Action Learning based on Developmental Body Perception

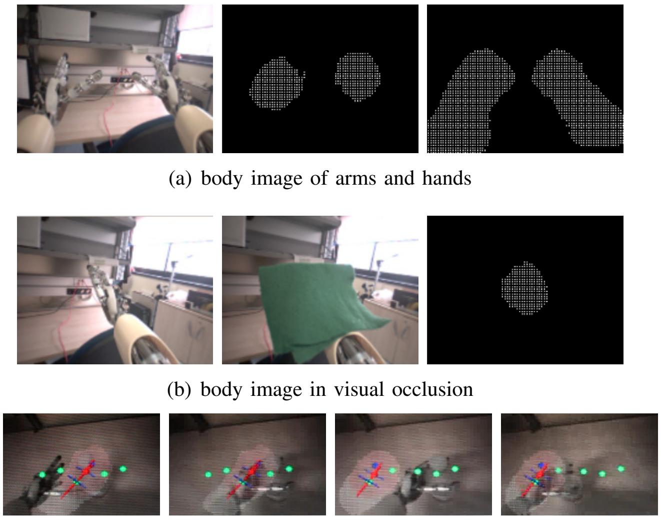
Ryo Saegusa, Giorgio Metta, Giulio Sandini and Lorenzo Natale

Abstract— **The paper describes a framework for action learning in anthropomorphic robots. The key idea of the framework is that the robot voluntarily generates movements to identify its own body, and refers to the identified body in learning of fixation, reaching and grasping actions for object operation. Developmental body perception is critical to identify changing body parts in unknown or non-stationary environments. The consolidation of action learning with developmental body perception allows for continuous body awareness and anticipation of self-generated action's results. We evaluated the proposed framework in experiments with a real robot. In experiments, the robot achieved autonomous body identification, learning of fixation, reaching and gasping, and anticipation-based action planning as well as its execution in object operation tasks.**

I. Introduction

How can a robot identify the body and associate it with its own actions? The body perception is fundamental for action learning in animals, whereas in robotics organic links between body perception and action learning have not been implemented yet. In past decades, physiological studies discovered interesting cognitive functions in nature; macaque monkeys are able to recognize their own bodies even when their bodies are experimentally modified or extended [1][2]. Also, they understand actions so as to mirror the actions 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 consolidate action learning with developmental body perception. Our claim for current robotic action learning systems is that the robots learn actions without developmental modeling of their own bodies. Therefore, self body perception in robots is not yet reconfigurable, and their actions for objects are not anticipated well in advance of execution. In this work, we will introduce a method for developmental body identification which allows for creation of perceptual images of multiple body parts in binocular vision. We will then apply the body identification for learning of manipulative actions such like fixation, reaching and grasping actions. Finally, we will present action anticipation as an advantage of the consolidation of action learning with developmental body perception (refer to Fig.1).


(a) body image of arms and hands

(b) body image in visual occlusion

(c) action anticipation

Fig. 1. Body image and action anticipation. (a) The arm and hand images. The reference, hand domains and forearm domains are presented from left to right. (b) The body image in occlusion. The reference before occlusion, reference after occlusion, and hand image while occluded are presented from left to right. (c) Anticipation of arm and hand locations in object operation. The hand and forearm visual appearances are represented with a pink and white cloud. The red dot with the red and blue segments represent the anticipated visual location, and the major and minor axes of the arm, respectively. The greed dots represent learned visual locations. The time course of the pictures is from left to right. In the first and second pictures, the robot is reaching for a bottle and grasps it. The first picture shows the expected location and shape of the arm/hand at the end of the movement. The second picture shows the arm/hand postures after the reachig and grasping. Similarly in the third and fourth pictures, but this time the arm is going back to the initial position.

In robotics, developmental sensory-motor coordination involving neuroscientific aspects and developmental psychology is well studied; e.g., sensorimotor prediction [6][7], mirror system [8][9], action-perception link [10], imitation learning [11] and gesture recognition [12] are representative studies. Body presentation plays an important role for a robot dealing with voluntary actions [13]. Hikita et al. proposed a visuo-proprioceptive representation of the end effector based on Hebbian learning [14]. Stoytchev proposed a visually-guided developmental reaching [15] which demonstrated tasks similar to those examined in [2]. Kemp et al. approached the robot hand discovery utilizing mutual information between arm joint angles and the visual location of an object [16].

