Implicit Sensorimotor Mapping of the Peripersonal Space by Gazing and Reaching

Eris Chinellato, *Member, IEEE*, Marco Antonelli, Beata J. Grzyb, and Angel P. del Pobil

***Abstract*—Primates often perform coordinated eye and arm movements, contextually fixating and reaching towards nearby objects. This combination of looking and reaching to the same target is used by infants to establish an implicit visuomotor representation of the peripersonal space, useful for both oculomotor and arm motor control. In this work, taking inspiration from such behavior and from primate visuomotor mechanisms, a shared sensorimotor map of the environment, built on a radial basis function framework, is configured and trained by the coordinated control of eye and arm movements. Computational results confirm that the approach seems especially suitable for the problem at hand, and for its implementation on a real humanoid robot. By exploratory gazing and reaching actions, either free or goal-based, the artificial agent learns to perform direct and inverse transformations between stereo vision, oculomotor, and joint-space representations. The integrated sensorimotor map that allows to contextually represent the peripersonal space through different vision and motor parameters is never made explicit, but rather emerges thanks to the interaction of the agent with the environment.**

***Index Terms*— Eye–arm coordination, humanoid robots, radial basis function networks, self-supervised learning, spatial awareness.**

I. Introduction

HUMANS and other primates build their perception of the surrounding space by actively interacting with nearby stimuli, mainly looking and reaching at them. Through active exploration, they construct a representation of the environment useful for further interactions. The main sensory information used to build such representation is retinotopic (visual data) and proprioceptive (eye, neck, and arm position). A critical issue in this process is to coordinate movements and associate sensory inputs, in order to obtain a coherent mental image of the environment. Indeed, eye and arm movements often go together, as we fixate an object before, or while, we reach it. Such combination of looking and reaching towards the same target is used to establish a consistent, integrated visuomotor representation of the peripersonal space.

In primates, areas within the dorsal visual stream of the primate brain, and more precisely, regions of the posterior parietal cortex (PPC), are in charge of performing the reference frame transformations required to map visual information to appropriate oculomotor and limb movements. These areas are the best candidates for the role of accessing and updating a visuomotor representation of the reachable space. It is often argued, and increasingly accepted by the neuroscientific community, that such ability is achieved through the use of gain fields and basis function representations, that permit to simultaneously represent stimuli in various reference frames. In fact, the basis function approach has the attractive feature that both head-centric representations for arm movements and retino-centric representations for gaze movements can be encoded concurrently in the same neural map [1]. In the basis function framework, explicit encoding of targets in retino-centric coordinates is enhanced via gain fields to hold in parallel an implicit encoding in other reference frames [2]. Such gain fields are found in retino-centric organized eye movement areas lateral intraparietal sulcus (LIP) [3], [4] and frontal eye field (FEF) [5] and, most importantly, in posterior parietal area V6A, as explain in Section II.

In this work, eyes and arms of a humanoid robot are treated as separate effectors that receive motor control via different movement vectors associated to specific spatial representations, i.e., oculomotor and arm joint space. These representations are combined to form a unique, shared visuomotor map of the peripersonal space. The exploratory behavior of the robot is based on a functional model of the tasks performed by the primate posterior parietal cortex. The main building block of such a model is a basis function framework that associate different reference frames, giving them mutual access to each other, during planning, execution and monitoring of eye and arm movements. We implement such framework by relying upon findings from human and monkey studies, especially from data on gaze direction and arm reaching movements in monkey area V6A [6].

Our system should finally be able to achieve a visuomotor knowledge of its peripersonal space in a dynamical way, through the practical interaction with the environment, using both stereoptic visual input and proprioceptive data concerning eye and arm movements. Following this approach, the robot should naturally achieve very good open-loop reaching and saccade capabilities towards nearby targets. This goal is represented in Fig. 1,

Manuscript received February 15, 2010; revised June 06, 2010; accepted December 23, 2010. Date of publication January 28, 2011; date of current version March 16, 2011. This work was supported in part by the European Commission's Seventh Framework Programme FP7/2007-2013, under Grant 217077 (EYESHOTS project), by the Ministerio de Ciencia y Innovación (DPI-2008-06636, FPU Grant AP2007-02565, and FPI Grant BES-2009-027151), by the Fundació Caixa-Castello-Bancaixa (P1-1B2008-51), and by the World Class University Program through the National Research Foundation of Korea, funded by the Ministry of Education, Science, and Technology (Grant R31-2008-000-10062-0).

E. Chinellato, M. Antonelli, and B. J. Grzyb are with the Robotic Intelligence Laboratory, Jaume I University, Castellón de la Plana 12071, Spain (e-mail: eris@uji.es; antonell@uji.es; grzyb@uji.es).

A. P. del Pobil is with the Robotic Intelligence Laboratory, Jaume I University, Castellón de la Plana, 12071, Spain, and he is also with the Department of Interaction Science, Sungkyunkwan University, Seoul, South Korea (e-mail: pobil@uji.es).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TAMD.2011.2106781

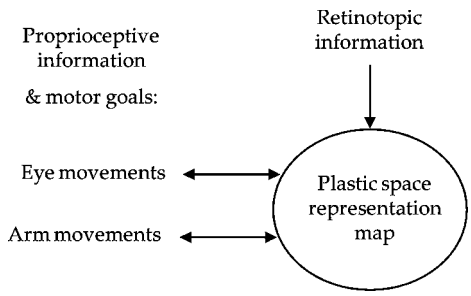

Fig. 1. Conceptual schema of the space representation map, with afferent and efferent signals.

which depicts a simple conceptual schema of how the space representation is generated and updated. The peripersonal space is represented by a plastic map, that constitutes both knowledge of the environment and a sensorimotor code for performing movements and evaluate their outcome. The map is accessed and modified by two types of information: retinotopic (visual) and proprioceptive (eye and arm movements). Contextually, eye and arm motor plans are devised in accordance to the map itself. This mechanism allows us to keep both eye and arm targeting in register and to establish a common spatial representation of the target's location. The proposed framework is therefore based on a common code for spatial awareness obtained by a learning procedure based on target errors of eye or arm movements to visual targets, as described in Section III. As a final goal, the agent should be able to purposefully build such visuomotor map of the environment in 3-D, simultaneously learning to look at and reach towards different visual targets.

The perception of the space and the related sensorimotor map is thus accessed and updated by visuomotor interaction, e.g., moving the gaze and the arm toward a goal position. More interestingly, the agent has to be able to keep learning during its normal behavior, by interacting with the world and contextually update its representation of the world itself. Such goal can be achieved by adjusting the weights of the basis functions according to the errors observed in the oculomotor and reaching movements as guided by the same basis functions. The concurrence of eye and arm movements should be sufficient to provide the appropriate error signals even when actual distances between target and final position of the action are not explicitly provided. We could call this approach as a "self-supervised learning" framework, in which the different modalities supervise each other, and eye and arm movements both improve, and obtain together a precise visuomotor representation of the surrounding space. Upon tactile feedback, confirming that the target object has been reached, the actual gazing and reaching directions can be compared with the expected ones and the 3-D visuomotor representation accordingly updated if necessary. The presence of a tactile response, as feedback for the correct execution of a reaching movement, provides a reliable "master" signal to ensure the accuracy of the global representation. The results presented in Section IV refers to the computational implementation of the model, and we are currently working on its development on our humanoid robot setup (see Section V).

II. Background

This research builds on neuroscience findings and insights, computational intelligence concepts and techniques, and engineering goals and constraints posed by the robotic implementation. In this section we introduce the fundamental inspiring concepts and relevant literature for each of these fields.

