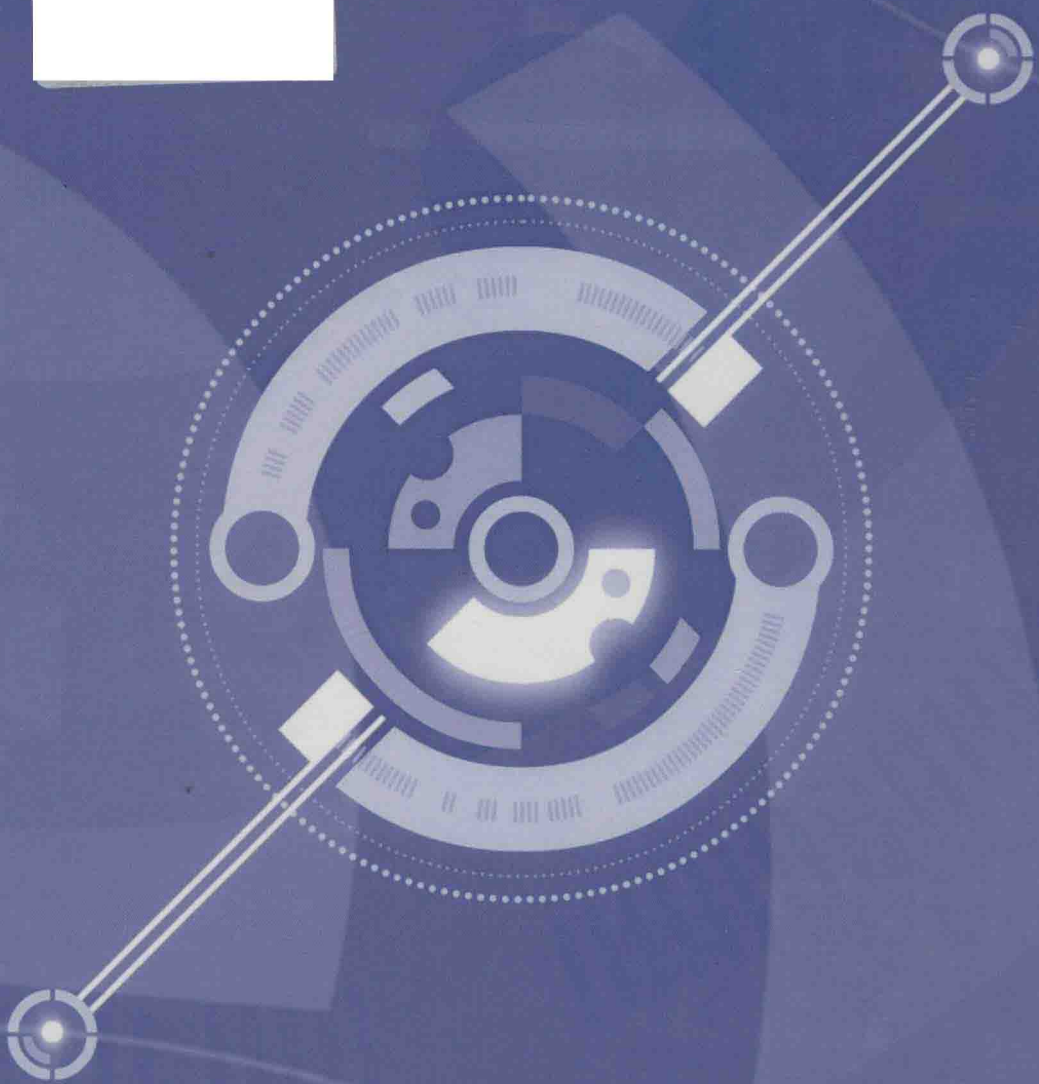

Machine Vision

Technology and Applications

Alexander Toker

Machine Vision: Technology and Applications

Edited by **Alexander Toker**

LANRYE INTERNATIONAL

New Jersey

Published by Clanrye International,
55 Van Reypen Street,
Jersey City, NJ 07306, USA
www.clanryeinternational.com

Machine Vision: Technology and Applications
Edited by Alexander Toker

International Standard Book Number: 978-1-63240-333-9 (Hardback)

Printed in China.

