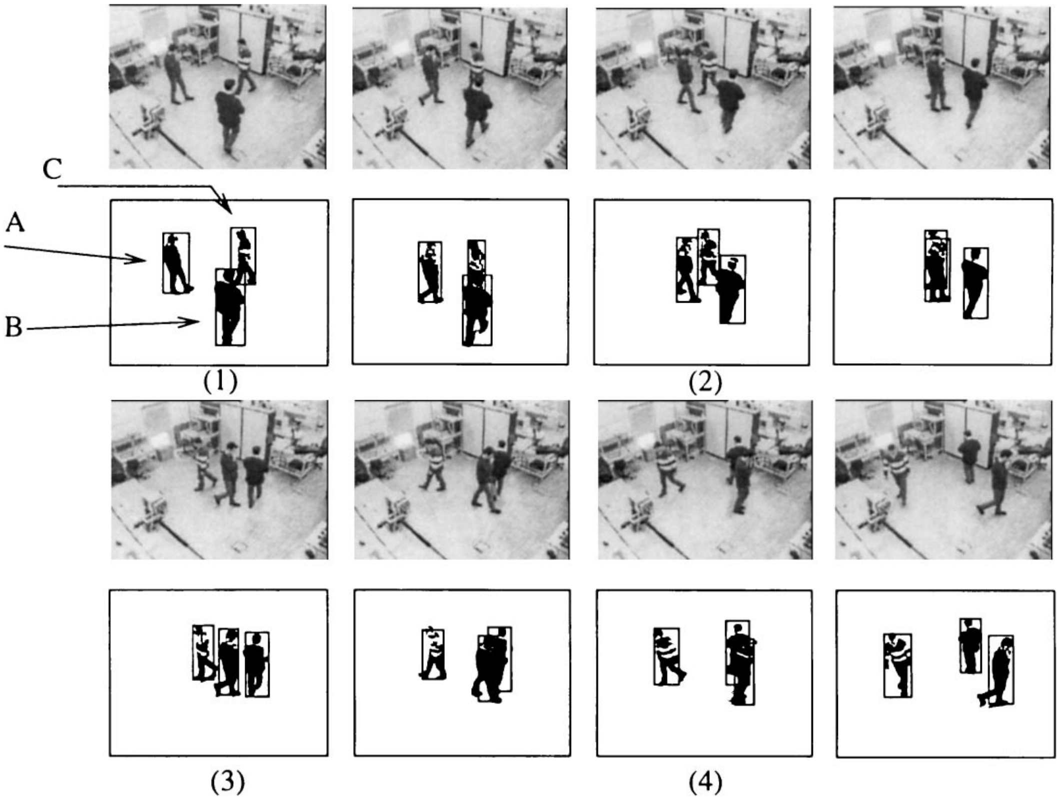


Fig. 2. Eight pictures taken from an image sequence.

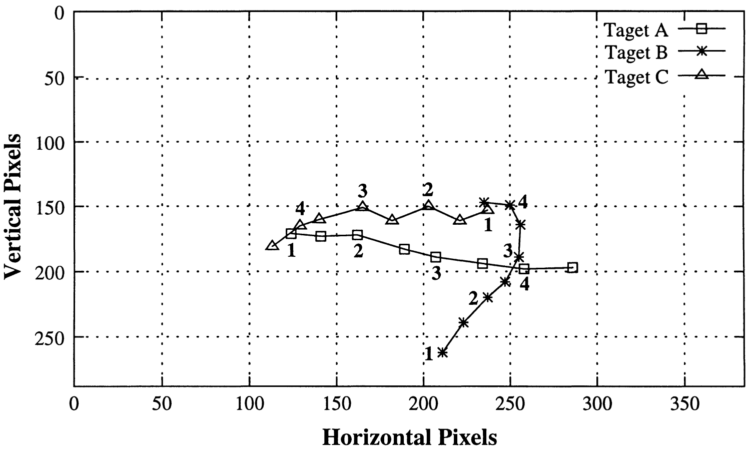


Fig. 3. The trajectories of the 3 targets on the image plane. The numbers reference the corresponding image in the sequence and indicate the temporal evolution of the trajectories.

2.2. 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 vergence (we are assuming that both vergence angles are equal) (see Fig. 4).

