Cognitive Robotics - Active Perception of the Self and Others -

Ryo Saegusa, Lorenzo Natale, Giorgio Metta, Giulio Sandini

Abstract— **An approach of active perception for robots is described. The main message of the paper is that a robot can understand human's action more naturally by comparing own action effects with human's action effects. The action perception is composed of three factors; (1) a robot voluntarily generates actions and discover the own body and objects to interact with. (2) the robot characterizes its own action based on the effect for objects. (3) the robot identifies the own action with the human action based on the effect for the object. Developmentally, a robot acquires motor intelligence from humans through the mirrored perception of the own action and the human's action.**

I. INTRODUCTOIN

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 [1] [2]. Monkeys are able to recognize their own body under various conditions, and extend their body schema while using a tool [3] [4]. Also, they associate others' behaviors with their own [5]. This kind of cognitive functions may have potential to break a limit in existing hand-coded intelligence of robots, and bring more interactive ability with humans.

Our goal is to realize a cognitive system which actively develops its perception of the self and others through sensorimotor interaction. The active perception which we are proposing is inspired from social interactions in humans and primates in which is emerging a question; why do primates imitate or mirror others' behaviors? Our inference for the question is that primates desire to transfer their intelligence by mirroring actions and sharing intention behind the actions. In this article, we are going to introduce a simple object interaction realized by a primate-like active perception. We challenge to generalize assumptions for a robot about the self and other objects to keep away from task-specific tricks; A robot starts to distinguish the self from other objects by making action, and associates the own actions with its effects. Then, humans actions are identified with an own action on the function level based on effects for objects. The approach potentially explores a clue as well for robots to understand intention behind human's actions.

II. METHOD

We shall define a physical object in our body scale as "what occupies space". In the context of object manipulation that we now focus on, however, it looks enough if we assume a manipulable entity as an object of interest. We sense objects in multi modal channels such as tactile, haptics, auditory, visual sensing, etc. We start discussion about an object perception mainly with visual cues, then consider the integration of the other modalities afterwards.

This work was supported by EU project CHRIS

R. Saegusa is with Robotics, Brain and Cognitive Sciences Dept. Italian Institute of Technology, Genoa, Italy ryos@ieee.org, {ryo.saegusa, lorenzo.natale, giorgio.metta, giulio.sandini}@iit.it

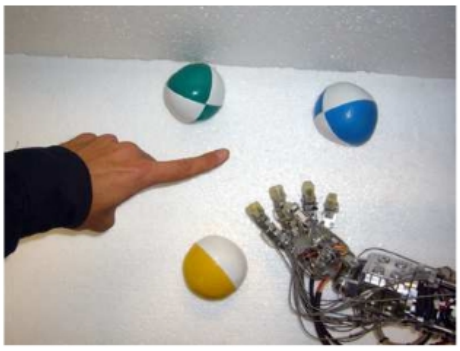

Fig. 1. Interaction among a robot, a person, and objects.

The visual scene is just a picture for a naive robot; in other words, the robot without any knowledge of objects perceives a visual scene just as a set of colored textures. Visual markers are practically useful to identify objects, and moreover the markers give the most important information; what an object is. However, this kind of tricks is weak for knowing new objects developmentally.

Our idea of object definition is based on a motion cue in the visual scene. The principal is simple; an object is what a robot can move. Something which occupies space but is hard to move (e.g. a table and wall) are now out of our scope of manipulation (but still it should be perceived for safety during manipulation). The bottom-up procedure of object perception is composed of three steps; motion detection, object detection, and object identification. In this phase, the robot body and the other entities are all objects for the robot. The own body perception is given next, and then action identification between a human and a robot is framed.

A. object perception

Fig.2 illustrates motion area detection. The absolute subtraction of a monochrome image $M(x,y,t)$ from the previous frame $M(x,y,t-\delta t)$ gives a flicker image $F(x,y,t)$ as follows,

$$F(x,y,t) = |M(x,y,t) - M(x,y,t-\delta t)|, \quad (1)$$

where x,y,t denotes the horizontal coordinate, and vertical coordinate, and the time when the image is captured. The n_f points on the flicker image are randomly sampled and linked

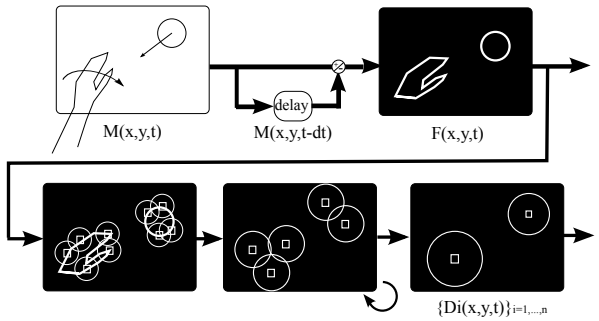

Fig. 2. Motion detection.

