Developmental Perception of the Self and Action

Ryo Saegusa, *Member IEEE,* Giorgio Metta, *Senior Member IEEE,* Giulio Sandini and Lorenzo Natale

***Abstract*—This paper describes a developmental framework for action-driven perception in anthropomorphic robots. The key idea of the framework is that action generation develops the agent's perception of its own body and actions. Action-driven development is critical for identifying changing body parts and understanding the effects of actions in unknown or non-stationary environments. We embedded minimal knowledge into the robot's cognitive system in the form of motor synergies and actions to allow motor exploration. The robot voluntarily generates actions and develops the ability to perceive its own body and the effect that it generates on the environment. The robot, moreover, can compose this kind of learned primitives to perform complex actions and characterize them in terms of their sensory effects. After learning, the robot can recognize manipulative human behaviors with cross-modal anticipation for recovery of unavailable sensory modality, and reproduce the recognized actions afterwards. We evaluated the proposed framework in experiments with a real robot. In the experiments, we achieved autonomous body identification, learning of fixation, reaching and grasping actions, and developmental recognition of human actions as well as their reproduction.**

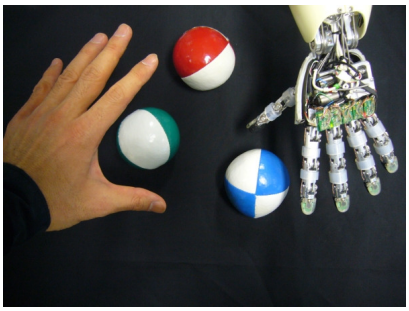
***Index Terms*—Self Perception, Action Perception, Manipulation, Action Learning, Mirror Neuron, Imitation.**

I. Introduction

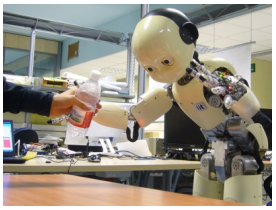
HOW can a robot identify the self and understand actions? Monkeys are able to recognize their own bodies even when they are experimentally modified or extended [1][2], and moreover, they understand actions so as to mirror them in observation and execution [3][4][5]. These kinds of cognitive functions may have the potential to break the limits of hand-coded machine intelligence.

The goal of this work is to create a cognitive ability which actively develops perception of the self and actions in non-stationary environments. Our claim for current cognitive systems is that robot actions are developed with perceptual information, but their perception is not adapted as the result of the explored action. In short, action-driven development of perceptual ability is missing in robot learning in non-stationary environments. Therefore, self-body perception in robots is not yet reconfigurable and the perception of actions demonstrated by robots and humans is not treated in the same way at a perceptual level.

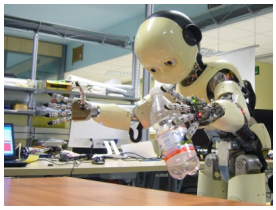
In this work, we introduce a method of primate-like developmental perception for manipulation tasks. A robot develops its ability to perceive by defining its own body with self-generated actions (motor exploration). The robot then learns primitive actions on fixation, reaching and grasping. Finally, the robot develops action perception based on observation of the results of self-generated actions. After learning, the robot can recognize human actions and also reproduce them.



(a) object manipulation



(b) action observation



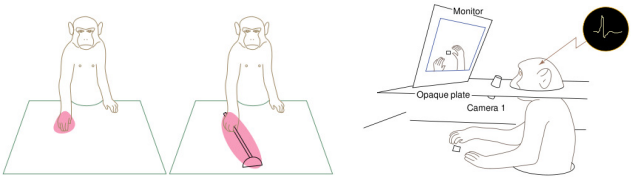
(c) action execution

Fig. 1. (a) A robot and a person interacting with objects. (b) The robot observes an experimenter's grasping action. (c) The robot then reproduces the recognized action.

Figure 1(a) shows a typical scene in which a robot and a person are interacting with objects. Questions here are how the robot can distinguish its own hand from others and how the objects are affected by actions. Neither the robot nor the person can conclude in advance whether the balls are manipulable (the balls may be fake pictures or stuck to the table). In our framework, the robot moves its hand and defines the object as its own hand if the visual movement of the object is correlated with its own motor sensing. The robot then demonstrates manipulative actions and characterizes the actions based on their effects on the objects. Effect-based action perception allows common identification of actions demonstrated by different agents (see Fig.1(b),(c)) in different body contexts (when the robot/human is using the hand or a tool).

This paper is organized as follows: Section II reviews the development of perception in biological and robotic systems.

R. Saegusa is with the Center for Human-Robot Symbiosis Research, Toyohashi University of Technology, 1-1 Hibarigaoka, Tempaku, Toyohashi, 441-8580, Japan. G. Metta and L. Natale are with the iCub Facility and G. Sandini is with the Robotics, Brain, and Cognitive Sciences Department, Istituto Italiano di Tecnologia, Via Morego 30, 16163, Genova, Italy.
This work was carried out at the Robotics, Brain and Cognitive Sciences Department, Istituto Italiano di Tecnologia.
E-mail to the corresponding author R.Saegusa: ryos@ieee.org



(a) tool use (b) visual guidance

Fig. 2. Body perception in monkeys. (a) Visual receptive field of the bimodal neurons (left: before tool use, right: after tool use). The monkey perceives a tool as an extended body part [1]. (b) Video-guided manipulation. After training, the monkey recognizes the hands projected on the monitor as its own hands [2] (the figures were reproduced from [6] under permission).

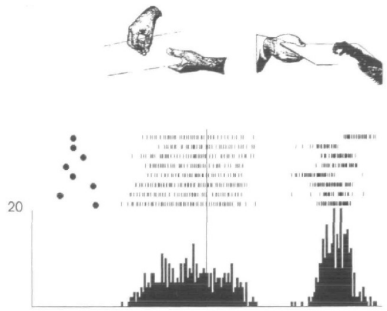


Fig. 3. Grasping mirror neurons in the premotor cortex of a monkey [3] [4]. The neurons were activated when the monkey observed a grasping action (left) and also when the monkey executed the grasping action (right) (the illustration was reproduced from [4] under permission).

Section III introduces a principle of body definition. Section IV describes a learning method of primitive actions. Section V describes the developmental perception of manipulative behaviors with humans. Section VI gives a comparison of the proposed method with other robotic and biological systems. Section VII concludes the proposed work and outlines some future tasks.

II. Development of Perception

In this section, we review the development of perception in biological systems and propose a corresponding framework for a robotic perception system. The detailed comparison of the proposed framework to other related robotic systems is presented later in Section VI.

A. Biological systems

Body image is fundamental for manipulation and it is extremely adaptive in animals. Iriki et al. found visuo-somatosensory neurons (bimodal neurons) in monkey intra-parietal cortex that incorporated a tool into a mental image of the hand [1]. The neurons respond to stimuli in the visual receptive field (reachable area) and the tactile receptive area (the surface of the hand or the shoulder). After tool use, the visual receptive field of these neurons is extended to include the tool (see Fig. 2(a)). In [2], the authors trained a monkey to recognize the image of the hand in a video monitor and demonstrated that the visual receptive field of these bimodal neurons was projected onto the video screen (see Fig. 2(b)). The experimental results suggested that the coincidence of movements between the real hand and the video-image of the hand led the monkey to use the video image for guiding hand movements. In summary, both experiments suggest that the monkey's body perception is developed through motor learning and then adapted for situations in operation.

Kaneko et al. investigated the perception of self-agency in chimpanzees [7]. They reported that chimpanzees are able to make a distinction between the self and others in external events that they are monitoring. This shows evidence of the ability in chimpanzees to perceive self-agency based on self-generated actions and their effects.

Rizzolatti et al. found visuomotor neurons (mirror neurons) in the premotor cortex of monkeys, which were activated when the monkeys performed a certain action and they also observed a similar action demonstrated by human experimenters [3] [4]. In the experiments, mirror neurons responded to the action of grasping, holding, placing, manipulating and two hand interaction. The activation of grasping mirror neurons is shown in Fig.3. As illustrated in the figure, the same set of neurons is activated during both observation (left) and execution (right) of the grasping action.

Interestingly, activation of the mirror neurons is selective for the type of actions, but the neurons are not responsive to mimicry actions without a target object. For example, Fogassi et al. found that the neurons in the Inferior Parietal Lobule (IPL) showed different activation for a specific act (e.g. grasping) when observed as part of different actions (e.g. eating, placing). The authors suggested that the IPL neurons and their connections encode not only the observed motor act but also the context of the act [5].

Beyond the experiments with monkeys, learning of contingency between their actions and events has been investigated in infant development studies [8]. The results of experiments with infants suggest that 2-month-olds can acquire and retain general body movements that induce contingent changes in a mobile object, while 3- and 4-month-olds form memories that serve as a constraint, enabling highly specific movements of their arms to effectively activate a mobile object.

B. A proposed robotic system

We introduce a framework of action-driven development for the self and action perception. The framework covers the construction of all perceptual systems in this work on body definition, motor control and action perception. Figure 4 illustrates a schematic presentation of action-driven development: a robot generates an action and associates the action with the effect that is perceived as a sensory event.

An original idea of the framework as compared to other related methods is that self-generated action drives the development of perception. In the initial phase, actions are randomly generated by a motor repertoire that includes simple reciprocal and ballistic movements like a Lévy process [9]. The generated actions create stimuli to the self's sensing system through the environment, and this sensory feedback develops the self's perception system.

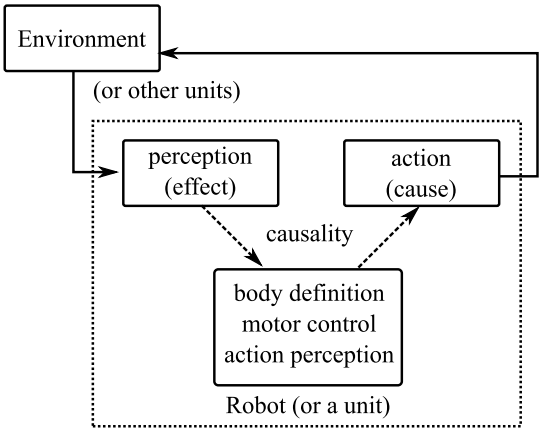


Fig. 4. Schematic presentation of action-driven developments. A robot generates an action and associates the action with the perceived sensory event. The causal relation constructs body definition, motor control and action perception.

Another new principle is that perceptual abilities are developed in an incremental manner. First, the robot identifies its own body with simple movements, and then it develops its body image and motor skills (primitive actions). Finally, the learned motor skills are combined as more complex manipulative behaviors and the robot develops action perception by demonstrating the behaviors with humans.

An overview of the system architecture is illustrated in Figure 5. Each bounding box in the figure represents a unit of sensory-motor functions that run independently in the networks. The whole system includes the sub-systems of vision, proprioception, tactile sensing, sensory integration, motor recognition and motor execution. Each function of the sub-systems is given in the following sections.

III. Body definition

Coincidence in vision and proprioception offers important clues for robots to build their body images. In a previous study we proposed a method for robots to learn their body image based on visuomotor correlation [10]. This section describes an extended method of body definition that allows multiple body segmentation in binocular vision. We first define motion-based visual extraction of a target, and then introduce a technique for body definition based on visuomotor correlation. At the end of the section, we present the results of experiments that demonstrate body definition.

A. Visual motion

A robot generates motor exploration with the arm motor synergy. The synergy in this work means coordination in the movements of multiple joint motors (detailed in Section III-B). Based on motor exploration, the robot identifies its own body using vision and proprioception. We use visual motion cues to segment the robot's body parts from the background, since visual motion cues prove the target's independence from the environment [11] and the cues are direct evidence of the self's motor controllability.

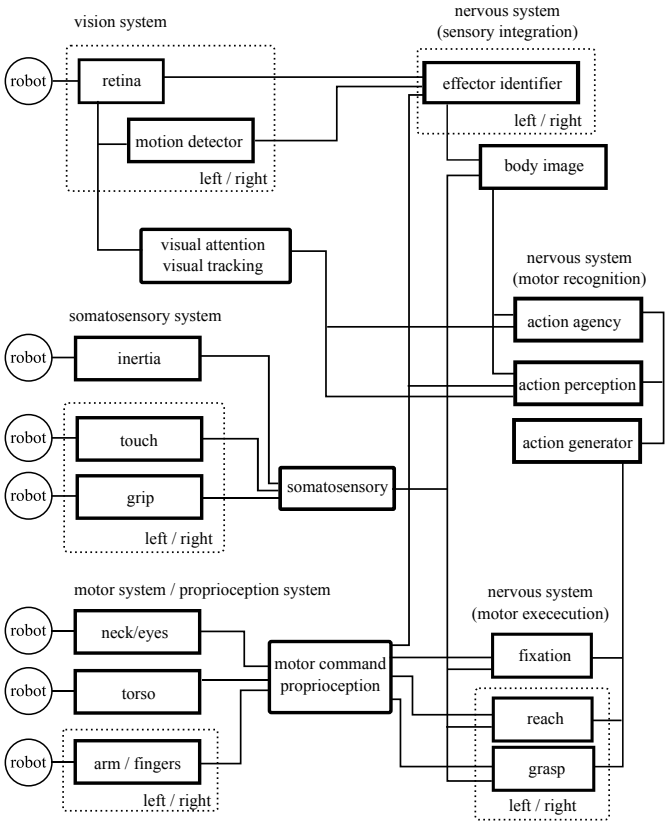


Fig. 5. A diagram of sensory-motor signal flows. The computations of sensory-motor modules are distributed in the networks.

Figure 6 illustrates visual motion detection. In the figure, we assume that there are some moving objects, and here the task is to extract visual blobs corresponding to moving objects. It is not critical that some objects that do not belong to the body are detected in the frame, since objects that move independently from the body will be filtered based on the visuomotor correlation in the next step (see Section III-B).

