Visual Behaviours for Binocular Tracking *

Alexandre Bernardino José Santos-Victor
Instituto de Sistemas e Robótica
Instituto Superior Técnico
1096 Lisboa Codex, Portugal
{ajmb,jasv}@isr.ist.utl.pt

Abstract

This paper presents a binocular tracking system based on the integration of visual behaviours. Biologically motivated behaviours, Vergence and Pursuit, cooperate as parallel, complementary and highly coupled processes in the tracking system, simplifying the acquisition of perceptual information and system modeling and control. The use of a space variant image representation and low-level visual cues as feedback signals in a closed loop control architecture, allow real-time and reliable performance for each behaviour, despite the low precision of the algorithms and modeling errors. The behaviours are integrated and the overall system is implemented in a stereo head running at real-time (12.5 Hz), without any specific processing hardware. Results are presented for objects of different shapes and motions, illustrating that tracking can be robustly achieved by the cooperation of purposively designed behaviours, tuned to specific subgoals.

1 Introduction

Many computer and robot vision problems can improve their performance by tracking objects in the visual field. These include trajectory reconstruction [3], egomotion estimation [10], object recognition [6] and navigation, which are important subjects in current robotics research.

In structured conditions, where objects have constrained shapes or motions, or move in simple backgrounds, some tracking systems have been successfully employed [13, 1]. However, in complex and dynamic environments, where restrictions about the kind of objects or motions cannot be made, most systems lack robustness, reactivity and flexibility.

In nature, biological systems interact constantly with unstructured environments, and show high reliability in executing most of the tasks that we would like to implement in artificial systems. This fact led some researchers to search for motivation on biological systems that could help on the design of the required solutions [19].

The tracking system presented in this paper is implemented on an active vision stereo head (Medusa [20]) and follows some existing solutions on the visual system of humans and other animals. These solutions are applied to the current problem showing several advantages over other more straightforward approaches.

Binocularity: Although many researchers have faced the tracking problem as a monocular one [15, 16], binocularity can provide major benefits. Not only it is possible to recover depth from the tracked object but also binocular fusion greatly simplifies figure-ground segmentation [8].

Ocular Movements: The most influent ocular movements in biological systems are saccadic, vergence and smooth pursuit movements [18]. Saccadic and smooth pursuit movements use retinal position and velocity errors, respectively, in order to compensate lateral motion of the target. Vergence movements, instead, are mainly based on disparity cues to compensate motion in depth. This decomposition leads to a simplification of the sensorimotor processes, which proved useful in the present application.

Space Variant Image Resolution: One of the main reasons for the existence of ocular movements is the space variant density of photo-receptors in the retina, distinguishing between a central high resolution area – the fovea – and the low resolution peripheral zone [7]. Combining this feature with the ability to move the eyes over the visual field, it is possible to obtain the information needed to perform the desired tasks, without considering all the data contained in uniformly sampled images. Hence, the processing effort is concentrated in the foveal area.

Two main visuomotor behaviours are developed and integrated in the tracking system: the vergence behaviour and the pursuit behaviour. The vergence behaviour acquires depth cues from the stereo images and controls camera angles in order to keep the target in the center of the images, despite motions in depth. The pursuit behaviour extracts target position and velocity errors from the stereo images and controls the observation direction. These behaviours are separable as they acquire different stimuli and control different motions but are highly coupled in the sense that each one depends on the performance of the other.

In both behaviours, visual sensing is made on log-polar images. The log-polar mapping [21] provides a non-uniform resolution geometry which is similar to the distribution of photo-receptors in the human retina, resulting in both algorithmic and perceptual advantages over the usual cartesian representation [3, 4].

The focus of this paper is on the design, integration

*Partially funded by projects PRAXIS/3/3.1/TPR/23/94 and JNICT-PBIC/TPR/2550/95.

and test of these behaviours. The design includes sensor and motor (control) aspects, but special attention is given to sensoring strategies. Control issues are currently well established and can be formalized in several ways. The visual servoing approach [9] provides ways for motor control with visual feedback from position or pose of known objects in simple backgrounds. However, under more general conditions and complex environments, visual sensing strategies often fail to provide adequate inputs to the control system.

