August 1994

Implementing Selective Attention in Machines: The Case of Touch-Driven Saccades

Michele Rucci
University of Pennsylvania

Ruzena Bajcsy
University of Pennsylvania

University of Pennsylvania Department of Computer and Information Science Technical Report No. MC-CIS-94-44.

Implementing Selective Attention in Machines: The Case of Touch-Driven Saccades

Abstract

Recent paradigms in the fields of robotics and machine perception have emphasized the importance of selective attention mechanisms for perceiving and interacting with the environment. In the case of a system involved in operations requiring a physical interaction with the surrounding environment, a major role is played by the capability of attentively responding to tactile events. By performing somatosensory saccades, the nature of the cutaneous stimulation can be assessed, and new motor actions can be planned. However, the study of touch-driven attention, has almost been neglected by robotics researchers. In this paper the development of visuo-cutaneo coordination for the production of somatosensory saccades is investigated, and a general architecture for integrating different kinds of attentive mechanisms is proposed. The system autonomously discovers the sensorymotor transformation which links tactile events to visual saccades, on the basis of multisensory consistencies and basic, built-in, motor reflexes. Results obtained both with simulations and robotic experiments are analyzed.

Keywords

selective attention, active perception, autonomous robots, machine learning

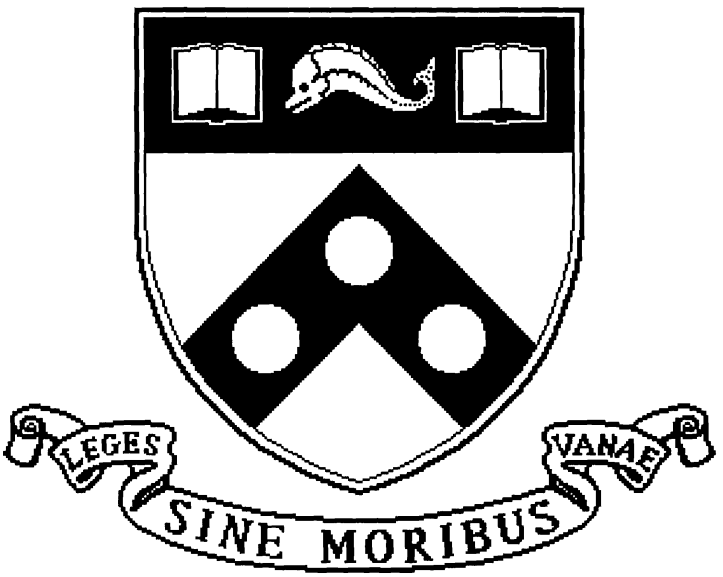
Comments

University of Pennsylvania Department of Computer and Information Science Technical Report No. MC-CIS-94-44.

Implementing Selective Attention in Machines: The Case of Touch-Driven Sccades

MS-CIS-94-44
GRASP LAB 379

Michele Rucci
Ruzena Bajcsy



University of Pennsylvania
School of Engineering and Applied Science
Computer and Information Science Department
Philadelphia, PA 19104-6389

August 1994

Implementing Selective Attention in Machines: The Case of Touch-Driven Saccades

Michele Rucci and Ruzena Bajcsy

Abstract

Recent paradigms in the fields of robotics and machine perception have enphasized the importance of selective attention mechanisms for perceiving and interacting with the environment. In the case of a system involved in operations requiring a physical interaction with the surrounding environment, a major role is played by the capability of attentively responding to tactile events. By performing somatosensory saccades, the nature of the cutaneous stimulation can be assessed, and new motor actions can be planned. However, the study of touch–driven attention, has almost been neglected by robotics researchers. In this paper the development of visuo-cutaneo coordination for the production of somatosensory saccades is investigated, and a general architecture for integrating different kinds of attentive mechanisms is proposed. The system autonomously discovers the sensorymotor transformation which links tactile events to visual saccades, on the basis of multisensory consistencies and basic, built–in, motor reflexes. Results obtained both with simulations and robotic experiments are analyzed.

KEYWORDS: Selective attention, active perception, autonomous robots, machine learning.

1 Introduction

During the last decade, researchers in the field of robotics and artificial intelligence have changed the way to approach the classical problem of developing machines which are able to perceive their environment. After the disillusion following the expectations generated by earlier approaches [Marr, 1982], new paradigms have been proposed in the second half of 1980s and in 1990s [Aloimonos *et al.*, 1988; Bajcsy, 1988; Ballard, 1991]. Even if they differ in several aspects, these paradigms agree in enphasizing the role played by an active interaction of the system with the environment, and by the task that the system has to perform. According to these approaches, sophisticated hardware tools, such as stereoscopic head–eye robotic systems [Krotkov, 1989; Pahlavan *et al.*, 1992], and ccd sensors mimicking the space-variant layout of the receptors of the retina [Sandini and Dario, 1989] have been developed, and their use is currently investigated.

In this context, selective visual attention has recently received an increasing interest, due to the role it plays in controlling human eye movements. By providing the capability of selectively processing simultaneous sources of information, attention mechanisms allow to allocate system computational resources to the process of relevant data. This is a basic step toward the goal of achieving real–time perfomances. An overview of the major contributions regarding the implementation of selective attention processes in machines is given in the following section.

Whereas selective visual attention is receiving larger and larger interest, attentive mechanisms related to sensory modalities different from vision have been much less studied. However, as research in the biological world has pointed out, the importance of these mechanisms may be comparable to that of the visual ones in many living beings. For example, attention processes based on auditive stimuli seems to have a higher priority in animals such as the barn owl, which shows also a high capability of spatially localizing auditive stimuli [Konishi, 1993].

A particular class of attentive mechanisms which has so far received very little interest, is the one related to the sense of touch. Touch–driven attention is crucial for all the systems that physically interact with the surrounding environment, and *somatosensory saccades*, i.e. visual saccades triggered by tactile stimuli, should be considered in basic interactive operations such as manipulation and navigation. By means of somatosensory saccades visual processing can be focused on obstacles that have been involuntarily hit during a motor action and new motor control strategies can be planned (see Fig. 1). In general, touch–driven attention mechanisms are powerful tools for dealing with *a priori* unknown environments.

We, humans, can switch our attention toward cutaneous stimuli without any conscious effort. Somatosensory saccades can occur whenever an unexpected tactile stimulus (such as, for example, the one produced by an insect moving on our arm) is sensed. However, the mechanisms involved in the process are not clear, and only recently researchers in neurophysiology and psychophysics have begun to investigate the point. [Groh, 1993].

The integration of attentive mechanisms which belong to different sensory modalities implies the analysis of several basic problems: first, the issue of *homogeneity* must be considered. In order to decide whether to attend to a specific cue with respect to others, comparisions among the saliency

of the cues according to the attended task and the current state of the system have to be performed. Thus, a need exists to represent attentive cues in a common frame which is independent on the original sensory modality. How to organize this representation and what exactly represents are major points to be solved.

The need for a central homogeneous representation entails the accomplishment of a series of *coordinate transformations*. Input data which are expressed in sensory reference frames (activation of the receptors of the eyes or the cameras for vision, activation of specific tactile receptors for touch, etc.), must be expressed in a new reference frame suitable for the execution of visual saccades. That is, stimuli initially encoded in a cutaneous, auditory or visual reference frame need to be converted into corresponding final activation of the muscles (motors) of the eyes (cameras).

Furthermore, the problem of *attention control* (sometimes indicated as the *where–to–look–next* problem [Rimey and Brown, 1991]) needs to be analyzed. That is, given a set of simultaneous stimuli, a strategy of scene exploration must be formulated in order to determine gaze direction at any time. It has been assessed that the task attended by the system plays a crucial role in the control strategy. Thus, a method for implementing a dependency on the task should be developed.

In this paper, a general architecture for integrating attentive mechanisms belonging to different modalities is proposed, and the implementation for the case of vision and touch is described. Sensory inputs contribute to activate a common map which represents the visual environment in a head-centered reference frame. The activation of locations of the map gives a measure of the conspicuousness of the corresponding stimulus and cues produced by sensory stimuli in different modalities can then be compared. The saliency of a given stimulus is evaluated by means of parameters whose values change in dependence on the current task and state of the system, thus producing a task dependent control of the attention flow. Sensorymotor transformations are carried out by following an approach based on autonomous system learning. Instead of explicitly modeling the structure and the functional relationships of the components, the system builds its own models on the basis of consistencies among different sensory data, and between sensory data and motor actions. Such models are continuosly refined during normal operation, so as to adapt to possible alterations of the functional parameters. Adaptability and the capability of recovering from partial damages and failures are extremely important issues for the development of autonomous robotic systems.

The work described in this paper combines a number of interesting aspects regarding the implementation of attentive processes in machines and the development of sensorymotor coordinations. Furthermore, is one of the few works in the context of visual saccades. Whereas, most of the research has been focused on other visual motor processes, not much work has been carried out on the implementation of saccades.

The paper is organized as follows: next section briefly review the state-of-the-art on the development of attention processes in machines. Section 3 gives a general overview of the proposed architecture. In section 4 the problem of developing visuo–cutaneo coordination is investigated and the resulting implementation is described in section 5. Results obtained with this approach, both in the case of simulations and with applications to real robotic systems are analyzed in section 6.

Finally, conclusions are drawn in section 7.

2 Selective attention in humans and machines

A large number of psychophysical experiments suggest that two separate stages can be singled out in the process of human visual perception [Treisman and Gelade, 1980]. In a first *preattentive stage*, a number of basic features are processed in parallel over the entire visual field. In a second *attentive stage*, the processing of visual data located in particular regions of the visual field seem to occur more rapidly than the others.

Selective attention is the capability of processing differentiately simultaneous sources of sensory information [Johnston and Dark, 1986]. Thanks to selective attention, is is possible to focus on specific inputs in the flow of incoming sensory information, while being able of switching toward salient stimuli at the occurance (cocktail-party problem [Cherry, 1953]) Some basic characteristics of attention, such as, for example, some of the factors contributing to drawing attention [James, 1890], and, for the visual case, the capability of focusing on parts of the visual field different from the fixation point [von Helmholtz, 1866], were already pointed out during last century.

In the last two decades several models of visual attention have been proposed, which attempt to explain different characteristics. Here, only the ideas which are somehow relevant for the work described in this paper are briefly considered. A general review can be found in [Kinchla, 1992].

One of the most well–known theories supports the idea that selective visual attention can be identified with a limited extent *attentional spotlight* in the visual field [Posner, 1980]. According to this model, processing of data included in the spotlight is facilitated with respect to the others. Psychophysical experiments have suggested that the spotlight can be shifted throughout the visual field and can vary in size [Tsal, 1983].

More recently, it has been proposed that a better way to think about selective visual attention is provided by the *zoom–lens* metaphor [Eriksen and James, 1986]. In the zoom–lens model, a trade-off exists between the extension of the attended area of the visual field and the resolution of detail at which the area is analyzed. As a result, even a wide range field can be attentively covered, but the resulting level of resolution is poor.

A similar width of field–resolution relationship is present in the theoretical framework proposed by Nakayama [Nakayama, 1991]. In this model, early visual processing stages are organized in a pyramidal structure, and they are linked to a fragmentary visual memory by means of a limited information bandwidth channel, the *iconic bottleneck*. Due to the limited information capability of the channel, larger fields of view can only be analyzed at lower resolution, which means that visual information is sampled at higher levels of the input pyramids.

Whereas almost all the proposed theories concern mainly with abstract theoretical issues, and less efforts are made for explaining possible implementations of the models, Koch and Ullman have proposed an interesting architecture with particular consideration for implementative aspects in the brain [Koch and Ullman, 1987]. The architecture, which attempts to explain how shifts of attention occur in humans, is based on a *saliency map*, which code an abstract measure of saliency

in the visual field. Mechanisms for switching among attentive cues are described, and a review of the neurophisiological supporting evidence is also provided.

Due to the space-variant sampling structure of the human retina, which is organized in a high-resolution, small central *fovea*, and a *periphery* whose resolution linearly decreases with eccentricity, visual exploration in humans occurs by actively shifting the fixation point, so as to exploit the detail capabilities of the fovea [Yarbus, 1967]. It is every–day experience to link gaze control to attentional mechanisms: saccades are performed toward moving objects in the periphery of the visual field or toward unexpected and salient visual stimuli. Even if attention can be shifted covertly, by means of eye movements selected stimuli can fall on regions of the retina characterized by a higher acuity. A number of theories have been proposed which link the processes of visual recognition with the sequence of eye movements performed and with shifts of attention [Noton and Stark, 1971; Nakayama, 1991].

It has been often claimed that attentive mechanisms allow to selectively allocate the computational resources of the system. It is evident that, due to the fact that the processing power of current computers is still by far lower than that of the brain, the use of attentive mechanisms is extremely appealing in the development of computer vision systems. Machine vision applications are often hampered by the need of processing huge amounts of data. In the past, this led to the common belief that the major bottlenecks were the computing resources and image acquisition facilities [Jain and Binford, 1991]. Yet, only a small fraction of the raw image data may be relevant to the task at hand. That is, vision systems usually do not need to understand the surrounding scenes, but they only need to extract the information required to accomplish specific tasks [Burt, 1988]. The idea of a system that purposively selects among visual data the significant information and ignores irrelevant details is common to several recently proposed machine vision paradigms (e.g.[Aloimonos, 1991; Ballard, 1991]), and it is crucial when real-time performances are required. A general review of the most important contributions to the implementation of selective attention in machine vision can be found in [Abbott, 1992].