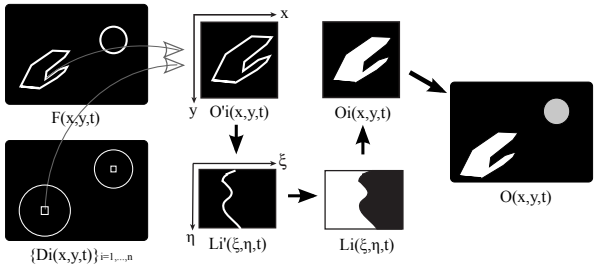

Fig. 3. Object detection.

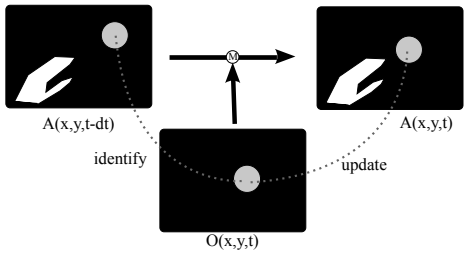

Fig. 4. Object identification.

each other, if the points are closer than the initial radius r_f. The linked points are abstracted by a domain of a circle D_i with a center (x_i, y_i) of the linked points and a radius r_i of the average deviation from the center to the linked points. The domain is formulated as;

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

where index i denotes the label of the set of the linked points. The area abstraction is repeated while a new link can be generated with the update radius. As a summary, given the sequential two monochrome frames, the motion detector returns giometrically-grouped moving areas $\{Di(x,y,t)\}_{i=1,\ldots,n}$. The parameters of the motion detector are the number of sampling points n_f and the initial radius r_f, i.e. resolution of detection and the minimum object size to detect.

Fig.3 illustrates object detection. The moving areas D_i on the flicker image is projected onto the log-polar coordinates (ξ, η), where ξ and η denotes the log-scale radius and the angle of the target from the center on the original Cartesian coordinates (x, y) as follows;

$$\xi = \log\{\sqrt{x^2 + y^2} + 1\}, \quad (3)$$

$$\eta = \arctan y/x, \quad (4)$$

where Cartesian coordinates are normalized as a point inside the unit circle, i.e. $\sqrt{x^2 + y^2} \leq 1$, which limits the domain of the log-scale radius as $\xi \in [0, \log 2]$. The domain of the angle which arctangent returns is modified as $\eta \in [0, 2\pi]$ after the transformation. The fragments of on the log-polar coordinates is interpolated as a function $\eta = f(\xi)$ by local weight regression with a Gaussian smoothing kernel;

$$f(\eta) = \frac{\sum_i \xi_i k(\eta, \eta_i)}{\sum_i k(\eta, \eta_i)}, \quad (5)$$

$$k(\eta, \eta_i) = \exp \frac{|\eta - \eta_i|^2}{2\sigma^2}, \quad (6)$$

where the interpolated curve corresponds to an outline of the object in the Cartesian space. The log-polar image is binarized with the interpolated curve then, it is inversed to the original Cartesian coordinates $O_i(x,y,t)$, where i denotes the identity of a blob of moving object. The object extraction is performed for all moving areas. These areas are integrated on an object image $O(x,y,t)$. As a summary, given a flicker image and motion areas, the object detector gives identified object areas. The parameters of the object detector are the deviation of the smoothing kernel σ, which controls the sensitivity for convex of the object outline.

Fig.4 illustrates object identifier. The object identifier idenfies instanteneous objects $\{O_i(x,y,t)\}_{i=1,\ldots,n}$ (given by the object detector) as objects $\{A_i(x,y,t)\}_{i=1,\ldots,n}$ in the attention memory and merge them, if location and texture of areas are enough similar. Otherwise, the instantaneous object is added as a new object in the attention memory. Initial attention memory is given by the instantaneous objects detected first.

