Developmental Action Perception for Manipulative Interaction

Ryo Saegusa, Giorgio Metta, Giulio Sandini, Lorenzo Natale

Abstract—The paper describes a developmental framework of action-driven perception in anthropomorphic robots. The key idea of the framework is that action develops the agent’s perception of the own body and its action. In this framework, a robot voluntarily generates movements, and then develops the ability to perceive its own body and the effects of action primitives. The robot, moreover, demonstrates manipulative actions composed of the learned primitives, and characterizes the actions based on their sensory effects. After learning, the robot can predictively recognize humans’ manipulative actions with cross-modal recovery of unavailable sensory information and reproduce the recognized actions. We evaluated the proposed framework in experiments with a real robot. In the experiments, we achieved developmental recognition of human actions as well as their reproduction.

I. INTRODUCTION

Action perception is fundamental for robots in manipulative collaboration with humans. In the context of developmental manipulation, however, motor learning for task operations are more focused than perceptual development of generated actions. In physiological studies, it is known that monkeys have brain area to link execution of actions and observation of actions [1] [2] [3]. The goal of this work is to realize a cognitive system which actively develops a perceptual model of actions which can be used for perception of actions demonstrated by other action agents (e.g., humans). 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. Therefore, the perception of actions demonstrated by robots and humans is not handled equivalently at a perceptual level. In this paper, we introduce a developmental action perception system with voluntarily learned action primitives such as fixation, reach and grasp actions.

In robotics, developmental sensory-motor coordination involving neuroscientific aspects and developmental psychology is well studied; e.g., sensorimotor prediction [4][5], mirror system [6][7], action-perception link [8], and imitation learning [9][10] are representative studies. Object affordance (or possible actions to operate an object) plays an important role in manipulation [11][12][13]. In literature on robotic object manipulation, Natale et al. proposed a developmental grasping system that allows self hand recognition [14]. Montesano et al. proposed a learning model of object affordance using Bayesian networks [15]. In this work, the probabilistic links among action, effect, and object allow plausible action imitation [7]. Oztop et al. proposed a biologically comparable model of mirror systems [16][17]. Castellini et al. studied an effect of object affordance in object recognition [18] 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, our framework supports learning of action primitives (fixation, reach and grasp) based on [19]. Manipulative actions for objects are built up from the action primitives. The robot, then, develops a perceptual model with self-generated actions. The learned knowledge is applicable to perception of actions demonstrated by other agents (humans). Compared to the predictive recognition system in [16], we implemented the system on an actual robot and demonstrated the action perception in the real world. The Bayesian approach for action perception in [7][15] 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.

II. LEARNING OF PRIMITIVE ACTIONS

We introduce action-driven prototyping of the self and actions. Figure 1 shows schematic presentation of the action-driven development. The robot generates an action (fixation, reach, grasp and their combined actions) and observes the effect of the action. The causality relationship between an action and the effect characterizes the action as a perceptual model. We studied an autonomous body definition based on this framework in [19]. Here, we focus on development of motor control and action perception.

Based on autonomous definition of body parts, a robot can perceive effects of self-generated actions. We assume the following motor units and corresponding action primitives: 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. 2). In the paper, we do not explain learning procedures of action primitives because of the limit of space. It is based on Jacobian learning with motor babbling and object attention (refer to the details in [20]).
III. 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 (detailed later). 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 [1][2][21] 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

Figure 3 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 $\tau$ 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).

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

$$\delta z^t(t) = z^t(t) - z^t(t^a),$$

$$\delta d(t) = d(t) - d(t^a),$$

$$d(t) = |z^e(t) - z^t(t)|,$$

where $\delta z^t$ and $\delta d$ represent the change in the target position and the change in distance between the target and the hand, respectively. $t^a$ 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.

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

\[ \delta z^a(t) = \hat{z}^a(t) - \hat{z}^a(t^s), \]
\[ \delta d^f(t) = d^f(t) - d^f(t^s), \]
\[ \hat{z}^a(t) = \hat{q}^a(t), \]
\[ d^f(t) = |q^f(t) - \hat{q}^f(t)|, \]

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 \( q^f(T) \) corresponding to the visually identified object \( T \) to be grasped. Note that \( q^f(T) \) is afforded by the object \( T \) and the grasping posture is learned in a previous phase.

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

\[ \tau(t) = \max_i \tau_i(t), \]

where \( \tau_i \) denotes the maximum tactile intensity of all fingers. \( \tau_i \) denotes the summation of all tactile sensor values on the \( i \)-th finger tip. This maximization relaxes ambiguity of contact conditions.

### B. Action perception

We will now set forth action perception based on the above-defined multi-modal sensory features \( \{f^v, f^p, f^t\} \). The proposed action perception system is illustrated in Fig. 4. 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 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 the mirror neuron model of [16]. (AC3) represents the robot’s action execution in a blind condition. After a one-shot visual detection of a target object in 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 [1].

