Influence of Action Production on Perception in Human Infants: A Computational Approach

Jorge L. Copete (Osaka University) Jimmy Baraglia (Osaka University) Yukie Nagai (Osaka University) and Minoru Asada (Osaka University)

1. Introduction

In early infancy, humans are not yet able to detect the goal of others' actions. Later on, infants undergo a developmental process that allows them to perceive others' actions as goal-directed. Several studies have been carried out to reveal when and how infants start understanding goal-directed actions. Woodward [1] and Sommerville et al. [2] reported that young infants with goal-directed action experience showed a stronger novelty response to test events that varied the goal of the actions than test events that varied the motion path of the actions (e.g., the motion path). In contrast, infants did not show differentiated responses between both test events when the action was not recognized as goal-directed. This constitutes evidence that goal-directed action execution alters the perception of similar actions performed by other individuals.

In this study we build a computational model to clarify the underlying mechanism that accounts for the influence of the action production on the perceptual system. We argue that experience of action production enables infants to detect the goal in others' actions. Further, we want to explain the connection of this mechanism to the visual attention, in accordance to [2] where infants' experience of action production produced changes in their visual attention.

2. Hypothesis

Sommerville et al. [2] reported that infants' action experience alters their perception when observing others' actions. Specifically, experience apprehending objects initially increased infants' attention to similar reaching events performed by another person (Figure 2 in [2]), and increased more attention to events changing the action goal(Figure 3 in [2]).

We argue that when infants observe others' actions they make predictions of others based on the sensory information they perceive (Fig. 1-a). On the other hand, when infants produce actions they acquire own action experience through the process of integrating motor and sensory information. In our hypothesis infants use the joint representation from that sensorimotor integration to predict others' actions (Fig. 1-b). In other words, both action perception and action production share a common predictor, which we consider accounts for the influence that action production has on action perception. Further, we claim that the motor information contains a representation of the action goal [4]. Based on this argumentation, we hypothesize that the motor information alters the sensorimotor representation in terms of the goal, which allows infants to detect the goal in others' actions.

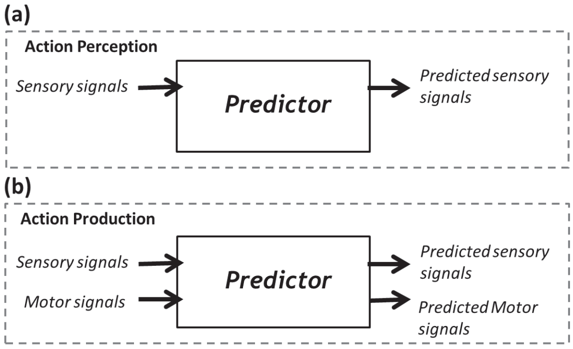

Fig.1 Our hypothesis. (a) During the action observation infants receive sensory information and predict sensory information. (b) During the action observation infants receive sensory and motor information and predict sensory and motor information.

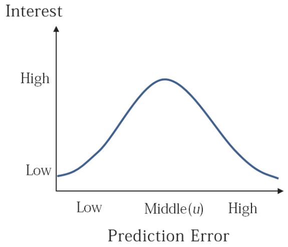

Fig.2 Visual attention. Curve of interest value in function of the prediction error.

During this learning process a prediction error arises between the predicted sensory information and the actual one [3]. The magnitude of the prediction error depends on the action experience. Here, we hypothesize that the prediction error modulates the level of attention to external stimulus through an interest function shown in Fig. 2 [5].

3. Computational Model

We propose a computational model based on our hypothesis which consists of four modules: the visual module, the motor module, the sensorimotor integration, and the visual attention module.

3.1 Motor Module

The motor module generates:

1. the motor primitives $P = [p_1, ..., p_m]$ represented as a vector of m binary signals whose components take values 0 or 1, where m is the number of action primitives,
2. and the target of the ongoing action $G = [g_1, ..., g_n]$ as a vector of n binary and mutually exclusive signals whose components take values 0 or 1, where n is the number of objects.

For the case of two objects (n=2) and two motor primitives (m=2): arm reaching primitive and arm retracting primitive, the motor module will output a vector $\mathbf{M}$ composed of four activation signals,

$$\mathbf{M}(t) = [g_1(t), g_2(t), p_1(t), p_2(t)], \qquad (1)$$

where t represents the time. The choice of variables is based on the idea that infants' actions are goal-directed (see goal babbling theory [6]).