The absolute subtraction between the successive frames of monochrome image $I^m(\boldsymbol{x}, t)$ results in a difference image $I^f(\boldsymbol{x}, t)$ as follows;

$$I^f(\boldsymbol{x}, t) = |I^m(\boldsymbol{x}, t) - I^m(\boldsymbol{x}, t - \tau)|, \tag{1}$$

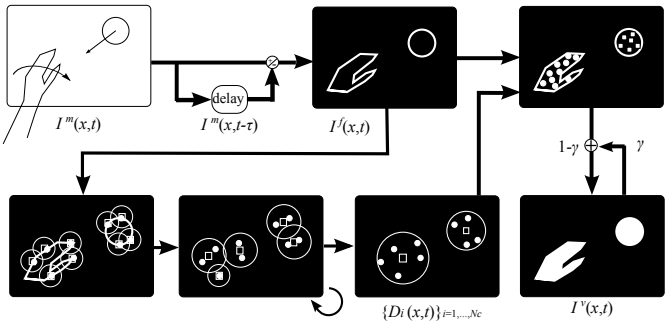
where $\boldsymbol{x} = (\xi, \eta)$ denotes the horizontal coordinate and vertical coordinate on the image. t and τ denote a sampling time and the interval of the frames.

We will now define a procedure for clustering different blobs and filling in the area. First, motion points/pixels are grouped in clusters. A set of points is randomly sampled from the high intensity points on I^f. Each sample point is given a small disk. The disk of the i-th point x_i is represented as follows;

$$D_i(\boldsymbol{x}) = \{\boldsymbol{x} | |\boldsymbol{x} - \boldsymbol{x}_i| \leq r_i\}, \tag{2}$$

where r_i denotes the radius of the disk. The neighbor disks are grouped as a new disk, if the disks intersect. The intersection of disk $D_i(\boldsymbol{x})$ and $D_j(\boldsymbol{x})$ are detected, when

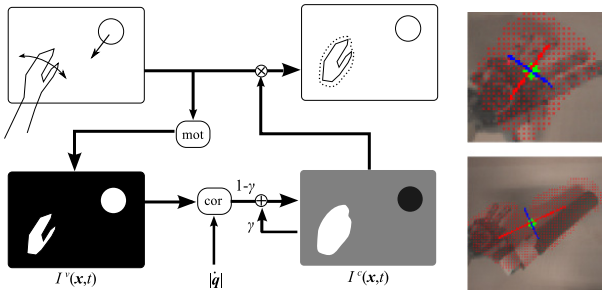
$$|\boldsymbol{x}_i - \boldsymbol{x}_j| < |r_i + r_j|. \tag{3}$$



(a) motion detection

(b) intermediate images

Fig. 6. Visual motion detection. (a) Detection procedure. The motion area is integrated in a bottom-up manner. (b) Intermediate images; the reference, difference, sampled points and filled blobs are presented from left to right.



(a) visuomotor correlation

(b) body parts

Fig. 7. Body identification. (a) The visual motion area is identified as a body part if its motion is correlated with proprioceptive motion. (b) Examples of identified body parts (top: inherent body; bottom: extended body).

The new disk takes all member points of the merged disks as its own. The new center and the radius of the merged disks are the average and distance deviation of the member points. This integration is repeated while a new disk appears.

After clustering, the set of disk centers $\{\boldsymbol{x}_i\}_{i=1,\ldots,n_c}$ is used for segmentation to obtain dense blobs of pixels which correspond to individual objects in the scene. The number of visual blobs, denoted as n_c, is dynamically given by the result of the area integration. High intensity points on the difference image are assigned to the nearest disk center, which are mostly along the outlines of motion areas. Random interpolation of these points gives a set of points that fill the motion area as follows:

$$\boldsymbol{x}'_k = a\boldsymbol{x}_i + (1-a)\boldsymbol{x}_j \tag{4}$$

where $\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{x}'_k$ denotes the i-th and j-th member point, and k-th interpolated point, respectively. The number of (i,j) couples, which corresponds to interpolation density, is selected empirically. Rate a is selected from uniform distribution in $[0,1]$. The interpolated points of the blob are blurred spatially by the Gaussian kernel and accumulated temporally as follows;

$$I^v(\boldsymbol{x},t) = \gamma I^v(\boldsymbol{x},t-\tau) + (1-\gamma)\sum_k K(\boldsymbol{x},\boldsymbol{x}'_k), \tag{5}$$

$$K(\boldsymbol{x},\boldsymbol{x}'_k) = A\exp\Big\{-\frac{|\boldsymbol{x}-\boldsymbol{x}'_k|^2}{2\sigma^2}\Big\}, \tag{6}$$

where $I^v(\boldsymbol{x},t)$ denotes the result image and x_k denotes the image coordinates of the k-th member point. $\gamma \in [0,1]$ is a decay rate. K is the Gaussian kernel. Image $I^v(\boldsymbol{x},t)$ forms clouds of labeled motion area. The parameters to be given are the initial r_i (connection scale), γ (sensitivity scale) and σ (blur scale). We set $A = 255$ for 8 bit intensity coding. Figure 6(b) shows the intermediate images of motion detection. $I^v(\boldsymbol{x},t)$ is initialized as zero each time before starting a body movement. We do not normalize I^v in Eq.5, since I^v is a positive value less than 255, and the positive summation in the second term can be controlled to be less than 255 by changing parameter A.

B. Body identification

We will now introduce the body identification procedure that allows a robot to segment its body from the environment. The assumption is that the causal relation between a self-generated action and its effect defines the body of the agent. The robot monitors the visuomotor correlation between proprioceptive and visual motion. When the robot detects the visuomotor correlation, the visually moving object is identified as a part of the body.

We have improved the single body part identification of the previous system [10] to allow multiple body part identification as follows: the robot generates actions with each motor unit (e.g., the wrist or shoulder of the left or right side), and associates the sensory event with the actuated motor unit individually. Multiple body part identification enables the robot to perceive its own body parts and link them to the corresponding motor units. The robot performs repetitive movements to exclude other objects from body identification. Figure 7 illustrates this procedure. The advantage of this technique is that the action-driven perception generalizes the body identification in which the body can be modified or extended by a grasped tool as demonstrated in Fig. 7(b).

The robot generates a movement:

$$\boldsymbol{u} = \boldsymbol{q} + \delta\boldsymbol{q}, \tag{7}$$

where $\boldsymbol{u}$ denotes the motor command of the motor unit, $\boldsymbol{q}$ denotes the reference encoder values of the motor unit, and $\delta\boldsymbol{q}$ denotes a variation. We consider here motor units of wrists and shoulders in the left and right arm. For example, when the robot generates a left wrist motor movement, the identified body part is coupled with the left wrist motor unit.

TABLE I
BODY DEFINITION, EXPERIMENTAL CONDITIONS

item	parameter	notation
motor unit	arm	$\boldsymbol{q} \in R^7$
exploration part	shoulder, wrist	S($\boldsymbol{u}_s \in R^3$), W($\boldsymbol{u}_w \in R^3$)
hand state	free, grasp	{V, H, N}, {Gf, Ga, Gb}

The visuomotor correlation map $I^c(\boldsymbol{x}, t)$ is given by the following equations;

$$I^c(\boldsymbol{x}, t) = \gamma I^c(\boldsymbol{x}, t-\tau) + (1-\gamma)c(\boldsymbol{x}, t), \quad (8)$$

$$c(\boldsymbol{x}, t) = \begin{cases} \Delta & \text{if } |\dot{\boldsymbol{q}}(t)| > p_0,\ I^v(\boldsymbol{x}, t) > I_0, \\ -\Delta & \text{if } |\dot{\boldsymbol{q}}(t)| < p_0,\ I^v(\boldsymbol{x}, t) > I_0, \\ 0 & \text{otherwise,} \end{cases}$$

where I^v represents the motion image and $\dot{\boldsymbol{q}}$ denotes the velocity of the joint angle vector of the motor unit. Δ, I_0, p_0 are positive constants. γ is a constant that determines the smoothing factor. For visualization, the baseline of I^c is set as the center of the intensity range (128 in [0,255]).

I^c is reset as the center value when the robot starts to send motor commands, and visuomotor correlation values are accumulated during the movements. After the repetitive movements, the system refers to the accumulated values of the correlation map I^c and extracts highly correlated areas by simple thresholding (see the dotted area in Fig. 7(a)). This repetitive approach filters out non-body moving objects in the frames, since the timing of the movement is uncorrelated with that of the body.

Segmented body parts are stored in memory as a set of tuples of visual and proprioceptive data. The visual information of a body part includes a blob image (extracted color patch), the blob's silhouette (extracted binary patch) and the blob's geometrical information (area, location, distortion and orientation). The proprioceptive information is the joint angle vector of the corresponding motor unit taken at the time the body is detected.

In the following experiments, we show that body identification is performed separately for different arm orientations. However, the robot can learn a general mapping of different postures in a continuous manner; this requires learning of the Jacobian matrix of the joints. We will explain the details of continuous body image reconstruction in Section IV-B.

C. Experiments

We performed experiments to evaluate the proposed body identification. In the experiments, we investigated inherent body identification, extended body identification and the effect of the magnitude of movements. Here we use the term 'inherent body' to identify the situation in which the body is not modified (no tools are attached to it). The term 'extended body', on the other hand, is used to identify an inherent body plus some extensions such as a tool or object that the agent grasps.

We used the iCub robot platform [12] [13] for the following experiments. The joint link structure is shown in Fig. 8. Table I summarizes the experiment conditions. The arm motor synergy has 7 DOF for each side of the body. q denotes the joint angle vector given by motor joint encoders (the values were normalized in [-1,1]).

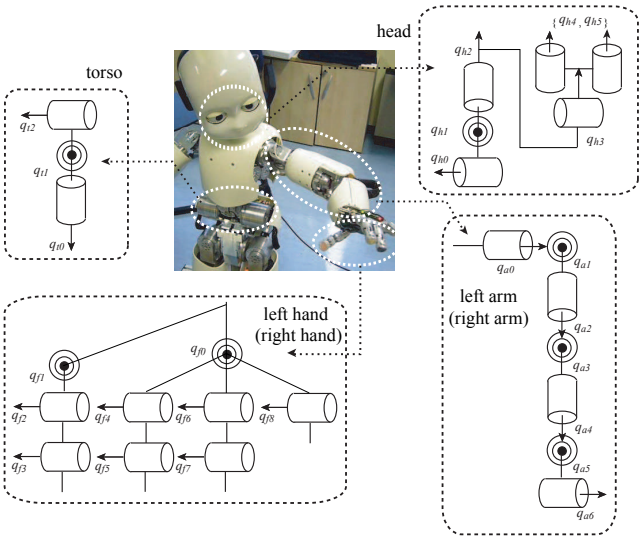


Fig. 8. The iCub robot platform [12]. The left side of the body is presented in the figure (the arm and hand on the right side are identical to those on the left side).

We define the shoulder and wrist movement as:

$$\boldsymbol{u}_s = \boldsymbol{q}_s + \delta\boldsymbol{q}_s, \quad (9)$$

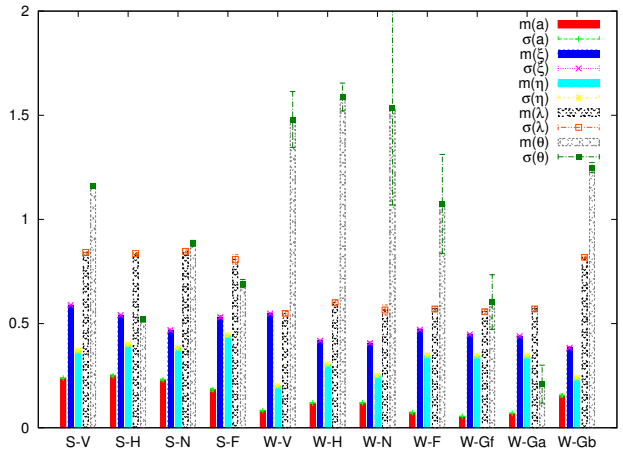
$$\boldsymbol{u}_w = \boldsymbol{q}_w + \delta\boldsymbol{q}_w, \quad (10)$$

where $\delta\boldsymbol{q}_s = (\delta q_0, \delta q_1, \delta q_2)$ and $\delta\boldsymbol{q}_w = (\delta q_4, \delta q_5, \delta q_6)$, respectively. The suffix number corresponds to the joint number of the arm $\boldsymbol{q}_a$ in Fig. 8. In the experiments, we actually performed repetitive movements of a back-and-forth movement ($\delta\boldsymbol{q}$ and $-\delta\boldsymbol{q}$) for body identification.

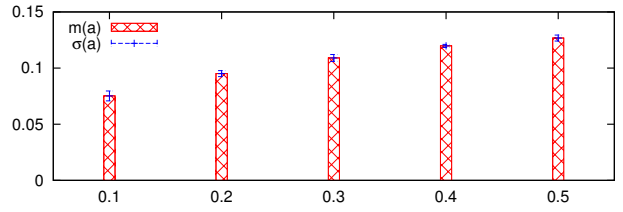
We investigated the visual features of the identified body parts in terms of visual area (how much space the body part occupies in the view field), location (where the body part is located in relation to the view field), distortion (how linear the body part is in shape) and orientation (in which direction the body part is oriented). In Fig.9, the variables $a, \boldsymbol{x} = (\xi, \eta), \lambda, \theta$ represent the area, location, distortion and orientation of the body part. With the term 'distortion' we identify the degree of similarity to a line segment (or dissimilarity from a circle) in shape. a is normalized as the frame area to be 1.0. $\boldsymbol{x} = (\xi, \eta)$ is normalized as the length of the diagonal segment of the frame to be 1.0. λ is given as follows: $\lambda = \lambda_1/(\lambda_1 + \lambda_2)$ where λ_1, λ_2 are the eigenvalue of the major and minor axes of the detected body part. θ is the orientation of the body part by radian; $\theta = \arctan(e_2/e_1)$ where $[e_1, e_2]^T$ denotes the vector of the major axis.

