2nd Workshop on
Biological Motivated Computer Vision,
Tuebingen, Germany, Nov. 2002.

A Binocular Stereo Algorithm for Log-polar Foveated Systems

Alexandre Bernardino and José Santos-Victor

Instituto Superior Técnico
ISR - Torre Norte, Piso 7
Av. Rovisco Pais, 1049-001 Lisboa, Portugal
{alex,jasv}@isr.ist.utl.pt

Abstract. Foveation and stereopsis are important features on active vision systems. The former provides a wide field of view and high foveal resolution with low amounts of data, while the latter contributes to the acquisition of close range depth cues. The log-polar sampling has been proposed as an approximation to the foveated representation of the primate visual system. Although the huge amount of stereo algorithms proposed in the literature for conventional imaging geometries, very few are shown to work with foveated images sampled according to the log-polar transformation. In this paper we present a method to extract dense disparity maps in real-time from a pair of log-mapped images, with direct application to active vision systems.

1 Introduction

Stereoscopic vision is a fundamental perceptual capability both in animals and artificial systems. At close ranges, it allows reliable extraction of depth information, thus being suited for robotics tasks such as manipulation and navigation. In the last decades a great amount of research has been directed to the problem of extracting depth information from stereo imagery (see [25] for a recent review). However, the best performing techniques are still too slow to use on robotic systems which demand real-time operation. The straightforward way to reduce computation time is to work with coarse resolution images but this restricts the acquisition of detailed information all over the visual field. A better solution, inspired in biological systems, is the use of ocular movements together with foveated retinas. The visual system of primates has a space-variant nature where the resolution is high on the fovea (the center of the retina) and decreases gradually to the periphery of the visual field. This distribution of resolution is the evolutionary solution to reduce the amount of information traversing the optical nerve while maintaining high resolution in the fovea and a wide visual field. Moving the high resolution fovea we are able to acquire detailed representations of the surrounding environment. The excellent performance of biological visual systems led researchers to investigate the properties of foveated systems. Many active vision systems have adopted this strategy and since foveated images contain less information than conventional uniform resolution images, one obtains important reductions on the computation time.

We may distinguish between two main methods to emulate foveated systems, that we denote by *multi-scale uniform sampling* methods and *non-uniform sampling* methods. *Uniform* methods preserve the cartesian geometry of the representation by performing operations at different scales in multi-resolution pyramids (e.g. [17],[10],[13]). Sampling grids are uniform at each level but different levels have different spacing and receptive field size. Notwithstanding, image processing operations are still performed on piecewise uniform resolution domains. *Non-uniform* methods resample the image with non-linear transformations, where receptive field spacing and size are non-uniform along the image domain. The VR transform [2], the DIEM method [19], and several versions of the *logmap* [30], are examples of this kind of methods.

The choice of method is a matter of preference, application dependent requirements and computational resources. *Uniform methods* can be easier to work with, because many current computer vision algorithms can be directly applied to these representations. However, *non-uniform* methods can achieve more compact image representations with consequent benefits in computation time. In particular the *logmap* has been shown to have many additional properties like rotation and scale invariance [31], easy computation of time-to-contact [28], improved linear flow estimation [29], looming detection [23], increased stereo resolution on verging systems [14], fast anisotropic diffusion [11], improved vergence control and tracking [7, 3, 4].

Few approaches have been proposed to compute disparity maps for foveated active vision systems, and existing ones rely on the foveated pyramid representation [17, 27, 6]. In this paper we describe a stereo algorithm to compute dense disparity maps on *logmap* based systems. Dense representations are advantageous for object segmentation and region of interest selection. Our method uses directly the gray/color values of each pixel, without requiring any feature extraction, making this method particularly suited for non-cartesian geometries, where the scale of analysis depends greatly on the variable to estimate (disparity).

To our knowledge, the only work to date addressing the computation of stereo disparity in *logmap* images is [15]. In that work, disparity maps are obtained by matching laplacian features in the two views (zero crossing), which results in sparse disparity maps.

