Visuomotor Associative Learning under the Predictive Coding Framework: a Neuro-robotics Experiment

Jungsik Hwang (P)[1,2], and Jun Tani[2]

1 Korea Advanced Institute of Science and Technology, Republic of Korea
2 Cognitive Neurorobotics Unit, Okinawa Institute of Science and Technology, Japan
E-mail: {jungsik.hwang, tani1216jp}@gmail.com

Abstract—This study aims to introduce our approach to build cognitive agent based on predictive coding of visual and proprioceptive signals. We assume that a robot can develop cognitive skills by learning sensorimotor experience end-to-end in a hierarchical neural network model. The results from the neuro-robotics experiment illustrated the role of visuomotor learning in achieving cognitive behaviors and highlighted the importance of prediction error minimization, supporting predictive coding account of mirror neuron system.

***Keywords*—Predictive Coding, Visuomotor Learning, Cognitive Neuro-robotics**

1. Introduction

An ability to predict the incoming sensory stimulus is one of the important features in the predictive coding framework [1]. In this paper, we introduce our approach to build a cognitive agent based on predictive coding of visual and proprioceptive signals. We assume that the robot can build a predictive internal model of the world by learning sensorimotor experience end-to-end. Particularly, we investigate the role of minimizing prediction error (PE) in inferring intention latent in observation as well as in recalling visuomotor representation obtained from consolidative learning of sensorimotor experience. We argue that visuomotor learning under the predictive coding framework can be built on a hierarchical neural network model for learning multimodal sensorimotor experience as well as a prediction error minimization mechanism.

2. Dynamic Neural Network Model

The neural network model discussed in this paper is called Predictive Visuomotor Deep Dynamic Neural Network (P-VMDNN) introduced in [2]. The model is an extension of the earlier version [3] which can associate visual perception and action generation. In [2], the model was extended under the predictive coding framework [1] so that the model could generate visual and proprioceptive predictions simultaneously.

The model (Figure 1) consists of two pathways for processing different modalities – vision and proprioception. The visual pathway consists of a set of layers maintaining the spatial and temporal features and it generates the visual prediction in the pixel-level images. The proprioceptive pathway consists of a set of multiple timescales recurrent neural network [4] layers and it generate proprioceptive predictions specified as the robot's joint position values.

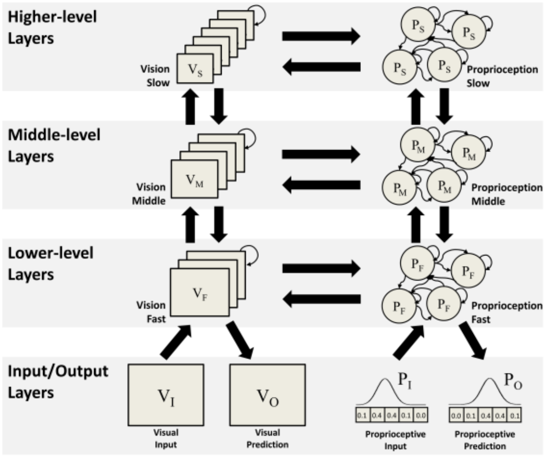

Figure 1. The predictive visuomotor deep dynamic neural network (P-VMDNN) model introduced in [2]. It consists of the visual pathway (left) and the proprioceptive pathway (right), and those pathways are connected through the lateral connections (horizontal arrows).

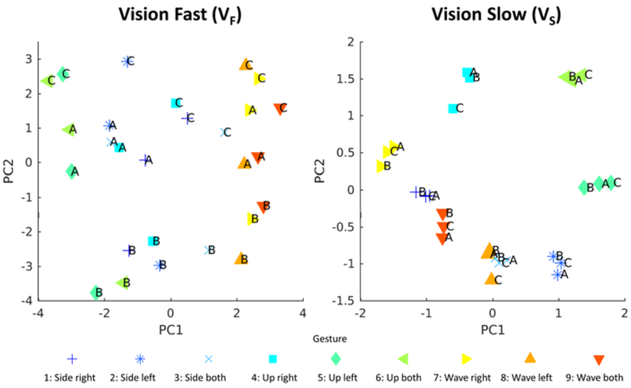

Figure 2. Principal component analysis on the initial states of the neurons in the Vision Fast (V_F) and Vision Slow (V_S) layers in [2]. The colors represent the type of the gestures and the letters indicate the human subject index.