Assuming that all target points considered in the static camera image lie on the ground plane then any pair of gaze angles (pan angle and vergence angle)

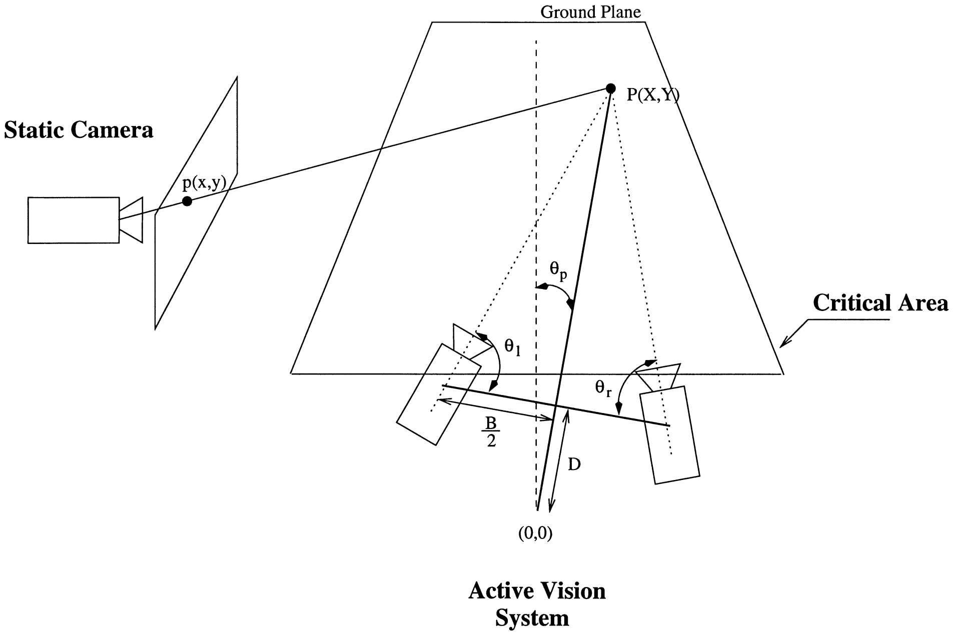


Fig. 4. Correspondence between ground plane points.

corresponding to the fixation into a given point P can be mapped to a point p in the image plane of the static camera using a homography. The relationship between each point $P_i(X_i, Y_i)$ and the pan and vergence rotation angles can be derived directly from the geometry of the active vision system

$$X_i = \sin\theta_{\mathrm{p}_i}\left(\frac{B}{2}\tan\theta_{\mathrm{v}_i} + D\right),$$
$$Y_i = \cos\theta_{\mathrm{p}_i}\left(\frac{B}{2}\tan\theta_{\mathrm{v}_i} + D\right) \qquad (1)$$

with B the baseline distance and $\theta_{\mathrm{v}} = \theta_{\mathrm{l}} = \theta_{\mathrm{r}}$.

Since each target corresponds to a blob and not all the blob pixels are on the ground plane one has to define which pixels should be used to compute the target position on the ground plane. In our case (due to our specific geometric configuration) we use the blob pixels closest to the lower part of the image. This assumption is valid as long as the target is not occluded.

2.3. Mutual cross-checking

Situations of partial target occlusion, and others can be dealt with by using the information available to both systems. If partial occlusion occurs then there will be an error in the mapping of the target position on the ground plane. If the active vision system is directed to track that specific target it may not find it (or it may find it beyond its expected position). In that case the peripheral vision system will be informed of such an occurrence and a new estimate of the position can be computed. Other typical situations are the cases when the active vision system starts tracking a specific target and changes to a different one (due to, for example, a frontal crossing of two intruders). This situation and other that are similar can be accounted for by cross-checking the location of the target on both systems.

At each frame the active vision system reports the coordinates of the actual target to the peripheral system (using inverse mapping). This position is then cross-checked against the detected target position in the peripheral system. If the positional error is above a certain threshold then a new fixation is forced. Cross-checking is possible in our case because both systems are fully synchronized. Since the images are synchronized and time-stamped all the events are referenced in the same time base.