1) Inherent body identification: We performed 20 trials of shoulder and wrist motor exploration for each different posture condition. In these experiments, we set the range of angular movement as $\delta\boldsymbol{q}_s = \delta\boldsymbol{q}_w = 0.1$: (we will show the results of the different range value later in Section III-C3). We will now present the results of the right arm.

Figures 9(a) and 10 show the mean (m) and standard deviation (σ) of the visual features of the identified body

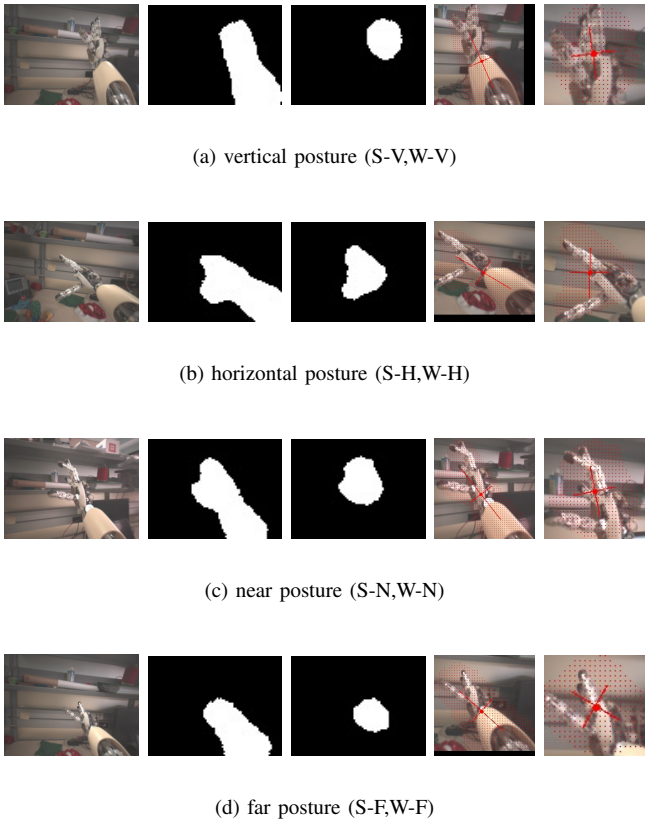


(a) inherent and extended body parts



(b) visual volume of body parts

Fig. 9. Visual features of body parts. (a) visual features of inherent body parts (S-V, S-H, S-N, S-F, W-V, W-H, W-N, W-F), visual features of extended body parts (W-Ga, W-Gb, W-Gc), (b) visual volume and magnitude of movements (the angle range is normalized as 1.0).



(a) vertical posture (S-V,W-V)

(b) horizontal posture (S-H,W-H)

(c) near posture (S-N,W-N)

(d) far posture (S-F,W-F)

Fig. 10. Inherent body identification; the reference frame, body part (shoulder), body part (wrist), body texture (shoulder), and body texture (wrist) are presented from left to right.

parts given by the shoulder and wrist movements. In the figures, S and W denote the shoulder and wrist that the robot moves. V, H, N and F denote the condition of the arm; vertical, horizontal, near and far, respectively. These are the fixed positions in joint space, which show four different representative arm postures. The reference frame was fixed for simplicity here, while in the learning phase of motor control the robot varies its neck and eyes (refer to Section IV).

The results of the experiments are summarized as follows;

- area, location, and distortion of the body parts were reliably detected (in the sense of the deviation value σ),
- orientation was comparably reliable for the shoulder part, but not for the wrist part because the major and minor motor axes can be easily swapped,
- area average $m(a)$ characterized the distance to the motor effector, and
- distortion average $m(\lambda)$ showed that the shape of the body parts defined by the shoulder and wrist movements were linear (close to 1.0) and circular (close to 0.5), respectively.

2) Extended body identification: We performed identification of the wrist motor unit in the case that an object is in the hand. Figures 9(a) and 11 show the mean (m) and standard deviation (σ) of the visual features. The items; Gf, Ga and Gb denote the type of grasp, free grasp, ball grasp and bottle grasp, respectively. The results of the experiments are summarized as follows;

- area average $m(a)$ characterized the volume of the extended body part, and
- distortion $m(\lambda)$ characterized a linear shape when grasping a bottle (Gb) compared to the free and ball grasp (Gf, Ga) that gave much less distortion in the hand shape.

In these demonstrations, the robot succeeded in identifying extended parts as its own body. The visual features of extended body parts are combined with proprioceptive information (described in Section IV-C).

3) Volume of body parts: We investigated the effects of the magnitude of the movements for body identification. Figures 9(b) and 12 show visual features of the body parts resulting from wrist movements. The norm of the variation vector $|\delta q_w|$ was set from 0.1 to 0.5 with step 0.1. As shown in the figures, we can conclude that

- the area average $m(a)$ is higher when the magnitude of movements is greater,

or, in other words, the variation term δq in the identification has to be small in order for the detected blob to fit the body part well.

(a) free grasping (Gf)

(b) object grasping A (Ga)

(c) object grasping B (Gb)

Fig. 11. Extended body identification
; the reference frame, body part (wrist), and body texture (wrist) are presented from left to right.

Fig. 12. The effect of the magnitude of movements; the reference frame and body part with different magnitude of movements, 0.1, 0.2, 0.3, 0.4, 0.5, are presented from left to right.

IV. Learning of Primitive Actions

Body identification allows the robot to learn primitive actions. In this section, we define learning of fixation, reaching and grasping actions, which will later be used as the building blocks of more complex manipulative actions. Figure 5 illustrates a diagram of sensory-motor signal flows. We assume the following motor units and corresponding primitive actions;

- head motor unit (fixation),
- arm motor unit (reaching), and
- finger motor unit (grasping).

The motor units give coordinated movements of multiple joints driven by activation signals from an action generator. The action generator is a module in a high-level motor execution system (refer to the module and relation to other modules in Fig. 5). We will detail its function in Section V. Our robot platform has two arms with hands. We therefore assume two independent arm and finger motor units.

The robot demonstrates motor exploration to learn the primitive actions in each motor unit. Motor exploration consists of two movements: one is a stroke movement to move joints to different angles, and the other is a repetitive short range movement for identification of a visual target (which can be the body part of another target). The robot first generates a random stroke movement in a motor unit to move the body part into a certain posture, and then generates perturbative movements to identify the body. After the movements, the robot associates the visual and tactile data with the proprioceptive data from the sampling posture.

After learning the data, the robot can estimate visual and tactile information from proprioceptive information and is also able to estimate the information in the opposite direction. This estimation is implemented as a look-up table in which the nearest data sample to input in visual/proprioceptive space is referred and this sample is used as the reference for a local linear interpolation that offers continuous data association.

In the following sections, we will describe the learning procedure for a primitive action in each motor unit (head, arm and finger motor units).

A. Head motor unit

1) Head motor exploration: The robot performs motor exploration with the head motor unit and it associates the resulting observed visual variation of the target. The head motor unit consists of motor joints in the neck and eyes. We mainly use neck pitch, neck yaw and eye vergence to localize a target in three-dimensional space. We do not discuss details here about the other DOF of the eyes for saccadic movements, however the robot can learn the movements in the same way.

We formulate the egocentric three-dimensional visual location of a target $\boldsymbol{z}$ as follows:

$$\boldsymbol{z} = (\xi^L, \eta^L, \xi^R - \xi^L), \tag{11}$$

where $\boldsymbol{x}^L = (\xi^L, \eta^L)$ and $\boldsymbol{x}^R = (\xi^R, \eta^R)$ denote the image coordinates of the target in the left and right images. We use the left frame as the reference. $\xi^R - \xi^L$ corresponds to the parallax.

The visual effect of the head motor exploration is given as follows:

$$\delta \boldsymbol{z} = J^h(\boldsymbol{q}, \boldsymbol{z})\delta \boldsymbol{q}, \tag{12}$$

where $\delta \boldsymbol{z}$ and $\delta \boldsymbol{q}$ denote a variation of the visual target location and the head posture, respectively. J^h represents the transformation matrix between them. The robot generates a posture variation $\boldsymbol{u} = \boldsymbol{q} + \delta \boldsymbol{q}$ and associates it with the observed visual position variation $\delta \boldsymbol{z}$. We assume a single joint variation:

$$\delta \boldsymbol{q}_i = (0, \cdots, dq_i, \cdots, 0), \tag{13}$$

for each i-th component. Therefore, the exploration result directly gives the i-th column of the transformation:

$$\boldsymbol{J}^h_i(\boldsymbol{q}, \boldsymbol{z}) = (1/dq_i)\delta \boldsymbol{z}_i, \tag{14}$$

where $\boldsymbol{J}^h_i$ and $\delta \boldsymbol{z}_i$ denote the i-th column vector of J^h and the observed vector of the visual variation.

Learning action-effect causality in the head motor unit allows bidirectional associations; vision to head proprioception (visual projection) and head proprioception to vision (visual fixation).

TABLE II
HEAD MOTOR UNIT, EXPERIMENTAL CONDITIONS

item	parameter	notation
motor unit	head	$\boldsymbol{q} \in R^6$
exploration part	neck with eyes	$\boldsymbol{u}_h \in R^3$
head state	down, front, right	Hd, Hf, Hr
arm state	near, far	An, Af

2) Visual projection: Visual projection aims at mapping memorized locations onto a view frame with a different viewpoint. This is effective for representing memorized visual locations taken at different viewpoints in a current frame. Given the current head joint posture $\boldsymbol{q}$, the location of $\boldsymbol{z}_i$ is estimated in the current frame as follows;

$$\hat{\boldsymbol{z}}(\boldsymbol{q}_i, \boldsymbol{z}_i; \boldsymbol{q}) = \boldsymbol{z}_i + \hat{J}^h_k(\boldsymbol{q} - \boldsymbol{q}_i), \quad (15)$$

$$\hat{J}^h_k = J^h(\boldsymbol{q}_k), \quad (16)$$

$$k = \arg\min_j |\boldsymbol{q} - \boldsymbol{q}_j|, \quad (17)$$

where $(\boldsymbol{q}_i, \boldsymbol{z}_i)$ denotes a set of head posture and visual location in the memory (learned sample). $(\boldsymbol{q}, \hat{\boldsymbol{z}}(\boldsymbol{q}_i, \boldsymbol{z}_i; \boldsymbol{q}))$ denote the current head posture and the estimated visual location in the current frame. $\hat{J}^h_k$ represents the estimated transformation at $\boldsymbol{q}_k$.

3) Visual fixation: The opposite association gives visual fixation, that is, the coordinated neck and eye movement to bring a target to the center of the view frame. Given the desired location $\boldsymbol{z}^d$ (the center of the view frame), the head joint posture to allow for visual fixation is estimated as follows;

$$\hat{\boldsymbol{q}}(\boldsymbol{q}, \boldsymbol{z}; \boldsymbol{z}^d) = \boldsymbol{q} + \hat{J}^{h\#}_k(\boldsymbol{z}^d - \boldsymbol{z}), \quad (18)$$

$$\hat{J}^h_k = J^h(\boldsymbol{q}_k), \quad (19)$$

$$k = \arg\min_j |\boldsymbol{q} - \boldsymbol{q}_j|, \quad (20)$$

where $(\boldsymbol{q}, \boldsymbol{z})$ denotes the current head posture and the visual location of the target, and $(\hat{\boldsymbol{q}}(\boldsymbol{q}, \boldsymbol{z}; \boldsymbol{z}^d), \boldsymbol{z}^d)$ denotes the estimated head posture and the goal location to bring the target into. In visual fixation, we assign the coordinates of the center of the view frame for $\boldsymbol{z}^d$, although the goal location is not limited to this (i.e. in theory, the robot can bring the target into any location of the view frame). $\hat{J}^{h\#}_k$ represents the generalized inverse $\hat{J}^h_k$ at $\boldsymbol{q}_k$.

4) Experiments: We examined visual projection and fixation with the head motor unit. Table II summarizes the experiment conditions. The head motor unit has 6 DOF. $\boldsymbol{q} \in R^6$ denotes a joint angle vector given by the motor encoders (the values were normalized in [-1,1]). The variation is defined as $\delta\boldsymbol{q} = (\delta q_0, \delta q_1, \delta q_5)$. The suffix of variables corresponds to the joint number in Fig. 8. We used the body parts as a visual target in head motor exploration. We believe the use of body parts for learning to be a natural solution for the following reasons; the reachable area is the most important area for the robot to learn; the appearance of the robot's body parts can be visually unique in the view frame and the robot is able to move the location of its own body parts autonomously while learning.

(a) Visual projection: In this experiment, we evaluate visual projection ability at each of four different joint postures. First, the robot performed head motor exploration (body identification and learning of transformation $J^h(\boldsymbol{q}, \boldsymbol{z})$) at a single joint posture $\boldsymbol{q}$, and then the robot randomly moved the joints of its head motor unit around the learned joint posture in order to sample tuples of a head posture and target location $\{\boldsymbol{q}_i, \boldsymbol{z}_i\}_{i=1,\cdots,n}$ for the evaluation. The estimation of multiple joint postures with a learned single joint posture does not result in a loss of generality because the location is estimated locally at the nearest learned joint posture (refer to Eq. 17), and the estimation is independent from other learned joint postures. The test tuples were sampled as follows:

$$\boldsymbol{u} = \boldsymbol{q} + \delta\boldsymbol{q}, \quad (21)$$

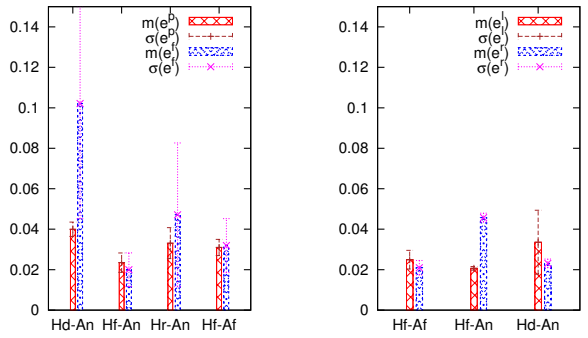
where $\boldsymbol{u}, \boldsymbol{q}$ and $\delta\boldsymbol{q}$ denote the head motor command, head joint angle and its variation. Each component of $\delta\boldsymbol{q}$ was given from the uniform distribution in $[-\alpha, \alpha]$ where α is a positive constant. In the following experiments, we used the value $\alpha = 0.2$, corresponding to a variation of 40 % of range from the learned joint posture. The robot sampled 10 test tuples. We used the right hand of the robot as a visual target.

