Proceedings of the 2002 IEEE/RSJ
Intl. Conference on Intelligent Robots and Systems
EPFL, Lausanne, Switzerland • October 2002

Segmentation of Dense Depth Maps using Inertial Data
A real-time implementation

Jorge Lobo, Luis Almeida and Jorge Dias

Institute of Systems and Robotics,
University of Coimbra, Portugal,
{jlobo,laa,jorge}@isr.uc.pt

Abstract

In this paper we propose a real-time system that extracts information from dense relative depth maps. This method enables the integration of depth cues on higher level processes including segmentation of structures, object recognition, robot navigation or any other task that requires a three-dimensional representation of the physical environment.

Inertial sensors coupled to a vision system can provide important inertial cues for the ego-motion and system pose. The sensed gravity provides a vertical reference. Depth maps obtained from a stereo camera system can be segmented using this vertical reference, identifying structures such as vertical features and levelled planes.

In our work we explore the integration of inertial sensor data in vision systems. Depth maps obtained by vision systems, are very point of view dependant, providing discrete layers of detected depth aligned with the camera. In this work we use inertial sensors to recover camera pose, and rectify the maps to a reference ground plane, enabling the segmentation of vertical and horizontal geometric features.

The aim of this work is a fast real-time system, so that it can be applied to autonomous robotic systems or to automated car driving systems, for modelling the road, identifying obstacles and roadside features in real-time.

1 Introduction

One of the very important tasks in computer vision is to extract depth information of the world. Stereoscopy is a technique to extract depth information from two images of a scene taken from different view points. This information can be integrated on a single entity called dense depth map.

In humans and in animals the vestibular system in the inner ear gives inertial information essential for navigation, orientation, body posture control and equilibrium. In humans this sensorial system is crucial for several visual tasks and head stabilisation. It is well known that the information provided by the vestibular system is used during the execution of visual movements such as gaze holding and tracking, as described by Carpenter [1]. Neural interactions of human vision and vestibular system occur at a very early processing stage [2].

In this work we use the vertical reference provided by the inertial sensors to perform a fast segmentation of depth maps obtained from a stereo real time algorithm.

Nowadays micromachined low cost inertial sensors can be easily incorporated in computer vision systems. These sensors can perform as an artificial vestibular system, providing valuable data to the vision system. The motivation might be stronger in applications such as walking or flying robots, but in automobiles, due to suspension and system compliance, it is also beneficial to have inertial sensors coupled to the vision system cameras.

In our previous work on inertial sensor data integration in vision systems, the inertial data was directly used with the image data [3][4][5]. In this work we use the inertial data to perform a fast segmentation of pre-computed depth maps obtained from the vision system.

1.1 Related Work

The aim of stereo systems is to achieve an adequate throughput and precision to enable video-rate dense depth mapping. The throughput of a stereo machine can be measured by the product of the number of depth measurements per second (pixel/sec) and the range of disparity search (pixels); the former determines the density and speed of depth measurement and the later the dynamic range distance measurement [7], [8], [9], [10]. The group of T. Kanade at CMU [11] succeeded in producing a video-rate stereo machine based on the multi-baseline stereo algorithm to generate a dense range map. SRI has

0-7803-7398-7/02/$17.00 ©2002 IEEE

developed an efficient implementation of area correlation stereo, the SRI Stereo Engine. The standard development environment for the SRI Stereo Engine is the Small Vision System (SVS), which runs on PCs under Linux or MS Windows. The PC implementation is a efficient solution, with support for camera calibration, 3D reconstruction, and effective filtering [12]. We are using this system to obtain real-time depth maps.

Viéville and Faugeras proposed the use of an inertial system based on low cost sensors for mobile robots [13] and studied the cooperation of the inertial and visual systems in mobile robot navigation by using the vertical cue taken from the inertial sensors [14] [15] [16]. An inertial sensor integrated optical flow technique was proposed by Bhanu *et al.* [17]. Panerai and Sandini used a low cost gyroscope for gaze stabilization of a rotating camera, and compared the camera rotation estimate given by image optical flow with the gyro output [18] [19]. Mukai and Ohnishi studied the recovery of 3D shape from an image sequence using a video camera and a gyro sensor [20].