2 Visuomotor Behaviours

When designing a visual servoing system we are faced with two main problems:

- What is the relevant perceptual information to extract from the images? (the sensor problem).
- How can this information be used to control the motion of our device? (the motor problem).

We define a visuomotor behaviour as a process that deals with the sensor problem and the motor problem as one. Sensing strategies and control strategies are purposively designed to the task at hand and may suffer constraints from each other. Each visuomotor behaviour extracts only the relevant information and controls only the degrees of freedom needed for a particular purpose. A coordination module may be required to integrate the contribution of each behaviour in the motion of the device (Figure 1).

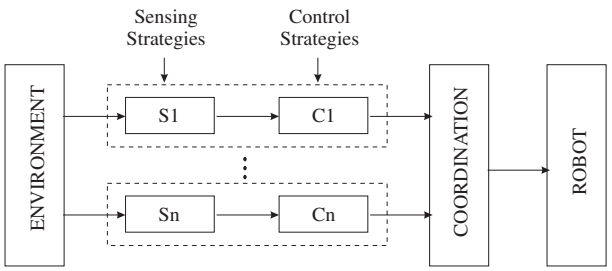

Figure 1: The visuomotor behaviours (inside dashed boxes) are composed by a sensing strategy and a control strategy.

3 Vergence Behaviour

The purpose of the vergence movements is to control the angle between the optic axes, in such a way that a fixation point along some specified gaze direction is kept at the center of the visual field. When this is the case, we say that the target is "foveated". Figure 2 illustrates the vergence angle ν, in a binocular system.

As the vergence angle is directly related to the target distance, depth cues can be used to help the vergence system in compensating depth changes. Depth cues such as binocular disparity and accommodation are important inputs to the vergence control system, and humans rely also on other cues such as motion and shading [7].

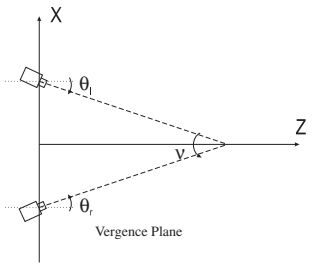

Figure 2: Stereo cameras verging at a point. The vergence angle, $\nu = \theta_l + \theta_r$, is shown.

Currently, only target disparity measures are used to control the vergence angle. For symmetrical vergence ($\theta_l = \theta_r$), small variations in vergence angle have an approximate linear relation with disparity for a large range of depth [3]. Thus, a variational kinematics model for vergence is given by:

$$\delta\nu \simeq \gamma_d \cdot d \tag{1}$$

where d is the target horizontal disparity and γ_d depends on system calibration parameters.

Traditional algorithms to calculate explicit disparity measures include cepstral filtering [17] and phase correlation [11], but these techniques are very time consuming, and therefore, not very adequate for real-time applications. By relying on closed loop control, it is still possible to achieve vergence with low precision algorithms and coarse system calibration.

3.1 Sensing Strategy

In order to obtain estimates of target disparity, correlation between the stereo images is used. The Sum of Squared Differences [2] is a frequently used measure of similarity, given by:

$$SSD\left[I_1\left(i,j\right), I_2\left(i,j\right)\right] = \sum_{i,j}\left[I_1\left(i,j\right) - I_2\left(i,j\right)\right]^2$$

and is minimal for the higher similarity between images. If d_0 is the average amount of horizontal disparity between images, then we expect the correlation function:

$$corr\left(I_1, I_2; d\right) = SSD\left[I_1\left(i + \frac{d}{2}, j\right), I_2\left(i - \frac{d}{2}, j\right)\right]$$

to have its minimum for $d = d_0$. Thus, disparity estimation can be considered as the search of the value d for which the correlation function is minimal. This search can be restricted to a discrete set of disparities, $D = \{d_1, ..., d_n\}$, whose cardinality depends on the requirements of precision versus computation time for the desired application. Figure 3 shows a block diagram of the sensoring process, composed by a set of channels that translate the current images by a pre-determined disparity and compute the correlation value for those images. The estimated disparity is given by the channel whose output is minimal.