A. Sensorimotor Transformations in the Posterior Parietal Cortex

The visual cortex of the primate brain is organized in two parallel channels, called "dorsal" and "ventral" streams. The former elaborates visual data with the main purpose of endowing the subject with the ability of interacting with his/her environment, and its tasks are often synthesized as "vision for action." The latter is dedicated to object recognition and conceptual processing, and thus performs "vision for perception." Although the interaction between the two streams is necessary for most everyday tasks, dorsal stream areas are more strictly related to the planning and monitoring of reaching and grasping actions [7]. In fact, dorsal visual analysis is driven by the absolute dimension and location of target objects, requiring continuous transformations from retinal data to effector-based frames of reference. The frame of reference used for arm movements is body-centered, e.g., fixed on the shoulder. Considering a dexterous primate, or a humanoid robot endowed with a pan-tilt-vergence head, head movements are also body-centered, whilst gaze movements are head-centered, usually referred to the cyclopean eye, ideal midpoint between the eyes. Visual information is retinocentric, and there are two different retinal reference frames for a stereo system. All these different reference frames are brought in register to each other by coordinated movements to the same target.

The ventral stream maintains instead a contextual coding of objects in the environment, based on their identity and meaning. Spatially, such coding can be defined as object-centered, as it is mainly concerned with the relative location of objects with respect to each other. This sort of coding is not used at this stage of our research.

The hypothesis of parallel visuomotor channels within the dorsal stream dedicated respectively to the transport and the preshaping components of the reach-to-grasp action is well recognized [8]. Anatomically, these two channels fall both inside the dorsal stream, and are sometimes named dorso–medial and dorso–lateral visuomotor channels [9]. For what concerns proximal joint movements, focus of interest of this research, and according to a well established nomenclature, the most important reach-related cortical areas are V6A and medial intraparietal sulcus (MIP), both receiving their main input from V6 and projecting to the dorsal premotor cortex [9]–[12].

Considering the functional role of the dorso–medial stream, information regarding eye position and gaze direction is very likely employed by area V6A in order to estimate the position of surrounding objects and guide reaching movements toward them. Two types of neurons have been found in V6A that allow us to sustain this hypothesis [6]. The receptive fields of neurons of the first type are organized in retinotopic coordinates,

but they can encode spatial locations thanks to gaze modulation. The receptive fields of the second type of neurons are organized according to the real, absolute distribution of the subject peripersonal space. In addition, V6A contains neurons that arguably represent the target of reaching retinocentrically, and others that use a spatial representation [13]. This strongly suggests a critical role of V6A in the gradual transformation from a retinotopic to an effector-centered frame of reference. Moreover, some V6A neurons appear to be directly involved in the execution of reaching movements [9], indicating that this area is in charge of performing the visuomotor transformations required for the purposive control of proximal arm joints, integrating visual, somatosensory, and somatomotor signals in order to reach a given target in the 3-D space.

B. The Basis Function Approach to Sensorimotor Transformations

Basis functions are building blocks that, when combined linearly, can approximate any nonlinear function, such as those required to map between different neural representations of the peripersonal space (retinotopic, head-centered, arm-centered). Basis function networks have been proposed as a computational solution especially suitable for modeling the kind of sensorimotor transformations performed by the posterior parietal cortex [1], [14], [15]. Networks of suitable basis functions are in fact able to naturally reproduce the gain-field effects often observed in parietal neurons [16]. It was suggested that positions of object in the peripersonal space are coded through the activity of parietal neurons that act as basis functions, and any coordinate frame can be read out from such population coding according to the task requirements [14].

Several different transfer functions can be used as basis functions, the only requirements are that they are nonlinear, that their interaction is also nonlinear (e.g., product versus sum), and that they cover all the possible input range. The most used functions, for their convenience and biological plausibility, are Gaussian and sigmoid functions. For example, retinotopic maps are often modeled by Gaussian basis functions, and eye position by sigmoid, or logistic, functions [14]. Learning in basis function networks is composed of two stages, the first for choosing the shape and location of the basis functions and the second to map them to the output representation. The first step is usually unsupervised, the second depend on errors observed during the sensorimotor interaction with the world.

C. Connectionist Sensorimotor Transformations in Robotics

Although the use of artificial neural networks in robotics is very diffuse and not at all novel [17], just few works concern visuomotor transformations involving arm movements, and especially rare is the coordinate control of gazing and reaching movements. Visuomotor arm control has been usually tackled with the use of self-organizing maps (SOM). Some works [18], [19] apply them to the coordination between visual information and arm control, modeling two cameras (although not in a classical stereo head configuration) and three degrees of freedom manipulators. Neither of the above papers consider eye movements, and their applicability to real robot setups is limited, respectively by the huge number of required learning steps [18], and by the discrete sampling of the space which makes any target reachable only up to a certain precision, and only after a number of steps dependent on the sampling [19]. A recent extension of these works [20] makes use of alternative SOM maps linked to different cameras, in order to deal with occlusion, and takes also into account the issue of obstacle avoidance. Especially interesting is the flexibility of their system to changes in the geometry of the effector, which we are able to reproduce with our approach. Fuke *et al.* [21] have used a SOM to model the relation between arm movements and the perception of a subject own face, as supposedly performed by area ventral intraparietal sulcus (VIP) of the primate brain.

Whilst the use of SOM networks is hence relatively common, the employment of biologically inspired radial basis function (RBF) networks remains relatively unexplored. Although RBF have been successfully applied to the computation of inverse kinematics, alone [22] or together with SOM [23], to the best of our knowledge only two papers describes the use of RBF networks for visuomotor transformations. The system of Marjanovic *et al.* [24] firstly learns the mapping between image coordinates and the pan/tilt encoder coordinates of the eye motors (the saccade map), and then the mapping of the eye position into arm position (the ballistic map). A similar learning strategy is employed by Sun and Scassellati [25], which use the difference vector between the target and the hand position in the eye-centered coordinate system without any additional transformational stages. Despite the similarity of their approaches to ours, some major differences can be pointed out: first of all, we exploit stereo vision, realizing a coordinated control of vergence and version movements, moreover, the saccade map in [24] is fixed and mainly used to provide visual feedback during the ballistic map learning. On the other hand, our sensorimotor transformations are bidirectional, so that our system learns to gaze towards its hand but also to reach where it is looking at. This skill is trained through a self-supervised learning framework, in which the different modalities supervise each other, and both improve contextually their mapping of the space. The distribution of the RBF centers also differs from the cited works, as we place the neural receptive fields according to findings from neurophysiological studies on monkeys.

A few attempts to tackle the problem of coordinate control of gazing and arm movements by using neural networks, but not RBFs, have also been reported. Schenk *et al.* [26] employ a feedforward neural network for learning to saccade toward targets, and a recurrent neural network is employed for executing the transformation carrying from the visual input to an appropriate arm posture, suitable for reaching and grasping a target object. The reaching model of Nori *et al.* [27] consists in learning a motor–motor map to direct the hand close to the fixated object, and then activate a closed loop controller that using visual distance between the hand and the target improves reaching accuracy. Eye gazing control is not adaptive, and they do not consider the importance of contextually maintaining a series of representations in different body reference frames, as suggested by neuroscience findings, especially those regarding posterior parietal area V6A.

III. Sensorimotor Mapping of the Peripersonal Space: Conceptual Framework

As mentioned in Section I and considering the theoretical description of Section II, the main sources of inspiration for our model are the basis function approach [14], [16] and neuroscience experiments on the role of posterior parietal cortical areas (especially V6A) during gazing and reaching actions [6], [9], [13].

