Proceedings of the 6th IEEE International Conference on Development and Learning and on Epigenetic Robotics September 19-22, 2016

Motor development facilitates the prediction of others' actions through sensorimotor predictive learning

Jorge Luis Copete, Yukie Nagai, Minoru Asada
Department of Adaptive Machine Systems, Graduate School of Engineering, Osaka University, Japan
Email: {jorge.copete,yukie,asada}@ams.eng.osaka-u.ac.jp

***Abstract*—Recent studies in psychology revealed that the emergence of infants' ability to predict others' action goals is correlated with the development of their motor ability to produce similar actions. In this regard, studies in neuroscience suggest that perception and production of actions share the same neural architecture (i.e., mirror neuron system). However, it is not yet clear what learning mechanisms are involved in the co-development of these abilities. Here we proposed a computational model to explain the development of prediction of others' action goals in synchronization with the development of action production. We adopted the concept of predictive learning of sensorimotor information as the key mechanism for it. Predictive learning intends to associate motor signals with sensory signals in a predictive manner during action executions. Thus, sensory signals perceived during action observations induce corresponding motor signals obtained in previous experiences. Our experimental results showed that our approach facilitated a robot to develop the ability to predict action goals in synchrony with the development of action production. Furthermore, our experiments demonstrated that the integration of goal-directed motor signals improved the accuracy to predict sensory signals and consequently action goals.**

I. Introduction

Humans engage in social relations which demand complex cognitive skills. One of those fundamental skills is the ability to understand intentions in the actions of other individuals. However, how humans acquire this ability remains an open question. Research studies dealing with this issue have introduced experimental paradigms to bring insight into the phenomena of action understanding [1]. Pioneering studies in psychology designed a paradigm of looking-time measures to investigate when and how infants start getting involved in goal-directed actions. The paradigm of looking-time measures consists of two steps: First, habituating infants to a certain action by presenting it repeatedly; and second, investigating whether infants' attention recovers in response to a change in the action either by a new goal or by a new means. A remarkable work conducted by Woodward [2] demonstrated that infants distinguish and process goals and means of human actions in a differentiated manner. Their experiments showed that, after habituation, infants exhibited a stronger novelty response to test events that varied the goal (e.g., the grasped object) than to events that varied the physical properties of the action (e.g., the motion path). Sommerville et al. [3] evaluated young infants' response to changes by employing a similar approach as in [2]. However, in contrast to [2], they divided infants into two groups. The infants in one group were allowed to experience object manipulation by themselves prior to visual habituation, while the infants in the other group were not. Their experiments indicated that only the infants with prior action experience exhibited differentiated responses to the new goal and new means events. Subsequent studies introduced a paradigm of anticipatory looking measures to investigate infants' ability to predict goal-directed actions. In that paradigm infants' eye movements are tracked in order to measure which object infants expect a person to grasp, and which path they expect a moving agent to take [1]. Kanakogi and Itakura [4] provided important evidence in relation to the developmental link between action prediction and action production. Their results demonstrated that infants' ability to predict reaching actions develops in synchrony with the development of their motor skills to produce similar actions.

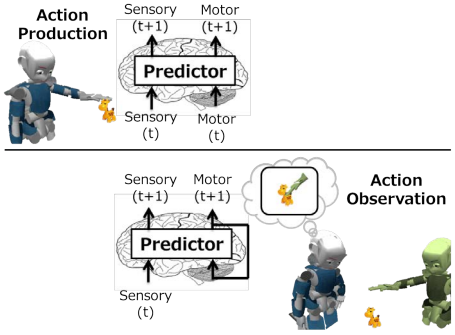


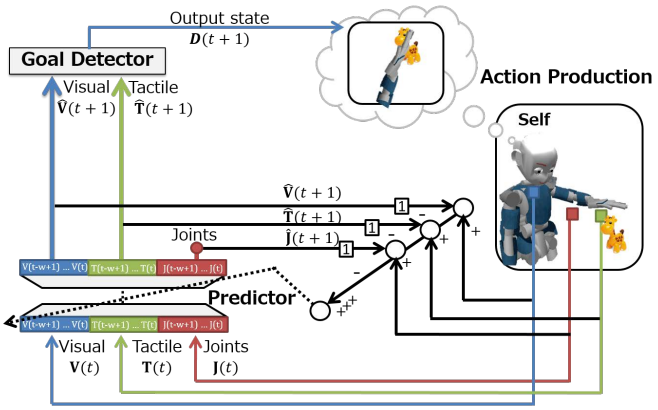
Fig. 1: Production of own actions and observation of others' actions mediated by sensorimotor predictive learning.