Previously we proposed an own body definition system based on visuomotor correlation [17], while the system was limited in monocular hand definition and action learning was not considered in the framework. In the following sections, we present an autonomous action learning system consolidated with an extended body definition system for multiple body part perception in binocular vision.

R.Saegusa is with the Center for Human-Robot Symbiosis Research, Toyohashi University of Technology, 1-1 Hibarigaoka, Tempakucho, Toyohashi 441-8580 Japan. G.Metta, G.Sandini and L.Natale are with the Department of Robotics, Brain and Cognitive Sciences, Italian Institute of Technology, Via Morego 30, Genova, 16163 Italy. E-mail: `ryos@ieee.org`

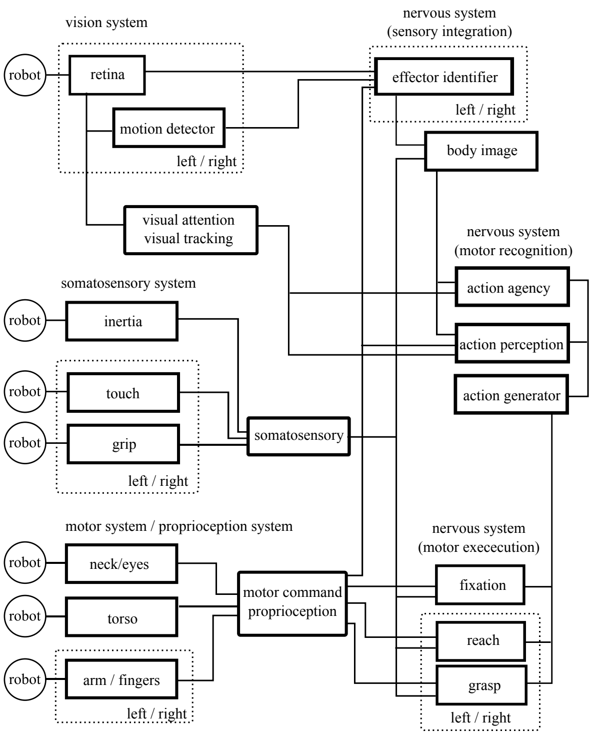
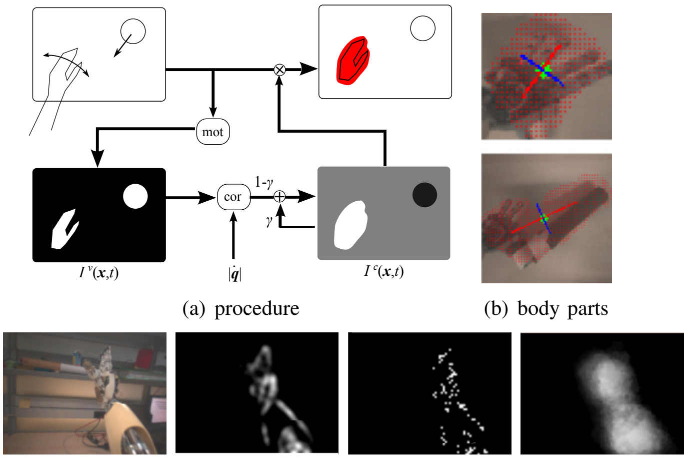


Fig. 2. A diagram of sensory-motor signal flows. The computations of sensory-motor modules are distributed in the networks.



(a) procedure

(b) body parts

(c) Motion detection (reference, difference, sampled points, filled blobs)

Fig. 3. Extraction of a motor-correlated visual blob. (a) Extraction procedure. (b) Identified body parts (top: inherent body; bottom: extended body). (c) Intermediate images in motion detection.