3. Active vision system visual routines

The active vision system is responsible for the pursuit of a specific target on the scene. This 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 and in the second one the target is tracked.

3.1. Fixation

The fixation process is realized using gross fixation solutions, defined as saccades, followed by a fine adjustments in both eyes in order to achieve vergence.

Having the target information given by the peripheral vision system, redirecting gaze can be regarded as controlling the geometry of the head so that the images of the target will fall into the fovea after the fixation. Because of the rapid movement of the MDOF (multi-degree of freedom) head joints during saccades, visual processing cannot be used. Therefore, saccades must be accomplished at the maximum speed or in the minimum possible time. Saccade motion is performed by means of position control of all degrees of freedom involved. The neck and eyes are moved in order to guarantee that the cyclopean eye is looking forward to the target and that both eyes have symmetric vergence. In order to achieve perfect gaze of both eyes in the moving target, and since the center of mass is probably not the same in both retinas for non rigid objects, after the saccade a fine fixation adjustment is performed. A grey level cross-correlation tracker is used to achieve perfect fixation of both eyes.

Some tests have been made in order to measure the ability of the active vision system to precisely fixate on the requested target. Fig. 5 represents the mapping error in terms of pan and vergence angles in function of the position of the target in the peripheral camera image. Only points in the critical area (a rectangle with $4\,\mathrm{m} \times 5\,\mathrm{m}$) were considered. These errors justify the use of a fine fixation adjustment to guarantee vergence necessary for the smooth pursuit process.

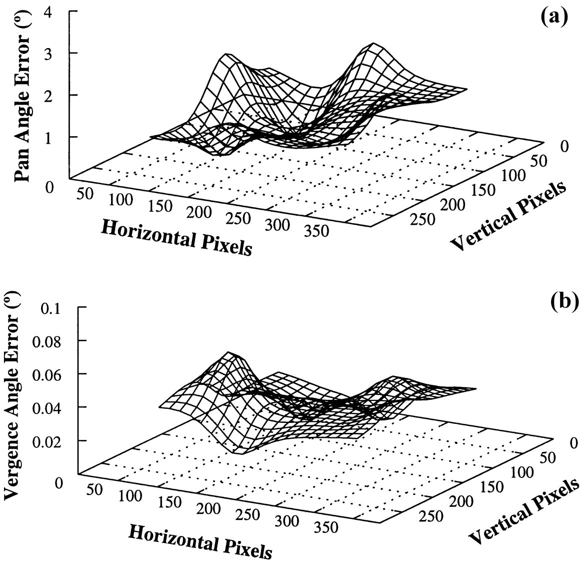


Fig. 5. Pan and vergence angle error.

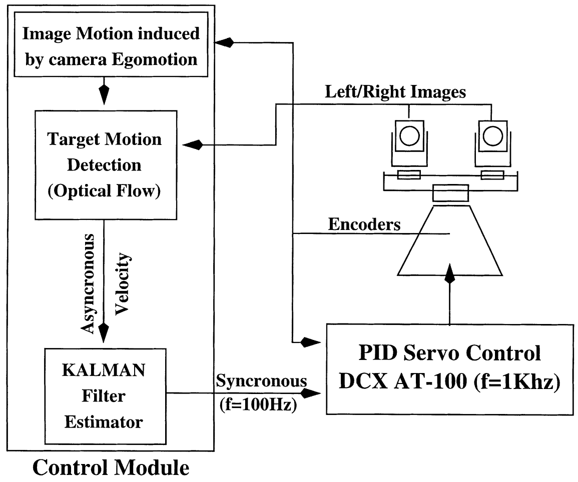


Image Motion induced by camera Egomotion
Left/Right Images
Target Motion Detection (Optical Flow)
Encoders
Asyncronous
Velocity
KALMAN Filter Estimator
Syncronous (f=100Hz)
PID Servo Control DCX AT-100 (f=1Khz)
Control Module

Fig. 6. The active vision system high level gaze controller.

3.2. Smooth pursuit using optical flow

During this process, the motion of the head must satisfy two basic requirements: stabilize the images of the target on both retinas and maintain fixation on the target.