2 Real-Time Log-polar Mapping

The log-polar transformation, or *logmap*, $\mathbf{l}(\mathbf{x})$, is defined as a conformal mapping from the *cartesian* plane $\mathbf{x} = (x, y)$ to the *log-polar* plane $\mathbf{z} = (\xi, \eta)$:

$$\mathbf{l}(\mathbf{x}) = \begin{bmatrix} \xi \\ \eta \end{bmatrix} = \begin{bmatrix} \log(\sqrt{x^2 + y^2}) \\ \arctan \frac{y}{x} \end{bmatrix} \tag{1}$$

Since the *logmap* is a good approximation to the retino–cortical mapping in the human visual system [26, 12], the cartesian and log-polar coordinates are also called "retinal" and "cortical", respectively. In continuous coordinates, a

cortical image I^{cort} is obtained from the corresponding retinal image I by the warping:

$$I^{cort}(\mathbf{z}) = I(\mathbf{l}^{-1}(\mathbf{x}))$$

A number of ways have been proposed to discretize space variant maps [5]. We have been using the *logmap* for some years in real-time active vision applications [3, 4]. To allow real-time computation of *logmap* images we partition the retinal plane into *receptive fields*, whose size and position correspond to a uniform partition of the cortical plane into *super-pixels* (see Fig. 1). The value of a *super-pixel* is given by the average of all pixels in the corresponding receptive field.

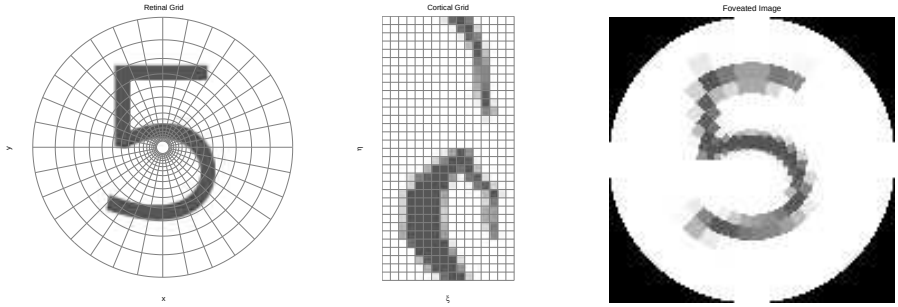

Fig. 1. The log-polar sampling scheme is implemented by averaging the pixels contained within each of the *receptive fields* shown in the left image. These space-variant receptive fields are angular sections of circular rings corresponding to uniform rectangular *super-pixels* in the cortical image (center). To reconstruct the retinal image, each receptive field gets the value of the corresponding *super-pixel* (right).

3 Disparity map computation

We start describing an intensity based method to find the likelihood of stereo matches in usual cartesian coordinates, $\mathbf{x} = (x, y)$. Then we show how the method can be extended to cope with *logmap* images. Finally we describe the remaining steps to obtain the disparity maps.

Let I and I' be the left and right images, respectively. For depth analysis, we are interested in computing the horizontal disparity map, but since we consider a general head vergence configuration, vertical disparities must also be accounted for. Therefore, disparity is a two valued function defined as $\mathbf{d}(\mathbf{x}) = (d_x, d_y)$. Taking the left image as the reference, the disparity at point $\mathbf{x}$ is given by $\mathbf{d}(\mathbf{x}) = \mathbf{x}' - \mathbf{x}$, where $\mathbf{x}$ and $\mathbf{x}'$ are the locations of matching points in the left and right images. If a pixel at location $\mathbf{x}$ in the reference image is not visible in the right image, we say the pixel is occluded and disparity is undefined ($\mathbf{d}(\mathbf{x}) = \varnothing$).