After sampling, the robot estimated the visual location $\boldsymbol{z}_i$ at each head posture $\boldsymbol{q}_i$. The estimated location is noted as $\hat{\boldsymbol{z}}(\cdot, \cdot; \boldsymbol{q}_i)$. In evaluating the estimations, one sample was used as a ground-truth sample, and other samples were used for estimation. The estimation error of the i-th ground-truth sample e_i is formulated as follows:

$$e_i = \sum_{j=1,\cdots,n, i\neq j}^{n} |\boldsymbol{z}_i - \hat{\boldsymbol{z}}(\boldsymbol{q}_j, \boldsymbol{z}_j; \boldsymbol{q}_i)| / (n-1), \quad (22)$$

where $m(e) = 1/n \sum_{i=1}^n e_i$ and $\sigma(e) = 1/n \sum_{i=1}^n |e_i - m(e)|$ denote the average and deviation of the estimation error.

Figures 13(a) and 14(a) show the results of the visual projection. In Fig. 13(a), $m(e_p)$ and $\sigma(e_p)$ denote the average and deviation of the estimation in the visual projection. The labels Hd, Hf and Hr denote the head joint posture corresponding to down, front and right. The labels An and Af denote the arm joint posture posing as positioned near and far from the head, respectively. In the experiments, we evaluated

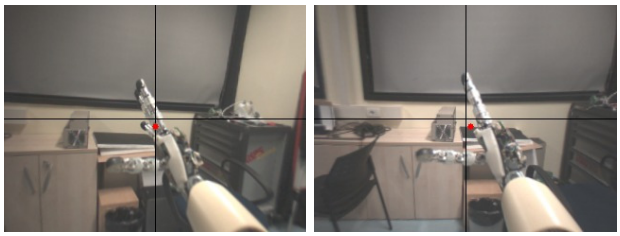


(a) visual projection and fixation

(b) arm localization and reaching

Fig. 13. Estimation error. (a) visual projection and fixation after head motor exploration. (b) arm localization and arm reaching after arm motor exploration.

(a) Visual projection (left and right sight)



(b) Visual fixation (left and right sight)

Fig. 14. Results of head motor exploration. (a) Visual projection of a target (the robot hand). Red dots are estimated locations, and green dots are ground-truth locations. (b) Visual fixation of a target (own hand). The red dot is the ground-truth location of the target sampled by wrist body identification after fixation. (a) and (b) present the results for the Hd-An condition. The results for other conditions, Hf-An, Hr-An, and Hf-Af, are similar to these (we have not presented the pictures in order to save space in the paper).

the head-arm posture combinations of Hd-An, Hf-Ad, Hr-An and Hf-Af. We believe these four types of combinations represent the most typical and different posture relations of the head and arm. The robot collected the corresponding transformation values ($J^h(\boldsymbol{q},\boldsymbol{z})$) for each head joint posture and visual location pair $(\boldsymbol{q},\boldsymbol{z})$. As explained above, the robot learned the transformation at each head-arm joint posture and evaluated an estimation of the visual location of the arm with variations within 40% of the range of each joint angle. As we can see in the figures, the estimated samples were projected quite close to the ground-truth sample with small deviations in different target conditions. We can easily improve the accuracy of the estimations by increasing the number of head-arm joint postures from which the robot learns the linear transformation.

(b) Visual fixation: After learning visuo-proprioceptive association, the robot performed visual fixation at the target locations sampled in the previous experiment. The desired visual location is $\boldsymbol{z}^d = (w/2r, h/2r, 0)$ where w, h, r denote the width, height and diagonal length of the view frame, respectively.

At the i-th tuple $(\boldsymbol{q}_i, \boldsymbol{z}_i)$, the robot estimated the head joint posture $\hat{\boldsymbol{q}}_i = \hat{\boldsymbol{q}}(\boldsymbol{q}_i, \boldsymbol{z}_i; z^d)$ to fixate the target, and commanded this posture as $\boldsymbol{u}_n = \hat{\boldsymbol{q}}_i$. After fixation, the robot performed wrist motor exploration to re-sample the target location $\boldsymbol{z}'_i$ at the same head posture $\hat{\boldsymbol{q}}_i$. Therefore, $\boldsymbol{z}'_i$ gives the ground-truth location of the target. The estimation error of the i-th sample is formulated as follows:

$$e_i = |\boldsymbol{z}^d - \boldsymbol{z}'_i(\hat{\boldsymbol{q}}_i)|. \tag{23}$$

Figures 13(a) and 14(b) show the results of the visual fixation. In the table, $m(e_f)$ and $\sigma(e_f)$ denote the average and deviation of the estimation in the visual fixation. As shown in the figures, targets in different configurations are fixated with high precision.

B. Arm motor unit

1) Arm motor exploration: The robot uses the arm motor unit to generate motor exploration and associates the observed visual variation of the body part with the action. This aims at building arm image and motor control in visual space. We formulate arm motor exploration in the same framework as head motor exploration, as follows:

$$\delta\boldsymbol{z} = J^a(\boldsymbol{q}, \boldsymbol{z})\delta\boldsymbol{q}, \tag{24}$$

where $\delta\boldsymbol{z}$ and $\delta\boldsymbol{q}$ denote a variation of the target's visual location and the arm posture, respectively. J^a represents a transformation between them. The robot generates a posture variation $\boldsymbol{u} = \boldsymbol{q} + \delta\boldsymbol{q}$ and associates it with the observed visual variation of position $\delta\boldsymbol{z}$. The exploration schema and visual coordinates of the body parts are formulated by Eq.13 and 14 substituting J^a for J^h. We mainly use shoulder pitch, yaw and roll, and elbow pitch in exploration.

Learning action-effect causality in the arm motor unit allows for bidirectional associations from vision to arm proprioception (arm image) and from arm proprioception to vision (arm reaching). This motor exploration supposes that the hand is not occluded while learning, for if the hand were occluded, the robot would not be able to construct a correct visuo-proprioceptive association. However, learning is driven by body identification. If the robot identifies its own hand, it will memorize sampled data of the association. If not, the robot will not memorize sampled data and randomly vary its arm posture to reattempt body identification in a different position. This procedure minimizes the situations that the hand is not visible in the image. In this work, we simply used uniform distribution for random exploration, but more sophisticated approaches, such as performance-biased random exploration in [14], could be applied.

2) Arm image: The arm image aims at mapping the body silhouette onto a view field. This function permits estimation of positions and visual appearances of the body parts from proprioception.First, the system recalls its own visual features corresponding to the current arm posture (the look-up procedure is similar to the one described previously for the head visual projection in Section IV-A2). The binocular visual location of the arm image is then estimated as follows:

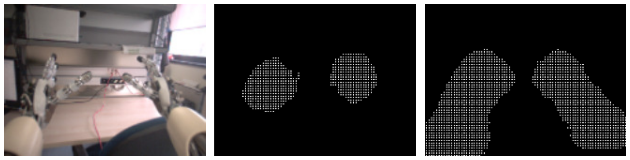
$$\hat{\boldsymbol{z}}(\boldsymbol{q}_i, \boldsymbol{z}_i; \boldsymbol{q}) = \boldsymbol{z}_i + \hat{J}^a_k(\boldsymbol{q} - \boldsymbol{q}_i), \tag{25}$$

$$\hat{J}^a_k = J^a(\boldsymbol{q}_k), \tag{26}$$

$$k = \arg\min_j |\boldsymbol{q} - \boldsymbol{q}_j|, \tag{27}$$

where $(\boldsymbol{q}_i, \boldsymbol{z}_i)$ denotes a set of arm posture and visual location in the memory. $(\boldsymbol{q}, \hat{\boldsymbol{z}}(\boldsymbol{q}_i, \boldsymbol{z}_i; \boldsymbol{q}))$ denotes the arm posture and estimated visual location.

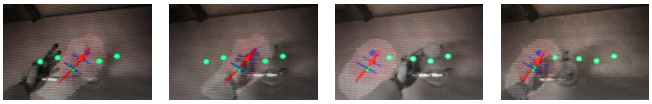
This procedure compensates for translation only. In theory, it could easily be extended using the affine transformations to handle rotations, but this would require high dimensional



(a) arm image

(b) visual occlusion

Fig. 15. Arm image. (a) the reference, hand domains and forearm domains are presented from left to right. (b) the reference before occlusion, the reference after occlusion and the hand image while occluded are presented from left to right.



(a) (b) (c) (d)

Fig. 16. Anticipation of arm and hand locations in object operation. The hand and forearm visual appearances are presented as a pink and white transparent cloud. The red dot with the red and blue segment represent the anticipated visual location, and the major and minor axes of the arm, respectively. The green dots represent learned visual locations. The time course of pictures is from left to right. In (a) and (b), the robot is reaching for a bottle and grasping it. (a) shows the expected location and shape of the arm/hand at the end of the movement. (b) shows the arm/hand postures after the reaching and grasping. Similarly in (c) and (d), but this time the arm is going back to its initial position.

J^a. Instead, we simply sampled different arm postures and interpolated the body image based on the sampled ground-truth locations with low dimensional J^a. This is practical for implementations in real robot platforms and supposes local linearity around the sampled points. In fact, it was successful in the following experiments.

Figure 15(a) shows the arm image estimated after learning. Four body parts (left hand, left forearm, right hand and right forearm) are projected in Fig. 15(a). In general, it is not easy to visually identify the left and right hand in the same frame, since their appearances are similar. On the other hand, the proprioceptive identification in Fig. 15(a) is distinctive and it even works for building an arm image when the arm is occluded as shown in Fig. 15(b). Note that in theory, we can assume the number of arm images to be equivalent to the number of motor units in the robot.

3) Arm reaching: Arm reaching aims at motor control of the arm to move the hand to a destination in three dimensional space. Given a current head posture $\boldsymbol{q}^h$ and a desired location $\boldsymbol{z}^d$, a motor command of the arm posture $\hat{\boldsymbol{q}}$ is estimated as follows:

$$\hat{\boldsymbol{q}}(\boldsymbol{q},\boldsymbol{z};\boldsymbol{z}^d) = \boldsymbol{q} + \hat{J}^{a\#}(\boldsymbol{q})(\boldsymbol{z}^d - \boldsymbol{z}), \tag{28}$$

$$\hat{J}^a_k = J^a(\boldsymbol{q}_k), \tag{29}$$

$$k = \arg\min_j |\boldsymbol{q} - \boldsymbol{q}_j|, \tag{30}$$

where $(\boldsymbol{q},\boldsymbol{z})$ denotes a reference arm posture and reference location. $(\hat{\boldsymbol{q}}(\boldsymbol{q},\boldsymbol{z};\boldsymbol{z}^d),\boldsymbol{z}^d)$ denote an estimated arm posture and the desired visual location. $\hat{J}^{a\#}_k$ represents the generalized inverse $\hat{J}^a_k$ at $\boldsymbol{q}_k$.

The reference $(\boldsymbol{q},\boldsymbol{z})$ can be given either in feed-forward or feedback mode. The feed-forward mode is a memory-based ballistic reaching that moves the arm into sight. The references are given by the memory as follows:

$$(\boldsymbol{q},\boldsymbol{z}) = (\boldsymbol{q}_i,\boldsymbol{z}_i), \tag{31}$$

$$i = \arg\min_j |\boldsymbol{q}^h - \boldsymbol{q}^h_j|, \tag{32}$$

where $(\boldsymbol{q}^h_i,\boldsymbol{q}_i,\boldsymbol{z}_i)$ denotes the head posture, the arm posture and visual location in the memory. In the feedback mode, the current arm posture and the identified visual location at the current head posture are given as reference $(\boldsymbol{q},\boldsymbol{z})$. The difference between the feed-forward and the feedback mode is that the former uses a memorized hand location and the latter uses the current hand location identified by visual recognition or visual-proprioceptive body identification. Note that the visual location depends on the head posture $\boldsymbol{q}^h$. In all the above computation, the visual location $\boldsymbol{z}$ is adjusted to fit the current head posture by visual projection, as described in Section IV-A2.

Figure 16 shows the anticipation of arm and hand locations in object operation. When the robot identifies an object of interest (the bottle, in this case), it anticipates the reaching and grasping postures. The robot, then, executes the task. Consequently, it selects the visual location towards the object to be moved. Using Eq.28, the robot estimates the arm posture from which it predicts the expected final appearance of the arm and hand in the visual field (Eq.25), substituting $\hat{\boldsymbol{q}}$ for $\boldsymbol{q}$ with compensation of the head posture using Eq.15. Grasping posture anticipation with visual object recognition is detailed in Section IV-C.

4) Experiments: We examined arm localization for arm image and arm reaching with the arm motor unit. Table III summarizes the experimental conditions. The arm motor unit has 7 DOF for each arm. $\boldsymbol{q} \in R^7$ denotes a joint angle vector given by motor encoders (the values were normalized in [-1,1]). The variation is defined as $\delta\boldsymbol{q}^a = (\delta q^{a0}, \delta q^{a1}, \delta q^{a2}, \delta q^{a3})$ where the suffix of variables corresponds to the joint number in Fig. 8. In the experiments, we evaluated the head-arm posture combinations of Hf-Af, Hf-An and Hd-An where Hf and Hd denote the front and down head postures, and Af and An denote the far and near arm postures, respectively.

