

Active Perception for Action Mirroring

Ryo Saegusa, Lorenzo Natale, Giorgio Metta, Giulio Sandini

Abstract—The paper describes a constructive approach on active perception for anthropomorphic robots. The key idea is that a robot tries to identify a human’s action as an own action based on the observation of action effects for objects. In the proposed framework, the active perception is decomposed into the three phases; First, a robot voluntarily generates actions to discover the own body and objects. Second, the robot characterizes its own action with the effect for the objects. Third, the robot identifies the human action with the own action. The mirrored perception of the own action and the human’s action allows the robot to share the goal-directed behavior with humans. The proposed framework of active perception was experimentally validated with the integrated sensory modalities of vision, proprioception and touch.

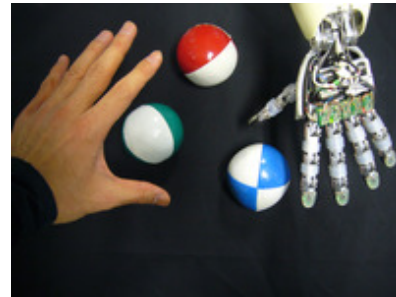
I. INTRODUCTION

HOW can a robot identify the self, and associate it with others? This is a fundamental question for the early life of primates and also embodied intelligence. Monkeys are able to recognize their own body, and extend their body schema while using a tool [1]. This kind of cognitive functions may have the potential to break a limit in existing hand-coded intelligence of autonomous robots and allows the understanding of intentions behind actions.

Our goal is to realize a cognitive system which actively develops its perception of the own body and action effect through interaction. The active perception which we are proposing here is inspired from action mirroring system found in the premotor cortex of the primates [2] [3] [4]. It is not clear why the primates acquired the ability to imitate or mirror actions, while we guess that the ability of the action mirroring was naturally selected in straggling to socialize action repertoires of individuals among the species.

In this paper, we introduce a framework to simulate primate-like active perception in the context of object manipulation. The concept is illustrated in Fig.1. The key idea of the action mirroring is to use objects as a medium to characterize the action. The action can be commonly characterized by an effect for the object regardless of the demonstrator’s body structure.

The process of the cognitive development is summarized as follows; A robot starts to build a basement of perception by exploratory actions. The auto-generated action allows the robot to define the own body and explore the environment to find manipulable objects. The robot, then, associates the own action with an effect for objects by visual and proprioceptive sensing. During interaction with humans, the robot interprets



(a)



(b)



(c)

Fig. 1. (a) Interaction among a robot, a person, and objects. Neither robots nor humans can conclude in advance of interaction whether the balls are manipulable (the balls might be fake pictures.) (b) The robot observes an experimenters grasping action. (c) The robot executes a similar grasping action. The executed actions are associated from the observed action by a mirroring system.

humans actions with an own action vocabulary and demonstrates the observed action. We expect that the paradigm of the action mirroring potentially gives a clue to understand intention in the action.

II. RELATED WORK

Rizzolatti et al. found the neurons (mirror neurons) in the premotor cortex of monkeys, which got activated both of when the monkey performed a certain action and observed a similar action demonstrated by a human experimenter [2] [3]. In their studies, each mirror neuron coded a single manipulative action or a combination of them. The type of actions which the mirror neuron responded to in the experiment was grasping, holding, placing, manipulating, and two hands interaction. Activation of grasping mirror neurons is shown in Fig.2. The same neurons were activated during observation (left) and execution (right) of the grasping action. Interestingly, the activation of the mirror neurons was neither selective for the type of objects which the monkey interacted with, and nor activated by an action without a target object.

Fogassi et al. found that the inferior parietal lobule (IPL) neurons coding a specific act (e.g., grasping) showed markedly different activation when this act was part of different actions (e.g., for eating or for placing). They concluded

This work is partially supported by EU FP7 project CHRIS (Cooperative Human Robot Interaction Systems FP7 215805).

R. Saegusa, L. Natale, G. Metta, G. Sandini are with the Department of Robotics, Brain and Cognitive Sciences, Italian Institute of Technology, Genoa, Italy (email: ryoas@ieee.org, ryo.saegusa@iit.it).