Assuming smooth trajectories for the target motion, disparity remains small most of the time, and the set

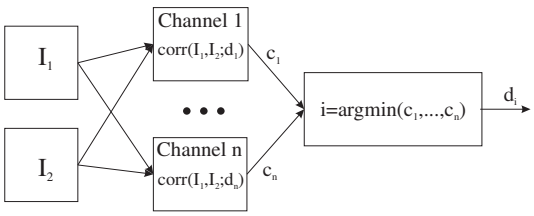

Figure 3: Vergence sensing strategy.

D should cover small disparities in a dense manner. To cope with faster displacements, large disparities should also be considered but in a coarse manner, such that the required computational power is kept within limits[1]. In situations of large displacements, convergence is achieved in two phases: initially, the vergence angle is roughly attracted to the correct angle and after that, a fine adjustment is performed.

For small objects, the existing background mismatches between the two images can lead to low values of similarity even in correct vergence situations. The non-uniform resolution representation is an elegant solution to this problem. Having an higher resolution in the center of the images, where the target is expected to be, the areas belonging to the target are dominant in the computation of the correlation index, reducing the negative influence of background elements [3, 5].

3.2 Control Strategy

The disparity estimates obtained by the sensing strategy are integrated in a standard feedback architecture. They are acquired at each time step and used as input to a motor strategy. Its output is used as position commands to motor servoing control.

Considering that motor position servoing is much faster than the visual sampling frequency (about 20x), the discrete dynamic model for the vergence angle can be approximately described by:

$$\begin{aligned} \delta\nu\,(i+1) &= \delta\nu\,(i) + u\,(i) \\ y\,(i) &= \tfrac{\delta\nu(i)}{\gamma_d} \end{aligned}$$

where i is the discrete time variable, u is the position reference given to motor servoing control and y is the system output, given by Equation (1).

In the current implementation, a simple proportional controller is used in the motor strategy:

$$u\,(i) = -k \cdot y\,(i)$$

and, introducing Equation (1), the closed loop system description is given by:

$$d\,(i+1) = \left(1 - \frac{k}{\gamma_d}\right) \cdot d\,(i)$$

which is typical of a first order system.

[1] $D = \{0, \pm1, \pm2, \pm3, \pm5, \pm8, \pm12, \pm17, \pm23\}$ *pixels*, in the present case.

4 Pursuit Behaviour

In the last section, we saw how fixation depth along the gaze direction is controlled by the vergence behaviour. In this section we are concerned in controlling the gaze direction by controlling the pan (θ_p) and tilt (θ_t) angles which correspond to the azimuth and elevation of the gaze direction. Figure 4 illustrates this geometry.

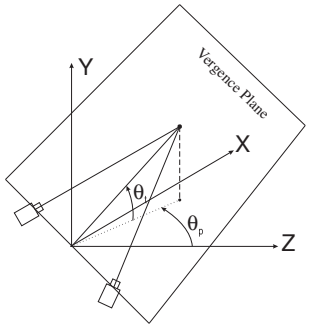

Figure 4: Gaze direction is controlled by pan and tilt motors

We assume that for a centered target, small variations of pan and tilt angles produce linear changes in image plane position, which a good approximation for the range of tilt angles attainable by the stereo head Medusa, $(-45^o, 45^o)$ [3]:

$$\begin{cases} \delta\theta_p \simeq \gamma_x\ x \\ \delta\theta_t \simeq \gamma_y\ y \end{cases} \tag{2}$$

In biological visual systems, pursuit movements are accomplished by saccadic and smooth pursuit movements that respond to target position and velocity errors in the retina [7]. For our purposes, we consider the pursuit behaviour composed by a sensoring module that extracts from the images target position and velocity estimates, and an appropriate control module. The next sections focus on these problems.