(a) Arm localization: First, the robot performed arm motor exploration as described in Section IV-B1, and learned

TABLE III
ARM MOTOR UNIT, EXPERIMENTAL CONDITIONS

item	parameter	notation
motor unit	arm	$\boldsymbol{q} \in R^7$
exploration part	shoulder, elbow	$\boldsymbol{u}_a \in R^4$
head state	front down	Hf, Hd
arm state	near, far	An, Af

the visuo-proprioceptive association. The robot then sampled tuples of an arm posture and hand location, $\{\boldsymbol{q}_i, \boldsymbol{z}_i\}_{i=1,\cdots,n}$ to evaluate the learning results. 20 tuples were sampled by generating random arm postures around the learned arm posture $\boldsymbol{q}$ as follows:

$$\boldsymbol{u}^a = \boldsymbol{q} + \delta\boldsymbol{q}, \tag{33}$$

where $\boldsymbol{u}^a$ and $\delta\boldsymbol{q}$ denote the arm motor command and the arm variation. Each component of $\delta\boldsymbol{q}$ was given from the uniform distribution in $[-\alpha, \alpha]$ where α is a positive constant. We used the right hand as a target body part. The head posture was fixed in each condition (while it is variable in experiment (b)).

The robot estimated visual location $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{z}}(\cdot,\cdot;\boldsymbol{q}_i)$ from head posture $\boldsymbol{q}_i^h$. $\boldsymbol{z}_i$ was used as the ground-truth location of the hand as follows:

$$e_i^l = |\hat{\boldsymbol{z}}_i - \boldsymbol{z}_i|, \tag{34}$$

where e_i^l denotes the error in localization for the i-th sample.

Figures 17(a) and 13(b) show the results of arm localization with $\alpha = 0.2$. $m(e^l)$ and $\sigma(e^l)$ denoting the average and deviation of the localization error, respectively. As shown in the table, the arm was localized with a high level of precision given the maximum variation of 0.2 between arm and reference postures.

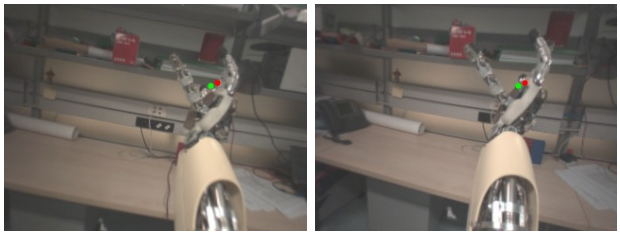
(b) Arm reaching: We performed arm reaching coordinated with visual projection. In contrast to experiment (a), the head posture was variable. After learning the visuo-proprioceptive association with the head and arm motor unit, the robot performed arm reaching.

The task of the action was to move the hand to an anonymous fixation point. The robot generated a set of random head postures $\{\boldsymbol{q}_i^h\}_{i=1,\cdots,n}$ based on Eq.21, and estimated the arm posture $\hat{\boldsymbol{q}}_i$ to reach the view center $\boldsymbol{z}^c$ for each head posture. The i-th error of effector reaching is defined as follows:

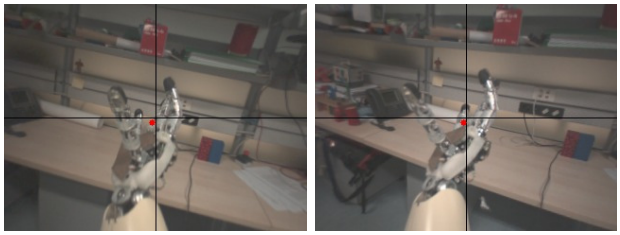
$$e_i^r = |\boldsymbol{z}_i'(\hat{\boldsymbol{q}}_i) - \boldsymbol{z}^c|. \tag{35}$$

where $\boldsymbol{z}_i'$ denotes the ground-truth location of the hand sampled after reaching. The average and deviation of the error are denoted as $m(e^r)$ and $\sigma(e^r)$, respectively.

As shown in Fig. 17(b) and Fig. 13(b), arm reaching towards fixated points was successful in the different head postures. Note that this method does not use the external world coordinates to accomplish three dimensional reaching, but manages it with internal coordinates (horizontal and vertical components in the left frame with vergence of the left and right frames). Therefore, arm localization and arm reaching are achieved with both left and right cameras.



(a) Arm localization



(b) Arm reaching

Fig. 17. Results of arm motor exploration. (a) Arm localization. The green and red dots represent the estimated and ground-truth locations of the hand, respectively. (b) Arm reaching. The red dot represents the ground-truth location of the hand. The black lines indicate the horizontal and vertical center. (a) and (b) represent the result for the Hf-An condition. The results for the other conditions, Hf-Af, Hd-An, are similar to this; (we have not presented the pictures in order to save space in the paper).

C. Finger motor unit

1) Finger motor exploration: The robot uses the finger motor unit to perform motor exploration with an object, and associates the observed somatosensory event with the features of action and the object. The objects are detected by the visual attention system in advance (detailed as in Section IV-C2).

We define finger motor exploration based on grip sensing as follows:

$$u_i^f = \begin{cases} q_i + \delta q_i & \text{if } \ g_i < g_0, \\ q_i & \text{otherwise,} \end{cases} \tag{36}$$

where $\boldsymbol{u}^f$ and $\boldsymbol{q}$ denote the motor command vector and encoder value vector of finger joint angles, respectively. $\boldsymbol{g}$ denotes the reaction grip (as defined below). The suffix i corresponds to the finger joint number in Fig. 8. The robot continues to fold each finger joint unless the corresponding reaction grip reaches a limit g_0. When all the joints stop folding, the finger posture vector is memorized. The reaction grip should be given by a torque sensor. Our robot platform, however, is not equipped with such sensors in the finger joints, though the joints are mechanically compliant. We employed this compliance to simulate the reaction grip; the reaction grip is defined as the difference between the motor command and the joint position as follows:

$$g_i(t) = |u_i^f(t) - q_i(t)|. \tag{37}$$

Note that there is an elastic coupling (spring) between the motor and the corresponding finger joint so that the current

position of the motor and the finger joint are different in the presence of a contact force that deflects the spring.

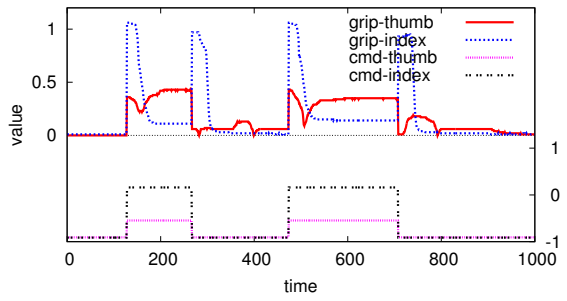
2) Visual attention: When the robot completes a trial in finger motor exploration, a tuple consisting of the object's visual blob I_b and the final finger posture $\boldsymbol{q}$ is stored. After learning, the robot visually identifies the object and grasps it with the associated finger posture. This visuo-proprioceptive association of objects and finger postures allows feed-forward finger shaping to grasp the object.

We implemented simple visual attention based on a motion cue that detects objects. In computer vision, visual attention is implemented based on various cues such as edges, colors and shapes [15]. In man-robot interaction, however, motion cues are considered more useful; we can easily inform a robot of an object of interest by moving it, while the other visual cues cannot be easily controlled by human partners.

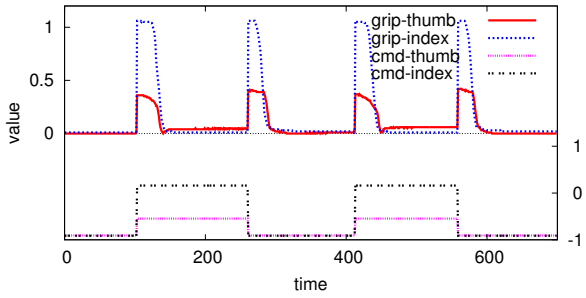
3) Experiments: We examined reaction grip detection with the finger motor unit and visual attention for objects to be grasped.

(a) Reaction grip: Figure 18 shows profiles of the reaction grip in relation to the motor command while grasping and releasing. We let the robot perform successive grasps and releases two times: i.e., demonstration of grasp, release, grasp, release and stop. We compared the results for two different conditions. The first condition was action with an object (object grasp); the second condition was action without an object (free grasp). During the initial and third phase of motor commanding in Fig. 18(a), we can see the non-zero $g_i(t)$ caused by object grasping. On the other hand, during those periods in Fig. 18(b) $g_i(t)$ converged to zero. As shown in the figures, the reaction grip $g_i(t)$ differentiates the two conditions correctly.

(b) Visual attention: Figure 19 shows motion-based visual attention and detected objects. This attention system detects an object that maintains motion for several seconds (set as 3s in the experiments). A similar mechanism can inform the robot of the experimenter's hand as shown in Fig. 19(d). This function is also used for the perception of human actions in Section V.

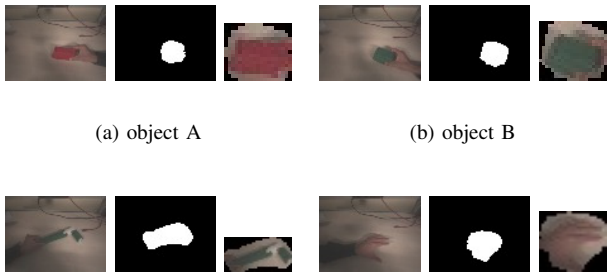


(a) object grasp



(b) free grasp

Fig. 18. Profiles of reaction grip. (a) object grasping and releasing. (b) free grasping and releasing. In each figure, two profiles of grip force (upper half) and two profiles of motor command (lower half) are presented. The profiles correspond to the joints in the thumb and index finger.



(a) object A (b) object B

(c) object C (d) object d

Fig. 19. Motion-based visual attention. The reference frame, attracted domain and detected object are presented from left to right in each target object.

V. Perception of Manipulative Behaviors

We will now propose a series of actions that can be shared between humans and robots. We characterize manipulative behaviors based on their effects on the geometrical relation of the hand and object (as defined in Section V-A). The characteristics of the proposed action perception system are summarized as follows:

- the action perception system is developed by observing the robot's self-generated actions,
- the motor repertoire is constructed incrementally by combining learned primitives,
- the sensory effect of an action is encoded in multi-modal sensory space,
- human actions are predictively recognized via intermediate evaluation of the sensory effect, and,
- action perception allows cross-modal sensory anticipation and action reproduction.

Some features of the action perception system are consistent with mirror systems in nature [3][4][7] and allow for more complex manipulative behaviors (e.g. a sequential combination of grasp, hold and drop). In the following sections, we formulate the processes of visual, proprioceptive and tactile sensing and a multi-sensory action perception system.

A. Sensory effects of actions

We will now propose a way to quantify the effects of actions. Actions of interest in this work are those that both

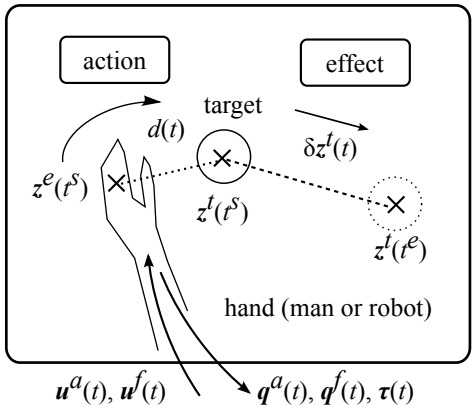


Fig. 20. A schematic representation of action perception. The motor command $\{u^a, u^f\}$, proprioceptive feedback $\{q^a, q^f\}$ and tactile feedback τ are available when the action is self-generated. The visual location of the hand and target z^e and z^t are available regardless of the action agent (human or robot).

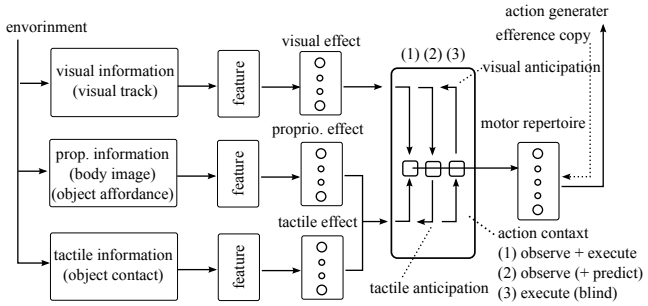


Fig. 21. Action perception system. The sensory features are classified into a visual, proprioceptive and tactile class. An action can be recognized either from all or one of the available modality classes. The cross-modal anticipation is computed for any missing sensory signals (refer to Section V-A). The parameters are learned from the self-generated actions.

robots and humans can perform using some objects in a similar manner. We assume that the effect of an action toward an object can be quantified by change in perception before and after the action. We use visual cues to quantify changes in the geometry of the action effector (i.e. hand) and operated object. We also use proprioceptive and tactile cues to quantify changes in sensing due to physical interaction with objects.

Figure 20 is a schematic representation of action perception. In the figure, z^e and z^t denote the position of a hand of either a robot or an human experimenter and a target in the view frame, respectively. Motor command of the arm and fingers $\{u^a, u^f\}$, proprioceptive feedback of those joint postures $\{q^a, q^f\}$, and tactile feedback τ are available, when the agent of the action is the robot itself. The superscript a and f indicate the joint angle vector of the arm and finger motor unit, respectively.

To detect locations z^e and z^t, we used the visual tracking system proposed previously in [10], which tracks hands and objects based on color and edge features. Here we let the robot memorize the appearance of a human hand by using a visual attention system (Section IV-C2). We then let the robot perform body identification (Section III-B) and forward the resulting appearance of its own hand to the visual attention system as input to memorize. In the following experiments, hand locations were detected successfully against small variations of hand shapes of the robot and human experimenter during movements of the arm and grasping of an object. We believe that successful detection resulted from less variation of edge information for the robot hand and less variation of color information for the human hand.