In our previous work we have explored the integration of inertial sensor data in vision systems. Using the vertical reference provided by the inertial sensors, the image horizon can be determined. Using just one vanishing point we can recover the camera's focal distance [3]. In a typical indoor corridor scene the vanishing point can also provide an external bearing for the robots navigation frame. Knowing the geometry of a stereo rig, and its pose from the inertial sensors, the homography of level planes can be recovered, providing enough restrictions to segment and reconstruct vertical features [5] and levelled planar patches [4].

2 Depth Maps

In order to describe the tasks involved in depth map construction, let us consider the following geometric model, figure 1, of our stereo vision system (see figure 2). The diagram shows the top view of a stereo system composed of two pinhole cameras. The left and right image planes are coplanar and represented by the segments I_l and I_r respectively. C_l and C_r are the centers of projection. The optical axes are parallel: for this reason, the *fixation point* defined as the point of intersection of the optical axes, lies infinitely far from the cameras.

The way in which stereo determines the position of $P_{(x,y,z)}$ in space is by *triangulation*, that is by intersecting the rays defined by the centers of projection and the images of P, p_l and p_r. Using the left camera as reference and solving for (x, y, z) gives:

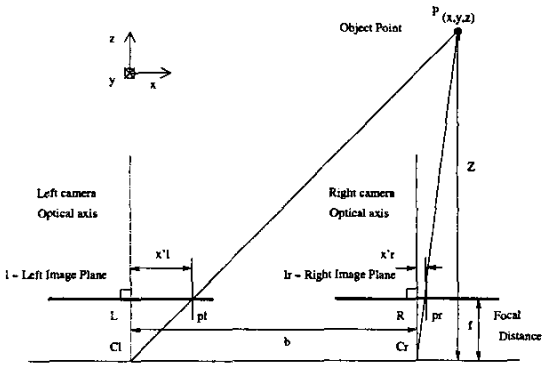

Figure 1: *Geometry of front-parallel stereo setup.*

$$\begin{cases} x'_l/f = x/z \\ x'_r/f = (x-b)/z \\ y'_l/f = y'_r/f = y/z \end{cases} \quad \begin{cases} x = x'_l * b/(x'_l - x'_r) \\ y = y'_l * b/(x'_l - x'_r) \\ z = f * b/(x'_l - x'_r) \end{cases} \quad (1)$$

The classical approach to estimate disparities uses two techniques: feature matching and correlation. In a feature-based algorithm, a set of complex tokens is extracted from each left and right images, and then combined according to some constraints. The second technique uses a measure of similarity, correlation for example, to find matching points in two images composing the stereo pair. For each point of the reference image, the corresponding point is selected in the other image by searching for a maximum in similarity measure.

Many stereo camera configurations have vergence and do not comply with the front-parallel geometric model. In that case the disparity measurement can be related with the *horopter*. Coombs [21] characterizes horopter as being the surface in three dimensional space defined by the points that stimulates exactly corresponding points (i.e., that has zero stereo disparity) in two cameras. Small disparities correspond to small deviation in depth from horopter. Such disparities can be measured by simple local neighbourhood operators, to build up a dense surface map of the environment near the horopter [22][23]. A stereo configuration with vergence angle can be considerably simplified when the images of interest have been rectified, i.e., replaced by two projectively equivalent pictures with a common image plane parallel to the baseline joining the two optical centers, and equivalent to a front-parallel system. The *rectification* process can be implemented by projecting the original pictures onto the new image plane [24]. With an appropriate choice of coordinate system, the rectified images have scanlines parallel to the baseline and the front-parallel geometry of figure 1 can be applied.