Selective processing in computer vision have been mainly investigated in the context of pyramidal image representations and space-varian sensing. Attentive processes are intrinsically present in the data selection control strategies used with multi–resolution image [Rosenfeld, 1984; Burt and Adelson, 1983]. Differential processing with such hierarchical structures is the result of a coarse-to-fine search through selected paths of the pyramids [Burt, 1988; Culhane and Tsotsos, 1992; Olshausen, 1992]

Also the space-variant structure of the human retina has received the interest of researchers in the field of computer vision, and a number of studies based on its simulations have been carried out [Weiman, 1989; R. Jain and O'Brien, 1987]. Furthermore, a hardware sensor mimicking the geometry of the human eye has been designed and developed, and it has been applied to 2D pattern recognition and motion estimation [Sandini and Tagliasco, 1980; Sandini and Dario, 1989; Tistarelli and Sandini, 1990].

More recently, with the development of fast moving head–eye systems, visual attention has been analyzed in the context of eye movements and selective fixations. Whereas a large number of

works have focused on the control issues involved in visual tracking [Papanikolopoulos *et al.*, 1993; Feddema and Mitchell, 1989], much less efforts have been carried out in the analysis of other kinds of eye movements, mainly due to the underlying theoretical difficulties. In particular, not much research has focused on the implementation of visual saccades. Notable exceptions are the work of Clark and Ferrier, who have implemented the idea of a saliency map in the case of a robotic system [Clark and Ferrier, 1992], and the work of Rimey and Brown, who focused on the application of augmented hidden Markov models and Bayes nets for the production of visual saccades [Rimey and Brown, 1991; Rimey and Brown, 1994].

To the best of our knowledge, almost no research has been carried out toward the implementation of mechanisms of selective visual attention based on non-visual cues.

3 An architecture for implementing attention in machines

One of the major goal of this paper is to propose a general system architecture which is able of integrating attention mechanisms operating in different sensory modalities. Basic design requirements were modularity and generalization possibilities, and the capability of producing real–time perfomances without sophisticated hardware tools.

The global scheme of the architecture is shown in Fig. 2. The system is organized in a sensory-motor 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 the system is currently attending to. After that the shift of gaze has been executed, a new set of interesting directions is generated and is added to the previous one.

A basic assumption in the proposed scheme is that shifts in selective attention are always followed by corresponding changes in gaze directions, that is, the system tends to keep the focus of attention centered in the visual field. This is clearly what a system provided with space–variant sensing capabilities would like to achieve, so that, when a stimulus is selected for fixation, it can be examined with the higher resolution capabilities provided by the fovea. However, the proposed architecture is general, and adapts also to the case of traditional visual sensors with constant spatial resolution. Even if in this case the visual resolution does not change, other advantages are present since new parts of the environment can be brought into the visual field.

In the experiments described in this paper, the application of the system to the case of conventional uniform resolution cameras is considered. A description of a preliminary implementation with space-variant retina-like sensors including only visual processes is given in [Colombo *et al.*, 1994].

The logical center of the architecture is located in a modified version of the saliency map [Koch and Ullman, 1987]. In the proposed version, the saliency map $\mathcal{S}$ can be seen as a matrix whose element s_{ij} represents the saliency of a specific visual direction (ϕ_i, ψ_i) in a head–centered reference frame. All possible visual directions are represented on the saliency map. 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. As will be shown later, it is not required for the system to know exactly which visual direction corresponds to a specific map location. The actual transformation can be learned on the basis of visual feedback so as to compensate for possible misallignaments of the camera and for motor inaccuracies.

Let $D(t) = \{d_1, \ldots, d_M\}$, be the input data to the system at time t. As illustrated in Fig. 2, sensory data are analyzed by a set of time–continuous processes $\{\mathbf{h}^1(t), \ldots, \mathbf{h}^N(t)\}$

$$\mathbf{h}^k(D^k) = (\mathrm{l}^k, \mathbf{U}^k) = \begin{pmatrix} l_1^k, & \mathbf{u}_1^k \\ l_2^k, & \mathbf{u}_2^k \\ \vdots & \vdots \\ l_{N_k}^k, & \mathbf{u}_{N_k}^k \end{pmatrix} \tag{1}$$

$$\mathbf{h}^k : D^k \subset R^k \rightarrow L \times B^k \qquad L = [0,1] \subset R, \;\; B^k \subset S \subset R^2$$

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_i^k for a set of locations $\mathbf{u}_i^k = (x_i, y_i)$ on the saliency map. The locations $\mathbf{u}_i^k$ with a nonzero saliency value l_i^k are the *attentive cues* generated by process $\mathbf{h}^k$.

It should be noted that, given the current posture of the system $P = \{p_1, \ldots, p_L\}$ both the generic process $\mathbf{h}^k$, and its range $L \times B^k$ are dependent on the position of some of the parts, that is $\mathbf{h}^k = \mathbf{h}^k(D^k, P^k), B^k = B^k(P^k)$. 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.

Each attentive process carries out the coordinate transformation necessary for activating a head–centered saliency map starting from data expressed in a sensory reference frame. For example, the visual routines transform retinotopic inputs in visual direction on the basis of the focal length (sensor plane-eye coordinate transformation) and of proprioceptive data.

As illustrated in Fig. 2, all the sensory processes contribute to activate the saliency map, so that the final value assumed by element s_{ij} is given by

$$S_{ij}(t) = F_s(\sum_k w_T^k \sum_q f_{ij}(\mathbf{h}_q^k)) \tag{2}$$

where F_s is a nonlinear monotonic function (in the experiments, a sigmoidal function has been applied) in $R[0,1]$, $\mathbf{h}_i^k$ is the i-th component of process $\mathbf{h}^k$, and

$$f_{ij}(l_k, \mathbf{u}_k) = \begin{cases} l_k & \text{if } \mathbf{u}_k = (i,j) \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

The set of visual processes $\{\mathbf{h}^1, ..., \mathbf{h}^N\}$ produces a set of attentive cues on the saliency map, which are candidates for drawing attention. However other cues can be easily included into the system. For example, semantic cues produced from knowledge-based expectations can also contribute to the activation of the saliency map. Semantic expectations may arise in recognition processes if the system incorporates a fragmentary representation of the objects to recognize [Rucci and Dario,

1993]. By means of such representation, the identification of a feature which characterizes a known object can stimulate the system to look in different directions in order to find other peculiar features, that can contribute to better assess object identity. A number of theories of visual recognition based on the sequence of gaze directions has been proposed [Noton and Stark, 1971; Nakayama, 1991; Yarbus, 1967].

In general, due to the organization of the system, no distinction is formally made between sensory and semantic cues, and both can contribute to attention control. Apart from some speculative considerations [Burt, 1988], researchers have usually focused either on only sensory (*bottom-up*) or semantic (*top-down*) characteristics of the input data, while much less work has been carried out toward their integration into a single architecture [Califano *et al.*, 1989].

As regard the control of attention, the actual direction of gaze can be selected as the location on the map with the maximum value of activation. Several other rules can be implemented, based, for example, on the distance from the current gaze direction [Koch and Ullman, 1987]. The analysis of these control strategies is beyond the scope of this paper. In the experiments illustrated in section 6, the winner rule has been adopted.

The dependence of control of attention on the task at hand is produced by the task weights w_T^k. As shown in eq. 2, the cues of each sensory process $\mathbf{h}^k$ are weighted in the saliency map by a set of time–varying values

$$\mathbf{w}_T(t) = (w_T^1(t), w_T^2(t), \ldots, w_T^N(t))^T \qquad w_T^k \in R[0,1] \tag{4}$$

which are adjusted accordingly to the current task, so that at every time

$$\sum_{k=1}^{N} w_T^k = 1 \tag{5}$$

In this way, on the basis of the task, a priority degree can be assigned to different sensory features. By properly arranging the weight values, it is possible to select a sensory feature with respect to others and/or to inhibit irrelevant cues. As a result, the attended task changes the way the system reacts to sensory stimuli and the way it interacts with the external world.

It is worth noting that, even if biological plausibility has not been a basic requirement in the design of the architecture, it accounts for some psychophysical phenomena, such as the *pop–out effect* [Treisman and Gelade, 1980]. Let suppose that the system includes a separate process for the identification of each significant visual feature (*e.g.* different colors, horizontal bars, vertical bars, etc.). If the selected task is to find a stimulus which differs for a single feature with respect to the others (e.g. a blue bar among red bars), than a task weight vector $\mathbf{w}_T$ where all the components are zero except for the interesting feature, will provide the activation of a single location in the saliency map, independently of the number of distractors in the image. On the contrary, if the target stimulus differs for a combination of features, (that is an horizontal red bar, among horizontal blue bars and vertical red and blue bars), the only possible value for $\mathbf{w}_T$ is to equally set both the weights corresponding to the relevant features. Depending on the saturation value of function

$F_s(x)$ in eq. 2, a set of possible locations is simultaneously active on the saliency map, and the search time will be proportional to the number of stimuli.

4 Autonomous development of visuo–cutaneo coordination

The geometry of somatosensory saccades is depicted in Fig. 3. Let's suppose that a tactile event is monitored by a receptor located on the n-th joint of a robotic manipulator, in position (p, q) in a cutaneous reference frame. That is, the stimulated receptor is the p-th, q-th element of a bidimensional array of tactile sensors spaced by δ_p and δ_q along the two dimensions. The spatial position of the tactile event in a reference frame S_n fixed with joint n-th is given by function $\mathcal{F}$, which is dependent on the structure of the array and on how the array is located on the joint. For the case illustrated in the figure,

$$(x, y, z)^T = \mathcal{F}(p, q) = (p\delta_p, L\cos(\frac{q\delta_q}{L}), L\sin(\frac{q\delta_q}{L}))^T \tag{6}$$

where L is the radius of the joint, which is supposed to be a cylinder. The spatial position of the stimulus with respect to the camera reference frame centered on the perspective focus can then be evaluated as

$$(x_e, y_e, z_e)^T = T_h^e\ T_a^h\ T_{j_1}^{j_0} \ldots T_{j_n}^{j_{n-1}}\ \mathcal{F}(p, q) \tag{7}$$

where

T_h^e = homogeneous matrix transformation from head to eye reference frame
T_a^h = homogeneous matrix transformation from arm to head reference frame
$T_{j_i}^{j_{i-1}}$ = homogeneous matrix transformation from arm joint j to joint $j-1$

Finally, given $(x_e, y_e, z_e)^T$, the angles which define gaze direction are given by

$$\begin{pmatrix} \phi \\ \psi \end{pmatrix} = \begin{pmatrix} \arctan(\frac{x_e}{y_e}) \\ \arctan(\frac{-z_e}{\sqrt{x_e^2+y_e^2}}) \end{pmatrix} \tag{8}$$

as illustrated in Fig. 3.

The direct application of eq. 6- 8 to the execution of somatosensory saccades has the remarkable advantage of using a well defined mathematical formulation which allows a clear analysis of system performances. However, at least two basic limitations are present: first, possible inaccuracies in the model describing the system may produce significant performance errors, and a particular attention is required in the evaluations of the elements of matrices T_h^e, T_a^h, $T_{j_1}^{j_0}, \ldots, T_{j_n}^{j_{n-1}}$ and in the estimate of function $\mathcal{F}$. In addition, it should be considered that even an initially accurate model can shortly produce poor results if the capability of adapting to physical and functional changes of the system is not included. Modifications of model parameters due to changed environmental conditions, aging of the components and damages, can easily occur, and adaptability is crucial for preserving good

performances in time. The capability of recovering from damages and adapting to new functional conditions by updating the robot model is a basic requirement in the development of autonomous robotic systems.

A way to periodically updating the model of the system, is provided by the procedure of calibration [Horn, 1986; Bennet *et al.*, 1991]. In the case of visuomotor coordination, calibration allows the estimation of model parameters by solving a set of nonlinear equations derived by positioning the system in *a priori* known locations. However, the adptability provided by calibration is not real–time since the procedure cannot be executed while the system is operating. In addition, the use of calibration is limited by the fact that is usually time consuming operation and that, due to the mathematical methods used for solving nonlinear equation systems, an initial good estimate of the model is often required. It should also be observed that the direct extension of visuomotor calibration to the sense of touch is not immediate. Tactile information is not passively irradiated by the environment as visual information, and the tactile stimulation of a number of receptors in known positions of the system is difficult to accomplish without an external operator. Thus, the method is not suited for the case of autonomous systems operating in unstructured environments.

The approach followed in this paper is based on the autonomous development of sensorymotor coordinations by means of learning. The system develops its own models by learning all the functional relationships between sensory and motor frames, so as to perform saccades on stimuli detected by different modalities. Due to the adopted learning methods, no need exists for a training phase separate from the real operative one. Learning occurs continuously while system is operating and robot internal models are updated in real–time. As a result, no need exist for external interventions and the system can operate in a completely autonomous manner.