The use of a robot hardware constitutes a possible complication in the realization of a model of cortical mechanisms, and some issues that would easily be solved in simulated environments have to be dealt with more accurately considering the real world implementation. The approach we follow is epigenetic, being the robot endowed with an innate knowledge of how to move in its environment, which is later developed and customized through exploration and interaction with visual and tactile stimuli. Following this idea, all transformations are first implemented on a computational model, which final configuration represents the genetic component of the developmental process, and is then used as a bootstrap condition for the actual experimental learning process by the robot.

Although in principle one representation should be enough for all the required transformations, the number of neurons necessary to contextually code for n different signals is given by the size of the signals to the power of n. It is easy to see that a representation maintaining both eye visual and proprioceptive signals, and arm joint information would be computationally unfeasible, even for the brain itself. A more logical structure is one in which a central, body-centered representation is accessed and updated both by limb sensorimotor signals on the one hand and visual and oculomotor signals on the other hand (see Fig. 2). Indeed, this seems to be how the problem is solved within the brain, in which different areas or populations of neurons in the same areas are dedicated to different transformations. Most importantly, this approach is consistent with the findings related to area V6A, which contains neurons that code for the retinocentric position of targets, others that code for their limb-related position, and even others that seem to maintain both codings and look thus especially critical for performing sensorimotor transformations. In this way, different representations of the same target can be maintained contextually, and used to act on the target if required. It is the relation between these representations that is accessed and modified by a conjunction of gaze and reach movements to the same target. The global structure of our model follows this principles, and is thus modular, separating the retinal to body-centered transformation and the body-centered to arm-joint transformations (left and right sides of Fig. 2).

The quality and the nature of the sensory stimuli to be introduced in the schema is quite varied, as the process of 3-D localization requires the integration of information coming from various sources and different modalities. Such integration can be modeled with different levels of detail and considering alternative data sources and formats. In our case, we include visual information about potential targets and proprioceptive data on eye position and arm position. Several possible alternatives for representing the above information can be employed. Among possible alternatives for representing binocular information we

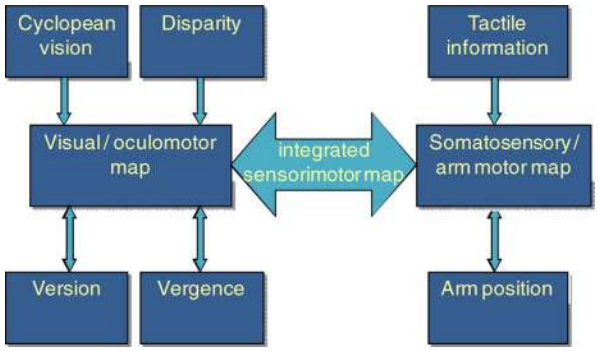

Fig. 2. Building blocks of the global space representation. The central body-centered map constitutes the integrated sensorimotor representation of the peripersonal space.

favor the composition of a cyclopean image representation with a disparity map (under the assumption that the correspondence problem is already solved), over the option of having separate left and right retinotopic maps (see left side of the schema of Fig. 2). Similarly, considering that we are modeling extrastriate and associative visual areas, it is plausible to assume that gazing direction is represented by version and vergence angles instead of the two explicit eye positions. This scheme allows us to transform ocular movements and stereoptic visual information to a body/head-centered reference frame and also, when needed, elicit the eye movements that are necessary to foveate on a given visual target. On the right hand of the conceptual schema of Fig. 2 we find the somatosensory/arm–motor map, required to code arm movements. Such map is modified by proprioceptive and tactile feedback, and allows to execute reaching actions toward visual or remembered targets. The integrated map, built of the two sides of the schema, is thus accessed and updated upon requirements, as described in Section IV-A.

The exploration of the environment through saccades and reaching movements constitutes the basic behavior that is employed to build the visuomotor representation of the peripersonal space. Building such representation is done incrementally, through subsequent, increasingly complex interactions. The learning sequence is inspired by infant development [28]. As a first step, the system learns the association between retinal information and gaze direction (i.e., proprioceptive eye position). This can be done simply by successive foveation on salient points of the binocular images. The subject looks around and focuses the eyes on certain stimuli, thus learning the association between retinal information and vergence and version parameters. Then, gaze direction is associated to arm position, e.g., moving the arm randomly and following it with the gaze, so that each motor configuration of the arm joint is associated to a corresponding configuration of the system for eye motor control. In this case, proprioceptive information regarding arm position is included in the computation, and the vectors corresponding to reaching movements can be extracted similarly to what is done for ocular movements. This process make the subject learn a bidirectional link between different sensorimotor systems. The subject can look where its hand is but also reach a point in space he is looking at. Later on, visual targets are shown to the system, which is required to perform

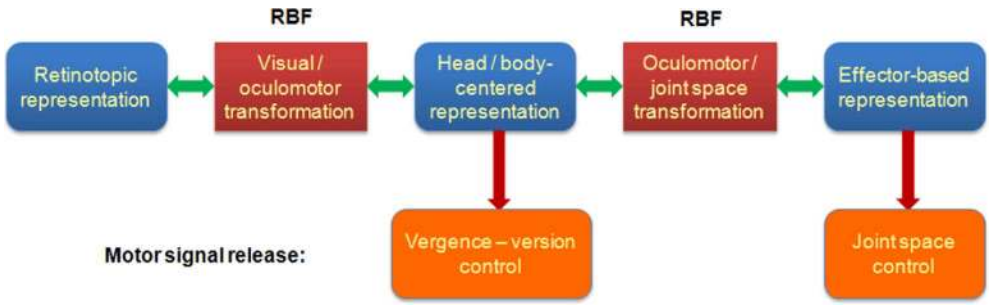

Fig. 3. Computational framework of the visuomotor integration model. Two transformations allow to code a stimulus contextually in visual, oculomotor, and arm–motor frames of reference.

both saccadic and arm reaching movements toward them. This requires the use of both direct and inverse transformations, and allow to fine-tune the sensorimotor representation of the space. Tactile feedback can be used as a master signal, for confirming that the target has been reached, making all the process substantially self-supervised.

IV. Sensorimotor Transformations With Radial Basis Functions

So far, the model has been implemented in a simulated environment, taking always into account the final application on our humanoid robotic setup, currently under development. The computational framework is depicted in Fig. 3, which is a simplification of the conceptual schema of Fig. 2, in which the neck is fixed, and thus body-centered corresponds to head-centered. Also, there is no tactile feedback for the moment, and the control of arm movements is based purely on proprioception.

The visual input regarding a potential target is expressed with its location in a cyclopean visual field accompanied by information on binocular disparity; the output is the correspondent head/body center representation, built of a potential vergence/version movement required to foveate on the target. This transformation has been implemented with a RBF network, described below. The second transformation, also implemented with radial basis functions, is used instead to maintain a contextual coding of stimuli in both a body-centered and an effector based frame of reference. It is used to recode oculomotor coordinates in arm joint space and vice versa.

Each of the two codings of the space corresponds to a potential movement, so that, thank to this second transformation, the agent is able to reach where it is looking at (direct transformation) and to foveate on the position of the hand (inverse transformation). If one of the potential motor signals is not released, eye and arm movements can be decoupled, i.e., the system can for example reach a peripheral visual target without directing the gaze toward it, only using the body-centered representation as an intermediate step to recode visual input to arm motor response. Similarly, arm movements can also be planned but not executed, e.g., waiting for a cue signal in an experimental protocol of delayed reaching. Details regarding the computational implementation of the two transformations are given next.

