4H1-2

Modeling Development of Action Recognition in Synchrony with Development of Action Production

Jorge Luis Copete*1 Yukie Nagai*1 Minoru Asada*1

*1 Graduate School of Engineering, Osaka University

The human ability to recognize actions executed by other individuals was found to develop in synchrony with the development of action production. We argue that the concept of predictive learning of sensorimotor integration explains the co-development of action recognition and action production. We proposed a computational model of development of prediction of goal based on predictive learning. Our results showed that our model was able to develop the ability to predict the action goal through the development of action production. Furthermore, we demonstrated that the integration of goal-directed motor signals improves the prediction performance.

1. Introduction

Explaining how humans recognize others' actions is a challenging question to address. Understanding its underlying mechanism may also bring novel perspectives for development of robot cognition. Psychological studies found that the emergence of infants' ability to predict action goals are correlated with the development of their motor skills to produce similar actions [Kanakogi 11, Sommerville 05]. These findings are in line with studies in neuroscience suggesting that observation and execution of actions share the same neural architecture (i.e., mirror neuron system) [Rizzolatti 04]. Computer scientists have attempted to propose models with similar characteristics to the mirror neuron system. Haruno et al. [Haruno 01] proposed MOSAIC model. Oztop et al. [Oztop 05] proposed the Mental State Inference (MSI) model to account for how mental state inference can be achieved with one's own motor system. Tani and Ito [Tani 03] proposed a type of recurrent neural network with Parametric Bias (PB). Similar computational models have been adopted for problems on action production and recognition in robots like imitation [Ogata 09] and object manipulation [Noda 13]. The study by Baraglia et al. [Baraglia 15] showed that action production alters action perception. Copete et al. [Copete 14] proposed a model for development of the ability to predict others' goal in terms of visual information. However, there are no computational studies to explain the development of the ability to predict others' goal in synchrony with development of action production. We proposed a computational model of the co-development of action prediction and action production and carried out experiments for reaching actions. We adopted the concept of predictive learning of sensorimotor information as the key mechanism that accounts for the co-development of action recognition and action production. The concept of predictive learning of sensorimotor information [Nagai 15] establishes that the predictor that is recruited for predicting actions performed by others is the same predictor that learns sensorimotor information associated to own action production. Therefore, the ability to predict others' actions relies on the experience of own action production.

Contact: Jorge Copete, Graduate School of Engineering, Osaka University, jorge.copete@ams.eng.osaka-u.ac.jp

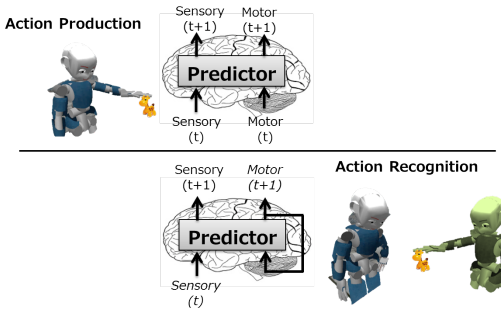
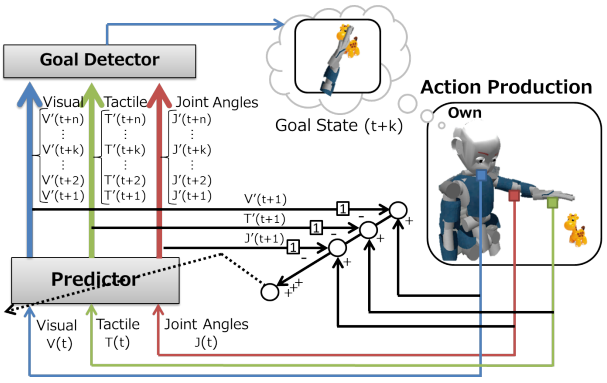


Figure 1: Scenario of reaching actions for action production and action recognition