At the beginning the system is provided with basic motor reflexes, which generate specific motor actions when input stimuli are detected. In particular, as will be illustrated in the following section, a shift of gaze direction is produced whenever a visual or a tactile stimulus is monitored. This is accomplished by means of the hard-wired connections existing between locations of the saliency map and positions of the visual system.

Two simultaneous learning processes are included in the system: *before* the execution of a saccade, corrections to the emerging coordinations can be carried out on the basis of incosistencies among the cues produced by separate processes operating on different sensory features of the same physical event. *After* the execution of the motor action, learning can occur by making use of the new sensory detections. Both the processes contribute to the final result, and account for a more robust behavior and to shorter times of adaptation.

In general, the effects of the two learning methods on an attentive process $\mathbf{h}(t)$ can be expressed as

$$\begin{aligned} \frac{\partial \mathbf{h}^k}{\partial t} &= \mathcal{L}_m(\mathbf{h}^k(t), \epsilon_m(D^k, \mathbf{h}^k(t))) \\ \frac{\partial \mathbf{h}^k}{\partial t} &= \mathcal{L}_s(\mathbf{h}^k(t), \epsilon_s(D^1, \ldots, D^N, \mathbf{h}^1(t), \ldots, \mathbf{h}^N(t))) \end{aligned} \tag{9}$$

where ϵ_m and ϵ_s are the errors which allow the development of $\mathbf{h}^k$. In the first case, sensorymotor coordination is modified on the basis of the retinotopic error $\epsilon_m(D^k, \mathbf{h}^k(t))$ recorded after the execu-

tion of a saccade. Sensory feedback-based lerning has been applied to several problems, such as the atuonomous development of invariant visual representations [Kuperstein, 1988a] and cutaneomotor coordination [Rucci and Dario, 1994]. In general, by relating the result of a motor interaction to the modifications in the perceptual scenario cause-effect relationships can be discovered.

In the second learning process, corrections occur on the basis of a comparision ϵ_s among the results of several processes. For example, it can happen that the contact with an external surface is monitored by both the visual and tactile modalities. As a consequence, the cutaneomotor and the visuomotor processes produce corresponding cues, whose accuracies depend on the current stages of development. The distances between these cues are a measure of the consistency of the processes and can be used for refining them. In particular, one of the modality can be assumed as dominant and it can be used to supervise the others. This is similar to what seems to occur in animals, such as the barn owl, which learn to capture the prey in complete obscurity only after they have practiced in illuminated environments [Konishi, 1973].

The association and consistency among separate perceptual frames is a powerful tool for developing a coherent behavior [Reeke *et al.*, 1990] [Kuperstein, 1988b]. Usually, physical events are detectable with more than a single sensory modality. In this way, invariants in the multisensory stimuli patterns can be extracted and models of the world developed. Furthermore, if intrisic sensory modalities, such as proprioception, are considered, it is possible to develop dynamic models of the system functional structure, that is, the system can discover its organization while interacting with the environment. [Kuperstein, 1991; Mel, 1990; Grossberg, 1988].

The system considered in the following of this paper is derived from the general architecture described in section 2. Two attentive processes are included: a visual process $\mathbf{h}^v$ and a cutaneous one $\mathbf{h}^c$. 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 dominant in the consistency-based learning process.

5 Learning visuomotor and somatosensory saccades

The system composed by the visual and cutaneous processes $\mathbf{h}^c$ and $\mathbf{h}^v$ has been implemented by means of neural networks techniques. As illustrated in Fig. 4, the visual and tactile systems have a similar organization. The architecture is composed of several maps whose activation code the current state of the system. Each map is composed of a set of units sensitive to the same kind of inputs, but with varying intensity. Even if all the maps are represented in Fig. 4 as bidimensional, they can have a different number of dimensions depending on the actual system implementation.

The following notation is used: given a map $\mathcal{M}$, the units of the map are indicated as m_{ij}, where the pedices $_{ij}$ locate the unit in the map (a bidimensional map is considered as an example), and the activation of unit m_{ij} is indicated as M_{ij}.

At the input level, the *sensory maps* $\mathcal{C}$ and $\mathcal{R}$ encode 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 position of the components of the visual sub–system (the cameras). 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. That is, the activation of unit m_{ij} in a generic map $\mathcal{M}$ for an input $\mathbf{x} = (x, y)$ is given by

$$M_{ij} = A_m \ \exp - \frac{(\tilde{c}_i - x)^2}{2\sigma_{mx}^2} \frac{(\tilde{c}_j - y)^2}{2\sigma_{my}^2} \tag{10}$$

where the constant $\tilde{\mathbf{c}}_{ij} = (\tilde{c}_i, \tilde{c}_j)$ depends on the adopted mapping function for map $\mathcal{M}$ $\tilde{\mathbf{c}}_{ij} = \mathbf{f}_{\mathcal{M}}(i, j)$. A common mapping function adopted in many of the maps of the experiments is the linear mapping

$$\begin{cases} \tilde{c}_i = i \ \frac{m_{sup}^x - m_{inf}^x}{M_x} + m_{inf}^x \\ \tilde{c}_j = j \ \frac{m_{sup}^y - m_{inf}^y}{M_y} + m_{inf}^y \end{cases} \tag{11}$$

where the constants $m_{sup}^x, m_{inf}^x, m_{sup}^y, m_{inf}^y$ define the range of sensed input data in a map composed of $M_x \times M_y$ units. As illustrated in Fig. 4 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:

$$\begin{aligned} V_{ij} &= d_{ij} R_{ij} \\ V_{ij}^{\phi} &= F_{\tau}(V_{ij}) \ (\textstyle\sum_{pq} w_{pq}^{\phi} M_{pq}^{v} + y_{ij}^{\psi}) \\ V_{ij}^{\psi} &= F_{\tau}(V_{ij}) \ (\textstyle\sum_{pq} w_{pq}^{\psi} M_{pq}^{v} + y_{ij}^{\psi}) \end{aligned} \tag{12}$$

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

$$\begin{cases} a_{pq}^{ij} = 1 & \text{if } \|(v_{ij}^{\phi}, u_{ij}^{\psi}) - (i/N_s^{\psi}, j/N_s^{\psi})\| < \tau_v \\ a_{pq}^{ij} = 0 & \text{otherwise} \end{cases} \tag{13}$$

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.

Learning occurs by properly modifying the weights y_{ij} and w_{ij}. Being vision the dominant sensory modality, only the learning process $\mathcal{L}_m$ of eq. 9 is applied. That is, weights are updated on

the basis of the retinotopic error $\epsilon = (\epsilon_x, \epsilon_y)$ registered after the execution of a visuomotor saccade. System weights are modified as follows:

$$\begin{aligned} y^{\phi}_{ij}(t+1) &= y^{\phi}_{ij}(t) + k^{\phi}_{y}\epsilon_x V_{ij} \\ w^{\phi}_{ij}(t+1) &= w^{\phi}_{ij}(t) + k^{\phi}_{m}\epsilon_x M^{v}_{ij} \end{aligned} \tag{14}$$

$$\begin{aligned} y^{\psi}_{ij}(t+1) &= y^{\psi}_{ij}(t) + k^{\psi}_{y}\epsilon_y V_{ij} \\ w^{\psi}_{ij}(t+1) &= w^{\psi}_{ij}(t) + k^{\psi}_{m}\epsilon_y M^{v}_{ij} \end{aligned} \tag{15}$$

In the visual case 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 nonliner 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. 4 the activation of the units t_{ij}, t^{ϕ}_{ij} and t^{ψ}_{ij} in the three layers $\mathcal{T}$, $\mathcal{T}^{\phi}$ and $\mathcal{T}^{\psi}$, is given by

$$\begin{aligned} T_{ij} &= q_{ij} C_{ij} \\ T^{\phi}_{ij} &= F_{\tau}(T_{ij})\,(\textstyle\sum_{pq} z^{\phi}_{pqij} M^{c}_{pq}) \\ T^{\psi}_{ij} &= F_{\tau}(T_{ij})\,(\textstyle\sum_{pq} z^{\psi}_{pqij} M^{c}_{pq}) \end{aligned} \tag{16}$$

where C_{ij} is the activation of unit i, j in the cutaneotopic sensory map $\mathcal{C}$ and M^{c}_{pq} the activation of unit p, q 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

$$\begin{cases} b^{pq}_{ij} = 1 & \text{if } (\|(t^{\phi}_{ij}, t^{\psi}_{ij}) - (i/N^{\phi}_{s}, j/N^{\psi}_{s})\| < \tau_c \\ b^{pq}_{ij} = 0 & \text{otherwise} \end{cases} \tag{17}$$

In the tactile system adaptation is provided by changes in the weights $z^{\phi}_{pqij}, z^{\psi}_{pqij}$. In this case, both the learning processes of eq.9 contribute to update the connection weights. If the tactile stimulus has a visual counterpart which happens to be in the visual field, then vision acts as a dominant 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 the learning equations. In both the cases weights are updated as

$$\begin{aligned} b^{\phi}_{ijpq}(t+1) &= b^{\phi}_{ijpq}(t) + k^{\phi}_{b}\delta_x M^{c}_{ij} \\ b^{\psi}_{ijpq}(t+1) &= b^{\psi}_{ijpq}(t) + k^{\psi}_{b}\delta_y M^{c}_{ij} \end{aligned} \tag{18}$$

where $\delta = (\delta_x, \delta_y)$ can be the retinotopic or the angular 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 also been found in humans [Groh, 1993].

6 Experimental Results

The system has been tested both with simulations and real robotics experiments. In both the cases the proposed architecture has been implemented according to the considered number of sensory dimensions and degrees of freedom. The experiments performed gave different validations to the model. Simulations, allowed to test how the system can recover from sudden alterations of some functional parameters and how it can adapt to changed conditions. Robotic experiments tested the approach in a real environment in the presence of noise in the sensory data and nonlinearities in the functional characteristics of the components, which make them differ significantly from the theoretical idealizations.

6.1 Simulation experiments

The system considered in the simulations is the planar 1 d.o.f. head-eye, 2 d.o.f. arm illustrated in Fig. 5. As shown in the figure, the eye is centered on the origin of the reference frame which is fixed to the head. In this way, the position of the eye with repect to the head is determined by the angle $\alpha \in (-\pi/2, \pi/2)$ between the gaze direction and the y axis. The position of the arm in the space is specified by the two angles $\theta_1 \in (0, 2\pi/3)$ and $\theta_2 \in (0, 2\pi/3)$, which determine the orientations of the two joints.

In order to develop visuo–cutaneo coordinations in the system, a random initial position of the arm and the head is selected and a stimulus, which can be either in only one or both the sensory modalities, is applied in a random spatial position. Visual stimuli were simulated as bright spots in the visual field, and tactile stimuli were assumed to have always a visual counterpart, as shown in Fig. 5. Only tactile stimuli on the second joint were considered in the experiments.

For a given position $c \in [0 - L_2]$ of the tactile stimulation the position that the visual system should assume in order to foveate on the point is given by:

$$\phi = \arctan\left(\frac{L_1 \cos\theta_1 + c\cos(\theta_1 + \theta_2)}{L_1 \sin\theta_1 + c\sin(\theta_1 + \theta_2)}\right) \tag{19}$$

The adaptation of the general system architecture to the considered robotic system implies that all the maps are monodimensional, excepted $\mathcal{M}^c$ which is bidimensional. The activation of the units in the cutaneotopic map code the position of the tactile stimulus c, whereas the retinotopic map code the position of the visual stimulus v on the monodimensional retina. The proprioceptive maps represent the position α of the eye and the posture (θ_1, θ_2) of the arm. Also the saliency map is monodimensional, and code the saliency of the visual directions ϕ.

In order to simulate the spatio–variant layout of the receptors of the retina, a cubic mapping function has been adopted for the visual sensory map. That is, the activation of unit r_i is given by

$$R_i = \exp(\frac{(v-\tilde{r}_i)^2}{2\sigma_r^2})$$
$$\tilde{r}_i = \frac{4V_f}{N_R^3}(i - \frac{N_R}{2})^3 \tag{20}$$

where N_R is the number of units of the retinotopic map and V_f is the width of the visual field. A linear mapping has been used for the cutaneotopic map $\mathcal{C}$ and for the input motor maps $\mathcal{M}^v$ and $\mathcal{M}^c$. The cutaneous and visual processes $h^v(\alpha, v)$ and $h^c(\theta_1, \theta_2, c)$, are implemented by means of the columnar organization shown in section 5. In this case, only two layers are present since the visual system has a single degree of freedom.

As explained in sections 4 and 5, learning proceeds simultaneously in both the sensory modalities. Whenever a stimulus appears in the visual field, a visuomotor saccade is attempted on the basis of the current visuomotor coordination, and weights are updated by means of the resulting retinal error. If a tactile stimulus occurs in the visual field, visuo-cutaneo coordination is developed by associating the tactile and visual cues on the saliency map, otherwise a somatosensory saccade is performed and the foveation error is used. Task weights in the initial developmental stage were set to $w_T^v = 0.6$ and $w_T^c = 0.4$ (exploration task). At the beginning visual stimuli occured with a higher frequency than tactile ones. The probabilities of occurance of visual and tactile stimuli were respectively

$$p_v = max(0.95 - 0.45\frac{t}{N_{it}}, 0.5)$$
$$p_c = 1 - p_v \tag{21}$$

After the exposure to N_{it} stimuli, the probability of the two events were equal (typically a N_{it} around 2-3000 was used).

