Learning Visuo-Tactile Coordination in Robotic Systems

M. Rucci

Scuola Superiore S.Anna
Pisa, Italy
mickey@nsi.edu

R. Bajcsy

University of Pennsylvania
Philadelphia, PA
bajcsy@grip.cis.upenn.edu

Abstract

As it occurs in humans, robotic systems should be able to respond to unexpected tactile events by orienting their visual attention toward the location of the stimuli. This implies two basic problems: first it is necessary to develop a general method for integrating attentive processes which belong to different sensory modalities according to the attended task. Then, for the specific case of touch-driven shift of gaze, a sensorymotor transformation needs to be identified, which links the stimulation of tactile receptors to the spatial position of the camera, via the current posture of the system. In this paper we describe a general framework for integrating multimodal attentive mechanisms, and we show how the visuo-tactile coordination can be autonomously learnt on the basis of sensory consistency and feedback. After the general presentation of the method, we consider the case of a robotic system composed of a 2 d.o.f. arm and a 2 d.o.f. head. Experiments with this system show that it discovers its own functional model without any external intervention and adapts it continuously during normal operation. The approach gives good results while presenting the advantages of autonomy and adaptability.

1 Introduction

Useful autonomous robotic systems need to operate in real environments, dealing with unpredictable situations and huge amount of redundant sensory data. Selective attention and learning are critical concepts for the development of these systems: from one side, attentive processes, as mechanisms which select the relevant information for the accomplishment of the task at hand, allow to deal with large flows of incoming data, and they can be crucial for overcoming classical perceptual difficulties and achieving real-time performances. From the other, learning capabilities allow an adaptation to the surrounding environment, and even the discovering of system specific functional and structural characteristics, which may compensate for possible alterations of the system components due to damages or aging. In order to be effective, learning should occur throughout all the operative "life" of a system, and algorithms which require the existence of a separate learning phase from the operative one cannot be used.

Whereas selective attention, after having been emphasized by several recent machine perception paradigms [1, 2, 3], is currently the subject of an increasing number of studies in the field of computer vision [4, 5], attentive mechanisms related to sensory modalities different from vision have been much less studied. In particular, the class of processes related to the sense of touch has been so far only superficially investigated, in spite of the importance of tactile events for all the systems that physically interact with the surrounding environment.

A typical situation where a *somatosensory saccade* -i.e. a shift of the gaze direction triggered by a tactile event- is needed in a robotic system, is when, during a motor operation, an external obstacle is hit. In this case, the visual system needs to be focused on the spatial location of the collision, in order to assess the nature of the obstacle and determine a new motor path. The process by which the final gaze direction is estimated involves the transformation of the tactile event in a somatosensory reference frame into a corresponding activation of the motors of the camera. Even if this transformation can be estimated by means of an accurate modelization process, the performance obtained in this way is fixed and does not adapt to possible alterations of the system. Furthermore, the results are highly dependent on the accuracy of the model, and the method is not appliable if a good mathematical

This work was developed at the GRASP Laboratory of the University of Pennsylvania. M. Rucci is currently at The Neurosciences Institute, La Jolla, CA 92037.

IEEE International Conference
on Robotics and Automation
0-7803-1965-6/95 $4.00 ©1995 IEEE

model of the system cannot be produced, such as, for example, with highly nonlinear systems.

In this paper, we analyze touch-based attentive mechanisms and we investigate their integration with attentive processes belonging to other sensory modalities. In particular, we consider the production of somatosensory saccades in a robotic system composed of a monocular head and a tactile sensitive arm, by following an approach based on machine learning. The system autonomously discovers the sensorymotor transformation which links tactile events to visual saccades, on the basis of multisensory consistencies, sensory feedback, and basic, built-in, motor reflexes. In this way, it builds, without any external intervention, a functional model of itself, which is continuously updated and adapted during normal operation, through on-line learning capabilities. In addition, there is no need of having an initial model to refine, as in other autonomous calibration method [6].

In the following section the architecture for integrating multimodal attentive mechanisms is briefly presented (a more complete description can be found in [7]). The neural network implementation of the sensorymotor coordination for the case of vision and touch is presented in section 3, and robotic results are shown in section 4. Finally, conclusions are drawn in section 5.