3.1 Bayesian formulation

To obtain dense representations, we use an intensity based method similar to [32]. We formulate the problem in a discrete bayesian framework. Having a finite set of possible disparities, $D = \{\mathbf{d}_n\}, n = 1 \cdots N$, for each location $\mathbf{x}$ we define a set of hypothesis, $H = \{h_n(\mathbf{x})\}, n = 0 \cdots N$, where $h_0(\mathbf{x})$ represents the occlusion condition ($\mathbf{d}(\mathbf{x}) = \varnothing$), and the other h_n represent particular disparity values, $\mathbf{d}(\mathbf{x}) = \mathbf{d}_n$. Other working assumptions are the following:

1. Object appearance does not vary with view point (lambertian surfaces) and cameras have the same gain, bias and noise levels. This corresponds to the *Brightness Constancy Assumption* [16]. Considering the existence of additive noise, we get the following stereo correspondence model:

$$I(\mathbf{x}) = I'(\mathbf{x} + \mathbf{d}(\mathbf{x})) + \eta(\mathbf{x}) \tag{2}$$

2. Noise is modeled as being independent and identically distributed with a certain probability density function, f. In the unoccluded case, the probability of a certain gray value $I(\mathbf{x})$ is conditioned by the value of the true disparity $\mathbf{d}(\mathbf{x})$ and the value of I' at position $\mathbf{x} + \mathbf{d}(\mathbf{x})$:

$$Pr(I(\mathbf{x})|\mathbf{d}(\mathbf{x})) = f(I(\mathbf{x}) - I'(\mathbf{x} + \mathbf{d}(\mathbf{x})))$$

We assume zero-mean gaussian white noise, and have $f(t) = 1/\sqrt{2\pi\sigma^2}e^{-t^2/2\sigma^2}$ where σ^2 is the noise variance.

3. In the discrete case we define the *disparity likelihood* images as:

$$L_n(\mathbf{x}) = Pr(I(\mathbf{x})|h_n(\mathbf{x})) = f(I(\mathbf{x}) - I'_n(\mathbf{x})) \tag{3}$$

where $I'_n(\mathbf{x}) = I'(\mathbf{x} + \mathbf{d_n})$ are called *disparity warped images.*

4. The probability of a certain hypothesis given the image gray levels (posterior probability) is given by the Bayes' rule:

$$Pr(h_n|I) = \frac{Pr(I|h_n)Pr(h_n)}{\sum_{i=0}^{N} Pr(I|h_i)Pr(h_i)} \tag{4}$$

where we have dropped the argument $\mathbf{x}$ since all functions are computed at the same point.

5. If a pixel at location $\mathbf{x}$ is occluded in the right image, its gray level is unconstrained and can have any value in the set of M admissible gray values,

$$Pr(I|h_0(\mathbf{x})) = \frac{1}{M} \tag{5}$$

We define a prior probability of occlusion with a constant value for all sites:

$$Pr(h_0) = q \tag{6}$$

6. We do not favor any *a priori* particular value of disparity. A constant prior is considered and its value must satisfy $Pr(h_n) \cdot N + q = 1$, which results in:

$$Pr(h_n) = (1 - q)/N \tag{7}$$

7. Substituting the priors (5), (6), (7), and the likelihood (3) in (4), we get:

$$Pr(h_n|I) = \begin{cases} \frac{L_n(I)}{\sum_{i=1}^{N} L_i(I)+qN/(M-qM)} & \Leftarrow n \neq 0 \\ \frac{qN/(M-qM)}{\sum_{i=1}^{N} L_i(I)+qN/(M-qM)} & \Leftarrow n = 0 \end{cases} \tag{8}$$

The choice of the hypothesis that maximizes (8) leads us to the MAP (*maximum a posteriori*) estimate of disparity[1]. However, without any further assumptions, there may be many ambiguous solutions. It is known that in the general case, the stereo matching problem is under-constrained and ill-posed [25]. One way to overcome this fact is to assume that the scene is composed by piece-wise smooth surfaces and introduce spatial interactions between neighboring locations to favor smooth solutions. Later we will describe a cooperative spatial facilitation method to address this problem.