Several tests were performed and learning has been studied with several combinations of the parameters. In particular, changes in the coding characteristics of the maps (mapping functions, σ) and in the training parameters (k^ψ, k^ϕ) were analyzed, as well as in the level of noise superimposed to the inputs. The system has proved to be robust, by converging in a broad range of parameters values.

Typical performance values were around 1.5% of the visual field for visuomotor saccades, and 10% of the manipulatory range for somatosensory saccades. Good performances were usually achieved after few thousand iterations.

Fig. 6 illustrates the accuracy of visuomotor and somatosensory saccades at different developmental stages. All the adptive connection weights were initialized to a constant value both in the visual and tactile systems. Fig. 7 shows the final values assumed by the weights of the visual system after that learning has occurred. As it should be expected, the visual and motor weights reflect the functional models of the two maps, and the spatial organization of weights y_k replicates the adopted cubic model (eq. 20).

The capabilities of the system to recover from damages and sudden changes of the functional parameters are illustrated in Fig. 8. The system has proved to be able to self-reorganize so as to compensate to changes both in the sensory and proprioceptive models. The left graph of Fig. 8

shows how visuomotor saccade accuracy is restored when the visual model is transformed from cubic to linear, that is the values $\tilde{r}_i$ in eq. 20 are replaced by

$$\tilde{r}_i = V_f(\frac{i}{N_R} - \frac{1}{2}) \tag{22}$$

This case simulates the replacement of space-variant visual sensors with common raster cameras. In general, in a robotic system significant modifications of the visual model occur all the time that a different camera or lens are used. The right half of the figure, illustrates the accuracy of somatosensory saccades following a change in the proprioceptive perception of the position of the arm. After 10000 iterations, the perceived angle values coded by the cutaneomotor map were set to

$$\begin{aligned} \theta_1' &= \tfrac{3}{2\pi}\theta_1^2 \\ \theta_2' &= \tfrac{3}{2\pi}\theta_2^2 \end{aligned} \tag{23}$$

instead of θ_1, θ_2 as before. This experiment simulates a sudden change of the characteristics of robot encoders. As illustrated by the two curves, after a drastic decrease in performances, in both the cases the system learns the new functional relationships, so that good accuracies are soon reached again.

6.2 Robotic experiments

The proposed approach has been tested in a real, not simulated, environment by means of two robotic manipulators PUMA 500, as illustrated in Fig. 9. One of the two PUMA 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.

The system architecture illustrating the scheme of communication among the components, is shown in Fig. 10. Two VME buses, one for each PUMA, connect the manipulators and the sensors to two Sun SparcStations. Communication between the workstations is also performed through ethernet connection. Both the robots were controlled in real–time by means of the RCCL routines [Lloyd and Heyward, 1993].

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. In general, if the tool is a cylinder of radius D and length L, the location of the stimulus (x_f, y_f, z_f) in a reference system centered on the F/T sensor, satisfies the following set of equations

$$\begin{cases} M_x = y_f F_z - z_f F_y \\ M_z = x_f F_y - y_f F_x \\ x_f^2 + y_f^2 = D^2 \end{cases} \tag{24}$$

where $\mathbf{F} = (F_x, F_y, F_z)^T$ and $\mathbf{M} = (M_x, M_y, M_z)^T$ are the registered force and torque vectors.

In the experiments, a basic assumption was that the stimulation occured on a plane perpendicular to the tool, that is $F_z = 0$. In this way, the location of the contact z_f could be immediately evaluated as

$$z_f = \frac{M_y}{F_x} = -\frac{M_x}{F_y} \tag{25}$$

Without affecting the generality of the approach, the use of a single cutaneous value and the application of a planar force allow to easily evaluate the tactile location and reduce the time required for training the system.

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).

The system was implemented with a bidimensional visual map composed of 100x100 units and a monodimensional tactile input map of 20 units. Both the motor maps were bidimensional: the position of the camera and of the tactile tools were encoded by a 100x100 and a 50x50 map, respectively. The saliency map included 100x100 units. All the intermediate maps in the visual subsystem were composed of 100x100 units, while the maps of the tactile subsystem included 20 units.

Typical values of the parameters used during the experiments were 3% of the sensed fields for the variance of all the maps, 0.2 for all the ks and 0.01 for τ_v andτ_c.

In order to reduce the time required to train the system, learning was performed separately on visuomotor and somatosensory saccades. That is, in an initial phase only visual stimuli were provided, so that first visuomotor coordination was developed. Once a good accuracy in visuomotor saccades was achieved, also tactile stimuli were presented to the system.

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

The execution of a somatosensory saccade is illustrated in Fig. 11. At the beginning, gaze direction is not centered on the tool, but after that a tactile input is registered, the camera moves towards the spatial location of the tactile event.

The effect of learning on the system can be appreciated by analyzing the spatial organizations assumed by the connection weights. Fig. 12 shows sections of weights in the visual system. The contribution produced by a specific location in each input map is highly position dependent. A

spatial organization can also be observed in the weights connecting the motor map $\mathcal{M}^c$ in the tactile subsystem to the units of the cutaneotopic maps $\mathcal{T}^\phi$ and $\mathcal{T}^\psi$, as illustrated in Figgs. 13 and 14, respectively. The development of the spatial organization of weights with learning is illustrated in the case of the tactile subsystem in Figgs. 15 and 16 for the two degrees of freedom.

An example of interaction of different attentive mechanisms is provided by the sequence in Fig. 17. The considered task was a grasping operation, and a fixed task weighting was adopted so that maximum priority ($w_c^T = 0.7$) was given to tactile events, which are potentially the most dangerous, followed by visual events ($w_v^T = 0.3$). This ordering of the events is the opposite of the one adopted in the exploration task. In order to simulate grasping, a basic motor procedure was also included in the system: whenever a visual event was detected, the arm was moved so as to bring the tactile probe over the object. As illustrated in the figure, at the beginning system attention is drawn by the bright spot appearing in the visual field, and a visuomotor saccade toward the spot is performed. A motor command is then sent to the other manipulator, which starts moving toward the object. Due to the priority assigned to tactile events, if a tactile stimulation is detected the motion of the arm is interrupted and a somatosensory saccade is performed. This situation is illustrated in the bottom part of Fig. 17, where the camera is moved so as to fixate on the location of the tactile event. After a delay of time (which could correspond for example to the formulation of a new motor strategy), the location of the tactile event in the saliency map is inhibited and a memory–driven saccade toward the location of previous visual stimulus is performed, while the reaching operation is restarted.

7 Conclusions

One of the major efforts of robotics research is the development of systems, which can autonomously operate in real, unknown and unstructured, environments. According to the goal to accomplish, these systems should be able to perceive the surrounding environment and consequently plan motor interactions, without requiring any external intervention.

An enormous number of applications, which range from the complete substitution of human operators in hazardous environments to human assistence in many fields, require autonomy. Furthermore, the development of behaviors in artificial systems is also of interest for neurosciences, by providing suggestions and, sometimes, validations of ideas. However, a number of basic problems, mainly related to perceptual capabilities contribute to make the development of fully autonomous systems an incredibly hard goal.

Selective attention can play a crucial role in order to overcome classical perceptual difficulties and achieving real–time capabilities. Thanks to attention mechanisms, the system is able to select, among the flow of incoming data, the information which is relevant for the accomplishment of the task at hand, and can discard huge amounts of irrelevant data. In a system where movements of the visual sensors are possible, attentive mechanisms control the direction of gaze. The spatial orientations of the cameras can be changed so as to fixate on the selected stimuli. This is accomplished through a set of sensorymotor transformations which convert the sensory locations of the

stimuli in corresponding positions of the visual sensors.

In order to avoid time degradation of system performances, sensorymotor coordinations must adapt to possible changes of system characteristics. Learning is basic issue in the development of autonomous robotic systems: from one side, learning capabilities can allow the system to adapt to the surrounding environment. For example, by discovering correlations among data in different sensory modalities and cause–effect relationships between motor actions and changes in the perceived scenario, models of the environment and strategies of interaction can be developed. From the other, by linking perceptual and proprioceptive frames, even system specific characteristics can be discovered This allows the system to adapt its internal model in order to compensate to alterations of the functional parameters due to damages, partial failures or aging of the components.

In order to be effective, learning should occur throughout all the operative "life" of the system. Learning algorithms which require the existence of separate learning and operative phases cannot be used with fully autonomous systems. Also supervised learning techniques are not feasible, if they required the intervention of an external operator. However, supervised learning algorithms can still be applied if the supervision is somehow provided by the system itself, for example by using the results of different sensory modalities.

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, as shown in previous sections, learning capabilities can be naturallly included in the system 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. As regards the implementation of attention in artificial systems, it has already been pointed out that very little research has been carried out on non visual attentive processes. This is particularly true for the case of touch, in spite of its importance for interactive operations such as grasping and manipulation. In the architecture described in this paper, visual and non visual processes operate in the same way and no formal distinction is required. Thus the system performs somatosensory and visuomotor saccades without any major difference. The same occurs in principle for sensory and sematic shifts of attention, since the results of all these processes are represented in a common reference frame.

An other major source of interest regards the learning capabilities of the system. Also in this case, research on sensorymotor coordination has mainly focused on vision, whereas other sensory modalities have been much less studied. Furthermore, most of the works described in the literature show only simulations of the proposed approaches and the final validation provided by a real robotic system is seldomly carried out. The described system develops simultaneously visuomotor and cutaneomotor coordinations. However, the approach is general and appliable to any kind of sensorymotor transformation. In addition, results both in the case of simulations and real robotic applications have been described. Due to the coexistence of several learning processes which are active during normal operative phases, the system can quickly adapt to changes in any of the

functional relationships of the parts. It has been shown how it can recover both from alterations in the functional characteristics of the motors/ encoders and of the sensors.

Due to the wide range of issues considered, 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.

8 Acknowledgements

This work has been supported by ARPA Grant N00014-92-J-1647, ARO Grant DAAL03-89-C-0031PRI, 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

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

[Aloimonos *et al.*, 1988] J. Aloimonos, I. Weiss, and A. Bandyopadhyay. Active vision. *International Journal of Computer Vision*, 1(4):333–356, 1988.

[Aloimonos, 1991] J. Aloimonos. Purposive and qualitative active vision. In Y.A. Feldman and A. Bruckstein, editors, *Artificial Intelligence and Computer Vision*. Elsevier, 1991.

[Bajcsy, 1988] R. Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8):996–1005, 1988.

[Ballard, 1991] D.H. Ballard. Animate vision. *Artificial Intelligence*, 48:57–86, 1991.

[Bennet *et al.*, 1991] D.J. Bennet, D. Geiger, and J.M. Hollerbach. Autonomous robot calibration for hand-eye coordination. *International Journal of Robotics Research*, 10(5):550–559, 1991.

[Burt and Adelson, 1983] P.J. Burt and E.H. Adelson. The laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31(4):532–540, 1983.

[Burt, 1988] P.J. Burt. Smart sensing within a pyramid vision machine. *Proceedings of the IEEE*, 76(8):1006–1015, 1988.

[Califano *et al.*, 1989] A. Califano, R. Kjeldsen, and R.M. Bolle. Data and model driven foveation. *IBM Research Report RC*, 15096, 1989.

[Cherry, 1953] E.C. Cherry. Some experiments on the recognition of speech, with one and two ears. *J. Acoustical Society of Anerica*, 25:975–979, 1953.

[Clark and Ferrier, 1992] J.J. Clark and N.J. Ferrier. Attentive visual servoing. In A. Blake and A.Yuille, editors, *Active Vision*. MIT Press, Cambridge, MA, 1992.

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

[Culhane and Tsotsos, 1992] S.M. Culhane and J.K. Tsotsos. An attentional prototype for early vision. In *Proceedings of the 2nd European Conference on Computer Vision*, pages 551–560, S. Margherita Ligure, Italy, 1992.

[Eriksen and James, 1986] C.W. Eriksen and J.D. St. James. Visual attention within and around the field of focal attention: A zoom–lens model. *Perception and Psychophysics*, 40(4):225–240, 1986.

[Feddema and Mitchell, 1989] J.T. Feddema and O.R. Mitchell. Vision–guided servoing with feature–based trajectory generation. *IEEE Trans. on Robotics and Automation*, 5(5):691–700, 1989.

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

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

[Horn, 1986] B.K.P. Horn. *Robot Vision*. MIT Press, Cambridge, MA, 1986.

[Jain and Binford, 1991] R.C. Jain and T.O. Binford. Ignorance, myopia, and naiveté in computer vision systems. *Computer Vision, Graphics and Image Processing: Image Understanding*, 53(1):112–117, 1991.

[James, 1890] W. James. *The Principles of Psychology*. Harvard University Press, Cambridge, MA (1983), 1890.

[Johnston and Dark, 1986] W.A. Johnston and V.J. Dark. Selective attention. *Annual Review of Psychology*, 37:43–75, 1986.

[Kinchla, 1992] R.A. Kinchla. Attention. *Annual Review of Psychology*, 43:711–742, 1992.