2 An architecture for integrating multimodal attention processes

Fig. 1 illustrates a general system architecture for the real–time, task–based integration of attention mechanisms operating in different sensory modalities [7]. The system is organized in a sensorimotor loop: the analysis of the incoming data produces a set of possible gaze directions. Within this set, the actual direction is selected on the basis of the absolute strength of the stimulus and of its importance in the context of the task. After that the shift of gaze has been executed, a new set of interesting directions is generated and is added to the previous one.

The logical center of the architecture is a modified saliency map $\mathcal{S}$ [8], whose location s_{ij} represents the saliency of a specific visual direction (ϕ_i, ψ_i) in a absolute head–centered reference frame. All possible visual directions are represented on the saliency map and a monotonic mapping exists between the saliency map and the motor of the cameras, so that, given a specific location of the map, corresponding positions of the cameras are determined. It is not required to

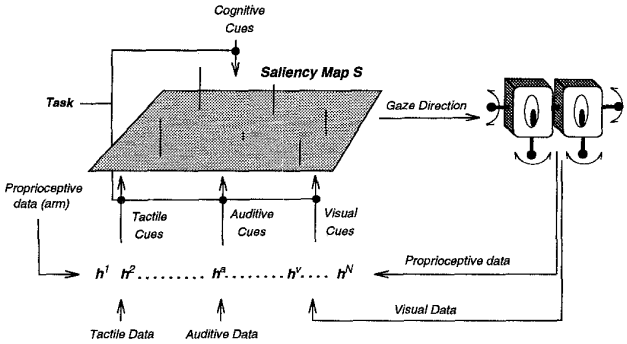

Figure 1: The proposed architecture.

specify exactly which visual direction corresponds to a specific map location. The actual transformation is autonomously learned by the system so as to compensate for inaccuracies and alterations of the mechanical and optical systems.

Let $D = \{d_1, \ldots, d_M\}$, be the input perceptual data to the system at time t and P the current posture of the system $P = \{p_1, \ldots, p_L\}$. As illustrated in Fig. 1, sensory data are analyzed by a set of continuous–time processes $\{\mathbf{h}^1, \ldots, \mathbf{h}^N\}$, for each of the sensory modalities.

$$\mathbf{h}^k(D^k, P^k) = \{l_{ij}^k\} = \begin{pmatrix} l_{11}^k & \cdots & l_{1,C}^k \\ \vdots & \ddots & \vdots \\ l_{R,1}^k & \cdots & l_{R,C}^k \end{pmatrix} \quad (1)$$

where each process acts on a subset of the input data D^k –typically belonging to a single sensory modality–, and gives the saliency $l_{ij}^k \in [0, 1]$ for each location i, j on the saliency map. Here R and C are the number of units in each dimension for the saliency map. The $l_{ij}^k > 0$ are the *attentive cues* generated by process $\mathbf{h}^k$.

Note that, in addition to the perceptual information, $\mathbf{h}^k$ depends also on the posture of some parts of the system, that is $\mathbf{h}^k = \mathbf{h}^k(D^k, P^k)$. For example, all the processes that operate on visual data provide cues located in the visual field, and the projection of the visual field on the saliency map changes with the position of the eyes with respect to the head. In general, each attentive process carries out the coordinate transformation necessary for activating the head–centered saliency map starting from data expressed in a sensory reference frame.

All the processes contribute to activate the saliency map, so that the final value assumed by element s_{ij} is given by

$$S_{ij} = F_s(\sum_k w_T^k I_{ij}^k) \quad (2)$$

where F_s is a nonlinear monotonic function (in the experiments, a sigmoidal function has been applied) in $[0, 1]$. The attentive cues produced by processes $\{\mathbf{h}^1, ..., \mathbf{h}^N\}$ are candidates for drawing attention. Several rules can be implemented for selecting the actual direction of gaze: the simplest is to choose the direction corresponding to the unit with maximum activation.

The dependence on the task at hand is produced by the task weights w_T^k,

$$\mathbf{w}_T = (w_T^1, w_T^2, \ldots, w_T^N)^T \qquad w_T^k \in [0, 1] \quad (3)$$

which modulate the cues of each sensory process $\mathbf{h}^k$ accordingly to the attended task, so that at every time

$$\sum_{k=1}^{N} w_T^k = 1 \quad (4)$$

In this way, a priority degree can be assigned to different sensory features. based on the task. By properly arranging the weight values, it is possible to select a sensory stimulus with respect to others and/or to inhibit irrelevant cues.

