Overt Visual Attention for a Humanoid Robot

Sethu Vijayakumar, Jörg Conradt, Tomohiro Shibata, and Stefan Schaal
Computer Science & Neuroscience
University of Southern California, Los Angeles, CA 90089-2520, USA
Institute of Neuroinformatics
University/ETH Zurich, Winterthurerstr. 190, CH-8057 Zurich
Kawato Dynamic Brain Project, ERATO
Japan Science & Technology Corp., Kyoto 619-0288, Japan

Abstract

The goal of our research is to investigate the interplay between oculomotor control, visual processing, and limb control in humans and primates by exploring the computational issues of these processes with a biologically inspired artificial oculomotor system on an anthropomorphic robot. In this paper, we investigate the computational mechanisms for visual attention in such a system. Stimuli in the environment excite a dynamical neural network that implements a saliency map, i.e., a winner-take-all competition between stimuli while simultenously smoothing out noise and suppressing irrelevant inputs. In real-time, this system computes new targets for the shift of gaze, executed by the head-eye system of the robot. The redundant degrees-of- freedom of the head-eye system are resolved through a learned inverse kinematics with optimization criterion. We also address important issues how to ensure that the coordinate system of the saliency map remains correct after movement of the robot. The presented attention system is built on principled modules and generally applicable for any sensory modality.

1 Introduction

Visual attention involves directing a "spotlight" of attention [12] to interesting areas, extracted from a multitude of sensory inputs. Most commonly, attention will require to move the body, head, eyes, or a combination of these in order to acquire the target of interest with high-resolution foveal vision, referred to as 'overt' attention, as opposed to covert attention which does not involve movement. In order to provide high-resolution vision simultaneously with large-field peripheral vision, our humanoid robot employs two cameras per eye, a foveal camera and wide-angle camera – this strategy mimics the log-polar retinal resolution of numerous biological species. Similar to biology, overt visual attention is a prerequisite in such a system in order to move the cameras such that a target can be inspected in the foveal field of view.

There has been extensive work in modeling attention and understanding the neurobiological mechanisms of generating the visual "spotlight" of attention [15], both from a top-down[16] and a bottom-up perspective [9, 10] - albeit mainly for static images. From the perspective of overt shift of focii, there has been some work on saccadic eye motion generation using spatial filters [17], saccadic motor planning by integrating visual information [13], social robotics [4], and humanoid robotics [6]. In contrast to this previous work, our research focus lies on creating a biologically inspired approach to visual attention and oculomotor control by employing theoretically sound computational elements that were derived from models of cortical neural networks, and that can serve for comparisons with biological behavior. We also emphasize real-time performance and the integration of the attention system on a full-body humanoid robot that is not stationary in world coordinates. As will be shown below, these features require additional computational consideration such as the remapping of a saliency map for attention after body movement. In the following sections, we will first give an overview of the attentional system's modules, then explain the computational principles of each module, before we provide some experimental evaluations on our humanoid robot.

2 An Overt Visual Attention Control System

The computations involved in an overt visual attentional mechanism can be modularized into broadly