Machine Vision: Technology and Applications

Preface

Vision plays a significant role for human beings by permitting them to communicate with their surroundings in an effective and efficient way. The major aim of machine vision is to bestow artificial systems with enough efficiency to manage without prior predetermined conditions. For this conclusion, the computing constraints of the hosting architectures and the particulars of the tasks to be performed have to be taken into consideration, along with constant adapting and optimizing of the visual processing techniques. However, by exploiting the low cost computational power of off-the-shell computing devices, machine vision is not restricted to industrial environments, where conditions and tasks are made easy and specific, but it is now pervasive to support system solutions of everyday life difficulties.

The information contained in this book is the result of intensive hard work done by researchers in this field. All due efforts have been made to make this book serve as a complete guiding source for students and researchers. The topics in this book have been comprehensively explained to help readers understand the growing trends in the field.

I would like to thank the entire group of writers who made sincere efforts in this book and my family who supported me in my efforts of working on this book. I take this opportunity to thank all those who have been a guiding force throughout my life.

Editor

Contents

	Preface	**VII**
Chapter 1	**Bio-Inspired Active Vision Paradigms in Surveillance Applications** Mauricio Vanegas, Manuela Chessa, Fabio Solari and Silvio Sabatini	**1**
Chapter 2	**Fast Computation of Dense and Reliable Depth Maps from Stereo Images** M. Tornow, M. Grasshoff, N. Nguyen, A. Al-Hamadi and B. Michaelis	**23**
Chapter 3	**Stereo Matching Method and Height Estimation for Unmanned Helicopter** Kuo-Hsien Hsia, Shao-Fan Lien and Juhng-Perng Su	**49**
Chapter 4	**Rotation Angle Estimation Algorithms for Textures and Their Implementations on Real Time Systems** Cihan Ulas, Onur Toker and Kemal Fidanboylu	**73**
Chapter 5	**Real-Time Processing of 3D-TOF Data in Machine Vision Applications** Stephan Hussmann, Torsten Edeler and Alexander Hermanski	**91**
Chapter 6	**Characterization of the Surface Finish of Machined Parts Using Artificial Vision and Hough Transform** Alberto Rosales Silva, Angel Xeque-Morales, L.A. Morales-Hernandez and Francisco Gallegos Funes	**111**
Chapter 7	**Methods for Ellipse Detection from Edge Maps of Real Images** Dilip K. Prasad and Maylor K.H. Leung	**135**

Chapter 8 **Detection and Pose Estimation of Piled Objects Using Ensemble of Tree Classifiers** 163
Masakazu Matsugu, Katsuhiko Mori, Yusuke Mitarai and Hiroto Yoshii

Chapter 9 **Characterization of Complex Industrial Surfaces with Specific Structured Patterns** 177
Yannick Caulier

Chapter 10 **Reflectance Modeling in Machine Vision: Applications in Image Analysis and Synthesis** 205
Robin Gruna and Stephan Irgenfried

Chapter 11 **Discontinuity Detection from Inflection of Otsu's Threshold in Derivative of Scale-Space** 225
Rahul Walia, David Suter and Raymond A. Jarvis

Chapter 12 **Towards the Optimal Hardware Architecture for Computer Vision** 247
Alejandro Nieto, David López Vilarino and Víctor Brea Sánchez

Permissions

List of Contributors

1

Bio-Inspired Active Vision Paradigms in Surveillance Applications

Mauricio Vanegas, Manuela Chessa, Fabio Solari and Silvio Sabatini
The Physical Structure of Perception and Computation - Group, University of Genoa
Italy

