Dealing with Large-Scale Spatio-Temporal Patterns in Imitative Interaction between a Robot and a Human by Using the Predictive Coding Framework

Jungsik Hwang, Jinhyung Kim, Ahmadreza Ahmadi, Minkyu Choi and Jun Tani

***Abstract*—This study aims to investigate how adequate cognitive functions for recognizing, predicting and generating a variety of actions can be developed through iterative learning of action-caused dynamic perceptual patterns. Particularly, we examined the capabilities of mental simulation of one's own actions as well as the inference of others' intention because they play a crucial role, especially in social cognition. We propose a dynamic neural network model based on predictive coding which can generate and recognize dynamic visuo-proprioceptive patterns. The proposed model was examined by conducting a set of robotic simulation experiments in which a robot was trained to imitate visually perceived gesture patterns of human subjects in a simulation environment. The experimental results showed that the proposed model was able to develop a predictive model of imitative interaction through iterative learning of large-scale spatio-temporal patterns in visuo-proprioceptive input streams. Also, the experiment verified that the model was able to generate mental imagery of dynamic visuo-proprioceptive patterns without feeding the external inputs. Furthermore, the model was able to recognize the intention of others by minimizing prediction error in the observations of the others' action patterns in an online manner. These findings suggest that the error minimization principle in predictive coding could provide a primal account for the mirror neuron functions for generating actions as well as recognizing those generated by others in a social cognitive context.**

***Index Terms*— Cognitive robotics, dynamic neural network, predictive coding, social cognition, cognitive system architectures and implementations.**

I. Introduction

RECENTLY, studies on how various cognitive functions can be developed through the experience of acting and perceiving in the environment have been attracting more researchers in the fields of cognitive neuroscience and cognitive robotics [1-4]. In neuroscience, the brain functions for perception, action and their association have been widely studied. A representative study illustrating possible links between perception, action, and cognition might be the one on the mirror neurons system (MNS) [5]. Mirror neurons were first discovered in area F5 of the monkey's premotor cortex [6, 7] and it was reported that mirror neurons were activated while executing own actions as well as observing the same ones performed by others [8-11]. The MNS has been reported to be involved in several cognitive processes, including action understanding and intention recognition [5]. Other previous studies have also illustrated the roles of the perception-action link in many cortical functions, including working memory, attention, and in social interaction [2, 9, 12, 13]. In [2], the authors argued that "the brain basis of cognition can be understood in terms of interlinked action perception representations".

In this study, we investigate how the cognitive functions of agents for generating and recognizing actions can be developed from learning causal models between one's own intentions and the resultant sensory outcomes perceived in dynamic visuo-proprioceptive patterns in the course of iterative interactions between the agents and the environment. Particularly, we focus on how cognitive competency for mental simulation and intention recognition can be developed as they play an important role, particularly in social cognition [2, 13-23]. Let us consider mutual imitation between two agents as an example for such social cognitive tasks. Imitation is closely interlinked with cognitive development [24] and it is important in acquiring sensorimotor skills as well as in social learning [25, 26]. Imitation is not only simply copying other's action but it requires a set of cognitive skills. That is, an agent is required to recognize the other's intention by observing their behavior and also to anticipate the consequences of own actions to the others [17, 23, 26, 27]. Therefore, it would be desirable if the agent can extract meaningful features from sensory observation and predict other's action as well as its own action [24].

This work was supported by the ICT R&D program of MSIP/IITP [2016(R7117-16-0163), Intelligent Processor Architectures and Application Softwares for CNN-RNN] in Korea and Okinawa Institute of Science and Technology Graduate University in Japan.

Jungsik Hwang is with the school of Electrical Engineering in Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea and Okinawa Institute of Science and Technology (OIST), Okinawa, Japan (e-mail: jungsik.hwang@gmail.com). Jinhyung Kim is with KAIST, Ahmadreza Ahmadi is with KAIST and OIST, Minkyu Choi is with OIST. Jun Tani is a corresponding author and he is with OIST (e-mail: tani1216jp@gmail.com).

We propose a dynamic neural network model called P-VMDNN (Predictive Visuo-Motor Deep Dynamic Neural Network). The proposed model is capable of learning large-scale visuo-proprioceptive patterns in a holistic manner by means of the hierarchically coupled multi-modal structure. In our previous studies [28-30], we have shown that the deep dynamic neural network model was able to extract latent features in dynamic visuo-proprioceptive patterns and to associate visual perception with proprioception by introducing multiple-scales spatio-temporal structure. In our recent work [31], the model has been extended under the predictive coding framework [32-35] to endow the model with the capability of acquiring a predictive internal model of others which is essential in social interaction [1, 23, 33, 35]. Consequently, the model was able to not only perceive the dynamic visuo-proprioceptive patterns but also predict them. In this study, we extend the previous model further based on the perception-action circuits found in the mammalian brain [12], so that vision and proprioception could be more tightly integrated (See Section II.A for detail).