Studies in neuroscience have focused on a special class of brain cells, namely mirror neuron system (MNS), that were found to activate both when producing own actions and when observing similar actions executed by another individual. Studies by Pellegrino et al. [5] and Gallese et al. [6] showed that there exists a relation between observed actions that these brain cells respond to and motor signals they code. This indicates that perception and production of actions share the same neural architecture. In line with this, the MNS in monkeys has been reported to be involved in cognitive processes like action understanding [7] and intention understanding [8]. In

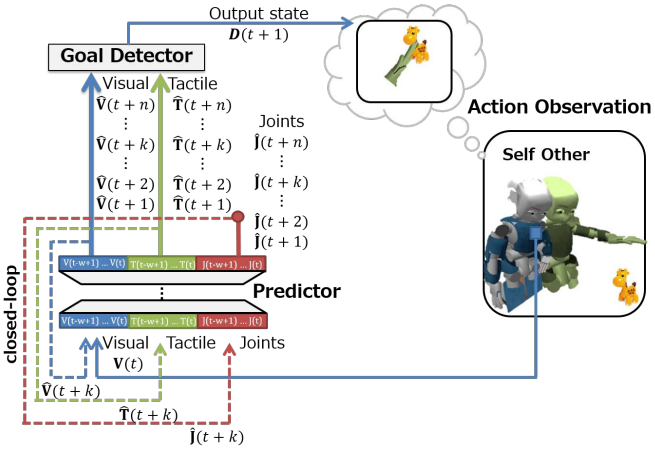
humans, the MNS was suggested to play a similar role in action understanding [7].

Numerous studies in computer science have focused on the concept of internal models to account for human motor abilities [9]. Internal models are neural mechanisms consisting of complex structures of forward and inverse models. The forward models predict sensory consequences associated with executed motor commands. The inverse models produce the motor commands required for achieving a desired sensory state. Accordingly, models with similar characteristics to the MNS have been proposed by coupling inverse and forward models [10] [11] [12]. These models have been adopted in robotics to study cognition [13]. Ogata et al. [14] and Demiris and Khadhouri [15] showed that reusing own forward-inverse models (i.e., own sensorimotor experience) when observing others was effective for imitation tasks. Baraglia et al. [16] proposed a model to explain how action production alters action perception. However, none of these studies have provided insight on the co-development of action prediction and action production as reported in [4].

We aim to model the development of prediction of others' action goals in synchronization with the development of action production. We adopt the concept of predictive learning of sensorimotor information as the key mechanism for the co-development. The concept of predictive learning [17] establishes that a predictor learns sensorimotor information when an agent produces own actions, and that the same predictor estimates sensorimotor information when the agent observes actions performed by others (see Fig. 1). A challenge here is that when observing others' actions (bottom), the agent perceives only sensory signals (e.g., vision) and the corresponding motor signals are missing. Of importance here is that the predictor can be employed to retrieve the missing signals by recalling the sensorimotor experience of own-produced actions. As a result, the experience of producing actions facilitates more accurate prediction of others' actions. We implement a computational model based on this idea and replicate the experiment by Kanakogi and Itakura [4]. Experiments are conducted to first analyze the co-development of prediction of others' action goals and action production, and then to assess the influence of integrating motor information on the development of goal prediction.



(a) Action Production



(b) Action Observation

Fig. 2: Computational model for action production and action prediction based on sensorimotor predictive learning. (a) A robot learns to reach for objects. The dotted line represents the error between the predicted data for $t+1$ and the actual data at $t+1$, which is fed back to the neural network. Through minimizing the prediction error, the system learns sensorimotor information that later can be used to predict others' actions. (b) The robot (self) observes others' reaching actions. The colored dotted lines represent the feedback of predicted signals used to predict future signals through imaginary simulations.

II. Key Idea and Computational Model

A. Scenario and key idea

We assume an experimental scenario using a robotic platform as shown in the right side of Fig. 2. The task for the robot is to acquire the ability to predict the goal of others' actions (e.g., reaching) while it learns to produce the same actions. In the phase of action production, the robot learns sensorimotor signals (e.g., visual $\mathbf{V}$, tactile $\mathbf{T}$ and motor $\mathbf{J}$) as shown in Fig. 2(a). In the phase of action observation (Fig. 2(b)), the robot makes predictions of others' actions based only on the perceived sensory signals (e.g., visual $\mathbf{V}$). A challenge here is the limited information that is perceived by the robot during the action observation, which makes the prediction difficult.

To tackle this problem, we adopt the concept of predictive learning of sensorimotor information [17]. Predictive learning intends to associate motor signals with sensory signals in a predictive manner. Thus, the robot learns to predict sensorimotor signals when executing actions. The important point here is that, even though the robot perceives only some sensory signals when observing actions, the predictive learning enables the robot to evoke the missing sensorimotor signals based on previous execution experiences. We hypothesize that evoking and integrating sensorimotor signals from own experiences facilitates better prediction of others' actions.

