An attentional system for a humanoid robot exploiting space variant vision

Giorgio Metta

MIT – AI-Lab, Humanoid Robotics Group
200, Technology Square – Cambridge MA – USA
e-mail: pasa@ai.mit.edu

Abstract

This paper describes the implementation of the attentional system of a humanoid robot based completely on space variant vision (in particular log-polar). The aim is that of providing the robot with a suitable measure of position, speed and saliency of possibly interesting objects for saccading and tracking. The major advantage of log-polar based imaging is related to the reduced number of pixels while maintaining a large field of view. This arrangement is very well suited for motor control, where the high-resolution center (fovea) allows precise positioning and, at the same time, the coarse resolution periphery permits detection of potential targets. Algorithms for color processing, optic flow, and disparity computation were developed within this architecture. The attentional modules are intended as the first layer of a more complicated system, which shall include learning of object recognition, trajectory tracking, and naïve physics understanding during the natural interaction of the robot with the environment.

Keywords: biologically inspired robots, log-polar vision, attention

Introduction

Besides the studies on artificial neural networks, substantial effort is devoted worldwide to build physical models of parts of biological systems with the aim of suggesting new solutions to robotics but more importantly with the ultimate goal of gaining a better understanding of how the brains of living systems solve the same sort of problems. Examples of this approach can be found in [1-6]. Although the "robotic" models are thus far only crude approximations of real living organisms, the motivations of the approach are rooted in the belief that constructing a real system might reveal problems and subtleties that a mere analysis could not.

Along this line of research we developed an attentional system for a humanoid robot. The unique aspect of this work is in the use of space variant vision. In particular we employed *log-polar* images, which, as described later on, model how photoreceptors are distributed in our retinas. Although, on a first inspection, it might seem that space variance poses more challenges than traditional rectangular imaging, we will show that very simple strategies might be employed to adapt the algorithms to the log-polar geometry. On the other hand, we gain (from the space variant sampling) the possibility to maintain a large field of view and at the same time process a limited number of pixels. This is advantageous since it allows the system to be simultaneously maximally responsive to new events and maximally precise in its movements (highest resolution in the image center). The robot exploits color, motion, and binocular cues to derive information about potentially interesting targets for pursuit and saccading. Other interesting aspects, borrowed from biology, are related to the use of inertial information to stabilize the visual world in spite of the movement of the robot or external disturbances.

The experimental setup is a seven degree-of-freedom robot head, with human-like performance in terms of speed and acceleration (see Figure 1). For the scope of this paper, the sensory system consists of a pair of cameras (standard CCD; sub-sampling to log-polar is carried out in software), an inertial sensor (InterSense IS300) which measures the roll, pitch, and yaw angles, and high resolution motor encoders providing the position of each joint. Visual processing and control are carried out by a set of PCs connected through a fast network and running QNX – a real-time OS. Video signals are synchronized, split, and sent in parallel to many nodes for parallel processing. Nine Pentium class processors are employed in the present implementation.

Log-polar vision

Among the many possible space variant sub-sampling procedures, the one we used is known as *log-polar*. The log-polar mapping resembles the distribution of the photoreceptors in the primate retina as well as the geometrical transform following the projection of these neurons into the primary visual cortex [7-10].

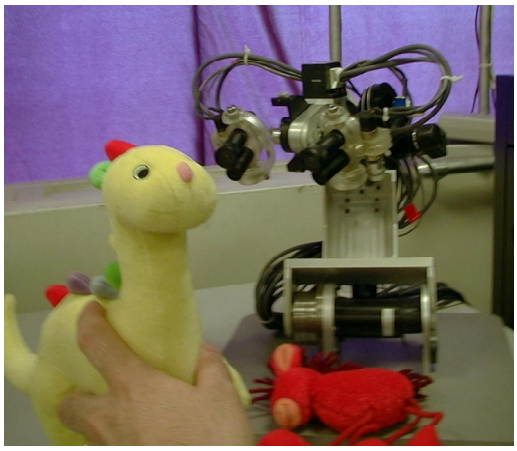

Figure 1: The robot setup "Lazlo". The robot head mounts four cameras (only two are used at the moment). It has vergence and three independent tilt joints, a pan at the level of the neck and a roll. On of the tilts and the roll movements are obtained by means of a differential joint.