3 A neural network-based implementation

Attentive processes $\mathbf{h}^k$ link sensory reference frames (for example, the activation of specific tactile receptors or of pixels in the input visual image) to motor reference frames (gaze directions, i.e. corresponding positions of the visual system). Of course, such sensorymotor transformations depend on the actual physical structure of the considered system and on the current position of the mobile components.

In this section, we describe how these processes can be autonomously developed by means of unsupervised learning mechanisms, by considering the specific case of a system which includes a visual ($\mathbf{h}^v$) and a tactile ($\mathbf{h}^c$) process.

Two different learning mechanisms act simultaneously in the system: *before* the execution of a shift of gaze, the consistency among the cues produced by different processes is used for updating the sensorymotor coordinations. If a physical event is monitored with different sensory modalities, all the visual directions produced by the corresponding processes should obviously be coincident. Their mismatches can be used as a measure, which is internally available to the system, for refining the processes. *After* the execution of a motor action, learning occurs on the basis of sensory feedback. For example, the foveation error detected after the execution of a visual saccade can be used for updating the sensorymotor transformations. Examples of application of consistency-based [9, 10, 11, 12], and feedback-based [13, 14] learning mechanisms are common in the neural network literature [15].

In the initial stage of development, an *exploration task* has been selected which gives priority to visual stimuli with respect to the tactile ones, that is the task weights are set so that the weight for visual cues is larger than the other. In this way, visuomotor coordination can be developed faster than cutaneomotor, and the visual sensory modality can be used as a reference in the consistency-based learning process.

In the proposed implementation, two input *sensory maps* $\mathcal{C}$ and $\mathcal{R}$ code the incoming tactile and visual stimuli in a somatotopic and retinotopic reference frame, respectively. That is, both the maps show a topological organization where units close to each other are sensitive to stimuli occurring in adjacent locations of the receptors layout. Two input *motor maps* $\mathcal{M}^v$ and $\mathcal{M}^c$, code at each time the position of the system, as detected by proprioceptive data (the data provided by robot encoders). In particular, $\mathcal{M}^c$ represents the posture of the parts of the system which have tactile capabilities, and $\mathcal{M}^v$ code the camera position. The units of all the input maps are characterized by gaussian receptive fields, so that the activation value of each unit is a gaussian function of the distance between the input and a specific value for the unit. As illustrated in Fig. 2, in both the sensory modalities the input sensory and motor maps activate the units of a three-layered sensorytopic columnar organization. In the visual sensory modality each column is composed of three units v_{ij}, v_{ij}^{ϕ}, v_{ij}^{ψ} located in the maps $\mathcal{V}$, $\mathcal{V}^{\phi}$ $\mathcal{V}^{\psi}$, respectively. Their activation is given by:

$$\begin{aligned} V_{ij} &= d_{ij} R_{ij} \\ V_{ij}^{\phi} &= F_\tau(V_{ij}) \, (\textstyle\sum_{pq} w_{pq}^{\phi} M_{pq}^v + y_{ij}^{\psi}) \\ V_{ij}^{\psi} &= F_\tau(V_{ij}) \, (\textstyle\sum_{pq} w_{pq}^{\psi} M_{pq}^v + y_{ij}^{\psi}) \end{aligned} \quad (5)$$

where F_τ is a step function with threshold τ, R_{ij} is the activation of unit r_{ij} in the retinotopic sensory map $\mathcal{R}$, and M_{pq}^v is the activation of unit m_{pq}^v in the visual motor map $\mathcal{M}^v$. The units of the bottom map $\mathcal{V}$ are fully connected with the units of the Saliency map $\mathcal{S}$. However, the strength of the connections are weighted as a function of the activation of the other two units of

the same column $< i, j >$, so that a spatial inhibitory organization is present in the connection scheme. The connection weight between units v_{pq} and s_{ij} is given by

$$\begin{cases} a_{pq}^{ij} = 1 & \text{if } \|(V_{ij}^{\phi}, V_{ij}^{\psi}) - (i/N_s^{\phi}, j/N_s^{\psi})\| < \tau_v \\ a_{pq}^{ij} = 0 & \text{otherwise} \end{cases} \tag{6}$$

where τ_v is a *a priori* set threshold, and N_s^{ϕ} and N_s^{ψ} are the numbers of units along the two directions of the Saliency Map. What happens is that the activation of units v_{ij}^{ϕ}, v_{ij}^{ψ} determines where to project on the saliency map the retinotpoic input in position (i, j).