4.1 Target Segmentation

To measure target position and velocity in the image planes, the figure-ground segmentation process must be addressed. We should be able to discriminate between target points and points belonging to the background or other objects in the visual field. When the target is verged, its location in each of the stereo images is approximately the same, and the segmentation problem can be addressed by zero disparity filtering [8]. Algorithms for the extraction of zero disparity points in the images are developed based on vertical gradient matching and care is taken to cope with the log-polar representation of images. Figure 5 shows the output of the segmentation process in a pair of real images.

4.2 Motion estimation

Once points belonging to the target have been extracted, we can estimate its motion. Position and velocity estimates are taken from the segmented areas and, although obtained from log-polar images, cartesian coordinates will be used for their representation,

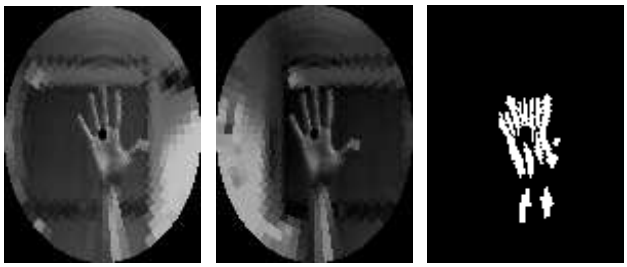

Figure 5: Figure-ground segmentation using zero disparity filtering in log-polar images. From left to right: left camera image, right camera image and filter output.

because pan and tilt angles depend directly in the horizontal and vertical cartesian coordinates of target motion, as expressed by Equation (2).

Position Error The position error is defined by the difference between the position of the object and the coordinates of the center of the image. However, this definition is ambiguous for objects occupying more than one point in the image. Thus, we use the centroid of the points belonging to the object to determine its position:

$$\overline{\mathbf{p}} = \frac{\iint\limits_{x,y} I(x,y)\ \mathbf{p}\, dx\, dy}{\iint\limits_{x,y} I(x,y)\ dx\, dy}$$

where $\mathbf{p}$ are position error coordinates.

Velocity Error Velocity estimation could be simply done by differencing the position data. However, position estimates have a considerable amount of noise which would be amplified by this strategy. Other well known technique for velocity estimation is the optical flow [14] which provides velocity estimates for each point in the image. The integration of the normal flow vectors[2] for each point belonging to the target, provides a stable velocity measure and good noise rejection. Thus, our sensoring strategy relies on the computation of the normal flow vectors in the segmented area, $\mathbf{v} = (v_x, v_y)$, followed by the computation of their average:

$$\overline{\mathbf{v}} = \frac{\iint\limits_{x,y} \mathbf{v}\ dx\ dy}{\iint\limits_{x,y} dx\ dy}$$

4.3 Pursuit Controller

Target position and velocity estimates are used as the input to a dynamic controller. We assume that head dynamics is negligible (which happens if motor drivers have high disturbance rejection). Thus, dynamic effects are due only to motor dynamics, which we assume to be first order. To have smoother trajectories, motors will be controlled in velocity mode. With this assumptions, and introducing Equation (2), a variational continuous state space model for the pan/tilt joints is given by:

$$\dot{\mathbf{x}}_{p/t}(t) = \begin{bmatrix} 0 & 1 \\ 0 & -1/\tau_{p/t} \end{bmatrix} \mathbf{x}_{p/t}(t) + \begin{bmatrix} 0 \\ 1/\tau_{p/t} \end{bmatrix} u_{p/t}(t)$$
$$\mathbf{y}_{p/t}(t) = \begin{bmatrix} \frac{1}{\gamma_{x/y}} & \frac{1}{\gamma_{x/y}} \end{bmatrix} \mathbf{x}_{p/l}(t)$$

where $\mathbf{x}_{p/t}(t) = \begin{bmatrix} \delta\theta_{p/t}(t) & \delta\omega_{p/t}(t) \end{bmatrix}$ is the state variable composed by pan/tilt motor position and velocity, $u_{p/t}(t)$ is the velocity command to pan/tilt motors and $\tau_{p/t}$ is the time constant of pan/tilt motor dynamics.