3.2 Vision Module

Here, we introduce the term relations which refer to the relative dynamic between objects and the moving effector. For example the moving effector getting closer to (or getting away from) an object is considered a relation. The visual module receives an input image and outputs:

1. the position (x, y, z) of the moving effector,
2. the matrix $\mathbf{R} = [r_{11}, ..., r_{1m}; r_{21} ..., r_{2m}; ...; r_{n1} ..., r_{nm}]$ of $n \times m$ possible combinations between the moving effector and n objects for m relations, whose components take values 0 or 1,
3. and the vector $\mathbf{S} = [s_1, ..., s_m]$ of m possible relations between the moving effector and any object (e.g., the relation getting closer takes value 1 if the moving effector is getting closer to any object), whose components take values 0 or 1.

This choice is justified by the fact that infants can be expected to distinguish between objects and actors (see [1]), and therefore to be potentially able to recognize dynamic relations between them. Note that the vector $\mathbf{S}$ guarantees a differentiated representation of the dynamic of the moving effector regardless of the identity of the targeted object.

Thus, for the case of two objects (n=2) and two relations (m=2), getting closer and getting away, the vision module will output a vector $\mathbf{V}$ made of nine signals,

$$\begin{aligned}\mathbf{V(t)} = [&x(t), y(t), z(t), r_{11}(t), r_{12}(t),\\ &r_{21}(t), r_{22}(t), s_1(t), s_2(t)],\end{aligned} \qquad (2)$$

3.3 Sensorimotor Integration Module

3.3.1 Sensorimotor Integration

Here, we take advantage of the structure and functionality of the Elman Recurrent Neural Network (RNN) [7]. The inputs of the neural network $\mathbf{I}(t)$ are the outputs from the visual module and the motor module. We used 13 neurons in the input and output units, and 50 neurons in the hidden and context units, which was empirically decided as the minimum number of neurons for the network to converge.

$$\mathbf{I(t)} = [\mathbf{V(t)}, \mathbf{M(t)}], \qquad (3)$$

and the outputs $\mathbf{O(t)}$ are the predicted visual and motor data,

$$\mathbf{O(t+1)} = [\mathbf{V_p(t+1)}, \mathbf{M_p(t+1)}], \qquad (4)$$

where $\mathbf{V_p(t+1)}$ is the predicted visual information, and $\mathbf{M_p(t+1)}$ is the predicted motor information. The internal composition of $\mathbf{V_p(t+1)}$ and $\mathbf{M_p(t+1)}$ is equivalent to $\mathbf{V(t)}$ and $\mathbf{M(t)}$, respectively. The neural network is trained using the back propagation through time method to minimize the learning error of visual and motor data.

3.3.2 Prediction Error

The prediction error $u(t+1)$ when observing others performing an action is calculated as,

$$u(t+1) = |\mathbf{V_p(t+1)} - \mathbf{V(t+1)}|, \qquad (5)$$

where $\mathbf{V_p(t+1)}$ is the predicted sensory data, and $\mathbf{V(t+1)}$ is the actual sensory data.

3.4 Visual Attention Module

We adopted the findings of Kidd et al. [8] who suggested that infants allocate their attention in order to maintain an intermediate level of complexity. Here the complexity is represented by the prediction error. Accordingly, the visual attention is assumed to be proportional to an interest value q (Fig. 2). The interest value q is defined as follows,

$$q(t) = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot e^{\frac{-(u-w)^2}{2 \cdot \sigma^2}} \qquad (6)$$

where α is a scaling factor, σ is the variance and w is the intermediate value of the prediction error, respectively. The interest function is maximized when the prediction error is moderate, that is, when the observed action is not too predictable (i.e., prediction error is low) or not too unpredictable (i.e., prediction error is high).

4. Experiments

4.1 Experimental settings

We reproduced similar experimental settings to those described in [2]. Our experiment procedure is summarized in Fig. 3. We conducted experiments with the simulated version of the humanoid robot iCub. The experiments considered two scenarios: the watch-first and the reach-first condition. For each experiment, the robot was placed 40 centimeters away of two objects, separated from each other by

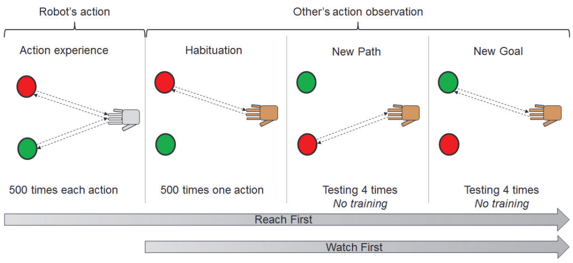

Fig.3 Procedure of the experiment

another 40 centimeters. In the watch-first scenario, the system first observed another individual reaching for one of the objects from the same location (perspective) as the robot, as in [1] [2]. This phase is called the visual habituation. Then, the position of the objects was swapped and the system observed two more actions: reaching for the other object (new goal event) and reaching for the same object (new path event). In the reach-first scenario, the same process was repeated, but this time the system previously experienced reaching for both objects in the action task, before the visual habituation. The three experiments (action task, habituation and new event) were repeated 20 times with random initialization of the weights of the neural network.