Learning occurs by properly modifying the weights y_{ij} and w_{ij}. Weights are updated on the basis of the retinotopic error $\epsilon = (\epsilon_x, \epsilon_y)$ registered after the execution of a visuomotor saccade (sensory feedback learning).:

$$\begin{aligned} y_{ij}^{\phi}(t+1) &= y_{ij}^{\phi}(t) + k_y^{\phi}\epsilon_x V_{ij} \\ w_{ij}^{\phi}(t+1) &= w_{ij}^{\phi}(t) + k_m^{\phi}\epsilon_x M_{ij}^{v} \end{aligned} \tag{7}$$

$$\begin{aligned} y_{ij}^{\psi}(t+1) &= y_{ij}^{\psi}(t) + k_y^{\psi}\epsilon_y V_{ij} \\ w_{ij}^{\psi}(t+1) &= w_{ij}^{\psi}(t) + k_m^{\psi}\epsilon_y M_{ij}^{v} \end{aligned} \tag{8}$$

In the visual system a linear model can be adopted by adding separately the visual and motor contributions, since they can always be considered indipendent for every position of the cameras and the stimuli. In the tactile system a similar linear separation is not feasible: foveation angles are a nonlinear function of the position of the tactile stimulus in the cutaneotopic reference frame and of all the angles defining arm position. Thus, in the columnar organization in Fig. 2 the activation of the units t_{ij}, t_{ij}^{ϕ} and t_{ij}^{ψ} in the three layers $\mathcal{T}$, $\mathcal{T}^{\phi}$ and $\mathcal{T}^{\psi}$, is given by

$$\begin{aligned} T_{ij} &= q_{ij} C_{ij} \\ T_{ij}^{\phi} &= F_{\tau}(T_{ij}) \left(\textstyle\sum_{pq} z_{pqij}^{\phi} M_{pq}^{c}\right) \\ T_{ij}^{\psi} &= F_{\tau}(T_{ij}) \left(\textstyle\sum_{pq} z_{pqij}^{\psi} M_{pq}^{c}\right) \end{aligned} \tag{9}$$

where C_{ij} is the activation of unit i, j in the cutaneotopic sensory map $\mathcal{C}$ and M_{pq}^{c} the activation of unit m_{pq}^{c} in the tactile motor map $\mathcal{M}^{c}$. Also in the tactile sub–system, the units of the bottom map $\mathcal{T}$ are fully connected with the units of the Saliency map, and the connections are inhibited by the activation of the units of the other two layers. The connection weight between units t_{pq} and s_{ij} is given by

$$\begin{cases} b_{ij}^{pq} = 1 & \text{if } (\|(T_{ij}^{\phi}, T_{ij}^{\psi}) - (i/N_s^{\phi}, j/N_s^{\psi})\| < \tau_c \\ b_{ij}^{pq} = 0 & \text{otherwise} \end{cases} \tag{10}$$

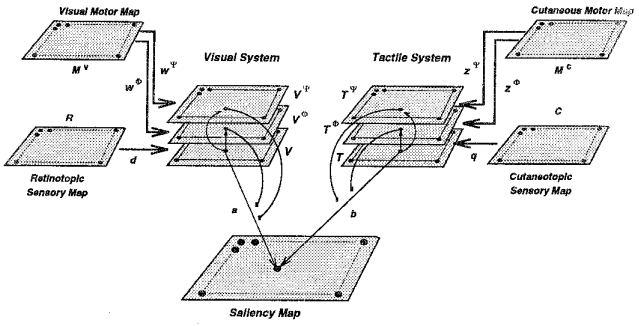

Figure 2: In both the sensory modalities, the activation of the motor–proprioceptive maps and the sensory maps are combined in a sensorytopic columnar organization which produces the corresponding cues for the saliency map.