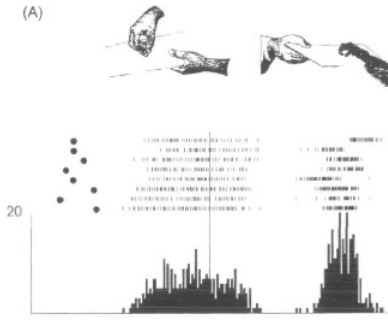


Fig. 2. Grasping mirror neurons in a premotor cortex of a monkey. The neurons were activated both when the monkey observed a grasping action (left) and the monkey executed a grasping action (right). The illustration was reproduced from [3] under the permission.

that the connection of the IPL neurons coded not only the observed motor act but also allow to understand the demonstrator’s intentions [4].

Iriki et al. found bimodal neurons (somatosensory and visual neurons) in the monkeys intraparietal cortex, which incorporated a tool into a mental image of the hand [1]. This group of the neurons responds to the both stimuli from the visual receptive field and the somatosensory receptive field. After the tool use, the visual receptive field of these neurons is extended as to include the tool. In [5], they trained a monkey to recognize the image of their hand in a video monitor, and demonstrated that the visual receptive field of these bimodal neurons was projected onto the video screen. The experimental results suggested that the coincidence of the movement between the real hand and the video-image of the hand seemed to be essential for the monkey to use the video-image to guide their hand movements.

In robotics, developmental sensorimotor coordination is well studied involving neuroscientific aspects and developmental psychology; e.g., sensorimotor prediction [6] [7], mirror system [8] [9], action-perception link [10], and imitation learning [11] [12].

The body presentation plays an important role for a robot to connect the own body to voluntary action and its effect [13]. Hikita et al. proposed a bimodal representation (visual and somatosensory representation) of the end effector based on Hebbian learning [14], which simulated the experiments with monkeys in [1]. Stoytchev proposed developmental video-guided reaching [15] which demonstrates similar tasks examined in [5]. Kemp et al. approached the robot hand discovery utilizing the mutual information between the arm joint angles and the visual location of an attractive object [16]. Saegusa et al. proposed an own body definition system based on visuomotor correlation, which builds a visuomotor memory of the own body regardless of body appearances or kinematic structure [17].

In the literature of object manipulation, Natale et al. proposed a developmental grasping system which allows the own hand recognition [18]. Object affordance, which is the possibilities of action toward the object, plays an important

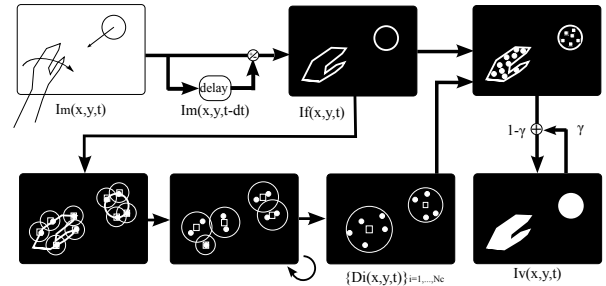


Fig. 3. Motion detection. Small motion area is integrated as large motion area in the bottom-up manner.

role for object interaction [19]. Montesano et al. proposed a learning model of object affordance using Bayesian networks [20]. The probabilistic links among action, effect, and object allows action imitation.

The framework which we are proposing is rather oriented for exploring the own body in an active manner and the mirroring actions as monkeys perform [2] [3] [4].

III. VISUO-PROPRIOCEPTIVE PERCEPTION

A. Definition of Objects

We explore a way to imprint a general sense of objects on a robot. Our idea for the definition of object is simple; we define an object as a manipulable thing, which includes what moves by itself (e.g., a robot hand and a human hand) and what the robot and human can move (e.g., a ball and a box). The perception of environmental objects (e.g., a table and wall) are important for safety in manipulation, while we shall keep this type of objects for later discussion, and here concentrate on the perception of manipulative objects.

The object definition based on manipulability was studied in [8]. In our framework, we generalize the idea to allow the action identification among different action demonstrators. An advantage to rely on the manipulability is that it is a reasonable solution in order to know the physical entity; i.e., its independence from the environment. A visual scene is just a picture for a robot in an initial phase. The robot cannot conclude in advance of interaction whether an object in sight is real or fake (Fig.1). The interaction gives a physical segmentation of the object itself.

B. Perception of Motion

Motion detection is illustrated in Fig.3. The absolute subtraction of a monochrome image $I_m(x, y, t)$ from the image $I_m(x, y, t - \tau)$ in the previous frame gives a flicker image $I_f(x, y, t)$ as follows,

$$I_f(x, y, t) = |I_m(x, y, t) - I_m(x, y, t - \tau)|, \quad (1)$$

where x, y, t denotes the horizontal coordinate, vertical coordinate, and the time when the image is sampled. τ denotes the time interval for image sampling.