A. Visual to Oculomotor Transformation

Learning the transformation from binocular visual data to eye position consists in identifying visual targets and foveating them with both eyes, in order to associate appropriate version and vergence movements to retinal locations. Either left and right retinal images or a cyclopean visual field accompanied by a disparity map can be used as visual input, and we employed the latter. Since visual processing is not the focus at this stage of development, visual stimuli are just point-like features, similar to the LEDs used in monkey experiments. The transformation was implemented with an RBF network, for the theoretical reasons explained above.

We tested Gaussian and sigmoid neural activation functions, for both cyclopean visual input and disparity, and try the corresponding nets with different spreads. The best performance was achieved for Gaussian-shaped units for both inputs, described by the following equation, which is the vectorial equivalent of a single variable Gaussian function

$$h_i(\mathbf{x}) = e^{-(\mathbf{x}-\mathbf{c}_i)^T \Sigma_i^{-1} (\mathbf{x}-\mathbf{c}_i)} \tag{1}$$

where $\mathbf{x} = [x_1, x_2 \ldots n_i]$ is the input vector, $\mathbf{c_i}$ the center of the ith bidimensional basis function, Σ_i the diagonal matrix with the spreads of the unit in each of its dimensions, and h_i is the resulting activation of the ith unit. The two outputs of the network, i.e., vergence and version movements required to foveate on the given input, are computed by a linear combination of the basis functions, according to

$$u_k = \sum_{i=1}^{n_h} w_{i,k} \cdot h_i(\mathbf{x}) \tag{2}$$

in which n_h is the number of units, $u_k (k = 1, 2)$ is the value of the kth output and $w_{i,k}$ is the weight that connects the ith RBF unit with the kth output. The learning process consists in finding the weights that best fit input with output datasets.

We decided to employ fixed centers, which receptive fields can not move according to the input data, favoring biological plausibility over potentially better performance. For this reason, we distributed the radial basis functions according to a retinotopic-like criterion (input to V6A is, at least partly, retinotopic), following a logarithmic distribution of the centers. For what concerns cyclopean visual input, a logarithmic organization of the neural receptive fields is suitable for modeling foveal magnification, whilst for disparity it corresponds to a finer coding for smaller disparities, actually observed in the primate visual cortex [29]. It is hence not surprising that the logarithmic organization of the centers allowed to obtain results about ten times as good as by using a homogeneous distribution (0.10 mm against

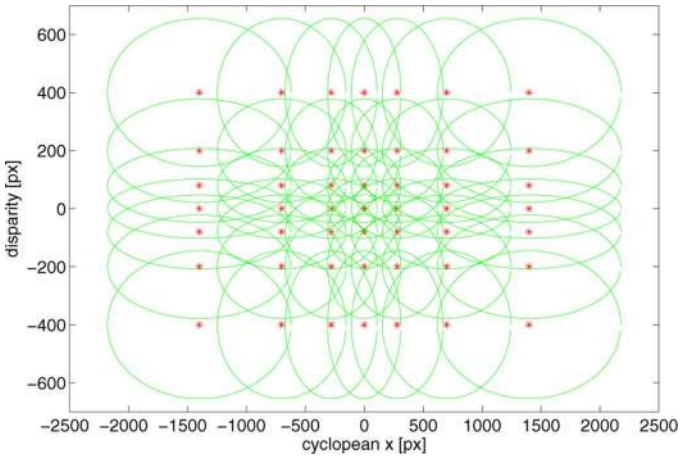

Fig. 4. Distribution and spread of the radial basis functions of the visual to oculomotor network on the cyclopean/disparity space.

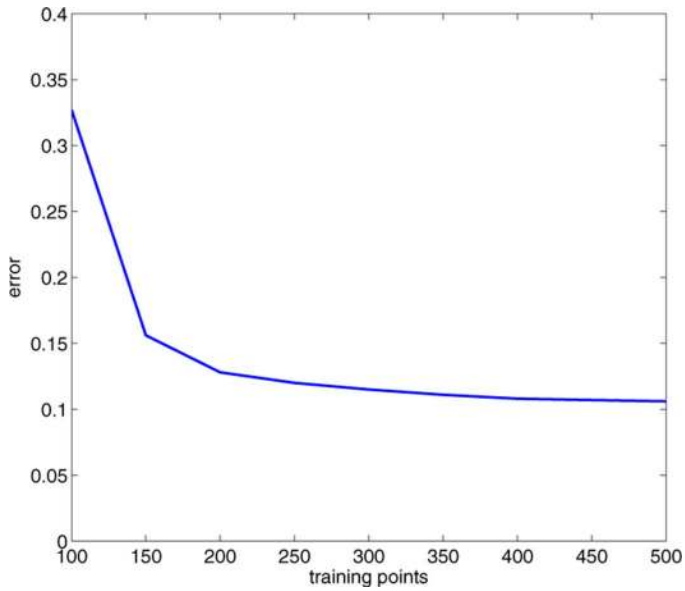

Fig. 5. Oculomotor positioning error (mm) of the network trained with datasets of different size.

1.1 mm). For setting the number of neurons, we defined an arbitrary threshold of 0.10 mm error for $p = 1000$ training points, that was achieved by a 7×7 neural lattice in the cyclopean/disparity input space (thus $n_h = 49$; see Fig. 4).

The training points are constituted by input–output pairs $\{\mathbf{x}_j, \mathbf{y}_j\}_{j=1}^{p}$, and are provided by the simulated execution of saccadic movements and the estimation of the target visual displacement. More exactly, an oculomotor behavior is simulated in which the agent: 1) is fixating a given point in space close to a visual stimulus; and 2) performs a random ocular movement toward a different location. The input-output pairs are composed by the visual position (cyclopean-disparity) of the initial stimulus as seen from the current ocular position, and the vergence–version components of the performed movement. This solution allows to learn a visual to oculomotor mapping without previous focusing skills. The underlying assumption is that of a rich visual environment, in which visual stimuli are always available close to the ocular fixation point. This assumption is reasonable for biological agents, and useful for the modeled system, but the robot has to follow a different strategy, as explained in Section V, due to its limited visual skills in dealing with cluttered environments.

An initial setting of the weights is done employing the batch learning technique of linear pseudo-inverse solution, typical for RBF networks [30]

$$\mathbf{w}_0 = (\mathbf{H}^T \cdot \mathbf{H})^{-1} \cdot \mathbf{H}^T \cdot \mathbf{Y} \tag{3}$$

where $\mathbf{H}$ is a $p \cdot n_h$ matrix containing the output of each RBF neuron for all inputs $\mathbf{x}_j$, while $\mathbf{Y} = \left[\mathbf{y}_1^T \ldots \mathbf{y}_p^T\right]^T$ is the output matrix. We tested this first learning step on datasets of different size p, obtaining the error evolution depicted in Fig. 5. Looking for a trade-off between performance and size of the dataset, we have chosen $p = 200$ (providing error $= 0.127$) as the number of training points to employ in this stage of the process.

To simulate the sort of learning process that would be performed by robot, we executed a step by step learning sequence on the next 200 points. This second learning stage is performed by applying the delta rule gradient descent technique, to update the weights of (2) on each time step t

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \alpha \cdot \mathbf{h}_t^T \cdot (\mathbf{y}_t - \mathbf{u}_t) \tag{4}$$

where $\mathbf{h}_t$ is the output of the neurons for input $\mathbf{x}_t$, $\mathbf{u}_t$ is the actual output of the network, $\mathbf{y}_t$ the expected output and α the learning rate (which we maintain constant). The use of an incremental learning rule such as in (4), instead of a batch learning as in(3), allows to keep the system flexible to possible changes in visual accuracy and body kinematics. In principle, applying the delta rule should allow to adapt to unavoidable hardware asymmetries, image distortions, and also to deceptive sensory information, as described below.