In the tactile system adaptation is provided by changes in the weights $z_{pqij}^{\phi}, z_{pqij}^{\psi}$. In this case, both the consistency and feedback learning processes contribute to updating the connection weights. If the tactile stimulus has a visual counterpart which happens to be in the visual field, then vision acts as a reference sensory modality, and the difference between the visual and tactile cues on the saliency map is used as a target error for improving performances. If only a tactile stimulus is present, a somatosensory saccades is attempted on the basis of the current status of the system, and the resulting retinotopic error is then used. In both the cases weights are updated as

$$\begin{aligned} z_{pqij}^{\phi}(t+1) &= b_{pqij}^{\phi}(t) + k_b^{\phi}\delta_x M_{pq}^{c} \\ z_{pqij}^{\psi}(t+1) &= b_{pqij}^{\psi}(t) + k_b^{\psi}\delta_y M_{pq}^{c} \end{aligned} \tag{11}$$

where $\delta = (\delta_x, \delta_y)$ can be the retinotopic (feedback-based) or the angular (consistency-based) error, depending on which learning process is applied. In the tactile subsystem it is worth noting that, even if the full connectivity of the adaptive layer may induce to suppose that a large number of connections is required, this is not necessarily the case. In general, a high accuracy of somatosensory saccades is not necessary, thus a smaller number of units in both the motor and sensory maps can be employed. A lower accuracy of somatosensory saccades with respect to visuomotor ones has been found also in humans [16].

4 Experimental Results

In the experiments, two robotic manipulators PUMA 500 were used. One of the two manipulators

was used as a head/eye visual system, with a b/w camera mounted as an end–effector. Only the last two joints of the manipulator were allowed to move, so that the visual system was provided with two degrees of freedom ψ (pan) and ϕ (tilt). On the other PUMA a tactile sensitive probe was mounted as end–effector. For this purpose, a Force/Torque sensor was used, and the location of contact was derived by the monitored data values, under the assumption that only a single contact occurred at any time. Also the manipulator holding the tactile probe was allowed of 2 d.o.f. corresponding to movements along the first two joints of the arm

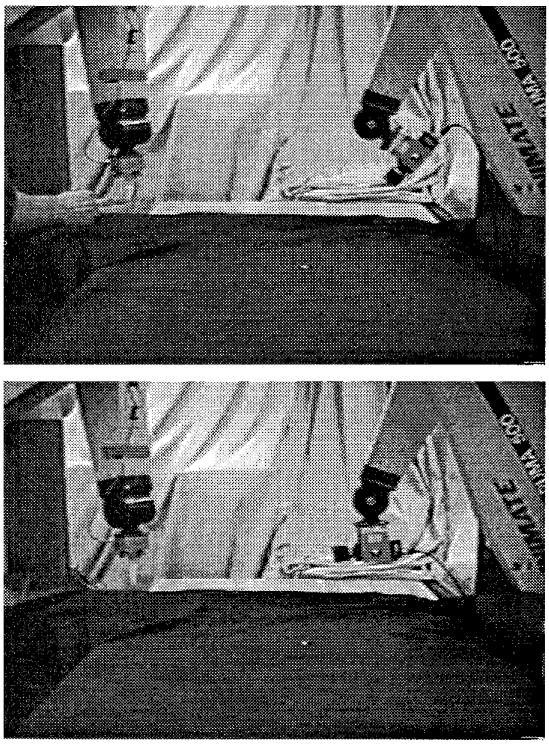

Figure 3: Execution of a somatosensory saccade: system attention is first drawn by a visual stimulus, then by a tactile one.

Preprocessing was carried out in both the visual and tactile systems. As regards tactile data, the activation of the input cutaneotopic map coded the position of the tactile event on the tool, evaluated as the distance z_f from the bottom of the tool. Preprocessing in the visual system allowed the evaluation of the position of the contact between the tactile probe and external tool. This was achieved by thresholding the image and using suitably colored tools (both the end–effector and the tip of the tool used for providing stimulation were painted so as to be differentiated in the grey-level hystogram).

Fig. 3 shows the execution of visuomotor and somatosensory saccades. System performances at different levels of learning are shown in table 1 and 2. The values show that accuracy improves gradually with experience, and training times are not long. In both the cases, good performances were achieved in less than two hours (600 stimuli).

Foveation Error		
it	*mean*	σ^2
50	0.17	0.26
200	0.08	0.05
400	0.05	0.03
600	0.02	0.01

Table 1: Accuracy of visuomotor saccades at different learning levels (percentage of the visual field)

Foveation Error		
it	*mean*	σ^2
100	0.20	1.66
300	0.10	0.59
600	0.07	0.14

Table 2: Accuracy of Somatosensory Saccades at different learning levels (percentage of the visual field)

5 Conclusions

The system described in this paper provides an example of autonomous adaptive system with multisensory attentive capabilities. The proposed architecture is specifically designed for integrating attentive mechanisms belonging to different sensory modalities, and for providing an intrinsic dependence on the task at hand. In addition, learning capabilities have been included so as to build adaptive sensorymotor coordinations. As a result, the system develops its own functional models, and changes the way it interacts with the world according to the goal to accomplish.