1. Introduction

Visual perception was described by Marr (1982) as the processing of visual stimuli through three hierarchical levels of computation. In the first level or *low-level* vision it is performed the extraction of fundamental components of the observed scene such as edges, corners, flow vectors and binocular disparity. In the second level or *medium-level* vision it is performed the recognition of objects (e.g. model matching and tracking). Finally, in the third level or *high-level* vision it is performed the interpretation of the scene. A complementary view is presented in (Ratha & Jain, 1999; Weems, 1991); by contrast, the processing of visual stimuli is analysed under the perspective developed by Marr (1982) but emphasising how much data is being processed and what is the complexity of the operators used at each level. Hence, the low-level vision is characterised by large amount of data, small neighbourhood data access, and simple operators; the medium-level vision is characterised by small neighbourhood data access, reduced amount of data, and complex operators; and the high-level vision is defined by non-local data access, small amount of data, and complex relational algorithms. Bearing in mind the different processing levels and their specific characteristics, it is plausible to describe a computer vision system as a modular framework in which the low-level vision processes can be implemented by using parallel processing engines like GPUs and FPGAs to exploit the data locality and the simple algorithmic operations of the models; and the medium and high-level vision processes can be implemented by using CPUs in order to take full advantage of the straightforward fashion of programming these kind of devices.

The low-level vision tasks are probably the most studied in computer vision and they are still an open research area for a great variety of well defined problems. In particular, the estimation of optic flow and of binocular disparity have earned special attention because of their applicability in segmentation and tracking. On the one hand, the stereo information has been proposed as a useful cue to overcome some of the issues inherent to robust pedestrian detection (Zhao & Thorpe, 2000), to segment the foreground from background layers (Kolmogorov et al., 2005), and to perform tracking (Harville, 2004). On the other hand, the optic flow is commonly used as a robust feature in motion-based segmentation and tracking (Andrade et al., 2006; Yilmaz et al., 2006).

This chapter aims to describe a biological inspired video processing system for being used in video surveillance applications; the degree of similarity between the proposed framework

and the human visual system allows us to take full advantage of both optic flow and disparity estimations not only for tracking and fixation in depth but also for scene segmentation. The most relevant aspect in the proposed framework is its hardware and software modularity. The proposed system integrates three cameras (see Fig. 1); two active cameras with variable-focal-length lenses (binocular system) and a third fixed camera with a wide-angle lens. This system has been designed to be compatible with the well-known iCub robot interface[1]. The cameras movement control, as well as the zoom and iris control run on an embedded computer PC/104. The optic flow and the disparity algorithms run on a desktop computer equipped with a processor *Intel Core 2 Quad @ 2.40GHz* and a memory RAM of about 8 GB. All system components, namely the desktop computer, the embedded computer PC/104, and the cameras, are connected in a gigabit Ethernet network through which they can interact as a distributed system.

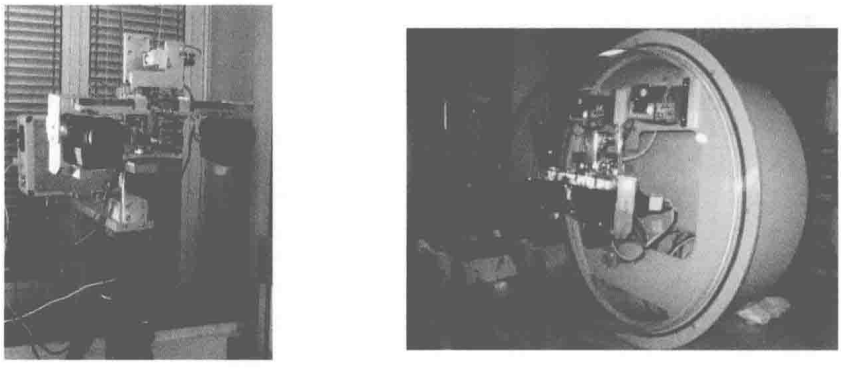

Fig. 1. Trinocular robotic head with 5 degrees of freedom, namely a common tilt movement, and independent zoom-pan movements for left and right cameras, respectively.

The general features of the moving platform are compiled in Table 1. Likewise, the optic features of the cameras are collected in Table 2. Lastly, it is important to mention that the binocular system has a baseline of 30 cm.