The inverse transformation from oculomotor to visual data has also been implemented, using the same parameters of the direct one. Its role is to estimate the expected retinal position of a fixated point after a given saccadic (vergence + version) movement. We plan to use it in the real robot to detect possible discrepancies between expected and observed visual feedback. We considered most plausible for this transformation to employ the same parameters of the direct one, such as the retinal organization of the centers, and probably for this reason it is not as precise as the visual to oculomotor transformation (the average error is about 0.5 mm).

To validate the model, we are comparing its behavior with some psychophysical effects described in the literature regarding the tasks it executes. For example, we are checking the model behavior in the case of the deceptive visual feedback, such as in typical experiments of saccadic adaptation [31]. This is done by eliciting a saccade (based on vergence/version eye movement control) toward a given visual target, and providing a fictitious error on the final reached position. For the computational model, this is achieved by adding an offset to the output. On the robot, the same effect will be obtained moving

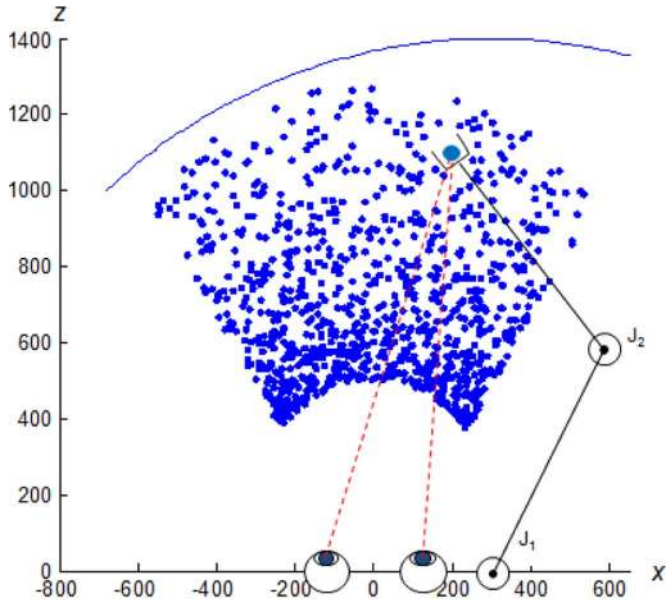

Fig. 6. Gazing and reaching schema. At each training step the artificial agent, either model or robot, is required to move its hand and gaze toward the same point, and update its sensorimotor representation using the observed error.

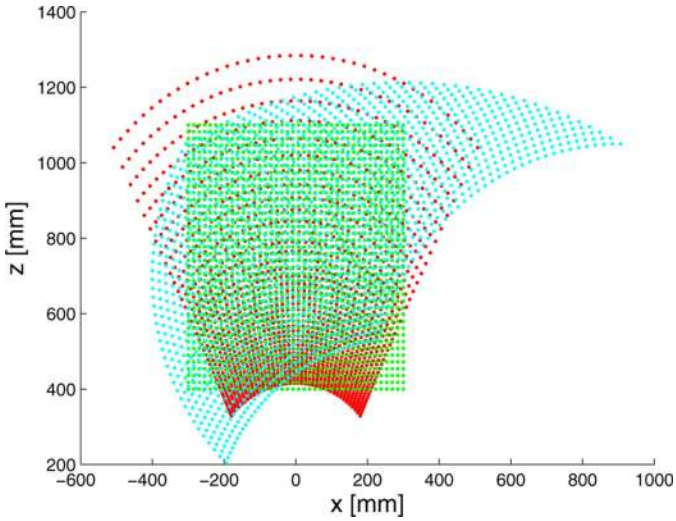

Fig. 7. Mapping of the space according to uniform distributions in a vergence/version oculomotor space (red), in a J1/J2 joint space (cyan) and in a standard horizontal/depth (x/z) Cartesian space (green).

the visual target as for human subjects. Analysis of how (as in the saccadic adaptation protocol) such artificial displacement of the target affects the artificial agent oculomotor and arm motor abilities can serve as a validation of the underlying model, and may help in advance hypotheses on saccadic adaptation mechanisms in humans and monkeys. So far, we were able to verify that our model do exhibit saccadic adaptation, altering its ability to perform correct saccades according to the deceptive feedback. The analysis of error distributions around the target point and of error vectors is also providing interesting information that we are currently studying with more detail, together with collaborators from cognitive sciences.

B. Oculomotor to Arm–Motor Transformation

In this learning phase, arm movements are introduced, as exemplified in Fig. 6. This phase is further subdivided in two stages, respectively free and goal-based. The free exploration consists of random arm movements and subsequent saccades toward the final hand position, which allows to learn the transformation from joint space to oculomotor space and vice versa. To generate the training points constituted by oculomotor/arm joint couples, a correct behavior of the visuomotor to oculomotor network in gazing movements is required, so the second transformation can only be trained after the first one. In the goal-oriented exploration a target object in space has to be foveated and reached.

The choice of how to distribute the basis function neurons is less straightforward for this second network. Automatic placing driven by the training points is a standard solution [25], but again we favor biological plausibility over performance. Our main inspiration is on neuroscience findings regarding the posterior parietal cortex, and especially area V6A. In a previous work, we showed that a population of V6A neurons is properly modeled by a basis function approach [32]. As anticipated, this area includes neurons having only visual response, neurons apparently involved mainly in motor actions and mixed neurons, activated in all phases of sensorimotor processes. With our model we wanted to check what computational advantages could be given by such responsiveness pattern. For simplicity at this stage, only two arm joints were used, and no tilt movements of the eyes, so that the accessible environment is a 2-D space placed horizontally in front of the subject, as in Fig. 6. This is anyway consistent with most of the monkey experiments in which activity in V6A was registered.

At this stage of the model development, we want thus to achieve good performances in the learning of the transformations between oculomotor and arm motor space, while respecting, and trying to emulate, the responsiveness pattern observed in area V6A. We simulated the different types of neurons of V6A with populations of radial basis function neurons uniformly distributed in the vergence/version space (representing oculomotor neurons) and in arm joint space (representing arm–motor neurons). Homogeneous distributions are used in this case instead of logarithmic ones, because the reachable space has to be covered all with the same precision. Again, we tested with both Gaussian and sigmoid functions, finding slightly better results for the former, as for the first network.

In order to check their suitability to model the transformations performed by V6A neurons, we trained RBF networks having the centers distributed as in Fig. 7, red and cyan graphs, for vergence/version and joint space respectively. V6A and nearby areas perform all the transformations required for a correct gazing and reaching, and for this reason, an important requirement is that the same pool of artificial neurons, centers of the radial basis functions, have to be used in the direct and inverse transformations, so we included both transformations in the comparison. To avoid biasing toward one or the other distribution, training (again 400 points) and test sets were taken randomly from a Cartesian space. As depicted in Fig. 7, the

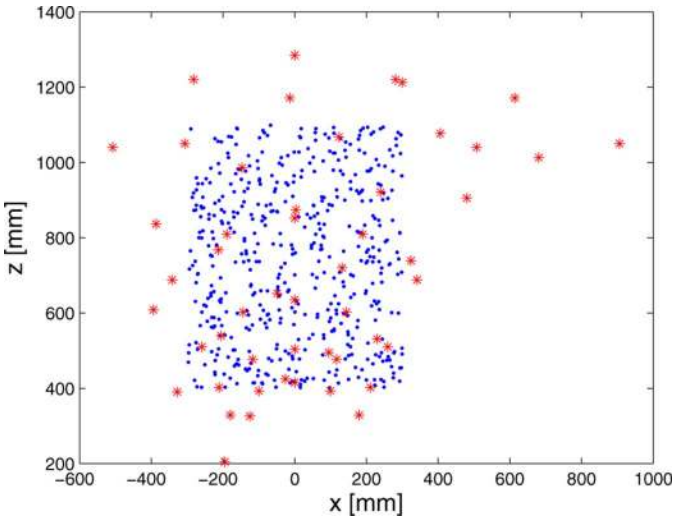

Fig. 8. Radial basis functions distributed according to a mixed criterion (25 oculomotor $+$ 25 arm motor, red stars), visualized over a typical Cartesian training set (blue dots).