Now, discretization of this model can be made, and proceed with the design of the state estimator, $\widehat{\mathbf{x}} = E(\mathbf{y}, \overline{\mathbf{p}}, \overline{\mathbf{v}})$, and the control law, $u = C(\widehat{\mathbf{x}})$. In the present case, a simple pole-placement technique was used but other techniques are available in the literature [12].

4.4 Coordination

In this section we have presented the visuomotor behaviours that compose the tracking system. The vergence behaviour and the pursuit behaviour extract different measures from the images and control different motor actions, requiring a low coordination effort. Basically, the two behaviours run in a parallel fashion most of the time, and the only internal dependency arises when the vergence behaviour can no longer guarantee correct vergence in the target. In this situation, the output of the segmentation process is not reliable and motion estimates cannot be extracted. The vergence behaviour inhibits the pursuit behaviour. Despite this low internal dependency, vergence and pursuit are highly coupled in an external sense. First, pursuit brings the target to the central area of the images, enhancing the performance of the vergence behavior. Second, vergence provides binocular fusion on the target, enabling good target segmentation and motion estimation, required for proper pursuit. In this sense, the tracking problem is an example of how separable behaviours, tuned to different goals, cooperate towards a common purpose.

5 Results

In this section we show the results of some experiments made with the proposed tracking system. The system runs at 12.5 Hz (half video rate) and all processing routines are implemented in a PENTIUM-200 computer. To illustrate the reliability and generality of the system we show results for targets with different shapes and motions.

In a first set of experiments, the target moved with constant velocity in a linear trajectory. The position of all joints was collected for all time steps, allowing to reconstruct the target trajectory in a qualitative way[3]. Figures 6, 7 and 8 show the time evolution of each of the joint angles and the 3D trajectory reconstruction. Figure 6 corresponds to a motion in depth. As expected

[2] Normal flow is the projection of the optical flow vector in the direction of the image gradient. Is the only observable component of the optical flow due to the aperture problem [14].

[3] system calibration and target motion were made in a coarse fashion.

only the vergence angle has significant variations. In Figure 7 the target moves horizontally in a direction perpendicular to the initial gaze direction. Most of the tracking motion is achieved by the pan joint. Figure 8 shows a diagonal trajectory where tracking requires the contribution of all joints. In all these cases, the reconstructed trajectory is close to the expected, in a qualitative sense.

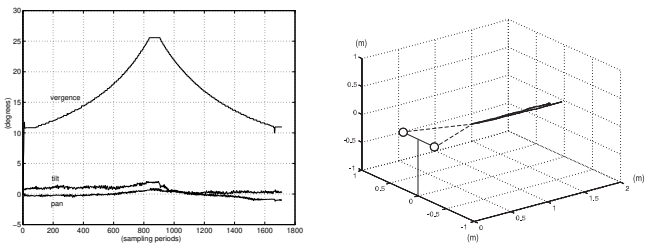

Figure 6: Depth tracking. Time evolution of the joint angles (left) and trajectory reconstruction (right).

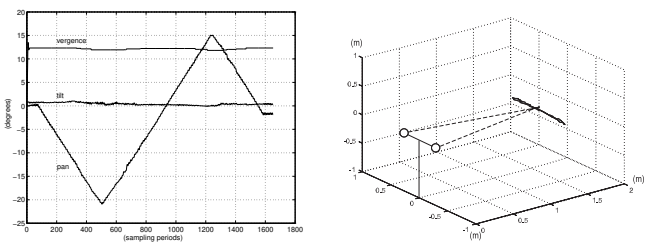

Figure 7: Fronto-parallel tracking. Time evolution of the joint angles (left) and trajectory reconstruction (right).