Features	Pan Movement	Tilt Movement
Limits:	$\pm 30^\circ$ (Software limit)	$\pm 60^\circ$ (Software limit)
Acceleration:	$5100^\circ/sec^2$	$2100^\circ/sec^2$
Max. Speed:	$330^\circ/sec$	$73^\circ/sec$
Resolution:	0.03°	0.007°
Optical Encoder:	512 pulses/revolution	512 pulses/revolution
Motor Voltage:	12 V	12 V
Gear Ratio:	1:80	1:80
Motor Torque:	0.59 Nm	0.59 Nm

Table 1. General features of the moving platform.

Most of the video surveillance systems are networks of cameras for a proper coverage of wide areas. These networks use both fixed or active cameras, or even a combination of both, placed

[1] The iCub is the humanoid robot developed as part of the EU project RobotCub and subsequently adopted by more than 20 laboratories worldwide (see `http://www.icub.org/`).

Features	Active Cameras	Fixed Camera
Resolution:	11392 x 1040 pixels	1624 x 1236 pixels
Sensor Area:	6.4 x 4.8 mm	7.1 x 5.4 mm
Pixel Size:	4.65 x 4.65 μm	4.4 x 4.4 μm
Focal Length:	7.3 $\sim$ 117 mm, FOV 47° $\sim$ 3°	4.8 mm, FOV 73°

Table 2. Optic features of the cameras.

at not predetermined positions to strategically cover a wide area; the term *active* specifies the camera's ability of changing both the angular position and the field of view. The type of cameras used in the network has inspired different calibration processes to find automatically both the intrinsic and extrinsic camera parameters. In this regard, Lee et al. (2000) proposed a method to estimate the 3D positions and orientations of fixed cameras, and the ground plane in a global reference frame which lets the multiple cameras views to be aligned into a single planar coordinate frame; this method assume approximate values for intrinsic cameras parameters and it is based on overlapped cameras views; however, others calibration methods have been proposed for non-overlapped cameras views (i.e. Kumar et al., 2008). In the case of active cameras, Tsai (1987) has developed a method for estimating both the matrices of rotation and translation in the Cartesian reference frame, and the intrinsic parameters of the cameras. In addition to the calibration methods, the current surveillance systems must deal with the segmentation and identification of complex scenes in order to characterise them and thus to obtain a classification which let the system to recognise unusual behaviours into the scene. In this regard, a large variety of algorithms have been developed to detect changes in scene; for example the application of a threshold to the absolute difference between pixel intensities of two consecutive frames can lead to the identification of moving objects, some methods for the threshold selection are described in (Kapur et al., 1985; Otsu, 1979; Ridler & Calvar, 1978). Other examples are the adaptive background subtraction to detect moving foreground objects (Stauffer & Grimson, 1999; 2000) and the estimation of optic flow (Barron et al., 1994). Our proposal differs the most of the current surveillance systems in at least three aspects: (1) the use of a single camera with a wide-angle lens to cover vast areas and a binocular system for tracking areas of interest at different fields of view (the wide-angle camera is used as the reference frame), (2) the estimation of both optic flow and binocular disparity for segmenting the images; this system feature can provide useful information for disambiguating occlusions in dynamic scenarios, and (3) the use of a bio-inspired fixation strategy which lets the system to fixate areas of interest, accurately.

In order to explain the system behaviour, two different perspectives were described. On the one hand, we present the system as a bio-inspired mathematical model of the primary visual cortex (see section 2); from this viewpoint, we developed a low-level vision architecture for estimating optic flow and binocular disparity. On the other hand, we describe the geometry of the cameras position in order to derive the equations that govern the movement of the cameras (see section 3). Once the system is completely described, we define an angular-position control capable of changing the viewpoint of the binocular system by using disparity measures in section 4. An interesting case study is described in section 5 where both disparity and optic flow are used to segment images. Finally, in section 6, we present and discuss the system's performance results.