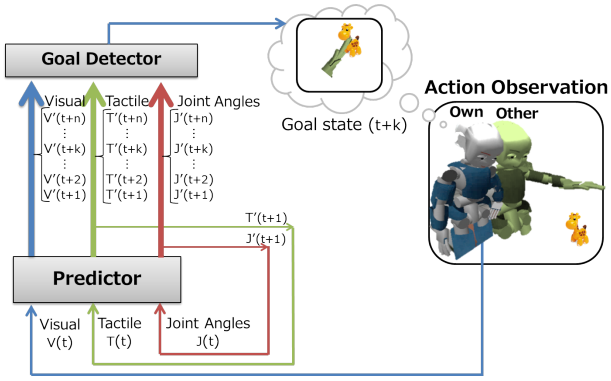
2. Our Proposed Approach

We hypothesize that the sensorimotor information that humans learn during own action production helps to recognize actions when observing others. We propose to employ the concept of predictive learning of sensorimotor information to model this hypothesis [Nagai 15]. According to this concept, sensorimotor information (e.g., visual, tactile, motor) are learned through a predictor during the development of own action production, as depicted in Fig. 1. Then, when observing others' actions, even though the sensory information that can be perceived is limited (e.g., visual), the predictor can reconstruct missing sensorimotor information associated to those actions based on the sensorimotor information learned during own action production. As a consequence, this information will help to recognize others actions. Another important aspect for action recognition is the ability to identify the goal of those actions. In this regard, we proposed that changes in the flow of tactile information can be employed to identify the goal of reaching actions.

We proposed a computational model for action recognition and action production as shown in Fig. 2. The model



(a) Action Production



(b) Action Observation

Figure 2: Computational model for action production and action recognition based on predictive learning of sensorimotor information.

is composed of two modules: predictor and goal detector.

2.1 Sensorimotor Predictor

This module is in charge of integrating and learning to predict sensorimotor information. We employed deep autoencoders based on previous work by Noda et al. [Noda 13] reporting their advantages for predicting temporal data sequences. A deep autoencoder learns to reproduce the same values of the input layer in the output layer. The main advantage of the autoencoder is that after training the network can be connected in a closed-loop manner to predict future states of a temporal sequence of data. The inputs are visual signals $\mathbf{V}(t)$, tactile signals $\mathbf{T}(t)$, and joint angles $\mathbf{J}(t)$. A sequence of input data from the last w steps (i.e., time window) are input to the deep autoencoder. The deep autoencoder predicts sensorimotor signals from time $t+1$ to $t+n$ through a closed-loop feedback, where n is the time ahead for prediction. When the system produces actions, the predictor learn by comparing the predicted sensorimotor signals at $t+1$ with the real signals at $t+1$, and feeding back the error to the neural network (Fig. 2(a)). When the system observes others' actions, only visual signals are available but the deep autoencoder predicts tactile signals and joint angles which are fed back in a closed-loop manner (Fig. 2(b)). The module output are $[\mathbf{V'}(t), \mathbf{J'}(t), \mathbf{T'}(t), ..., \mathbf{V'}(t+n-1), \mathbf{J'}(t+n-1), \mathbf{T'}(t+n-1)]$.

2.2 Goal Detector

This module receives sensorimotor information from the predictor, detects the goal state and outputs the sensorimotor information corresponding to the goal state. The goal of the action corresponds to the state at which significant changes in sensory information occur. Tactile signals were employed for goal detection in reaching actions in our current implementation. For detecting significant tactile changes, we calculate the norm of the vectorial difference between the predicted values of tactile signal at time $t+k-d$ and time $t+k$,

$$c(t) = ||\hat{\mathbf{T}}(t+k) - \hat{\mathbf{T}}(t+k-d)|| \quad (1)$$

Then, the action goal G corresponds to

$$\mathbf{G} = \begin{cases} \mathbf{V'}(t+k) & c(t)>h \\ \mathbf{V'}(t+1) & (\text{otherwise}) \end{cases} \quad (2)$$

where k ($k \leq n$) represents the time ahead in frames of prediction, h is a threshold to account for abrupt changes, and d is a constant value to account for a span of time within which significant changes can be detected. The module outputs the sensorimotor information corresponding to the goal state.