We conducted a set of synthetic robotic experiments to examine the proposed model. In our experiment, a robot was trained to imitate gesture patterns of the human subjects in the simulation environment. We first examined how the proposed model could proactively reconstruct the learned visuo-proprioceptive primitives without the external inputs but with a given intention through the top-down process. Then, we examined the role of minimizing prediction error in recognizing the intention in the observed visuo-proprioceptive patterns.

There have been a few studies showing the implication of perception-action models on building embodied cognitive systems [19, 36-41]. Previous studies, however, often postulated separate learning processes for generation and recognition of actions [41] or a single pathway for multimodal patterns [39]. In addition, the computational model of the predictive coding framework which can handle large-scale pixel level visual stream patterns has not yet been fully addressed. The model proposed in the current study, however, can mirror generation and recognition processes for dealing with complex spatio-temporal patterns in visuo-proprioceptive input streams by using the predictive model developed from consolidative learning of the patterns.

II. Dynamic Neural Network Model

A. Model Overview

In this study, we extend an artificial neural network model called Predictive Visuo-Motor Deep Dynamic Neural Network (P-VMDNN) introduced in [31]. P-VMDNN is a dynamic neural network model which can build a predictive internal model of the environment through learning of large-scale spatio-temporal patterns of different modalities (vision and proprioception). For example, sequential patterns obtained from different sources (e.g., cameras and encoders embedded in a robot) can be learned in a holistic manner without any separate processing. P-VMDNN is an extension of our previous model [28-30] which consisted of the visual and proprioceptive pathways. In our previous studies [28-30], we showed how visual perception and proprioceptive information could be abstracted and associated in a spatio-temporally hierarchical structure. In our recent work [31], we extended the previous model under the predictive coding framework [32-35] to endow the model with the capability of predicting visuo-proprioceptive patterns. In this study, the model was improved to tightly integrate the visual and the proprioceptive pathways based on the findings in perception-action circuits in the mammalian brain [12].

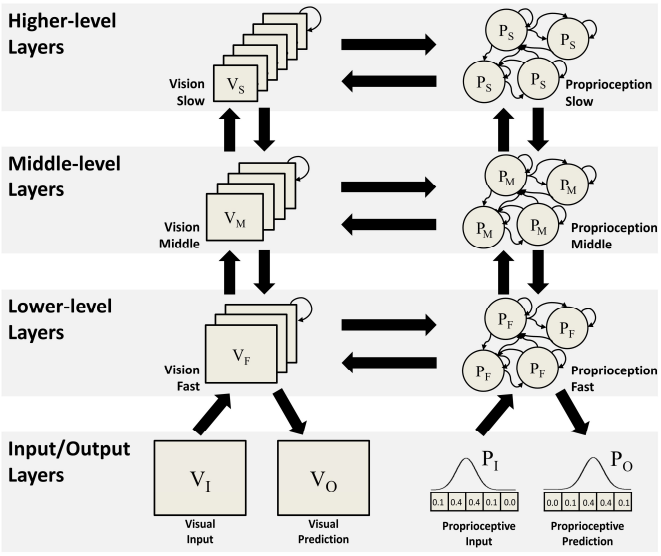


Fig. 1. The proposed model consists of the visual pathway (left) and the proprioceptive pathway (right). The visual pathway consists of Vision Input (V_I), Vision Output (V_O), Vision Fast (V_F), Vision Middle (V_M) and Vision Slow (V_S) layers. The proprioceptive pathway consists of Proprioception Input (P_I), Proprioception Output (P_O), Proprioception Fast (P_F), Proprioception Middle (P_M) and Proprioception Slow (P_S). The proposed model is an extension of our previous model [31] which has the lateral connection (horizontal arrows) at the higher-level layers only. The proposed model has been extended to have additional lateral connections at the middle-level and the lower-level layers based on the findings in perception-action circuits in [12].

The proposed model (Fig. 1) consists of the visual and proprioceptive pathways for perceiving and predicting the dynamic visual images and the proprioceptive signals (the perceptual outcome of the robot's actions), respectively. Each pathway is composed of a set of layers and the layers at the same level are connected reciprocally, allowing the bidirectional flow of the visuo-proprioceptive signals. Note that the lateral connection existed only at the higher-level layers in the previous model [31] whereas the proposed model is equipped with the lateral connections at every level of the hierarchy. By means of those lateral connections, vision and proprioception can be associated within the model by learning the visuo-proprioceptive patterns simultaneously in the tightly coupled structure.