B. Computational model

We propose a computational model composed of two modules, the sensorimotor predictor and the goal detector (see the left side of Fig. 2). The predictor integrates and learns sensorimotor signals through a fully connected neural network which enables the different signals to interact with and influence each other (see Section II-B-1 for more details). We take advantage of the duality of the predictor, i.e., the predictor learns sensorimotor information of own actions during action production (e.g., visual **V**, tactile **T** and motor **J** in Fig. 2(a)), and the same predictor is recruited during action observation to predict sensorimotor information (Fig. 2(b)). We hypothesize that the missing modalities during action observation (e.g., tactile **T**, motor **J**) can be reconstructed through recalling the sensorimotor information learned during own action production. Then, the integration of the reconstructed signals in compensation for the missing ones will improve the accuracy of the sensorimotor predictions. In addition, the predictor also implements imaginary predictions for future time steps through a repeated feedback of the predicted sensorimotor signals in closed-loop manner (the dashed arrows in Fig. 2(b)). Finally, the goal detector detects the sensorimotor state which may correspond to the goal among the predicted sensorimotor states. We define that the goal corresponds to the sensorimotor state where abrupt changes in the tactile modality occurs. This is in accordance with studies in humans indicating that tactile inputs provide transition markers for manipulation [18].

There are two assumptions in our current approach. First, we assume that the motor development corresponds to the process of learning to predict sensorimotor signals of own actions. The sequence of signals to be learned are given through kinesthetic teaching (i.e., the robot does not perform exploratory learning). We then investigate how the ability of action prediction improves as learning advances. The second assumption concerns the perspective difference. Usually the vision observing the own actions and that observing others' actions look different. However, our current study focuses on the learning of own sensorimotor signals as a crucial factor for reconstructing missing signals during action observation. Therefore, we assume that the observer and the executer share the same visual perspective. This problem will be again discussed as a future issue.

1) Sensorimotor predictor: The predictor integrates and learns to predict visual, motor and tactile signals. Here, we adopt a deep autoencoder to implement the predictor. A deep autoencoder can be regarded as two deep neural networks stacked in a mirrored structure, in which the output layer reproduces the same values as the input layer. In the autoencoder, consecutive layers are fully connected regardless of the input modalities. Therefore, different modalities can interact with and influence each other. A reason for adopting a deep autoencoder is that it was shown to yield efficient performance for predicting sensorimotor signals [19]. Particularly, we take advantage of its ability to process high-dimensional data like images without explicitly preprocessing them. The function of the autoencoder is formulated as:

$$\mathbf{U} = F(\mathbf{S}), \tag{1}$$

$$\hat{\mathbf{S}} = F^{-1}(\mathbf{U}), \tag{2}$$

where $\mathbf{S}$, $\mathbf{U}$, and $\hat{\mathbf{S}}$ are the input vector, the vector at the central hidden layer, and the output vector, respectively. $F(.)$ represents the transformation mapping from the input layer to the central hidden layer, and $F^{-1}(.)$ represents the mapping from the central hidden layer to the output layer.

In action learning mode (Fig. 2(a)), the robot perceives visual signals $\mathbf{V}(t)$, tactile signals $\mathbf{T}(t)$, and joint angles $\mathbf{J}(t)$ at time t:

$$\mathbf{I}(t) = [\mathbf{V}(t), \mathbf{T}(t), \mathbf{J}(t)]. \tag{3}$$

Then, the input $\mathbf{S}$ to the autoencoder is designed as a contiguous segment (time window w) of the signals:

$$\mathbf{S}(t) = [\mathbf{I}(t-w+1), ..., \mathbf{I}(t-2), \mathbf{I}(t-1), \mathbf{I}(t)]. \tag{4}$$

The output $\hat{\mathbf{S}}(t)$ of the autoencoder is a vector with the same structure as $\mathbf{S}(t)$, where $\hat{\mathbf{V}}(t)$, $\hat{\mathbf{T}}(t)$ and $\hat{\mathbf{J}}(t)$ are the predicted signals for time t. The difference between the predicted signals $\hat{\mathbf{V}}(t)$, $\hat{\mathbf{T}}(t)$ and $\hat{\mathbf{J}}(t)$ and the actual signals $\mathbf{V}(t)$,$\mathbf{T}(t)$ and $\mathbf{J}(t)$ represents the prediction error. This error is fed back to the neural network in order to update the connecting weights of the autoencoder using back-propagation. We argue that learning to predict sensorimotor signals leads to the development of action production.