4.2 Action Task

During the action task, the robot's arm moved toward and touched one of two objects, then came back to the initial position and repeated the same action for the same or the other object, randomly. During the action task, the neural network was trained with vision and motor data for 500 reaching actions, each one composed of 175 steps (i.e., 87500 action steps). Fig. 4 (a) depicts the mean error of the action task over all training trials. The mean error u_m was calculated as the average of the prediction errors in a time window of size of 50 steps (chosen empirically) in order to attenuate the noise due to the dynamic of the reaching actions, which is not the target of our study.

4.3 Visual Habituation

The neural network was the same as the one trained during the action task for the reach-first condition. The motor inputs were fixed to 0 and the backpropagation was disabled for both the motor inputs and outputs so that the network does not unlearn the previously acquired motor prediction abilities (in the reach-first condition). During the habituation, the neural network was trained with only the vision for 500 reaching actions, each one composed of 175 steps. Fig. 4 shows the experimental results. Here, the maximum prediction error value was 0.365 (in watch-first condition), and the intermediate error w (Eq. 6) used for the interest function q was defined as half of the maximum prediction error. Hereafter the intermediate error stands as a reference value in our discussion regarding visual attention. The variance σ (Eq. 6) was arbitrarily defined to be 0.7 for illustration purposes since it does not alter the relation between high and low errors u and high and low interest q.

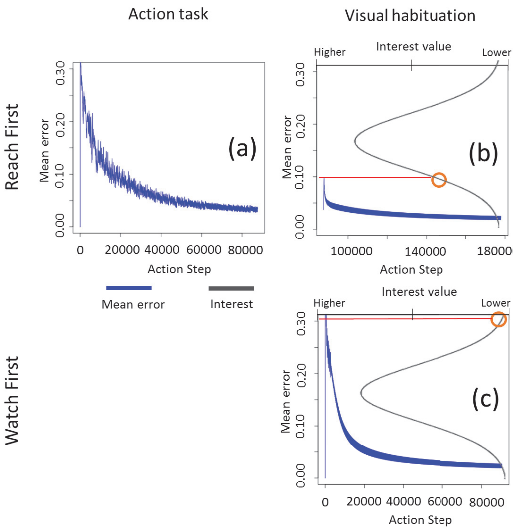

Fig.4 The bottom horizontal axis represents the action step, the vertical axis represents the mean error and the top horizontal axis represents the interest value. The blue line is the mean error in function of the action step and the gray line is the interest value in function of the mean error (see Eq. 6), respectively. The red line and the red point represent the intersection of the mean error with the curve of interest. (a) Action task for reach-first; (b) Habituation task for reach-first; (c) Habituation task for watch-first

We can observe from Fig. 4 (b) and (c) that the error for the reach-first condition was significantly lower than the error in the watch-first condition. Fig. 4 shows that the interest value (grey line) for the reach-first condition was higher than the interest value for the watch-first condition. This result suggests that own visuomotor experience contributes to make others' actions more predictable. Although, since othersf actions are not yet fully predictable, the prediction error that arises from observing others' actions causes a change in the visual attention.

4.4 New Path and New Goal

Here we measured the mean error when the goal or the trajectory were changed after the habituation, namely new goal and new path event, respectively. During the new path and new goal tests, the neural network was tested with the vision for 4 reaching actions for each test condition. The graphs of the mean error and the interest value for watch-first condition and reach-first condition are shown in Fig. 5.