3. Experimental Settings

We employed the simulated version of the humanoid robot iCub. The system receives and processes visual, motor and tactile signals from the robot. The input signals to the predictor are a 20x15 RGB image; 4 joint angles of the left arm (shoulder yaw, shoulder pitch, shoulder roll, elbow); and 3 binary tactiel signals with identical value. The time window w of the predictor is 30 steps and the threshold h of tactile change for goal detection was 0.8. The prediction module is composed of two deep autoencoders. One autoencoder makes dimension reduction of input images, and the second autoencoder makes sensorimotor prediction. Each autoencoder has 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 activation functions are linear functions for the hidden layers and logistic functions for the output layer. We adapted available implementations for Deep Neural Networks based on Theano [Chapelle 11]. The autoencoder for dimension reduction transforms the dimensionality of the input images from 900 to 30 dimensions (i.e., encoding), or vice versa (i.e., decoding). The autoencoder for sensorimotor prediction learns the encoded visual signals, the motor signals, the tactile signals. The weights of both autoencoders were initialized to random values and the training process was carried out using a stochastic gradient descent algorithm.

The experimental task consisted in the robot reaching for three objects during training (i.e., moving the left hand forward and back continuously) and observing the arm reaching for objects during testing (i.e., only visual signals). The experiments included two conditions for action learning:

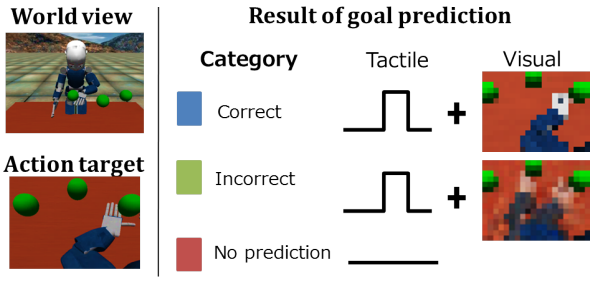


Figure 3: Example of the three categories of prediction

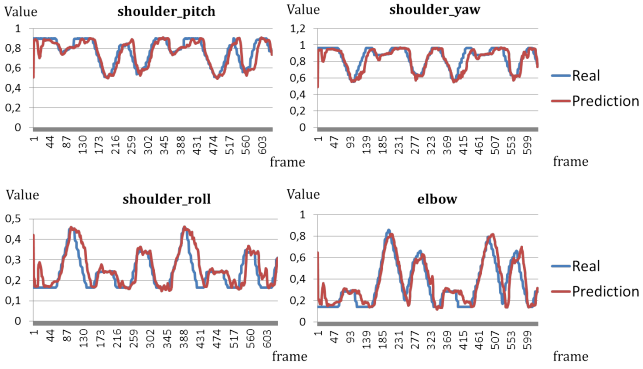


Figure 4: Example of reconstruction of motor signals from visual signals

sensory information and motor information are integrated (i.e., visual, tactile and joint angles); motor information is not integrated (i.e., visual and tactile). The problem of visual perspective difference is out of our current target and will be discussed as a future issue. For one session of learning we ran 15 trainings and 15 testings alternately, so each training and testing accounted for one developmental stage. The prediction results were classified into three categories: correct, incorrect and non-prediction, in order to assess the development of the goal prediction ability. Fig. 3 shows examples of classification of prediction results. The correct category indicates that the system predicted the correct visual and tactile information corresponding to the target object before the robot hand touched the object. The incorrect category indicates that the predicted visual information did not correspond to the correct one. The non-prediction category indicates that the system did not predict tactile changes before touching the object.