[Koch and Ullman, 1987] C. Koch and S. Ullman. Shifts in selective visual attention: Toward the underlying neural circuitry. In L.M. Vaina, editor, *Matters of Intelligence*. D. Reidel Pub. Comp., 1987.

[Konishi, 1973] M. Konishi. How the owl tracks its prey. *American Scientist*, 61:414–424, 1973.

[Konishi, 1993] M. Konishi. Listening with two ears. *Scientific American*, pages 34–41, 1993.

[Krotkov, 1989] E.P. Krotkov. *Active Computer Vision by Cooperative Focus and Stereo.* Springer Verlag, Berlin, Germany, 1989.

[Kuperstein, 1988a] M. Kuperstein. An adaptive neural model for mapping invariant target position. *Behavioral Neuroscience*, 102(1):148–162, 1988.

[Kuperstein, 1988b] M. Kuperstein. Neural network model for adaptive hand–eye coordination for single postures. *Science*, 239:1308–13011, 1988.

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

[Lloyd and Heyward, 1993] J. Lloyd and V. Heyward. Real–time trajectory generation in multi-rccl. *Journal of Robotic Systems*, 10(3):369–390, 1993.

[Marr, 1982] D. Marr. *Vision.* W.H. Freeman and Co., San Francisco, CA, 1982.

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

[Nakayama, 1991] K. Nakayama. The iconic bottleneck and the tenuous link between early visual processing and perception. In C. Blakemore, editor, *Vision: Coding and Efficiency.* University Press, 1991.

[Noton and Stark, 1971] D. Noton and L. Stark. Eye movements and visual perception. *Scientific American*, 224(6):34–43, 1971.

[Olshausen, 1992] B. Olshausen. A neural model of visual attention and invariant pattern recognition. *Tech. Rep. CalTech, CNS Memo 18*, September 1992.

[Pahlavan *et al.*, 1992] K. Pahlavan, T. Uhlin, and J.O. Eklundh. Integrating primary ocular processes. In *Proceedings of the 2nd European Conference on Computer Vision*, pages 526–541, S. Margherita Ligure, Italy, 1992.

[Papanikolopoulos *et al.*, 1993] N. Papanikolopoulos, P.K. Kosla, and T. Kanade. Visual tracking of a moving target by a camera mounted on a robot: A combination of control and vision. *IEEE Trans. on Robotics and Automation*, 9(1):14–35, 1993.

[Posner, 1980] M. Posner. Orienting of attention. *Quarterly Journal of Experimental Psychology*, 32:3–25, 1980.

[R. Jain and O'Brien, 1987] S.L. Bartlett R. Jain and N. O'Brien. Motion stereo using ego-motion complex logarithmic mapping. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 9(3):356–365, 1987.

[Reeke *et al.*, 1990] G.N Reeke, O. Sporns, and G.M. Edelman. Synthetic neural modeling: The "darwin" series of recognition automata. *Proceedings of the IEEE*, 78(9):1498–1530, 1990.

[Rimey and Brown, 1991] R.D. Rimey and C.M. Brown. Controlling eye movements with hidden markov models. *Int. J. Computer Vision*, 7(1):47–65, 1991.

[Rimey and Brown, 1994] R.D. Rimey and C.M. Brown. Control of selective perception using bayes nets and decision theory. *Int. J. Computer Vision*, 12(2):173–207, 1994.

[Rosenfeld, 1984] A. Rosenfeld. *Multiresolution Image Processing and Analysis.* Springer Verlag, Berlin, Germany, 1984.

[Rucci and Dario, 1993] M. Rucci and P. Dario. Selective attention mechanisms in a vision system based on neural networks. In *Proc. IEEE/RSJ Int. Conf.on Intelligent Robots and Systems*, Yokohama, Japan, 1993.

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

[Sandini and Dario, 1989] G. Sandini and P. Dario. Active vision based on space–variant sensing. In *5-th International Symposium on Robotics Research*, pages 408–417, Tokyo, Japan, 1989.

[Sandini and Tagliasco, 1980] G. Sandini and V. Tagliasco. An anthropomorphic retina-like structure for scene analysis. *Computer Graphic and Image Processing*, 14(3):365–372, 1980.

[Tistarelli and Sandini, 1990] M. Tistarelli and G. Sandini. Estimation of depth from motion using an anthropomorphic visual sensor. *Image and Vision Computing*, 8(4):271–278, 1990.

[Treisman and Gelade, 1980] A. Treisman and G. Gelade. A feature–integration theory of attention. *Cognitive Psychology*, 12:97–136, 1980.

[Tsal, 1983] Y. Tsal. Movements of attention across the visual field. *J. Exp. Psychol.: Hum. Percept. Perform.*, 9:523–530, 1983.

[von Helmholtz, 1866] H. von Helmholtz. *Psychological Optics.* J.P.C. Sothall, Dover, NY (1925), 1866.

[Weiman, 1989] C.F.R. Weiman. Tracking algorithms using log-polar mapped image coordinates. In *Proc. SPIE Intelligent Robots and Computer Vision VIII: Algorithms and Techniques*, pages 843–853, Philadelphia, PA, 1989.

[Yarbus, 1967] A.L. Yarbus. *Eye movements and vision.* Plenum Press, 1967.

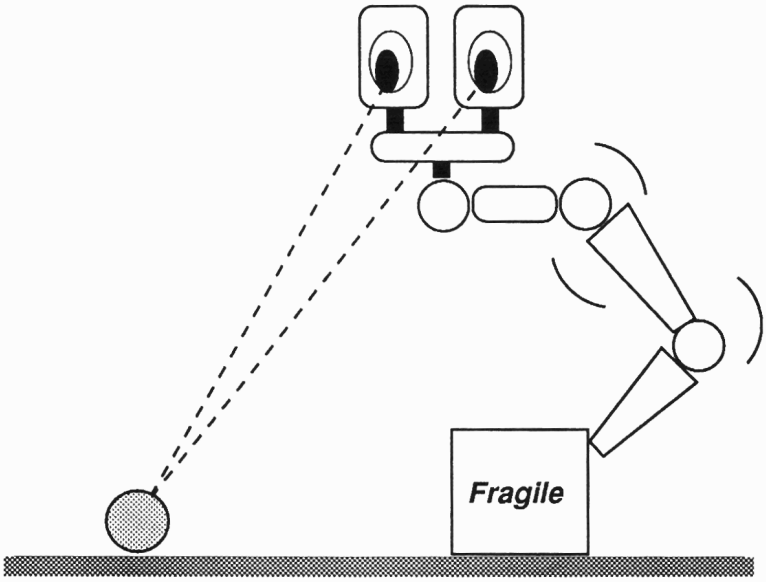


Figure 1: Attention can be drawn by stimuli belonging to different sensory modalities. If an object is hit while the system is trying to reach the ball, visual attention should be directed toward the location of the tactile event, in order to find a collision free path.

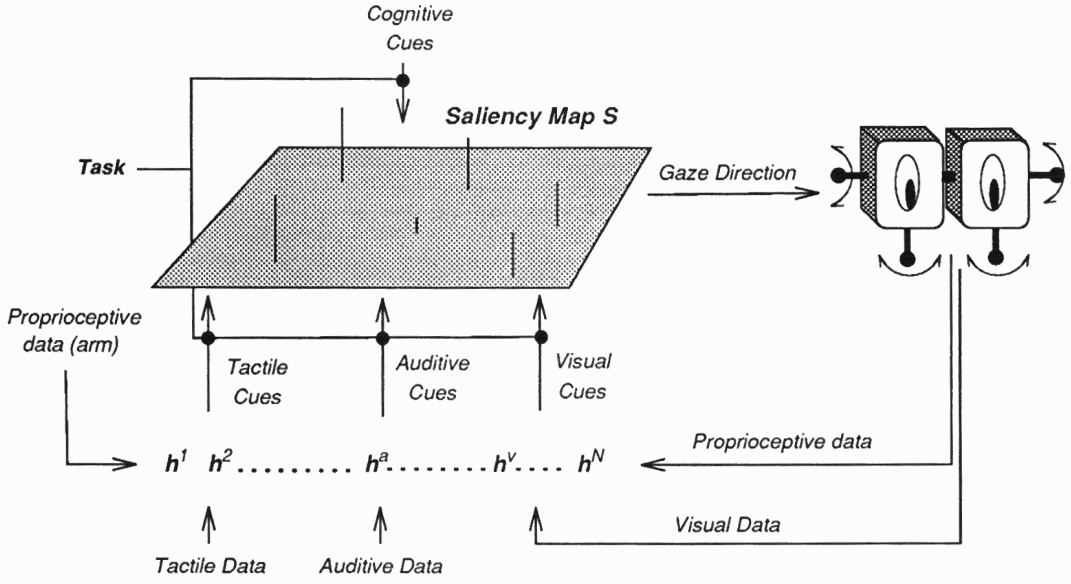


Figure 2: The proposed architecture. Data belonging to different sensory modalities are separately processed by dedicated modules so as to activate corresponding locations of a common head-centered saliency map. The location with maximum value of activation indicates next gaze direction. Cues are differently weighted in dependence on the current task.

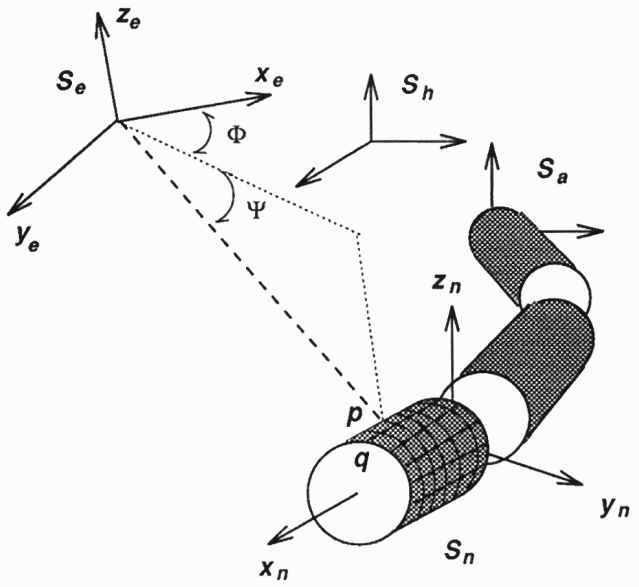


Figure 3: The geometry of somatosensory saccades. The location of a tactile receptor in a cutaneo–centered reference frame should be converted in the corresponding pan and tilt of the cameras.

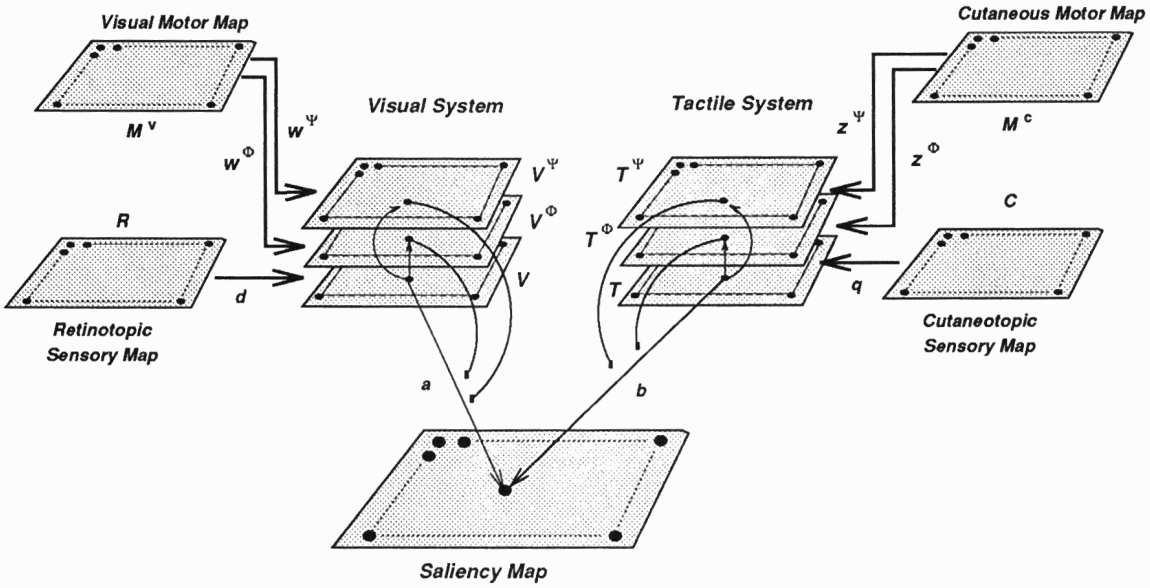


Figure 4: In both the sensory modalities, the activation of the motor–proprioceptive maps and the sensory maps are combined in a sensorytopic columnar organization which produces the corresponding cues for the saliency map.