3.2 Cortical Likelihood Images

While in cartesian coordinates the *disparity warped images* can be obtained by shifting pixels by an amount independent of position, $\mathbf{x}' = \mathbf{x} + \mathbf{d}_n$, in cortical coordinates the disparity shifts are different for each pixel, as shown in Fig.2. Thus, for each cortical pixel and disparity value, we have to compute the corresponding pixel in the second image. Using the *logmap* definition (1), the cortical correspondences can be obtained by:

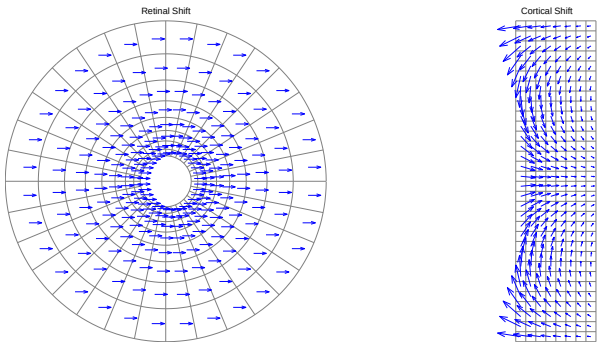

Fig. 2. A space invariant shift in retinal coordinates (left) corresponds to a space variant warping in the cortical array.

$$\mathbf{z}'_n(\mathbf{z}) = \mathbf{l}\Big(\mathbf{l}^{-1}(\mathbf{z}) + \mathbf{d_n}\Big) \tag{9}$$

This map can be computed off-line for all cortical locations and stored in a look-up table to speed-up on-line calculations. To minimize discretization errors, the

[1] The terms in the denominator are normalizing constants and do not need to be computed explicilty

weights for intensity interpolation can also be pre-computed and stored. A deeper explanation of this technique can be found in [22].

Using the pre-computed look up tables, the cortical *disparity warped images* can be efficiently computed on-line:

$$I_n^{cort'}(\mathbf{z}) = I^{cort'}(\mathbf{z}_n(\mathbf{z}))$$

From Eq. (3) we define $N+1$ *cortical likelihood images*, $L_n^{cort}(\mathbf{z})$, that express the likelihood of a particular hypothesis at cortical location $\mathbf{z}$:

$$L_n^{cort}(\mathbf{z}) = f(I^{cort}(\mathbf{z}) - I_n^{cort'}(\mathbf{z}))$$

Substituting this result in Eq. (8) we have the cortical posterior probabilities:

$$Pr^{cort}(h_n|I^{cort}) \propto \begin{cases} L_n^{cort}(I) & \Leftarrow n \neq 0 \\ qN/(M - qM) & \Leftarrow n = 0 \end{cases} \tag{10}$$

3.3 Cooperative spatial facilitation

The value of the likelihood images L_n^{cort} at each cortical location $\mathbf{z}$ can be interpreted as the response of disparity selective neurons, expressing the degree of match between corresponding locations in the right and left images. When many disparity hypothesis are likely to occur (e.g. textureless areas) several neurons tuned to different disparities may be simultaneously active. In a computational framework, this "aperture" problem is usually addressed by allowing neighborhood interactions between units, in order to spread information from and to non-ambiguous regions. A bayesian formulation of these interactions leads to Markov Random Fields techniques [33], whose existing solutions (annealing, graph optimization) are still computationally expensive. Neighborhood interactions are also very commonly found in biological literature and several cooperative schemes have been proposed, with different facilitation/inibhition strategies along the spatial and disparity coordinates [18, 21, 20]. For the sake of computational complexity we adopt a spatial-only facilitation scheme whose principle is to reinforce the output of units at locations whose coherent neighbors (tuned for the same disparity) are active. This scheme can be implemented very efficiently by convolving each of the *cortical likelihood images* with a low-pass type of filter, resulting on $N + 1$ *Facilitated Cortical Likelihood Images*, F_n^{cort}. We use a fast IIR isotropic separable first order filter, which only requires two multiplications and two additions per pixel. We prefer filters of large impulse response, which provide better smoothness properties and favor blob like objects, at the cost of missing small or thin structures in the image. Also, due to the space-variant nature of the cortical map, regions on the periphery of the visual field will have more "smoothing" than regions in the center.

