A hierarchical system for a distributed representation of the peripersonal space of a humanoid robot

Marco Antonelli, Agostino Gibaldi, Frederik Beuth, Angel J. Duran, Andrea Canessa, Manuela Chessa, Fabio Solari, Angel P. del Pobil, Fred Hamker, Eris Chinellato, Silvio P. Sabatini

***Abstract*—Reaching a target object in an unknown and unstructured environment is easily performed by human beings. However, designing a humanoid robot that executes the same task requires the implementation of complex abilities, such as identifying the target in the visual field, estimating its spatial location, and precisely driving the motors of the arm to reach it. While research usually tackles the development of such abilities singularly, in this work we integrate a number of computational models into a unified framework, and demonstrate in a humanoid torso the feasibility of an integrated working representation of its peripersonal space. To achieve this goal, we propose a cognitive architecture that connects several models inspired by neural circuits of the visual, frontal and posterior parietal cortices of the brain. The outcome of the integration process is a system that allows the robot to create its internal model and its representation of the surrounding space by interacting with the environment directly, through a mutual adaptation of perception and action. The robot is eventually capable of executing a set of tasks, such as recognizing, gazing and reaching target objects, which can work separately or cooperate for supporting more structured and effective behaviors.**

***Index Terms*—implicit distributed representation, sensorimotor learning, humanoid robot, visual cortex, object recognition.**

I. Introduction

THE introduction of robotic agents in humans' daily life requires the development of autonomous systems able to interact with changing and unstructured environments. In such conditions, conceiving and selecting the most suitable action to accomplish the desired task can be done only if the robot is endowed with adequate sensors, to access its internal state with respect to the state of the environment. Visual processing by itself is usually not sufficient, because extracting information from the environment is generally an ill-posed problem [1]. However, an agent that is endowed with the capability of voluntarily moving itself can simplify the perceptual process by means of active exploration [2], paying the price of an increasing complexity of the architecture.

On the one hand, the robot can exploit the interaction with the peripersonal space, both to calibrate its own internal model and to create its egocentric representation of the environment. From this perspective, the representations of both the environment and the internal model are intertwined for achieving an action, and hence they should be developed in parallel [3]. On the other hand, active exploration can be fulfilled only with a cognitive agent that flexibly integrates the perception of the environment with the performed actions.

Biological systems provide an important source of inspiration for developing cognitive abilities on robots, because of their capability of adaptation. In particular, neuroscientific and psychophysical findings provide new models of brain functions that can be adapted for implementation on robotic systems.

Thus, composing an integrated system is instrumental to investigate the existing interactions between vision and motor control, and to study how to exploit these interactions. Even if some authors have developed integrated robotic systems [4]–[6], they typically rely on a computer vision approach, without taking inspiration from computational neuroscience. Otherwise, several single models developed by the computational neuroscience community have been successfully implemented on robotic setups, but few works have focused on the integration of such models [7]–[10].

To achieve the above goals, we propose a framework that hierarchically integrates computational models based on the brain areas subtending visual attention, object recognition and localization, and sensorimotor learning [11], [12]. Grounding on the functional structure of the visual cortex, the architecture is composed of two parallel and interacting pathways inspired by the ventral and the dorsal streams [13] of the primate cortex. Both streams rely on a common visual front-end that models the primary visual cortex (V1), which provides a cortical-like (*i.e.* distributed) representation of the binocular visual signal [14], [15].

Ventral stream processing ("vision for perception") is based on an object recognition module (OR) that employs a distributed representation of objects as observed in the primate cortex [16]. This representation, which encodes oriented edges, contrast differences and retinal disparity, allows the robot to identify and localize known objects.

Dorsal stream processing ("vision for action"), includes

M. Antonelli, A. P. del Pobil (corresponding authors) and A. J. Duran are with Robotic Intelligence Lab, Universitat Jaume I, Spain. e-mail: {antonell,pobil,abosch}@uji.es. A.P. del Pobil is also with Sungkyunkwan University, Seoul, Korea.

F. Beuth and F. Hamker are with Department of Computer Science, Chemnitz University of Technology, Germany. e-mail: {beuth,fhamker}@hrz.tu-chemnitz.de

A. Gibaldi, A. Canessa, M. Chessa, F. Solari and S.P. Sabatini are with Department of Informatics, Bioengineering, Robotics and System Engineering, University of Genoa, Italy. e-mail: {agostino.gibaldi,silvio.sabatini}@unige.it

E. Chinellato is with University of Leeds, UK. e-mail: e.chinellato@leeds.ac.uk

This work was supported in part by EC (Project FP7-ICT-217077 EYESHOTS), by Ministerio de Ciencia y Innovación (FPI grant BES-2009-027151, DPI2011-27846), by Generalitat Valenciana (PROMETEO/2009/052) and by Fundació Caixa-Castello-Bancaixa (P1-1B2011-54).

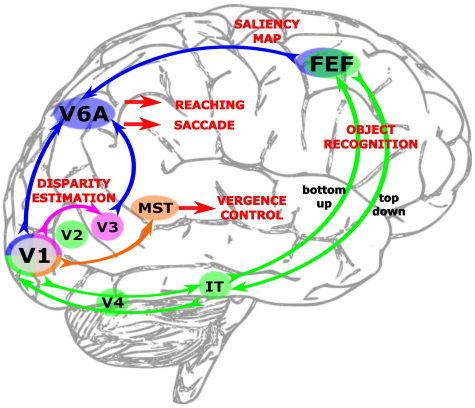

Fig. 1. Brain areas and their interconnections involved in the active exploration of the peripersonal space. The colors denote their functionality: vergence movements (orange), disparity estimation (pink), object recognition (green), gazing and reaching actions (blue). Abbreviations: V1: primary visual cortex; V2-V4: visual areas in the extrastriate cortex; IT: inferior temporal cortex; MST: medial superior temporal area; V6A: visuomotor medial posterior parietal area V6A; FEF: frontal eye field.

both eye vergence control [17], [18] and gazing and reaching movements [19], [20]. The system constructs a sensorimotor egocentric representation of the space which is based on different sources of information, like binocular visual cues (disparity map), the visual position of target objects (coded by the frontal eye field), signals from the oculomotor system (eye joint position) and signals related to the reaching movements performed by the arm (arm joint position). The robot learns to integrate these cues through gaze and reach movements by comparing the outcome of the performed action with the prediction of the internal model [19]. The ventral stream ability to recognize objects, and thus identify the correct target, is instrumental to calibrate the dorsal stream, in accordance with the psychophysical and neurological evidence for the interaction between the two streams [14].

The integration of all modules results in a system that achieves brilliant sensorimotor skills, flexibly integrating sensory and motor information, and using specialized components that interact with each other and tune themselves to the task at hand.

The remainder of the paper is organized as follows. In Section II we introduce the neuroscience background upon which the proposed model, described in Section III, is inspired. In Section IV we describe the robotic setup and provide the details of the control system, and we illustrate the achieved results in Section V. Section VI offers discussion and conclusion, and a comparison of our results with related works. Mathematical details of the models are reported in the appendices.

II. Neuroscience Background

In primates, the observed scene is perceived by the vision system through 2D projections on the left and right retinas. The retinal ganglion cells are connected to lateral geniculate nucleus in the thalamus which project directly to the primary visual cortex (V1).

V1 is characterized by *simple* and *complex cells*. Both types of cells have small receptive fields, and are sensitive to monocular and binocular visual stimuli, such as oriented edges, moving bars or gratings, and binocular disparities [21], [22]. However, complex cells responses, differently from simple cells, exhibit a specific invariance to the phase of the stimulus [21], which makes them perfectly suitable to unambiguously encode binocular disparity (from an implementation point of view, V1 can be seen as the substrate that encodes the raw binocular information provided by the retinas into a feature-based space). Downstream from V1, visual processing splits into two parallel streams.

The ventral stream [16] performs object recognition, and consists mainly of the visual cortical areas V1, V2, V4 and inferior temporal cortex (area IT) (green regions in Fig. 1). Each of these areas is sensitive to specific features that get increasingly complex and invariant against affine transformations. For example, V1 cells respond to edges, V2 cells to a combination of edges (e.g. corners), V4 cells to small parts of an object and IT cells to a whole object or to some of its views [23]. Other features like color and disparity are hierarchically encoded in the ventral stream, and integrated with texture features. Apart from the main bottom-up projections from V1 to IT, there is also a top-down bias on visual processing which specifies which features are most important for the task at hand, e.g. for the attentive search of target objects in the scene [11].

While the ventral stream detects target objects, the dorsal stream estimates their spatial location and their size (blue regions in Fig. 1). The dorsal stream is also in charge of planning eye movements such as vergence and saccades. Vergence movements are used to change the fixation distance in order to simultaneously foveate the visual target with both eyes, so as to restore and/or maintain the singleness of vision. In this process, disparity information provided by V1 is interpreted by the medial superior temporal area (MST) [24] to gather a signal proportional to the disparity to be reduced. Thus, vergence is a close-loop movement and is usually relatively slow ($\approx$ 60deg/s).

Saccades are fast, ballistic movements that are used to gaze at visual stimuli. Once the target stimulus is detected in the visual field, the oculomotor system triggers a saccade to shift the gaze to the target, and therefore eye version and vergence are both changed [25]. The movement can be as fast as 900 deg/s and its execution is not modified by visual perception, which is suppressed during saccadic movements [26]. Considering the open-loop nature of this movement, it is important for the brain to have both a good knowledge of the oculomotor plant and a good estimation of the target object location. Indeed, several adaptive mechanisms maintain the saccadic generator system calibrated [27]. The target of a saccade emerges from the interconnectivity between several cortical areas, such as the superior colliculus [28], the basal ganglia, the posterior parietal cortex, the frontal eye field (FEF), the cerebellum and the brainstem [29]. The FEF [30], core center of this network of areas, The FEF contains a retinotopically-organized map of *visual*, *visuo-movement* and *movement* cells [31]. While the *visual* cells respond to the onset of visual stimuli, the *movement* cells respond to the onset of a saccade,

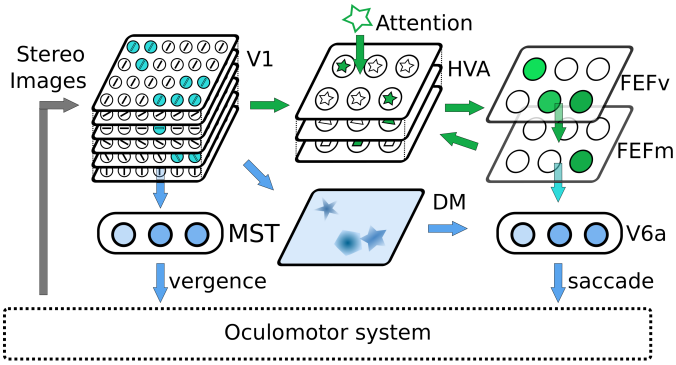

Fig. 2. The artificial vision architecture that simulates the brain's visual cortex. Green connections simulate the ventral stream while blue connection the dorsal stream. The arrow between the FEF and V6A represents the connection between the two parallel streams. Abbreviations: V1: primary visual cortex; HVA: higher visual area, which incorporates cells of areas V4 and IT; MST: medial superior temporal area; DM: disparity map; V6A: visuomotor medial posterior parietal area V6A; FEF*v* and FEF*m*: visual and movement cells in the frontal eye field.

and thereby encode expected landing position of the eyes after the saccade. The FEF is bidirectionally connected to area V4 (ventral stream) [32], [33] and to the lateral intraparietal area (LIP) in the dorsal stream. From these areas it receives information regarding visual features of the target (V4) and its spatial location (parietal regions as V6A [34]), and projects back its retinotopic position. These interactions are not limited to eye movements, but to reentrant processing in general, *e.g.* to the deployment of visual attention [11]. In this way, the visual system maintains a consistent object representation in all cortical areas (frontal eye field, ventral and dorsal streams).

The dorsal stream is also in charge of computing the sensorimotor transformations required to perform arm movements. These transformations, as those required to control the eyes, are likely to be performed through the gain field effect of neurons in the posterior parietal cortex. In particular, neurons in V6A area have been found to contextually encode different representations of the target position, allowing for easy reference frame transformations In V6A, neurons with retinotopically organized receptive fields are modulated by gaze direction in order to encode spatial positions [35], [36]. Moreover, V6A has proprioceptive properties, and is directly involved in the execution of reaching and grasping movements [37], in accordance with its role in reference frame transformations between eye-centered, body-centered, and hand-centered representations [12], [38], [39]. Once these frames of reference (f.o.r.) have been computed, gazing and/or reaching actions are executed by the basal ganglia and the cerebellum through the motor cortex.

III. Cognitive Architecture

This section describes the cognitive architecture that we have developed, inspired by the concepts described in the previous section.