We define visual feature $f^v = \{\delta z^t, \delta d\}$ as follows:

$$\delta z^t(t) = z^t(t) - z^t(t^s), \tag{38}$$
$$\delta d(t) = d(t) - d(t^s), \tag{39}$$
$$d(t) = |z^e(t) - z^t(t)|, \tag{40}$$

where δz^t and δd represent the change in the target position and the change in distance between the target and the hand, respectively. t^s is the time the action starts. The feature f^v encodes the visual effect on the hand and object state caused by an action. We assume that the human hand and target are visually tracked. The action agency is confirmed by the following visuo-proprioceptive contingency:

$$s = \sum_{t=t^s}^{t^e} |z^e(t) - \hat{z}^a(t)|, \tag{41}$$

where $\hat{z}^a(t)$ denotes the location of the self's hand estimated from arm proprioception using Eq.15 and Eq.25. t^e denotes the time the action ends. The system recognizes self agency when the value of s is below a threshold.

We define a proprioceptive feature $f^m = \{\delta\hat{z}^a, \delta d^f\}$ as follows:

$$\delta\hat{z}^a(t) = \hat{z}^a(t) - \hat{z}^a(t^s), \tag{42}$$
$$\delta d^f(t) = d^f(t) - d^f(t^s), \tag{43}$$
$$\hat{z}^a(t) = \hat{z}(q^a(t)), \tag{44}$$
$$d^f(t) = |q^f(t) - \hat{q}^{f(T)}|, \tag{45}$$

where $\hat{z}^a(t)$ denotes the estimated hand location (defined above), and $d^f(t)$ represents the distance between the current finger posture q^f and the finger posture $\hat{q}^{f(T)}$ corresponding to the visually identified object T to be grasped (see Section IV-C2).

We define a tactile feature $f^\tau = \{\tau(t^s), \tau(t)\}$ as follows:

$$\tau(t) = \max_i \tau_i(t), \tag{46}$$

where τ denotes the maximum tactile intensity of all fingers. τ_i denotes the summation of all tactile sensor values on the i-th fingertip. This maximization relaxes ambiguity of contact conditions.

Note that perception of all of the features mentioned above is based on previously developed systems of body perception, visual attention and motor skills. In particular, learned primitive actions (reaching and grasping) play an important role in realizing complex behaviors for manipulation to be perceived developmentally in the action perception phase.

B. Action generation and action perception

We will now set an action perception based on the above-defined multi-modal sensory features $\{f^v, f^m, f^\tau\}$. The pro-

posed action perception system is illustrated in Fig. 21. In action perception, we assume three action contexts;

(AC1) observation and execution,
(AC2) predictive observation, and
(AC3) blind execution.

(AC1) represents the robot's action execution and simultaneous observation of the action. This context is used in the action learning phase and reproduction phase to perform recognized actions. (AC2) represents predictive observation of actions performed by human experimenters. Predictive action perception is made possible by intermediate evaluation of the sensory effect, which was inspired by [16]. (AC3) represents the robot's action execution in a blind condition. After a one-shot visual detection of a target object in the work space, the robot executes an action without visual information. This property simulates monkeys' mirror neurons that are active while grasping an object in a blind condition [3].

In the learning phase, the sensory features $\{\boldsymbol{f}^v,\boldsymbol{f}^m,\boldsymbol{f}^\tau\}$ at the end of the actions are stored. When a certain number of sensory features have been learned, the system updates the clustering parameters. Clustering sensory features aids in reducing computations in action recognition, and discretization by clustering allows for the application of a naive Bayesian estimation.

We used the k-means algorithm [17] for unsupervised clustering of each sensory feature as follows:

$$v_i = \begin{cases} 1 & (|\boldsymbol{f}-\boldsymbol{w}_i| \leq |\boldsymbol{f}-\boldsymbol{w}_j|, \forall j) \\ 0 & (\text{otherwise}) \end{cases} \quad (47)$$

where $\boldsymbol{f}$ denotes the input vector (either of $\{\boldsymbol{f}^v,\boldsymbol{f}^m,\boldsymbol{f}^\tau\}$), and $\boldsymbol{v}$ denotes the output vector following the winner-takes-all rule. $\{\boldsymbol{w}_i\}_{i=1,\cdots,n_c}$ denotes a set of prototype vectors; (n_c denotes the number of the classes). A single component of $\boldsymbol{v}$ is activated (i.e. the best match class), and the other component values are deactivated. Consequently, the sensory effect class is defined as $e = \arg_i\{v_i = 1\}$. e^v, e^p and e^τ denote the visual, proprioceptive and tactile effect class, respectively. In the following, e_i represents either of $\{e^v, e^p, e^\tau\}$. For learning, we used a standard learning rule detailed in [17].

Action perception is modeled based on the causal relation between an action and the corresponding effect. We represented the causal relation with the Bayesian rule as follows:

$$\begin{aligned} &\hat{a}(E = (\cdots, e_i, \cdots)) \\ &= \arg\max_a \; p(A = a) \prod_{i=1}^{n} p(E_i = e_i | A = a), \end{aligned} \quad (48)$$

where a denotes the action class, which corresponds to a category of actions in the motor repertoire. When an action is executed, its action class is given by the action generator, like an efference copy of a motor command in biological systems. The efference copy is known as a neural signal of a motor command originating in the central nervous system in motor control domains [18]. When another agent's action is observed, the action class is estimated from the sensory effect classes. An action is a single continuous movement composed of the reaching and grasping primitive learned in the earlier phase. In our implementation, the action generator module (refer to illustration of the module in Fig. 5) decodes the i-th action class a_i as a sequence of primitive actions $\{a_i^0, a_i^1, \cdots\}$, and sends signals to corresponding primitive action modules in the same order. For example, the grasping action class (detailed in Section V-C) is composed of the grasping primitive action and the reaching primitive action, and the action generator sends execution commands in that order.

TABLE IV
ACTION PERCEPTION, EXPERIMENTAL CONDITIONS.

action	number of trials (agent)	initial
grasp	10 (robot), 10 (human)	free
place	10 (robot), 10 (human)	grasp
hold	10 (robot), 10 (human)	grasp
drop	10 (robot), 10 (human)	grasp
poke	10 (robot), 10 (human)	free

E_i and A represent corresponding random variables. Probabilities are given by a set of learned tuples composed of the efference copy of action, visual, proprioceptive and tactile effect class. The data set is learned in (AC1). In (AC2), only the visual effect class is used as the sensory effect, while in (AC3) the proprioceptive and tactile effect class are used. For simplicity, we assume that each E_i is conditionally independent of every other E_j for $j \neq i$.

Multi-modal action perception allows for the estimation or recovery of unavailable sensory modality information during action observation and execution. We propose cross-modal sensory image (sensory anticipation) as follows:

$$\begin{aligned} &\hat{e}'(E = (\cdots, e_i, \cdots)) \\ &= \arg\max_{e'} \; p(E' = e') \prod_{i=1}^{n} p(E_i = e_i | E' = e'). \end{aligned} \quad (49)$$

In (AC2), e' denotes the tactile class (e^τ), which gives a tactile anticipation from visual observation of an experimenter's action. In (AC3), e' denotes the visual effect class (e^v), which gives a visual anticipation from the self's action execution in a blind condition.

C. Experiments

We performed experiments to evaluate the perceptual ability of the action perception system. An experimenter and robot performed manipulative behaviors. The types of actions and the number of trials are listed in Table IV. In the table, the *grasp action* denotes an action to reach an object and grasp it. The *place action* denotes an action to release an object and retract the hand. The *hold action* denotes an action to hold up a grasped object. The *drop action* denotes an action to release an object when holding it up. The *poke action* denotes an action to side-push an object. All actions were composed of the fixation, reaching and grasping primitives learned in the earlier phase. In the experiments, the actions were performed by both the robot and a human experimenter.

In the learning phase, we let the robot generate actions in the motor repertoire and simultaneously observe the sensory effect of the actions. Tuples of the actions and the sensory effects were used to develop action perception. After the learning

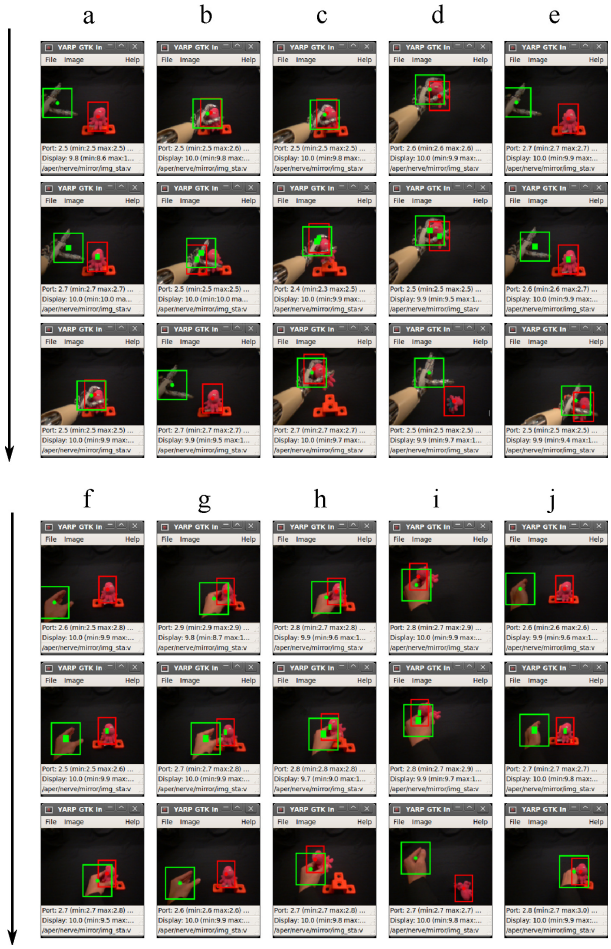


Fig. 22. Snapshots of actions performed by a robot and an experimenter. All of the actions are observed by the robot's vision system. (a) to (e) present the grasp, place, hold, drop and poke actions performed by the robot. (f) to (j) present the actions performed by an experimenter. The arrow indicates the time course.

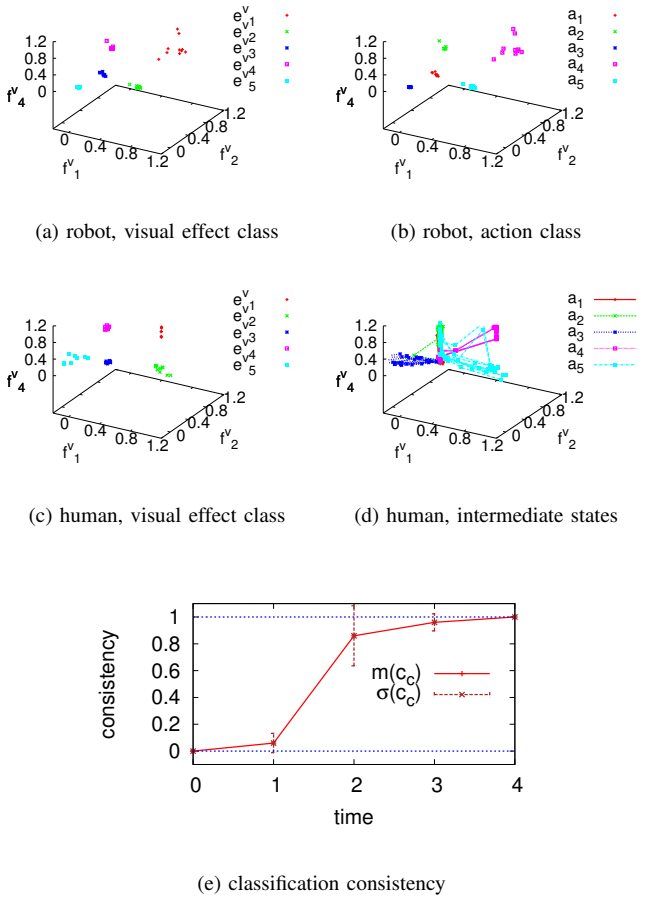


Fig. 23. Visual features of actions. The actions were performed by the robot and a human experimenter. The features are labeled the visual effect class in (a) and (c), and the action class in (b) and (d). (e) plots consistency in the predictive classification of the experimenter's actions. The horizontal and vertical axes indicate consistency in time and classification.

phase, an experimenter performed the actions, and the robot recognized the observed actions.

When the robot observed an action performed by itself, the robot was aware of the timing of the start and end of actions from its own proprioceptive signals. When the robot observed an action performed by an experimenter, we manually informed the robot of the timings for simplicity. The system, however, has an autonomous mode to detect action timing by monitoring increases and decreases in the area of visual motion in the view frame and segmenting a sequence of the action.

We excluded failed actions from the evaluation in order to focus the evaluation on action perception rather than motor control (although the failure rate was less than 10% of all trials). In the experiments, the location of the target was not precisely controlled, but the robot adapted its actions to the environment.

Snapshots of actions performed by a robot and an experimenter are shown in Fig. 22. All of the actions are observed by the robot's vision system. In preliminary experiments, we performed actions with different objects (that were acceptable for dropping) and got similar results in sensory effects. Below, we present and discuss the results obtained with a single object to eliminate noise from the comparison of perception in different modalities.

1) Visual effect: We analyzed the experimental results of visual effect classification and action recognition. The actions were performed by either the robot or the human experimenter and in both cases the robot recognized the actions using vision only (without proprioception and tactile information) in order to compare the results with different action agents in the same condition.