2. The system: a low-level vision approach

The visual cortex is the largest, and probably the most studied part of the human brain. The visual cortex is responsible for the processing of visual stimuli impinging on the retinas. As a matter of fact, the first stage of processing takes place in the lateral geniculate nucleus (LGN) and then the neurons of the LGN relay the visual information to the primary visual cortex (V1). Then, the visual information flow hierarchically to areas V2, V3, V4 and V5/MT where visual perception gradually takes place.

The experiments carried out by Hubel & Wiesel (1968) proved that the primary visual cortex (V1) consists of cells responsive to different kinds of spatiotemporal features of the visual information. The apparent complexity with which the brain extracts the spatiotemporal features has been clearly explained by Adelson & Bergen (1991). The light filling a region of space contains information about the objects in that space; in this regard, they proposed the *plenoptic function* to describe mathematically the pattern of light rays collected by a vision system. By definition, the plenoptic function describes the state of luminous environment, thus the task of the visual system is to extract structural elements from it.

Structural elements of the plenoptic function can be described as oriented patterns in the plenoptic space, and the primary cortex can be interpreted as a set of local, Fourier or Gabor operators used to characterise the plenoptic function in the spatiotemporal and frequency domains.

2.1 Neuromorphic paradigms for visual processing

Mathematically speaking, the extraction of the most important aspects of the plenoptic function can emulate perfectly the neuronal processing of the primary visual cortex (V1). More precisely, qualities or elements of the visual input can be estimated by applying a set of low order directional derivatives at the sample points; the so obtained measures represent the amount of a particular type of local structure. To effectively characterise a function within a neighbourhood, it is necessary to work with the local average derivative or, in an equivalent form, with the oriented linear filters in the function hyperplanes. Consequently, the neurons in V1 can be interpreted as a set of oriented linear filters whose outputs can be combined to obtain more complex feature detectors or, what is the same, more complex receptive fields. The combination of linear filters allow us to measure the magnitude of local changes within a specific region, without specifying the exact location or spatial structure. The receptive fields of complex neurons have been modelled as the sum of the squared responses of two linear receptive fields that differ just in phase for 90° (Adelson & Bergen, 1985); as a result, the receptive fields of complex cells provide *local energy measures*.

2.2 Neural Architecture to estimate optic flow and binocular disparity

The combination of receptive fields oriented in space-time can be used to compute local energy measures for optic flow (Adelson & Bergen, 1985). Analogously, by combining the outputs of spatial receptive fields it is possible to compute local energy measures for binocular disparity (Fleet et al., 1996; Ohzawa et al., 1990). On this ground, it has been recently proposed a neural architecture for the computation of horizontal and vertical disparities and optic flow (Chessa, Sabatini & Solari, 2009). Structurally, the architecture comprises four processing stages (see

Fig. 2): the distributed coding of the features by means of oriented filters that resemble the filtering process in area V1; the decoding process of the filter responses; the estimation of the local energy for both optic flow and binocular disparity; and the coarse-to-fine refinement.

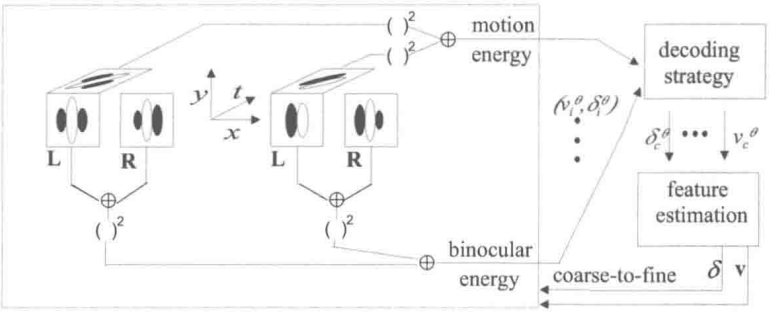

Fig. 2. The neural architecture for the computation of disparity and optic flow.