There are several key features in the proposed model. First, the proposed model can learn high-dimensional visuo-proprioceptive patterns in a holistic manner. In our previous studies [28-30], it was shown that end-to-end learning on the hierarchical model enabled the development of the functional hierarchy, such that the higher-level and lower-level layers

encoded the abstract and specific information of the patterns respectively. Second, the proposed model is able to generate visuo-proprioceptive prediction proactively with a given intention through the top-down process. Similarly to our previous model [31], the proposed model is also capable of mentally simulating the possible incoming dynamic visuo-proprioceptive patterns without the external input from the environment [42-47]. The mental simulation capability is considered one of the important cognitive skills [2, 13, 43, 45-48] and it is essential to successfully interact with a dynamic environment [17, 48, 49]. Third, the proposed model provides a mechanism for updating the internal states through minimizing the prediction error, which results in recognition of the underlying intention latent in the perceived dynamic visuo-proprioceptive patterns. Recognizing others' intentions by observing their action is an essential skill required for social cognition [15, 16, 20, 21]. Minimizing prediction error is at the essence of predictive coding [32-34] and Kilner, Friston and Frith [34] argued that the underlying cause of the observed action could be inferred by minimizing the prediction error. Similarly, the proposed model provides an online prediction error minimization mechanism by which the intention behind the observed visuo-proprioceptive patterns can be inferred by updating the neurons' internal states in the direction of minimizing the prediction error. According to [34], the important aspect of predictive coding is that the same structure is employed in action generation as well as in action recognition. The proposed model utilizes the same neural architecture to generate the visuo-proprioceptive patterns and also to infer the cause of the perceived patterns. Finally, the lateral connections at all levels of the hierarchy were introduced in the proposed model to enable a tight coupling of vision and proprioception which is an essential component in cognitive development [3, 41, 50]. By means of the lateral connections in the proposed model, the visuo-proprioceptive information can flow bi-directionally at all levels of the hierarchy. As mentioned in [51], such a tight sensory-motor mapping can be simplified to a sensor-actuator function which can implement direct perception in the study on affordance [27, 52-54]. That is, situated behavior can be generated without complicated calculation, but by maintaining perceptual coordination [4].

B. Visual Pathway

Through the visual pathway, the model perceives and predicts the dynamic pixel-level visual images. To construct the visual pathway, we employed the predictive coding-based recurrent neural network model called P-MSTRNN (Predictive Multiple Spatio-Temporal Scales Recurrent Neural Network) which could perceive and predict the dynamic pixel-level images [55]. Instead of employing the separate feature maps and the context maps as in [55], the visual pathway in the proposed model consists of a single type of feature maps equipped with the recurrent connections.

In the proposed model, the visual pathway is composed of five layers: Vision Input and Output (V_I, V_O), Vision Fast (V_F), Vision Middle (V_M) and Vision Slow (V_S). Each layer consists of a group of 2-dimensional feature maps retaining spatial and temporal information and those layers are imposed with different spatio-temporal constraints. A previous study [55] has emphasized the importance of the progressively slower dynamics (from the lower to the higher-level) in achieving the functional hierarchy. Similarly, the smaller time constants are assigned on the lower-level layers and the bigger time constants are assigned on the higher-level layers. The layers in the visual pathway are connected bi-directionally from the I/O layers (V_I, V_O) to the highest-level layer (V_S). Also, the feature maps in each layer are equipped with the recurrent connections between the feature maps within the same layer.

At each time step t, the internal states u_i^{txy} and the activations v_i^{txy} of the neural units in each layer are computed as follows:

$$u_i^{txy} = \left(1-\frac{1}{\tau_i}\right)u_i^{(t-1)xy} + \begin{cases} \frac{1}{\tau_i}\left(\sum_{j\in V_M \vee V_S}\left(k_{ij} * v_j^{t-1}\right)^{xy} + \sum_{k\in P_S}\left(k_{ik} * y_k^{t-1}\right)^{xy} + b_i\right) & if\ i \in V_S \\ \frac{1}{\tau_i}\left(\sum_{j\in V_F \vee V_M \vee V_S}\left(k_{ij} * v_j^{t-1}\right)^{xy} + \sum_{k\in P_M}\left(k_{ik} * y_k^{t-1}\right)^{xy} + b_i\right) & if\ i \in V_M \\ \frac{1}{\tau_i}\left(\sum_{j\in V_F \vee V_M}\left(k_{ij} * v_j^{t-1}\right)^{xy} + \sum_{k\in P_F}\left(k_{ik} * y_k^{t-1}\right)^{xy} + \sum_{l\in V_I}\left(k_{il} * V_l^{t}\right)^{xy} + b_i\right) & if\ i \in V_F \\ \frac{1}{\tau_i}\left(\sum_{j\in V_F}\left(k_{ij} * v_j^{t}\right)^{xy} + b_i\right) & if\ i \in V_O \end{cases} \quad (1)$$