The initial analytical formulation based on studies on the primates' visual pathways is due mainly to Schwartz [11]; his model can be roughly summarized as follows:

- The distribution of the photoreceptors in the retina is not uniform. They lay more densely in a central region called the fovea, while they are sparser in the periphery. Consequently, the resolution also decreases moving away from the fovea toward the periphery. The retina has a radial symmetry, which can be approximated by a polar distribution.
- The projection of the photoreceptor array into the primary visual cortex can be described by a log-polar distribution mapped onto an almost rectangular surface (the cortex).

From the mathematical point of view, the log-polar mapping can be expressed as a transformation between a polar plane (ρ,θ) (retinal plane) and a Cartesian plane (ξ,η) (cortical plane), as follows:

$$\begin{cases} \eta = q\vartheta \\ \xi = k_\xi \ln_a \rho/\rho_0 \end{cases} \qquad (0.1)$$

where ρ_0 is the radius of the innermost circle, *1/q* is the minimum angular resolution of the log-polar layout, and (ρ,θ) are the polar coordinates. k_ξ is a linear scaling parameter: this has been added to the original formulation in order to fit the mapping into a fixed size squared image (which is determined by the frame grabber characteristics). (ρ,θ) are related to the conventional Cartesian reference system by:

$$\begin{cases} x = \rho\cos\vartheta \\ y = \rho\sin\vartheta \end{cases} \qquad (0.2)$$

A pictorial example is shown in Figure 2, where the leftmost panel (a) shows a Cartesian or retinal image (before sub-sampling) and the corresponding log-polar (or cortical) image on the right (b).

(a) (b)

Figure 2: An example of log-polar mapping: note as radial structures in the flower (petals) map to horizontal structures in the log-polar image. Circles, on the other hand, map to vertical patterns.

Optic flow

Optic flow as described by Horn is "*the apparent motion of pixels in the image plane*" [12]. Horn proposed also a continuity constraint for the optic flow involving the spatio-temporal derivatives of the image intensity:

$$\frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v + \frac{\partial I}{\partial t} = 0 \qquad (0.3)$$

where I is the image intensity and u,v the flow components. The two components cannot be recovered from equation (0.3) alone. If we assume that the flow is well represented by an affine model such as:

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} u_0 \\ v_0 \end{bmatrix} + \begin{bmatrix} D+S_1 & S_2-R \\ R+S_2 & D-S_1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} \qquad (0.4)$$

then a combination of equation (0.3) and (0.4) allows computing the parameters of the affine model provided we estimate it in at least six points – usually a least square approach is taken and more points are used. The parameters have the meaning of translation, divergence, curl, and shear. The approach is similar to that proposed by Koenderink et al. [13]. To take into account the log-polar geometry we have to transform further equation (0.3) into:

$$-\frac{\partial I}{\partial t} = \begin{bmatrix} \gamma_1 & \gamma_2 & g_\xi & g_\eta & \gamma_3 & \gamma_4 \end{bmatrix} \cdot \begin{bmatrix} u_0 & v_0 & D & R & S_1 & S_2 \end{bmatrix} \qquad (0.5)$$

where the constants γ and g represent the matrix product of the image derivative with the log-polar Jacobian; for a complete derivation see [6].

Further, by processing the optic flow we can determine which parts of the image are moving, and consequently segment the target from the background (in those cases when they are moving differently). Roughly speaking, this is accomplished by computing the expected optic flow due to the movement of the camera and subtracting it from the actual optic flow. Where the two differ enough (by a suitable measure) the pixels could be identified and labeled as an independent moving object.
The expected flow is determined by using a constant approximation of the image Jacobian:

$$\begin{bmatrix} u \\ v \end{bmatrix} = J(\mathbf{q})\dot{\mathbf{q}} \simeq J\dot{\mathbf{q}} \qquad (0.6)$$

The matrix J is estimated by incremental least squares and by collecting example pairs of the joint speed $\dot{\mathbf{q}}$ versus the measured optic flow. An appropriate delay line takes care of synchronizing the two signals.
The actual segmentation algorithm in this case develops on the Horn equation. It suffices to note that equation (0.3) is satisfied when the flow vector are in "agreement" with the spatio-temporal gradient of I. On the other hand, where the equation is not satisfied it means that the expected flow is not correct for that pixel. Consequently, by identifying the regions where the expected flow causes:

$$\left\| \frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v + \frac{\partial I}{\partial t} \right\| > \varepsilon \qquad (0.7)$$

we can segment the ego-motion component from the moving object(s). ε is a suitable threshold.
It is fair to say that various reasons not necessarily related to the presence of an object might cause the flow not to respect the Horn constraint. These include the presence of strong edges (where the spatial gradient is high) or fast movements of the head, for which the linear prediction model is prone to failure.