A prerequisite to using pursuit planning is that the target is not far from the fixation point of the head (the images of the target on the retinas must not be far from the center of the foveal window). Otherwise, a saccade must be started prior to the smooth-pursuit process. This means that saccades have higher priority than pursuit. After fixating on the target the pursuit process is started by computing the optical flow. During the pursuit process velocity control of the degrees of freedom is used instead of position control as in the case of the saccade.

Assuming that the moving object is inside the fovea after a saccade, the smooth pursuit process starts a Kalman filter estimator, which takes the estimated image motion velocity of the target as an input. Using this approach, the smooth pursuit controller generates a new prediction of the current image target velocity, and this information is sent to the motion servo controller every 10 ms (see Fig. 6). The smooth-pursuit controller assumes that the moving target is always located on the horopter and the cyclopean eye is pointing straight to the target. With this assumption, the motion induced on the retina by the moving target is almost the same in both eyes [1]. Two Kalman filters were used to filter the estimated image motion velocities and the velocity used to control the neck pan and tilt joints was considered as the average value of both velocities (cyclopean eye velocity). Maintaining the target on the horopter is accomplished by the vergence process.

Fig. 7 shows an example of a frame taken during the smooth pursuit process. Both the static camera image and the images taken by the active vision system can be seen.

3.2.1. The vergence process

To fixate and verge on a target means to keep the images of the target at the image center (center of the fovea). If the image positions of the target on both eyes are known, the 3D position of the target can be recovered using the inverse kinematics of the head.

Based on the fact that the baseline distance is relatively small compared to the target distance, and that the target is not far from the center of the fovea, if the target is moving along the horopter its location is almost the same in both retinas [1]. For this particular situation, within the field of view of the foveal area, the horopter and the circle of equal cyclopean depth are almost coincident, which means that targets moving along the horopter maintain a constant cyclopean

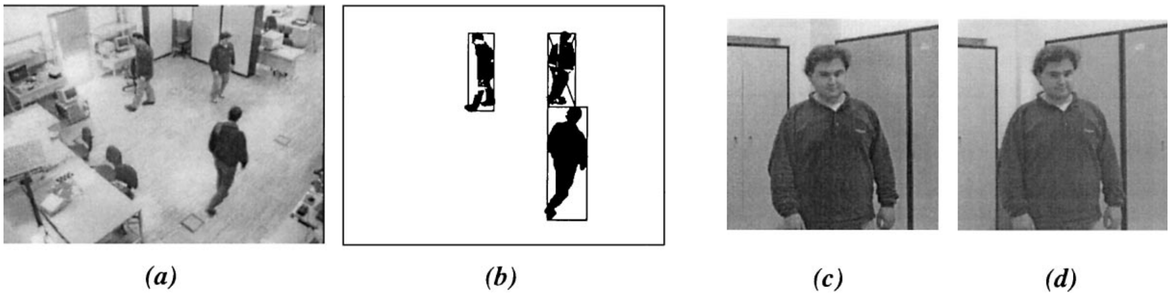


Fig. 7. Target detection and tracking. (a) Image acquired with the peripheral vision system; (b) target segmentation; (c), (d) active vision system left and right images.

depth and no vergence control is required. However, if the target is moving outside the horopter, then its location on the retinas is no longer the same. This displacement is used to control the vergence angles and redirect the head gaze.

Consider the existence of a point P_{c} with coordinates $(X_{\mathrm{c}}, Y_{\mathrm{c}}, Z_{\mathrm{c}})$ in the cyclopean eye coordinate system, moving with velocity $V_{\mathrm{c}} = -\Omega_{\mathrm{c}} \times P_{\mathrm{c}} - t_{\mathrm{c}}$, with $\Omega_{\mathrm{c}} = [\Omega_1, \Omega_2, \Omega_3]^{\mathrm{T}}$ being the angular velocity and $t_{\mathrm{c}} = \left[t_x, t_y, t_z\right]^{\mathrm{T}}$ the translational velocity. This velocity, V_{c}, can be expressed in each one of the retina coordinate systems, $V_{\mathrm{left/right}}$, by

$$V_{\mathrm{l/r}} = -R_{\mathrm{l/r}}\,[\Omega_{\mathrm{c}} \times P_{\mathrm{c}} + t_{\mathrm{c}}], \tag{2}$$

with $R_{\mathrm{l/r}}$ representing the rotation matrix between the cyclopean and the left/right retinal coordinate systems.

