Experiments in Visual-based Navigation with an Omnidirectional Camera

Niall Winters[1,2], José Gaspar[2], Etienne Grossmann[2] and José Santos-Victor[2]

[1]Computer Vision and Robotics Group,
Department of Computer Science,
University of Dublin, Trinity College,
Dublin 2 - Ireland.
Niall.Winters@cs.tcd.ie

[2]Instituto de Sistemas e Robótica,
Instituto Superior Técnico,
Av. Rovisco Pais, 1,
1049-001 Lisboa - Portugal.
{jasv,etienne,jag}@isr.ist.utl.pt

Abstract

This paper overviews experiments in autonomous visual-based navigation undertaken at the Instituto de Sistemas e Robótica. By considering the precise nature of the robot's task, we specify a navigation method which fulfills the environmental and localization requirements best suited to achieving this task. On-going research into task specification using 3D models, along with improvements to our topological navigation method are presented. We show how to build 3D models from images obtained with an omnidirectional camera equipped with a spherical mirror, despite the fact that it does not have a single projection centre.

1. Introduction

The problem of navigation is central to the successful application of mobile robots. Autonomous visual-based navigation can be achieved using catadioptric vision sensors. Significantly, such a sensor can generate diverse views of the environment. We have utilized omnidirectional, panoramic and bird's-eye views (i.e. orthographic images of the ground plane) to achieve a variety of navigation tasks [8].

The motivation for our research comes from studies of animal navigation. These suggest that most species utilize a very parsimonious combination of perceptual, action and representational strategies [3] that lead to much more efficient solutions when compared to those of today's robots. For example, when entering one's hall door one needs to navigate with more precision than when walking along a city avenue. This observation of a *path distance/accuracy trade-off* is a fundamental aspect of our work [8, 17], with each mode requiring a specific environmental representation and localization accuracy.

For traveling long distances within an indoor environment our robot does not rely upon knowledge of its exact position [18]. Instead, fast qualitative positioning is undertaken in a *topological* context by matching the current *omnidirectional image* to an a priori acquired set, i.e. an appearance based approach to the problem using PCA [11]. For tasks requiring very precise movements, e.g. docking, we developed a method termed *Visual Path Following* [7] for tracking visual landmarks in *bird's-eye views* of a scene. We can easily integrate these approaches in a natural manner. In one experiment the robot navigates through the Computer Vision Lab, traverses the door, travels down a corridor and then returns to its starting position.

It is important to note that for successful completion of such a task, user-specified information is required. While our robot is autonomous it is not *independent*. This constraint lead us to investigate a method whereby a user could specify particular subtasks in as *intuitive* a manner as possible. This is the subject of on-going research. We wished to solve this problem by designing an effective human-robot interface using only omnidirectional images. In this paper, we show how to construct this complimentary human-robot interface module and how it fits into our existing navigation framework. Our solution is to reconstruct a 3D model of the environment from a *single* omnidirectional image. An intermediate step in this construction requires the generation of perspective images. Thus, we show how our sensor can be modeled by the central projection model [9], despite the fact that it does not have a single centre of projection [1].

Additionally, this paper presents an improvement to our topological navigation module. In our previous work, entire omnidirectional images where used for localization. Using a statistical method termed *Information Sampling* [19], we can now select the most dis-

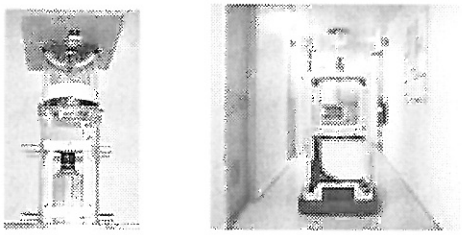

Figure 1: Left: the omnidirectional camera. Right: the camera mounted on the mobile robot.

criminatory data from a set of images acquired a priori. This not only gives us a speed increase but is also beneficial when dealing with such problems as partial occlusion and illumination change.

This paper is outlined as follows: Section 2 introduces the Information Sampling method and presents associated results. Section 3 details our human-robot interface in addition to showing how the central projection model can model our spherical sensor. Results are also given. In Section 4 we present our conclusions and the future direction of our research.