In the action observation mode, the robot has to predict the sensorimotor signals from $t+1$ to $t+n$, where n is a time ahead for prediction ($1 \leq k \leq n$ in Fig. 2(b)). When observing others' actions at time t, the only available inputs are the visual signals $\mathbf{V}(t)$. Of importance here is that the autoencoder predicts the missing tactile $\hat{\mathbf{T}}(t+1)$ and joint angle $\hat{\mathbf{J}}(t+1)$ signals for $t+1$ by recalling the sensorimotor information learned during own actions. Then, the predicted signals for $t+1$ are fed back to the input of the predictor to compensate for the missing ones at $t+1$. We consider that this process of reconstructing missing signals based on own action experience is crucial to become able to predict others' actions. Next, the robot predicts the sensorimotor signals from $t+2$ to $t+n$. In this case, the robot feeds back $\hat{\mathbf{V}}(t+k)$, $\hat{\mathbf{T}}(t+k)$ and $\hat{\mathbf{J}}(t+k)$ in closed-loop manner (the dashed arrows in Fig. 2(b)) at each iteration k to make the prediction of the signals from $t+2$ to $t+n$. Now, in relation to our implementation, if the robot perceives the signals $\mathbf{V}(t)$ at time t, then

$$\mathbf{I}(t) = [\mathbf{V}(t), \hat{\mathbf{T}}(t), \hat{\mathbf{J}}(t)], \tag{5}$$

where $\hat{\mathbf{T}}(t)$ and $\hat{\mathbf{J}}(t)$ are the reconstructed signals at the previous time step. Thus, the input $\mathbf{S}_1$ to the autoencoder is,

$$\mathbf{S}_1 = [\mathbf{I}(t-w+2), \mathbf{I}(t-w+3), ..., \mathbf{I}(t-1), \mathbf{I}(t), \mathbf{I}(t)]. \tag{6}$$

where the last slot of $\mathbf{S}_1$ is filled with a copy of the input signals $\mathbf{I}(t)$, as shown in Fig. 3. Once the prediction is done, the last slot of the output sequence $\hat{\mathbf{S}}_\mathbf{1}$ corresponds to the predicted signals $\hat{\mathbf{I}}(t+1)$. Next, the input sequence is shifted

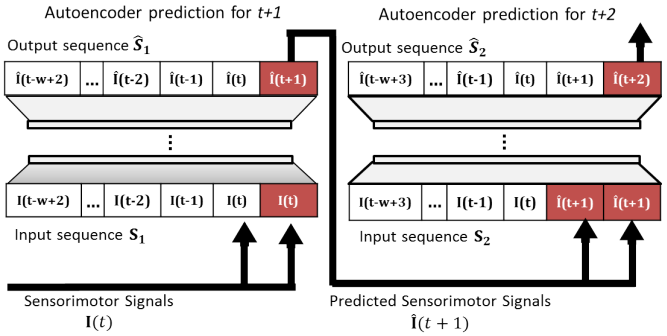


Fig. 3: Example of prediction ahead in time based on a closed-loop of predicted sensorimotor information.

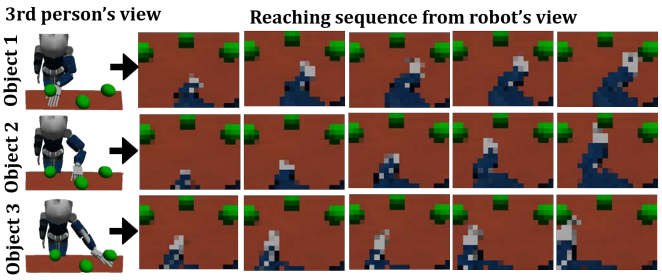


Fig. 4: Sequences of 20x15 RGB images showing the reaching trajectories towards the three objects.

and the last two slots of the input sequence $\mathbf{S}_2$ are again filled with the output of the last prediction, $\hat{\mathbf{I}}(t+1)$. The last slot of the output sequence $\hat{\mathbf{S}}_2$ is the predicted signals $\hat{\mathbf{I}}(t+2)$. The procedure in closed-loop manner is repeated n times until obtaining $\hat{\mathbf{I}}(t+n)$. This procedure generates the predicted sequence of future sensorimotor modalities.

2) Goal Detector: This module detects the goal from the predicted sequence of sensorimotor information. In our model the goal state $\mathrm{g}(t)$ is characterized by abrupt changes in the flow of tactile information. For detecting tactile changes, the module calculates the norm of the difference between predicted tactile signals at different time steps,

$$\Delta\mathrm{T}(t+k) = ||\hat{\mathbf{T}}(t+k) - \hat{\mathbf{T}}(t+k-d)|| \qquad (7)$$

for $d+1 \leq k \leq n$ and $d > 0$, where d is a constant value accounting for a time span. The goal state $\mathrm{g}(t)$ corresponds to the minimum value of k for which $\Delta\mathrm{T}(t+k) > h$ is satisfied, where h is a threshold for abrupt changes. This module outputs the visual information $\hat{\mathbf{V}}(t+\mathrm{g}(t))$ if $\mathrm{g}(t)$ exists.

III. Experimental Settings

Our study replicated the experiment in [4]. We employed the simulated version of the humanoid robot iCub. The tasks for the robot were reaching three objects during training and observing actions toward the same objects during testing. Fig. 4 shows examples of reaching actions to the three objects. We carried out two experiments to verify our hypothesis. The first experiment analyzed the co-development of action production and prediction of others' action goals. The second experiment assessed the influence of integrating motor information on the ability to predict goals. In this last experiment we contrasted two conditions: the condition where sensory modalities (i.e., visual, tactile) and motor information (i.e., joint angles) are integrated during action learning; and the condition where only sensory modalities are learned.