$$v_i^{txy} = 1.7159 \times \tanh\left(\frac{2}{3}u_i^{txy}\right) \quad (2)$$

i denotes the index of the feature map, x and y denote the horizontal and vertical location of the neural unit on the feature map, τ denotes the time constant, k_{ij} is the kernel connecting jth feature map in V_j with the ith feature map in the current layer, * is the convolution operator, b is the bias and V is an input visual image. Note that the transposed convolution operation is performed in cases where the size of the input feature map is smaller than the size of the output feature map, such as for the top-down connections from the higher-level layers and the lateral connections from the proprioceptive pathway. In our previous study [31], the lateral connection existed at V_S layer only. The proposed model is equipped with the additional lateral connections at V_M and V_F layers. To enhance the speed of convergence, the hyperbolic tangent recommended in [56] is used for the activation function (2).

C. Proprioceptive Pathway

The model perceives and predicts the perceptual outcomes of the robot's action (i.e. proprioception) through the proprioceptive pathway. Note that the proposed model predicts the perceptual outcomes of the action, not the actual action. The actual action (controlling the robot's joints) is accomplished by the motor control interface embedded in the robot which operates the robot's actuators based on the proprioceptive prediction (joint angle positions) given from the model. In this sense, the proprioceptive output of the model can be considered as the kinematic level representation of the action which describes the trajectories of the movement in space and time [34, 57].

To construct the proprioceptive pathway, a dynamic neural

network called Multiple Timescales Recurrent Neural Network (MTRNN) [58] is used. MTRNN is a hierarchical continuous time recurrent neural network consisting of a set of layers imposed with different temporal constraints. In this sense, MTRNN is similar to P-MSTRNN which is used to construct the visual pathway. However, it should be noted that P-MSTRNN imposes additional spatial constraints on neural activations so that it is more suitable to process pixel-level images of preserving local topology rather than a set of joint angles of robots without local topology. The distinctive feature of MTRNN is that it can self-organize a functional hierarchy in which the robot's action can be hierarchically represented [58, 59].

In the proposed model, the proprioceptive pathway is composed of five layers: Proprioception Input and Output (P_I, P_O), Proprioception Fast (P_F), Proprioception Middle (P_M) and Proprioception Slow (P_S). Each layer in the proprioceptive pathway is imposed with the different temporal constraints. More specifically, the progressively larger time constants from the lower-level layers to the higher-level layers are assigned as suggested in [28, 29, 58]. As a result, the neurons in the lower-level layers with the smaller time constants exhibit relatively faster dynamics compared to the ones in the higher-level layers. The P_I and P_O layers are composed of the softmax neurons representing the sparse representation of the robot's joint position values [42]. The neurons in the proprioceptive pathway have the recurrent connection between the neurons within the same layer. In addition, the neurons in each layer of the proprioceptive pathway have the bidirectional connection to the neurons in the neighboring layers as well as to the ones at the same level in the visual pathway (lateral connection).

At each time step t, the internal states p_i^t and the activations y_i^t of the neurons in each layer are computed as follows:

$$p_i^t = \left(1-\frac{1}{\tau_i}\right)p_i^{t-1} + \begin{cases} \frac{1}{\tau_i}\left(\sum_{j\in V_S} k_{ij} * v_j^{t-1} + \sum_{k\in P_S \vee P_M} w_{ik} y_k^{t-1} + b_i\right) & if\ i \in P_S \\ \frac{1}{\tau_i}\left(\sum_{j\in V_M} k_{ij} * v_j^{t-1} + \sum_{k\in P_F \vee P_M \vee P_S} w_{ik} y_k^{t-1} + b_i\right) & if\ i \in P_M \\ \frac{1}{\tau_i}\left(\sum_{j\in V_F} k_{ij} * v_j^{t-1} + \sum_{l\in P_I} w_{il} P_l^t + \sum_{k\in P_M \vee P_F} w_{ik} y_k^{t-1} + b_i\right) & if\ i \in P_F \\ \frac{1}{\tau_i}\left(\sum_{j\in P_F} w_{ij} y_j^t + b_i\right) & if\ i \in P_O \end{cases} \quad (3)$$

$$y_i^t = \begin{cases} \frac{\exp(p_i^t)}{\sum_{j\in P_O}\exp(p_j^t)} & if\ i \in P_O \\ 1.7159 \times \tanh\left(\frac{2}{3}p_i^t\right) & otherwise \end{cases} \quad (4)$$

i denote the index of the neuron, w_{ij} is the weight connecting the jth neuron to the ith neuron and P is a proprioceptive input. The convolution terms in (3) refer to the lateral connection from the visual pathway. The softmax activation function is used at the output layer (P_O) and the hyperbolic tangent recommended in [56] is used in the other layers as the activation function to enhance the speed of convergence.