2. Information Sampling for Navigation

Information Sampling is a statistical method to extract the most discriminatory data from a set of omnidirectional images acquired a priori. We rank this data from most to least discriminatory. For more details on ranking methods, see [19]. Then, using only the highest ranking data we build a *local appearance space* which the robot utilizes for autonomous navigation. Theoretically, Information Sampling can be applied on a *pixel-by-pixel basis to any type of image*. For computational reasons, we use windows instead of pixels, extracted from omnidirectional images. Our set up is shown in Figure 1.

2.1. Related Work

Ohba [12] presents a method to divide images into a number of smaller windows. Eigenspace analysis was then applied to each window. Thus, even if a number of windows become occluded, the remaining ones contain enough information to perform the given task. Unfortunately, this method requires storage of a very large number of image windows and the chances of one window, acquired at runtime being matched to a number of images from the a priori set is high. Therefore, it is highly desirable that only the most *discriminatory* windows are selected from each acquired image. A number of solutions have been proposed including

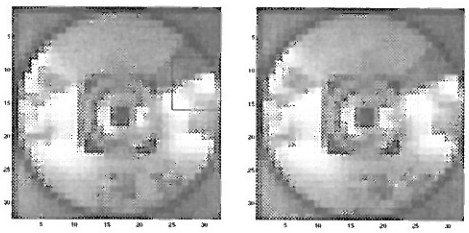

Figure 2: Left: A 32×32 omnidirectional image acquired at runtime and its most discriminatory 8×8 information window. Right: Its reconstruction using this information window.

the use of interest operators [14], pre-defined grids [4] or criteria such as detectability, uniqueness and reliability [12]. It should be noted that unlike the other solutions cited here, Information Sampling does not rely on eigenspace analysis or the use of interest operators.

2.2. Information Sampling: Overview

Essentially, Information Sampling can be described as a process for (i) reconstructing an entire image from the observation of a few (noisy) pixels and (ii) determining the *most relevant* image pixels, in the sense that they convey the most information about the image set.

Part (i) selects a number of pixels, d to test according to a selection matrix, S. The problem is to estimate an (entire) image based on these partial (noisy) observations of the pixels, d. Once we have this information, we can calculate the maximum a posteriori estimate of an image, $\hat{I}_{MAP}$ [13] as follows:

$$\hat{I}_{MAP} = \arg\max_{I} p(I|d) = \frac{(\Sigma_I^{-1} + S^T\Sigma_n^{-1}S)}{(\Sigma_I^{-1}\bar{I} + S^T\Sigma_n^{-1}d)} \quad (1)$$

In order to determine the best selection of pixels in the image, part (ii) involves computing the error associated with the above reconstruction. This error is calculated as the determinant of the error covariance matrix: $\Sigma_{error} = \mathrm{Cov}(I - \hat{I}_{MAP})$. In essence, we wish to determine the S that minimizes (in some sense) Σ_{error}. This will be our information window. Notice that the information criterion is based on the entire set of images and not, as with other methods, on an image-by-image basis. More details on Information Sampling can be found in [19].

Figure 2 (left) shows an omnidirectional image acquired at runtime and (right) its reconstruction using only the *most discriminatory* information window. This illustrates the power of our method. Results are detailed in the following section.

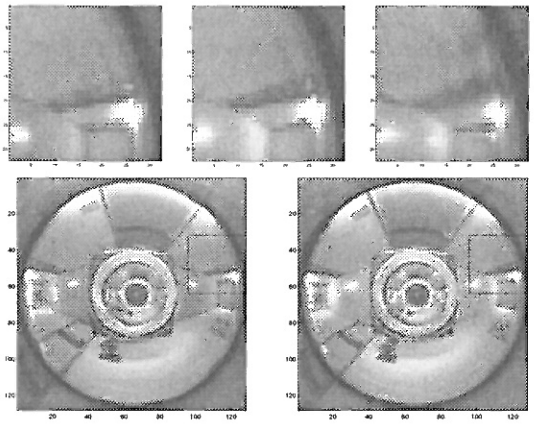

Figure 3: Top: Close-up of the 32 × 32 information windows: unknown (left), closest (middle) and reconstructed (right). Bottom: The position of the unknown and closest images in their respective omnidirectional images.