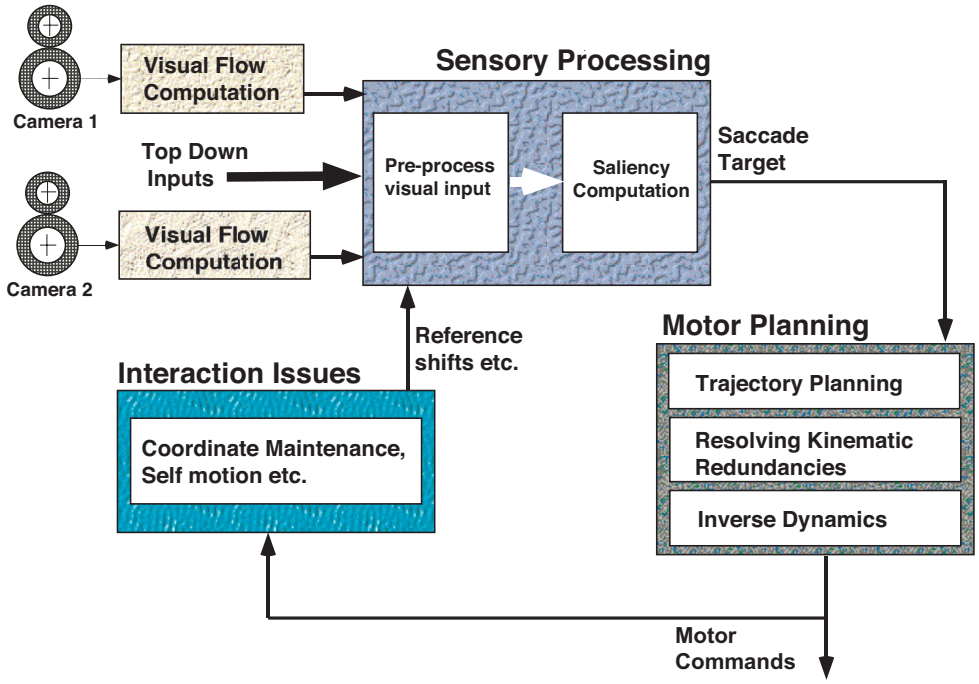


Figure 1: A schematic block diagram of the various modules involved in the system for implementing overt visual attention

three distinct subparts: the sensory processing module, the motor planning module and a module in charge of interaction issues. Fig. 1 represents a schematic block diagram based on these distinctions.

The *sensory processing module* receives as input raw bottom-up sensory signals from all availabe modalities, e.g., vision, audition, and haptics, and also top-town volitional inputs. After appropriate computations, this module outputs the new desired focus of attention in camera coordinates as a target for the next saccade.

The *motor planning module* takes the saccade target in camera coordinates and converts it into a sequence of motor commands necessary to drive the oculomotor system and the head to gaze at this location.

Finally, the *interaction issues module* is needed to take care of higher level issues of overt attention. For instance, after a saccade, it is necessary to re-map attentational saliency maps according to the amount of eye movement, self-motion needs to be canceled as a potential attentional target, or perturbations of the body need to be factored into the re-mapping of saliency maps.

In the following sections, we will provide the details in each module. As sensory input, only visual flow is currently employed since it can be computed reliably in real-time from dedicated hardware. Other sensory modalities could be handled in the same way as described for visual flow.

2.1 Sensor Pre-processing and Integration

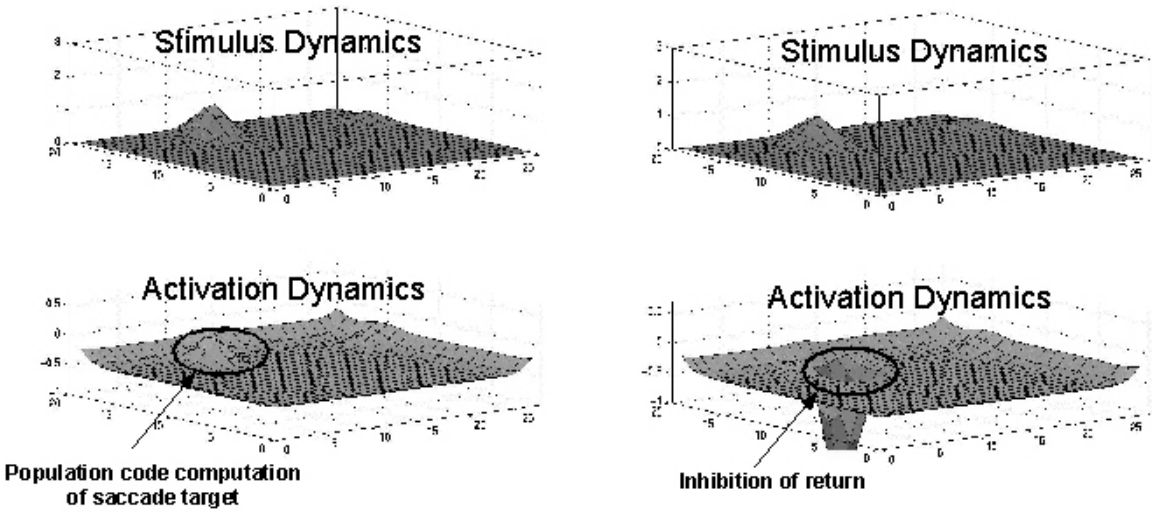
The key element of our Sensory Pre-Processing block (Fig. 1) is a competitive dynamical neural network, derived in Amari and Arbib's [1] *neural fields* approach for modeling cortical information processing. The goal of this network is to take as input spatially localized stimuli, have them compete to become the next saccade target, and finally output the winning target. For this purpose, the sensory input pre-processing stage takes the raw visual flow $V_F(\boldsymbol{x}, t)$ as inputs to the *stimulus dynamics*, a first order dynamical system. Using $\boldsymbol{x}$ to denote the position of a stimulus in camera coordinates, the stimulus dynamics is:

$$\dot{S}(\boldsymbol{x}) = -\alpha S(\boldsymbol{x}) + VisInp(\boldsymbol{x}, t) \qquad (1)$$

where

$$VisInp(\boldsymbol{x}, t) = \int_R G(\boldsymbol{x}, t) * exp(-\boldsymbol{x}^2/2\sigma^2) d\boldsymbol{x} \qquad (2)$$

$$G(\boldsymbol{x}, t) = V_F(\boldsymbol{x}, t) + \gamma * \lfloor \dot{V}_F(\boldsymbol{x}, t) \rfloor_+ \qquad (3)$$



(a) (b)

Figure 2: A snap shot of the stimulus and activation dynamics just (a) before and (b) after the saccade

Eq.(3) enhances the raw visual flow vector when it is increasing to emphasize new stimuli in the scene, while Eq.(2) implements a Gaussian spatial smoother of the stimuli to reduce the effects of noise. The top of Fig. 2a shows an example of a typical stimulus pattern in the two dimensional neural network due to a moving object at the top-left of the camera image. In general, we could have multimodal sensory inputs, e.g. from color detectors, edge detectors, audio input, etc., feeding into Eq.(3) as a sensory signal. As suggested by Niebur, Itti and Koch [9, 10], it would be useful to weight these inputs according to their importance in the scene, usually based on some top-down feedback or task-specific biasing (e.g., if we know that color is more important than motion).

This stimulus dynamics feeds into a *saliency map* [12], essentially a winner-take-all (WTA) network which decides on a winning stimulus from many simultaneous stimuli in the camera field. The winning stimulus will become the next saccade target or focus of overt attention. The WTA network is realized based on the theory of *neural fields*, a spatial neural network inspired by the dynamics of short range excitatory and long range inhibitory interactions in the neo-cortex [1, 2]. The activation dynamics $u(\boldsymbol{x}, t)$ of the saliency map is expressed as:

$$\begin{aligned} \tau \dot{u}(\boldsymbol{x}) &= -u(\boldsymbol{x}) + S(\boldsymbol{x}) + h \\ &+ \sum_{\boldsymbol{x}'} \boldsymbol{w}(\boldsymbol{x}, \boldsymbol{x}') \sigma(u(\boldsymbol{x}')) \end{aligned} \tag{4}$$

Here, h is the base line activation level within the field, $S(\boldsymbol{x}, t)$ is the external stimulus input (Eq.1), $\boldsymbol{w}(\boldsymbol{x}, \boldsymbol{x}')$ describes the coupling strength between all the units of the network, and $\sigma(u)$ controls the local threshold of activation. Depending on the choice of parameter h and the form of σ and $\boldsymbol{w}$, the activation dynamics of Eq.(4) can have various stable equilibrium points [1]. We are interested in a solution which has uniform activation at base line level in the absence of external stimuli, and which forms a unimodal activation pattern at the most significant stimulus in the presence of stimuli that are possibly dispersed throughout the spatial network. This is achieved by choosing a transfer function:

$$\sigma(u) = 1/(e^{(-cu)} + 1) \tag{5}$$

with constant $c >> 1$ and an interaction kernel with short range excitation and long-range inhibition term H_0:

$$\boldsymbol{w}(\boldsymbol{x}, \boldsymbol{x}') = k e^{-(\boldsymbol{x}-\boldsymbol{x}')^2/\sigma_w^2} - H_0 \tag{6}$$

The values of the constants H_0, k, σ_w^2, and c have to be decided based the magnitude of the stimulus dynamics $S(\boldsymbol{x}, t)$, as outlined in [1].