A. Implementation of the computational model

The input signals were 4 joint angles, **J**, of the left arm (shoulder yaw, shoulder pitch, shoulder roll, elbow); 3 binary tactile signals, **T**, with identical value; and one 30-dimensions visual vector **V**. The signals **V** and **J** were normalized in the range [0,1]. The 30-dimensions visual vector originated from a 320x240 RGB image from the iCub camera which was first resized to a 20x15 RGB image using OpenCV, and later compressed using an additional autoencoder from 900 dimensions (input vector **S** in Eq. 1) to 30 dimensions (central hidden layer vector **U** in Eq. 1). The purpose of reducing dimensionality is to compensate for the considerable difference in dimensionality between raw images and the other modalities. Once the prediction was done, the predicted visual vector $\hat{\mathbf{V}}$ was decompressed from 30 dimensions to 900 dimensions (output vector $\hat{\mathbf{S}}$ in Eq. 2). The iCub images were taken from a world-view camera located contiguous to the eyes so as to solve the narrowed panoramic view of the iCub's eyes.

The autoencoder for sensorimotor prediction (i.e., the predictor) had 12 hidden layers: 6 encoding hidden layers of 1000, 500, 250, 150, 80, and 30 neurons and 6 decoding hidden layers with 30, 80, 150, 250, 500 and 1000 neurons. The additional autoencoder for compressing images had the same structure. The activation functions were linear functions for the hidden layers and logistic functions for the output layer. We adapted implementations for deep neural networks based on Theano [20] [21]. The time window w for prediction was 30. The training was carried out using stochastic gradient descent by backpropagation. To alleviate CPU constraints, the system made predictions one time step ahead at every step and 20 steps ahead at every 10 steps. The threshold h for goal detection was 0.8. The time span d was 10.

B. Conditions for experimental analysis

The two experiments were repeated 13 times. The training/testing process of one experiment was divided into 15 stages. A contiguous pair of one training and one testing accounted for one stage (i.e., in total, 15 trainings and 15 testings). Each stage included 6 reaching action trials for learning/testing (two reaching trials for each object). One action trial took approximately 40 steps for reaching to the object, 40 steps for getting back to the home position and 25 steps for being static at home position. The prediction autoencoder (i.e., the predictor) was trained for 150 learning iterations at each training. The autoencoder for image reduction was fully trained for 1500 learning iterations since the first stage every

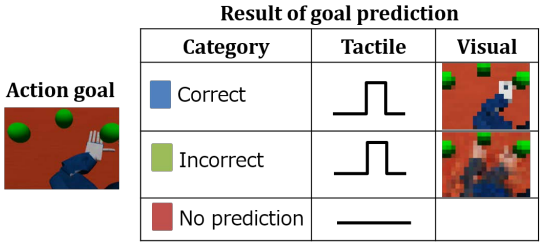


Fig. 5: Example of the three result categories

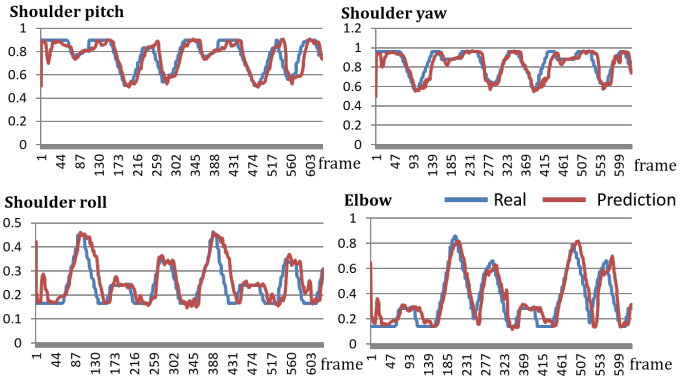


Fig. 6: Example of reconstruction of missing motor signals from visual signals

two stages. The weights of the autoencoders were randomly initialized at the first stage. For later stages the weights started with the weights as trained at the end of the previous stage.

The prediction results were classified into three categories: correct, incorrect and non-prediction. The purpose of this classification is assessing the goal prediction in terms of the sensorimotor information. Fig. 5 shows classification examples. The correct category indicates that the system predicted correctly the tactile and visual information of the goal state at the first attempt (i.e., first goal detection after moving from the home position). The incorrect category indicates that the system predicted a tactile change but the predicted visual information was not correct (i.e., average difference between the predicted and the correct images exceeded a threshold). The non-prediction category indicates that the system did not predict tactile changes before the hand touched the object.