D. Forward Dynamics

During the forward dynamics computation, the visual and proprioceptive predictions are generated with given inputs and the initial states of the model. The initial states refer to the internal states of the neural units at the beginning of the forward dynamics computation. To be more specific, the initial states (u_i^{oxy} and p_i^o) are given to every layer of the model at the onset of computation ($t = 0$). Then, at each time step t, the visual input (a pixel-level grayscale image) and the proprioceptive input (the robot's joint position values) are given to the vision input layer (V_I) and the proprioception input layer (P_I) respectively. Then, the internal states (u_i^{txy} and p_i^t) and the activations (v_i^{txy} and y_i^t) of the neural units at each layer are calculated using (1) ~ (4). Note that the visuo-proprioceptive information flows bi-directionally through the lateral connections during the forward dynamics computation.

In our study, two different methods of generating the visuo-proprioceptive predictions are considered. The first method is called an open-loop generation or the sensory entrainment [19]. In this method, the input to the model (the visual images and the joint position values) represents robot's current sensory perception obtained from the robot's cameras and the encoders, and this external sensory information drives the neural dynamics of the model.

Another method is called a closed-loop generation [19]. In this method, the input to the model is not from the external environment but from the model itself. That is, the visuo-proprioceptive prediction generated at the current time step is fed back to the input of the model in the next time step. Therefore, the closed-loop generation method does not require the external inputs from the environment, resulting in the mental simulation capability where the dynamic visuo-proprioceptive sequences can be anticipated [19, 42-48].

In our experiments, the closed-loop generation was used during the training process to achieve the robust mental simulation capability. During the testing process, two different methods were used to illustrate the key characteristics of the proposed model. Specifically, the closed-loop and the open-loop generation methods were used to examine the model 's performance with and without a prediction error minimization mechanism respectively (See Section II.F for a prediction error minimization mechanism).

E. Training the Model

The model is trained in a supervised end-to-end manner in which the visual and the proprioceptive pathways are trained simultaneously by directly learning the dynamic visuo-proprioceptive patterns [28]. In our experiment, the training data was collected during the tutoring process in which the robot was manually operated by the experimenter. This method is known as direct teaching or kinesthetic teaching [25, 60]. During the tutoring process, the experimenter demonstrated how to imitate the gestures by guiding the robot. While the robot was being guided by the experimenter, dynamic visual images perceived from the robot's camera were jointly

collected with the joint position values from the encoders in the robot's joints at each time step.

During the training, the model is trained to generate a one-step look-ahead visuo-proprioceptive prediction using backpropagation through time (BPTT) [61]. The model's learnable parameters, such as kernels (*k*), weights (*w*), biases (*b*) and the initial states (u_i^{0xy} and p_i^0) of the neurons are optimized to minimize the error defined as the sum of the errors in the visual pathway (E_V) and the proprioceptive pathway (E_P). Note that the initial states at every layer are obtained for each training sequence by computing the partial derivative of the error with respect to them ($u_i^{0xy} = \{u_{i1}^{0xy}, u_{i2}^{0xy}, \dots, u_{iN}^{0x}\}$, $p_i^0 = \{p_{i1}^0, p_{i2}^0, \dots, p_{iN}^0\}$ where N is the number of the training sequence) to generate the different visuo-proprioceptive patterns in the closed-loop manner. The error in each pathway is defined as the discrepancy between the predicted and the teaching signal (i.e. training data) as follows.

$$E = E_V + E_P \tag{5}$$

$$E_V = \sum_t \sum_y \sum_x \left(\bar{v}_i^{txy} - v_i^{txy}\right)^2 \tag{6}$$

$$E_P = \sum_t \sum_i \bar{y}_i^t log \frac{\bar{y}_i^t}{y_i^t} \tag{7}$$

Where $\bar{v}$ and $\bar{y}$ denote the visual and proprioceptive teaching signal respectively. Note that the error in the proprioceptive pathway is represented by the Kullback-Leibler divergence between the teaching signal $\bar{y}$ and the proprioceptive output *y* (7).