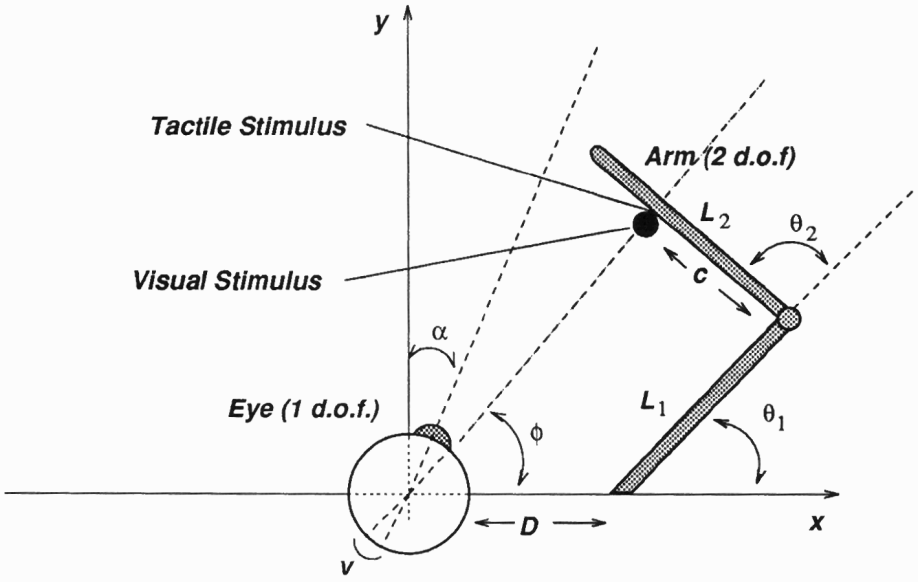


Figure 5: The system used in the simulations. The position of a 1 d.o.f. eye with respect to a head–centered reference system is given by the angle α, and the position of the 2 d.o.f. arm is specified by (θ_1, θ_2)

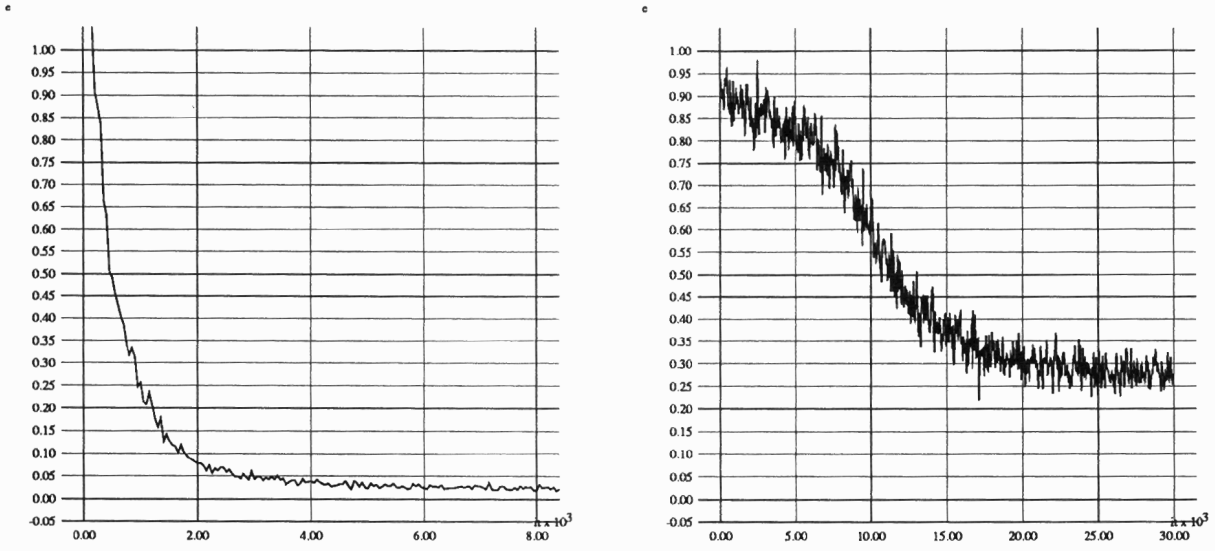


Figure 6: Performances of the system with learning. *(Left)* Visual performances: avarage foveation error when visual stimuli are applied. *(Right)* Tactile performances: avarage foveation error when cutaneous stimuli are applied. Both the avarage errors are evaluated on a fixed number of tests. The system was implemtented with 150 units in all the visual maps and 20 units in the tactile ones. All weights were updated with $k = 0.05$

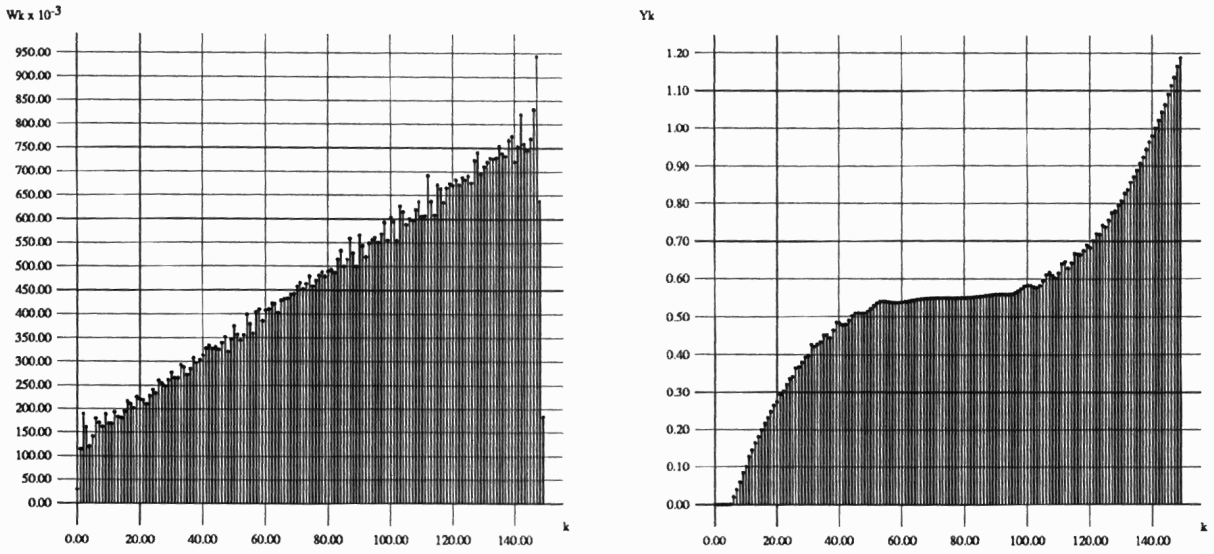


Figure 7: Weights of the system after several iterations of learning. *(Left)* Weights w_k.*(Right)* Weights y_k. In each graph the length of segment k is proportional to the $k-th$ weight.

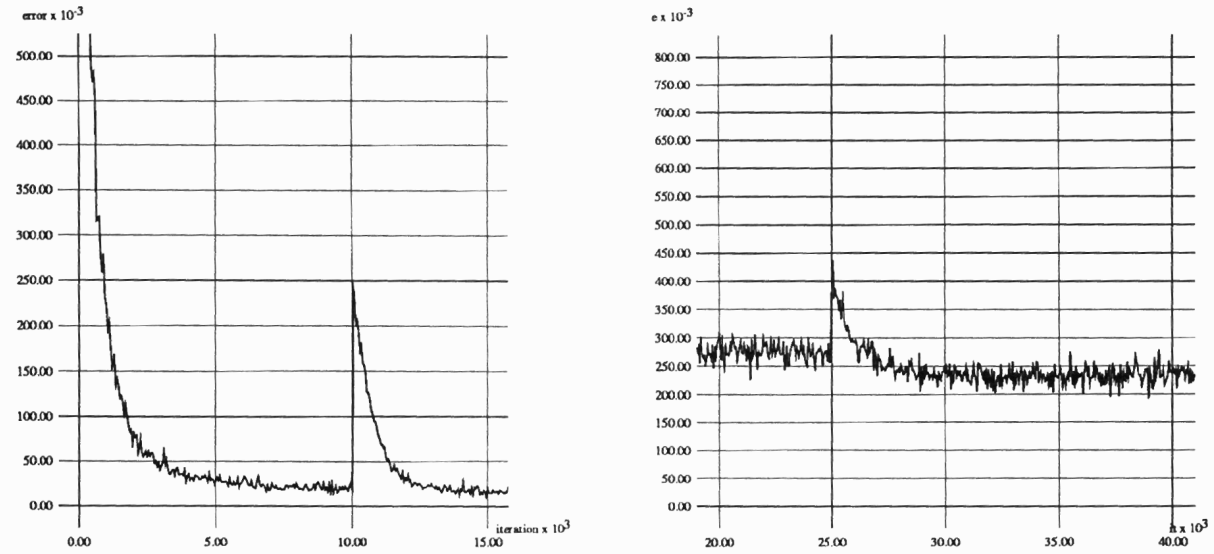


Figure 8: Recovering from alterations of the model (see text for details). *(Left)* Visual saccades accuracy when the visual model changes from cubic to linear *(Right)* Somatosensory saccades accuracy when the arm joint proprioception is changed.

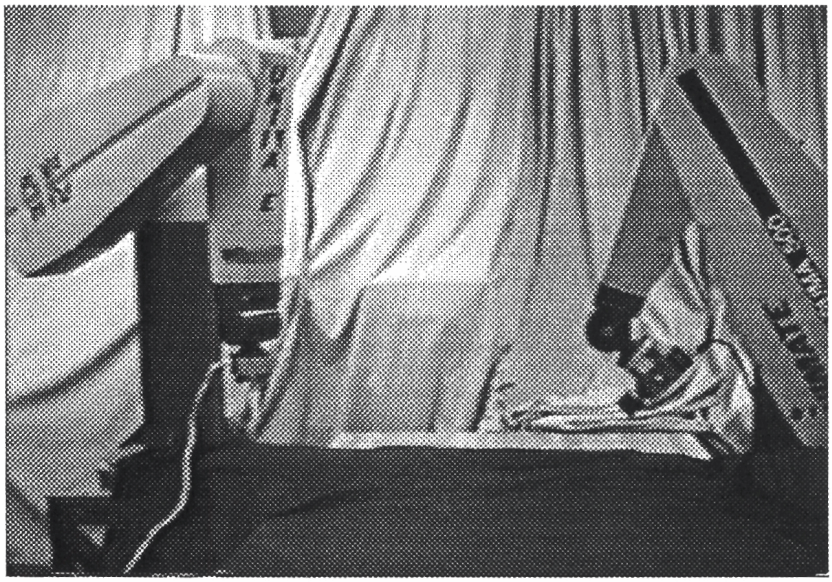


Figure 9: The experimental scenario

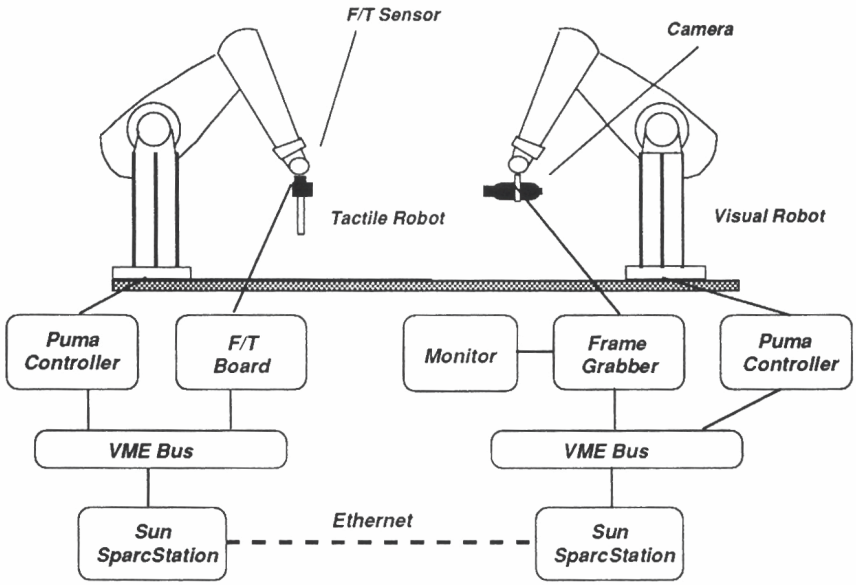


Figure 10: Scheme of the robotic system used for the experiments. Two Puma manipulators are used. Control is implemented on two workstations connected via two VME buses to the robots and the sensors.