The architecture is composed of several neural networks, each one modeling a brain's cortical region. The perceptive and proprioceptive information is encoded through a distributed approach by the response of the network's cells. This distributed representation is maintained and passed on from a module to another, and only when necessary, it is mapped into a closed loop motor command (vergence control), into the visual identity of the searched object (object recognition), or its spatial location (gazing and reaching).

Following this principle, instead of using an *artificial* Cartesian coordinate system, the position of target objects is maintained by a more *biological* coordinate system, consisting in the angular position of eyes and arm's joints. From this perspective, the robot becomes the measurement instrument that the neural architecture exploits to represent the surrounding space. The advantage of this *implicit* representation is two-fold. First, it is useful to avoid intermediate "decoding" stages, maintaining the system flexible for learning and robust to errors. Second, it is suitable to release motor commands and to calibrate the internal representation while the robot is interacting with the environment [19], [40].

Information flow in our schema follows the two pathways separation, downstream from the first processing module modeling V1. Now, we describe how we modeled different brain areas and their interconnections.

A. V1 Area

The computational model of area V1, grounding on the binocular energy model, encodes local, simple features of the visual stimuli like the orientation of edges [41], local contrast differences [42], and retinal disparity [43] (see Fig. 2). These features are available in a distributed representation [15] which provides a complete structural analysis of the visual signal [44].

On the basis of neurophysiological evidences [21], [22], the V1-like binocular complex cells [45] are modeled by the sum of the squared response of two quadrature pairs of simple cells. The model is conceived so that the response of the binocular energy unit models the mean firing rate of the complex cell [21]. Accordingly, each simple cell has a binocular Gabor-like receptive field characterized by a phase difference between the left and the right, that yields the sensitivity to a specific retinal disparity (for further details see Appendix B). In this way, the resulting complex cell is able to encode specific properties of the visual signal, defined by the disparity, size, orientation and shape of its component receptive fields.

We simulate a population of V1 simple and complex cells, divided into several sets, operating at different spatial resolutions and orientations of the stereoscopic images [15]. Each set, working at a specific spatial scale, is sensitive to a limited range of disparities, oriented edges and spatial frequencies. From this perspective, the implemented population, covering a sufficient number of oriented channels, phase shifts and frequency scales, yields a complete characterization of the local structure of the binocular visual signal. Such a representation is essential to gain a perception which is reliable (*i.e.* stable), dense and immune to lighting conditions, in order to ground the succeeding stages of perception and action at different levels of complexity.

B. Ventral Stream and Frontal Eye Field

The models of ventral stream and frontal eye field detect and localize the object of interest. Both models are a scaled version of a previously published anatomically and physiologically motivated model of attention, which were modified for processing real-world objects. Biological background can be found in [32]. All neurons use a rate coded model which describes the firing rate of a cell as its average spike frequency.

Objects are encoded as single views, i.e. as a specific visual appearance, in a *high visual area* (HVA). These view-tuned cells can be related to brain areas V4 and IT [23]. Using weight sharing, the HVA is organized in different retinotopic maps where each map encodes the feature selectivity (green shapes in Fig. 2). Given this retinotopic organization of the map, each cell of HVA is activated when a particular object is located at the retinal location underlying its receptive field [46]. Mimicking the V1 contrast response function [42], we used a binary step as activation function to saturate the response of the cells with an activation greater than a threshold. In such way, the response of the HVA neurons comes to be more robust to illumination changes. A single view of an object is encoded by the weights of the connection between V1 and HVA cells. Thus, HVA cells are selective for a specific patterns of V1 responses, i.e. for a specific pattern of oriented edges, disparities and local contrasts. With the aim of implementing an object recognition system able to recognize the objects by their 3D shape, color information is not used. These weights were learned during an off-line training phase using unsupervised learning. As this learning should lead to largely depth invariant object representations, our method relies on temporal continuity similar to those used to learn position invariance [47], [48].

In extension to common techniques in object recognition networks [49], we employed top-down and bottom-up processes for contextual feature enhancement. This strategy allows us to solve the dilemma of parallel segmentation and localization [46]: object segmentation depends on localization, that, in turn, requires the segmentation itself. When a particular object is searched, a top-down "feature-based attention signal" reinforces the activation of all HVA cells that encode a view of such an object [46] and suppresses the activation of the cells that encode unattended objects. This greatly reduces the false recognition due to similar views belonging to other objects. In our implementation, the suppressive field of an object contains the views of all other learned objects (Section V-B).

Spatial information is encoded in the frontal eye field (FEF module), simulated by two maps: FEFv indicates retinal locations of all objects (green dots in Fig. 2) whereby FEFm encodes only the saccadic target (single green dot in Fig. 2). Therefore the FEFv represents a perceptual map which is often referred to as a saliency map [50]. These maps represent the visual (FEFv) and movement cell (FEFm) types of the FEF [31], [32]. The former map is computed by choosing the maximum activation over all the features in HVA at each location. The latter one is calculated from FEFv by a Gaussian filter to reinforce neighboring locations and competition to suppress all others. This process is iterated until an area of activation exceeds a threshold which then triggers a saccade to foveate the searched object. The saccadic target is then used together with the disparity map to convert the visual location of the target into a gaze or an arm movement.

C. Dorsal stream

The dorsal stream detects the three dimensional structure of the scene using a representation that is suitable to support gazing and reaching actions.

The stream is composed of three modules, the computation of the disparity map (DM), used to gather a qualitative evaluation of the structure of the environment, the vergence control (VC) that reduces the local disparity of the central part of the visual field and brings the system within the working range of the DM, and V6A, which integrates the sensorimotor information in order to perform coordinated arm and eye movements towards given targets.

The horizontal disparity is commonly considered a highly informative cue for depth perception and vergence eye movements, both in neurophysiology [51] and computer and robot vision [52]. Working with a real and active robotic stereo head, the disparity pattern is not only composed of horizontal disparities as in the case of parallel optical axes [53]. The vertical component, arising from the vergent geometry and mechanical imprecision of the system, needs to be considered for a reliable and effective estimation of the horizontal one.

From this perspective, building a population of complex cells tuned to different spatial orientations is instrumental to recover the full disparity of the binocular image [15]. Exploiting the information encoded by the multi-frequency channels, the disparity estimation relies on a pyramidal decomposition combined with a coarse-to-fine refinement [44] (see also [54]). The features, obtained at a coarser level of the pyramid, are expanded and used to warp the spatially convolved images, on which the residual disparity is computed.

At the level of a single scale of frequency, we compute the component of the disparity along each orientation by applying a center of mass decoding strategy. Successively, the full disparity is obtained by an intersection of constraints [55], solving in such a way the aperture problem [56].

Since the disparity depends on the relation between the eye positions and the environment, the system might fall in a configuration where the disparity is outside the detectable range supported by the DM module. The VC is a fast active mechanism that works in a visual closed loop to bring and keep the disparity estimation module in its working range. Since a real-time behavior is needed, the module takes as input only the responses of the *complex cells* at an intermediate spatial frequency and directly converts them into effective vergence signals [17], [18] that trigger proper convergence or divergence of the robot's cameras (left side of Fig. 2).

The control signals are obtained by a weighted sum of the cell responses. The desired control should provide a sensitivity to the horizontal disparity and an insensitivity to the vertical one. The weights are computed by minimizing a functional that exploits the cells tuning curves (to the full vector disparity) to obtain the desired behavior [17]. The resulting VC signal

enables visually-guided eye movements that allow the robot to move the fixation point to the closest visible surface of the target object. The information conveyed by a coarse scale guarantees an effective trade-off between an adequate precision of the control and real-time performances. Moreover, relying on the distributed code of the disparity information, the VC is capable of providing an effective and stable signal even with noisy and changeable real world images, so as to cope with the imprecision of a real robot head [57].

In addition to vergence movements, the robot can perform gazing and reaching actions, that are voluntary, ballistic and usually goal-directed. The target is recognized and localized in the visual field by the *ventral stream* and is made available to the *dorsal stream* by means of the retinotopic FEFm map. In order to perform a correct gaze/reach movement, we need to solve the sensorimotor transformation problem, which in our application consists in converting the retinotopic (sensory information) position of the target into an eye/arm motor command.

We approached this problem by simultaneously maintaining the position of the target in three different f.o.r.: retinotopic, eye-centered and arm-centered. The retinotopic f.o.r. is defined by the location of the stimulus on the image (obtained from the FEF) and its disparity (obtained from the DM) (right side of Fig. 2). Instead of using a Cartesian space, we define the eye-centered and the arm-centered f.o.r. as motor spaces, which are determined by the angular positions of eye and arm joints, respectively. The angular positions of eyes and limbs are provided by the encoders of the robot, which replace the kind of information provided by proprioceptive cues in biological systems.

The transformations from one frame to another are computed by radial basis function networks (RBFN). This kind of network was chosen for its ability to approximate any kind of non-linear functions [58], which makes them especially suitable for the sensorimotor transformation problem. Therefore, RBFNs were used to simulate populations of V6A [12] neurons, and to simulate the gain modulation effect observed in the neurons of the parietal cortex [59], [60].The hidden layer of the RBFNs performs a non-linear transformation of the input and is then linearly combined to produce the output response. Herein, the hidden units are Gaussian functions with fixed parameters, while the weights of the linear combination are adapted on-line through a recursive least square algorithm [61]. In this way the robot can incrementally update its internal representation at each interaction with the environment by evaluating the error of the performed movement. The on-line algorithm provides the capability of modeling adaptive properties of the saccadic adaptation of human beings [62].

Taking into account that the error depends on the visual position of the target object or the robot hand, the learning stage is strongly dependent on the capability of the visual system to detect and localize such stimuli.

Summarizing, like in the cortex of primates, the integration between the dorsal and ventral pathways is necessary to develop complex behaviors in the robot. Indeed, vergence control (*dorsal stream*) reduces the range of the visual disparity to keep it in the working range of the object recognition system

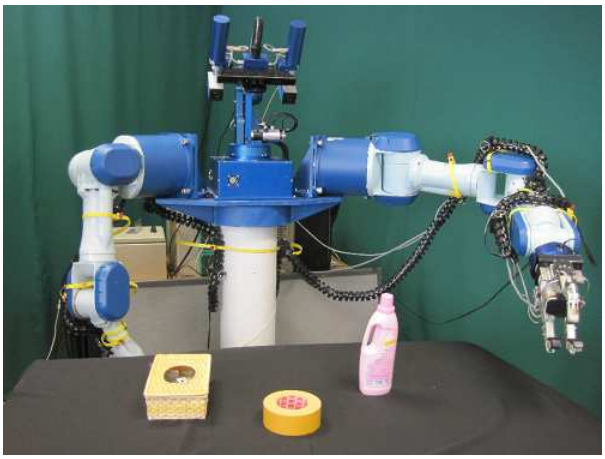

Fig. 3. The Tombatossals robot and its peripersonal workspace at the Robotic Intelligence Lab, Universitat Jaume I. Three objects have been employed in the experiments: a bottle, a box and an adhesive tape.

(*ventral stream*) and of the disparity computation. On the other hand, the retinal location of the object of interest, which is extracted by the interaction between HVA (*ventral stream*) and FEF, is directly involved in the execution of gaze and reach movements and in the subsequent calibration of the internal model (*dorsal stream*).

IV. Implementation

A. Robot

The cognitive architecture described in this paper is implemented and tested on the robotic torso *Tombatossals* (Catalan for *mountain-crasher*). The head is endowed with two cameras (resolution: 1024×768, frame rate: 30 fps) mounted at a baseline of $\approx$ 27 cm. In this work, the images acquired by the cameras were converted to gray scale and down-sampled to 320×240 pixels. We set the focal length of the cameras as short as possible ($\approx$ 5 mm) to obtain a broad field of view. In this condition, the images are strongly affected by lens distortion, however, since the proposed architecture is able to learn and to adapt from the visual signal, we do not compensate for it. The head is designed to have three degrees of freedom, *i.e.* a common tilt, and separate pan for the two cameras. This geometrical configuration allows for an independent control of gaze direction and vergence angle. The arms have seven degrees of freedom each, but we have employed three degrees only, *i.e.* shoulder pitch and roll, and the elbow yaw. Both the head and the arms are equipped with encoders that allow us to gain access to the motor positions with high precision.

B. Software Architecture

The implementation of the architecture described in the previous section is organized into several modules. Each module runs in parallel with the others and they interact with each other to produce behavior. We employ the YARP middleware to manage the communication among the modules [63] because it handles various types of communication interfaces,

such as TCP, UDP and shared memory, and can run on several operative systems. It is particularly suitable for "streaming" communication where the data sender is decoupled from the receiver, and it also supports "send/reply" communications where the sender and the receiver are tightly coupled.

We group the functional blocks that compose the system according to their scope: hardware interface, visual processing, memory and controller (see Fig. 4). The configuration of these blocks and their connection, is managed by a supervisor module, called *task manager* (TM), that changes the settings of the architecture by means of control messages that employ "send/reply" communications.