The clustering results of visual features are shown in Fig. 23. Figures 23(a) and (b) plot a set of visual features $\boldsymbol{f}^v(t^e)$ of the actions performed by the robot. Here, $\boldsymbol{f}^v = (\delta z_1, \delta z_2, \delta z_3, \delta d)$, but we present the following three components; $(f_1, f_2, f_4) = (\delta z_1, \delta z_2, \delta d)$ in the plot. The visual features were sampled at the end of action t_e. For comparison, we plotted the visual features with labels of the visual effect class $\{e_i^v\}_{i=1,\cdots,5}$ in Fig. 23(a) and then with labels of the action class $\{a_i\}_{i=1,\cdots,5}$ in Fig. 23(b). The visual effect classes were estimated using Eq.47. The number of visual classes was empirically set as five in the experiments.

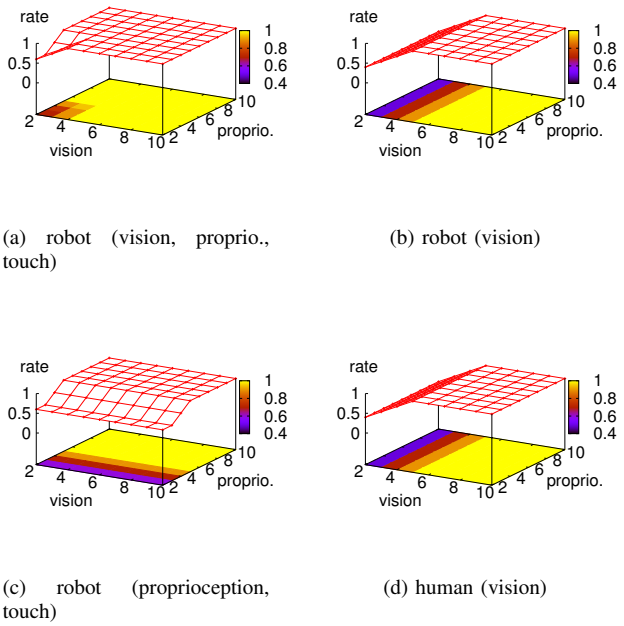
As we can see in these figures, visual features were classified similarly to action classes. Note that the clustering results of visual features are not necessarily similar to the action classes, since actions and visual effects are not always in one-to-one correspondence.

Figure 23(c) and (d) plot a set of the visual features $\boldsymbol{f}^v(t_e)$ of actions performed by an experimenter with labels of visual effect classes and action classes, respectively. Figure 23(d) also presents the trajectories of the intermediate visual features $\boldsymbol{f}_v(t)$ (they are referred to in classification consistency as detailed later). The visual effect of the experimenter's action was classified by the prototypes trained with self-generated actions. A comparison of Fig. 23(a) with (c) suggests that the visual features of the experimenter's actions were distributed similarly to those of the robot's actions. Therefore, the visual effects of the actions performed by the robot and the experimenter were similarly classified.

Figure 23(e) shows the consistency of the visual effect classification of the experimenter's actions at intermediate states (referred to in Fig. 23(d)). The horizontal and vertical axes indicate time and classification consistency, respectively. Here, classification consistency c_c represents the number of trials with an identical classification result at the present t and the end of the action t_e. The values in Fig. 23(e) were normalized with the number of trials (the maximum value is 1.0), and the time interval was normalized with five slices. Most of the original time intervals of the actions were around 3s. The error bar in the plot indicates the deviation of the values with respect to the action class. Figure 23(e) shows that predictive classification of the visual effect from observation of 1/2 of the action sequence had a consistency rate of 86%. In comparison, predictive classification upon observing 3/4 of the action sequence achieved 96% with an acceptable deviation.

2) Action perception: Figure 24 shows the results of action perception. In all of the graphs, the horizontal axis from left to right indicates the number of visual classes (2 to 10), and the horizontal axis from near to far indicates the number of proprioceptive effect classes (2 to 10). The number of classes corresponds to the resolution of the sensory effect in clustering. The vertical axis indicates the action recognition rate. The number of tactile classes was fixed at 3. We selected the best clustering results from 20 learning trials for each coupling of the visual and proprioceptive effect class numbers. The action recognition rate is the number of correctly recognized actions divided by the number of all trials. Note that all grid points in the graphs correspond to real values given by the experimental results (i.e. no interpolation technique was used for visualizing the grid surface).

Figures 24(a), (b) and (c) show the recognition results of the actions performed by the robot. The recognized action classes were given by Eq.48. Figure 24(a) shows the recognition results when the system used all sensory modalities (vision, proprioception and touch). Figure 24(b) shows the results when the system only used vision. Figure 24(c) shows the results when the system used proprioception and touch (i.e. the action was recognized in a blind condition). In these contexts, the action perception system was aware of the action classes because they were given by the action generator (refer to



(a) robot (vision, proprio., touch)

(b) robot (vision)

(c) robot (proprioception, touch)

(d) human (vision)

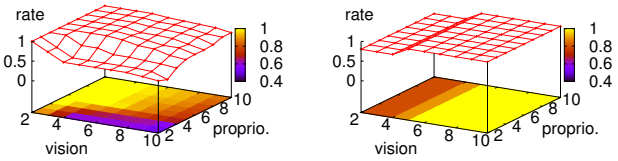
Fig. 24. Action recognition. The horizontal axes from left to right and from near to far indicate the number of classes (resolution of description) of visual and proprioceptive sensory effect, respectively. The vertical axis indicates the recognition rate. (a) presents the action recognition rate when the action agent was the robot and the sensory modalities used in recognition were vision, proprioception and touch. (b), (c) and (d) are labeled in the same manner.

efference copy presented in Fig. 21). Efference copies were used as ground-truth action classes to evaluate the estimations.

Figure 24(a) suggests that if the class number of either modality of vision or proprioception was five or more, action recognition rates were maximal. This means that a synergy of multi-modal sensing recovers low resolution of a member modality in action recognition. As shown in Fig. 24(b) and (c), when some sensory modalities are unavailable, the available modalities (vision in (b), and proprioception and touch in (c)) require high resolution to achieve a high action recognition rate. Figure 24(d) shows the recognition results of actions performed by the experimenter. The experimenter's actions were recognized well if the resolution of visual effect was high enough. This result was similar to the recognition of self-generated actions with vision-only in Fig. 24(b).

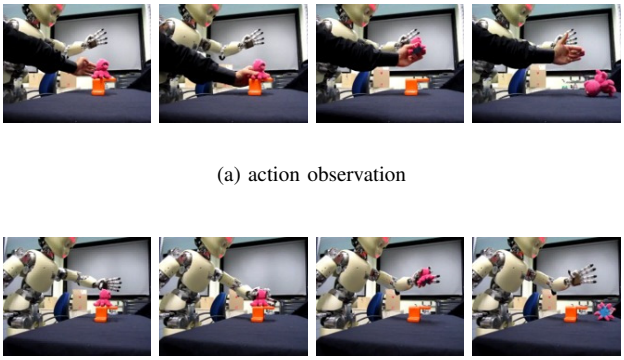
3) Cross-modal sensory anticipation: Figure 25 shows the results of cross-modal sensory anticipation. Estimations from Eq.49 and actual perception are compared.

The horizontal axes in both graphs are the same as those in Fig. 24. The vertical axis indicates the sensory match rate defined as the number of correctly estimated sensory effect classes divided by the number of all trials. Figure 25(a) shows visual sensory anticipation from the proprioceptive and tactile effect (visual anticipation) in (AC3). In this context, the actions were generated by the robot in a blind condition. Figure 25(b) shows tactile sensory anticipation from the visual effect (tactile anticipation) in (AC2). In this context, the actions were generated by the experimenter. To evaluate the sensory match rate, we used the corresponding visual and



(a) visual anticipation (robot) (b) tactile anticipation (human)

Fig. 25. Cross-modal sensory anticipation. The horizontal axes indicate the number of classes (resolution of description) of visual and proprioceptive sensory effect, respectively. The vertical axis indicates the sensory match rate. The graph labels present the estimated sensory modality and the action agent.



(a) action observation

(b) action reproduction

Fig. 26. Action reproduction by observation. The grasp, hold and drop actions are sequentially presented by an experimenter. The robot observed the actions and reproduced them in the order of recognition. The time course of the scenes is from left to right.

tactile effect classes observed in (AC1) as the ground-truth classes.

Figure 25(a) suggests that visual anticipation scored highly when the resolution of the visual effect was low and that of the proprioceptive effect was high. This means that (1) if visual resolution is low, visual anticipation matches easily; and (2) if the proprioceptive resolution is high, the input information is not lost and this then aids in reliable estimation. This experiment corresponds to the recovery of the visual sensory modality while executing an action in darkness. The results are also related to the behaviors of monkeys' mirror neurons in darkness [3].

Figure 25(b) suggests that tactile anticipation scored highly when resolution of the visual effect is high. Tactile anticipation is not affected by proprioceptive resolution, since only the visual sensory modality describes the experimenter's actions and no useful information comes from proprioception while observing them. Tactile anticipation is an interesting property of the proposed action perception; as we can see in the results, developments in action perception enabled the robot to generate internal sensory information of the experimenter (his touch sense) based on observation of human actions and the robot's sensory experience in its own action executions. We believe that action learning by robots set in human environments may increase the robots' sympathetic perception of humans.

4) Action reproduction by observation: We let the robot reproduce sequential actions from observation. Figure 26 presents scenes of action observation and action reproduction. An experimenter presented sequential actions to a robot. The action perception system buffered the recognition results and sent them to the action generator (see Fig. 5). The action generator then reproduced the actions in the buffered order. Figure 26 shows a demonstration composed of the grasp, hold and drop actions in the recognized order. At the end of each action, the experimenter paused between movements. This pause was used to segment the actions in the action perception system. As shown in the figure, the robot reproduced these actions in the same order as the experimenter's demonstration.

VI. Discussion

In this section, we compare the proposed method to related works in robotics and discuss the relation to infant development and biological mirror systems. We then present the limits of the proposed framework and possible solutions.

A. Comparison with robotic systems

In robotics, developmental sensory-motor coordination involving neuroscientific aspects and developmental psychology is well studied; e.g. sensorimotor prediction [19][20], mirror system [11][21], action-perception link [22], and imitation learning [23][24] are representative studies. Below, we will review literature that addresses body presentation and object affordance. These two aspects constitute the cornerstone of the research into body image and action perception implemented in this work.

Body presentation plays an important role for a robot dealing with voluntary actions [25]. Hikita et al. proposed a visuo-proprioceptive representation of the end effector based on Hebbian learning [26]. Stoytchev proposed a visually-guided developmental reaching [27] which demonstrated tasks similar to those examined in [2]. Kemp et al. approached robot hand discovery utilizing mutual information between arm joint angles and the visual location of an object [28]. Saegusa et al. proposed an own body definition system based on visuomotor correlation, and the system created a body representation regardless of body appearances or kinematic structures [10].

Object affordance (or possible actions to operate an object) plays an important role in manipulation [29]. In literature on robotic object manipulation, Natale et al. proposed a developmental grasping system that allows self hand recognition [30]. Montesano et al. proposed a learning model of object affordance using Bayesian networks [31]. In this work, the probabilistic links between action, effect, and object allow plausible action imitation [21]. Oztop et al. proposed a biologically comparable model of mirror systems [16][32]. Castellini et al. studied an effect of object affordance in object recognition [33] in which the authors experimentally showed that object recognition with visuomotor features gives higher

scores than a case with visual features.

In contrast to the previous studies, the framework proposed here is original in its developmental construction of the whole perception system (e.g. body identification, learning of motor control and learning of action perception) driven by self-generated actions. We hypothesize that only the results of actions can lead to reliable identity of the dynamically changing body and the meaning of actions in unknown or non-stationary environments. In previous work of [10], we proposed a body definition system based on visuomotor correlation that creates the body image of a single motor unit in monocular vision. The new system allows for creation of a more general body image with distinction of multiple motor units in binocular vision.

Moreover, the proposed system develops an incremental motor repertoire and action perception that is able to recognize human actions predictively. A simple action for humans such as picking up an object is rather complicated for robots. In the literature, Yokota et al. achieved action encoding and decoding with recursive network models [34]. Paine et al. proposed a model to decompose an action into motor primitives autonomously [35]. In our developmental scenario, we let a robot learn primitive actions (fixate, reach and grasp) and then construct more complex manipulative actions by combining them. This approach allows the motor repertoire to be built incrementally.

Compared to the predictive recognition system in [16], we implemented the system on an actual robot and demonstrated action perception in the real world. The Bayesian approach for action perception in [21][31] is related to the proposed work. We generalized the main idea of these studies to encompass cross-modal sensory association which yields sensory anticipation or compensation of unavailable sensory modalities when observing and executing actions. For example, the robot anticipates tactile sensory input when observing a human action, whereas the robot anticipates visual sensory input when executing an action blind. These are new functions compared to related methods. Compared to the latest studies in affordance learning [36] [37], the proposed method focuses on incremental ability in the development of perception from low level sensory-motor signals.

B. Comparison with biological systems

The findings of the study in [8] overlap the proposed procedure of learning from primitive to specific in this work. In the initial phase, the proposed system develops perception ability of the self's body from low-level visuomotor signals and proceeds to learn primitive actions (e.g. fixation, reaching and grasping) in the next phase. In the final phase, the robot develops the recognition of more specific, complex behaviors (e.g. grasp, hold and drop an object) based on the developed body image and primitive actions.

Some functions of the proposed action perception system are consistent with mirror systems in nature [3][4][7]. In particular, the proposed system supports the three action contexts, AC1, AC2 and AC3, for learning action perception, action execution and reproduction of recognized actions. These action contexts are equivalent to the experimental conditions with monkeys in [3]. In modeling action perception as well, the box of motor repertoire and connected signal flows in Fig. 21 correspond to the instance of mirror neurons in monkeys.

C. Limits of the proposed system

The described method proposes different phases in autonomous development of perception. However, we did not investigate how the transition between these phases could happen in a continuous developmental path. In the experiments reported in this paper, the human experimenter manually switched each learning phase (the learning phase of primitive actions and action perception). How to make this autonomous is an important problem to be investigated in the future. In addition, complex actions like a sequence of grasp, hold and drop were defined beforehand by selecting and combining together the learned primitives. Such actions could however be learned autonomously by the robot either in exploration [38] or observation [39].