A set of points is randomly sampled from high intensity points on the flicker image. Initially, a small disk is given to

each sampled point as it is set as the center. The i th disk D_i is represented as follows;

$$D_i(x, y, t) = \{x(t), y(t) | \sqrt{(x - x_i)^2 + (y - y_i)^2} \leq r_i\}, \quad (2)$$

where (x_i, y_i) and r_i denote the center and radius of the disk, respectively. Then, the neighbour disks are grouped as a new disk, if the disks have intersection. The D_i and D_j have intersection if the following formula holds;

$$\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} < |r_i + r_j|. \quad (3)$$

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

The final set of the disk centers $S_c = \{(x_i, y_i)\}_{i=1, \dots, N_c}$ is used for segmentation of the moving objects. N_c is not a constant, but dynamically given by the result of motion area integration. Practically, N_c is the similar number of the moving objects. The high intensity points on the flicker image are assigned for the nearest disk center and form outlines of segmented motion area. The random interpolation of the outlines gives a set of points which cover the motion area. The points are blurred by the Gaussian kernel and accumulated temporal, which forms clouds of motion area. The accumulated motion image $I_v(x, y, t)$ is formulated as follows;

$$I_v(x, y, t) = \gamma I_v(x, y, t - dt) + \alpha \sum_i K(x, y, x_i, y_i), \quad (4)$$

$$K(x, y, x_i, y_i) = \exp\left\{-\frac{(x - x_i)^2 + (y - y_i)^2}{2\sigma^2}\right\}, \quad (5)$$

where (x_i, y_i) denote the image coordinates of the i th point. $\gamma \in [0, 1]$ is the decay rate.

C. Perception of the Body

We studied the own body definition [17]. The own body can be defined with the motor correlation between vision and proprioception. The improved own body detection system is illustrated in Fig.4. The system functions as a proprioceptive image filter which selectively extracts the own body and the extended body.

When the robot moves its own hand in the view field, the visual motion of the hand is correlated with the proprioceptive motion. The body filter accumulates this visuo-proprioceptive correlation on the visual field and applies the correlation map to body extraction. The correlation map $VP(x, y, t)$ is given by the following equations;

$$I_{vp}(x, y, t) = \gamma I_{vp}(x, y, t - dt) + \alpha C(t), \quad (6)$$

$$C(t) = \begin{cases} 1 & \text{if } |dq/dt(t)| > p_u \text{ and } I_v(x, y, t) > v_u, \\ -1 & \text{if } |dq/dt(t)| < p_l \text{ and } I_v(x, y, t) > v_u, \\ 0 & \text{otherwise.} \end{cases}$$

where $|dq/dt(t)|$ denote the velocity norm of the joint angle vector. The baseline of I_{vp} is set as the center of the range (128 in $[0, 255]$). By the short term accumulation of the

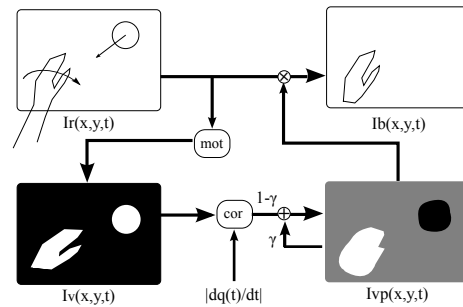


Fig. 4. Body filter. The visuo-proprioceptive correlation is projected on the visual field and accumulated for short time. The correlation map is applied to body extraction.

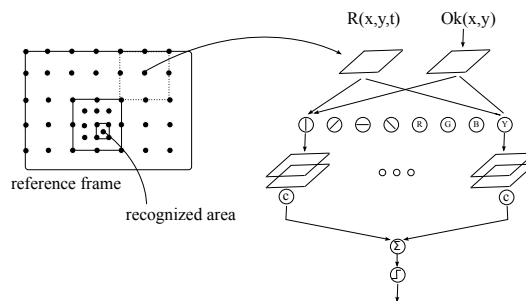


Fig. 5. Object identification. The object of interest is searched in the image frame, and tracked if founded. The image search is in the manner of recursive binary search with channel based matching.

correlation, the body part is marked in high intensity, while an independently moving object is marked in low intensity unless the object is synchronized with the body part in terms of motion. The joint angles and visual location are stored in the memory with the appearance of the extracted body.