Defining the retinal image velocity by

$$\begin{bmatrix} v_{\mathrm{l/r}_x} \\ v_{\mathrm{l/r}_y} \end{bmatrix} = \begin{bmatrix} \frac{f}{P_{\mathrm{l/r}} \cdot \hat{z}_{\mathrm{l/r}}} \left(V_{\mathrm{l/r}} \cdot \hat{x}\right) - \frac{f}{(P_{\mathrm{l/r}} \cdot \hat{z}_{\mathrm{l/r}})^2} P_{\mathrm{l/r}} (V_{\mathrm{l/r}} \cdot \hat{z}) \\ \frac{f}{P_{\mathrm{l/r}} \cdot \hat{z}_{\mathrm{l/r}}} \left(V_{\mathrm{l/r}} \cdot \hat{y}\right) - \frac{f}{(P_{\mathrm{l/r}} \cdot \hat{z}_{\mathrm{l/r}})^2} P_{\mathrm{l/r}} (V_{\mathrm{l/r}} \cdot \hat{z}) \end{bmatrix} \tag{3}$$

and representing $P_{\mathrm{l/r}}$ by $P_{\mathrm{l/r}} = R_{\mathrm{l/r}} P_{\mathrm{c}} + T_{\mathrm{l/r}}$, with $T_{\mathrm{l/r}}$ the translation between the cyclopean and the left/right retinal coordinate systems, the retinal image motion flow disparity is given by

$$\Delta_v = v_{\mathrm{l}} - v_{\mathrm{r}}. \tag{4}$$

After some mathematical manipulation, and considering that the target is verged with equal vergence angles ($\theta = \theta_{\mathrm{l}} = \theta_{\mathrm{r}}$), which means that its coordinates are $P_{\mathrm{c}} = [0, 0, Z_{\mathrm{c}}]$ and its image projections are $[0, 0]_{\mathrm{l/r}}$, the retinal image motion flow disparity is given by

$$\Delta_v = \begin{bmatrix} \Delta v_x \\ \Delta v_y \end{bmatrix} = \begin{bmatrix} \frac{t_z f \sin 2\theta}{Z_{\mathrm{c}}} \\ 0.0 \end{bmatrix} \tag{5}$$

with f being the focal length of both lenses.

For this particular geometry, the horizontal retinal motion disparity allows the computation of *time-to-contact* $Z_{\mathrm{c}}/t_z = (f \sin 2\theta)/\Delta v_x$.

Assuming that the target is verged with equal vergence angles on both retinas, $Z_{\mathrm{c}} = \frac{1}{2} B \tan\theta$, we get for the horizontal retinal motion disparity $\Delta v_x = 4(t_z f \cos^2\theta)/B$, that allows the computation of the Z component of the translational velocity performed by the target, t_z.

Differentiating Z_{c} with respect to time, we get

$$\frac{\partial Z_{\mathrm{c}}}{\partial t} = \frac{B}{2\cos^2\theta} \frac{\partial\theta}{\partial t}. \tag{6}$$

Considering that $\partial Z_{\mathrm{c}}/\partial t = t_z = B \Delta v_x/(4 f \cos^2\theta)$ and replacing $\partial Z_{\mathrm{c}}/\partial t$ in Eq. (6) we get

$$\frac{\partial\theta}{\partial t} = \frac{\Delta v_x}{2f} \tag{7}$$

which represents the angular velocity of the vergence joints to maintain vergence on the moving target. For this particular geometry, only the horizontal motion flow disparity is required to control the joints vergence velocity of both retinas.