At this point, it is worth noticing that since the 70's, biological studies show that neurons tuned to similar disparities are organized in clusters on visual cortex area V2 in primates [8], and more recently this organization has also been found on area MT [9]. Our architecture, composed by topographically organized maps of units tuned to the same disparity, agrees with these biological findings.

3.4 Computing the solution

Replacing in (10) the *cortical likelihood images* L_n^{cort} by their filtered versions F_n^{cort} we obtain $N+1$ *cortical disparity activation images*:

$$D_n^{cort} = \begin{cases} F_n^{cort}(I) & \Leftarrow n \neq 0 \\ qN/(M - qM) & \Leftarrow n = 0 \end{cases} \tag{11}$$

The disparity map is obtained by computing the hypothesis that maximizes the *cortical disparity activation* images for each location:

$$\hat{\mathbf{d}}(\mathbf{z}) = \arg\max_n (D_n^{cort}(\mathbf{z}))$$

In a neural networks perspective, this computation is analogous a winner-take-all competition between non-coherent units at the same spatial location, promoted by the existence of inhibitory connections between them [1].

4 Results

We have tested the proposed algorithm on a binocular active vision head in general vergence configurations, and on standard stereo test images. Results are shown on Figs. 3 and 4. Bright and dark regions correspond to near and far objects, respectively. The innermost and outermost rings present some noisy disparity values due to border effects than can be easily removed by simple post-processing operations.

Fig. 3. The image in the right shows the raw foveated disparity map computed from the pair of images shown in the left, taken from a stereo head verging on a point midway between the foreground and background objects.

Some intermediate results of the first experiment are presented in Fig. 5, showing the output of the cortical likelihood and the cortical activation for a particular disparity hypothesis. In the likelihood image notice the great amount of noisy points corresponding to false matches. The spatial facilitation scheme and the maximum computation over all disparities are essential to reject the false matches and avoid ambiguous solutions.

A point worth of notice is the blob like nature of the detected objects. As we have pointed out in section 3.3, this happens because of the isotropic nature

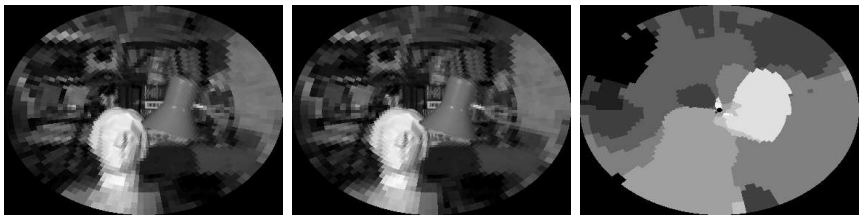

Fig. 4. The disparity map on the right was computed from the well known stereo test images from Tsukuba University. In the left we show the foveated images of the stereo pair. Notice that much of the detail in the periphery is lost due to the space variant sampling. Thus, this result can not be directly compared with others obtained from uniform resolution images.

and large support of the spatial facilitation filters. Also, the space variant image sampling, blurs image detail in the periphery of the visual field. This results in the loss of small and thin structures like the fingertips in the stereo head example and the lamp support in the Tsukuba images. However note that spatial facilitation do not blur depth discontinuities because filtering is not performed on the disparity map output, but on the likelihood maps before the maximum operation.

The lack of detail shown in the computed maps is not a major drawback for our applications, that include people tracking, obstacle avoidance and region of interest selection for further processing. As a matter of fact, it has been shown in a number of works that many robotics tasks can be performed with coarse sensory inputs if combined with fast control loops [24].