D. Perception of Objects

An object is sensed by its motion. A set of points are randomly sampled from the moving objects. The color and shape cues of the points characterize the object visually. We define the color cue c and shape cue e with four points in UV color space and four directions of the edge, respectively. Each point is characterized by the most similar color and shape cue. A histogram of the cues $(c_1, \dots, c_4, e_1, \dots, e_4)$ represents the object. This feature vectors are voted for short time, and the object which forms the principal cluster in feature space is stored in a memory. The memorized objects in the memory are identified during the manipulation. The object identifier is illustrated in Fig.5.

IV. ACTION MIRRORING

Imitation is a strong paradigm for transferring motor intelligence. A problem when realizing imitation of actions among humans and robots is how to characterize an observed action to be less dependent on the body context of the demonstrator. Apparently, a robot and a person have different kinematic structures and dynamics. Moreover, a way to demonstrate an action is not unique.

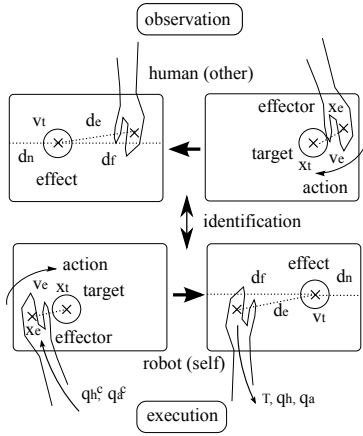


Fig. 6. A schematic representation of the features defined in the context of action observation and execution.

Our approach for action characterization is based on a physical effect for an object. What the robot should understand in object interaction is rather the results of the action than the trajectory of the effector. Object interaction is sensed in multi sensory modalities such as vision, touch, and force sensing. The tactile and force feedback is not available when observing the action, while visual feedback is available for both of when observing and executing the action. For this reason, we introduce the vision-dominant characterization of actions.

A. Visual Features

We define features given from vision system. A schematic representation of the visual features is illustrated in Fig.6. We define the following quantities; x_e is the position of the effector in the visual field (either a robot or an agent performing an action), and x_t is the position of a target in the visual field. x_n and x_f are what we call *near* and *far* sides, i.e., the fixed position in the image with respect to the agent and his effector. In the case of the robot and the left arm, the *near* side is a point at center-right of the image and the *far* side is the point at the center-left of the image.

We define the following visual features:

$$d_e(t) = |x_t(t) - x_e(t)|, \quad (7)$$

$$d_n(t) = |x_t(t) - x_n|, \quad (8)$$

$$d_f(t) = |x_t(t) - x_f|, \quad (9)$$

$$v_t(t) = |dx_t/dt|, \quad (10)$$

$$v_e(t) = |dx_e/dt|, \quad (11)$$

where d_e , d_n , d_f represent the visual distance of the effector with respect to the target, the visual distance of the target with respect to the near and far side respectively. v_t , and v_e are the velocity norm of the target and the effector.

When given a target, the robot regards what approaching the target as an effector. The effector can be a robot hand and an experimenter's hand, but the robot is able to recognize whether the effector is its own body or the other (refer to the previous section). When the distance d_e is less than

a threshold, it starts to recode the observing action. The experimenter selects a target from the objects that the robot recognizes. We think this assumption is natural, since in the experiments with monkeys, an experimenter assumes that a piece of food (i.e., the target for the monkeys) attracts the monkeys.

The distances defined above abstract the positions as the geometrical relation among the target, the effector, and the environment. Here, the absolute location of the near side and the far side are demonstrator dependent, because the near and the far is relative to the location of the demonstrator. In this framework, we defined the near side as the side of approach in which the arm moves close to the body. The far side is its opposite. The robot knows who is demonstrating an action from the own body perception and motor command generation. This information is used to interpret x_n and x_f .

B. Proprioceptive Features

We define features based on feedback from proprioceptive sensory systems.

$$T_i(t) = \sum_j T_{ij}(t), \quad (12)$$

$$v_{hi}(t) = |dq_{hi}/dt|, \quad (13)$$

$$v_{ai}(t) = |dq_{ai}/dt|, \quad (14)$$

where T_{ij} is the activation of the j th taxel of the i th finger, and T_i is the overall activation on the i th finger. q_{hi} and q_{ai} are the joint angle of the i th (left or right) hand and arm, respectively.