Since this process only guarantees that the target center of mass is located in the center of the fovea,

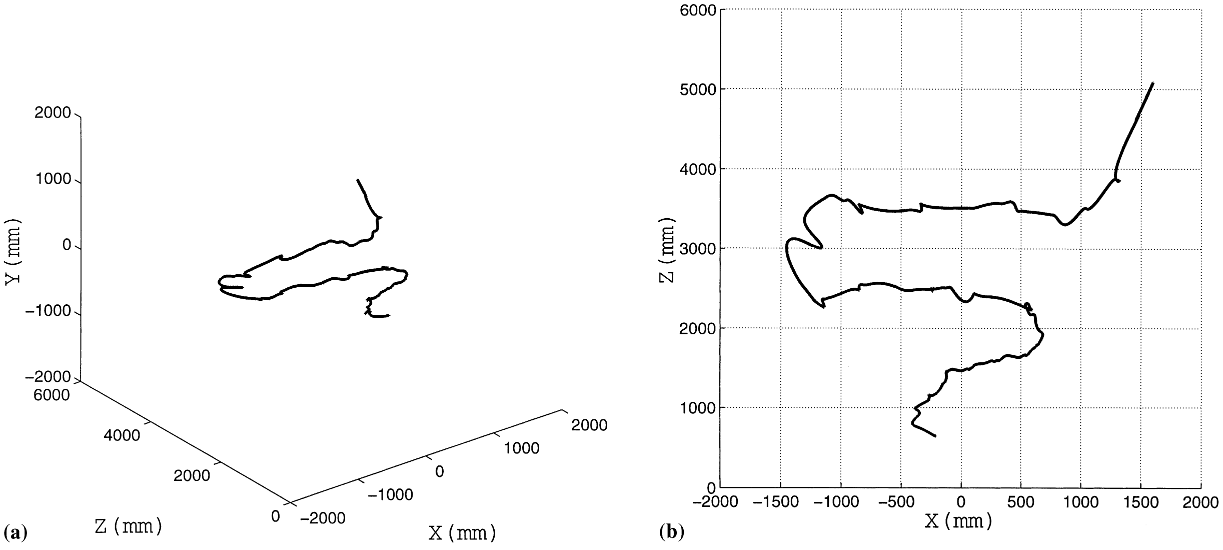


Fig. 8. 3D target motion obtained with the proposed vergence and smooth-pursuit process. The surface plots represent the top-view of the volume plot, showing the depth movement of the target.

we use a grey level cross correlation to adjust vergence and adjust the target depth, at a sample rate 10 times smaller (every 10th frame). The target depth is used to control the auto-focusing of both eyes, taking advantage of the pre-calibration of the focusing depth [2]. Fig. 8 shows the target depth obtained by triangulation using the proposed binocular vergence process.

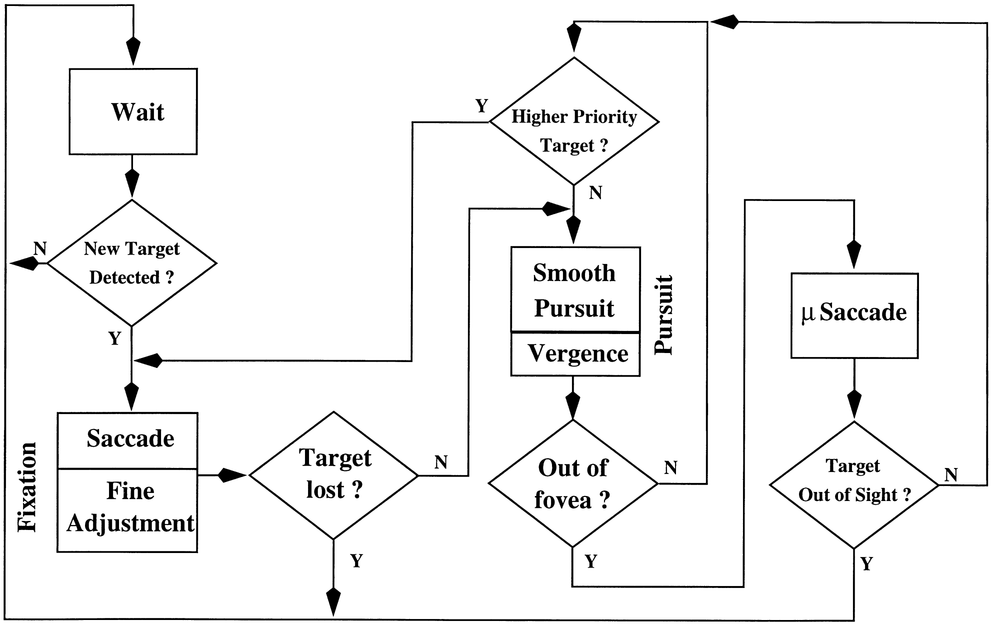


Fig. 9. State transition system.