The neuronal population is composed of a set of 3D Gabor filters which are capable of uniformly covering the different spatial orientations, and of optimally sampling the spatiotemporal domain (Daugman, 1985). The linear derivative-like computation concept of the Gabor filters let the filters to have the form $h(\mathbf{x},t) = g(\mathbf{x})f(t)$. Both spatial and temporal terms in the right term are comprised of one harmonic function and one Gaussian function. This can be easily deduced from the impulse response of the Gabor filter.

The mathematical expression of the spatial term of a 3D Gabor filter rotated by an angle θ with respect to the horizontal axis is:

$$g(x,y;\psi,\theta) = e^{\left(-\frac{x_\theta^2}{2\sigma_x^2}-\frac{y_\theta^2}{2\sigma_y^2}\right)} e^{j(\omega_0 x_\theta+\psi)}, \tag{1}$$

where $\theta \in [0, 2\pi)$ represents the spatial orientation; ω_0 and ψ are the frequency and phase of the sinusoidal modulation, respectively; the values σ_x and σ_y determine the spatial area of the filter; and (x_θ, y_θ) are the rotated spatial coordinates.

The algorithm to estimate the binocular disparity is based on a *phase-shift* model; one of the variations of this model suggests that disparity is coded by phase shifts between receptive fields of the left and right eyes whose centres are in the same retinal position (Ohzawa et al., 1990). Let the left and right receptive fields be $g^L(\mathbf{x})$ and $g^R(\mathbf{x})$, respectively; the binocular phase shift is defined by $\Delta\psi = \psi^L - \psi^R$. Each spatial orientation has a set of k receptive fields with different binocular phase shifts in order to be sensitive to different disparities ($\delta^\theta = \Delta\psi/\omega_o$); the phase shifts are uniformly distributed between $-\pi$ and π. Therefore, the left and right receptive fields are applied to a binocular image pair $I^L(\mathbf{x})$ and $I^R(\mathbf{x})$ according to the following equation:

$$Q(\mathbf{x}_0;\delta^\theta) = \int_{-\infty}^{\infty} g^L(\mathbf{x}_0-\mathbf{x})I^L(\mathbf{x})d\mathbf{x} + \int_{-\infty}^{\infty} g^R(\mathbf{x}_0-\mathbf{x})I^R(\mathbf{x})d\mathbf{x}, \tag{2}$$

so, the spatial array of binocular energy measures can be expressed as:

$$E(\mathbf{x};\delta^\theta) = |Q(\mathbf{x};\delta^\theta)|^2 = |Q^L(\mathbf{x};\delta^\theta) + e^{-j\Delta\psi}Q^R(\mathbf{x};\delta^\theta)|^2. \tag{3}$$

Likewise, the temporal term of a 3D Gabor filter is defined by:

$$f(t;\omega_t) = e^{\left(-\frac{t^2}{2\sigma_t^2}\right)} e^{j\omega_t t} 1(t), \tag{4}$$

where σ_t determines the integration window of the filter in time domain; ω_t is the frequency of the sinusoidal modulation; and $1(t)$ denotes the unit step function. Each receptive field is tuned to a specific velocity v^θ along the direction orthogonal to the spatial orientation θ. The temporal frequency is varied according to $\omega_t = v^\theta \omega_0$. Each spatial orientation has a set of receptive fields sensitive to M tuning velocities; M depends on the size of the area covered by each filter according to the Nyquist criterion.

The set of spatiotemporal receptive fields $h(\mathbf{x}, t)$ is applied to an images sequence $I(\mathbf{x}, t)$ according to the following equation:

$$Q(\mathbf{x_0}, t; v^\theta) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(\mathbf{x_0} - \mathbf{x}, t - \tau) I(\mathbf{x}, \tau) d\mathbf{x} d\tau, \tag{5}$$

so, the motion energy $E(\mathbf{x_0}, t; v^\theta)$ equals:

$$E(\mathbf{x_0}, t; v^\theta) = |Q(\mathbf{x_0}, t; v^\theta)|^2 = \left| e^{j\psi(t)} \int_0^t Q(\mathbf{x_0}, \tau; v^\theta) e^{-j\omega_t \tau} d\tau \right|^2. \tag{6}$$

where $\psi(t) = \psi + \omega_t t = \psi + \omega_0 v^\theta t$.