Color processing

Color processing comes in many flavors within this implementation: i) a general-purpose color segmentation algorithm, ii) a color "blob" detector, and iii) a skin tone detector.
The general-purpose segmentation is based on histograms. It is started by a motion sensitive cueing procedure. It subsequently builds a pair of histograms: one to represent the target (the moving object), and a second that contains the information about the background. The latter in continuously adapted and thus provides a sort of habituation to the color of the background. Histograms are constructed in the HSV color space; they have the form:

$$histo(H,S) = h(H,S) \Big/ \sum_{H,S} h(H,S) \qquad (0.8)$$

i.e. they are independent of the image intensity (V) and normalized to one. A pixel is assumed to belong to the object if its probability (an estimate of) computed as:

$$p(object \mid h,s) = histo(h,s) \qquad (0.9)$$

is greater than a threshold and its histogram does not overlap with that of the background.
The *blob* detector is based on a very standard region growing procedure. Areas of uniform color, as measured by taking into account *hue* and *saturation* only, are labeled. A further grouping and coherency test of the resulting regions is performed to eliminate spurious results (very likely due to noise). In spite of its simplicity the algorithm provides a very stable behavior.
Finally a skin tone detector has been implemented. It is based on the algorithm developed in [14]. It has been found to improve the robot's ability to interact with humans, although it is not sufficient, for example, to unambiguously detect faces.

Disparity computation

Binocular disparity is the strongest cue related to depth. In the context of sensori-motor coordination it can be used to control vergence. A suitable procedure to estimate the disparity of a target (or of a particular region of the image) is that of using cross-correlation (or another suitable distance measure) to find the difference in position between the left and right image of that particular region (representative of the target).
This procedure can be implemented by an exhaustive search. In formula:

$$d_{est} = \arg\max_{d} f_U\left(I_L(x,y), I_R(x+d,y)\right) \qquad (0.10)$$

where the function f is the pixel similarity measure, U the support of f, and I_L, I_R the left and right image respectively.
Equation (0.10) needs to be converted to the log-polar domain in order to take account of how pixels shift under a Cartesian translation d:

$$I_L(\xi,\eta) = I_R\left(lp_d(\xi,\eta)\right) \qquad (0.11)$$

It is easy to verify that the transformation lp_d itself does not depend on the actual images and thus can be computed beforehand [15].
In our case we chose the normalized cross-correlation as f in equation (0.10). As a consequence of the log-polar mapping, an explicit segmentation is not necessary and in fact U was chosen to be the whole image. The disparity is that of the target as long as it remains close to the foveae (since most of the pixels

belong to the target), otherwise the value of disparity would switch to that of the background.
It is worth mentioning that the current implementation assumes that the transform from the left to the right image is a pure translation along the horizontal direction. This in reality is unlikely to be the case. Disparity in fact is strictly horizontal only when the optical axes are parallel (and vertically aligned). A further limitation might arise because of the distortion of the lenses. In this case too the "pure translation" assumption would fail. This was not the case in our configuration.