C. Action Coding

A simple action such as picking up an object is still complicated in the level of the joint motor command. In the literature of dynamical systems implemented by a connectionist model, an action is encoded as wiring of networks, and its recursive computation decodes the action as a sequential motor command. There are some models to decompose an action into motor primitive [21], but it seems still difficult to reorganize the motor primitives for other motor tasks.

We simply encode an action as a sequence of motor primitives, and decode the code with a visuomotor context. In the similar manner of synergistic reflexes, we assume a set of motor synergy units and target states. We assume the following synergy units {left arm, right arm, left hand, right hand}. The robot is trained to demonstrate target states {release, grasp} for hand synergy units, and target states {home, target, near, far} for arm synergy units. The summary is listed in Table I. The functions of the arm and hand synergy units are identical for left and right (abbreviated as 'L' and 'R', respectively). The target state of the arm synergy is dynamically decoded based on the visuomotor context. At this aim the robot previously built a correspondence between the arm joint angles and the visual location of the hand. The grasping action of the hand synergy was simply embedded as a reflex action. A single action code is presented as a couple of binary vectors which identifies the synergy unit

TABLE I
SYNERGY UNIT AND TARGET STATE.

synergy unit	unite notation	target state	state notation
arm	LA / RA	home	HO
arm	LA / RA	target	TA
arm	LA / RA	near	NE
arm	LA / RA	far	FA
hand	LH / RH	release	RE
hand	LH / RH	grasp	GR

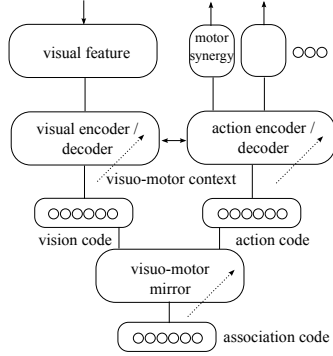


Fig. 7. Visuomotor mirror. The observation code (visual code) and execution code (action code) are mirrored.

and the target state. For example, the action of the left hand grasping is encoded as $\{LH, GR\}$ (left hand, grasp). A set of single action codes composes a sequential action code. For example, a sequential action of reaching and grasping with the left arm and hand is encoded as $\{\{LA, TA\}, \{LH, GR\}\}$.

The visual code (observation code) and action code (execution code) are linked together in the action representation as illustrated in Fig.7. The coupled representation is autonomously given by observing a self-executed action, and stored in a memory. The vision code is given based on the visual features by equation (7)-(11) after the action observation. The action code can be generated randomly in the manner of action babbling. In the following experiments, however, the action code was manually given, since we aimed at simulating similar actions demonstrated by the monkeys in [3].

When a robot recognizes an action by observation, it compares the vision code of the observed action with the ones that are learned before. The comparison is based on Euclid distances between the vision code and the ones that are stored before. Then, the action code coupled with the minimum distant vision code is recalled for action imitation.

V. EXPERIMENTS

We performed experiments to evaluate perceptual ability of the proposed system. In the preliminary experiments, we examined the visual perception systems. In the main experiment, we demonstrated action mirroring with humanoid robot iCub [22]. The body structure of the robot is illustrated in Fig.8.

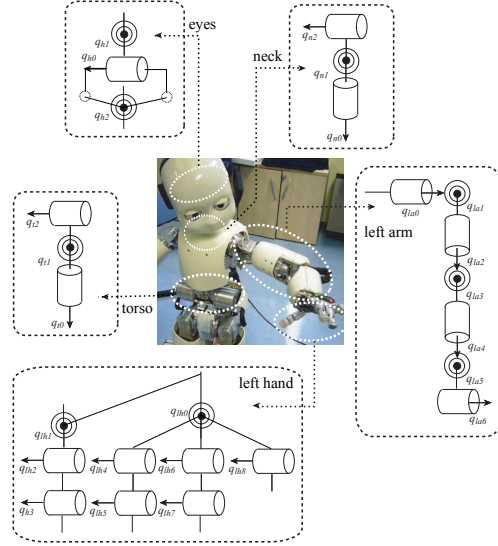


Fig. 8. The robot platform iCub [23]. The right parts of motor synergies are abbreviated in the figure (The arm and hand are identical in left and right).