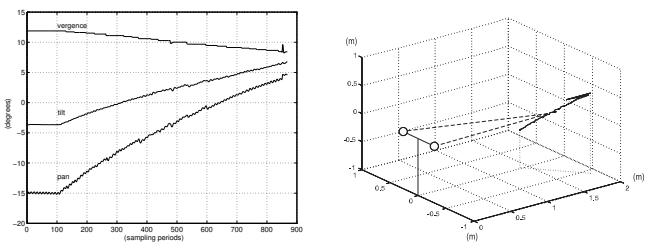

Figure 8: Diagonal tracking. Time evolution of the joint angles (left) and trajectory reconstruction (right).

A second set of experiments was made with unknown motion for the target. In this case results are presented by a sequence of images obtained during the tracking process. A good performance can be observed whenever the target is kept close to the center of the images. Both sequences are presented in Figure 9. In the first sequence we can observe the tracking of a human hand. In the beginning, fixation is stable in the door at the back of the room (images 1 and 2). Once the hand gets close to the gaze direction, the system starts verging and tracking the target. Notice that the hand is always kept very close to the center of the images despite background motion and target rotation and scaling. The second sequence shows a reliable tracking of a human face with scaling and pose changes.

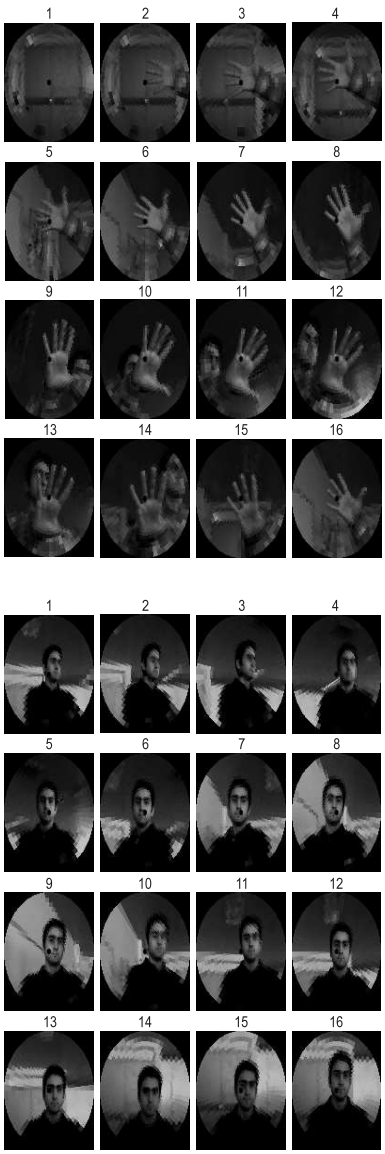

Figure 9: Hand tracking (left) and face tracking (right).

6 Conclusions

Tracking objects in the environment is an important subject in robotics research. This paper presented a real-time tracking system which copes with objects of different shapes and motions.

The approach is centered in the cooperation of two biologically motivated visuomotor behaviors (vergence and pursuit). Vergence performs tracking in depth along the current gaze direction and pursuit controls the gaze direction using motion estimates from the verged object. Each behaviour is composed by sensing

strategies and motor strategies. Sensing strategies use log-polar geometry to represent images, which proved adequate for the extraction of visual stimuli, for tracking purposes. This geometry focus attention in the center of the images, where the target is usually located, and distracting elements in the periphery become less influent in the overall system performance. The sensing algorithms, based on image correlation, optical flow and disparity filtering, do not assume any previous knowledge about target shape or motion. Furthermore, log-polar mapping reduces the size of the images, allowing faster algorithms and real-time functionality. Motor strategies are based on simplified behaviour models and a feedback control architecture, which compensates for the low precision of the sensing algorithms and modeling errors, and produces smooth motions even in the presence of considerable amounts of sensor noise. Integration and coordination of the visual behaviours is simplified by their separable nature. Separate stimuli and motions are addressed by each behaviour allowing low coordination effort. Their close external coupling permits a robust cooperation in a great diversity of situations and conditions – each behaviour enhances the perceptual requirements of the other.