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,F}, {Gf, Ga, Gb}

II. DEVELOPMENTAL BODY IDENTIFICATION

We introduce a developmental body identification system. The body identification system learns a representation that links together the visual features of a visual blob with the corresponding proprioceptive information (head, eyes and arm joint encoder values). This enables the robot to estimate the position and appearance of body parts from the proprioceptive information. The overview of the system structure is illustrated in Fig. 2.

A. *Body identification*

The system actuates a motor unit (motor exploration), and associates visual features of a motor-correlated blob in a frame with the proprioceptive features of the actuated motor unit. The motor unit is a local group of motor joints (e.g., a set of the pitch, yaw and roll joints), We assumed four motor units, left wrist, right wrist, left shoulder and right shoulder, in following experiments.

We implemented two types of random movements for the motor units; the first is a ballistic movement to transport the arm to a certain posture; the second is a perturbation movement to vary the arm and hand position for visual blob segmentation in the frame. A motor unit is randomly selected and given a motor command. An actuation of the motor unit is defined as follows:

$$\boldsymbol{u} = \boldsymbol{q} + \delta\boldsymbol{q}, \qquad (1)$$

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. A motor-correlated visual blob is extracted by this motor exploration. We skip explanations of this procedure (refer to [18]), and only present its schematic illustration in Fig. 3.

B. *Experiments*

We performed experiments with a real robot to evaluate the proposed body identification system. We used the iCub robot platform [19] for the experiments. The joint link structure of the robot is presented in [18]. Table I summarizes experimental conditions. We define the shoulder and wrist movement:

$$\boldsymbol{u}_s = \boldsymbol{q}_s + \delta\boldsymbol{q}_s, \qquad (2)$$

$$\boldsymbol{u}_w = \boldsymbol{q}_w + \delta\boldsymbol{q}_w, \qquad (3)$$

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 [18]. In the experiments, we performed repetitive back-and-forth movements ($\delta\boldsymbol{q}$ and $-\delta\boldsymbol{q}$) for body identification. In the experiments, we investigated the identification of 1) the inherent body and 2) the modified body.

1) Inherent body identification: The robot performed the shoulder and wrist motor explorations with four representative postures. Figure 4 shows the mean m and standard deviation σ of the visual features of the identified body (detailed later). Figure 5 shows the appearances of the identified body parts. In the figures, S and W denote the shoulder and wrist that the robot moves. V, H, N, F denote

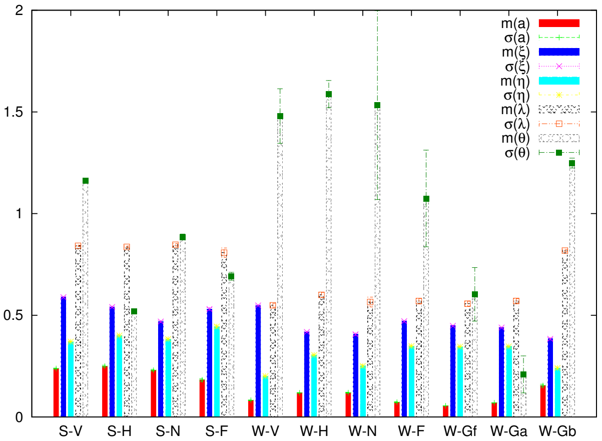


Fig. 4. Visual features of body parts. Visual features of inherent body parts (S-V,S-H,S-N,S-F,W-V,W-H,W-N,W-F), visual features of extended body parts (W-Gf,W-Ga,W-Gb).

the condition of the arm; vertical, horizontal, near and far, respectively. We performed 20 trials for each postures.