C. Functional blocks

Hardware interfaces: The *hardware interfaces* manage the communication with the input/output devices of the robot, i.e. robotic head, cameras and the manipulator arms. The camera modules acquire the images and provide them to the other modules. The modules connected to the arms and the head perform a bidirectional communication, as they provide the angular positions and the velocities of the joints, and simultaneously control the motors, by receiving a desired position/velocity from the control modules (blue blocks).

Visual Processing: The *visual processing* units described in Sec. III, perform image processing and operate in the retinotopic-domain. V1 encodes the low level visual features, and HVA extracts high level visual features and localizes the target on the image plane.

The V1 module elaborates the input images provided by the cameras to compute the complex cells population responses. This is composed of 280 retinotopic layers (5 scales $\times 8$ orientations $\times 7$ phases) that properly cover the 2D spatial frequency domain [44]. At each scale the resolution is halved with respect to the previous one, hence the finer scale is composed of $320\times240\times8\times7$ cells, and the coarser one of $20\times15\times8\times7$ cells.

The disparity map (DM) obtained is a dense map of the same size (320×240) as the input image, with sub-pixel resolution, that can be used both to interpret the structure of the scene, and to guess the combined saccade/vergence movement (for further details, see Appendix B).

The high-level visual area (HVA) binds together the complex cells' V1 responses to obtain the high-level features instrumental to represent the objects in the scene. These features are encoded by 10 retinotopically-organized layers. Each single HVA cell gathers input from all V1 layers in a spatial neighborhood of 72×72 pixels (first scale resolution) and from all the other cells of HVA. In addition to the cells of V1, the HVA can receive as input an internal model representation of a desired object, which is used as attentional signal to enhance the relevant views. The recursive loop that connects the HVA with the FEFv and FEFm maps is iterated until a threshold is exceeded in FEFm, hence until the desired object is detected. The centroid is then provided as the output of the module.

Memories: The representation of the objects is stored in a long-term memory, called *Object Memory* (OM), which binds the objects with their views in HVA. This memory is created

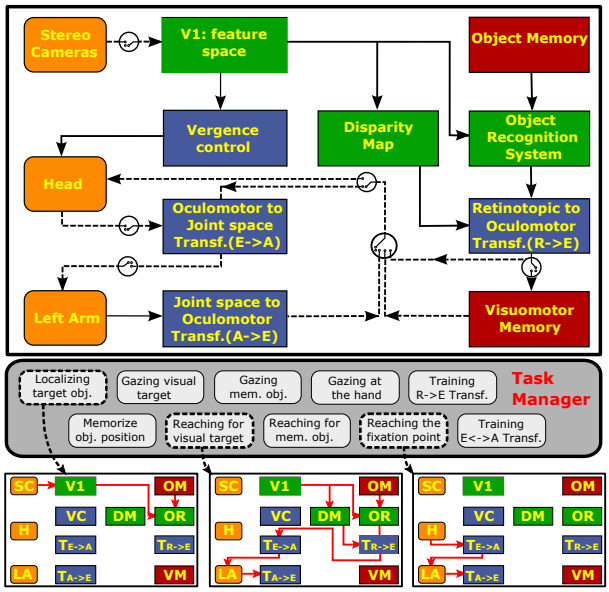

Fig. 4. The proposed control system. Modules are grouped by scope: *hardware interfaces* (orange), *visual processors* (green), *controllers* (blue) and *memories* (red). The data flow among the modules is ruled by the *task manager*, which employs *multiplexers* (displayed as circles) to enable and disable some connections (dashed line) depending on the required task. For example, for localizing a target object (bottom-left) V1 extracts low-level features from the binocular input and projects them to the object recognition system, which binds the features to create a high level description of the scene. Moreover, object recognition enhances the features that belong to the target object, so as to improve its localization. Once a target is localized, a sequence of transformations convert the retinotopic location of the target into a gaze direction and the latter one into an arm-joint configuration. Such a configuration is finally used to control the ballistic movement of the arm (center). On the right, the current gaze direction is converted into an arm-joint configuration which is eventually sent to the arm controller.

after the learning phase of HVA. By its connections to HVA it serves to guide visual search, as the top-down connections had been organized to activate the relevant HVA cells and to suppress those cells that are shared by distracting objects (see Sec. V-B). Thus, the top-down modulation from each OM unit consists of two binary vectors, each containing 10 entries that are linked to the corresponding 10 layers of HVA (Fig. 7). The OM cells receive as input the index of the target object.

The system is also equipped with a medium-term memory that keeps track of the positions of the objects in the head-centered frame of reference, namely *Visual-Motor Memory* (VMM): This module receives as input the head-centered position of a stimulus together with its view-based description. If the object has already been stored in the memory, its position is updated, otherwise it is added as a new object. The output of the module is the head-centered position of a selected target.

Controllers: The *controllers* allow the robot to close the perception-action loop by transforming their input cues into suitable motor commands.

The vergence control block converts the population response of the V1 complex cells into a vergence command. Using a single mid scale limits the required computational load and allows real time performance (≈ 40 fps). The resulting control is both able to follow an object moving in depth and to bring the fixation point on the surface of the gazed object [57], *i.e.*

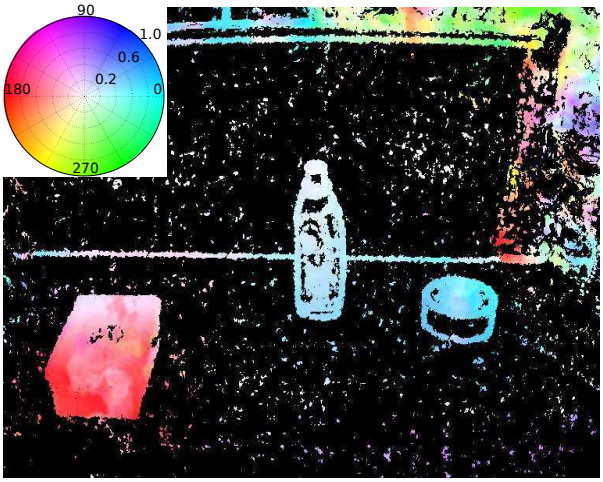

Fig. 5. An example of disparity map (in pixels) computed from the output of v1. The direction of the full vector disparity, arising from the vergence geometry (see [53]), is represented by the color, according to the color wheel. Red and cyan correspond to crossed and uncrossed horizontal disparities, while green and blue represent left and right vertical disparities. Seemingly, the magnitude is represented by the saturation, where a bright color stands for small disparities and a saturated color represents large ones. The robot is gazing at the top of the bottle in the middle of the image (zero disparity). The bottom of the bottle and the adhesive tape on the right are slightly farther (uncrossed disparity), whereas the box on the left is nearer (crossed disparity).

at zero disparity in a foveal area (see Fig. 5).

Three modules encode the sensorimotor transformations that convert the visual position of the target into an eye movement and an arm position, providing the reference positions for both the gazing and reaching movements. These transformations are $T_{R\mapsto E}$, $T_{E\mapsto A}$ and $T_{A\mapsto E}$. The $T_{R\mapsto E}$ module converts the retinal location of the saccadic target into the increment of the eye position that allows for gazing the target. The input is composed of the position of the target $\mathbf{x}$ in the left image and of its disparity $\boldsymbol{\delta}$, obtained by the DM. The transformation is encoded by a RBFN composed of 125 ($5\times5\times5$) Gaussian neurons that cover the input space $(\mathbf{x}, \boldsymbol{\delta})$ with a uniform distribution. The output of the network is the angular displacement of the motors (left pan, right pan and common tilt). This module, once it has received the visual position of the stimulus after the saccadic movement, updates and refines the encoded transformation on-line [40].

The $T_{E\mapsto A}$ module converts an eye position into the arm position which permits it to reach the fixation point, while $T_{A\mapsto E}$ allows the robot to gaze at its own hand, even though it may be out of the field of view. The gaze direction is encoded by the angular position of the eye motors (left and right pan, tilt), whereas the shoulder pitch and roll and the elbow yaw encode the arm position. Both transformations are encoded by a RBFN composed of Gaussian neurons uniformly distributed in the input space. The $T_{E\mapsto A}$ transformation employs 343 neurons ($7\times7\times7$) and the $T_{A\mapsto E}$ transformation employs 125 neurons ($5\times5\times5$). These modules update on-line the underlying transformation when the robot is looking at the hand [40], so that, at least for the learning process, visual feed-back is required.

The parameters of the networks and their performance (see section V-C) are reported in table II.

D. Dynamic connections and task manager

The implemented modules provide basic skills for interacting with the environment, even though individually they are not sufficient to perform a structured action. Indeed, the functional blocks need to be connected to create proper data flows that link perception to action. For example, if the robot recognizes an object, we can connect the $T_{R\mapsto E}$ block with the eye controller to gaze at it, as well as the $T_{R\mapsto E}$ and the $T_{E\mapsto A}$ blocks to the arm controller to reach it.

We introduced two kinds of modules in order to dynamically coordinate and configure the functional blocks: the *task manager* (TM) and the *multiplexer*. The multiplexer is a unit that receives multiple inputs and forward one or none of them to the output. The multiplexers can enable and disable connections on-line (dashed lines in Fig. 4) and they are placed in the system where multiple arrows enter the same block. The data flow implemented at each particular time depends on the selection command given by the TM. Indeed, it changes the configuration of the modules depending on the ongoing task, selecting the object of interest (acting on the memories), and changing the data flow (acting on the multiplexers).

Such an architecture allows the system to be flexible and adaptable, because a new action can be simply defined by its data flow (connections list). On the other hand, multiplexers avoid any interference or conflict among modules that send motor commands, because they forward one signal at a time. This makes the system more robust and maintainable.

In order to test the performance of the system, we created a collection of ten tasks that are shown in Fig. 4:

1) Localizing a visual stimulus (object or hand).
2) Training the visuo-oculomotor transformation ($T_{R\mapsto E}$).
3) Training the eye-arm coordination ($T_{E\leftrightarrow A}$).
4) Gazing a visual stimulus (object or hand).
5) Reaching for a visual stimulus (object or hand).
6) Reaching for the fixation point.
7) Gazing to the hand inside or outside the field of view exploiting the proprioception ($T_{E\leftrightarrow A}$).
8) Memorizing the position of a target object.
9) Gazing to a memorized objects' positions.
10) Reaching for memorized objects' positions.

When a gazing or reaching task is required, the TM selects the multiplexers that activate the appropriate input cues for the head or the arm controller. The vergence controller is always active in order to adjust the fixation point on the current target. When a gazing movement is ongoing, the stream of input images is interrupted in order to avoid the processing of inconsistent data. Moreover, the sensorimotor learning can be activated when the robot is looking at the hand ($T_{E\leftrightarrow A}$) or after a saccade towards a visual stimulus ($T_{R\mapsto E}$).

Bottom of Fig. 4 shows the connection of some tasks (tasks 1, 5, 6) that we used in the experiments. For example, in order to localize a target object, the TM connects the images acquired by the cameras with V1. On the left, V1 extracts low-level features and projects them to the high visual area, which binds the features to create a high level description of the scene. Finally, the HVA obtains the features that belong to the target object from the object memory, and enhances them

to localize the target. In order to reach a target, that can be either an object or the fixation point, the robot converts the eye-centered position of the target into an arm position that is used to control the joints of the limb. If the target is the fixation point, the target position is provided directly by the position of the eyes that are provided by proprioceptive cues (center of Fig. 4). If the target is an object, its position is provided by the conversion of the output of the high visual area (retinotopic-centered) into an eye-centered representation (right side of Fig. 4). The TM does not contain any cognitive skills to choose the action to execute, but receives commands given by a human user by means of a graphical user interface (GUI), or executes pre-defined tasks composed of sequences of elemental action components.

V. RESULTS

In order to assess the efficacy of the architecture in a real setup, we first tested separately the main modules (vergence control, object recognition and sensorimotor transformation), and then we tested the whole system with complex tasks.

The setup consisted of the robot platform with objects placed in its peripersonal space, *i.e.* within a reachable distance. The environment was composed of three objects placed on a table covered by a black cloth. The illuminance of the environment was approximately 465 lux and was originated mainly by fluorescent tube lights.

The three objects were a bottle, a box and an adhesive tape roll (Fig. 3) and were chosen because of their different three-dimensional structure. During the experiments, the objects were arbitrarily placed on the table, in a working area of about (50×50) cm^2, so that the farthest object was still reachable by the arm.

In the first test, we show the capability of the vergence control which is mandatory for the correct functioning of the other modules. Indeed, a correct fixation posture allows the V1 to work within its working range, ensuring a reliable information to the sensorimotor transformation and to the HVA module. The second test shows how the robot discriminates and localizes the three objects. This experiment involves V1, which provides a low level description of the scene, and HVA, which recognizes the target/selected object based on its view. The third test shows how the sensorimotor framework adapts the internal model during the interaction with the environment. This experiment involves V1 and HVA, which provide the location of the target and V6A that learns the transformations among the different coordinate systems (retinotopic-oculomotor-arm joint space). Finally, the last test shows how the connections among the modules provide the robot with the capability of achieving complex and articulated tasks in a real setup.