There are several key features of the model. For instance, the model is equipped with the connective pathways that tightly couple vision and proprioception. Consequently, the model can learn visuo-proprioceptive patterns in an end-to-end manner without separate processing of each modality. In addition, the model has a spatio-temporal hierarchy that plays a crucial role in achieving functional hierarchy [4]. For instance, the analysis on neural activation showed that the higher-level layers in the visual pathway were encoding abstract information such as the type of the gesture whereas the lower-level layers were

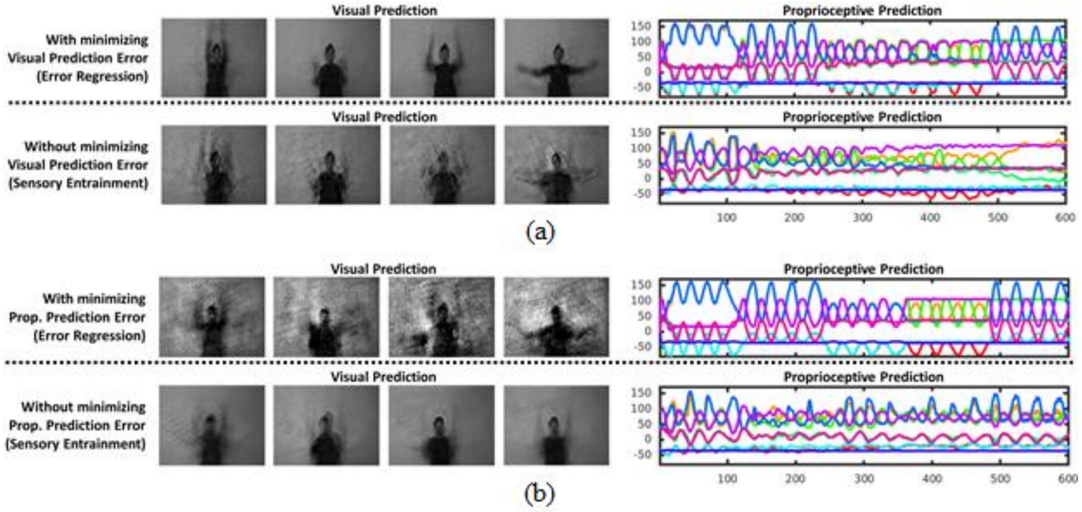

Figure 3. Experiment results from [2]. (a) Minimizing visual prediction error and (b) minimizing proprioceptive prediction error.

encoding specific information such as the appearance of the human subjects (Figure 2). Please see [2] for more details about the model.

3. A Neuro-robotic Experiment

The robot was trained to imitate the gesture of the human subjects displayed on the screen in the simulation environment. Nine types of the gestures performed by three human subjects were used during the training stage.

During testing, either visual or proprioceptive observation was presented to the model. Then, the model minimized the prediction error (PE) of the given sensory modality and generated the prediction of another modality simultaneously. During PE minimization, the model iteratively updates its internal states in the direction of minimizing discrepancy between prediction and observation. In this sense, perception in this model is an active cognitive process in which robot's prediction is consistently updated to match the actual observation.

3.1. Minimizing Visual Prediction Error

The result (Figure 3 (a)) showed that the model was able to minimize visual PE so that it could generate correct visual prediction about the gesture of the human subject. At the same time, the model was also able to generate corresponding proprioceptive prediction, resulting in successful imitation. The analysis on neural activation revealed that this capability could be achieved by inferring intention as well as recalling the corresponding visuomotor representation during PE minimization. Without minimizing visual PE, however, the model was not able to generate neither visual nor proprioceptive predictions. Consequently, the robot was not able to imitate the gesture of the human subjects.

3.2. Minimizing Proprioceptive Prediction Error

The model was able to minimize proprioceptive PE successfully (Figure 3 (b)). In addition, the model was able to generate corresponding visual prediction simultaneously. Although the visual prediction was a bit blurry, the type of the gesture was still identifiable. One of the reasons of blurry visual prediction might be due to that proprioceptive observation did not contain any information about the appearance of the human subject, but only the type of the gesture. Consequently, the model generated the blurry visual prediction in which the type of the gesture can be identified, but the appearance of the human subject cannot be precisely generated.

4. Conclusion

In this paper, we illustrate our approach to build a cognitive robot under the predictive coding framework [1]. The robot was able to learn high-dimensional sensorimotor experience using a dynamic neural network model in which action was hierarchically represented. It was also shown that the robot was able to adapt to a dynamic environment by minimizing prediction errors. The experimental results revealed that minimizing prediction error enabled the model to infer intention latent in observation as well as to recall the visuomotor representation acquired during training. In turn, these findings support the predictive coding account of mirror neuron system (MNS) [1].

References

[1] J. M. Kilner, K. J. Friston, and C. D. Frith, "Predictive coding: an account of the mirror neuron system," *Cogn. Process.*, vol. 8, no. 3, pp. 159-166, 2007.

[2] J. Hwang, J. Kim, A. Ahmadi, M. Choi, and J. Tani, "Dealing With Large-Scale Spatio-Temporal Patterns in Imitative Interaction Between a Robot and a Human by Using the Predictive Coding Framework," *IEEE Trans. Syst., Man, Cybern., Syst*, pp. 1-14, 2018.

[3] J. Hwang and J. Tani, "Seamless Integration and Coordination of Cognitive Skills in Humanoid Robots: A Deep Learning Approach," *IEEE Trans. Cogn. Develop. Syst.*, vol. PP, no. 99, 2017.

[4] Y. Yamashita and J. Tani, "Emergence of functional hierarchy in a multiple timescale neural network model: a humanoid robot experiment," *PLoS Computational Biology*, vol. 4, no. 11, p. e1000220, 2008.