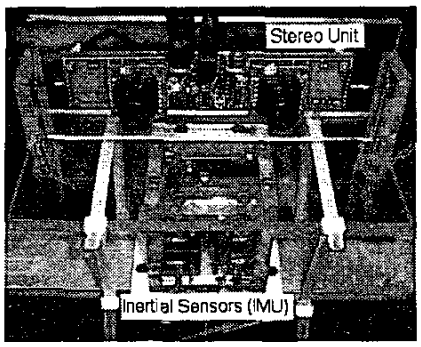

Figure 2: Cameras with Inertial Measurement Unit based on low cost sensors.

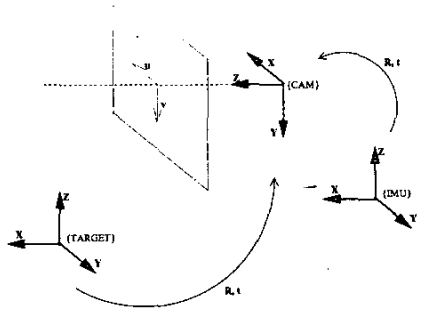

Figure 3: Camera and IMU referentials.

2.1 SRI Stereo Engine and the Small Vision Systems (SVS)

In order to compute range from stereo images we are using the SRI Stereo Engine [12]. It implements an area correlation algorithm for computing range from stereo images and it supports camera calibration, 3D reconstruction, and effective filtering. We are running an implementation of Stereo Engine for Linux *Small Vision System (SVS)* (version svs23c). SVS consists of a set of library functions, in C++, implementing the stereo algorithms optimized for PC's using MMX instruction. It can receive input stereo images from standard cameras and video capture devices. On this particular work we are using a small and compact stereo head developed by Videre Design [12] (see figure 2), the STH-V3. This analog vision head can also send a single video signal with the interlaced stereo image pair. STH-V3 consists of two synchronized CMOS cameras modules mounted on a baseboard with 320x240 pixels (NTSC) each one. The software is running on a Linux RH7.1 box (PII 350Mhz) with a Pinnacle Studio PCTV (Bt878-based) card as framegrabber. With a frame size of 160x120, searching 16 disparities and a search window size of 5 x 5 we achieved a frame rate of 30 Hz.

3 Inertial Data

An inertial measurement unit (IMU) coupled to a camera can provide valuable data about camera pose and movement. Figure 2 shows an inertial system prototype built at our lab [6] that was coupled to a stereo camera rig to perform the tests. Camera calibration was performed using a fixed target and moving the system, recovering the cameras' intrinsic parameters, as well as the target positions relative to the cameras.

By moving the cameras instead of the target, the cameras' position is determined relative to the fixed target. Since the IMU is rigidly connected to the camera, R and t shown in figure 3 can be determined from the set of camera positions obtained from the calibration and the corresponding data from the inertial sensors. By performing a trajectory with the target always in view, the camera calibration can reconstruct the camera pose and position, providing an error measure for the INS data.

If both IMU and cameras are perfectly aligned, we have a simple translation change of axis, and

$$ {}^{C}T_{\mathcal{CAM}} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ -1 & 0 & 0 & b/2 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (2) $$

where b is the baseline, $\{\mathcal{CAM}\}$ the camera referential and $\{C\}$ the camera system referencial, with origin at the center of the baseline.

Having determined the rigid transformation between the camera and the IMU, the sensed acceleration and rotation are mapped to the camera system referential.

3.1 Gravity Vector

The measurements $\boldsymbol{a}$ taken by the inertial unit's accelerometers include the sensed gravity vector $\boldsymbol{g}$ summed with the body's acceleration $\boldsymbol{a}_b$:

$$ \boldsymbol{a} = \boldsymbol{g} + \boldsymbol{a}_b \quad (3) $$

Assuming the system is motionless, then $\boldsymbol{a}_b = 0$ and the measured acceleration $\boldsymbol{a} = \boldsymbol{g}$ gives the gravity vector in the system's referential. So, with a_x,a_y and a_z being the accelerometer filtered measurements along each axis, the vertical unit vector will be given by

$$ \hat{\boldsymbol{n}} = -\frac{\boldsymbol{g}}{\|\boldsymbol{g}\|} = \frac{1}{\sqrt{a_x^2+a_y^2+a_z^2}} \begin{bmatrix} a_x \\ a_y \\ a_z \end{bmatrix} = \begin{bmatrix} n_x \\ n_y \\ n_z \end{bmatrix} \quad (4) $$

This vertical reference, given in the systems frame of reference, will be used in segmenting the depth maps obtained from the stereo algorithm.