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

Foveation Error		
it	*mean*	σ^2
100	0.20	1.66
300	0.10	0.59
600	0.07	0.14
1000	0.04	0.07

Table 2: Accuracy of somatosensory saccades at different learning levels

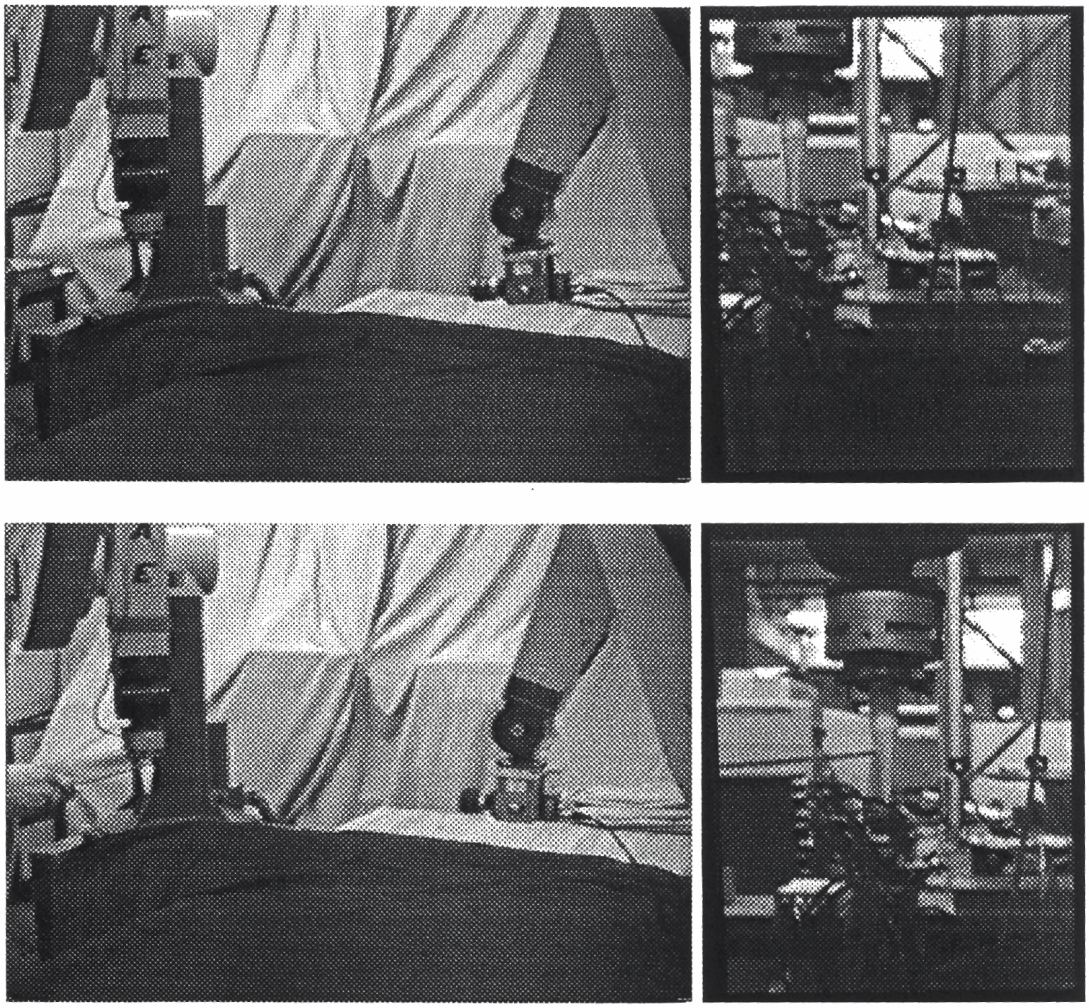


Figure 11: Execution of a somatosensory saccade. The images on the left are camera views of the scene. *Top* Before that foveation occurs, *Bottom* after the saccade.

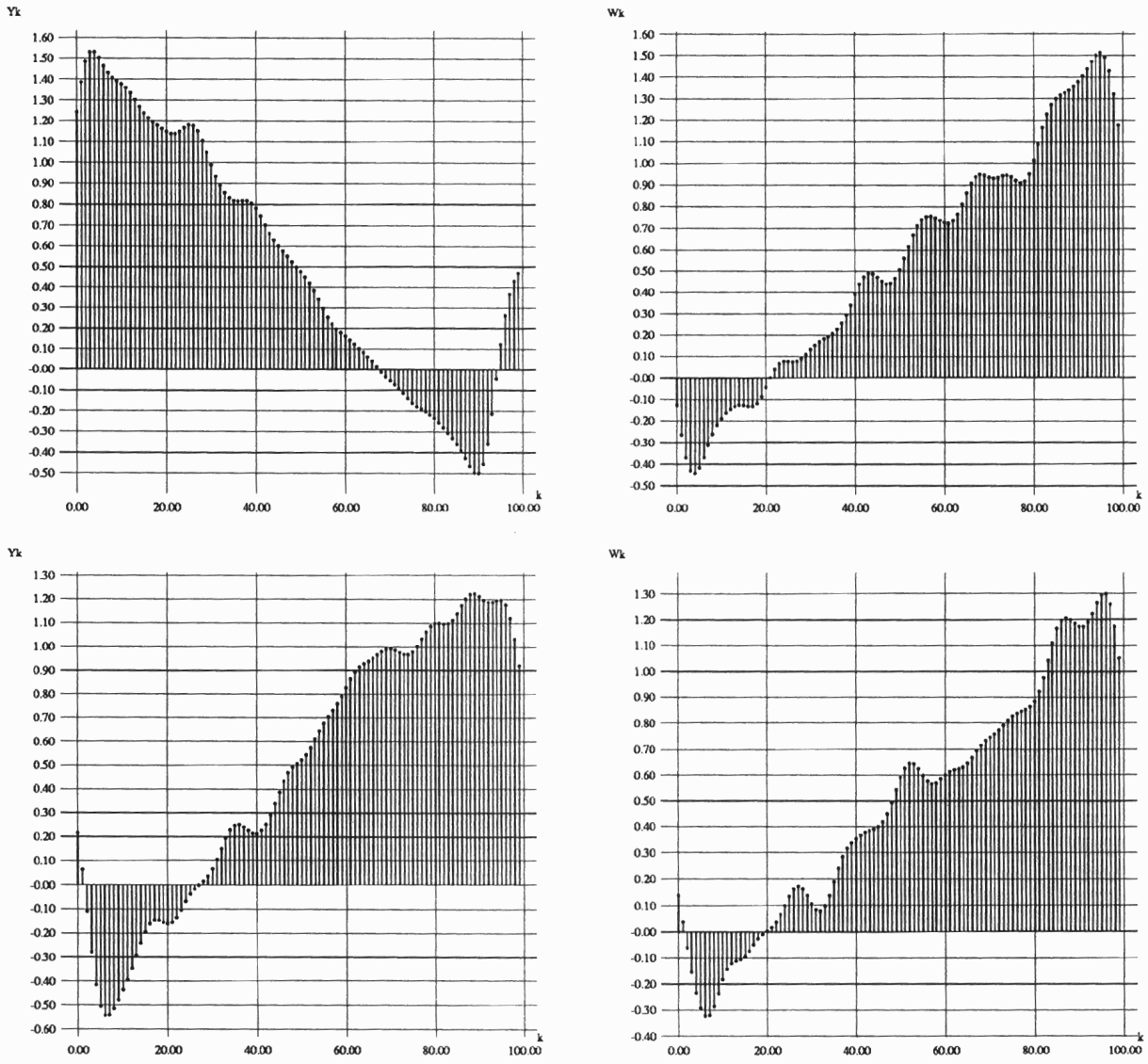


Figure 12: Some of the weights of the visual system after several iterations of learning. The sections correspond to the center units of $\mathcal{R}$ and $\mathcal{M}^v$ maps (weights y_{k50}, w_{k50}, $k = 0 - 99$). *(Top)* Weights from the visual (left) and the motor (right) input maps for the d.o.f. ϕ. *(Bottom)* Weights from the visual (left) and the motor (right) input maps for the d.o.f. ψ. In each graph the length of segment k is proportional to the k-th weight.

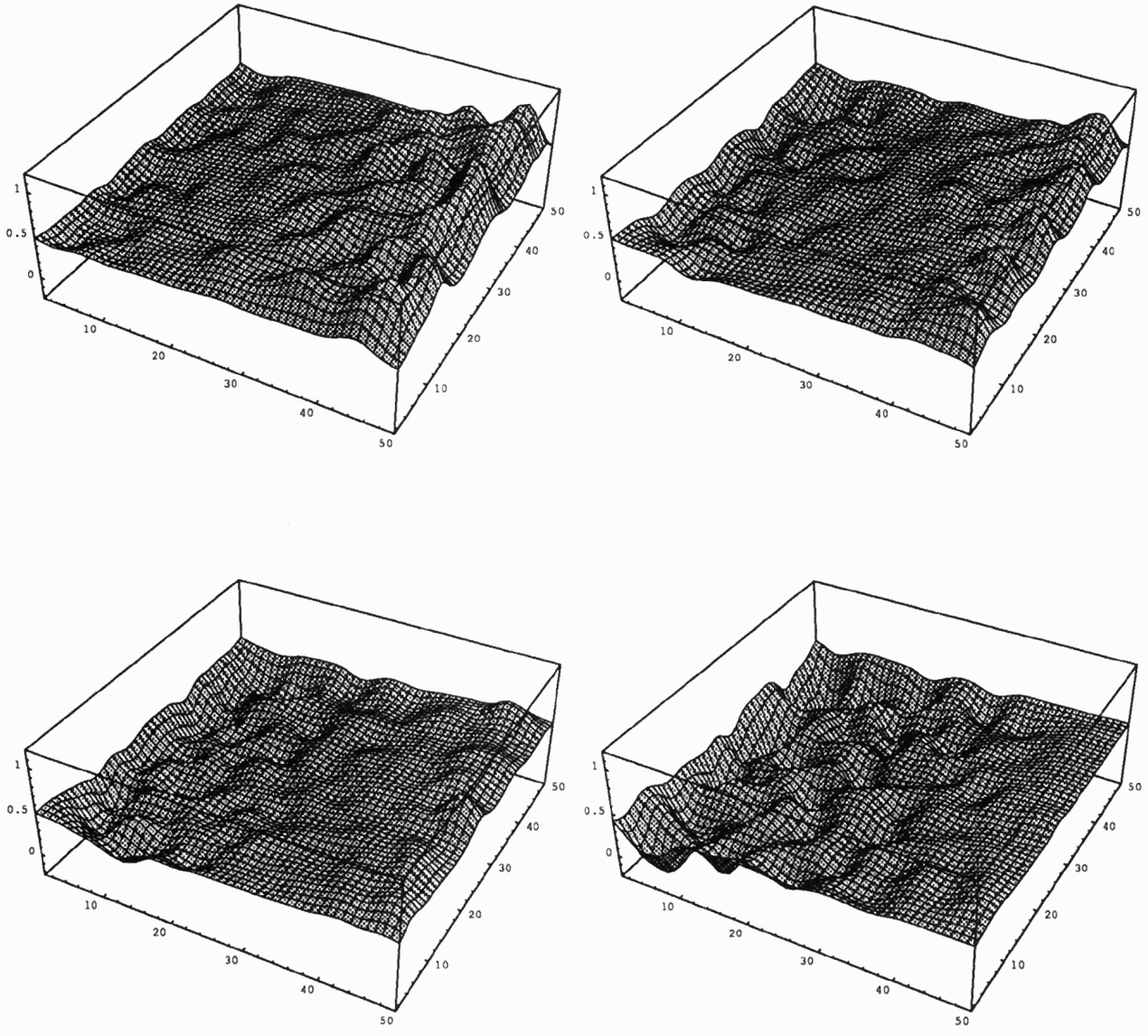


Figure 13: System weights values after the presentation of 1000 tactile stimuli for the d.o.f. ϕ. In each graph, the value in position $< i, j >$ is proportional to weight z^{ϕ}_{ijk} connecting units m^{c}_{ij} and t_k. The connections with different t_k are shown. *(Top left)* cutaneous level $k = 3$, *(Top right)* cutaneous level $k = 8$, *(Bottom left)* cutaneous level $k = 12$, *(Bottom right)* cutaneous level $k = 17$.