Additionally, a general and consistent learning algorithm applicable for all modules should be introduced into the proposed framework. At the moment, the learning modules use a memory system that indexes data using a nearest neighbor approach and interpolates the output locally. This approach can scale well to allow for long-term learning in which a large amount of data has to be processed, but it also has limitations due to the lack of topological maps representing the state space. Learning such topological maps was not investigated in this work, since, as mentioned in Section I, the main focus of this work is not on the development of motor control but rather on the development of sensory perception.

VII. Conclusion

We proposed a robot's developmental perception driven by active motor exploration. In the proposed framework, the robot discovers its own body (body image) through self-generated actions, the relationship between sensory states and motor commands (motor control), and the effects of actions on objects (action perception). In the development of perception, multi-modal sensing played an important role, since multi-modality allows cross-modal sensory anticipation.

We evaluated the proposed framework in repetitive experiments with an anthropomorphic robot. The robot developmentally achieved the following perceptual abilities: body image of multiple motor units, primitive motor skills of fixation, reaching and grasping, predictive human action recognition, and cross-modal sensory effect anticipation. Overall, the robot succeeded in recognizing actions performed by a human experimenter and in mapping the corresponding sensory feedback on its own internal sensory system.

Development ability is the most important aspect for robots or mobile intelligence targeted for work in non-stationary environments. A typical problem of non-stationary settings for robots is self-perception. As shown in the experiments, the self-perception system was able to adapt to drastic changes in body appearance as a result of object grasping. This perceptual

ability also helped the robot perceive actions performed by humans.

An ability lacking in the proposed system is the use of a tool. Tool use was beyond the scope of the current work, since we intended to focus on the robot's perceptual developments rather than those of motor control. However, in this work we demonstrated that the proposed system can adapt to changes in the kinematics and hand visual appearance resulting when the robot grasps a tool. Such a perceptual component is of critical importance for learning tool use.

ACKNOWLEDGMENTS

This work is partially supported by EU FP7 project CHRIS (Cooperative Human Robot Interaction Systems FP7 215805) and EU FP7 project Xperience (Robots Bootstrapped through Learning and Experience, FP7 97459).

REFERENCES

[1] A. Iriki, M. Tanaka, and Y. Iwamura, "Coding of modified body schema during tool use by macaque postcentral neurones," *Neuroreport*, vol. 7(14), pp. 2325–30., 1996.

[2] A. Iriki, M. Tanaka, S. Obayashi, and Y. Iwamura, "Self-images in the video monitor coded by monkey intraparietal neurons," *Neuroscience Research*, vol. 40, pp. 163–173, 2001.

[3] G. Rizzolatti, L. Fadiga, V. Gallese, and L. Fogassi, "Premotor cortex and the recognition of motor actions," *Cognitive brain research*, vol. 3, no. 2, pp. 131–141, 1996.

[4] V. Gallese, L. Fadiga, L. Fogassi, and G. Rizzolatti, "Action recognition in the premotor cortex." *Brain*, vol. 119, pp. 593–609, 1996.

[5] L. Fogassi, P. Ferrari, B. Gesierich, S. Rozzi, F. Chersi, and G. Rizzolatti, "Parietal lobe: from action organization to intention understanding," *Science*, vol. 308, no. 5722, p. 662, 2005.

[6] A. Maravita and A. Iriki, "Tools for the body (schema)," *Trends in Cognitive Sciences*, vol. 8(2), pp. 79–96, 2004.

[7] T. Kaneko and M. Tomonaga, "The perception of self-agency in chimpanzees(pan troglodytes)." *Proceedings of the Royal Society B: Biological Sciences*, 2011.

[8] H. Watanabe and G. Taga, "General to specific development of movement patterns and memory for contingency between actions and events in young infants," *Infant Behavior and Development*, vol. 29, pp. 402–422, 2006.

[9] G. Viswanathan, E. Raposo, and M. Da Luz, "Lévy flights and superdiffusion in the context of biological encounters and random searches," *Physics of Life Reviews*, vol. 5, no. 3, pp. 133–150, 2008.

[10] R. Saegusa, G. Metta, and G. Sandini, "Body definition based on visuomotor correlation," *IEEE Transaction on Industrial Electronics*, vol. 59, no. 8, pp. 3199–3210, 2012.

[11] P. Fitzpatrick and G. Metta, "Grounding vision through experimental manipulation," *Philosophical Transactions of the Royal Society: Mathematical, Physical, and Engineering Sciences*, vol. 361, no. 1811, pp. 2165–2185, 2003.

[12] N. Tsagarakis, G. Metta, G. Sandini, D. Vernon, R. Beira, F. Becchi, L. Righetti, J. Santos-Victor, A. Ijspeert, M. Carrozza, and D. G. Caldwell, "icub: the design and realization of an open humanoid platform for cognitive and neuroscience research," *Advanced Robotics*, vol. 21, no. 10, pp. 1151–?1175, 2007.

[13] G. Metta, L. Natale, F. Nori, G. Sandini, D. Vernon, L. Fadiga, C. Von Hofsten, K. Rosander, M. Lopes, J. Santos-Victor *et al.*, "The icub humanoid robot: An open-systems platform for research in cognitive development," *Neural Networks*, vol. 23, no. 8, pp. 1125–1134, 2010.

[14] R. Saegusa, G. Metta, and G. Sandini, "Active learning for multiple sensorimotor coordinations based on state confidence," in *the 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS2009)*, St. Louis, MO, USA, October 11-15 2009, pp. 2598–2603.

[15] L. Itti, C. Koch, and E. Niebur, "A model of saliencybased visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998.

[16] E. Oztop and M. Arbib, "Schema design and implementation of the grasp-related mirror neuron system," *Biological cybernetics*, vol. 87, no. 2, pp. 116–140, 2002.

[17] J. MacQueen, "Some methods for classification and analysis of multivariate observations," *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 281–297, 1967.

[18] D. Wolpert and J. Flanagan, "Motor prediction," *Current Biology*, vol. 11, no. 18, pp. R729–732, 2001.

[19] D. Wolpert, Z. Ghahramani, and M. Jordan, "An internal model for sensorimotor integration," *Science*, vol. 269, no. 5232, pp. 1880–1882, 1995.

[20] M. Kawato, "Internal models for motor control and trajectory planning," *Current Opinion in Neurobiology*, no. 9, pp. 718–727, 1999.

[21] G. Metta, G. Sandini, L. Natale, L. Craighero, and L. Fadiga, "Understanding mirror neurons: a bio-robotic approach," *Interaction Studies*, vol. 7, no. 2, pp. 197–232, 2006.

[22] P. Fitzpatrick, A. Needham, L. Natale, and G. Metta, "Shared challenges in object perception for robots and infants," *Infant and Child Development*, vol. 17, no. 1, pp. 7–24, 2008.

[23] S. Schaal, "Is imitation learning the route to humanoid robots?" *Trends in Cognitive Sciences*, vol. 3, pp. 233–242, 1999.

[24] S. Calinon, F. Guenter, and B. Aude, "On learning, representing and generalizing a task in a humanoid robot," *IEEE Transactions on system, man, and cybernetics, Part B*, vol. 37, no. 2, pp. 286–298, 2007.

[25] M. Hoffmann, H. Marques, A. Arieta, H. Sumioka, M. Lungarella, and R. Pfeifer, "Body schema in robotics: A review," *Autonomous Mental Development, IEEE Transactions on*, vol. 2, no. 4, pp. 304–324, 2010.

[26] M. Hikita, S. Fuke, M. Ogino, and M. Asada, "Cross-modal body representation based on visual attention by saliency," in *IEEE/RSJ International Conference on Intelligent Robotics and Systems (IROS)*, 2008.

[27] A. Stoytchev, "Toward video-guided robot behaviors," in *Proceedings of the Seventh International Conference on Epigenetic Robotics (EpiRob)*, L. Berthouze, C. G. Prince, M. Littman, H. Kozima, , and C. Balkenius, Eds., vol. Modeling 135, 2007, pp. 165–172.

[28] C. C. Kemp and E. Aaron, "What can i control?: The development of visual categories for a robot's body and the world that it influences," in *Proceedings of the Fifth International Conference on Development and Learning, Special Session on Autonomous Mental Development*, 2006.

[29] J. Gibson, *The ecological approach to visual perception*. Lawrence Erlbaum Associates, 1986.

[30] L. Natale, "Linking action to perception in a humanoid robot: A developmental approach to grasping." Ph.D. dissertation, LIRA-Lab, DIST, University of Genoa, 2004.

[31] L. Montesano, M. Lopes, A. Bernardino, and J. Santos-Victor, "Learning object affordances: From sensory–motor coordination to imitation," *Robotics, IEEE Transactions on*, vol. 24, no. 1, pp. 15–26, 2008.

[32] E. Oztop, M. Kawato, and M. Arbib, "Mirror neurons and imitation: A computationally guided review," *Neural Networks*, vol. 19, no. 3, pp. 254–271, 2006.

[33] C. Castellini, T. Tommasi, N. Noceti, F. Odone, and B. Caputo, "Using object affordances to improve object recognition," *Autonomous Mental Development, IEEE Transactions on*, vol. 3, no. 3, pp. 207–215, 2011.

[34] R. Yokoya, T. Ogata, J. Tani, K. Komatani, and H. Okuno, "Experience-based imitation using rnnpb," *Advanced Robotics*, vol. 21, no. 12, pp. 1351–1367, 2007.

[35] R. Paine and J. Tani, "Motor primitive and sequence self-organization in a hierarchical recurrent neural network," *Neural Networks*, vol. 17, no. 8-9, pp. 1291–1309, 2004.

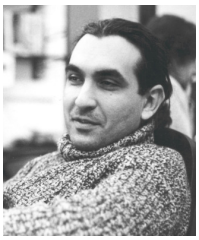
[36] S. Griffith, J. Sinapov, V. Sukhoy, and A. Stoytchev, "A behavior–grounded approach to forming object categories: Separating containers from non-containers," *IEEE Transactions on Autonomous Mental Development*, vol. 4, no. 1, pp. 54–69, 2012.

[37] E. Ugur, E. Oztop, and E. Sahin, "Goal emulation and planning in perceptual space using learned affordances," *Robotics and Autonomous Systems*, vol. 59, no. 7?-8, pp. 580–595, 2011.

[38] P. Kormushev, S. Calinon, R. Saegusa, and G. Metta, "Learning the skill of archery by a humanoid robot icub," in *2010 IEEE-RAS International Conference on Humanoid Robots (Humanoids2010)*, Nashville, TN, USA, December 6-8 2010, pp. 417–423.

[39] S. Lallée, U. Pattacini, J. Boucher, S. Lemaignan, A. Lenz, C. Melhuish, L. Natale, S. Skachek, K. Hamann, J. Steinwender *et al.*, "Towards a platform-independent cooperative human-robot interaction system: Ii. perception, execution and imitation of goal directed actions," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2011.

Ryo Saegusa has been a project associate proffessor with the Center for Human-Robot Symbiosis Research, Toyohashi University of Technology since 2012. He attained B.Eng., M.Eng. and D.Eng. degrees in applied physics from Waseda University, Tokyo, Japan, in 1999, 2001 and 2005. From 2004 to 2007, he was a research associate at the Department of Applied Physics at Waseda University. He was a postdoctoral researcher from 2007 to 2009 and a senior postdoctoral researcher from 2009 to 2012 with Robotics, Brain and Cognitive Sciences Department at the Istituto Italiano di Tecnologia, Genoa, Italy. His research interests include machine learning, computer vision, signal processing, cognitive robotics and health care robotics.



Giorgio Metta is a director of the iCub Facility at the Istituto Italiano di Tecnologia where he coordinates the development of the iCub robotic platform/project. He holds a MSc cum laude (1994) and PhD (2000) in electronic engineering both from the University of Genoa. From 2001 to 2002 he was a postdoctoral associate at the MIT AI-Lab. He was previously with the University of Genoa and since 2012 a professor of Cognitive Robotics at the University of Plymouth (UK). He is a deputy director of IIT delegate to the international relations and external funding. In this role he is a member of the board of directors of euRobotics aisbl, the European reference organization for robotics research. His research activities are in the fields of biologically motivated and humanoid robotics and, in particular, in developing humanoid robots that can adapt and learn from experience. He is an author of approximately 200 scientific publications. He has been working as a principal investigator and research scientist in about a dozen international as well as national funded projects.

Giulio Sandini is a director of the Robotics, Brain and Cognitive Sciences Department at the Istituto Italiano di Tecnologia and full professor of bioengineering at the University of Genoa. His main research interests are in the fields of computational and cognitive neuroscience and robotics with the objective of understanding the neural mechanisms of human sensorimotor coordination and cognitive development from a biological and artificial perspective. He graduated in electronic engineering (bioengineering) at the University of Genoa. He has been an assistant professor at the Scuola Normale Superiore in Pisa and Visiting Scientist in the department of neurology at Harvard Medical School and the Artificial Intelligence lab at MIT. Since 2006 he has been a director of research at the Istituto Italiano di Tecnologia where he heads the Robotics, Brain and Cognitive Sciences Department.

Lorenzo Natale attained an MSc degree in electronic engineering and a PhD in robotics from the University of Genoa, Genoa, Italy, in 2000 and 2004, respectively. Over the past ten years, he has worked with several humanoid platforms. He worked in the Laboratory for Integrated Advanced Robotics (LIRA-Lab) at the University of Genoa, and was then postdoctoral researcher at the MIT Computer Science and Artificial Intelligence Laboratory. At the moment he is a team leader at the iCub Facility at the Istituto Italiano di Tecnologia, Genoa, Italy. His research interests include the field of humanoid robotics and range from sensorimotor learning and perception to software architectures for robotics.