2.3. Position Estimation Results

Once the best window was determined using *Information Sampling*, we built a local appearance space [19, 4] using only the data selected with this window from each a priori image. This additionally compressed our data to only approximately one thousandth of the original 128 × 128 image data. Successful position estimation has been achieved using windows as small as 4×4 pixels in size.

We projected *only* the selected windows from each image into the eigenspace. This is an improvement on previous approaches, where all windows first had to be projected. Thus, we were able to immediately reduce the ambiguity associated with projection. Figure 3 top row, from left to right, shows the most relevant information window from an unknown image, its closest match from the a priori set of omnidirectional images and its reconstruction. The bottom row shows the information window in the unknown 128 × 128 image (left) and its closest match from the a priori set, obtained by projecting only the most relevant information window (right). We note here that we could in principal, given enough computing power, use equation (1) to reconstruct a 128 × 128 *image* using only the most relevant *window*.

Figure 4 shows the distance between images acquired at run time and those acquired a priori. The global minimum is the correct estimate of the robot's current topological position. Local minima correspond to images similar to the current one, some distance away from the robot's position. Figure 4 (left) shows the graph obtained using 16 × 16 images, while Figure 4 (right) that obtained using the most discriminating 4 × 4 image window. While different local minima are obtained by both methods, the global minimum is maintained. Thus, we have shown that effective navigation can be undertaken using only the most informative 16 pixels from each image.

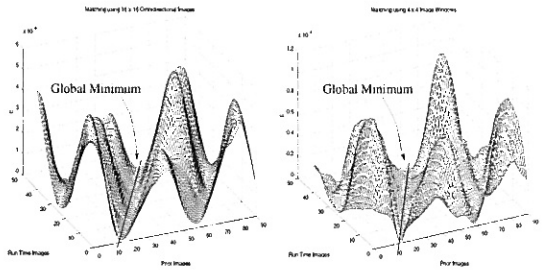

Figure 4: A 3D plot of images acquired at run time versus those acquired a priori using (top) 16×16 *images* and (bottom) 4 × 4 *information windows*, respectively.

3. Human Robot Interface

Once we had developed an effective method for autonomous qualitative robot navigation along a topological map, we turned our attention to developing an intuitive user interface from which to select subtasks. While final experiments have yet to be undertaken, in this section we show how to construct this interface.

Our interface consists of a 3D interactive reconstruction of a structured environment obtained from a *single* omnidirectional image. In order to build such a model, we first have to generate perspective images from our catadioptric sensor, despite the fact that it does *not* have a single centre of projection, since it utilizes a spherical mirror. The benefit of such an interface is that even though it does not provide very fine details of the environment, it provides the user with a sufficiently rich description that may easily be transmitted over a low bandwidth link.

3.1. Catadioptic Sensor Modeling

In [9] Geyer and Daniilidis present a central projection model for all catadioptric systems with a *single projection centre*. This model combines a mapping to a sphere followed by a projection to a plane. The projection of a point in space (x, y, z) to an image point (u, v) can be written as:

$$\begin{bmatrix} u \\ v \end{bmatrix} = \frac{l+m}{l \cdot r - z} \begin{bmatrix} x \\ y \end{bmatrix} = \mathcal{P}(x, y, z; l, m) \qquad (2)$$

$$r = \sqrt{x^2 + y^2 + z^2}$$

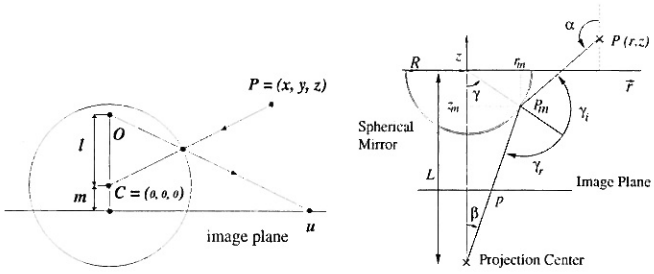

Figure 5: Left: Central Projection Model. Right: Spherical Mirror.

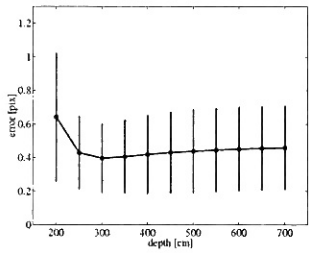

Figure 6: Mean absolute error between the central projection model and the projection with a spherical mirror. Vertical bars indicate the standard deviation.

where l and m, represent the distance from the sphere centre to the projection centre, O, and the distance from the sphere centre to the plane, respectively. This is graphically represented by Figure 5 (left).