So far, we have described the process of encoding both binocular disparity and optic flow by means of a $N \times M \times K$ array of filters uniformly distributed in space domain. Now, it is necessary to extract the component velocity (v_c^θ) and the component disparity (δ_c^θ) from the local energy measures at each spatial orientation. The accuracy in the extraction of these components is strictly correlated with the number of filters used per orientation, such that precise estimations require a large number of filters; as a consequence, it is of primary importance to establish a compromise between the desired accuracy and the number of filters used or, what is the same, a compromise between accuracy and computational cost.

An affordable computational cost can be achieved by using weighted sum methods as the *maximum likelihood* proposed by Pouget et al. (2003). However, the proposed architecture uses the centre of gravity of the population activity since it has shown the best compromise between simplicity, computational cost and reliability of the estimates. Therefore, the component velocity v_c^θ is obtained by pooling cell responses over all orientations:

$$v_c^\theta(\mathbf{x_0}, t) = \frac{\sum_{i=1}^{M} v_i^\theta E(\mathbf{x_0}, t; v_i^\theta)}{\sum_{i=1}^{M} E(\mathbf{x_0}, t; v_i^\theta)}, \tag{7}$$

where v_i^θ represent all the M tuning velocities; and $E(\mathbf{x_0}, t; v_i^\theta)$ represent the motion energies at each spatial orientation. The component disparity δ_c^θ can be estimated in a similar way.

Because of the aperture problem a filter can just estimate the features which are orthogonal to the orientation of the filter. So we adopt k different binocular and M different motion receptive fields for each spatial orientation; consequently, a robust estimate for the full velocity $\mathbf{v}$ and for

the full disparity δ is achieved by combining all the estimates v_c^θ and δ_c^θ, respectively (Pauwels & Van Hulle, 2006; Theimer & Mallot, 1994).

Finally, the neural architecture uses a coarse to fine control strategy in order to increase the range of detection in both motion and disparity. The displacement features obtained at coarser levels are expanded and used to warp the images in finer levels in order to achieve a higher displacement resolution.

3. The system: a geometrical description

In the previous section we presented the system from a biological point of view. We have summarised a mathematical model of the behaviour of the primary visual cortex and we have proposed a computational architecture based on linear filters for estimating optic flow and binocular disparity. Now it is necessary to analyse the system from a geometrical point of view in order to link the visual perception to the camera movements, thus letting the system to interact with the environment.

To facilitate the reference to the cameras within this text, we are going to refer the fixed camera as *wide-angle camera*, and the cameras of the binocular system as *active cameras*. The wide-angle camera is used for a wide view of the scene, and it becomes the reference of the system. In vision research, the cyclopean point is considered the most natural centre of a binocular system (Helmholtz, 1925) and it is used to characterise stereopsis in human vision (Hansard & Horaud, 2008; Koenderink & van Doorn, 1976). By doing a similar approximation, the three-camera model uses the wide-angle-camera image as the cyclopean image of the system. In this regard, the problem statement is not trying to construct the cyclopean image from the binocular system, but using the third camera image as a reference coordinate to properly move the active cameras according to potential targets or regions of interest in the wide range scenario.

Each variable-focal-length camera can be seen as a 3DOFs pan-tilt-zoom (PTZ) camera. However, the three-camera system constraints the active cameras to share the tilt movement due to the mechanical design of the binocular framework. One of the purposes of our work is to describe the geometry of the three-camera system in order to properly move the pan-tilt-zoom cameras to fixate any object in the field of view of the wide-angle camera and thus to get both a magnified view of the target object and the depth of the scene.