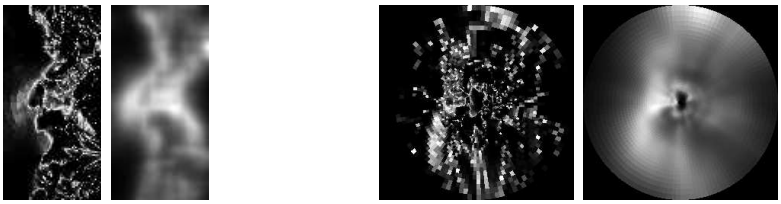

Fig. 5. Intermediate results for the experiment in Fig.3. This figure shows the cortical maps tuned to retinal disparity $d_i = 26$, for which there is a good match in the hand region. In the left group we show the likelihood images L_i^{cort} (left) and D_i^{cort} (right) corresponding to the cortical activation before and after the spatial facilitation step. In the right group, the same maps are represented in retinal coordinates, for better interpretation of results.

The parameters used in the tests are the following: log-polar mapping with 128 angular sections and 64 radial rings; retinal disparity range from -40 to 40 pixels (horizontal) and from -6 to 6 pixels (vertical), both in steps of 2; $q = 0.1$

(prior probability of occlusion); $M = 256$ (number of gray values); $\sigma = 3$ (white noise standard deviation); facilitation filtering with zero-phase forward/reverse filter $y(n) = 0.8y(n-1) + 0.2x(n)$.

The algorithms were implemented in C++ and take about three seconds to run in a PII 350MHz computer.

5 Conclusions

We have presented a real-time dense disparity estimation algorithm for foveated systems using the *logmap*. The algorithm uses an intensity based matching technique, which makes it easily extensible to other space variant sampling schemes. Some results were taken from an active stereo head and others obtained from standard test images. Many robots are currently equipped with foveated active vision systems and the availability of fast stereopsis will drastically improve their perceptual capabilities. Obstacle detection and tracking, region of interest selection and object manipulation are some possible applications.

Acknowledgements

This work was partially supported by EU project MIRROR: *Mirror Neurons based Robot Recognition, IST-2000-28159.*

References

1. S. Amari and M. Arbib. *Competition and Cooperation in Neural Nets*, pages 119–165. Systems Neuroscience. J. Metzler (ed), Academic Press, 1977.
2. A. Basu and K. Wiebe. Enhancing videoconferencing using spatially varying sensing. *IEEE Trans. on Systems, Man, and Cybernetics*, 38(2):137–148, Mar. 1998.
3. A. Bernardino and J. Santos-Victor. Binocular visual tracking : Integration of perception and control. *IEEE Trans. on Robotics and Automation*, 15(6):137–146, Dec. 1999.
4. A. Bernardino, J. Santos-Victor, and G. Sandini. Foveated active tracking with redundant 2d motion parameters. *Robotics and Autonomous Systems*, 39(3-4):205–221, June 2002.
5. M. Bolduc and M. Levine. A review of biologically motivated space-variant data reduction models for robotic vision. *CVIU*, 69(2):170–184, Feb. 1998.
6. T. Boyling and J. Siebert. A fast foveated stereo matcher. In *Proc. Conf. on Imaging Science Systems and Technology*, pages 417 – 423, Las Vegas, USA, 2000.
7. C. Capurro, F. Panerai, and G. Sandini. Dynamic vergence using log-polar images. *IJCV*, 24(1):79–94, Aug. 1997.
8. T. Wiesel D. Hubel. Stereoscopic vision in macaque monkey. cells sensitive to binocular depth in area 18 of the macaque mokey cortex. *Nature*, 225:41–42, 1970.
9. G. DeAngelis and W. Newsome. Organization of disparity–selective neurons in macaque area mt. *The Journal of Neuroscience*, 19(4):1398–1415, 1999.
10. S. Mallat E. Chang and C. Yap. Wavelet foveation. *J. Applied and Computational Harmonic Analysis*, 9(3):312–335, Oct. 2000.