On the other hand, catadioptric sensors which use a spherical mirror, shown in Figure 5 (right), are essentially modelled by the equation of reflection at the mirror surface, $\gamma_i = \gamma_r$ [7].

A camera with a spherical mirror cannot be exactly represented by the central projection model. In order to find an approximate representation, we compare the image projection error, instead of analyzing the projection centre itself.

Let $\mathcal{P}(x_i, y_i, z_i; \theta)$ denote the central projection defined in equation (2) and $\mathcal{P}_c$ be the projection with a spherical mirror. Grouping the geometric and intrinsic parameters together - θ and θ_c for the former and latter projections - we want to minimize the cost functional associated with the image projection error:

$$\hat{\theta} = \arg_\theta \min \sum_i \|\mathcal{P}(x_i, y_i, z_i; \theta) - \mathcal{P}_c(x_i, y_i, z_i; \theta_c)\|^2$$

The minimization of this cost functional gives the desired parameters, θ for the central projection model, $\mathcal{P}$ which approximates the spherical sensor characterized by $\mathcal{P}_c$ and θ_c. Despite the fact that projection using spherical mirrors cannot be represented by the central projection model, Figure 6 shows that for a certain operational range, this theory gives a very good approximation. The error is less than 1 pixel.

3.2. Forming Perspective Images

The acquisition of correct perspective images, independent of the scenario, requires that the vision sensor be characterised by a single projection centre [1]. By definition the spherical projection model has this property but due to the intermediate mapping over the sphere the images obtained are, in general, not perspective.

In order to obtain correct perspective images, the spherical projection must first be reversed from the image plane to the sphere surface and then, re-projected to the desired plane from the sphere centre. We term this reverse projection *back-projection* after Sturm in [15, 16]. The back-projection of an image pixel (u, v), obtained through spherical projection, yields a 3D direction $k \cdot (x, y, z)$ given by the following equations, derived from equation (2):

$$a = (l + m), b = (u^2 + v^2)$$

$$\begin{bmatrix} x \\ y \end{bmatrix} = \frac{la - sign(a)\sqrt{(1 - l^2)b + a^2}}{a^2 + b} \begin{bmatrix} u \\ v \end{bmatrix} \quad (3)$$

$$z = \pm\sqrt{1 - x^2 - y^2}$$

where z is negative when $|l + m| / l > \sqrt{u^2 + v^2}$, and positive otherwise. It is assumed, without loss of generality, that (x, y, z) lies on the surface of the unit sphere.

Note that the set $\{(x, y, z)\}$, interpreted as points of the projective plane, already define a perspective image. However for displaying or obtaining specific viewing directions further development is required.

Letting R denote the orientation of the desired (pinhole) camera relative to the frame associated with the results of back-projection, the new perspective image $\{(\lambda u, \lambda v, \lambda)\}$ becomes:

$$\begin{bmatrix} \lambda u \\ \lambda v \\ \lambda \end{bmatrix} = K \cdot R^{-1} \begin{bmatrix} x \\ y \\ z \end{bmatrix} \quad (4)$$

where K contains the intrinsic parameters and λ is a scaling factor. This is the pin-hole camera projection model [5], when the origin of the coordinates is the camera centre.

3.3. Interactive 3D Reconstruction

We now describe the method used to obtain 3D models given the following information: a perspective image (obtained from an omnidirectional camera with a spherical mirror), a camera orientation obtained from vanishing points [2] and some limited user input [6, 10].

Let $p = [u\ v\ 1]^T$ be the projection of a 3D point $[C\ C'\ C''\ 1]^T$ that we want to reconstruct. Then, if we consider a normalized camera [5], we have the following:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \lambda [r_1\ r_2\ r_3\ 0] \begin{bmatrix} C \\ C' \\ C'' \\ 1 \end{bmatrix} \quad (5)$$

$$= \lambda (Cr_1 + C'r_2 + C''r_3)$$

where r_1, r_2, r_3 are vanishing points. As is usual, we choose 0 as the origin of the coordinates for reconstruction. Next, we define lines towards vanishing points:

$$l_i = r_i \times [u\ v\ 1]^T, \quad i = 1 \ldots 3. \quad (6)$$

Then, using the cross and internal products property that $(r \times p)^T.p = 0$, we obtain:

$$l_i^T.r_1 C + l_i^T.r_2 C' + l_i^T.r_3 C'' = 0 \quad (7)$$

which is a linear system in the coordinates of the 3D point. This can be rewritten as:

$$\begin{bmatrix} 0 & l_1^T.r_2 & l_1^T.r_3 \\ l_2^T.r_1 & 0 & l_2^T.r_3 \\ l_3^T.r_1 & l_3^T.r_2 & 0 \end{bmatrix} \begin{bmatrix} C \\ C' \\ C'' \end{bmatrix} = 0. \quad (8)$$

Generalising this system to N points we again obtain a linear system:

$$A.\mathcal{C} = 0 \quad (9)$$

where $\mathcal{C}$ contains the $3N$ tridimensional coordinates that we wish to locate and A is block diagonal, where each block has the form shown in equation (8). Thus, A is of size $3N \times 3N$. Since only two equations from the set defined by equation (8) are independent, the co-rank of A is equal to the number of points N. As expected, this shows that there is an unknown scale factor for each point.

Now, adding some limited user input, in the form of co-planarity or co-linearity point constraints, a number of 3D coordinates become equal and thus the number of columns of A may be reduced. As a simple example, if we have 5 points we have 5×3 free coordinates, i.e the number of columns of A. Now, if we impose the constraint that points P_1, P_2, P_3 are co-planar, with constant z value and points P_4, P_5 are co-linear, with a constant (x, y) value, then the total number of free coordinates is reduced from the initial 15 to 12. Thus, the coordinates P_{2z}, P_{3z}, P_{5x}, P_{5y} are dropped from the linear system defined by equation (9).

Given sufficient user input, the co-rank of A becomes 1. In this case, the solution of the system will give a reconstruction with no ambiguity other than that - well known - of scale [10].

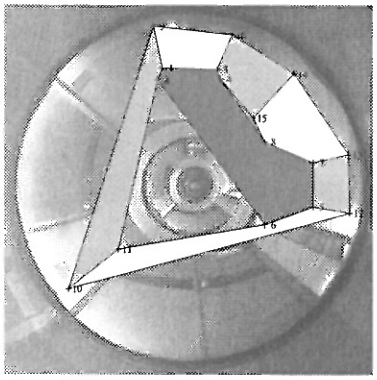

Figure 7: User defined planes orthogonal to the x axis (light gray), y axis (white) and z axis (dark gray).

Axis	Planes	Lines
x	$(1,2,9,10,11)$, $(3,4,14,15,16)$, $(5,7,12,13)$	
y	$(5,6,10,11,12)$, $(7,8,13,14,15)$, $(1,3,9,16)$	$(1,2)$
z	$(1,2,3,4,5,6,7,8)$, $(9,12,13,16)$	

Table 1: User-defined planes and lines. The numbers are the indexes of the image points shown in Figure 7. The first column indicates the axis to which the planes are orthogonal and the lines are parallel.

3.4. Developing the Human Robot Interface: Results

Figure 7 shows an omnidirectional image and superposed user input. This input consists of the 16 points shown, knowledge that sets of points belong to constant x, y or z planes and that other sets belong to lines parallel to the x, y or z axes. Table 1 details all the user-defined data. Planes orthogonal to the x and y axes are in light gray and white respectively, and one horizontal plane is shown in dark gray (the topmost horizontal plane is not shown as it would occlude the other planes). The coordinates in the original image were transformed to the equivalent pin-hole model coordinates and used for reconstruction. Figure 8 shows the resulting texture-mapped reconstructions. These results are interesting given that they required only a single image and limited user input to reconstruct the surroundings of the sensor.