Consider a point given in the camera system referential ${}^{C}\boldsymbol{P}$ that belongs to the ground plane. The plane equation is given by

$$ {}^{C}\hat{\boldsymbol{n}}.{}^{C}\boldsymbol{P} + d = 0 \tag{5} $$

where d is the distance from the origin to the ground plane, *i.e.*, the system height. In some applications it can be known or imposed by the physical mount.

4 Depth Maps in Inertial Reference Frame

In our experimental setup, the stereo algorithm provides depth maps in the left camera frame of reference. Using the vertical reference provided by the inertial sensors, $\hat{\boldsymbol{n}}$, the depth maps can be rotated and aligned with the horizontal plane. The points obtained in the camera referential, $\{C\}$, can be converted to a world frame of reference $\{W\}$. The vertical unit vector $\hat{\boldsymbol{n}}$ and system heigh d can be used to define $\{W\}$, by choosing ${}^{W}\hat{\boldsymbol{x}}$ to be coplanar with ${}^{C}\hat{\boldsymbol{x}}$ and ${}^{C}\hat{\boldsymbol{n}}$ in order to keep the same heading, we have

$$ {}^{W}\boldsymbol{P} = {}^{W}\boldsymbol{T}_{C}.{}^{C}\boldsymbol{P} \tag{6} $$

where

$$ {}^{W}\boldsymbol{T}_{C} = \begin{bmatrix} \sqrt{1-n_x^2} & \frac{-n_x n_y}{\sqrt{1-n_x^2}} & \frac{-n_x n_z}{\sqrt{1-n_x^2}} & 0 \\ 0 & \frac{n_z}{\sqrt{1-n_x^2}} & \frac{-n_y}{\sqrt{1-n_x^2}} & 0 \\ n_x & n_y & n_z & d \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{7} $$

System height d can be known *apriori* or inferred from the subsequent segmentation process, using an initial null value.

If a heading reference is available, then $\{W\}$ should not be restricted to having ${}^{W}\hat{\boldsymbol{x}}$ coplanar with ${}^{C}\hat{\boldsymbol{x}}$ and ${}^{C}\hat{\boldsymbol{n}}$, but use the known heading reference. Using a heading reference given by the unit vector $\hat{\boldsymbol{m}} = (m_x, m_y, m_z)$ we get

$$ {}^{C}\boldsymbol{T}_{W} = \begin{bmatrix} m_x & n_y m_z - n_z m_y & n_x & -n_x d \\ m_y & n_z m_x - n_x m_z & n_y & -n_y d \\ m_z & n_x m_y - n_y m_x & n_z & -n_z d \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{8} $$

We are therefor able to have $\{W\}$ coherent with the inertial vertical and the available scene heading. The gyros included in the inertial measurement unit can be used to keep a heading without external references, but they accumulate drift over time. Visual land marks or a magnetic compass provide an external heading reference to reset the drift. In scenes of man made environments image vanishing points from detected edge lines can also provide a heading reference. The segmented depth maps can also be used, by identifying features such as walls in the points mapped to the inertial reference frame and above the ground plane.

5 Segmented Depth Maps

Using the vertical reference, the depth maps can be segmented to identify horizontal and vertical features. The aim in on having a simple algorithm suitable for a real-time implementation. Since we are able to map the points to an inertial reference frame, planar levelled patches will have the same depth z, and vertical features the same xy, allowing simple feature segmentation using histogram local peak detection. Figure 4 summarizes the proposed depth map segmentation method.