4. Experiment 1

The first experiment analyzed the developmental process of goal prediction in synchronization with the development of action production. The experimental results are shown in Figs. 4, 5 and 6. Fig. 4 shows examples of motor signals retrieved from visual signals which demonstrate the ability of the system to reconstruct missing signals. In Fig. 5 and 6, the horizontal axis represents the learning stages, where each stage accounts for a set of one training phase and one testing phase. The left vertical axis of Fig. 5 represents the rate of each result category. The right vertical axis of Fig.

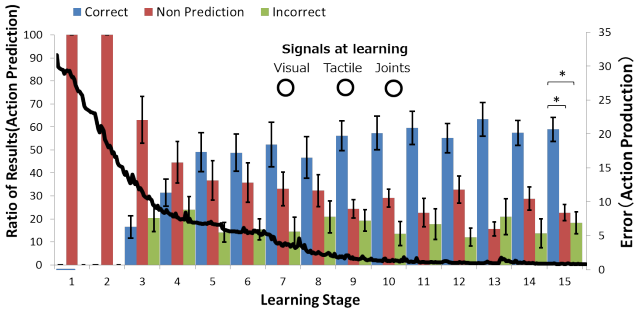


Figure 5: Development of goal prediction in synchronization with development of action production with integration of motor, visual and tactile information.

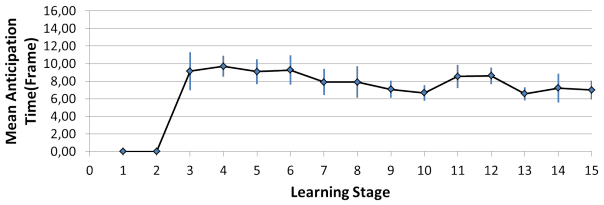


Figure 6: Anticipation time in experiment 1 with integration of motor, visual and tactile information.

5 represents the output error of the predictor. Both graphs show average values of prediction and their standard error. The vertical axis of Fig. 6 shows how early the correct predictions was, which help to examine if the system acquired the ability to predict the goal before the arm reaches to the object.

Fig. 5 shows that learning error for action production was initially high but decreased rapidly for the first three stages. Later, the error decreased until reaching its lower value after stage twelve. In parallel, the ratio of goal prediction in the initial four stages was dominated by the non-prediction category. However, since the third stage the correct and incorrect predictions showed a significant increase with a correspondent decrease in the non-prediction. Later, the correct prediction continued to increase until reaching an average of 58% at stage fifteen. Fig. 6 shows the average prediction time was around eight frames earlier before reaching the object. The experiment demonstrated that our computational model was able to develop the ability to predict the action goal through development of action production.

5. Experiment 2

The second experiment analyzed the influence that motor information have on the development of goal prediction. For that purpose, to contrast our hypothesis, we tested the condition in which motor information is not integrated with the sensory information during development of action production and action recognition. Therefore, we carried out an experiment of action learning without integrating motor signals (i.e., we set to zero the input layer corresponding to motor signals and integrated only visual and tactile signals).

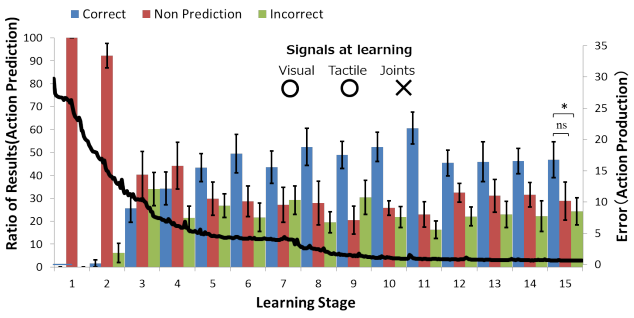


Figure 7: Development of goal prediction in synchronization with development of action production with integration of visual and tactile information.

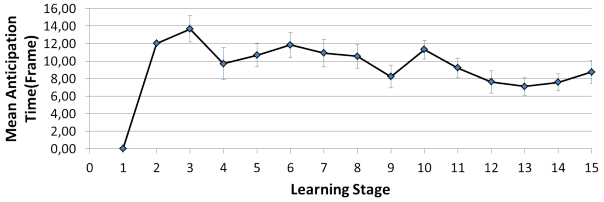


Figure 8: Anticipation time in experiment 2 with integration of visual and tactile information.