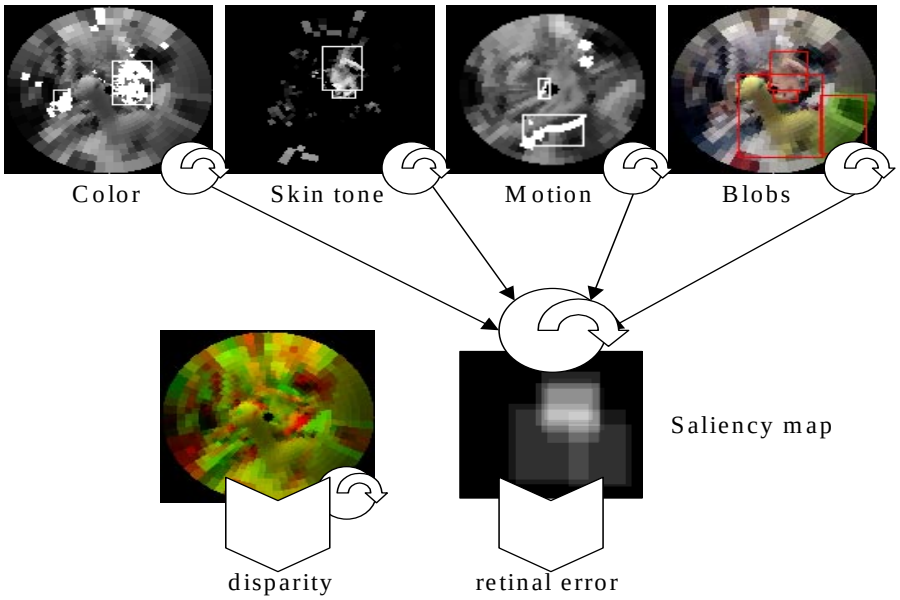

Figure 3: An example of visual processing and saliency map. Basic processing modules are combined to produce the retinal error and the disparity signals. Each separate processing module provides a list of probable targets and relative bounding boxes. A voting mechanism is used to build the saliency map. The position of the maximum of the saliency map is the retinal error.

Integration

In order to provide the controller with a reasonable reference values, the results provided by the various algorithms have to be integrated in a single percept.
The two quantities relevant to the control of gaze direction are the position of the target in the left-right, up-down directions and the depth with respect to the fixation point. These two quantities are related to two different control modes: version (same control values to both eyes) and vergence (opposite commands to the eyes). The first quantity, apt to control version, is estimated by a voting mechanism. Each algorithm provides a list of potential targets and their bounding boxes in retinal coordinates. The regions identified by the bounding boxes get their saliency increased. The increment of saliency is weighed to give more importance to particular aspects of the environment (e.g. skin tone versus motion). The position of the maximum of the saliency function determines what is tracked.

Although it is not a concern here, it is worth noting that the weights and shape of the attentional regions can be modified on-the-fly to give more or less importance to different aspects of the observed scene, and this can be carried out in relation to the task or internal status of the robot [16].

Control

Control is mostly constructed around the two quantities described in the previous section. The controller can be further divided into two sub-modules dealing respectively with gaze stabilization and saccadic behaviors (gaze shifting). This roughly reflects two distinct functional modes of the controller itself. Gaze stabilization is obtained by means of closed loop controllers (i.e. PID), while saccades are open loop.

The control of the eyes

In stabilizing the gaze, the eyes are controlled in order to zero the retinal error:

$$\begin{bmatrix} \dot{q}_1 \\ \dot{q}_2 \\ \dot{q}_3 \end{bmatrix} = PID\left(e_x, e_y\right) \qquad (0.12)$$

where $\dot{q}_1, \dot{q}_2, \dot{q}_3$ are the control variables: i.e. the speed of the eyes. e_x, e_y is the retinal error. This particular module does not change the vergence angle (which is adjusted by another control loop instead) and consequently $\dot{q}_2 = \dot{q}_3$.

A word of caution is necessary: the controller assumes that the dynamics of the system is negligible. This is only approximately true. While stability is not compromised (the control loop can be shown to be stable by applying, for example, the visual servoing theory [17]), the performance could be nevertheless affected [18]. In our case, the inertias involved are very small compared to the low-level PID gain and thus the system's dynamics are truly negligible.

Inertial stabilization

Gaze stabilization can be also obtained through other means. The general idea, borrowed from the biological vestibular stabilization mechanisms (the *vestibulo-ocular reflex*; see for example [19]), is that of using inertial sensing. In our case, three gyroscopes are employed arranged along three orthogonal directions. A simple controller can be formulated as follows:

$$\begin{bmatrix} \dot{q}_1 \\ \dot{q}_2 \\ \dot{q}_3 \end{bmatrix} = \begin{bmatrix} -k_1 & 0 \\ -k_1 & 0 \\ 0 & -k_2 \end{bmatrix} \cdot \begin{bmatrix} \omega_{yaw} \\ \omega_{pitch} \end{bmatrix} \qquad (0.13)$$

with k_1 and k_2 two suitable gains, and ω_{yaw}, ω_{pitch} the angular velocity measured by the gyros along the yaw and pitch direction. The intuitive description of the controller is that it counter-rotates the eyes in order to compensate for the movement of the head or body of the robot. A further loop exploits the ability to control the roll angle to maintain the eyes approximately aligned with respect to gravity.