Using the stereo depth algorithm we obtain a set of points ${}^{CAM}P_i$ in the left camera referential. Using the previous equations we can map them to the world referential as

$$ {}^{W}P_i = {}^{W}\boldsymbol{T}_{C}.{}^{C}\boldsymbol{T}_{CAM}.{}^{CAM}P_i \tag{9} $$

In order to detect the ground plane point, an histogram is performed for point depth.

$$ hist_z(n) = \sum (P_i \mid floor(z_{P_i}) = n) \tag{10} $$

The histogram's lower local peak z_{gnd} is used as the reference depth for the ground plane. The detected points can than be parsed and segmented as being a ground plane point, or some feature above ground. Points bellow the ground plane can be ignored or not, depending on the application.

$$ P_{gnd} = P_i \mid z_{gnd} - \delta \leq floor(z_{P_i}) \leq z_{gnd} + \delta \tag{11} $$

$$ P_{above} = P_i \mid floor(z_{P_i}) \geq z_{gnd} + \delta \tag{12} $$

were δ is the allowed tolerance. The points above ground can be projected in the XY plane, and further segmentation performed to identify vertical features.

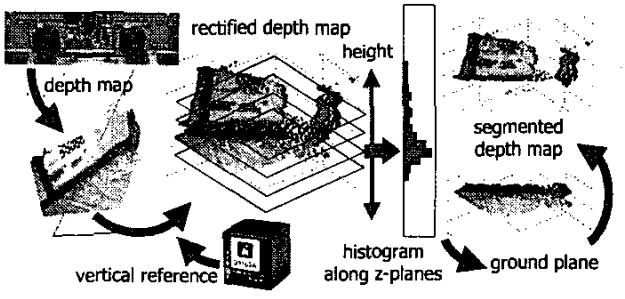

Figure 4: Summary of implemented method.

6 Results

A simple indoor scene was used to test our method. The stereo pair seen in figure 6 was obtained with the experimental setup shown in figure 5. Figure 7 shows the disparity image and reconstructed 3D points obtained with the SVS package [12].

Using the vertical reference provided by the inertial sensors, in this case $\boldsymbol{n} \approx (-0.456, -0.022, 0.890)$, the 3D points were transformed to a world aligned frame of reference as previously described.

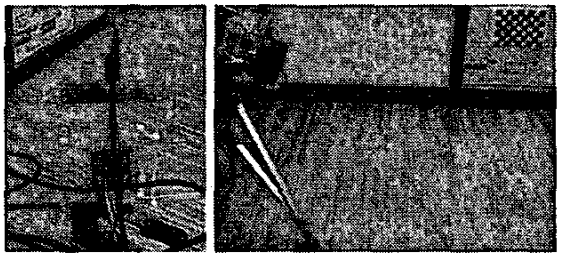

Figure 5: *Experimental setup with inertial sensors and vision system, and scene used for the test.*

Figure 6: *Stereo rectified image pair obtained with SVS [12] system.*

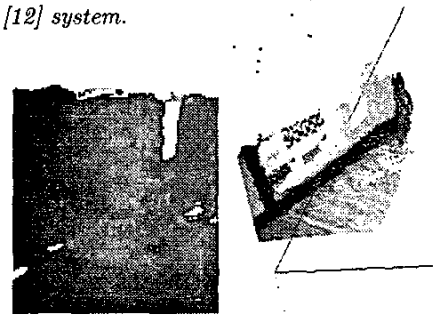

Figure 7: *Disparity image obtained with SVS [12], and reconstructed 3D points*

In order to detect the ground plane, an histogram was done for all depths, and the peak used as a reference value, as seen in figure 8. The points were than parsed and segmented as ground plane points and points above ground.

Figure 9 shows the graphical front-end of the implemented system working realtime at 10 frames per second. On the left the height histogram is shown.