In addition to the stimulus driven dynamics, we also supress the activation of the most recently attended location by adding a large negative activation in Eq.(3) at the location of the last saccade target. This strategy implements an *inhibition of return* [10] and ensures that the robot does not keep attending to the same location in the continuous presence of an interesting stimuli. The plots at the

bottom of Fig. 2(a)(b) illustrate the behavior of the activation dynamics just before and after an attention shift, including the effect of the negative activation after the saccade.

2.2 Planning and Generation of Motor Commands

Given a new saccade target, extracted from the saliency map, the direction of gaze needs to be shifted to the center of this target. Since fifth order splines are a good approximatin of biological movement trajectories [11, 3], we use this model to compute a desired trajectory from the current position $\boldsymbol{x}_0$ to the target $\boldsymbol{x}_f$, all expressed in camera coordinates:

$$disp = \boldsymbol{x}_f - \boldsymbol{x}_0 \quad (7)$$

$$\tau = t/T \quad (8)$$

$$\boldsymbol{x}(t) = \boldsymbol{x}_0 + disp * (15\tau^4 - 6\tau^5 - 10\tau^3) \quad (9)$$

$$\dot{\boldsymbol{x}}(t) = disp * (60\tau^3 - 30\tau^4 - 30\tau^2) \quad (10)$$

The movement duration T was chosen such that the maximal movement velocity was the same for each saccade, i.e., $T = disp * 0.2[s]$ in our implementation.

The camera-space trajectory is converted to joint space by inverse kinematics computations based on Resolved Motion Rate Control (RMRC). We assume that only head and eye motion is needed to shift the gaze to the visual target, an assumption that is justified given that the target was already visible in the peripherial field of view. For the time being, the inverse kinematics computation is performed for the right eye only, while the left eye performs exactly the same motion as the right eye. Thus, we need to map from a 2D camera space of the right eye to a 5D joint space, comprised of pan and tilt of the camera, and 3 DOFs of the robot's neck. To obtain a unique inverse, we employ Liegeois [14] pseudo-inverse with optimization:

$$\dot{\boldsymbol{\theta}} = \mathbf{J}^{\#}\dot{\boldsymbol{x}} + (\mathbf{I} - \mathbf{J}\mathbf{J}^{\#})\boldsymbol{k}_{null} \quad (11)$$

$$\text{where } \mathbf{J}^{\#} = \mathbf{J}^T(\mathbf{J}\mathbf{J}^T)^{-1}$$

$\boldsymbol{k}_{null}$ is the gradient of an optimization criterion w.r.t. to the joint angles $\boldsymbol{\theta}$. The second term of the Eq.(11) is the part that controls the movement in the null space of the head-eye system. Any contribution to $\dot{\boldsymbol{\theta}}$ from this term will not change the direction of gaze but will only change how much we use the head or eye DOFs to realize that gaze. As optimization criterion we chose:

$$L = \frac{1}{2}\sum_i w_i(\theta_i - \theta_{def,i})^2 \quad (12)$$

resulting in

$$k_{null,i} = \frac{\partial L}{\partial \theta_i} = w_i(\theta_i - \theta_{def,i}) \quad (13)$$

(14)

This criteron keeps the redundant DOFs as close as possible to a default posture $\boldsymbol{\theta}_{def}$. Adding the weights w_i allows giving more or less importance to enforcing the optimization criterion for certain DOF–this feature is useful to create natural looking head-eye coordination.