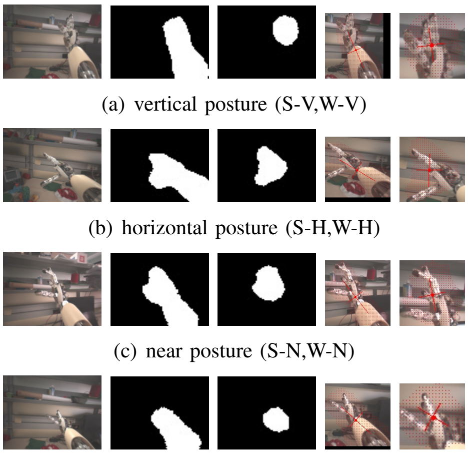
The variables a, x, λ, θ in Fig.5 present the area, location, distortion and orientation of the body part. a is normalized so as to let the frame area 1.0. x is normalized so as to let the length of the diagonal segment of the frame 1.0. λ is given as follows: $\lambda = \lambda_1/(\lambda_1 + \lambda_2)$ where λ_1, λ_2 are the eigenvalue of the major and the minor axes of the body part. θ is the orientation of the body part; $\theta = \arctan(e_2/e_1)$ where $[e_1, e_2]^T$ denotes the major axis.

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),
- the orientation was comparably reliable for the shoulder part, but not for the wrist part because the major and minor 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 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 the wrist motor exploration in the case that an object is in the hand. Figure 4 shows the visual features. Figure 6 shows the appearances of the extended body part. In the figure, 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.



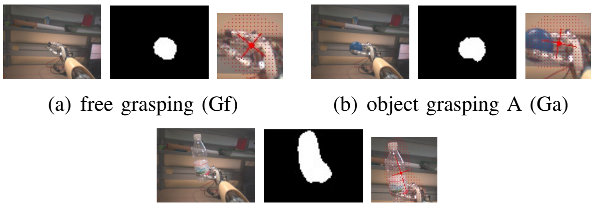
(a) vertical posture (S-V,W-V)

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

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

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

Fig. 5. 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.



(a) free grasping (Gf)

(b) object grasping A (Ga)

(c) object grasping B (Gb)

Fig. 6. Extended body identification; the reference frame, body part (wrist), and body textire (wrist) are presented from left to right.

In this demonstrations, the robot succeeded to identify extended parts as its own body. The visual features of extended body parts are combined with proprioceptive information.

III. Learning of Motor Skills

Body identification allows the robot in the next phase to learn motor skills for object operation. In this section, we define learning of fixation, reaching and grasping actions, which will be used as the building blocks of more complex manipulative actions afterwards. We assume the following motor units and corresponding actions; neck and eyes (fixation), left and right arm (reaching), left hand and right hand (grasping) (refer to the corresponding modules in Fig. 2). In the following sections, we will describe the learning procedure for the corresponding action in each motor unit.

A. Head motor unit

Learning in the head motor unit allows for visual projection (estimation from vision to head proprioception) and visual fixation (estimation from head proprioception to vision).

1) Head motor exploration: We formulate the egocentric three-dimensional visual location of a target z as follows:

$$z = (\xi^L, \eta^L, \xi^R - \xi^L), \tag{4}$$

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:

$$\delta \boldsymbol{z} = J^h(\boldsymbol{q}, \boldsymbol{z})\delta \boldsymbol{q}, \tag{5}$$

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:

$$\delta \boldsymbol{q}_i = (0, \cdots, dq_i, \cdots, 0), \tag{6}$$

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

$$\boldsymbol{J}_i^h(\boldsymbol{q}, \boldsymbol{z}) = (1/dq_i)\delta \boldsymbol{z}_i, \tag{7}$$

where $\boldsymbol{J}_i^h$ 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).

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;

$$\hat{\boldsymbol{z}}(\boldsymbol{q}_i, \boldsymbol{z}_i; \boldsymbol{q}) = \boldsymbol{z}_i + \hat{J}_k^h(\boldsymbol{q} - \boldsymbol{q}_i), \tag{8}$$

$$\hat{J}_k^h = J^h(\boldsymbol{q}_k), \tag{9}$$

$$k = \arg\min_j |\boldsymbol{q} - \boldsymbol{q}_j|, \tag{10}$$

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. $\hat{J}_k^h$ 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;

$$\hat{\boldsymbol{q}}(\boldsymbol{q}, \boldsymbol{z}; \boldsymbol{z}^d) = \boldsymbol{q} + \hat{J}_k^{h\#}(\boldsymbol{z}^d - \boldsymbol{z}), \tag{11}$$

$$\hat{J}_k^h = J^h(\boldsymbol{q}_k), \tag{12}$$

$$k = \arg\min_j |\boldsymbol{q} - \boldsymbol{q}_j|, \tag{13}$$

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. In visual fixation, we assign the coordinates of the center of the view frame for $\boldsymbol{z}^d$, though the goal location is not limited to that (i.e., in theory, the robot can bring the target in any location of the view frame). $\hat{J}_k^{h\#}$ represents the generalized inverse $\hat{J}_k^h$ at $\boldsymbol{q}_k$.