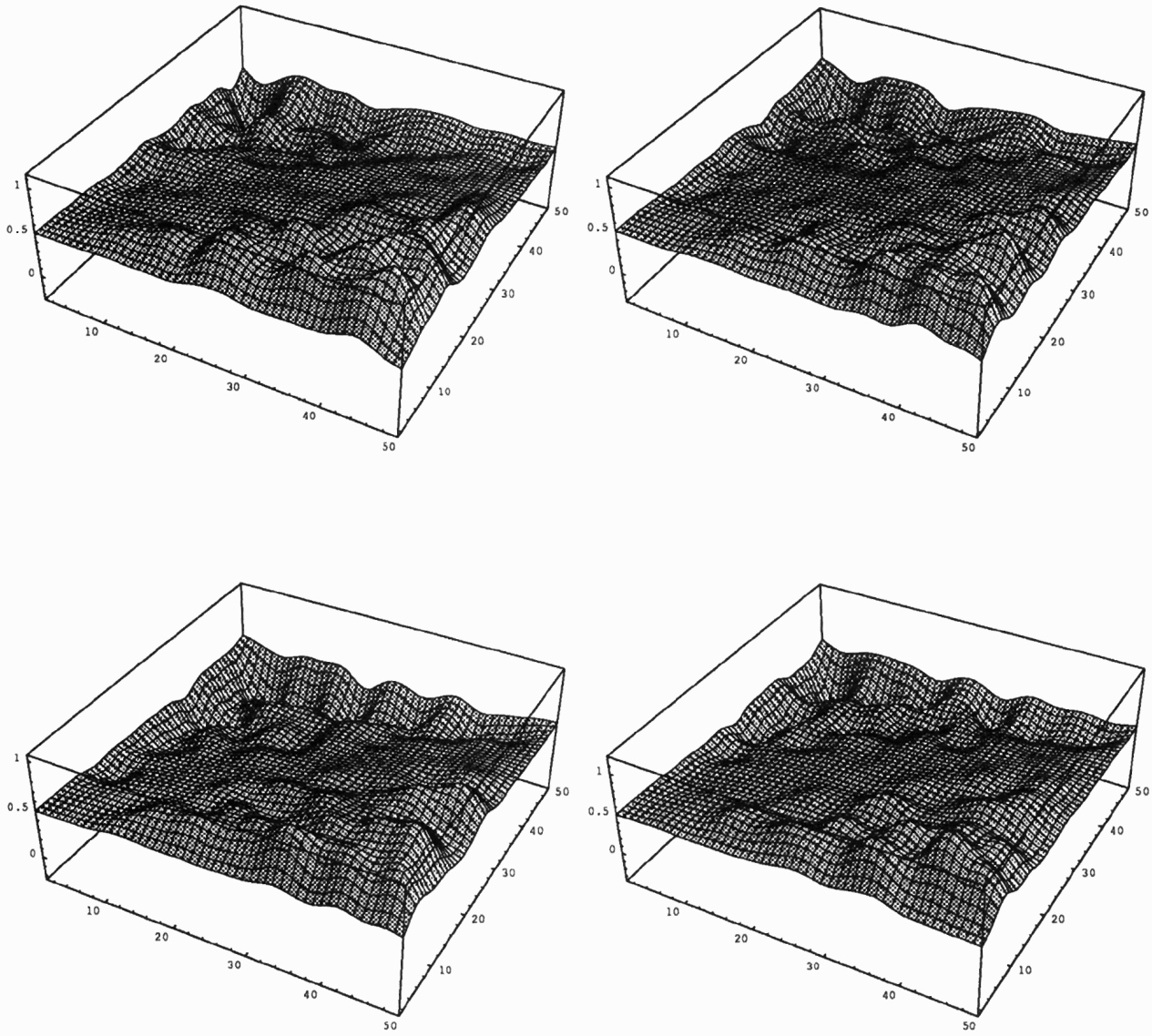


Figure 14: System weights values after the presentation of 1000 tactile stimuli for the d.o.f. ψ. In each graph, the value in position $< i, j >$ is proportional to weight z^{ψ}_{ijk} connecting units m^{c}_{ij} and t_k. The connections with different t_k are shown. *(Top left)* cutaneous level $k = 3$, *(Top right)* cutaneous level $k = 8$, *(Bottom left)* cutaneous level $k = 12$, *(Bottom right)* cutaneous level $k = 17$.

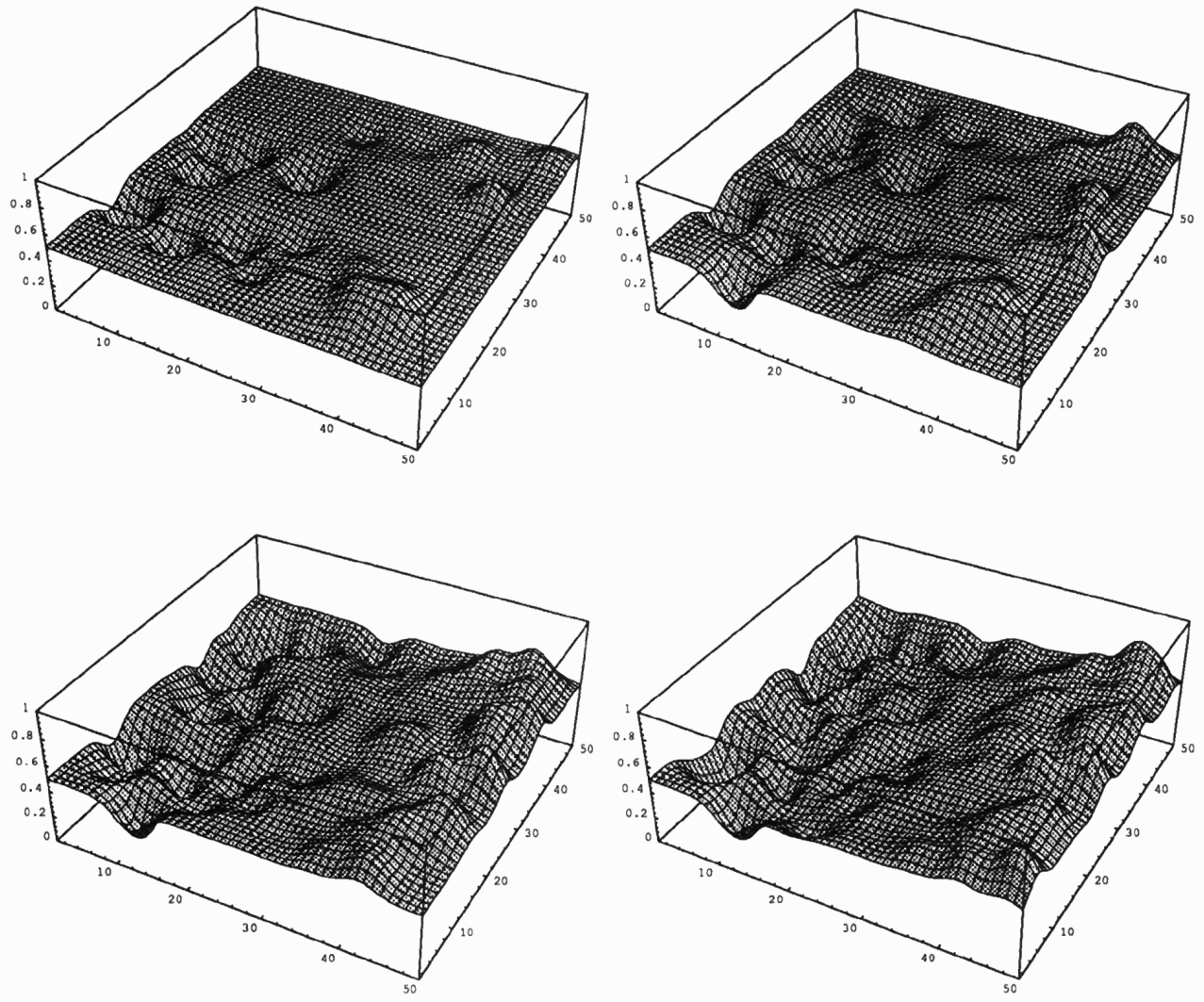


Figure 15: System weights z^{ϕ}_{ij10} at different stages of development. *(Top left)* after the presentation of 50, *(top right)* 100, *(bottom left)* 300, and *(bottom right)* 600 tactile events.

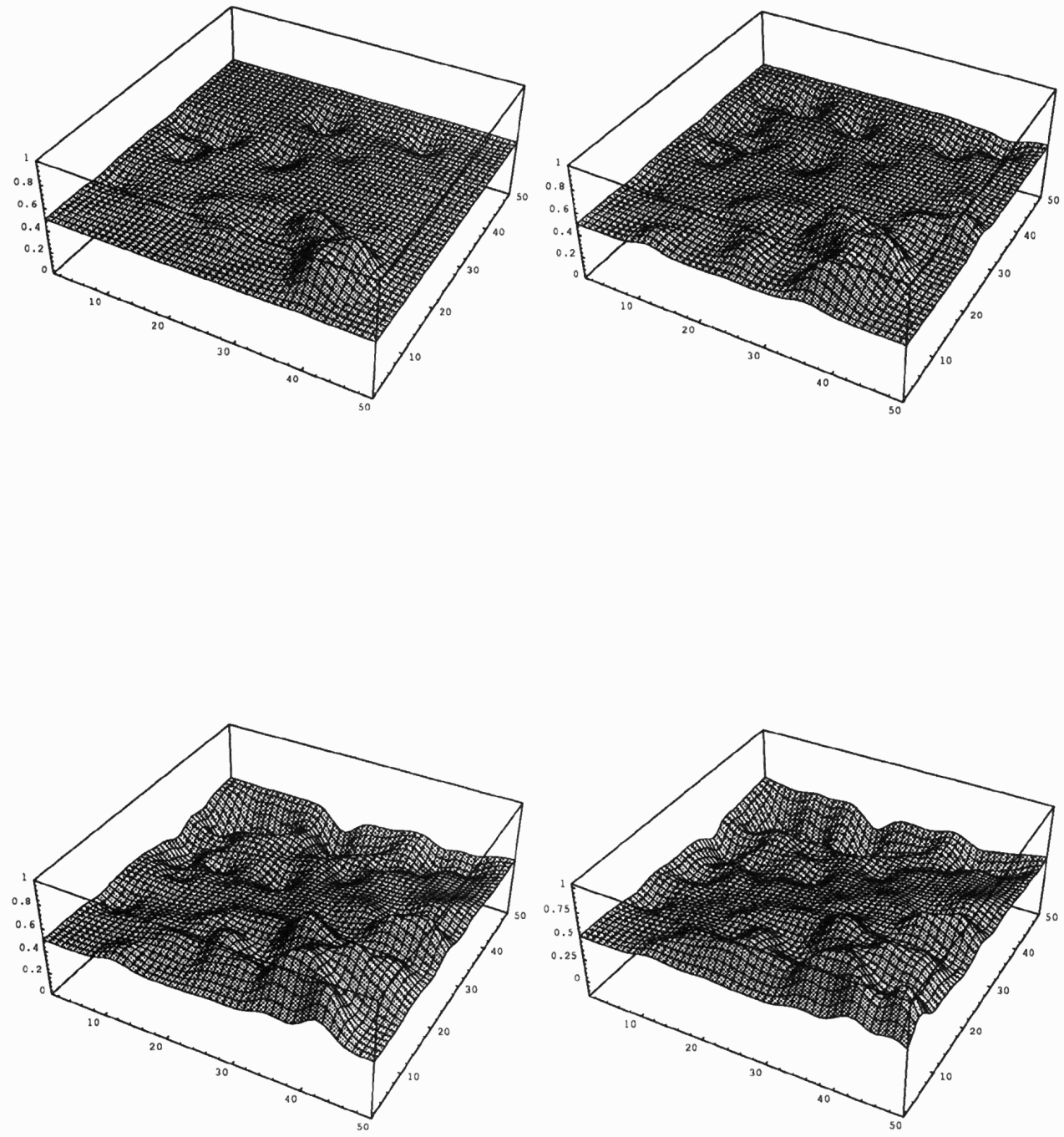


Figure 16: System weights z^{ψ}_{ij10} at different stages of development. *(Top left)* after the presentation of 50, *(top right)* 100, *(bottom left)* 300, and *(bottom right)* 600 tactile events.

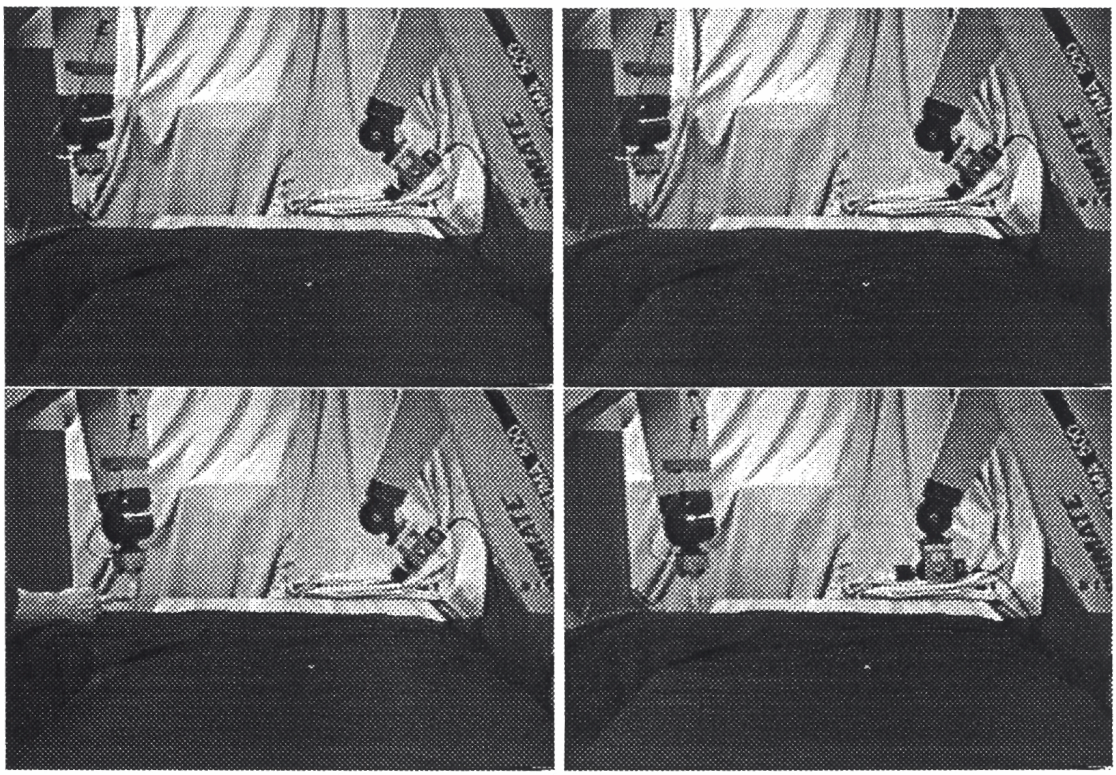


Figure 17: Interaction of visual and tactile attentive mechanisms in a grasping task. System attention switches from a visual to a tactile cue accordingly to the task weights (see text for details).