F. Inferring Internal States through Minimizing Prediction Error

Prediction error minimization is at the core of predictive coding [32-34]. Kilner, et al. [34] argued that one could infer the underlying cause of an observed action by minimizing the prediction error while observing the action. The proposed model provides a similar mechanism called an error regression scheme (ERS) [42, 62] by which the model minimizes the prediction error in an online manner.

Previous studies [42, 62] have shown that the higher-level intention in the observed sensorimotor patterns could be recognized by minimizing the prediction errors in an online manner. Note that the higher-level intention refers to internal cause enabling proactive generation of the visuo-proprioceptive patterns and they are specified as the internal states [62]. The previous studies have demonstrated the ERS with the single modality patterns [42] or in the single pathway that processed the multi-modal patterns [62]. In contrast, the ERS used in this study can be conducted with the visuo-proprioceptive patterns processed in the different pathways. In this sense, the ERS in our model is also different from the mental state inference (MSI) model [10, 63] which operates on the output in "visual-like coordinates".

The ERS consists of two distinct processes: the top-down and the bottom-up processes. In the top-down process, the model generates the visuo-proprioceptive predictions in the closed-loop manner with the given internal states representing the intention. In other words, the model predicts the perceptual consequence of the intended action as similar to the generative or the forward models [3, 34]. In the bottom-up process, the desired visuo-proprioceptive sequence is given and the prediction error between the desired and the predicted sequence is calculated. Then, the prediction error back-propagates from the output layers to the higher-level layers along the pathways, and the internal states of the neurons are updated in the direction of minimizing the prediction error at the output level. During the ERS, the top-down and the bottom-up processes are iteratively conducted to minimize the prediction error and to infer the possible cause of the observed visuo-proprioceptive patterns.

To implement the ERS, two hyper-parameters are required: the size of the temporal window and the number of iteration. The temporal window with the size of W represents the immediate past from the time step *t-W* to the current time step *t*. The number of iteration denotes the number of updates of the internal states conducted at each step during the ERS.

More precisely, at each time step *t*, the model generates the visuo-proprioceptive outputs (from *t-W* to *t*) with the initial states of the temporal window $u_{t\text{-}W}$ in the closed-loop manner (top-down). Note that the initial states of the temporal window $u_{t\text{-}W}$ refer to the internal states of the neurons in every layer at time step *t-W*. Then, the prediction error within the temporal window is computed (8) ~ (10) and back-propagates to update the initial states $u_{t\text{-}W}$ in the direction of minimizing the prediction error (bottom-up) with a learning rate η as illustrated in (11). As a result, the neural activation at all levels, as well as the visuo-proprioceptive predictions inside the temporal window, are updated. This process is iteratively conducted as specified by the number of iteration.

By means of the ERS, the proposed model is capable of updating the current intention to match the intention behind the perceived visuo-proprioceptive patterns through minimizing the perceptual prediction error generated in the immediate past. Note that only the initial states of the window $u_{t\text{-}W}$ are optimized during the ERS and the other learnable parameters are remained fixed during the ERS.

$$PE_t = PE_{V,t} + PE_{P,t} \tag{8}$$

$$PE_{V,t} = \sum_{s=t-W}^{t} \sum_y \sum_x \left(\bar{v}_i^{sxy} - v_i^{sxy}\right)^2 \tag{9}$$

$$PE_{P,t} = \sum_{s=t-W}^{t} \sum_i \bar{y}_i^s log \frac{\bar{y}_i^s}{y_i^s} \tag{10}$$

$$u_{t-W} = u_{t-W} - \eta \frac{\partial PE_t}{\partial u_{t-w}} \tag{11}$$

III. EXPERIMENTS

A. Experiment Settings

We conducted a set of experiments using the iCub simulator [64] to examine the proposed model. iCub [65] is a humanoid robot designed for cognitive and developmental robotics research and the iCub simulator provides a virtual environment where many of the robot's functionalities including perception and action can be examined. In addition, the iCub simulator and the real robot share the same controller interfaces, so that the model examined in the simulation environment can be easily extended to a real robot setting. Consequently, iCub and its