4. Conclusions & Future Work

This paper presented experiments in visual-based navigation using an omnidirectional camera. We showed improvements to our topological module by using Information Sampling. We explained how our sensor,

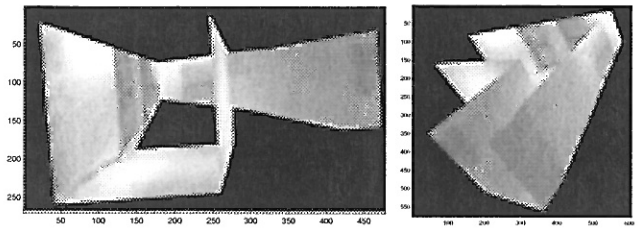

Figure 8: Views of the reconstructed 3D model.

despite the fact that it does not have a single projection centre, can be modeled by the central projection model. Then, using *single* omnidirectional images and some limited user input we showed how to build a 3D model of the environment. This acted as a basis for an intuitive human-robot interface.

In terms of our future work, we plan on extending the implementation of Information Sampling. Additionally, creation of large scene models shall be achieved by fusing different models together. We plan on carrying out extended closed-loop control experiments verifying the applicability of our navigation framework.

Acknowledgements

This work was partly funded by the European Union RTD - Future and Emerging Technologies Project Number: IST-1999-29017, Omniviews.

References

[1] S. Baker and S.K. Nayar. A theory of single-viewpoint catadioptric image formation. *International Journal of Computer Vision*, 35(2):175–196, 1999.

[2] B. Caprile and V. Torre. Using vanishing points for camera calibration. *International Journal of Computer Vision*, 4:127–140, 1990.

[3] T. S. Collett, E. Dillmann, A. Giger, and R. Wehner. Visual landmarks and route following in the desert ant. *J. Comp. Physiology A*, 170:435–442, 1992.

[4] V. Colin de Verdière and J. L. Crowley. Visual recognition using local appearance. In *5th European Conference on Computer Vision, (ECCV 1998)*, pages 640–654, Freiburg, Germany, June 1998.

[5] O. Faugeras. *Three Dimensional Computer Vision*. MIT Press, 1993.

[6] J. Gaspar, E. Grossmann, and J. Santos-Victor. Interactive reconstruction from an omnidirectional image. In *9th International Symposium on Intelligent Robotic Systems (SIRS'01)*, Toulouse, France, July 2001.

[7] J. Gaspar and J. Santos-Victor. Visual path following with a catadioptric panoramic camera. In *7th International Symposium on Intelligent Robotic Systems (SIRS'99)*, pages 139–147, Coimbra, Portugal, July 1999.

[8] J. Gaspar, N. Winters, and J. Santos-Victor. Vision-based navigation and environmental representations with an omni-directional camera. *IEEE Transactions on Robotics and Automation*, 16(6):890–898, December 2000.

[9] C. Geyer and K. Daniilidis. A unifying theory for central panoramic systems and practical applications. In *ECCV 2000*, pages 445–461, Dublin, Ireland, 2000.

[10] E. Grossmann, D. Ortin, and J. Santos-Victor. Reconstruction of structured scenes from one or more views: Nature of solutions. *Submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2001.

[11] H. Murase and S. K. Nayar. Visual learning and recognition of 3d objects from appearance. *International Journal of Computer Vision*, 14(1):5–24, January 1995.

[12] K. Ohba and K. Ikeuchi. Detectibility, uniqueness and reliability of eigen windows for stable verification of partially occluded objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(9):1043–1048, September 1997.

[13] A. P. Sage and J. L. Melsa. *Estimation Theory with Applications to Communications and Control*. McGraw-Hill, 1971.

[14] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5):530–535, May 1997.

[15] P. Sturm. A method for 3d reconstruction of piecewise planar objects from single panoramic images. In *1st International IEEE Workshop on Omnidirectional Vision at CVPR*, pages 119–126, 2000.

[16] P. Sturm and S. Maybank. A method for interactive 3d reconstruction of piecewise planar objects from single images. In *British Machine Vision Conference*, pages 265–274, 1999.

[17] N. Winters, J.Gaspar, G.Lacey, and J. Santos-Victor. Omni-directional vision for robot navigation. In *1st International IEEE Workshop on Omni-directional Vision at CVPR*, pages 21–28, 2000.

[18] N. Winters and J.Santos-Victor. Omni-directional visual navigation. In *7th International Symposium on Intelligent Robotics Systems (SIRS'99)*, pages 109–118, Coimbra, Portugal, July 1999.

[19] N. Winters and J.Santos-Victor. Information sampling for optimal image data selection. In *9th International Symposium on Intelligent Robotics Systems (SIRS'01)*, Toulouse, France, July 2001.