11. B. Fischl, M. Cohen, and E. Schwartz. Rapid anisotropic diffusion using space-variant vision. *IJCV*, 28(3):199–212, July/Aug. 1998.
12. G. Gambardella G. Sandini, C. Braccini and V. Tagliasco. A model of the early stages of the human visual system: Functional and topological transformation performed in the peripheral visual field. *Biological Cybernetics*, 44:47–58, 1982.
13. W. Geisler and J. Perry. A real-time foveated multi-resolution system for low-bandwidth video communication. In *Human Vision and Electronic Imaging, SPIE Proceedings 3299*, pages 294–305, Aug. 1998.
14. N. Griswald, J. Lee, and C. Weiman. Binocular fusion revisited utilizing a log-polar tessellation. *CVIP*, pages 421–457, 1992.
15. E. Grosso and M. Tistarelli. Log-polar stereo for anthropomorphic robots. In *Proc. 6th ECCV*, pages 299 – 313, Dublin, Ireland, June-July 2000.
16. B. Horn. *Robot Vision*. MIT Press, McGraw Hill, 1986.
17. W. Klarquist and A. Bovik. Fovea: A foveated vergent active stereo system for dynamic three-dimensional scene recovery. *IEEE Trans. on Robotics and Automation*, 14(5):755 – 770, Oct. 1998.
18. D. Marr and T. Poggio. Cooperative computation of stereo disparity. *Science*, 194:283–287, 1976.
19. M. Peters and A. Sowmya. A real-time variable sampling technique: Diem. In *Proc. ICPR*, pages 316–321, Brisbane, Australia, Aug. 1998.
20. S. Pollard, J. Mayhew, and J. Frisby. Pmf: A stereo correspondence algorithm using a disparity gradient limit. *Perception*, 14:449–470, 1985.
21. K. Prazdny. Detection of binocular disparities. *Biol. Cybern*, 52:93–99, 1985.
22. G. Metta R. Manzotti, A. Gasteratos and G. Sandini. Disparity estimation on log-polar images and vergence control. *CVIU*, 83:97–117, 2001.
23. G. Salgian and D. Ballard. Visual routines for vehicle control. In D. Kriegman, G. Hager, and S. Morse, editors, *The Confluence of Vision and Control*. Springer Verlag, 1998.
24. J. Santos-Victor and A. Bernardino. Vision-based navigation, environmental representations, and imaging geometries. In *Proc. 10th Int. Symp. of Robotics Research*, Victoria, Australia, Nov. 2001.
25. D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47(1):7–42, April-June 2002.
26. E. Schwartz. Spatial mapping in the primate sensory projection : Analytic structure and relevance to perception. *Biological Cybernetics*, 25:181–194, 1977.
27. J. Siebert and D. Wilson. Foveated vergence and stereo. In *Proc. of the 3rd Int. Conf. on Visual Search (TICVS)*, Nottingham, UK, Aug. 1992.
28. M. Tistarelli and G. Sandini. On the advantages of polar and log-polar mapping for direct estimation of the time-to-impact from optical flow. *IEEE Trans. on PAMI*, 15(8):401–411, April 1993.
29. H. Tunley and D. Young. First order optic flow from log-polar sampled images. In *Proc. ECCV*, pages A:132–137, 1994.
30. R. Wallace, P. Ong, B. Bederson, and E. Schwartz. Space variant image processing. *IJCV*, 13(1):71–90, Sep. 1995.
31. C. Weiman and G. Chaikin. Logarithmic spiral grids for image processing and display. *Comp Graphics and Image Proc*, 11:197–226, 1979.
32. R. Zabih Y. Boykov, O. Veksler. Disparity component matching for visual correspondence. In *Proc. CVPR*, pages 470–475, 1997.
33. R. Zabih Y. Boykov, O. Veksler. Markov random fields with efficient approximations. In *Proc. CVPR*, pages 648–655, 1998.