A more sophisticated control schema (optimal in the sense of image stabilization) has been investigated in [20] together with an on-line learning strategy, although not on this robot.

Head control

The goal of the head movements is simply that of repositioning the head after the eyes have lost their "central" position (symmetric vergence). Essentially, the controller drives the head joints in order to zero the deviation from the symmetric configuration, or, in the case of the tilt, from a resting configuration with the joints aligned. For example, for the pan at the level of the neck the controller is:

$$\dot{q}_6 = PID\left(q_2 - q_3\right) \qquad (0.14)$$

This strategy alone would very likely oscillate (or otherwise the movements must be kept very slow) because the head movements would disturb the movement of the eyes. A possible solution is that of compensating the movement of the head by a counter rotation of the eyes. This is exactly the inertial stabilization mechanism already described. There is evidence that a similar mechanism is employed by humans to coordinate the head and eyes [19]. This strategy is also efficient in the sense that it maximizes the range of movement of the eyes by maintaining positions far from the physical limits.

The control of vergence

Vergence control is provided by a completely separated loop. The disparity measurement process is separated; this reduces the chances of a conflict between the pursuit and the "verge" behavior. The controller in this case tries to keep the disparity *d* close to zero. Vergence, together with the control of tracking, assures that the object of interest is kept almost in the foveae (left and right eyes).

Saccades

Saccades neatly complement the pursuit controllers and gaze stabilization when the object moves too fast to be appropriately followed or when a rapid shift of attention is required (because a more salient target appeared). When performing a saccade, the gaze stabilization behaviors get temporarily inhibited. The precise computation of saccades would require the knowledge of the mapping between the retinal error and the appropriate motor command (a learning strategy has been investigated in [21]).

In our case we resorted to a simpler implementation by using a linear map, which has been tuned by hand (it is substantially only a gain matrix). A final note concerns the actual activation of saccades: the logic behind their generation checks whether the target is outside the fovea (defined by a threshold) and if a refractory period has elapsed. The latter is required to stabilize the system. In fact, saccades, acting as a very high-gain controller, might lead to unstable behaviors.

Figure 4, below, is intended to give the general flavor of how the different control loops are combined and organized (see caption for details). Figure 5 shows an example trajectory: it is possible to note the activation of the two control modes (open- and closed-loop).