A. Vergence Movements

To verify the precision of the vergence movement we implemented a test in which the robot gazes towards a frontoparallel plane at a fixed reference vergence of 8° ($\approx$ 1000 mm distance). The fixation point is moved trial by trial to a random position within $\pm4^\circ$ with respect to the reference vergence

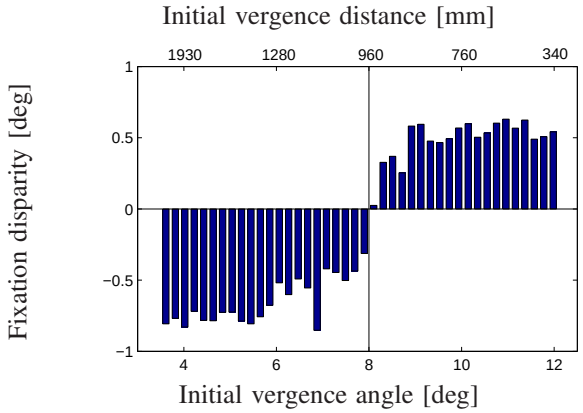

Fig. 6. Precision of the vergence control: the absolute error on the positioning of the fixation point, *i.e.* the fixation disparity, is plotted against its initial vergence angle and distance.

(*i.e.* $\approx [600-2000]$ mm), and we expect the vergence control to move it back onto the surface of the plane, accordingly. The vergence error is measured as fixation disparity, *i.e.* the residual binocular disparity at fixation, in order both to have a measure independent of the estimated motor position, and thus repeatable and reliable, and to establish a direct comparison with the human behavior. Fig. 6 shows the fixation disparity plotted against the initial vergence angle.

The vergence control, tested over 250 trials, shows a mean precision on the positioning of the fixation point of $0.6\pm0.17^\circ$, which is comparable to the magnitude of the actual fixation disparity measured in humans [64]. Moreover, such an error is almost constant and positive (0.5°) for diverging movements (plane farther with respect to the initial fixation point), whereas for converging movements is dependent on the starting position ($0.3 \sim 0.7^\circ$), showing a marked asymmetry of the vergence behaviour with respect to the initial position. From a mechanical point of view, this behaviour might be explained by the limited motors' sensitivity to small movements. Besides, the results obtained directly resemble the behavior observed in humans regarding the relation existing between fixation disparity, and the convergent/divergence dynamics in human eye movements (cf. our Fig. 6 with Fig. 6b in [64]), strengthening the validity of the vergence model.

B. Object recognition

The object recognition module provides the location of a target object for the other modules to enable complex behavior like reaching and the execution of saccades. In this section, we will first explain the learning procedure, then we test the module regarding recogntition accurancy, location accurancy, and robustness against distractors, and finally we discuss the general applicability of the object recognition approach.

Learning: The system was trained unsupervised and offline to create the internal representation of the considered objects ('Box', 'Bottle' and 'Tape'). We created a stereo image set (training data) which includes those objects under all considered transformations, i.e. different scales and disparities. Each object was placed alone in the scene and moved on the table to five different positions in depth. Additionally, the fixation

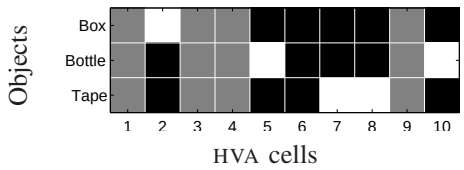

Fig. 7. Representation of the object memory for the three objects of the dataset. The relevance (white cells) and suppression (black cells) of each view-tuned HVA cell (y-axis) is plotted for each object id (x-axis).

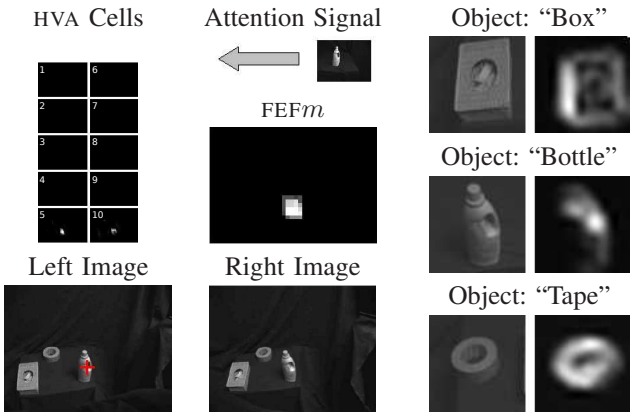

Fig. 8. (a) Example of a successful object recognition and localization in a typical scene. Each box in HVA (higher visual area) shows the activity of a single view in image coordinates. The system has to search for the target object 'bottle', so to focus the attention to the views 5 and 10. FEF (frontal eye field) encodes the spatial position of the object. The red cross (displayed only for left image) marks the position of the target object after the recognition. (b) Area HVA encoding some views of objects. For each object (left), the weights V1 $\rightarrow$HVA (right) of one exemplary HVA cell are illustrated as the maximum over all V1 features (i.e. phase, orientation, disparity) at a certain position. Brightness denotes weight strength.

point was set to 50 different fixation points on a cubic grid that was aligned on the table (5 points horizontally, 2 points vertically and 5 in depth). For each of these combinations, a stereo image pair was captured from the robotic cameras, providing 250 pairs. To avoid a learning process for each retinal position, we learned only in a single retinotopic location and we subsequently shared the weights with all the other locations, so to create a full map of HVA cells. This is practically achieved by cropping out the target object in each image, which creates a new 'cut' image data set on which learning took place.

We trained the system using a trace learning algorithm, which in turn, grounds on temporal continuity [47], [48]. Such an approach assumes that on average, movements of the target and saccades in the vicinity of the target are more likely than saccades to different objects. Thus we created an image sequence that resembles the temporal behavior of the retinal image stream: the object position changes randomly every 50 ms, the fixation point every 250 ms and the object type every 37.5 s. The training was performed off line on the image sequence until all objects were perfectly recognized on the training scenes. This required about $2 \cdot 10^5$ presented images or 10^4s simulation time. From this sequence, the system learns a population of view-tuned HVA cells that encode the statistically significant information of a certain view of an object (Fig. 8(b)).

The association of HVA cells to a certain object was designed via a manual mapping. This process provides both a feature-based memory, denoting which HVA cells (and hence which views) belong to a certain object, and a suppression field indicating which views of other objects could be confused with the current view and should be suppressed (Fig. 7).

Additionally to the three objects, the robot learns a representation of the hand that is used to avoid a false recognition of other objects. Indeed, as the system has the ability to avoid an object via the suppression field, the robot was instructed to ignore the visual input originating from its own hand.

Recognition results: The recognition process, as used in task 1, exploits visual attention to enhance the features belonging to the target object and to suppress those belonging to other objects in HVA (Fig. 8(b)). The loop through HVA, FEFv and FEFm processes the spatial scene information which finally results in a spatially selective signal in FEFm indicating the saccade target (section III-B and Appendix D).

The discriminative ability of the object recognition system was tested on a separate test set (denoted test set 1) of 27 scenes (Fig. 8(a) shows one example), each one containing three objects which were recognized individually, resulting in 81 object discrimination and localization tasks. This test set was separately created to ensure independence from the training data. Compared to the trainings data, the test set contains different objects positions, scales and disparities as object and fixation locations are chosen differently. In the test set, both are chosen completely arbitrary, whereby in the trainings set, both are chosen from a small set of grid positions as described in the learning procedure. Hence, the module has to generalize from fixed training samples to arbitrary selected ones. Object rotation was kept identical between training and test set as rotation invariance was not the focus of the work. As result, the system's object discrimination rate was 100%, while the localization rate was about 96% (see Tab. I). Localization is rated as correctly if the saccadic target point is located within the object borders. Concerning the maximum amount of mislocalization, the distance from the saccadic target point to the object border was 20 pixels (the image size was 320×240 pixels).

To further evaluate the robustness of the system against arbitrary distracting objects and different object scales, we introduced an additional unknown object in the test scene and we place arbitrary targets farther away. We created a new test set (test set 2) containing eight arbitrary distracting objects ('apple', 'bottle 2', 'car', 'drill', 'hammer', 'tea box', 'dumbbell', 'wooden figure') in four random spatial arrangements, resulting in 32 scenes and 96 recognition tasks. As result, the overall discrimination rate drops from 100% to 82% (Table Ib). We examined the misrecognitions: at first, errors resulted from the global normalization in V1, introducing incorrect V1 responses in cases of very salient distractors. At second, being the objects much smaller with respect to the training set, they were not recognized (expected as explained in the following discussion). At third, the recognition failed in case the target was very similar to the unknown object as the suppression field cannot be used for unknown objects.

TABLE I
A) IN TEST SET 1, LOCALIZATION RATES IN % AND MAXIMAL MISLOCALIZATIONS IN PIXEL ARE DENOTED FOR EACH OBJECT. B) IN TEST SET 2, THE DISCRIMINATION ABILITIES IN % ARE ILLUSTRATED BY A CONFUSION MATRIX ('D': DISTRACTOR). THE COLUMN DENOTES THE TARGET AND THE ROW THE DETECTED OBJECT.

(a)

Object	Rate	Mislocal.
Box	96.3	7
Bottle	96.3	20
Tape	96.0	6

(b)

Object	Box	Bot.	Tape	D
Box	94	0	0	6
Bot.	19	78	0	6
Tape	6	13	75	6

Discussion: In general, we proved that the object recognition module learned to recognize an object under many different transformations, due to its view-based representation (stage HVA) [46]. A view represents a specific visual appearance of an object and when a transformation changes the visual appearance, a different view cell is activated. The learning algorithm creates view cells for all transformations included in the learning set. As this recognition approach is completely independent of the kind of the transformation, it is in principle able to work for arbitrary ones. Here, we showed the approach coping with scaling/disparity transformations by including only these in the learning set. Other ones like rotations were not included, but could be easily added. In summary, the proposed system provides spatial invariance grounded on weight sharing, and scaling/disparity invariance grounded on the view-based representation.

In our setup, we simplified the environment by using three specified objects, only. This setup ensures that the object recognition module provides a very reliable target location to the other modules, thus allowing us to evaluate complex behaviors, such as reaching and gazing, independently of the object recognition performance. Previous work successfully used the same approach for more (ten) and harder to distinguished objects [46], hence we expect that a larger number of object can be supported. To further demonstrate the general validation of the solution, we additionally evaluated its robustness against unknown arbitrary distractors and we still obtained 82% accuracy. The main limitation of the module is that the approach can only cope with those transformations and objects that appear in the training set. Also, a deep evaluation of the object recognition approach is not reported here as this was not the focus of the work.

C. Learning the sensorimotor control

This experiment shows how the system can learn the associations among the visual location of the stimulus, the gazing[1] direction and the reaching position. These associations are encoded into three sensorimotor transformations, one that converts the retinotopic position of the target into an eye movement ($T_{R\mapsto E}$) and two involved into the eye-arm coordination ($T_{E\leftrightarrow A}$). Learning is the normal behavior of the agent that adapts the sensorimotor transformations after each movement. This process is self-supervised, since the robot compares the outcome of the performed movement with the predicted one and uses the mismatch to correct its internal model.

[1]In this work the head is kept still, so we use the terms "gaze" and "saccade" interchangeably.

With respect to our previous work on sensorimotor transformation [40] we substituted the delta rule algorithm with a recursive least square. This learning technique guarantees faster convergence of the learning process, so that, the sensorimotor transformations can be learned from scratch on the real robot, instead of bootstrapping it with simulated data.

Learning the visuo-oculomotor transformation: The learning of the visuo-oculomotor transformation $T_{R\mapsto E}$ (task 2) demonstrates the cooperation among visual attention, object recognition, depth estimation and sensorimotor learning. This behavior consists in the following sequence of tasks: 1) select and localize a target object (task 1); 2) compute the $T_{R\mapsto E}$ transformation and trigger a saccade to attempt to foveate it (task 4); 3) localize again the target (task 1); 4) use the visual displacement of the target due to the saccadic movement to train the transformation (task 2). Learning takes place after each saccade, so to allow the robot to keep up-to-date its internal model. The weights of the networks were initialized to zero and we instructed the robot to perform 2544 saccades from random starting position of the eyes. After 500 iterations, the mean visual error after a saccade was around 1 pixel and the system brought the target in the center of the image (≈ 2.5 pixels) with just one saccade in the 90% of the trials.