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

4) Experiments: We examined visual projection and fixation with the head motor unit. Table II summarizes the experimental 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 [18]. 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.

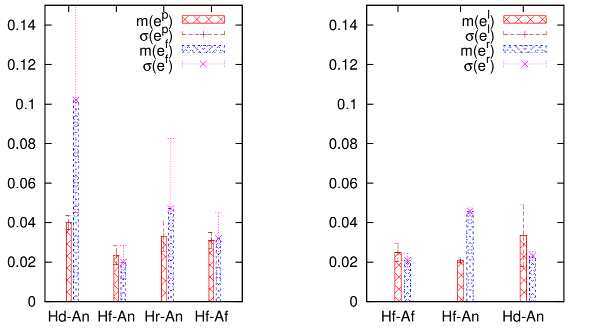
(a) Visual projection: In this experiment, we evaluate visual projection ability at each of four different joint postures. First, the robot performed head motor exploration (body identification and learning of transformation $J^h(\boldsymbol{q}, \boldsymbol{z})$) at a single joint posture $\boldsymbol{q}$, 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 lose generality because the location is estimated locally at the nearest learned joint posture (refer to Eq. 10), and the estimation is independent from other learned joint postures. The test tuples were sampled as follows:

$$\boldsymbol{u} = \boldsymbol{q} + \delta \boldsymbol{q}, \tag{14}$$

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:

$$e_i = \sum_{j=1,\cdots,n, i \neq j}^{n} |\boldsymbol{z}_i - \hat{\boldsymbol{z}}(\boldsymbol{q}_j, \boldsymbol{z}_j; \boldsymbol{q}_i)|/(n-1), \tag{15}$$



(a) visual projection and fixation (b) arm localization and reaching

Fig. 7. Estimation error. (a) Visual projection and fixation after head motor exploration. (b) Arm localization and arm reaching after arm motor exploration.

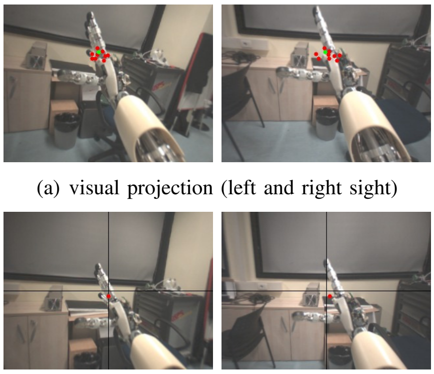
where $m(e) = 1/n\sum_{i=1}^{n} e_i$ and $\sigma(e) = 1/n\sum_{i=1}^{n} |e_i - m(e)|$ denote the average and deviation of the estimation error.

Figures 7(a) and 8(a) show the results of the visual projection. In Fig. 7(a), $m(e_p)$ and $\sigma(e_p)$ denote the average and deviation of the estimation in the visual projection. The label Hd, Hf and Hr denote the head joint posture corresponding to down, front and right. The label An and Af denote the arm joint posture posing as positioned near and far from the head, respectively. In the experiments, we evaluated 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:

$$e_i = |\boldsymbol{z}^d - \boldsymbol{z}_i'(\hat{\boldsymbol{q}}_i)|. \tag{16}$$



(a) visual projection (left and right sight)

(b) visual fixation (left and right sight)

Fig. 8. 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).