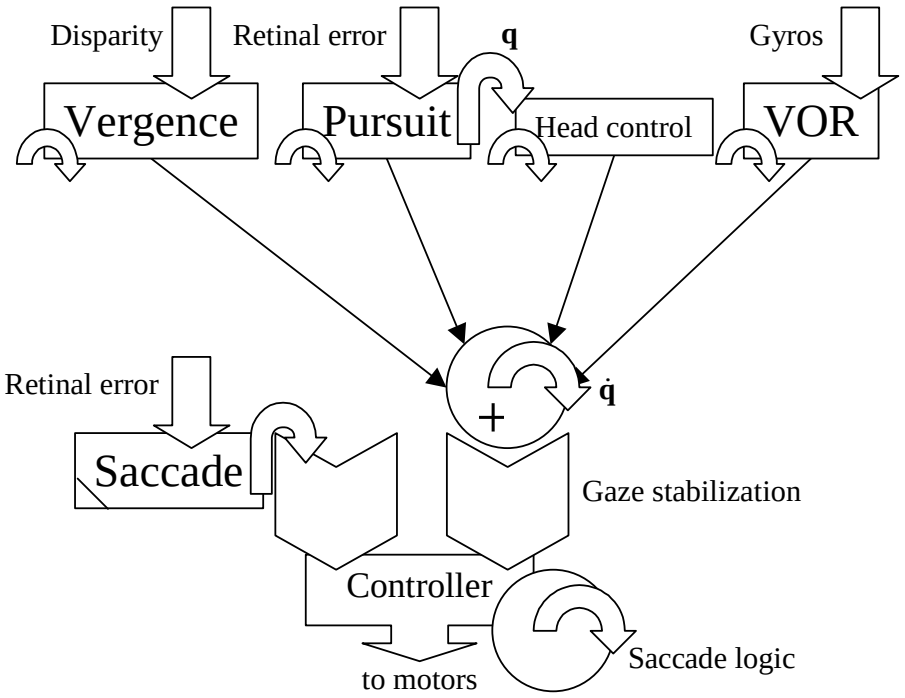

Figure 4: A schematic of the head controller. Different signals (top) are used to build different independent control loops. Each module generates a velocity command. For what concerns gaze stabilization, the velocity commands are combined by adding them together. Saccades are independently calculated and activated when needed by the saccade control logic. Finally velocity commands are sent to a low-level PID controller, which generates the appropriate signals to drive the motors.

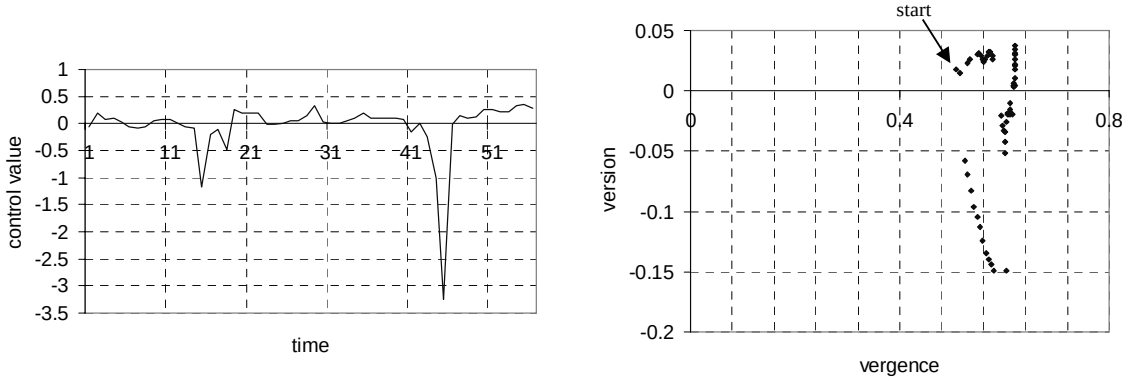

Figure 5: Robot behavior. The plot on the left shows the horizontal component of the eye movement command. Note the two negative peaks due to the generation of two saccades. The plot on the right shows instead the position of the fixation point in 2D (top view) corresponding to the same movement on the left.

Conclusions

This paper addressed the problem of designing and realizing a biologically inspired attentional system for a humanoid robot. We showed, through the use of space variant vision, that it is possible to maximally exploit the available computational power without compromising the ability to perform accurate movements. The benefits are still moderate at the present resolution: images were only 64x32 pixels. The ratio between the log-polar and the corresponding Cartesian would grow even further with the increase of the resolution. For instance a 33000 pixel log-polar image would correspond (assuming that the foveal resolution is the same) to a

million-pixel rectangular image. In terms of timing, the present implementation is eight times faster that the corresponding Cartesian version – part of the processing runs on 400MHz processors at frame-rate, the more demanding disparity computation runs on a 800Mhz Pentium. A hypothetical 33000 implementation would be 30 times faster than its Cartesian counterpart and it would be still manageable while the million pixels processing could be awkwardly complicated (in terms of storage, bus bandwidth, raw processing power, etc). A similar consideration applies to biological system: if the retina were uniform, the optic nerve would need to be 6 cm in diameter to deliver the information to the cortex; one might wonder what would be the size of the brain in that case.

We showed also how optic flow, color and depth cues could be estimated from log-polar images. Not less important, we showed how a simplified coordination schema of head and eye movements could be devised under the hypothesis that compensatory eye movement can be generated.

It is important to note that the architecture is completely bottom-up. We are aware that this is a biologically implausible simplification; in our view this has to be considered only the very first visuo-motor coordination layer. Furthermore, the integration mechanism (as described in the relative section) is not tuned on the basis of the current state of the robot or the task at hand. However, this issue has been already investigated in other contexts, for example by Scassellati [22], and it is likely to be inserted in this model as investigation proceeds.

Future work will include object recognition abilities. In this context, the multi-cue approach is extremely effective in driving the exploration of the environment, and thus in facilitating the acquisition of training samples for autonomous learning.