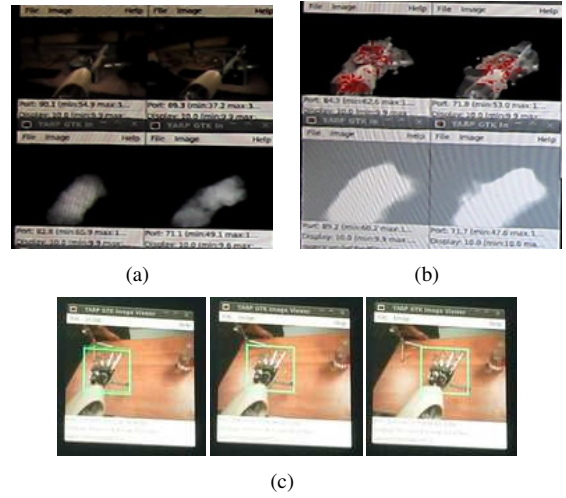


Fig. 9. Binocular own body perception by stochastic motor exploration. (a) The reference images (top) and motion area (bottom). (b) The extracted body (top) and visuo-proprioceptive correlation area (bottom). (c) After the body definition, the hand and arm are visually identified and tracked without using the motion cue.

A. Visual Perception

A result of own body perception is shown in Fig.9. The robot moved its own arm randomly for a few seconds, and generated a visuo-proprioceptive correlation map. Based on the map, the body part was extracted.

A result of object identification for a static image is shown in Fig.10(a). In this test, we gave a single reference image and a target template sampled from the reference. A human face bounded by a green box is the target and the grand truth of the image search as well. The search points converged to the grand truth. In the object identification, the target image was decomposed into eight channels (bottom images) and

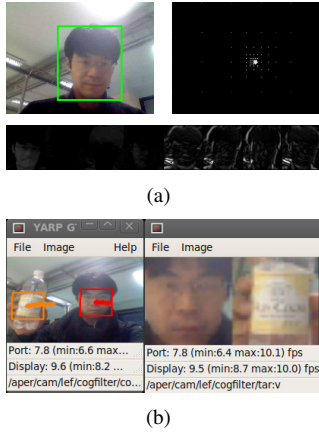


Fig. 10. (a) Object identification for a static image. Reference image with a target area (top left), search points on the reference frame (top right), the decomposed image channel of the target image (bottom). (b) Object identification for a video stream. Object tracking (left) and target objects (right).

TABLE II
ACTION DEMONSTRATED IN THE EXPERIMENTS.

action	tri.	demo.	action code	initial
hold	5	Rob	$\{\{LA,TA\},\{LH,GR\},\{LA,HO\}\}$	free
place	5	Rob	$\{\{LA,TA\},\{LH,RE\},\{LH,HO\}\}$	grasp
take	5	Rob	$\{\{LA,TA\},\{LA,NE\}\}$	free
hold	5	Man	N.A.	free
place	5	Man	N.A.	grasp
take	5	Man	N.A.	free

evaluated by integrating the matching scores of the channels.

Object recognition with a video stream is shown in Fig.10(b). A face and a bottle were given to the system as the targets to track on the reference frame. The targets were identified successfully when they were moving. The identification was robust for changes of the scale and orientation because of the channel-based image matching.

B. Setup of Actions

An experimenter and iCub performed sequences of actions. The demonstrated actions and the number of trials are listed in Table II. In the table, $\{LA,TA\}$ denotes reaching for a target with a left arm. $\{LH,GR\}$ denotes grasping of a target with a left hand. The action capability depends on the initial state of the effector and the target. As shown in Table II, we assume that the effector is free before executing the hold and take action, and the effector is grasping an object before executing the place action.

The variation of the actions are based on the monkeys experiment [3], in which the action of grasp, hold, place, manipulate, and bimanual interaction were examined. Here, we took the hold, place and manipulation. In the original literature, manipulation is an action to take a piece of food to possess it. In this sense, we renamed the action of manipulation as take. Grasping action is included in the actions which we defined here. Bimanual interaction was excluded in this experiment for simplicity.



Fig. 11. Grasping an object as performed by an experimenter and the robot. The left two columns show the scene of the action demonstration. The right two columns present the observed action from the viewpoint of the robot.

In the trials of the experimenter's demonstration, the robot observes the experimenter's action, and recognizes the type of the action. The parameters for visual features encoding can be learnt from the results from the robot's demonstration, though we set empirical values for the parameters to concentrate on the action mirroring.