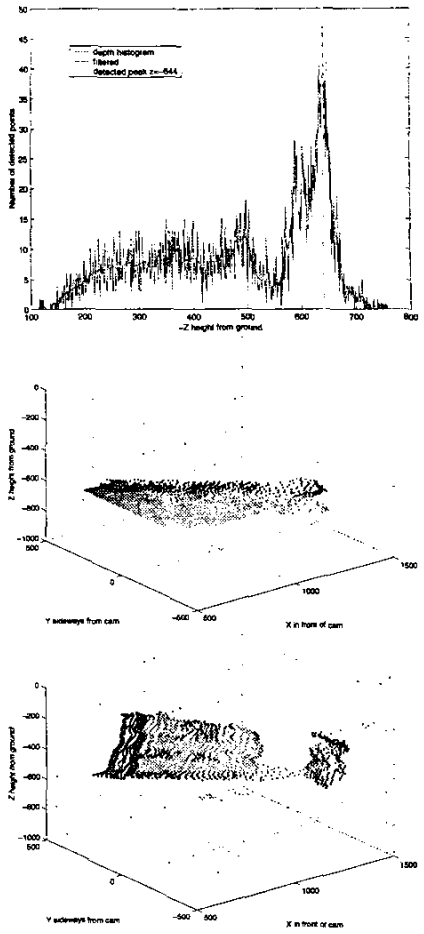

Figure 8: *Depth histogram with detected peak; Ground plane points; Points above the floor, walls or obstacles.*

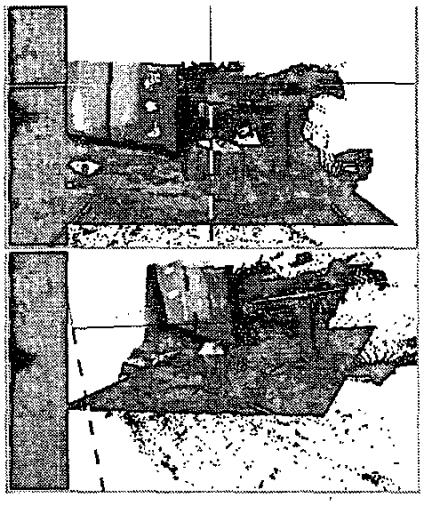

Figure 9: *Graphical front-end with height histogram and segmented depth map.*

A linear line fit was done using the points above ground from the data set in figure 8, ignoring their depth, to reconstruct the wall orientation in the test scene. Figure 10 shows the result. More complex scenes require a previous point clustering stage, so that a simplified world model can be built, but this only has to be done in 2D.

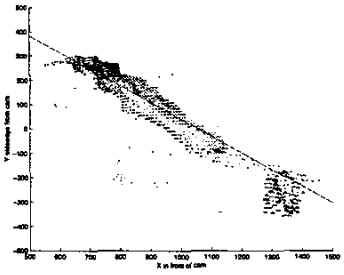

Figure 10: *Top view of all points above the floor, and line fit for wall orientation.*

Acknowledgments

This work was sponsored by the Tele3D project from *Fundação para a Ciência e Tecnologia-FCT*, Portugal) and the European Laboratory for Particle Physics (*Centre Européen pour la Recherche Nucléaire-CERN*).

7 Conclusions

Depth maps obtained from a stereo camera system were segmented using a vertical reference provided by inertial sensors, identifying structures such as vertical features and level planes. Rectifying the maps to a reference ground plane enables the segmentation of vertical and horizontal geometric features. Preliminary results were presented that show the validity of the method.

The aim of this work is a fast real-time system, avoiding 3D point clustering methods that are not suitable for real-time implementations. It can be applied to an automated car driving system, modelling the road, identifying obstacles and roadside features.

References

[1] H. Carpenter, *Movements of the Eyes*, London Pion Limited, 2nd edition, 1988, ISBN 0-85086-109-8.

[2] A. Berthoz, *The Brain's Sense of Movement*, Havard University Press, 2000, ISBN: 0-674-80109-1.

[3] J. Lobo and J. Dias, "Fusing of image and inertial sensing for camera calibration", in *Proc. Int. Conf. on Multisensor Fusion and Integration for Intelligent Systems*, Baden-Baden, Germany, August 2001, pp.103-108.

[4] J. Lobo and J. Dias, "Ground plane detection using visual and inertial data fusion", in *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, Victoria, Canada, October 1998, pp.912-917.