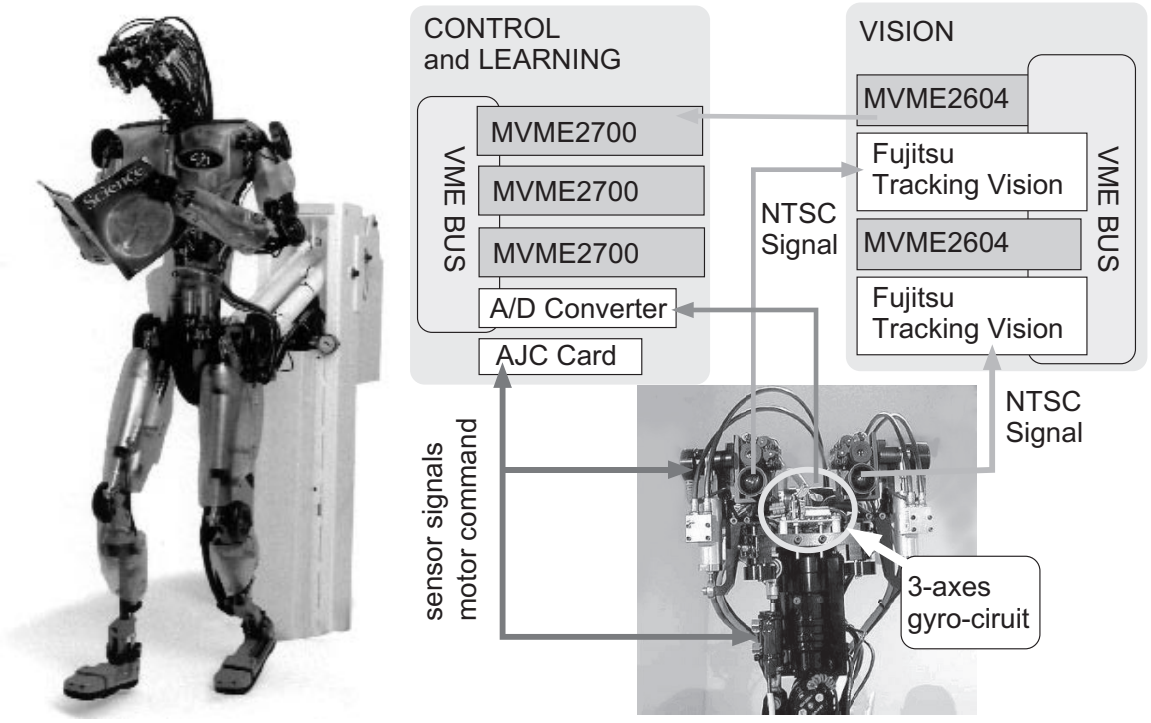
Once the desired trajectory is converted to joint space, it is tracked by an inverse dynamics controller using a learned inverse dynamics model [19].

3 Interaction Issues

Several issues of our visual attention system require special consideration. First, there is the problem of maintaining a frame of reference for the saliency map. When the robot makes an overt shift of attention, the camera coordinates are changed and the locations of the current stimulus and activation dynamics need be to updated accordingly. In our current implementation, we use a *camera-centric* frame of reference and shift all the stimulus and activation patterns relative to the center of the visual field. Locations that fall out of the saliency map are discarded. Obviously, such a remapping strategy cannot guarantee accurate re-mapping over a chain of movements. However, this inaccuracy are negliable on the time scale of the dynamics of the stimulus and saliency map dynamics.

Another important problem is to stabilize the image on the cameras when the body of the robot moves or there are external perturbations. In [18], we demonstrated how image stabilization can be achieved with learning approaches and we will integrate this strategy in our attentional system. As a result, re-mapping of the saliency map and stimulus dynamics will not be necessary when the robot head moves involuntarily.

A rather difficult issue for attention arises from the need to neglect self-motion stimuli, i.e., visual flow that is caused either by the motion of the oculomotor system or by body parts of the robot that are in the view of the camera eyes. Currently, we circumvent this problem by discarding stimuli during the time the robot moves its eyes and head. In future work, we will address to predict optical flow from self-motion and actively suppress such false stimuli.



(a) (b)

Figure 3: (a) The 30 DOF Humanoid robot used as a testbed for implementation (b) The experimental setup of the humanoid vision-head system

4 The experimental setup

Fig. 3(a) shows the 30 degree-of-freedom(DOF) humanoid robot that we use as our testbed. Each DOF of the robot is actuated hydraulically out of a torque control loop. Concentrating on the oculomotor specifications: each eye of the robot's oculomotor system consists of two cameras, a wide angle (100 degrees view-angle horizontally) color camera for peripheral vision, and a second camera for foveal vision, providing a narrow-viewed (24 degrees view-angle horizontally) color image. This setup mimics the foveated retinal structure of primates, and it is also essential for an artificial vision system in order to obtain high resolution vision of objects of interest while still being able to perceive events in the peripheral environment. Each eye has two independent degrees of freedom, a pan and a tilt motion.