IV. Experiment 1: Co-development of action production and action prediction

We analyzed the development of goal prediction in synchronization with the development of action production. The results of the experiment are shown in Figs. 6, 7 and 8. Fig. 6 shows examples of motor signals retrieved from visual signals. This result indicates that the system could reconstruct missing modalities based on the sensorimotor information learned during own action production.

Now we focus on the analysis of our main target, i.e., the goal prediction. The horizontal axis in Figs. 7 and 8 represents the learning stages (one stage accounts for one training and one testing). The blue, green and red bars in Fig. 7 represent the percentage (average and standard error) of correct prediction, incorrect prediction, and non-prediction, respectively. The black curve in Fig. 7 represents the mean learning error of the predictor. Fig. 7 shows that initially the action learning error was high and the prediction was dominated by the non-prediction category. However, while the learning error decreased through the learning stages (i.e., the motor ability develops), the correct and incorrect predictions increased significantly in correspondence with a decrease in the non-prediction. At the last stage, the learning error decreased significantly and the correct prediction reached an average of 58%. Fig. 8 represents how early in frames (i.e., anticipation time) the correct predictions were achieved. We examined the anticipation time to check whether the correct predictions were effectively achieved before the hand reached the objects. Fig. 8 shows that the anticipation time was around eight frames. The results showed that our model developed the ability to predict the action goal in synchrony with the development of action production.

V. Experiment 2: Influence of integrating motor information on goal prediction

The second experiment analyzed the influence of integrating motor information on the development of goal prediction. The purpose was to contrast our hypothesis according to which the integration of sensorimotor modalities improves the prediction accuracy. As we mentioned, visual, tactile and motor modalities interact with and influence each other in the hidden layers of the predictor. Thus, here our purpose is assessing, in terms of the goal prediction performance, the effect of not including motor signals during action learning.

We carried out the experiment of action learning without integrating goal-directed motor signals through the predictor (i.e., we set to zero the input vector corresponding to the motor signals). The bars in Fig. 9 represent the percentages of prediction results as in Fig. 7. The Fig. 10 represents the anticipation time as in Fig. 8. When comparing the graphs in Figs. 7 and 9, we can observe that the correct predictions were lower when goal-directed motor signals were not integrated. The average difference at the last stage was around 10% in favor of the system with integration of goal-directed signals. Also, in contrast to experiment 1, the experiment 2 showed that the prediction performance without integration of motor signals was less stable and had temporary increases (e.g., around stage 11 in Fig. 9). Nonetheless, further experiments under diverse settings are required in order to analyze this temporary increases. The graph in Fig. 10 shows that the anticipation time was around nine frames. We carried out a two-way ANOVA using two factors (1: presence of motor signal and 2: category) to compare the results between the two experiments at the last stage. The data was normally distributed for each category as determined by Chi-square test (P <0.05; $\chi^2 = 10.10 < 11.07 = \chi^2_{crit}$). The two-

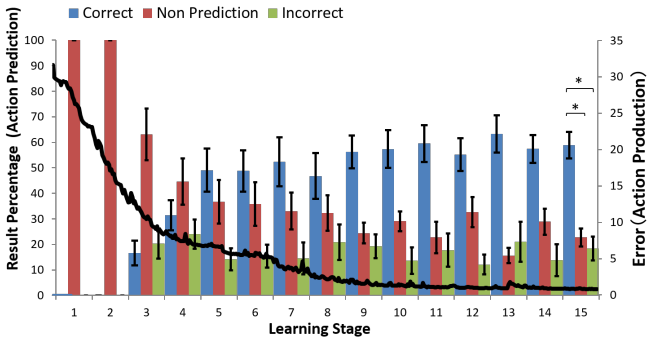


Fig. 7: Development of goal prediction in synchronization with development of action production when integrating motor, visual and tactile information.

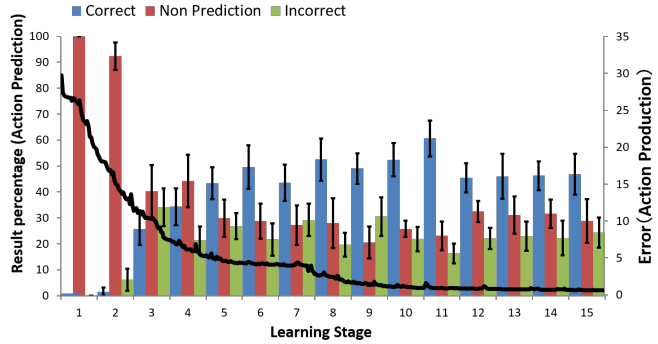


Fig. 9: Development of goal prediction in synchronization with development of action production in experiment integrating visual and tactile information.

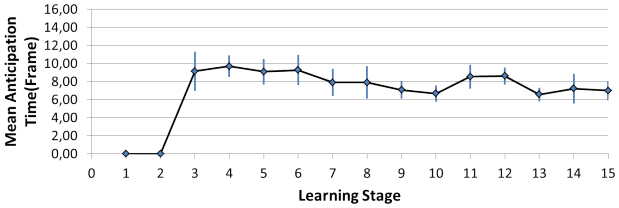


Fig. 8: Anticipation time when integrating motor, visual and tactile information.