In order to have a quantitative measure of the performance, we stopped the on-line learning and we test the system on 500 testing saccades. The movement error is shown in table II while the visual error is $1.02\pm\ 0.84$ pixels, which is comparable with the performance of other architectures ($2.5/(512\times 512)$ pixels reported by Bruske et al. [65]; $5.5/(640\times 480)$ pixels reported by Forssén et al. [66]).

Learning eye-arm coordination: Once the $T_{R\mapsto E}$ is learned, it is exploited to learn the $T_{E\leftrightarrow A}$ transformations (task 3) that deals with the coordination of the eyes and the arm, and adapts when the robot is gazing at the hand. To test the learning capabilities of the network, we initialized the weights of the $T_{E\leftrightarrow A}$ transformation to zero and we executed a learning behavior 7152 times. The behavior used to train the eye-arm coordination is the following: 1) move spontaneously the arm (motor babbling); 2) localize the hand using a visual marker (task 1); 3) gaze at the hand using $T_{R\mapsto E}$ (task 4); 4) train the direct and inverse transformations ($T_{E\leftrightarrow A}$). After the on-line training phase, we tested the abilities of gazing at the hand (task 6) and reaching for the fixation point (task 7) without employing information from the vision system. The Euclidean distance between the desired joint position and the computed one was $0.28\pm 0.27°$for $T_{A\mapsto E}$ and $2.78\pm 3.57°$ for $T_{E\mapsto A}$. We also stored the data acquired during the on-line exploration to test the system using the K-Fold Cross validation (K = 5). We observed similar quantitative results ($0.29\pm\ 0.29°$ for $T_{A\mapsto E}$ and $2.84\pm\ 3.61°$ for $T_{E\mapsto A}$) that demonstrate the generalizing capabilities of the networks.

Grasping the bottle: We designed a grasping setup in order to test the performance of the system on a real world scenario. The behavior is the following: 1) recognize and localize the bottle (task 1); 2) gaze at the visual target (task 4); 3) reach for the fixation point (task 6). Once the movement is terminated, the robot closes the hand and the outcome of the action is evaluated. The trial is marked as successful if the robot

TABLE II
PARAMETERS OF THE RBFNS AND THEIR PERFORMANCE. THE RADIUS OF THE GAUSSIAN IS REPORTED WITH RESPECT TO THE NORMALIZED INPUT. THE ERROR IS THE EUCLIDEAN DISTANCE EXPRESSED IN DEGREES.

Transf.	RBFN par.		N. points	Error[degree]	
	centers	radius(σ)		μ	σ
$T_{R\mapsto E}$	$5\times5\times5$	0.3	500	0.2	0.23
$T_{A\mapsto E}$	$5\times5\times5$	0.3	7152	0.29	0.29
$T_{E\mapsto A}$	$7\times7\times7$	0.22	7152	2.84	3.61

correctly grasps the bottle, or as failure otherwise.

For this experiment, tactile feedback was not used during grasping, in order to better assess the quality of the visual processing, and we have chosen the bottle as a target object because its symmetric shape reduces randomness effects, allowing for an easier statistical assessment of the system. The bottle was placed on a grid of 3 by 4 points that covered an (X, Z) region of 75×60 cm in front of the robot. The arm began each movement from a "home position" that allowed us to reach for the bottle without any collision. During the training of the $T_{E\leftrightarrow A}$ transformations, a marker was put in the center of the hand, so we expected that a correct arm movement would bring the center of the hand near the most salient feature of the bottle (near the top). The robot grasped correctly the bottle 11 times out of 12. The single failure happened because the bottle was positioned at the boundary of the training space and slightly too far to be grasped. However, by moving the bottle 2 cm toward the robot, it managed to grasp it correctly.

D. Complex behaviors

Grounding on the tasks defined in Sec.IV-D, we illustrate the performance of the integrated architecture in order to represent the capabilities of the system in interacting with the environment. Fig.9 shows the left camera image, the disparity map and the FEFm map for a sequence of tasks performed by the robot:

- A) the robot starts fixating to a "random" central point of the table and the VC controls the eyes to minimize the overall disparity (red cross). The robot is not gazing any specific object, and is asked to localize the box. The DM gives a clear hint of the three-dimensional structure of the three objects, while the FEFm shows the location of the upcoming saccade towards the recognized target object.
- B) the robot correctly gazes the previously localized object (A), and uses the FEFm to effectively verify if the box is in center of the visual field. After the movement, the visual displacement of the box is used to train the $T_{R\mapsto E}$ transformation (task 2).
- B-D) the robot localizes the three learned objects: the box (B), the tape (C) and the bottle (D). Regarding the effectiveness and the capability of the OR, it is worth noticing two aspects. First, even if the tape is only partially covered by the HVA cells (see FEFm), the module is equally able to recognize and localize it, and second, the bottle is recognized by its three-dimensional most salient and recognizable feature, *i.e.* the cap (Fig. 8(b)).
- E) the robot shows the effectiveness of the $T_{R\mapsto E}$ transformation in cascade with the $T_{E\mapsto A}$, reaching for an object that is not in the fixation point.
- F) the robot gazes a memorized object (the bottle), whose position was localized in (E).
- G) the robot, once that is has correctly reached the bottle, shows the capability of the $T_{A\mapsto E}$ transformation, gazing its own hand. Moreover, when the robot is reaching to and gazing at the same spatial location, the system updates the $T_{E\leftrightarrow A}$ transformations (task 3).

The behaviour of the robot accomplishing the tasks of localizing, gazing and reaching visual targets can be also seen in the video [67].

VI. DISCUSSION AND CONCLUSIONS

A. Comparison with the State of the Art

In this work, we have integrated different models of the visual and visuomotor cortex into a unified robotic framework.

Shibata et al. [7] also proposed a biologically plausible simulation of the oculomotor system, however they did not consider an active control of the vergence angle. Paying attention to the simulation of the saccadic system, they mainly focus on the generation of the eye trajectory given the target position, whereas our module generates the sensorimotor transformation to locate the target in the joint space. Therefore, from this point of view, the two approaches are complementary. More precisely, they compute the eye trajectory for a single eye and exploit the same trajectory for the other one (conjugate movement). Conversely, by including a vergence component, we obtain two different trajectories for the two eyes (conjugate and disconjugate movement). Finally, they obtain a saliency map [50] by a pure bottom-up processing that enhances the moving stimuli, whereas our cognitive architecture employes both bottom-up and top-down processing in the FEFv-HVA network.

In a recent work based on developmental theories, Mc Bride et al. [9] learned eye-arm coordination by means of the visuo-oculomotor [68] and oculo-arm motor transformations [69]. Both transformations were encoded by a topological map based on the nearest neighborhood [68], [69]. On the contrary, we approximate each transformation with a continuous function in order to obtain higher precision with a smaller amount of resources. Moreover, we describe the target arm position in the arm-joint space (implicit representation) whereas Lee et al. [68] [69] used the task space (explicit representation). The visual features were localized on the basis of a bottom-up saliency map and the relevance of previously gazed stimuli were reduced with an inhibition of return (IOR) mechanism [9]. The concurrent top-down and bottom-up processes, which we employ to generate the saliency map, can be extended to implement the IOR by adding a suppressing factor in the top-down signal that depends on the head-centered position of the gazed stimuli.

In the field of integrative robotic solutions, another active vision system is presented by Rasolzadeh and Björkman [4]. Even if their system is not bio-inspired, it shows capabilities similar to the ones of our architecture. Contrarily to our

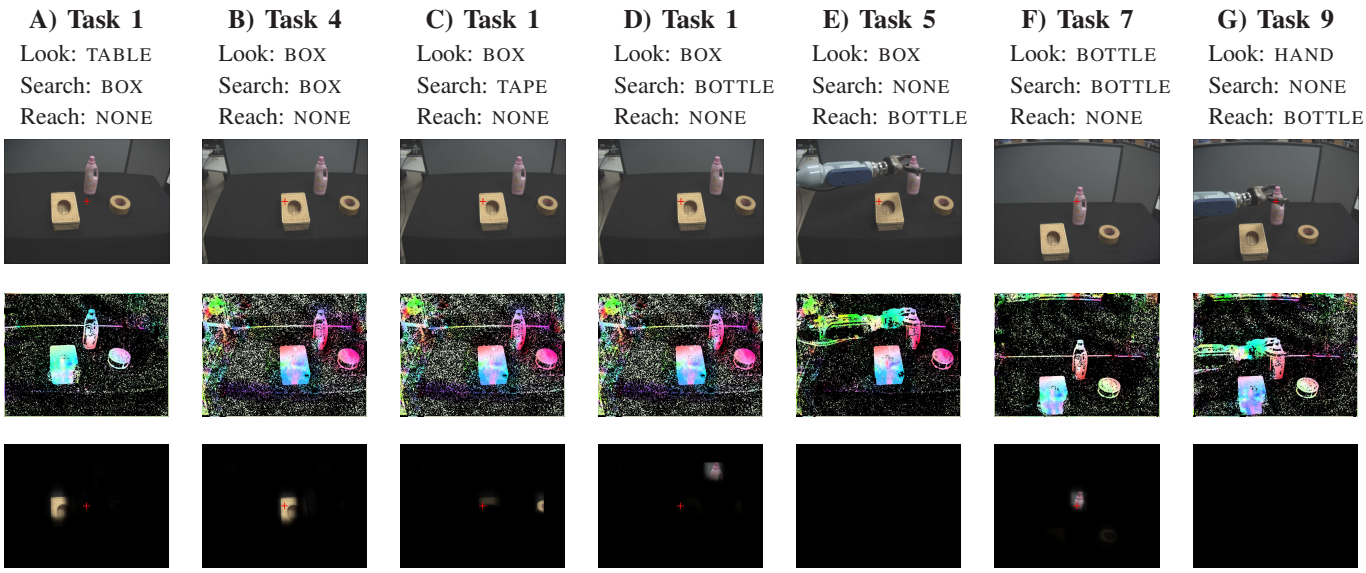

Fig. 9. Examples of complex tasks accomplished by the robot. For each task the panels show the left image (up), the disparity map (center) and the gazing target provided by the FEFm (bottom). For understanding the meaning of the disparity map, please see Fig. 5.

approach based on a visual front-end, their approach employs different representations depending on the task at hand. Indeed, they employed Harris' corner to calibrate the visual system, size and hue histogram to extract the saliency map, color and scale-rotation invariant features (SIFT, [70]) to recognize objects.

Rasolzadeh and Björkman encoded the objects by a color histograms and SIFT. Given that SIFT is based on the local maximum of Gaussian filters [70], their representation is roughly comparable to the one provided by the cells in the early area V1. As the calibration of the robotic system is concerned, they calibrate on-line the extrinsic parameters of the cameras by fitting a linear function, and assuming known the intrinsic camera parameters. We obtain the position of the cameras by proprioception and we train this subsystem on-line using a non-linear transformation, which not only copes with, but implicitly encodes both the intrinsic and extrinsic parameters of the cameras, thus avoiding the need for explicit calibration (see Section V-C).

B. Conclusions

The proposed architecture results in a repertoire of behaviors such as gazing and reaching target objects. More structured and effective behaviors are built on the integration of the different modules, which can work separately or cooperate together.

At the level of each single module, the information, both visual and proprioceptive, is encoded in a distributed way. Module V1 effectively encodes visual information both for *early* and *advanced* tasks. The flexibility of this approach allows us to exploit the population response at different levels of complexity, either directly for vergence control in an (early) visual closed loop, or to evaluate the depth map of the environment, or to build a hierarchical representation of the view and structure of a single object (advanced). Regarding the spatial representation of the three-dimensional space that surrounds the robot, the choice to implicitly encode it by the (distributed) proprioceptive information of both the eyes and the arm joints, provides an effective capability of interaction and a more immediate mapping among the different reference frames. All along the neural paths, the information is maintained distributed, postponing the decision as long as possible, in order to maintain a flexible use of visual and proprioceptive data.

At the level of the integrated architecture, avoiding a Cartesian frame and representing the peripersonal space by proprioceptive cues, allows for a continuous on-line learning in which the ventral stream ("vision for perception") is instrumental in calibrating the dorsal stream ("vision for action"). Thus, the cross-talk between the two streams can be considered at the base of "action for vision". In fact, without any *a priori* knowledge of the optical and geometrical characteristics of the robot, the architecture learns an implicit representation of the surrounding space and works out how to appropriately interact with it. Differently from a standard computer vision approach, our approach does not require an explicit calibration of the robot's parameters because they are learned implicitly by the system and are *embodied* in it.

Our contribution to the computational and experimental neuroscience communities has been presented elsewhere [15], [17], [19], [46], [62], [71]. Here, we have focused on the technological benefits obtained by implementing such models on a robotic platform. In previous studies, biological inspiration allowed us to create adaptive and robust algorithms; here, the integration of a set of these modules allows us to create a highly structured and coherent cortical architecture. Maintaining biological inspiration even at the integration level, makes the robotic agent able to accomplish complex tasks, like perceiving and interacting, in a continuous and mutual adaptation of perception with action, and vice-versa.