TABLE I
OCULOMOTOR AND ARM JOINT SPACE RANGES

	Min. [*rad*]	Max. [*rad*]
Vergence	$\pi/15$	$\pi/5$
Version	$-\pi/7$	$\pi/7$
	Min. [*rad*]	Max [*rad*]
J1	$\pi/6$	$\pi/2$
J2	$\pi/3$	$3\pi/4$

TABLE II
PERFORMANCE OF RBF NETWORKS WITH NEURONS DISTRIBUTED ACCORDING TO A VERGENCE/VERSION OCULOMOTOR SPACE (V), ARM JOINT SPACE (J), AND MIXED SPACE (M), FOR BOTH DIRECT AND INVERSE TRANSFORMATIONS OCULOMOTOR ⇔ ARM–JOINT

Neuron	V⇒J transformation		J⇒V transformation	
distribution	Error (mm)	St. dev.	Error (mm)	St. dev.
J	2.92	5.69	2.27	3.48
V	4.76	7.63	0.74	1.64
M	1.07	1.20	0.29	0.48
forward select	1.63	1.80	0.63	1.06

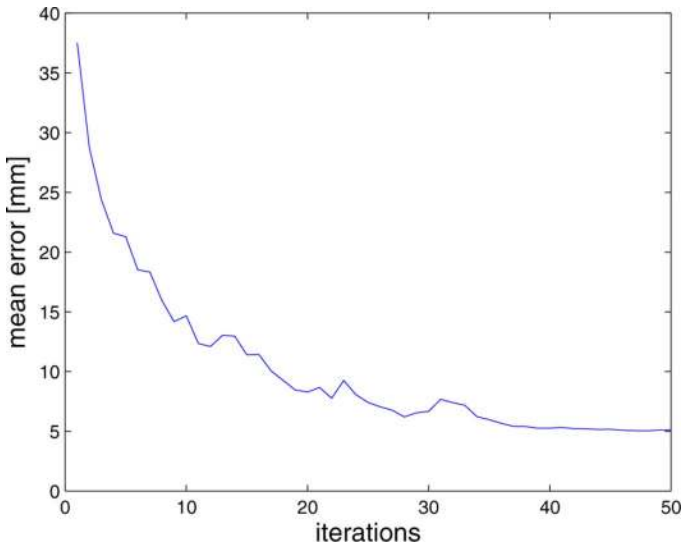

Fig. 9. Typical learning curve of the $J \rightarrow V$ transformation during the adaptation to the new parameters after the kinematics of the robot model has been changed.

ranges were taken so that the superposition between the center distributions and the training and test sets were equivalent between the oculomotor and the joint space. The exact ranges of vergence and version and of joints J1 and J2 employed in both direct and inverse transformations are provided in Table I. A further complication in the comparison between distributions is that different neuron placements and different transformations are optimized with different number of neurons and amplitudes. We tried to normalize the various solutions as much as possible in order to make them comparable. The number of neurons of the pure oculomotor and joint space distributions were 49 (7×7), to repeat the population of the first network, whilst for the mixed distribution we employed 50 neurons ($5 \times 5 \times 2$) to match the total number of neurons as close as possible, obtaining the placement shown in Fig. 8. For each configuration we searched for the values of the spreads which provide the smallest errors, to compare their best performances. The learning process did not differ from the visual to oculomotor case, and the same equations apply.

The results of the different tested configurations are shown in Table II. As it can be observed, the joint space distribution of neurons is reasonably good in both transformations, from oculomotor to joint space and inverse, whilst the vergence/version distribution is good only for the joint to oculomotor transformation, which in general seems to be easier to learn than its inverse. A mixed distribution, with both types of neurons, allows to obtain the best results in both transformations, much better than either distribution alone. As a further experiment, we tried to distribute the neurons according to a forward selection algorithm, that automatically place the centers to best fit the training data. As shown in Table II, the results are better than the single criterion distribution but not better than the mixed one. The stop condition for the forward select algorithm was to have 50 neurons, to allow for a fair comparison with the other methods.

Apart for the shear improvement in performance, the use of the mixed distribution should be especially suitable for modifying working conditions. To test this hypothesis, and to estimate the sort of results we could expect applying the computational framework to the robotic setup, we changed the kinematic parameters of the robot model, and start training the network with the old weights from the new configuration. The parameters included in the model are five: lengths of arm L_a and forearm L_f, interocular distance I, and relative position of shoulder and eyes (two parameters, supposing they are aligned in the z coordinate). Considering for example the more demanding oculomotor to joint-space transformation, we modified the first three of the above parameters, leaving unchanged the relative position of shoulder and eyes. Changing L_a and L_f both from 700 to 650 mm and I from 270 to 240 mm, the error first rises up to 40 mm, and drops back almost to the original precision only after about 50 trials, as shown in Fig. 9. This behavior shows the adaptability of the system to changes in working conditions, and supports its suitability for implementation on the robot. Moreover, this test shows that the RBF architecture constitutes a simple body-schema for the

robot [33], as it implicitly represents its internal parameters, and is able to plastically adapts to modified conditions and altered body parts.

Recent experiments [P.Fattori, unpublished data] show that the receptive fields of many V6A neurons seem to be indeed distributed according a vergence/version criterion. Less clear is the effect of arm joints, also because of our simplification of the actual joint space. In any case, our simulation supports the hypothesis that a mixed population of neurons such as that observed in V6A is especially suitable for a cortical area which contextually codes for different reference frames. From a pragmatic point of view, through the use of basis function neurons set according to what suggested by neuroscience data, we were able to learn very accurately direct and inverse transformations between oculomotor and joint space, in a way suitable for their application to the robotic setup.

V. Robotic Setup and Experimental Framework

On the robotics side, the final goal of this work is to provide the robot with advanced skills in its interaction with the environment, namely in the purposeful exploration of the peripersonal space and the contextual coding and control of eye and arm movements. On the other hand, the implementation on an actual sensorimotor setup is a potential source of additional insights for the computational model, hardly achievable with simulated data. Extensive experimentation with the robot is not yet available, and constitutes the bulk of our current work, which methodology is outlined below.

Our humanoid robot (Fig. 10) is endowed with a pan-tilt-vergence stereo head with coordinated vergence/version control of the eyes and a multijoint arm with a three finger Barrett Hand (not used in this work). The sort of stimuli the robot is able to recognize by visual processing are simple dots/crosses/blobs on a computer monitor, or custom visual markers placed on a visible section of the robot arm, independently from their position in the robot field of view. The workspace is first positioned at eye level, so that only 2-D eye and arm movements are required. After the 2-D transformation have been successfully applied to the robot according to the model described in the previous section, we plan to extend it to the 3-D space, introducing tilt movements of the head and at least one more joint for the arm. Preliminary studies with three-input RBF transformations were successful in this regard.