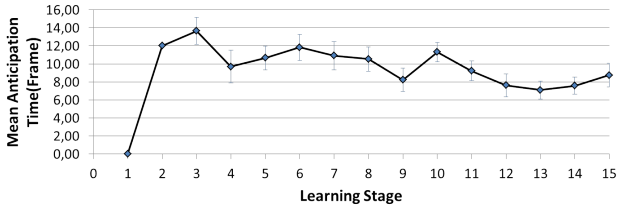


Fig. 10: Anticipation time in experiment integrating visual and tactile information.

way ANOVA indicated statistically no significant differences between the results for the presence of motor signals as determined by two-way ANOVA ($F(2,72) = 0.78$, $P > 0.05$). We carried out two separate one-way ANOVA to determine if there exists significant differences between the correct prediction, the non-prediction and the incorrect prediction categories within each experiment at the last stage. For the system without integration of motor signals, there were no statistically significant differences between group means ($F(2,36) = 2.609$, $P > 0.05$). However, for the system integrating goal-directed motor signals, there were statistically significant differences between group means as determined by one-way ANOVA ($F(2,36) = 23.374$, $P < 0.01$). We carried out post-hoc t-test to determine differences between the correct prediction and the non-prediction, and between the correct prediction and the incorrect prediction, for the system with motor integration. The t-test showed to be significant for both cases, ($t(24) = 6.18$, $p < .05$) and ($t(23) = 5.83$, $p < .05$), respectively. Though the difference between the two experimental results was not significant enough to be indicated by the two-way ANOVA, the one-way ANOVA and t-test confirmed that the presence of motor signals produced a significant increase of the correct predictions. Regarding the two-way ANOVA, we expect the difference becomes significant for conditions where there is more ambiguity or noise in the learning signals, as discussed in the next section. These results showed that integrating goal-directed motor signals improved the ability to make predictions.

VI. Discussion and Future Work

Our experiments demonstrated that the development of action production through predictive learning facilitates the ability to predict others' actions goals. The experiments assessing the integration of motor signals showed that action learning without integration of goal-directed motor signals affected negatively the ability to predict the action goal. This suggests that the integration of motor information during action production improves the accuracy of the sensorimotor prediction during action observation. We attribute this result to the fact that the integrated motor signals interact with other modalities in the hidden layers of the predictor. Thus, the motor signals help to guide the sensory prediction toward a set of sensory information learned during own action production. In relation to this, and in order to know how the motor signals interact with the sensory modalities, a future work to be addressed is analyzing the feature space in the hidden layers.

Regarding further roles for the motor signals, we expect that they become crucial to maintain predictive performance under, e.g., ambiguous or noisy conditions. In our current setting (conducted in a simulator) there is no significant ambiguity or noise. However, in a real setting, noises from external sources (e.g., visual signals) might become larger than those from internal sources (e.g., motor signals). Under this scenario we expect that the motor signals, which are less harmed by noise, will help to maintain the prediction accuracy.

Figs. 8 and 10 suggested that in the early stages the anticipation was earlier but less correct than in later stages in which the anticipation got slower but much more correct. One

possible reason for this pattern is that at early stages the closed loop (Fig. 3) using low accurate predictions produces larger sensorimotor changes between predicted states, and therefore the apparent goal state is reached earlier at the expense of low prediction accuracy (i.e., incorrect predictions). Further analysis is required to assess the trade-off between anticipation and accuracy.

Psychological findings pointed out that infants can predict actions of others even if they cannot produce them by themselves [1]. Accordingly, we plan to extend our model to explain this phenomenon. One of the limitation of our current model is that the goal detection relies on tactile information, which is not viable when only visual information is involved. This assumption is reasonable as a first step to verify our current hypothesis about the role of the motor signals. However, a future work is to propose a principle for goal detection that explains goals in terms of several sensorimotor modalities.

Another problem that must be addressed is modeling humans' ability to understand actions regardless the visual perspective. In terms of our model, the problem is to find the correspondence between the sensorimotor information learned during own actions and the one perceived when observing the same action executed by others (i.e., different visual perspective). Some methods have been proposed in robotics to account for the visual perspective [22], [23]. However, they assume explicit information about others' perspectives [22], or invariance of the sensorimotor information to spatio-temporal transformations so as to allow sensorimotor matching [23].

In informal tests we found that improvements are required in the autoencoder-based model for reaching to variable goal positions (i.e., learning generalization). Currently the model does not guarantee interpolation between raw images since each input pixel is treated as a different dimension, and therefore the network cannot learn pixel spatial relations.