C. Future work

The hierarchical and modular organization of the presented cognitive architecture allows us to easily adapt it to new

tasks by adding, rewiring or modifying single modules in the system. Considering the cross-talk between the ventral and the dorsal stream, in a future development of the work the system will exploit the OR to recognize and localize its own hand, without using a visual marker, so to train the $T_{E \leftrightarrow A}$ transformations. From the perspective of the eye movements, we will implement an on-line learning procedure of the vergence control [72], [73], and we will include a module for the control of smooth pursuit movements [74]. The proposed improvements, grounding on the same V1 architecture and on RBFs, respectively, are suited for a direct integration within the current hierarchical architecture. This will provide the system with new functionalities for more complex tasks, such as reaching for a moving object. From the perspective of the overall working capabilities of the robotic agent, the simplicity by which a new behavior can be created, i.e. by enabling new connections among modules, allows us to employ it in cognitive science and in human-robot interaction experiments [62], [75].

REFERENCES

[1] M. Bertero, T. A. Poggio, and V. Torre, "Ill-posed problems in early vision," *Proceedings of the IEEE*, vol. 76, no. 8, pp. 869–889, 1988.

[2] J. Aloimonos, I. Weiss, and A. Bandyopadhyay, "Active vision," *Int. J. Comput. Vision*, pp. 333–356, 1988.

[3] F. Wörgötter, A. Agostini, N. Krüger, N. Shylo, and B. Porr, "Cognitive agents: a procedural perspective relying on the predictability of Object-Action-Complexes (OACs)," *Robot. Auton. Syst.*, vol. 57, no. 4, pp. 420–432, Apr. 2009.

[4] B. Rasolzadeh and M. Björkman, "An active vision system for detecting, fixating and manipulating objects in the real world," *Int. J. Robot. Res.*, vol. 29, no. 2-3, pp. 1–40, 2010.

[5] P. Azad, T. Asfour, and R. Dillmann, "Stereo-based 6d object localization for grasping with humanoid robot systems," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2007, pp. 919–924.

[6] S. Calinon, F. Guenter, and A. Billard, "On learning, representing, and generalizing a task in a humanoid robot." *IEEE T. Syst. Man Cy. B*, vol. 37, no. 2, pp. 286–98, Apr. 2007.

[7] T. Shibata, S. Vijayakumar, J. Conradt, and S. Schaal, "Biomimetic oculomotor control," *Adapt. Behav.*, vol. 9, no. 3-4, pp. 189–207, 2001.

[8] H. Hoffmann, W. Schenck, and R. Möller, "Learning visuomotor transformations for gaze-control and grasping," *Biol. Cybern.*, vol. 93, no. 2, pp. 119–130, 2005.

[9] S. McBride, J. Law, and M. Lee, "Integration of active vision and reaching from a developmental robotics perspective," *IEEE Trans. Auton. Ment. Dev.*, vol. 2, no. 4, pp. 355—366, 2010.

[10] B. Grzyb, E. Chinellato, A. Morales, and A. P. del Pobil, "A 3d grasping system based on multimodal visual and tactile processing," *Ind. Robot*, vol. 36, no. 4, pp. 365–369, 2009.

[11] F. Hamker, "The emergence of attention by population-based inference and its role in distributed processing and cognitive control of vision," *Comput. Vis. Image Und.*, vol. 100, no. 1-2, pp. 64–106, Oct. 2005.

[12] E. Chinellato, B. J. Grzyb, N. Marzocchi, A. Bosco, P. Fattori, and A. P. del Pobil, "The dorso-medial visual stream: From neural activation to sensorimotor interaction," *Neurocomp.*, vol. 74(8), pp. 1203 – 1212, 2011.

[13] M. A. Goodale and A. D. Milner, "Separate visual pathways for perception and action." vol. 15, no. 1, pp. 20–25, Jan. 1992.

[14] M. Goodale and D. Westwood, "An evolving view of duplex vision: separate but interacting cortical pathways for perception and action," *Curr. Opin. Neurobiol.*, vol. 14, no. 2, pp. 203–211, 2004.

[15] M. Chessa, S. P. Sabatini, and F. Solari, "A fast joint bioinspired algorithm for optic flow and two-dimensional disparity estimation," in *ICVS*, 2009, pp. 184–193.

[16] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Robust object recognition with cortex-like mechanisms," *IEEE T. Pattern Anal.*, vol. 29, no. 3, pp. 411–426, 2007.

[17] A. Gibaldi, M. Chessa, A. Canessa, S. P. Sabatini, and F. Solari, "A cortical model for binocular vergence control without explicit calculation of disparity," *Neurocomp.*, vol. 73, pp. 1065–1073, 2010.

[18] Y. Wang and B. Shi, "Improved binocular vergence control via a neural network that maximizes an internally defined reward," *IEEE Trans. Auton. Ment. Dev.*, vol. 3, no. 3, pp. 247 –256, 2011.

[19] E. Chinellato, M. Antonelli, B. Grzyb, and A. P. del Pobil, "Implicit sensorimotor mapping of the peripersonal space by gazing and reaching," *IEEE Trans. Auton. Ment. Dev.*, vol. 3, pp. 45–53, Jan 2011.

[20] B. Girard and A. Berthoz, "From brainstem to cortex: computational models of saccade generation circuitry," *Progress in Neurobiology*, vol. 77, no. 4, pp. 215–251, 2005.

[21] I. Ohzawa, R. Freeman, and G. DeAngelis, "Stereoscopic depth discrimination in the visual cortex: Neurons ideally suited as disparity detectors." *Science*, vol. 249, pp. 1037–1041, 1990.

[22] S. Prince, B. Cumming, and A. Parker, "Range and mechanism of encoding of horizontal disparity in macaque v1." *J. Neurophysiol.*, vol. 87, pp. 209–221, 2002.

[23] N. K. Logothetis, J. Pauls, and T. Poggio, "Spatial reference frames for object recognition. tuning for rotations in depth," in *AI Memo 1533, Massachusetts Institute of Technology*. MIT Press, 1995.

[24] A. Takemura, Y. Inoue, Kawato K., C. Quaia, and F. Miles, "Single-unit activity in cortical area mst associated with disparity-vergence eye movements: Evidence for population coding." *J. Neurophysiol.*, vol. 85(5), pp. 2245–2266, 2001.

[25] J. T. Enright, "Changes in vergence mediated by saccades." *The Journal of Physiology*, vol. 350, pp. 9–31, 1984.

[26] E. Castet and G. S. Masson, "Motion perception during saccadic eye movements," *Nature Neurosci.*, vol. 3, no. 2, pp. 177–183, 2000.

[27] H. Deubel, "Separate adaptive mechanisms for the control of reactive and volitional saccadic eye movements." *Vision Res.*, vol. 35, no. 23-24, pp. 3529–3540, Dec. 1995.

[28] D. Munoz and R. Wurtz, "Saccade-related activity in monkey superior colliculus. I. Characteristics of burst and buildup cells," *Journal of Neurophysiology*, vol. 73, no. 6, pp. 2313–2333, 1995.

[29] U. Büttner and J. A. Büttner-Ennever, "Present concepts of oculomotor organization." *Progress in brain research*, vol. 151, pp. 1–42, Jan. 2006.

[30] C. J. Bruce and M. E. Goldberg, "Primate frontal eye fields. I. Single neurons discharging before saccades." *J. Neurophysiol.*, vol. 53, no. 3, pp. 603–35, Mar. 1985.

[31] J. D. Schall, "Neuronal activity related to visually guided saccades in the frontal eye fields of rhesus monkeys: comparison with supplementary eye fields." *J. Neurophysiol.*, vol. 66, no. 2, pp. 559–79, Aug. 1991.

[32] F. H. Hamker, "The reentry hypothesis: the putative interaction of the frontal eye field, ventrolateral prefrontal cortex, and areas V4, IT for attention and eye movement." *Cereb. cortex*, vol. 15(4), pp. 431–47, Apr. 2005.

[33] S. Szczepanski and Y. Saalmann, "Human fronto-parietal and parieto-hippocampal pathways represent behavioral priorities in multiple spatial reference frames," *Bioarchitecture*, vol. 3, no. 5, pp. 147–152, 2013.

[34] C. A. Buneo, M. R. Jarvis, A. P. Batista, and R. A. Andersen, "Direct visuomotor transformations for reaching." *Nature*, vol. 416, no. 6881, pp. 632–6, Apr. 2002.

[35] 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, 2008.

[36] A. Bosco, R. Breveglieri, E. Chinellato, C. Galletti, and P. Fattori, "Reaching activity in the medial posterior parietal cortex of monkeys is modulated by visual feedback," *The Journal of Neuroscience*, vol. 30, no. 44, pp. 14 773–14 785, 2010.

[37] C. Galletti, D. 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(2), pp. 158–170, 2003.

[38] P. Fattori, R. Breveglieri, V. Raos, A. Bosco, and C. Galletti, "Vision for action in the macaque medial posterior parietal cortex," *The Journal of Neuroscience*, vol. 32, no. 9, pp. 3221–3234, 2012.

[39] C. A. Buneo and R. A. Andersen, "Integration of target and hand position signals in the posterior parietal cortex: effects of workspace and hand vision," *Journal of neurophysiology*, vol. 108, no. 1, p. 187, 2012.

[40] M. Antonelli, E. Chinellato, and A. P. del Pobil, "On-line learning of the visuomotor transformations on a humanoid robot," in *Intelligent Autonomous Systems 12*, ser. Advances in Intelligent Systems and Computing, S. Lee, H. Cho, K.-J. Yoon, and J. Lee, Eds. Springer Berlin Heidelberg, 2013, vol. 193, pp. 853–861.

[41] D. H. Hubel and T. N. Wiesel, "Receptive fields and functional architecture of monkey striate cortex," *The Journal of Physiology*, vol. 195, no. 1, p. 215, 1968.

[42] D. Albrecht and D. Hamilton, "Striate cortex of monkey and cat: Contrast response function." *J. Neurophysiol.*, vol. 48, no. l, 1982.

[43] N. Qian, "Computing Stereo Disparity and Motion with Known Binocular Cell Properties," *Neural Computation*, vol. 6, no. 3, pp. 390–404, 1994.

[44] S. P. Sabatini, G. Gastaldi, F. Solari, K. Pauwels, M. V. Hulle, J. Diaz, E. Ros, N. Pugeault, and N. Krüger, "A compact harmonic code for early vision based on anisotropic frequency channels," *Comput. Vis. Image Und.*, vol. 114, no. 6, pp. 681–699, 2010.

[45] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *The Journal of Physiology*, vol. 160, no. 1, p. 106, 1962.

[46] F. Beuth, J. Wiltschut, and F. Hamker, "Attentive stereoscopic object recognition," in *Workshop New Challenges in Neural Computation 2010*, T. Villmann and F.-M. Schleif, Eds., 2010, p. 41.

[47] M. W. Spratling, "Learning viewpoint invariant perceptual representations from cluttered images." *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 5, pp. 753–61, May 2005.

[48] M. Teichmann, J. Wiltschut, and F. H. Hamker, "Learning invariance from natural images inspired by observations in the primary visual cortex." *Neural Comput.*, vol. 24, no. 5, pp. 1271–96, May 2012.

[49] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nature Neurosci.*, vol. 2, pp. 1019–1025, 1999.

[50] L. Itti and C. Koch, "Computational modelling of visual attention." *Nature Rev. Neurosci.*, vol. 2, no. 3, pp. 194–203, Mar. 2001.

[51] G. Masson, C. Busettini, and F. Miles, "Vergence eye movements in response to binocular disparity without depth perception," *Nature*, vol. 389, pp. 283–286, 1997.

[52] O. D. Faugeras, Q.-T. Luong, and T. Papadopoulo, *The geometry of multiple images - the laws that govern the formation of multiple images of a scene and some of their applications*. MIT Press, 2001.

[53] M. Hansard and R. Horaud, "Patterns of binocular disparity for a fixating observer," *Advances in Brain, Vision, and Artificial Intelligence*, pp. 308–317, 2007.

[54] Y. Chen and N. Qian, "A coarse-to-fine disparity energy model with both phase-shift and position-shift receptive field mechanisms," *Neural Comput.*, vol. 16, no. 8, pp. 1545–1577, 2004.

[55] W. M. Theimer and H. A. Mallot., "Phase-based binocular vergence control and depth reconstruction using active vision." *Comput. Vis. Image Und.*, vol. 60(3), pp. 343–358, 1994.

[56] M. J. Morgan and E. Castet., "The aperture problem in stereopsis." *Vision Res.*, vol. 37, pp. 2737–2744, 1997.

[57] A. Gibaldi, A. Canessa, M. Chessa, S. P. Sabatini, and F. Solari, "A neuromorphic control module for real-time vergence eye movements on the icub robot head," in *IEEE/RAS Int. Conf. Humanoid Robots*, 2011, pp. 1065–1073.