As explained above, the actual map of the peripersonal space is learned through active exploration, following the increasingly demanding sequence of tasks described in Section III. The training points for the visual to oculomotor transformation are not collected as for the model, because of the robot limited skills in the analysis of complex visual environments. The employed solution is to visualize at each step a new visual target, to which the robot has to try and perform a gazing movement using the visual to oculomotor map itself. The gaze is very likely not to land on the target, and the estimated residual distance of the target from the new fixation point is used to train the network. At the beginning, the residual error of the movement can be very high, and this is one of the main reasons which justifies the use of bootstrap learning with weights provided by the model, in order to start with an acceptable gazing performance, that is refined by online learning.

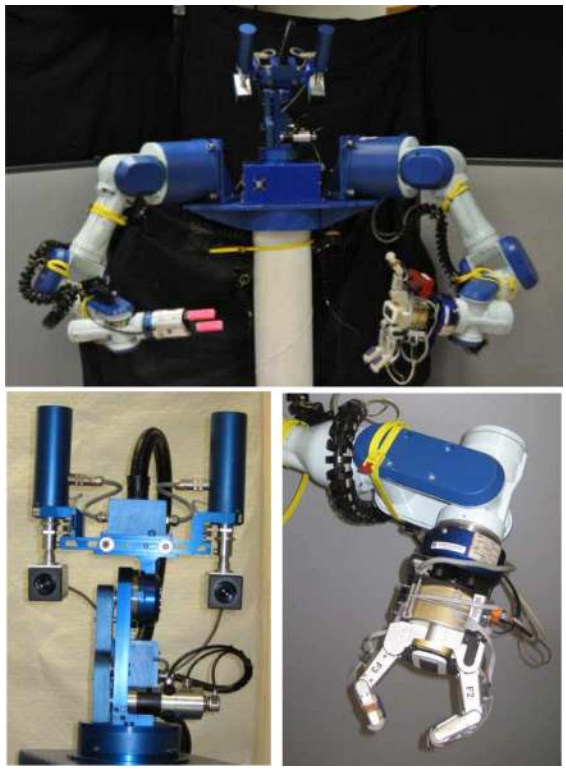

Fig. 10. Humanoid robot with detail of pan/tilt/vergence head and arm with hand.

Once the robot has learned to perform the visual to oculomotor transformation according to the above schema, no additional skills are required to train the oculomotor to joint space networks. The robot has only to be able to gaze toward the visual marker placed on its arm, which is moved in a random position at each step, using the first network. The eye/arm movement pairs are then used to train both direct and inverse transformations between eye and arm motor representations.

Following the above principles, the robot keeps improving its visuomotor and arm–motor skills in each gazing or reaching movement towards nearby goals. The use of tactile feedback upon object touching could finally constitute a master signal that allows to infer the exact magnitude of visual and motor errors. The adaptability of the RBF-based computational framework, highlighted by experiments on saccadic adaptation and with altered kinematics conditions, indicates that the learning framework seems indeed appropriate to be transferred to the real hardware.

VI. Conclusion

Experiments of concurrent reaching and gazing allow to generate an implicit representation of the peripersonal space obtained by matching head-center and arm-centered schemes. Such representation remains implicit, and far from being an actual map of the environment, it rather constitutes a skill of the robot in interacting with it. As a first implementation of the model, simulated experiments of coordinated reach/gaze actions have been performed, in which there is visual tracking

of the effector but not tactile feedback. The implemented RBF framework is capable of bidirectional transformations between stereo visual information and oculomotor (vergence/version) space, and between oculomotor and arm joint space. For our modeling purposes we used insights and functional indications coming from monkey and human studies, especially regarding the transformations and the contextual encoding of features in the peripersonal space performed by area V6A. The computational structure which allows to jointly represent oculomotor and joint space was defined in accordance to the above studies, supporting the hypothesis that a mixed population of neurons is the most suitable for performing different transformations. Additionally, the system is able to adapt to altered conditions, such as visual distortions or modified kinematic configurations, and experiments were performed in which the agent had its kinematics changed and was able to learn the correct actions associated to the new parameters.

The final, integrated representation of the peripersonal space emerges thanks to the simulated interaction of the agent with the environment. Such implicit representation allows to contextually represent the peripersonal space through different vision and motor parameters. Very importantly, the oculomotor/arm motor transformation is bidirectional, and the underlying representations are both accessed and modified by each exploratory action. The above schema is now being implemented on a real humanoid torso, in which coordinated reach/gaze actions are being used to integrate and match the sensorimotor maps. This learning process is the normal behavior of the agent, constituting the most fundamental component of its basic capability of interacting with the world, and contextually updating its representation of it.

REFERENCES

[1] A. Pouget and L. H. Snyder, "Computational approaches to sensorimotor transformations," *Nature Neurosci.*, vol. 3 Suppl, pp. 1192–1198, Nov. 2000.

[2] S. Deneve and A. Pouget, "Basis functions for object-centered representations," *Neuron*, vol. 37, no. 2, pp. 347–359, Jan. 2003.

[3] R. A. Andersen, R. M. Bracewell, S. Barash, J. W. Gnadt, and L. Fogassi, "Eye position effects on visual, memory, and saccade-related activity in areas LIP and 7A of macaque," *J. Neurosci.*, vol. 10, no. 4, pp. 1176–1196, 1990.

[4] F. Bremmer, C. Distler, and K.-P. Hoffmann, "Eye position effects in monkey cortex. II: Pursuit and fixation related activity in posterior parietal areas LIP and 7A," *J. Neurophysiol.*, vol. 77, pp. 962–977, 1997.

[5] C. R. Cassanello, A. T. Nihalani, and V. P. Ferrera, "Neuronal responses to moving targets in monkey frontal eye fields," *J. Neurophysiol.*, vol. 100, pp. 1544–1556, 2008.

[6] P. Fattori, D. F. Kutz, R. Breveglieri, N. Marzocchi, and C. Galletti, "Spatial tuning of reaching activity in the medial parieto-occipital cortex (area V6A) of macaque monkey," *Eur. J. Neurosci.*, vol. 22, no. 4, pp. 956–972, Aug. 2005.

[7] M. A. Goodale and A. D. Milner, *Sight Unseen*. London, U.K.: Oxford Univ. Press, 2004.

[8] M. Jeannerod, "Visuomotor channels: Their integration in goal-directed prehension," *Human Movement Science*, vol. 18, no. 2, pp. 201–218, Jun. 1999.

[9] C. Galletti, D. F. Kutz, M. Gamberini, R. Breveglieri, and P. Fattori, "Role of the medial parieto-occipital cortex in the control of reaching and grasping movements," *Exp. Brain Res.*, vol. 153, no. 2, pp. 158–170, Nov. 2003.

[10] P. Fattori, M. Gamberini, D. F. Kutz, and C. Galletti, "'Arm-reaching' neurons in the parietal area V6A of the macaque monkey," *Eur. J. Neurosci.*, vol. 13, no. 12, pp. 2309–2313, Jun. 2001.

[11] P. Dechent and J. Frahm, "Characterization of the human visual V6 complex by functional magnetic resonance imaging," *Eur. J. Neurosci.*, vol. 17, no. 10, pp. 2201–2211, 2003.

[12] R. Caminiti, S. Ferraina, and A. B. Mayer, "Visuomotor transformations: Early cortical mechanisms of reaching," *Current Opinion Neurobiol.*, vol. 8, no. 6, pp. 753–761, Dec. 1998.

[13] N. Marzocchi, R. Breveglieri, C. Galletti, and P. Fattori, "Reaching activity in parietal area V6A of macaque: Eye influence on arm activity or retinocentric coding of reaching movements?," *Eur. J. Neurosci.*, vol. 27, no. 3, pp. 775–789, Feb. 2008.