A number of innovative aspects are present in the system. First of all, it address the problem of implementing touch-driven attention mechanisms in machines. So far, only few works have investigated non-visual attentive processes. In the architecture described in this paper, visual and non visual processes

operate in the same way and no formal distinction is required. As regards learning, we have investigated the coexistence of different processes, which contributed to increasing robustness. In addition, we have shown results in the case of real robotic applications (and not simplified simulations).

The obtained results propose the approach as an alternative method for the calibration of complex robotic systems, with the advantages of autonomy and continuous adaptability. This is important when a sensory modality like touch, which require a physical contact, is included, since the extension of more traditional calibration methods to this case is not immediate and an external operator (or a highly structured environment) is usually required.

Several directions of future research are possible. From one side, it could be interesting to apply the approach to more sophisticated robotic systems and analyze more complex tasks with a large number of processes. For example, it could be interesting to apply the architecture to the implementation of touch-driven attention mechanisms in the context of manipulation with multifingered robotic hands. From the other, a number of theoretical issues can be further investigated, such as the autonomous evaluation of suitable task weights for performing specific tasks, or the inclusion in the architecture of other motor control procedures.

Acknowledgements

This work has been supported by ARPA Grant N00014-92-J-1647, ARO Grant DAAL03-89-C-0031 PRI, and NSF Grant CISE/CDA-88-22719. One of the authors (M. Rucci) has been supported by a fellowship from the Italian National Research Council.

References

[1] J. Aloimonos, I. Weiss, and A. Bandyopadhyay, "Active vision," *International Journal of Computer Vision*, vol. 1, no. 4, pp. 333–356, 1988.

[2] R. Bajcsy, "Active perception," *Proceedings of the IEEE*, vol. 76, no. 8, pp. 996–1005, 1988.

[3] D. Ballard, "Animate vision," *Artificial Intelligence*, vol. 48, pp. 57–86, 1991.

[4] P. Burt, "Smart sensing within a pyramid vision machine," *Proceedings of the IEEE*, vol. 76, no. 8, pp. 1006–1015, 1988.

[5] A. Abbott, "A survey for selective fixation control for machine vision," *IEEE J. of Control Systems*, vol. August, pp. 25–31, 1992.

[6] D. Bennet, D. Geiger, and J. Hollerbach, "Autonomous robot calibration for hand-eye coordination," *International Journal of Robotics Research*, vol. 10, no. 5, pp. 550–559, 1991.

[7] C. Colombo, M. Rucci, and P. Dario, "Attentive behavior in an anthropomorphic robot vision system," *Journal of Robotics Autonomous Systems*, 1994.

[8] C. Koch and S. Ullman, "Shifts in selective visual attention: Toward the underlying neural circuitry," in *Matters of Intelligence* (L. Vaina, ed.), D. Reidel Pub. Comp., 1987.

[9] G. Reeke, O. Sporns, and G. Edelman, "Synthetic neural modeling: The "darwin" series of recognition automata," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1498–1530, 1990.

[10] S. Grossberg, "Nonlinear neural networks: Principles, mechanisms, and architectures," *Neural Networks*, vol. 1, pp. 17–61, 1988.

[11] B. Mel, *Connectionist Robot Motion Planning.* San Diego, CA: Academic Press Inc., 1990.

[12] M. Kuperstein, "Adaptive visual-motor coordination in multijoint robots using parallel architecture," in *Proc. IEEE Int. Conf. on Robotics and Automation*, (Raleigh, NC), pp. 1595–1602, 1987.

[13] M. Kuperstein, "Infant neural controller for adaptive sensory–motor coordination," *Neural Networks*, vol. 4, pp. 131–145, 1991.

[14] M. Rucci and P. Dario, "Development of cutaneo–motor coordination in an autonomous robotic system," *Autonomous Robots Journal*, 1994. in press.

[15] A. Maren, C. Harston, and R. Pap, *Handbook of Neural Computing Applications.* San Diego CA: Academic Press Inc., 1990.

[16] J. Groh, *Coordinate Transformations, Sensorimotor Integration and the Neural Basis of Saccades to Somatosensory Targets.* University of Pennsylvania: PhD dissertation, 1993.