Fig. 7 represents the rate of each result category. Comparing graphs in Fig. 7 and Fig. 5 we can observe the correct predictions of the system without integration of motor information was lower, accounting for a difference of around 10% in the last stages in favor for the system with motor integration. We carried out two t-test in the last developmental stage to determine differences between the correct prediction category and the non-prediction category, and between the correct prediction category and the incorrect prediction category. 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. For the system without motor integration, the t-test showed no significant difference between the correct prediction and the non-prediction ($t(24) = 1.59$, $p < .05$), and significant difference with the incorrect prediction ($t(23) = 2.37$, $p < .05$). These results showed that the experience of goal-directed motor signals (in contrast to the non-goal-directed zero signals) improved the ability to make correct predictions, which could be due to the learned association between goal-directed motor signals and visual signals which make the prediction of sensory information more accurate.

6. Conclusions and Future Work

The experimental results showed that our model was able to develop the ability to predict the action goal through the development of action production. Our analysis demonstrated that the integration of goal-directed motor signals across the development improves the prediction performance through the association between goal-directed motor signals and visual signals which helps to make the sensory prediction more accurate. In addition, the role attributed to the tactile signals for goal detection was effective for reaching tasks. A next problem to address is explaining the mechanism and conditions that allow recognizing others actions regardless visual perspective. Additional future work includes studying the influence that action recognition has on action production and generalizing the mechanism for goal detection.

References

[Kanakogi 11] Kanakogi, Yasuhiro, and Shoji Itakura. "Developmental correspondence between action prediction and motor ability in early infancy." Nature communications 2 (2011): 341.,

[Rizzolatti 04] Rizzolatti, G., and Craighero, L. The mirrorneuron system. Rev. Neuroscience (2004)

[Haruno 01] Haruno, Masahiko, Daniel M. Wolpert, and Mitsuo Kawato. "Mosaic model for sensorimotor learning and control." Neural computation 13.10 (2001)

[Sommerville 05] Sommerville, Jessica A., Amanda L. Woodward, and Amy Needham. "Action experience alters 3-month-old infants' perception of others' actions." Cognition 96.1 (2005): B1-B11.

[Oztop 05] Oztop, Erhan, Daniel Wolpert, and Mitsuo Kawato. "Mental state inference using visual control parameters." Cognitive Brain Research 22.2 (2005)

[Tani 03] Tani, Jun, and Masato Ito. "Self-organization of behavioral primitives as multiple attractor dynamics: A robot experiment." Systems, Man and Cybernetics (2003),

[Ogata 09] Ogata, Tetsuya, et al. "Prediction and imitation of other's motions by reusing own forward-inverse model in robots." ICRA'09, 2009.,

[Copete 14] Copete, Jorge Luis, Yukie Nagai, and Minoru Asada. "Development of goal-directed gaze shift based on predictive learning." ICDL-Epirob, 2014

[Baraglia 15] Baraglia, Jimmy, et al. "Motor experience alters action perception through predictive learning of sensorimotor information." Development and Learning and Epigenetic Robotics (ICDL-EpiRob), 2015 Joint IEEE International Conference on. IEEE, 2015.

[Noda 13] K. Noda, H. Arie, Y. Suga, and T. Ogata. Multimodal integration learning of object manipulation behaviors using deep neural networks. (IROS), 2013

[Nagai 15] Nagai, Yukie, and Minoru Asada. "Predictive Learning of Sensorimotor Information as a Key for Cognitive Development." Proc. of the IROS 2015 Workshop on Sensorimotor Contingencies for Robotics. 2015.

[Chapelle 11] Chapelle, Olivier, and Dumitru Erhan. "Improved preconditioner for hessian free optimization." NIPS Workshop on Deep Learning and Unsupervised Feature Learning. Vol. 201. No. 1. 2011.