References

[1] R. Andersson. Dynamic sensing in a ping-pong playing robot. *IEEE J. Robotics and Automation*, 5(6):728–739, December 1989.

[2] P. Aschwanden and W. Guggenbühl. Experimental results from a comparative study on correlation–type registration algorithms. In W. Förstner and St. Ruwiedel, editors, *Robust Computer Vision*. Wichmann, 1992.

[3] A. Bernardino. Seguimento binocular de alvos móveis baseado em imagens log-polar. Master's thesis, Instituto Superior Técnico, Lisbon, Portugal, December 1996.

[4] A. Bernardino and J. Santos-Victor. Sensor geometry for dynamic vergence: characterization and performance analysis. In *Proc. of the Workshop on Performance Characteristics of Vision Algorithms*, Cambridge, UK, April 1996.

[5] A. Bernardino and J. Santos-Victor. Vergence control for robotic heads using log-polar images. In *Proc. of the 1996 IEEE/RJS International Conference on Intelligent Robots and Systems*, pages 1264–1271, Osaka, Japan, November 1996. IEEE Computer Society Press.

[6] K. Brunnström, J. Eklundh, and T. Uhlin. Active fixation for scene exploration. *Int. Journal of Computer Vision*, 17(2):137–162, February 1996.

[7] R. Carpenter. *Movements of the eyes*. Pion, London, 1988.

[8] D. Coombs and C. Brown. Real-time binocular smooth pursuit. *Int. Journal of Computer Vision*, 11(2):147–164, October 1993.

[9] B. Espiau, F. Chaumette, and P. Rives. A new approach to visual servoing in robotics. *IEEE J. Robotics and Automation*, 8(3):313–326, June 1992.

[10] C. Fermüller and Y. Aloimonos. The role of fixation in visual motion analysis. *Int. Journal of Computer Vision*, 11(2):165–186, October 1993.

[11] D. Fleet, A. Jepson, and M. Jenkin. Phase-based disparity measurement. *CVGIP: Image Understanding*, 53(2):198–210, March 1991.

[12] G. Franklin, J. Powell, and M. Workman. *Digital Control of Dynamic Systems*. Addison-Wesley, 1990.

[13] D. Gennery. Visual tracking of known three-dimensional objects. *International Journal of Computer Vision*, 7(3):243–270, April 1992.

[14] B. Horn and B. Shunk. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.

[15] D. Koller, K. Daniilidis, and H. Nagel. Model-based object tracking in monocular image sequences of road traffic scenes. *International Journal of Computer Vision*, 10(3):257–281, June 1993.

[16] D. Lowe. Robust model-based motion tracking through the integration of search and estimation. *International Journal of Computer Vision*, 8(2):113–122, August 1992.

[17] K. Ludwig, H. Neumann, and B. Neumann. Robust estimation of local stereoscopic depth. In W. Förstner and St. Ruwiedel, editors, *Robust Computer Vision*. Wichmann, 1992.

[18] D. Robinson. The oculomotor control system: A review. *Proc. of the IEEE*, 56(6), June 1968.

[19] J. Santos-Victor, G. Sandini, F. Curotto, and S. Garibaldi. Divergent stereo in autonomous navigation : From bees to robots. *Int. Journal of Computer Vision*, 14(2):159–177, 1995.

[20] J. Santos-Victor, F. van Trigt, and J. Sentieiro. Medusa - a stereo head for active vision. In *Proc. of the Int. Symposium on Intelligent Robotic Systems*, Grenoble, France, July 1994.

[21] E. Schwartz. Spatial mapping in the primate sensory projection : Analytic structure and relevance to perception. *Biological Cybernetics*, 25:181–194, 1977.