4. The global controller strategy

The strategy adopted by the gaze controller to combine saccade, smooth pursuit and vergence to track moving objects using the active vision system was based on a *State Transition System*. This controller defines four different states: *Waiting, Fixation (Saccade), Smooth Pursuit and μSaccade*. Each one of these states receives control commands from the previous state, and triggers the next state in a closed loop transition system. The peripheral vision system is used as the supervisor of the active vision system, triggering the transition between the different states. The global configuration of this state transition system is show in Fig. 9.

5. Conclusions

In this paper we described a surveillance system that integrates the information from two different vision systems in order to achieve a higher degree of robustness. Since the active vision system is only able to track a single target at each instant the peripheral sensor enables the monitoring of several moving objects by acting as a supervisor. Due to the mapping of the intruders positions into a common reference plane switching of the focus of attention is possible according to some scheme for priority assignment. By cross-checking the location of the target on both systems the robustness of the system can be improved since errors introduced by partial occlusion of the target can be overcome.

References

[1] J. Batista, P. Peixoto, H. Araújo, Real-time vergence and binocular gaze control, in: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'97), Grenoble, France, September 1997.

[2] J. Batista, P. Peixoto, H. Araújo, Real-time visual behaviors with a binocular active vision system, in: Proceedings of the IEEE International Conference on Robotics and Automation, Albuguergue, NM, April 1997.

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

[4] R.C. Bolles, K.G. Konolige, M.A. Fischler, Extra set of eyes, in: Proceedings of the DARPA Image Understanding Workshop, New Orleans, LA, Vol. 1, 1997, pp. 41–44.

[5] 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.

[6] M.J. Carlotto, Detection and analysis of change in remotely-sensed imagery with applications to wide area surveillance, Image Processing 6 (1) (1997) 89–202.

[7] D. Coombs, C.M. Brown, Real-time binocular smooth pursuit, International Journal of Computer Vision 11 (2) (1993) 47–165.

[8] J. Crowley, Coordination of action and perception in a surveillance robot, IEEE Expert 2 (1987).

[9] 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, January 1998, pp. 2–9.

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

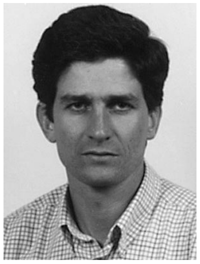
[11] E. Grimson, P. Viola, O. Faugeras, T. Lozano-Perez, T. Poggio, S. Teller, A forest of sensors, in: Proceedings of the DARPA Image Understanding Workshop, New Orleans, LA, 1997, Vol. 1, pp. 45–50.

[12] D. Hogg, Model-based vision: a program to see a walking person, Image and Vision Computing 1 (1983) 5–20.

[13] R. Howarth, H. Buxton, Visual surveillance monitoring and watching, in: Proceedings of ECCV96-II, 1996, pp. 321–334.

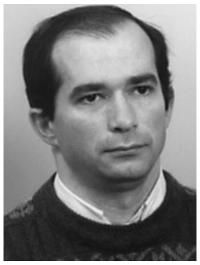
[14] T. Kanade, R.T. Collins, A.J. Lipton, P. Anandan, P. Burt, L. Wixson, Cooperative multisensor video surveillance, in: Proceedings of the DARPA Image Understanding Workshop, 1997, pp. 3–10.

[15] B. Rao, H. Durrant-Whyte, A decentralized bayesian algorithm for identification of tracked targets, IEEE Transactions on Systems, Man and Cybernetics 23 (6) (1993) 1683–1698.



Paulo Peixoto received the B.Sc. degree in Electrical Engineering, and the 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 a 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, and the M.S. degree in Systems and Automation from the University of Coimbra in 1986, and 1992, respectively. He is finishing his Ph.D. degree in Electrical Engineering at the University of Coimbra. He is a founding member of the Portuguese Institute for Systems and Robotics (ISR), where he is now a researcher. His research interests include camera calibration, computer vision, active vision and visual surveillance.



Helder Araújo is currently an Associate Professor in the Department of Electrical Engineering, University of Coimbra. He is 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.