Figures 7(a) and 8(b) show the results of the visual fixation. In the figures, $m(e_f)$ 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

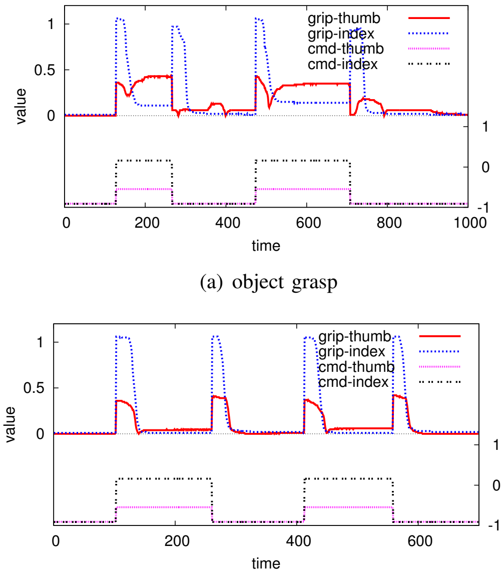
Learning in the arm motor unit allows for arm localization (estimation from vision to arm proprioception) and arm reaching (estimation from arm proprioception to vision). Unfortunately, we are forced to skip the presentation of the learning procedure of arm localization and arm reaching because of limited space in the paper. We will explain the procedure when we present this work, and here we just show the most important results in Fig. 7(b).

C. Finger motor unit

The robot uses the finger motor unit to perform motor exploration with an object, and associates the observed somatosensory event with the features of the action and object. The objects are learned with the visual attention system in advance. Unfortunately, we are forced to skip the presentation of the learning procedure of grasping and visual attention for hands and objects because of limited space in the paper. We will explain the procedure when we present this work, and here we just show the most important results in Fig. 9 and 10.

IV. Body image and action anticipation

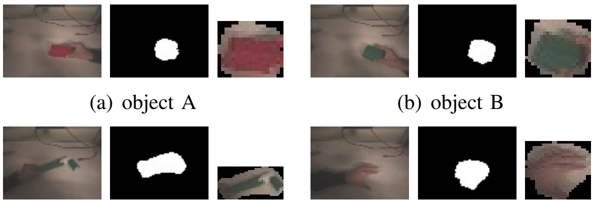
The learned body knowledge and motor skills were applied for proprioceptive body image and action anticipation in manipulation tasks. Figure 1(a) shows body image and action anticipation. Four body parts (left hand, left forearm, right hand and right forearm) are projected in Fig. 1(a). In general, it is not easy to visually identify the left and right hand in the same frame, since their appearances are similar. On the



(a) object grasp

(b) free grasp

Fig. 9. 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. 10. Motion-based visual attention. The reference frame, attracted domain and detected object are presented from left to right in each target object.

other hand, the proprioceptive identification if Fig. 1(a) is distinctive, and it works even for building an arm image when the arm is occluded as shown in Fig. 1(b).

Figure 1(c) shows the anticipation of arm and hand locations in object operation. When the robot identified an object of interest (the bottle, in this case), it anticipated the final arm posture and projected the expected final appearance of the arm and hand in the visual field. The robot then executed the task, and verified that its result matched the anticipation.

V. Conclusion

We proposed a method of action learning consolidated with developmental body perception for anthropomorphic robots. A robot autonomously identified multiple body parts based on visuomotor correlation, and the identified body parts were referred in voluntary learning of fixation, reaching and grasping actions. The synergistic development of the motor skills and perceptual body modeling provided the ability of dynamic body image and action anticipation in object operation.

The proposed method has an advantage in adaptation to body change, and this allows for body perception and action learning free from body modification. In the paper, we did not discuss autonomous classification of body extension or the effect of its extension on object operation. In future, we are focusing on these missing pieces and try to approach reasoning problems for autonomous body modification towards tool use.

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

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

[8] 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.

[9] 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.

[10] 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.

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

[12] A. Chaudhary, J. Raheja, and S. Raheja, "A vision based geometrical method to find fingers positions in real time hand gesture recognition," *Journal of Software*, vol. 7, no. 4, pp. 861–869, 2012.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] ——, "Own body perception based on visuomotor correlation," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS2010)*, Taipei, Taiwan, October 18-22 2010, pp. 1044–1051.

[19] 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.