B. Perception of the self

We have discussed the own body definition in the literatures [6] [7]. The basic idea of the own-body definition is to use the correlation between the visual and proprioceptive sensory feedback as an indicator of the own generated movements. Fig.5 shows the own-body definition based on the correlation between the motion in vision and proprioception. The motor information of the ith object in sight is simplified as the velocity norm of the object center denoted as $v_{xi}(t) = |dx_i/dt(t)|$. The motor information of the proprioception is simplified as the velocity norm of the arm joint angles denoted as $v_q(t) = |dq/dt(t)|$. The correlation of the motor information is;

$$c(i,t) = \frac{\sum_k v_{xi}(t - k\tau) v_q(t - k\tau)}{\sum_k v_{xi}(t - k\tau) \sum_k v_q(t - k\tau)}, \quad (7)$$

which is a canonical correlation normalized as $c \in [-1, 1]$. The object is perceived as the "self" (an own body), if the correlation is greater than a certain threshold. Otherwise, the objects are perceived as the "others".

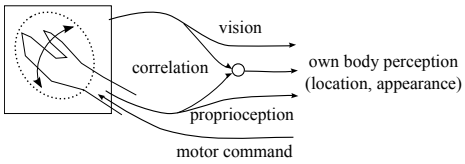

Fig. 5. Own-body definition.

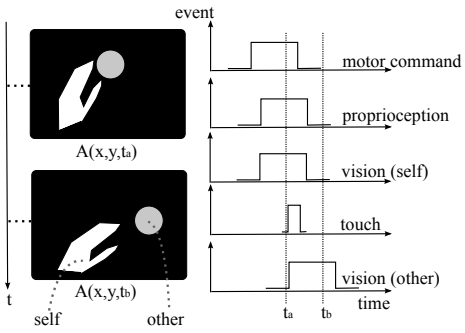

Fig. 6. Cause and effect in interaction.

C. Cause and effect in interaction

How can a robot learn interaction? Though interaction understading in general sense is too broad, we challenge to break down the problem based on the sensorimotor ability of the robot. Our idea is that a robot makes an action and then, couples the action with its effect observed by its own sense. The discovery of a cause-effect rule behind observed events is connected to a natural understanding of the interaction, then.

Fig.6 shows sensorimotor events in interaction. The robot commands to move its arm, then a movement is sensed in proprioception and vision. When a tip of the body part arrives in the surface of an object, the tactile feedback is given as well as the the visual movement of the object. Here, we simplify complex sensorimotor signals as a single binary value which represents an onset and offset of an event. Then, the sensorimotor events in the different modality are coupled, if the events are oberserved in a short time delay. The event of the motor command, proprioception and vision are given by thresholding the norm of the velocity profile in each modality. The event of the touch is given as an integrated signal of a sensor array with thresholding.

D. Action identification

Imitation is a strong paradim to transfer intelligence. A problem in imitation is how actions are characterize and identified. A redundant kinematic structure such as a dextrose arm of the robot gives huge variation of actions. Even if an action is specified, the imitation of the action with the different kinematic structure (e.g. imitation of a human's arm movement by a robot arm) is not trivial.

Our idea of the imitation is to characterize an action based on the physical effect for objects. The function of the action can be categorized in the resolution of the results, which is sensible as the effect for objects. Apparently, we make

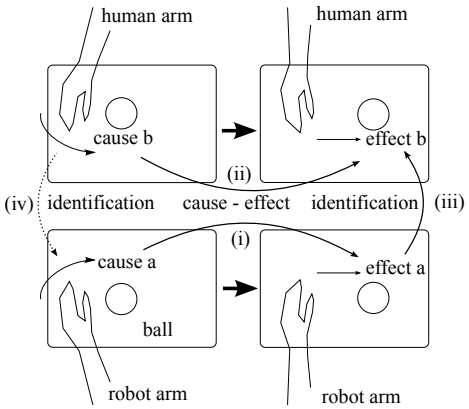

Fig. 7. Action identificaiton.

actions which are not directed for objects such as walking, watching and speaking; however, these variation of actions can be characterized with the effect in cognitive state space. For a manipulation task, it seems natural to frame an action with its result for objects. Moreover, effect for objects is free from a demonstrator of the action and the context of the body; i.e. the actions by different body structures can be easily identified by observing the effect rather than observing the action itself.

Fig.7 illustrates a scenario to identify a pushing action of objects. Based on the self identification, the robot can distinguish the own arm from the ball. Then, it extracts a cause-effect rule in pushing action, which couples the arm movement with the object movement (i). On the other hand, an experimentor demonstrates a similar pushing action (ii), then the robot identifies the person's arm action (where the person's arm is also a normal object as well as the ball for the robot) with the own arm action (iv) through the observation of the same effect for the ball (iii).

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

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

[6] R. Saegusa, G. Metta, and G. Sandini, "Self-body discovery based on visuomotor coherence," in *Proc. of 3rd International Conference on Human System Interaction (HSI10)*, May 3-8 2010, pp. 356–362.

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