In the learning phase, the sensory features \( \{f^v, f^p, f^t\} \) 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 [22] for unsupervised clustering of each sensory feature. \( e^v \), \( e^p \) and \( e^t \) denote the visual, proprioceptive and tactile effect class, respectively. In the following, \( e_i \) represents either of \( \{e^v, e^p, e^t\} \).

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:

\[ \hat{a}(E = (\cdots, e_i, \cdots)) = \arg \max \limits_a p(A = a) 3 \prod \limits_{i=1} p(E_i = e_i | A = a), \]

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 [23]. When an other 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 the illustration of the module in Fig. 2) decodes the \( i \)-th action class \( a_i \) as a sequence of primitive actions \( \{a_i^v, a_i^p, \cdots\} \), and send signals to corresponding primitive action modules in the same order. For example, the grasping action class is composed of the grasping primitive action and the reaching primitive action, and the action generator sends execution commands in that order.

\( E_i \) and \( A \) represent corresponding random variables. Probabilities are given from 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 \( j \neq i \). This assumption seemed realistic, since the estimation was successful in experiments.

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:

\[ e'(E = (\cdots, e_i, \cdots)) = \arg \max \limits_{e'} p(E' = e') 3 \prod \limits_{i=1} p(E_i = e_i | E' = e'). \]
TABLE I
<table>
<thead>
<tr>
<th>ACTION PERCEPTION, EXPERIMENTAL CONDITIONS.</th>
</tr>
</thead>
<tbody>
<tr>
<td>action</td>
</tr>
<tr>
<td>grasp</td>
</tr>
<tr>
<td>place</td>
</tr>
<tr>
<td>hold</td>
</tr>
<tr>
<td>drop</td>
</tr>
<tr>
<td>poke</td>
</tr>
</tbody>
</table>

In (AC2), \( e' \) denotes the tactile class \( (e^t) \), which gives a tactile anticipation from the 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.

IV. 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 I. 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 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 (though 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. 5. 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.

We excluded failed actions from the evaluation in order to focus on the evaluation on action perception rather than motor control (though 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. 5. 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.

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 only using vision (without proprioception and tactile information) in order to compare the results with different action agents in the same condition.

Figure 6(a) with (b) 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 6(c) shows the consistency of the visual effect classification of the experimenter’s actions at intermediate states. Here, classification consistency \( c_e \) represents the number of trials with an identical classification result at the present \( t \) and the end of the action \( t_e \). Because of the high consistency in the intermediate phase of actions, the robot was able to predict the type of human actions from observation of the 1/2 of the actions with rate 86% (at time 2 in the figure).

2) Action perception: We will analyse the effect of the number of visual, proprioceptive classes on recognition rate. Figures 7(a), (b) and (c) show the recognition results of the actions performed by the robot. The recognized action

with a single object to eliminate noise from the comparison of perception in different modalities.
classes were given by Eq.9. Figure 7(a) shows the recognition results when the system used all sensory modalities (vision, proprioception and touch). Figure 7(b) shows the results when the system only used vision. Figure 7(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 efference copy presented in Fig. 4). Efference copies were used as the ground-truth action classes to evaluate the estimations.

Figure 7(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. 7(b) and (c), when some sensory modalities are unavailable, the available modalities (vision in (b), and proprioception and touch in (c)) should have high resolution to achieve a high action recognition rate. Figure 7(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. 7(b).

3) Cross-modal sensory anticipation: We will analyse the results of cross-modal sensory anticipation. Estimations from Eq.10 and actual perception are compared.

Figure 8(a) suggests that visual anticipation marked high scores, 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 [1].

Figure 8(b) suggests that tactile anticipation marks high scores, when resolution of the visual effect was 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 image 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 setin human environments may raise the robots’ sympathetic perception of humans.

4) Action reproduction by observation: We let the robot reproduce sequential actions from observation. Figure 9 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. 2). The action
generator then reproduced the actions in the buffered order. As shown in the figure, the robot reproduced the actions in the same order as the experimenter’s demonstration.

V. CONCLUSION

We proposed a robot’s developmental perception driven by active motor exploration. In the proposed framework, through self-generated actions the robot discovers the effects of actions on objects and recognizes the actions. In the development of perception, multi-modal sensing played an important role, since multi-modality allows cross-modal sensory anticipation.

In the proposed framework, 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 or observation. Another limitation in experiments was that the point of view of the experimenter was set similarly to that of the robot. Although the visual effect of actions are coded with the point-of-view independent features (object displacement and hand-object distance), the robot should explicitly understand differences in geometrical coordinates of action agents.

REFERENCES