Acknowledgments

This work was funded by DARPA as part of the "Natural Tasking of Robots Based on Human Interaction Cues" project under contract number DABT 63-00-C-10102, and by the Nippon Telegraph and Telephone Corporation as part of the NTT/MIT Collaboration Agreement. I wish to thank also Dr. Brian Scassellati and Aaron Edsinger for their useful comments during the preparation of the manuscript.

References

1. Brooks, R. *Behavior-Based Humanoid Robotics*. in *IEEE/RSJ IROS'96*. 1996.
2. Sandini, G. *Artificial Systems and Neuroscience*. in *Proc. of the Otto and Martha Fischbeck Seminar on Active Vision*. 1997.
3. Pfeifer, R. and C. Scheier. *Representation in Natural and Artificial Agents: an Embodied Cognitive Science Perspective*. in *Natural Organisms, Artificial Organisms, and Their Brains*. 1998. Bielefeld, Germany: Verlag.
4. Kuniyoshi, Y. and G. Cheng, *Complex Continuous Meaningful Humanoid Interaction: A Multi Sensory-Cue Based Approach*. 1999, Personal Communication.
5. Voetglin, T. and P.F.M.J. Verschure, *What Can Robots Tell Us About Brains? A Synthetic Approach Towards the Study of Learning and Problem Solving.* Reviews in the Neurosciences, 1999. **10**: p. 291-310.
6. Metta, G., *Babyrobot: A Study on Sensori-motor Development*, in *Dipartimento di Informatica, Sistemistica e Telematica*. 1999, University of Genoa: Genova, Italy.
7. Daniel, M. and D. Whitteridge, *The Representation of the Visual Field on the Cerebral Cortex in Monkeys.* Journal of Physiology, 1961(159): p. 203-221.
8. Cowey, A., *Projection of the Retina on to Striate and Prestriate Cortex in the Squirrel Monkey (Saimiri Sciureus).* Journal of Neurophysiology, 1964(27): p. 266-293.
9. Allman, J.M. and J.H. Kaas, *Representation of the Visual Field in Striate and Adjoining Cortex of the Owl Monkey (Aotus Trivirgatus).* Brain Research, 1971. **35**: p. 89-106.
10. Hubel, D.H. and T.N. Wiesel, *Functional architecture of macaque monkey cortex.* Proceedings of the Royal Society of London, 1977(198): p. 1-59.
11. Schwartz, E.L., *Spatial Mapping in the Primate Sensory Projection: Analytic Structure and Relevance to Perception.* Biological Cybernetics, 1977. **25**: p. 181-194.
12. Horn, B.K.P., *Robot Vision*. 1986: MIT Press.
13. Koenderink, J. and J. Van Doorn, *Affine Structure from Motion.* Journal of the Optical Society of America, 1991. **8**(2): p. 377-385.
14. Scassellati, B. *Eye Finding via Face Detection for a Foveated, Active Vision System*. in *AAAI 1998*. 1998. Madison WI.
15. Manzotti, R., et al., *Disparity in log polar images and vergence control.* Computer

Vision and Image Understanding, 2001(To appear in 2001).
16. Scassellati, B., *Foundations for a Theory of Mind for a Humanoid Robot*, in *Department of Computer Science and Electrical Engineering*. 2001, MIT: Cambridge MA.
17. Espiau, B., F. Chaumette, and P. Rives, *A New Approach to Visual Servoing in Robotics.* IEEE Transactions on Robotics and Automation, 1992. **8**(3): p. 313-326.
18. Bernardino, A. and J. Santos-Victor, *Binocular Visual Tracking: Integration of Perception and Control.* IEEE Transactions on Robotics and Automation, 1999. **15**(6): p. 1080-1094.
19. Miles, F.A., *Visual stabilization of the eyes in primates.* Current Opinion Neurobiology, 1997(7): p. 867-871.
20. Panerai, F., G. Metta, and G. Sandini. *Learning VOR-like stabilization reflexes in robots*. in *8th European Symposium on Artificial Neural Networks*. 2000. Bruges, Belgium.
21. Metta, G., et al. *An Incremental Growing Neural Network and its Application to Robot Control*. in *International Joint Conference on Neural Networks*. 2000. Como, Italy.
22. Scassellati, B. *Theory of Mind for a Humanoid Robot*. in *First IEEE/RAS International Conference on Humanoid Robotics*. 2000. Cambridge MA.