Fig. 3(b) shows the oculomotor system in detail. Two subsystems, a control (and learning) subsystem and a vision subsystem, are setup in each VME rack and carry out all necessary computations out of the real-time operating system VxWorks. Three CPU boards (Motorola MVME2700) are used for the movement control and the sensory processing subsystem, and two CPU boards (Motorola MVME2604) are dedicated to the vision subsystem. In the movement control/sensory processing subsystem, CPU boards are used, respectively, for: i) sensory processing and saccade target generation ii) dynamics learning and task execution (behaviors like overt shifts of attention), and iii) low level motor control of head, eye and other body joints of our robot (compute torque mode). All communication between the CPU boards is carried out through the VME shared memory communication which, since it is implemented in hardware, is very fast.

In the vision subsystem, each CPU board controls one Fujitsu tracking vision board in order to calculate the visual flow. We use optical flow calculations based on the block-matching method [8] which is performed by the Fujitsu Tracking Vision board in real-time. For our experiments, the visual flow is computed on a grid of 25x25 nodes spread evenly over the entire peripheral visual field. This resolution was decided based on real time data transmission and computation bounds of our current setup, although scaling it up would just require a faster processor and faster data transmission. At

each of the flow computation nodes, an 8x8 pixel window is compared for the best fit at surrounding neighbouring locations in the next video frame, and the vision tracking hardware gives us an optical flow vector (direction and magnitude) based on the best fit and also a matrix of confidence bounds distributions. Confidence information helps us getting rid of noise and ambiguities arising from plain non-textured background.

NTSC video signals from the binocular cameras are synchronized to ensure simultaneous processing of both eyes' vision data. Raw extracted vision data (in our case, the optical flow) are sent via a serial port (115200 bps) to the control (and learning) module. This is where the sensory processing, described in detail in Section 2.1 take place. For the experimental demonstrations of this paper, the image from only one peripheral camera is processed for visual flow computations and the motion of the two eyes are coupled together in both its horizontal (pan) and vertical (tilt) degrees-of-freedom. Multiple degrees of freedom per camera, and multiple eyes just require a duplication of our circuits and a correction for vergence under small focal length. In order to mimic the semicircular canal of biological systems, we attached a three-axis gyro-sensor circuit to the head (Murata Manufacturing). From the sensors of this circuit, the head angular velocity signal is acquired through a 12 bit A/D board such that active image stabilization can be performed when the head is perturbed. The oculomotor and head control loop runs at 480 Hz, while the vision control loop runs at 30 Hz.

Figure 4: Snap shots of the robot's peripheral view before, during, and after an attentional head-eye saccade, taken at 30 Hz sampling rate. Superimposed on the images is the visual flow field.

5 Results and Discussion

We implemented the visual attention system on our humanoid robot. The stimulus dynamics and saliency map had 44x44 nodes, i.e., twice the length and width of the 22x22 nodes of the visual flow grid of the peripheral vision. This extended size assured that after a saccade, the remapping of the saliency map and stimulus dynamics could maintain stimuli outside of the peripheral vision for some time. The Jacobian needed for the inverse kinematics computation was estimated with linear regression from data collected from moving the head-eye system on randomized periodic trajectories for a few minutes. Due to the limited range of motion of the eye and head DOFs, the Jacobian could be assumed to be constant throughout the entire range of motion of head-eye system, which was confirmed by the excellent coefficient of determination of the regression of the Jacobian. In an alterntive approach, we also directly learned the inverse kinematics as described in D'Souza et al. [7], which yielded equally good results as the analytical method using the regressed Jacobian. The saliency map was able to determine winning targets at about 10Hz, which is comparable to the capabilities of the human attentional system.