[5] J. Lobo, C. Queiroz, and J. Dias, "Vertical world feature detection and mapping using stereo vision and accelerometers", in *Proc. 9th Int. Symp. on Intelligent Robotic Systems*, Toulouse, France, July 2001, pp.229-238.

[6] J.-Lobo and J. Dias. "Integration of Inertial Information with Vision towards Robot Autonomy", in *Proc. IEEE Int. Symp. on Industrial Electronics*, Guimaraes, Portugal, July 1997, pp 825-830.

[7] H.K. Nishihara, "Real-time implementation of a signcorrelation algorithm for image-matching", (Draft) Teleos Research, February, 1990.

[8] J. Webb, "Implementation and Performance of Fast Parallel Multi-baseline Stereo Vision", in *Proc. Image Understanding Workshop*, 1993, pp.1005-1012.

[9] L.H. Matthies, "Stereo vision for planetary rovers: stochastic modelling to near real time implementation", *Int. Journal of Computer Vision*, 1992, 8(1), pp.71-91.

[10] O. Faugeras et al, "Real time correlation based stereo: algorithm, implementations and applications", Research Report 2013, INRIA Sophia-Antipolis, 1993.

[11] T. Kanade, H. Kano, and S. Kimura, "Development of a video-rate stereo machine", in *Proc. Int. Robotics and System Conf.*, Pittsburg (PA), 1995.

[12] K. Konolige, "Small Vision Systems: Hardware and Implementation", *8th Int. Symp. on Robotics Research*, Hayama, Japan, October 1997.

[13] T. Viéville and O.D. Faugeras, "Computation of Inertial Information on a Robot", in H. Miura and S. Arimoto, editors, *5h Int. Symp. on Robotics Research*, MIT-Press, 1989, pp.57-65.

[14] T. Viéville and O.D. Faugeras, "Cooperation of the Inertial and Visual Systems", in T.C. Henderson, editor, *Traditional and NonTraditional Robotic Sensors*, volume F 63 of *NATO ASI*, SpringerVerlag, 1990, pp.339-350.

[15] T. Viéville, et. al., "Autonomous navigation of a mobile robot using inertial and visual cues", in M. Kikode, T. Sato, and K. Tatsuno, editors, *Intelligent Robots and Systems*, Yokohama, 1993.

[16] T. Viéville, E. Clergue, and P.E.D. Facao, "Computation of ego-motion and structure from visual an inertial sensor using the vertical cue", in *ICCV93*,1993, pp. 591-598.

[17] B. Bhanu, B. Roberts, and J. Ming, "Inertial Navigation Sensor Integrated Motion Analysis for Obstacle Detection", in *Proc. IEEE Int. Conf. on Robotics and Automation*, Cincinnati, Ohio, USA, 1990, pp.954-959.

[18] F. Panerai and G. Sandini, "Oculo-Motor Stabilization Reflexes: Integration of Inertial and Visual Information". *Neural Networks*, 11(7-8), 1998, pp.1191-1204.

[19] F. Panerai, G. Metta, and G. Sandini, "Visuo-inertial stabilization in space-variant binocular systems". *Robotics and Autonomous Systems*, 30(1-2), 2000, pp.195-214.

[20] T. Mukai and N. Ohnishi, "Object Shape and Camera Motion Recovery Using Sensor Fusion of a Video Camera and a Gyro Sensor", *Information Fusion*, 1(1), 2000, pp.45-53.

[21] D.J. Coombs, "Real-time Gaze Holding in Binocular Robot Vision", Ph.D. Dissertation, Dept. Computer Science, Univ. Rochester, June, 1992.

[22] M.Jenkin, J. Tsotos, and G. Dudek, "The horopter and active cyclotorsion", in *Proc. IEEE International*, 1994.

[23] C.F.M. Weiman, "Log-polar vision system", Technical report, NASA,1994.

[24] G. Xu and Z. Zhang, *Epipolar Geometry in Stereo, Motion and Object Recognition - A unified approach*, Kluwer Academic publishers, 1996, ISBN 0-7923-4199-6.