[58] J. Park and I. W. Sandberg, "Universal approximation using radial-basis-function networks," *Neural Comput.*, vol. 3(2), pp. 246–257, Jun. 1991.

[59] 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.

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

[61] N. B. Karayiannis and A. N. Venetsanopoulos, "Fast learning algorithms for neural networks," *IEEE T. Circuits-II*, vol. 39(7), pp. 1–22, Jul 1992.

[62] E. Chinellato, M. Antonelli, and A. P. del Pobil, "A pilot study on saccadic adaptation experiments with robots," in *Biomimetic and Biohybrid Systems*, ser. LNCS, T. J. Prescott, N. F. Lepora, A. Mura, and P. F. Verschure, Eds. Springer Berlin Heidelberg, 2012, vol. 7375, pp. 83–94.

[63] G. Metta, P. Fitzpatrick, and L. Natale, "Yarp: Yet another robot platform," *Int. J. Adv. Rob. Syst.*, vol. 3, pp. 043–048, Jan 2006.

[64] S. Patel, B. Jiang, and H. Ogmen, "Vergence dynamics predict fixation disparity," *Neural computation*, vol. 13, no. 7, pp. 1495–1525, 2001.

[65] J. Bruske, M. Hansen, L. Riehn, and G. Sommer, "Biologically inspired calibration-free adaptive saccade control of a binocular camera-head," *Biological Cybernetics*, vol. 77, no. 6, pp. 433–446, 1997.

[66] P. Forssén, "Learning saccadic gaze control via motion prediciton," *Computer and Robot Vision (CRV), 2007. Fourth Canadian Conference on*, pp. 44–54, 2007.

[67] RobinLab, "A hierarchical system for a distributed representation of the peripersonal space of a humanoid robot," https://www.youtube.com/watch?v=gE9cTqj-16g, 2014.

[68] F. Chao, M. Lee, and J. Lee, "A developmental algorithm for ocular-motor coordination," *Robot. Auton. Syst.*, vol. 58(3), pp. 239–248, 2010.

[69] S. McBride and M. Lee, "Fast learning mapping schemes for robotic hand–eye coordination," *Cognitive Computation*, Jan 2010.

[70] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[71] M. Zirnsak, F. Beuth, and F. H. Hamker, "Split of spatial attention as predicted by a systems-level model of visual attention." *Eur. J. Neurosci.*, vol. 33, no. 11, pp. 2035–45, Jun. 2011.

[72] A. Gibaldi, A. Canessa, M. Chessa, F. Solari, and S. P. Sabatini, "Population coding for a reward-modulated hebbian learning of vergence control," in *Neural Networks (IJCNN), The 2013 International Joint Conference on*. IEEE, 2013, pp. 1–8.

[73] ——, "How a population-based representation of binocular visual signal can intrinsically mediate autonomous learning of vergence control," *Procedia Computer Science*, vol. 13, pp. 212–221, 2012.

[74] M. Antonelli, A. J. Duran, and A. P. Del Pobil, "Application of the visuo-oculomotor transformation to ballistic and visually-guided eye movements," in *Neural Networks (IJCNN), The 2013 International Joint Conference on*. IEEE, 2013, pp. 1–8.

[75] A. Stenzel, E. Chinellato, M. A. T. Bou, A. P. del Pobil, M. Lappe, and R. Liepelt, "When humanoid robots become human-like interaction partners: Corepresentation of robotic actions." *J. Exp. Psychol. Human.*, vol. 38, no. 5, pp. 1073–1077, 2012.

[76] P. Földiák, "Forming sparse representations by local anti-Hebbian learning," *Biol. Cybern.*, vol. 64, pp. 165–170, 1990.

[77] J. Heinzle, "A model of the local cortical circuit of the frontal eye fields," Ph.D. dissertation, Swiss Federal Institute of Technology, 2006.

[78] A. Segraves and E. Goldberg, "Functional properties of corticotectal neurons in the monkey's frontal eye field," *J. Neurophysiol.*, vol. 58, no. 6, pp. 1387–1419, 1987.

[79] J. Schall, K. Thompson, N. Bichot, A. Murthy, and T. Sato, "Visual processing in the macaque frontal eye field," in *The Primate Visual System*, J. Kaas and C. Collins, Eds., 2004, pp. 205–230.

[80] R. P. Hasegawa, B. W. Peterson, and M. E. Goldberg, "Prefrontal neurons coding suppression of specific saccades." *Neuron*, vol. 43, no. 3, pp. 415–25, Aug. 2004.

[81] J. Schall, "Neural basis of saccade target selection," *Rev. Neuroscience*, vol. 6, pp. 63–85, 1995.

APPENDIX A
NOTATION

This section describes the mathematical notation used in the appendices. The firing rates of all neurons are labeled with $\boldsymbol{r}$, superscripts denote the cortical area or the function, whereas subscripts denote the neuron indexes (e.g. $\boldsymbol{r}^{\text{V4}}_{\boldsymbol{i},\boldsymbol{x}}$). The index $\boldsymbol{x}$ refers to the spatial location of the receptive field in retinal coordinates $(\boldsymbol{x_1}, \boldsymbol{x_2})$. The second multi-index $\boldsymbol{i}$ refers to specific features of the neuron at a certain location. Weights matrices which connected *area1* with *area2* are termed as $\boldsymbol{w}^{\text{area1-area2}}_{\boldsymbol{x},\boldsymbol{x'}}$ with the current post-synaptic neuron $\boldsymbol{x}$ and the pre-synaptic neuron $\boldsymbol{x'}$. Weights matrices connecting different features laterally are termed as $\boldsymbol{w}^{\text{area/i}}$.

APPENDIX B
PRIMARY VISUAL CORTEX (V1)

The local geometry of the visual stimulus in the neighborhood of a given point on the image plane $\boldsymbol{x}$ is represented in the harmonic space (amplitude, phase, and orientation), through filtering operations with complex-valued 2D band-pass kernels, *i.e.* a complex Gabor function, defined by:

$$\boldsymbol{h}(\mathbf{x};\boldsymbol{\theta},\boldsymbol{\psi}) = \boldsymbol{\eta} e^{\left(-\frac{1}{2\sigma^2}\mathbf{x}_{\boldsymbol{\theta}}^T\mathbf{x}_{\boldsymbol{\theta}}\right)} e^{\boldsymbol{\jmath}(\boldsymbol{k_0}\mathbf{x}_{\boldsymbol{\theta}}+\boldsymbol{\psi})} \tag{1}$$

where $\mathbf{x}_{\boldsymbol{\theta}}$ is the rotated coordinate system by an angle $\boldsymbol{\theta}$, $\boldsymbol{k_0}$ is the radial peak frequency, $\boldsymbol{\eta}$ is a normalization constant, and $\boldsymbol{\psi}$ is the phase value that characterizes the receptive field profile.

Formally, the response of a binocular simple cell of area V1, centered in $\mathbf{x}$, with a phase shift of $\boldsymbol{\Delta\psi}$ and oriented along $\boldsymbol{\theta}$,

can be written as the scalar product between the image $\boldsymbol{I_{L/R}}$ and the cell's receptive field profile $\boldsymbol{h_{L/R}}$:

$$r^{\text{V1}}_{s,\theta,\Delta\psi,\mathbf{x}} = \boldsymbol{I_L^T}\cdot\boldsymbol{h_L}(\mathbf{x};\boldsymbol{\theta},\boldsymbol{\psi_L}) + \boldsymbol{I_R^T}\cdot\boldsymbol{h_R}(\mathbf{x};\boldsymbol{\theta},\boldsymbol{\psi_R}) \quad (2)$$

while the response of a binocular complex cell is:

$$r^{\text{V1}}_{c,\theta,\Delta\psi,\mathbf{x}} = \left| r^{\text{V1}}_{s,\theta,\Delta\psi,\mathbf{x}} + r^{\text{V1}}_{s,\theta,\Delta\psi+\frac{\pi}{2},\mathbf{x}} \right|^2 \quad (3)$$

where $\boldsymbol{\Delta\psi} = \psi^L - \psi^R$.

From the population response, assuming it to have $\boldsymbol{N_p}$ different phase values along each orientation, the *component disparity* along the orientation $\boldsymbol{\theta}$ in the retinal position $\mathbf{x}$ can be obtained by applying a center of mass decoding strategy:

$$\boldsymbol{\delta}_{\boldsymbol{\theta},\mathbf{x}} = \frac{\sum_{\boldsymbol{\Delta\psi_i}}^{\boldsymbol{N_p}} \frac{\boldsymbol{\Delta\psi_i}}{\boldsymbol{k_0}} r^{\text{V1}}_{c,\theta,\Delta\psi_i,\mathbf{x}}}{\sum_{\boldsymbol{\Delta\psi_i}}^{\boldsymbol{N_p}} r^{\text{V1}}_{c,\theta,\Delta\psi_i,\mathbf{x}}} \quad (4)$$

Assuming it to have $\boldsymbol{N_o}$ orientations, the full disparity vector is obtained by an intersection of constraints [55], thus solving the aperture problem:

$$\boldsymbol{\delta}_{\mathbf{x}} = \underset{\boldsymbol{\delta}_{\mathbf{x}}}{\mathbf{argmin}} \sum_{\boldsymbol{\theta_j}}^{\boldsymbol{N_o}} \left(\boldsymbol{\delta}_{\boldsymbol{\theta_j},\mathbf{x}} - \frac{\mathbf{k_j}^T}{\boldsymbol{k_0}} \boldsymbol{\delta}_{\mathbf{x}} \right)^{\mathbf{2}}, \quad (5)$$

where $\mathbf{k_j}$ is the 2D peak frequency vector of the Gabor function. Following this approach, it is possible to obtain a full, dense and robust disparity map of the observed scene.

Appendix C
Vergence Control (VC)

The output of the vergence control $\boldsymbol{r}^{\text{MST}}$ is obtained by a weighted sum of the cell responses $\boldsymbol{r}^{\text{V1}}$:

$$\boldsymbol{r}^{\text{MST}} = \sum_{\mathbf{x}\in\boldsymbol{\Omega}} \boldsymbol{G}(\mathbf{x}) \sum_{\boldsymbol{\Delta\psi_i},\boldsymbol{\theta_j}}^{\boldsymbol{N_p}\times\boldsymbol{N_o}} \boldsymbol{w}^{\text{V1-MST}}_{\boldsymbol{i,j},\mathbf{x}} \boldsymbol{r}^{\text{V1}}_{\boldsymbol{c},\boldsymbol{\theta_j},\boldsymbol{\Delta\psi_i},\mathbf{x}} \quad (6)$$

where $\boldsymbol{G}(\mathbf{x})$ is a Gaussian weighting profile centered in the fovea. The weights $\boldsymbol{w_{i,j}}$ are obtained by minimizing a functional that includes a sensitivity to the horizontal component of the vector disparity $\boldsymbol{\delta_H}$ and an insensitivity to the vertical component $\boldsymbol{\delta_V}$:

$$\begin{aligned} \boldsymbol{E}(\boldsymbol{w}) = & \left\| \textstyle\sum_{\boldsymbol{\Delta\psi_i},\boldsymbol{\theta_j}}^{\boldsymbol{N_p}\times\boldsymbol{N_o}} \boldsymbol{\varphi_{c,i}}(\boldsymbol{\delta_H} - \boldsymbol{\delta_{H,i}}) \boldsymbol{w}^{\text{V1-MST}}_{\boldsymbol{i,j}} - \boldsymbol{\upsilon_H} \right\|^2 + \\ + & \left\| \textstyle\sum_{\boldsymbol{\Delta\psi_i},\boldsymbol{\theta_j}}^{\boldsymbol{N_p}\times\boldsymbol{N_o}} \boldsymbol{\varphi_{c,i}}(\boldsymbol{\delta_V})(\boldsymbol{\varphi_i} - \mathbf{1}) \right\|^2 \end{aligned} \quad (7)$$

Considering $\boldsymbol{\varphi_{c,i}}(\boldsymbol{\delta})$ the response surface of a cell to the vector disparity, $\boldsymbol{\varphi_{c,i}}(\boldsymbol{\delta_H})$ and $\boldsymbol{\varphi_{c,i}}(\boldsymbol{\delta_V})$ are the tuning curves derived for horizontal and vertical disparity, *i.e.* the horizontal and vertical cross-sections of the surface. $\boldsymbol{\upsilon_H}$ is the desired (i.e., imposed) behavior of the horizontal vergence control.