A part of the action sequence (reach and grasp) is shown in Fig.11. The reaching and grasping are the primitives performed by the arm and hand synergies. The action was executed by an experimenter and the robot. The robot observed the both actions.

C. Action Mirroring

The observation of full actions is shown in Fig.12. The recognized target and effector are presented with a blue and red box respectively. In the experiment, the near side (i.e., the own side) of the robot and the experimenter were defined as the right and left side, respectively. The far side is its opposite. The hold and place action are characterized by the relation between the effector and the object. In the final phase of the hold action, the robot keeps the object in the hand, while in case of the place action, the robot moves the hand away from an object. On the other hand, the take action is characterized by the relative distance of the object from the "near" side (the right side of the image).

Profiles of visual features are shown in Fig.13. In the figure, the horizontal axis indicates the sampling time, and the vertical axis indicates the distance. For the calculation of the distance, the image coordinates were normalized with respect to the center and the diagonal line of the image. The image center was set as the origin of the coordinates (0,0), and the coordinates were scaled as the length from the image center to an image corner is 1.0. Therefore, the domain of the above distances is [0,2]. For example, the distance from the target at the center to the near and far sides should be 1.0 or less.

The red plot in the hold action clearly shows that the effector approached the target and kept the target in hand.

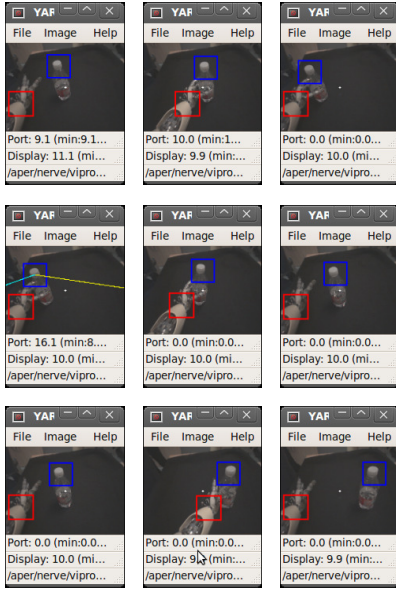


Fig. 12. Three demonstrated actions. Hold, place, and take action are presented in the top, middle, and bottom row, respectively. The time course in each row is from left to right. Blue and red boxes represent a target and an effector.

The same plot in the place action shows that after the effector released the target, the effector went away from the target. In the take action, the red line shows the "U" shape profile, since the distance between the effector and the target changed as far-close-far. It approached the target, pushed the target towards the near side (which corresponds to the "own" side), and retracted afterwards. The green and blue line denote the distance from the near and far side. The take action is characterized as the action that shortens the distance of the target to the near side. Fig.13 shows that the visual features of the action demonstrated by the robot were very similar to the ones demonstrated by the human experimenter, even though the own and other side were flipped (the near side and far side are dynamically adapted with respect to the direction of approach of the effector).

The visual features are encoded as the parameters in Table III. A visual distance in the initial and final part of the profile was encoded by averaging and applying a threshold. Here, the initial and final part were the 10% of the profiles which included the first sample and the last sample, respectively. We used value 0.5 and 0.85 for thresholding the distance between target-effector and target-near/far, respectively. The threshold is operated by the step function. The threshold can be determined automatically, but in this experiment they were optimized experimentally. The initial and final code were coupled together, and used for coding the visual observation.

The profiles of the touch feedback during object interaction is shown in Fig.14. Interestingly, the touch feedback was corresponding to the profile of the visual feature d_e (see Fig.13). Indeed, activation of the touch sensation was detected when the distance between the effector and the