An illustration of the working of the attentional system is provided in Fig. 4. The top image shows the robot's right eye peripheral view of the lab, focussing on the person in the middle of the image. At the bottom left part of the image, another person was waving a racket to attract the robot's attention. This motion elicited a saccade, recognizable from the middle image of Fig. 4 which shows

the visual blur that the robot experienced during the movement. The bottom image of Fig. 4 demonstrates that after the saccade, the robot was correctly focusing on the new target. Note that the three images were sampled at 30Hz, indicating that the robot performed a very fast head-eye saccade of about 60ms duration, which is again comparable to human performance. Our future work will augment the attentional system with more sensory modalities, including the learning of weighting of sensory modalities for different tasks.

References

[1] S. Amari. Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, 27:77–87, 1977.

[2] S. Amari and M.A. Arbib. Competition and cooperation in neural nets. In Metzler, editor, *Systems neuroscience*, pages 119 – 165. Academic Press, New York, 1977.

[3] G. R. Barnes. Visual-vestibular interaction in the control of head and eye movement: the role of visual feedback and predictive mechanisms. *Progress in Neurobiology*, 41:435–472, 1993.

[4] C. Breazeal, B. Scassellati. A context dependent attention system for a humanoid robot. *Proceedings of IJCAI-99*, 1146–1151, 1999.

[5] T. Bergener, C. Bruckhoff, P. Dahm, H. Jansen, F. Joublin, R. Menzner, A. Steinhage, and W. von Seelen. Complex behaviour by means of dynamical systems for an anthropomorphic robot. *Neural Networks*, 12:1087–1099, 1999.

[6] J.A. Driscoll, R.A. Peters II, K.R. Cave. A visual attention network for a humanoid robot. In *Proceedings of International Conferfence on Intelligent Robots and Systems (IROS-98)*, pages 1968–1974, 1998.

[7] A. D'Souza, S. Vijayakumar, S. Schaal. Learning inverse kinematics. Submitted to the *International Conferfence on Intelligent Robots and Systems (IROS-2001)*.

[8] H. Inoue, M. Inaba, T. Mori, and T. Tachikawa. Real-Time Robot Vision System based on Correlation Technology. In *Proceedings of International Symposium on Industrial Robots (ISIR)*, pages 675–680, 1993.

[9] L. Itti and C. Koch. A comparison of feature combination strategies for saliency-based visual attention systems. In *Proc. SPIE Human Vision and Electronic Imaging IV (HVEI'99), San Jose, CA*, volume 3644, pages 473–82, January 1999.

[10] L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10-12):1489–1506, 2000.

[11] M. Kawato. Internal models for motor control and trajectory planning. *Current Opinion in Neurobiology*, 9:718–727, 1999.

[12] C. Koch and S. Ullman. Selecting one among the many: A simple network implementing shifts in selective visual attention. A.I. Memo 770, Massachusetts Institute of Technology, January 1984.

[13] K. Kopecz and G. Schoner. Saccadic motor planning by integrating visual information and pre-information on neural dynamic fields. *Biological Cybernetics*, 73:49–60, 1995.

[14] A. Liegeois. Automatic supervisory control of the configuration and behavior of multibody mechanisms. *IEEE Transactionsn on Systems, Man, and Cybernetics*, 7:868–871, 1977.

[15] E. Neibur and C. Koch. Computational architectures for attention. In R. Parasuraman, editor, *The Attentive Brain*, pages 163–186. MIT Press, Cambridge, MA, 1998.

[16] R. Parasuraman. *The Attentive Brain*. MIT Press, Cambridge, MA, 1998.

[17] R. Rao and D. Ballard. Learning saccadic eye movements using multiscale spatial filters. In G. Tesauro, D. S. Touretzky, and T. K. Leen, editors, *Advances in Neural Information Processing Systems 7 (NIPS 94)*, pages 893–900. MIT Press, 1995.

[18] T. Shibata and S. Schaal. Biomimetic gaze stabilization based on feedback-error-learning with nonparametric regression networks. *Neural Networks*, 14(2), 2001.

[19] S. Vijayakumar and S. Schaal. LWPR : An O(n) algorithm for incremental real time learning in high dimensional space. In *Proc. International Conference on Machine Learning (ICML 2000)*, pages 1079–1086, 2000.