We can see that the prediction error was higher for new goal event than for new path event for both

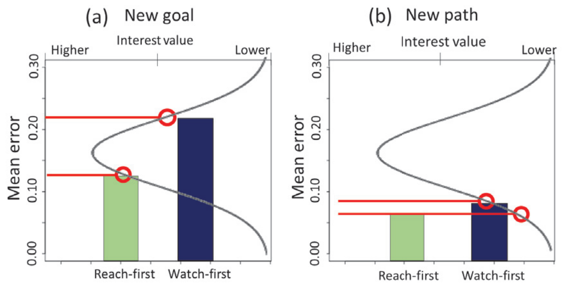

Fig.5 The bottom horizontal axis represents the condition, the vertical axis represents the mean error and the top horizontal axis represents the interest function. The green and blue bars represent the mean error for the reach-first condition and the watch-first condition, respectively. The gray line, whose independent axis is the top horizontal axis, represents the interest value in function of the mean error. The red line and point represent the intersection of the mean error with the curve of interest. (a) New goal event, (b) New path event.

the system with action experience (i.e., the reach-first condition) and the system without motor experience (i.e., the watch-first condition). We must notice that depending on the assignment of the visual signals the error could become higher for new goal and lower for new path condition, or vice versa. Thus, here we consider the difference of the error between new path and new goal as reference. Then, we can see in Fig. 5 that the difference in error between new goal and new path was higher for watch-first condition than for new goal condition. It means that in the reach-first condition the experience of motor signals and visual signals encoded in the visuomotor representation of the system was used to predict other's actions. Regarding the goal detection, the results revealed that due to the lack of visual experience of the system in the watch-first condition, it is not possible to make a comparison between reach-first and watch-first conditions, and therefore the results are not yet conclusive about whether the system detected the action goal.

5. Discussion

Our experimental results in terms of patterns of prediction error demonstrated a clear influence of the action experience on the perception. The predictor acquired through visuomotor experience of own actions was used to predict visual information of others' actions. Here we employed a Gaussian-shaped curve and the middle value of the prediction error to establish a relation between prediction error and visual attention, and the experimental results demonstrated to be in favor of our selection. Nonetheless, we consider that tuning those parameters, including σ (Eq. 6), requires additional evidence from future psychological studies.

We must say that our results are not conclusive regarding the detection of action goals. Our experiments reproduced similar experimental settings to those in [2], but the experiments indicated that it is necessary to distinguish between the influence of the motor and the visual information when comparing the reach-first and the watch-first conditions. Thus, further experiments must be carried out allowing the system in the watch-first condition to acquire the visual component of action experience in the first experiment (i.e., before the habituation phase). This setting will allow to make comparisons between reach-first and watch-first conditions, and therefore to measure the influence of the motor experience on goal detection.

6. Conclusion

We proposed a computational model to explain findings showing that action production alters the perception of other's actions in infants [2]. Our results demonstrated that the sensorimotor integration of own actions led to distinctive patterns of prediction error depending on own action experience that altered perception of others' actions

Acknowledgment

This work is partially supported by MEXT/JSPS KAKENHI (Research Project Numbers: 24119003, 24000012, 25700027) and JSPS Core-to-Core Program, A. Advanced Research Networks.

Bibliography

[1] A. L. Woodward, "Infants selectively encode the goal object of an actor's reach," *Cognition*, vol. 69, no. 1, pp. 1–34, 1998.

[2] J. A. Sommerville, A. L. Woodward, and A. Needham, "Action experience alters 3-month-old infants' perception of others' actions," *Cognition*, vol. 96, no. 1, pp. B1–B11, 2005.

[3] H. E. Den Ouden, P. Kok, and F. P. De Lange, "How prediction errors shape perception, attention, and motivation," *Frontiers in psychology*, vol. 3, 2012.

[4] G. Rizzolatti, L. Cattaneo, M. Fabbri-Destro, and S. Rozzi, "Cortical mechanisms underlying the organization of goal-directed actions and mirror neuron-based action understanding," *Physiological reviews*, vol. 94, no. 2, pp. 655–706, 2014.

[5] Y. Nagai, "A model of infant preference based on prediction error: How does motor development influence perception?," in *the Biennial Meeting of the Society for Research in Child Development*, March 2015.

[6] M. Rolf and J. J. Steil, "Goal babbling: a new concept for early sensorimotor exploration," *Osaka*, vol. 11, p. 2012, 2012.

[7] J. L. Elman, "Finding structure in time," *Cognitive science*, vol. 14, no. 2, pp. 179–211, 1990.

[8] C. Kidd, S. T. Piantadosi, and R. N. Aslin, "The goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex," *PloS one*, vol. 7, no. 5, p. e36399, 2012.

[9] E. N. Cannon and A. L. Woodward, "Infants generate goal-based action predictions," *Developmental science*, vol. 15, no. 2, pp. 292–298, 2012.