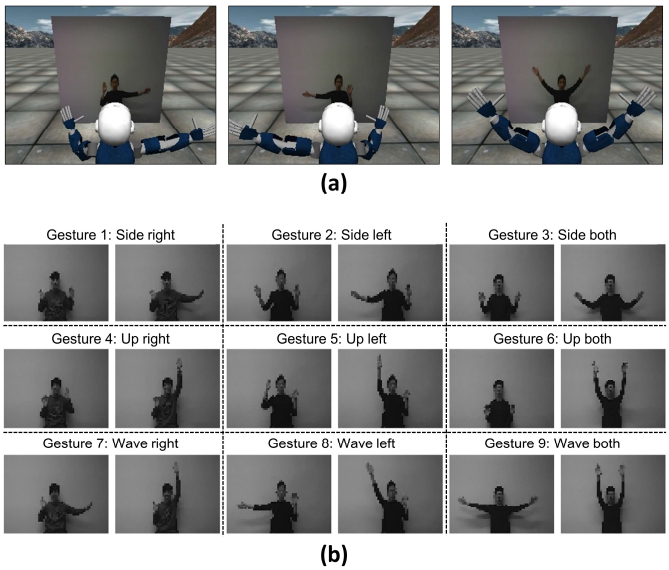


Fig. 2. The experiment setting. (a) The iCub simulator environment showing the human gesture on the screen and the robot. (b) The example of the nine gestures used in the imitation task.

simulator have been widely used in research on cognitive robotics and autonomous systems [36, 38, 66-69].

In our experiments, the robot was trained to imitate the gestures of the human subjects displayed on the screen (Fig. 2 (a)). In the imitation task, the model predicted not only its own movement (proprioceptive prediction) but also the movement of the human subject on the screen (visual prediction). The training data was composed of a set of the visuo-proprioceptive patterns collected from the tutoring process prior to training. During the tutoring process, the robot was operated manually by the experimenter to imitate the gestures of the human subjects on the screen (i.e. kinesthetic teaching [60]). At each step of the tutoring process, the visual images perceived from the robot's camera showing the gestures of the human subjects were jointly collected with the joint position values from the encoders in the robot's joints.

Regarding the robot's visual perception, we used the camera embedded in the left eye of the robot and the obtained visual images were converted to grayscale, resized to 64 (w) × 48 (h) and normalized to −1 to 1. Regarding the robot's behavior, we used the five joints (shoulder's pitch, roll, yaw, elbow, wrist's pronosupination) in each arm (a total number of 10 joints). To enhance learning, we used the sparse representation of the joint position values as illustrated in [42]. Each joint position value was converted into a sparse form represented by the 10 softmax neurons. Accordingly, there were 100 softmax neurons in the P_I and P_O layers that consisted of 10 groups, each representing a joint position value. Each group was composed of 10 softmax neurons.

Table I shows the values of the network's hyper-parameters including the number and the size of the feature maps and the neurons, kernels, weights and the time constants. Those values were found empirically in our preliminary experiments and they were used throughout our experiments. Regarding the time constant settings, we assigned progressively larger time constants from the lower levels to the higher levels in each pathway as suggested in [28, 29, 55, 58].

TABLE I
THE PARAMETER SETTING USED IN OUR EXPERIMENTS

		Visual Pathway				
		V_I	V_O	V_F	V_M	V_S
Time Constants		1	1	2	4	8
Feature Maps	*Number*	1	1	4	8	12
	Size	64×48	64×48	60×44	29×21	13×9
Top-Down Kernel	*Size*	-	5×5	4×4	5×5	-
	Stride	-	1,1	2,2	2,2	-
Bottom-Up Kernel	*Size*	-	-	5×5	4×4	5×5
	Stride	-	-	1,1	2,2	2,2
Recurrent Kernel	*Size*	-	-	2×2	2×2	2×2
	Stride	-	-	1,1	1,1	1,1
Lateral Kernel	*Size*	-	-	60×44	29×21	13×9
	Stride	-	-	1,1	1,1	1,1
		Proprioceptive Pathway				
		P_I	P_O	P_F	P_M	P_S
Time Constants		1	1	2	4	8
Number of Neurons		100	100	30	20	10
Top-Down Weights		-	30×100	20×30	10×20	-
Bottom-Up Weights		-	-	100×30	30×20	20×10
Recurrent Weights		-	-	30×30	20×20	10×10
Lateral Kernel	*Size*	-	-	60×44	29×21	13×9
	Stride	-	-	1,1	1,1	1,1

B. Experiment 1. Mental Simulation of the Visuo-Proprioceptive Patterns

1) Top-down Proactive Generation of the Visuo-proprioceptive Patterns

In the first experiment, we examined the model's mental simulation capability [43-48]. During the training, the robot was trained to imitate nine different types of gesture demonstrated by three human subjects (Fig. 2 (b)). Consequently, a total number of 27 visuo-proprioceptive sequences were used in the training. Those gestures consisted of the different arm movements: side right, side left, side both, up right, up left, up both, wave right, wave left and wave both. Each human subject showed slight differences in appearance including amplitude and speed of the gestures. Tensorflow [70] was used during the training and the model was trained for 100,000 epochs using the ADAM optimizer [71] with the learning rate of 0.0001. At the beginning of the training, the learnable parameters were initialized with the neutral values. Note that the different initial states were obtained for each training sequence, resulting in 27 initial states for each training data. After the training, the model generated the trained sequences in the closed-loop manner with the given initial states obtained during the training.