VII. Conclusion

We introduced a computational model for the co-development of action prediction and action production. The results showed that our model based on the concept of predictive learning was effective to account for this co-development. Furthermore, our experiments demonstrated that the integration of goal-directed motor information improves the prediction accuracy. We consider that the integration of motor signals improves prediction by guiding the sensory prediction toward a set of sensory information learned during own action production. These results support our claim that predictive learning may explain the underlying mechanism for co-development of action prediction and action production.

Acknowledgment

This work is partially supported by MEXT/JSPS KAKENHI (Research Project Numbers: 24119003, 24000012, 25700027).

References

[1] S. Hunnius and H. Bekkering, "What are you doing? how active and observational experience shape infants' action understanding," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 369, no. 1644, p. 20130490, 2014.

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

[3] 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.

[4] Y. Kanakogi and S. Itakura, "Developmental correspondence between action prediction and motor ability in early infancy," *Nature communications*, vol. 2, p. 341, 2011.

[5] G. Di Pellegrino, L. Fadiga, L. Fogassi, V. Gallese, and G. Rizzolatti, "Understanding motor events: a neurophysiological study," *Experimental brain research*, vol. 91, no. 1, pp. 176–180, 1992.

[6] V. Gallese, L. Fadiga, L. Fogassi, and G. Rizzolatti, "Action recognition in the premotor cortex," *Brain*, vol. 119, no. 2, pp. 593–610, 1996.

[7] G. Rizzolatti and L. Craighero, "The mirror-neuron system," *Annu. Rev. Neurosci.*, vol. 27, pp. 169–192, 2004.

[8] L. Fogassi, P. F. Ferrari, B. Gesierich, S. Rozzi, F. Chersi, and G. Rizzolatti, "Parietal lobe: from action organization to intention understanding," *Science*, vol. 308, no. 5722, pp. 662–667, 2005.

[9] M. Kawato, "Internal models for motor control and trajectory planning," *Current opinion in neurobiology*, vol. 9, no. 6, pp. 718–727, 1999.

[10] M. Haruno, D. M. Wolpert, and M. Kawato, "Mosaic model for sensorimotor learning and control," *Neural computation*, vol. 13, no. 10, pp. 2201–2220, 2001.

[11] E. Oztop, D. Wolpert, and M. Kawato, "Mental state inference using visual control parameters," *Cognitive Brain Research*, vol. 22, no. 2, pp. 129–151, 2005.

[12] J. Tani and M. Ito, "Self-organization of behavioral primitives as multiple attractor dynamics: A robot experiment," *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 33, no. 4, pp. 481–488, 2003.

[13] M. Asada, K. F. MacDorman, H. Ishiguro, and Y. Kuniyoshi, "Cognitive developmental robotics as a new paradigm for the design of humanoid robots," *Robotics and Autonomous Systems*, vol. 37, no. 2, pp. 185–193, 2001.

[14] T. Ogata, R. Yokoya, J. Tani, K. Komatani, and H. G. Okuno, "Prediction and imitation of other's motions by reusing own forward-inverse model in robots," in *IEEE International Conference on Robotics and Automation, 2009. ICRA'09.*, pp. 4144–4149, IEEE, 2009.

[15] Y. Demiris and B. Khadhouri, "Hierarchical attentive multiple models for execution and recognition of actions," *Robotics and autonomous systems*, vol. 54, no. 5, pp. 361–369, 2006.

[16] J. Baraglia, J. L. Copete, Y. Nagai, and M. Asada, "Motor experience alters action perception through predictive learning of sensorimotor information," in *Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob), 2015*, pp. 63–69, IEEE, 2015.

[17] Y. Nagai and M. Asada, "Predictive learning of sensorimotor information as a key for cognitive development," in *Proc. of the IROS 2015 Workshop on Sensorimotor Contingencies for Robotics*, 2015.

[18] R. S. Johansson and J. R. Flanagan, "Coding and use of tactile signals from the fingertips in object manipulation tasks," *Nature Reviews Neuroscience*, vol. 10, no. 5, pp. 345–359, 2009.

[19] K. Noda, H. Arie, Y. Suga, and T. Ogata, "Multimodal integration learning of robot behavior using deep neural networks," *Robotics and Autonomous Systems*, vol. 62, no. 6, pp. 721–736, 2014.

[20] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio, "Theano: A cpu and gpu math compiler in python," in *Proc. 9th Python in Science Conf*, pp. 1–7, 2010.

[21] O. Chapelle and D. Erhan, "Improved preconditioner for hessian free optimization," in *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

[22] R. Nakajo, S. Murata, H. Arie, and T. Ogata, "Acquisition of viewpoint representation in imitative learning from own sensory-motor experiences," in *2015 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics*, pp. 326–331, IEEE, 2015.

[23] F. Schrodt, G. Layher, H. Neumann, and M. V. Butz, "Embodied learning of a generative neural model for biological motion perception and inference," *Frontiers in computational neuroscience*, vol. 9, 2015.