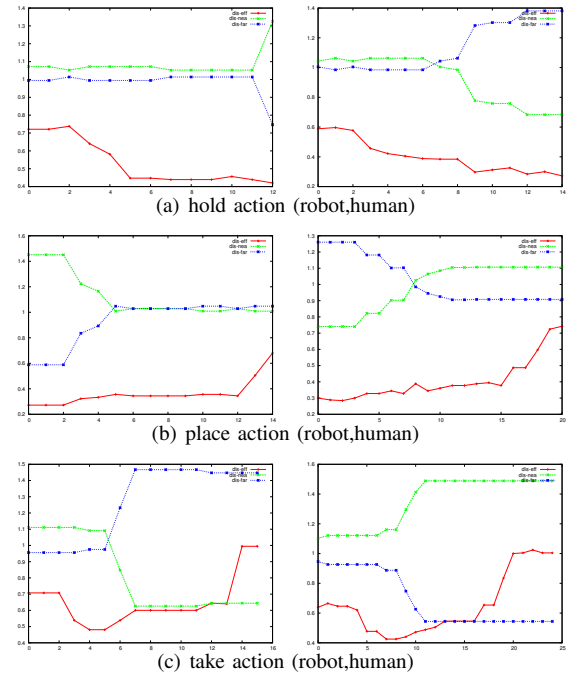


Fig. 13. Profile of visual features during action observation. Left: observation of self-performed actions by a robot. Right: observation of actions performed by a human demonstrator. The shapes of the blue and green plots are flipped (robot vs human), because the point of view of the observer is different. However, this is dynamically adapted with respect to the action demonstrator.



Fig. 14. Features associated to tactile feedback during observation of self-performed actions.

target was short. In general, contact detection by vision is quite difficult, since the system is required to recognize the surface of an object precisely. In this sense, touch feedback plays an important role for the robot to verify the contact in vision.

The visual code is universal for action demonstrators. Therefore, when the action observation was successful, the robot could mirror the action presented by a human experimenter. The success rate of action mirroring is shown in Table IV. Overall 15 actions were demonstrated to the robot,

TABLE III
VISUAL CODE.

action (phase)	target-effector	target-near	target-far
hold (initial)	1	1	1
hold (final)	0	1	0
place (initial)	0	1	0
place (final)	1	1	1
take (initial)	1	1	1
take (final)	1	0	1

TABLE IV
SUCCESS RATE IN ACTION MIRRORING.

action	success/trial	detection rate
hold	3/5	60%
place	3/5	60%
take	4/5	80%

in 10 cases the robot was able to detect and mirror the correct action (the accuracy was 66%). In the current experiment set-up, the start and end of the action observation were given to the robot, and the target position was set in quasi center in sight. For the future demonstration, however, the target position is not important because of the visual context based action decoding. In the proposed system, the action code and vision code are presented with binary values. The babbling on the action code (combination of the primitive actions) allows the high level action exploration rather than the primitive-level exploration.

We examined the learning based action mirroring system which associates the observed actions with the self-generated actions. The association of observation and execution in this work still remains in a reflex like response, in which a robot faithfully demonstrates the observed action. In the future work, we will bias association of actions with reward stimuli. This schema allows the robot to respond to an observed action by demonstrating another action in the learned action vocabularies. For example, when an opponent points out the object, the robot can respond to this action by taking an object. We expect that understanding of the intention behind an action can be achieved by action response learning.

VI. CONCLUSION

We proposed a framework of action mirroring. The robot discovers the body through self-generated movements, and distinguishes it from the other objects based on the visuomotor correlation. After the trials of exploratory actions, the robot learns primitive reaching and grasping actions. In the object interaction with humans, the robot characterizes manipulative action with the effect for an object, which is universal for action demonstrators. This manner of action coding allows the robot to demonstrate an observed action.

The developed system was experimentally evaluated with a humanoid robot using vision, touch, and proprioceptive sensing. This attempt is a robotic simulation of the mirroring function in a premotor cortex of monkeys [2] [3] [4]. In a future work, we are extending the framework towards the mirroring of action "relations". The extension could realize an action based language among robots and humans.

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] 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.
- [3] V. Gallese, L. Fadiga, L. Fogassi, and G. Rizzolatti, "Action recognition in the premotor cortex." *Brain*, vol. 119, pp. 593–609, 1996.
- [4] 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.
- [5] 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.
- [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] 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.
- [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, "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.
- [18] 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.
- [19] J. Gibson, *The ecological approach to visual perception*. Lawrence Erlbaum Associates, 1986.
- [20] 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.
- [21] 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.
- [22] G. Metta, G. Sandini, D. Vernon, L. Natale, and N. F., "The icub humanoid robot: an open platform for research in embodied cognition;" in *Proceedings of the 8th Workshop on Performance Metrics for Intelligent Systems*, Washington DC, USA, 2008, pp. 50–56.
- [23] G. Metta, P. Fitzpatrick, and L. Natale, "Yarp: Yet another robot platform," *International Journal on Advanced Robotics Systems*, vol. 3, no. 1, pp. 43–48, 2006.