[14] A. Pouget and T. J. Sejnowski, "A new view of hemineglect based on the response properties of parietal neurones," *Philos. Trans. Roy. Soc. B: Biol. Sci.*, vol. 352, no. 1360, pp. 1449–1459, Oct. 1997.

[15] A. Pouget, S. Deneve, and J.-R. Duhamel, "A computational perspective on the neural basis of multisensory spatial representations," *Nat. Rev. Neurosci.*, vol. 3, no. 9, pp. 741–747, Sep. 2002.

[16] E. Salinas and P. Thier, "Gain modulation: A major computational principle of the central nervous system," *Neuron*, vol. 27, no. 1, pp. 15–21, Jul. 2000.

[17] , G. A. Bekey and K. Y. Goldberg, Eds., *Neural Networks in Robotics*. Norwell, MA: Kluwer Academic, 1993.

[18] T. M. Martinetz, H. J. Ritter, and K. J. Schulten, "Three-dimensional neural net for learning visuomotor coordination of a robot arm," *IEEE Trans. Neural Netw.*, vol. 1, no. 1, pp. 131–136, Mar. 1990.

[19] M. Jones and D. Vernon, "Using neural networks to learn hand-eye co-ordination," *Neural Comput. Appl.*, vol. 2, no. 1, pp. 2–12, 1994.

[20] M. Han, N. Okada, and E. Kondo, "Coordination of an uncalibrated 3-d visuo-motor system based on multiple self-organizing maps," *JSME Int. J. Series C Mech. Syst., Mach. Elements Manufact.*, vol. 49, no. 1, pp. 230–239, 2006.

[21] S. Fuke, M. Ogino, and M. Asada, "Acquisition of the head-centered peri-personal spatial representation found in vip neuron," *IEEE Trans. Autonom. Mental Develop.*, vol. 1, no. 2, pp. 131–140, Aug. 2009.

[22] P.-Y. Zhang and T.-S. L.-B. Song, "RBF networks-based inverse kinematics of 6r manipulator," *Int. J. Adv. Manufact. Technol.*, vol. 26, no. 1–2, pp. 144–147, Jul. 2005.

[23] S. Kumar, L. Behera, and T. McGinnity, "Kinematic control of a redundant manipulator using an inverse-forward adaptive scheme with a ksom based hint generator," *Robot. Autonom. Syst.*, vol. 58, no. 5, pp. 622–633, Mau 2010.

[24] M. Marjanovic, B. Scassellati, and M. Williamson, "Self-taught visually-guided pointing for a humanoid robot," in *Proc. Int. Conf. Simulation Adapt. Behav. (SAB) 1996*, 1996.

[25] G. Sun and B. Scassellati, "A fast and efficient model for learning to reach," *Int. J. Human. Robot.*, vol. 2, no. 4, pp. 391–413, 2005.

[26] W. Schenck, H. Hoffmann, and R. Möller, F. Schmalhofer, R. M. Young, and G. Katz, Eds., "Learning internal models for eye-hand coordination in reaching and grasping," in *Proc. EuroCogSci 2003*, 2003, pp. 289–294.

[27] F. Nori, L. Natale, G. Sandini, and G. Metta, "Autonomous learning of 3D reaching in a humanoid robot," in *Proc. IEEE Int. Conf. Intell. Robot. Syst.*, San Diego, CA, 2007, pp. 1142–1147.

[28] K. E. Adolph and A. S. Joh, "Motor development: How infants get into the act," in *Introduction to Infant Development*, A. Slater and M. Lewis, Eds. London, U.K.: Oxford Univ. Press, 2007, pp. 63–80.

[29] G. F. Poggio, F. Gonzalez, and F. Krause, "Stereoscopic mechanisms in monkey visual cortex: Binocular correlation and disparity selectivity," *J. Neurosci.*, vol. 8, no. 12, pp. 4531–4550, Dec. 1988.

[30] C. M. Bishop, *Neural Networks for Pattern Recognition*. London, U.K.: Oxford Univ. Press, 1995.

[31] T. Collins, K. Dore-Mazars, and M. Lappe, "Motor space structures perceptual space: Evidence from human saccadic adaptation," *Brain Res.*, vol. 1172, pp. 32–39, Aug. 2007.

[32] E. Chinellato, B. J. Grzyb, N. Marzocchi, A. Bosco, P. Fattori, and A. P. del Pobil, "Eye-hand coordination for reaching in dorsal stream area V6A: Computational lessons," in *Bioinspired Applications in Artificial and Natural Computation*, J. Mira, J. M. Ferrez, J.-R. Alvarez, F. de la Paz, and F. J. Toledo, Eds. Berlin, Germany: Springer-Verlag, 2009, vol. LNCS 5602, pp. 304–313.

[33] M. Hoffmann, H. Marques, A. Arieta, H. Sumioka, M. Lungarella, and R. Pfeifer, "Body schema in robotics: A review," *IEEE Trans. Autonom. Mental Develop.*, vol. 2, no. 4, pp. 304–324, Dec. 2010.

Eris Chinellato (S'03–M'08) received the B.Sc. degree in industrial engineering from the Universitá degli Studi di Padova, Padova, Italy, in 1999, the M.Sc. degree in artificial intelligence, with the Best Student Prize, from the University of Edinburgh, Edinburgh, U.K., in 2002, and the Ph.D. degree in intelligent robotics from Jaume I University, Castellón de la Plana, Spain, in 2008.

His interdisciplinary research is mainly focused on the use of visual information for reaching and grasping actions in natural and artificial systems. He has published in influential journals and proceedings in robotics, neuroscience, and computational neuroscience, and has served as reviewer and program committee member for international journals and conferences.

Marco Antonelli received the M.Sc. degree in computer engineering, with an industrial automation specialism, from Universitá degli Studi di Padova, Padova, Italy, in 2008. He is currently working towards the Ph.D. degree at the Robotic Intelligence Lab, Universitat Jaume I, Castellón de la Plana, Spain.

He is collaborating with the EU-FP7 Project "Eyeshots." His research is mainly focused on biologically inspired humanoid robotics. He has published in international conferences and journals and participated to specialized courses on the subject. He is now involved in the development of a model of multisensory egocentric representation of the 3-D space based on binocular visual cues, and oculomotor and arm–motor signals.

Beata J. Grzyb received the M.Sc. degree in computer science from Maria Curie-Sklodowska University, Lublin, Poland. She is currently working towards the Ph.D. degree in the Robotic Intelligence Lab, Jaume I University, Castellón de la Plana, Spain.

In her research, she follows the approach of cognitive developmental robotics, and tackles problems related to body representation, peripersonal space representation, and perception of body effectivities, by means of synthesizing neuroscience, developmental psychology, and robotics, and she has already published in several journal and proceedings.

Angel P. del Pobil received the B.Sc. degree in physics in 1986, and the Ph.D. degree in engineering robotics in 1991, both from the University of Navarra, Navarra, Spain. His Ph.D. dissertation was the winner of the 1992 National Award of the Spanish Royal Academy of Doctors.

He is currently a Professor of Computer Science and Artificial Intelligence at Jaume I University, Castellón de la Plana, Spain, and is the founding Director of the Robotic Intelligence Laboratory. He is author or coauthor of over 120 research publications, and has been invited speaker of 34 tutorials, plenary talks, and seminars. His past and present research interests include motion planning, visually guided grasping, service robotics, mobile manipulators, visual servoing, learning for sensor-based manipulation, and the interplay between neurobiology and robotics.