We used three coordinates systems to describe the relative motion of the active cameras with respect to the wide-angle camera (see Fig. 3). The origin of each coordinate system is supposed in the focal point of each camera and the Z-axes are aligned with the optical axes of the cameras. The pan angles are measured with respect to the planes $X_L = 0$ and $X_R = 0$ respectively; note that pan angles are positive for points to the left of these planes ($X_L > 0$ or $X_R > 0$). The rotation axes for the pan movement are supposed to be parallel. The common tilt angle is measured with respect to the horizontal plane; note that the tilt angle is positive for points above the horizontal plane ($Y_L = Y_R > 0$).

The point $P(X, Y, Z)$ can be written in terms of the coordinate systems shown in Fig. 3 as follows:

$$(X, Y, Z) = (X_L, Y_L, Z_L) - O_L, \tag{8}$$

$$(X, Y, Z) = (X_R, Y_R, Z_R) - O_R, \tag{9}$$

where $O_L = (dx_L, dy_L, dz_L)$ and $O_R = (-dx_R, dy_R, dz_R)$ are the origin of the coordinate system of the left and right cameras with respect to the wide-angle camera coordinate system.

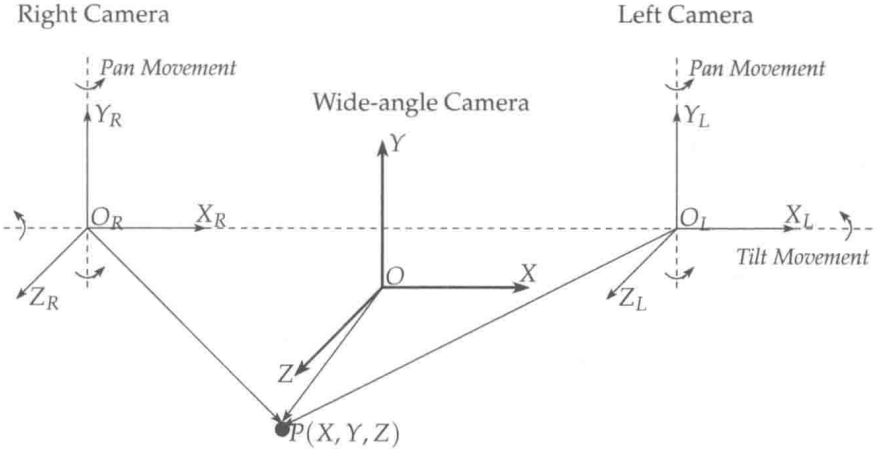

Fig. 3. The coordinate systems of the three cameras in the binocular robotic head.

It is considered f_w as the focal length of the wide-angle camera and f as the focal length of the active cameras. The Equations 8 and 9 can be written in terms of the image coordinate system of the wide-angle camera if these equations are multiplied by factor $\frac{f_w}{Z}$:

$$\frac{f_w}{Z}(X_L, Y_L, Z_L) = (x, y, f_w) + \frac{f_w}{Z}(dx_L, dy_L, dz_L), \tag{10}$$

$$\frac{f_w}{Z}(X_R, Y_R, Z_R) = (x, y, f_w) + \frac{f_w}{Z}(-dx_R, dy_R, dz_R). \tag{11}$$

Now, it is possible to link the image coordinate system of the wide-angle camera to the image coordinate system of the active cameras by multiplying the Equations 10 and 11 by the factors $\frac{f}{Z_L}$ and $\frac{f}{Z_R}$, respectively:

$$\frac{f_w}{Z}(x_L, y_L, f) = \frac{f}{Z_L}(x, y, f_w) + \frac{f_w f}{Z_L Z}(dx_L, dy_L, dz_L), \tag{12}$$

$$\frac{f_w}{Z}(x_R, y_R, f) = \frac{f}{Z_R}(x, y, f_w) + \frac{f_w f}{Z_R Z}(-dx_R, dy_R, dz_R). \tag{13}$$

Assuming that the position of the origin with respect to the Z-axis is small enough compared to the distance of the real object in the scene, it can be done the next approximation $Z \approx Z_L$ and $Z \approx Z_R$. Accordingly, the Equations 12 and 13 can be rewritten to obtain the *wide-to-active*