Appendix D
High Visual Area (HVA)

This section describes the object recognition system. Each HVA cell ($\boldsymbol{r}^{\text{HVA}}$), over its receptive field, gains excitation (eq. 11) from a weighted sum of V1 cells and is inhibited by all other HVA cells (eq. 8):

$$\begin{aligned} \boldsymbol{\tau_R}^{\text{HVA}} \frac{\boldsymbol{\partial} \boldsymbol{r}^{\text{HVA}}_{\boldsymbol{i,x,k}}}{\boldsymbol{\partial t}} = & -\boldsymbol{r}^{\text{HVA}}_{\boldsymbol{i,x,k}} + \boldsymbol{e_{i,x}} \cdot \boldsymbol{a_{i,x,k}} \\ & - \boldsymbol{d_{nl}} \cdot \sum_{\boldsymbol{i'},\boldsymbol{i'\neq i}} \boldsymbol{f}\left(\boldsymbol{w}^{\text{HVA/i}}_{\boldsymbol{i,i'}} \boldsymbol{r}^{\text{HVA}}_{\boldsymbol{i',x,k}} \right) \end{aligned} \quad (8)$$

$$\boldsymbol{cn}(\boldsymbol{u},\boldsymbol{m}) = \begin{cases} \frac{\boldsymbol{u}}{\boldsymbol{m}}\ \boldsymbol{u} > \mathbf{0.1}\,\boldsymbol{m} \\ \boldsymbol{u}\ \boldsymbol{u} \leq \mathbf{0.1}\,\boldsymbol{m} \end{cases} \quad (9)$$

$$\begin{aligned} \boldsymbol{a_{i,x,k}} = & \left(\mathbf{1} - \max_{\boldsymbol{i'}} \left(\boldsymbol{r}^{\text{OM}}_{\boldsymbol{i',k}} \right) + \boldsymbol{r}^{\text{OM}}_{\boldsymbol{i,k}} \right) \\ & \cdot \begin{cases} \mathbf{1} - \max_{\boldsymbol{x'}} \left(\boldsymbol{r}^{\text{FEFm}}_{\boldsymbol{x'}} \right) + \boldsymbol{r}^{\text{FEFm}}_{\boldsymbol{x}} & \boldsymbol{k} = \mathbf{1} \\ \mathbf{1} & \boldsymbol{k} = \mathbf{2} \end{cases} \end{aligned} \quad (10)$$

$$\boldsymbol{e_{i,x}} = \sum_{\boldsymbol{i',x'}\in \text{V1}} \boldsymbol{w}^{\text{V1-HVA}}_{\boldsymbol{i',x',i}} \cdot \boldsymbol{cn}\left(\boldsymbol{r}^{\text{V1}}_{\boldsymbol{i',x'}}, \max\left(\boldsymbol{r}^{\text{V1}}_{\boldsymbol{i',x'}} \right) \right) \quad (11)$$

where $\boldsymbol{\tau_R} = \mathbf{12}$ is the time constant, $\boldsymbol{d_{nl}} = \mathbf{0.8}$ modulates the competition between HVA cells, and $\boldsymbol{f}(\cdot)$ is a non-linear processing stage: $\boldsymbol{f}(\boldsymbol{u}) = \boldsymbol{log}(\frac{\mathbf{1}+\boldsymbol{u}}{\mathbf{1}-\boldsymbol{u}})$. The selectivity of a single HVA cell to a specific object is thus given by $\boldsymbol{w}^{\text{V1-HVA}}$ and $\boldsymbol{w}^{\text{HVA/i}}$, that are the weights for the feedforward and the lateral connections, respectively. The term $\boldsymbol{e}$ denotes the excitation from V1 and the term $\boldsymbol{a}$ the spatial and the feature-based attention from object memory (OM). The function $\boldsymbol{cn}$ increases the contrast inside each receptive field separately to improve robustness against different stimuli contrasts and illumination conditions. The HVA and FEF are split into two parts, the normal one driven by feature-based attention ($\boldsymbol{k} = \mathbf{1}$) and one driven by feature-based suppression ($\boldsymbol{k} = \mathbf{2}$).

Connection weights between V1 and HVA ($\boldsymbol{w}^{\text{V1-HVA}}$) are learned using the trace learning approach. The activation of a pre-synaptic cell is combined with the post-synaptic activation of the previous stimulus:

$$\begin{aligned} \boldsymbol{\tau_w} \frac{\boldsymbol{\partial} \boldsymbol{w}^{\text{V1-HVA}}_{\boldsymbol{i',i}}}{\boldsymbol{\partial t}} = & (\boldsymbol{r}^{\text{V1}}_{\boldsymbol{i'}} - \boldsymbol{\theta}^{\text{V1}})_t \, [\boldsymbol{r}^{\text{HVA}}_{\boldsymbol{i}} - \boldsymbol{\theta}^{\text{HVA}}]^+_{\boldsymbol{t-1}} \\ & - \boldsymbol{\alpha_w}\, \boldsymbol{w_{i',i}}\, (\boldsymbol{r}^{\text{HVA}}_{\boldsymbol{i}} - \boldsymbol{\theta}^{\text{HVA}})^{\mathbf{2}}_{\boldsymbol{t-1}} \end{aligned} \quad (12)$$

where $\boldsymbol{\alpha_w} = \mathbf{350}$ constrains the weights, $\boldsymbol{\tau_w} = \mathbf{2 \cdot 10^5}$ is the time constant which controls the speed of the learning process, and $[\boldsymbol{x}]^+$ stands for $\mathbf{argmax}(\boldsymbol{x}, \mathbf{0})$. The term $\boldsymbol{\theta}^{\text{V1}} = \bar{\boldsymbol{r}}^{\text{V1}}$ is the mean activation of the whole population of V1 while $\boldsymbol{\theta}^{\text{HVA}} = \mathbf{max}(\mathbf{0.95 \cdot max}(\boldsymbol{r}^{\text{HVA}}), \bar{\boldsymbol{r}}^{\text{HVA}})$. A weight sharing approach was used to analyze the whole visual scene in parallel, hence, the detection of the objects is independent of their spatial location.

Lateral connections among HVA cells ($\boldsymbol{w}^{\text{HVA/i}}$) were learned by Anti-Hebbian learning in order to increase the competition among them. That is, inhibition is strengthened when both cells are activated simultaneously:

$$\boldsymbol{\tau_c} \frac{\boldsymbol{\partial} \boldsymbol{w}^{\text{HVA/i}}_{\boldsymbol{i',i}}}{\boldsymbol{\partial t}} = (\boldsymbol{r}^{\text{HVA}}_{\boldsymbol{i'}} - \boldsymbol{\theta}^{\text{c}}) \cdot (\boldsymbol{r}^{\text{HVA}}_{\boldsymbol{i}} - \boldsymbol{\theta}^{\text{c}}) - \boldsymbol{\alpha_c} \cdot \boldsymbol{w}^{\text{HVA/i}}_{\boldsymbol{i',i}} \cdot (\boldsymbol{r}^{\text{HVA}}_{\boldsymbol{i}} - \boldsymbol{\theta}^{\text{c}}) \quad (13)$$

where $\boldsymbol{\alpha_c} = \mathbf{0.1}$ constrains the weights, $\boldsymbol{\tau_c} = \mathbf{10^6}$ is the time constant and $\boldsymbol{\theta}^{\text{c}} = \mathbf{0.33}$ is a fixed threshold. Anti-Hebbian learning leads to de-correlated responses and a sparse code of the cell population [76].

The frontal eye field is highly involved in planning eye movements [32], [77]; its neurons show activities [31] related to 1) the visual selectivity of a saccade target (called visual cells [78], [79]), to 2) eye motor activity (motor cells [30], [78]), to 3) suppression of saccades (fixation cells [80]) and to 4) visual and motor signals [81].

The part FEF$\boldsymbol{v}$ (eq. 14) simulates the visual cells; therefore it represents a perceptual map which is often referred to as a saliency map. This map receives afferents from level V4pool at the same retinotopic location, irrespective of the feature information and thus, encodes the conspicuity of locations.

The part FEF$\boldsymbol{m}$ (eq. 15) represents the motoric cells which encode saccade target information. They receive excitatory signals for attended objects and inhibitory signals for unattended ones. The visuomotor cells are not simulated, but their function is simulated by the interaction of FEF$\boldsymbol{v}$ and FEF$\boldsymbol{m}$. The fixation cells are roughly represented by the second channel of FEF$\boldsymbol{v}$ which suppress indirect saccades to certain locations.

$$r^{\text{FEFv}}_{x,e} = \max_{i'} \left(r^{\text{HVA}}_{i',x,e} \right) \tag{14}$$

$$r^{\text{FEFm}}_{x} = h \left[\max_{x' \in \text{FEFv}} \left((r^{\text{FEFv}}_{x',e=1})^p \right) - w^{\text{Inh}} \cdot \max_{x' \in \text{FEFv}} \left((r^{\text{FEFv}}_{x',e=2})^p \right) \right] \tag{15}$$

$$h(r) = \left(\frac{r}{\max(r)}(1+c) - c \right) \cdot \max_{x'} \left(r^{\text{FEFv}}_{x',e=1} \right) \tag{16}$$

where $w^{\text{Inh}} = 0.5$ regulates the amount of feature-based suppression. The power rule factor $p = 1.4$ controls the amount of competition and facilitates functionally a contrast increase. The function $h(r)$ preserves the maximum value and decreases the minimum by a global inhibition mechanism, controlled by the term $c = 0.1 \cdot \max(r^{\text{FEFm}})$. The processing is stopped when the FEF$\boldsymbol{m}$ encodes a valid saccade target, that is, when its maximum firing rate reaches a chosen threshold (0.96).

Appendix E
Visuomotor transformations in the posterior parietal cortex

The main source of inspiration for the sensorimotor framework [19] are the basis function approach [60] and the neuroscience experiments on the role of the V6A area during gazing and reaching action [35]–[37]. The computational framework is composed by three radial basis function networks (RBFNs).

The visuo-oculomotor transformation ($\boldsymbol{T_{R \mapsto E}}$) receives as input the retinotopic position of the target ($\boldsymbol{t_r}$) provided by the FEF$\boldsymbol{m}$ area and its disparity which is provided by V1 through the disparity estimation, that is $\boldsymbol{t_r} = [r^{\text{FEFm}}_{x}\ r^{\text{FEFm}}_{y}\ \delta_x]^T$. The receptive fields of V6A neurons are modeled using Gaussian funtions to make them selective to specific retinotopic positions. Each neuron is characterized by its own center of activation ($\boldsymbol{\mu_j}$) while the shape of the activation, *i.e.* the width of the Gaussian ($\boldsymbol{\Sigma}$), is the same for every units:

$$r^{\text{V6A}}_{r,j}(t_r) = e^{-(t_r - \mu_j)^T \Sigma^{-1} (t_r - \mu_j)}. \tag{17}$$

The centers of the neurons cover the whole input space, so that, creating a distributed representation of the retinotopic location of the target. Such a representation can be converted into the head-centered f.o.r. (common tilt, left and right pan) by means of a weighted sum of the population response:

$$r^{\text{V6A}}_{h,i}(t_r) = \sum_j z_{j,i} \cdot r^{\text{V6A}}_{r,j}(t_r) \tag{18}$$

where $z_{j,i}$ is the weight that links the j-th unit of the population with the output i.

The plasticity of the transformation resides in the weighted connections. These connections can be changed by using bio-inspired gradient descent techniques such as the delta rule [60]. Conversely to previous work [40], we replaced the delta rule with a recursive least square algorithm [61], in order to speed up the learning process on the robotic setup. At each step, the algorithm updated the information matrix $\mathbf{P}$ and the weights $z_{i,j}$, using the following equations:

$$\alpha = [1 + r^{\text{V6A}}_{r}(t_r)^T \cdot \mathbf{P} \cdot r^{\text{V6A}}_{r}(t_r)]^{-1} \tag{19}$$

$$\Delta \mathbf{z} = \alpha \cdot [\bar{r}^{\text{V6A}}_{h}(t_r) - r^{\text{V6A}}_{h}(t_r)] \cdot r^{\text{V6A}}_{r}(t_r)^T \cdot \mathbf{P} \tag{20}$$

$$\mathbf{P} = \mathbf{P} - \alpha \cdot \mathbf{P} \cdot r^{\text{V6A}}_{r}(t_r) \cdot r^{\text{V6A}}_{r}(t_r)^T \cdot \mathbf{P} \tag{21}$$

where $\bar{r}^{\text{V6A}}_{h}(t_r)$ is the desired head position for the target t_r.

The transformations that link the eyes and arm position ($\boldsymbol{T_{E \leftrightarrow A}}$) have the same structure of the visuo-oculomotor transformation. The main difference is that the inputs are directly provided by the proprioceptive sensors of the eye and limb motors and not by the visual cortex. The inputs of $\boldsymbol{T_{E \mapsto A}}$ are the eye positions and the outputs are the arm joint positions required to reach the fixation point. On the other hand, the inputs of $\boldsymbol{T_{A \mapsto E}}$ are the current arm joint position and the outputs are the eye positions required to gaze at the hand.