The result verified the mental simulation capability of the proposed model. Fig. 3 depicts the visuo-proprioceptive predictions generated in the closed-loop manner (see the supplementary video also). Note that the trajectories of the 10 joints in the model's proprioceptive predictions are depicted on the same scale although they have different ranges of joint

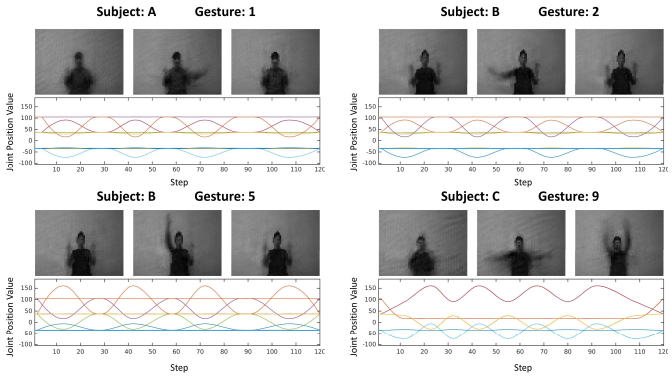


Fig. 3. The visuo-proprioceptive predictions generated in the closed-loop method (mental simulation).

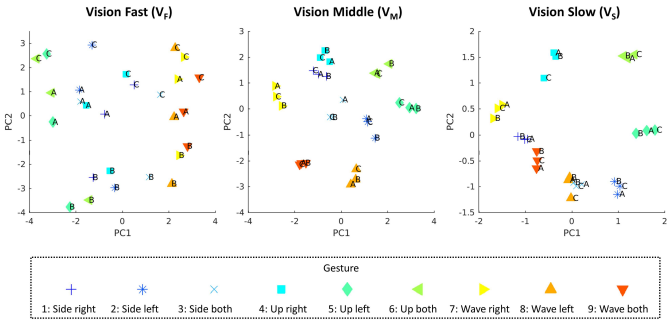


Fig. 4. PCA plot on the initial states of each primitive in the visual pathway (V_F, V_M and V_S). The horizontal and the vertical axes indicate the first and the second principal component respectively. The alphabet character denotes the human subject and the colors and the shapes indicate the type of gesture.

angles. With the given initial states, the model was able to generate a set of the dynamic visuo-proprioceptive patterns proactively in a top-down manner without the external inputs. That is, based on the model's own intention specified as the initial states, the model anticipated the consequence of its own action (proprioceptive prediction) as well as of other's action (visual prediction). Furthermore, the visual prediction generated during the mental simulation was in synchrony with the proprioceptive prediction, implying the coordinated and the tightly coupled vision and proprioception. In addition, as can be seen from Fig. 3, the different shapes between the human subjects were preserved in the visual prediction.

To investigate how the robot's 'intention' was encoded within the model, we conducted a principal component analysis (PCA) on the initial states. Fig. 4 illustrates the internal representation of the initial states of the visual pathway (V_F, V_M and V_S). In Fig. 4, the X and the Y axes indicate the first and the second principal components respectively. The colors denote the types of gesture and the alphabet character indicates the human subject.

The PCA results showed that each layer encoded the different level of the representation. In the lowest-level layer (V_F), the internal representations of the training sequences belonging to the same human subject were distributed closely. In V_M, those representations started to form the clusters. Finally, the clusters reflecting the type of the gesture appeared in the highest-level layer (V_S). This implies that the abstract information, such as the type of the gesture was encoded in the higher-level layer whereas the specific information, such as the shape of a specific human subject was encoded in the lower-level layer. In turn, this result suggests that the functional hierarchy was self-organized within the model.

2) Mental Simulation of the Sequential Visuo-proprioceptive Patterns

To investigate the functional hierarchy of the model, we trained the model further with an additional training data. A total number of 27 visuo-proprioceptive patterns were used and those patterns were generated by concatenating the three visuo-proprioceptive sequences (primitives) randomly. That is, each sequential pattern in this experiment contained randomly selected three visuo-proprioceptive patterns of a randomly selected human subject. The model's learnable parameters were initialized with the ones obtained from the previous training. Then, the network was trained for 50,000 epochs in the closed-loop manner using the ADAM optimizer [71] with the learning rate of 0.0001. Similar to the previous training, the different initial states were obtained for each training data during the training. After the training, the model generated the trained sequences in the closed-loop manner with the given initial states obtained during the training (mental simulation).

Fig. 5 illustrates some examples of the visuo-proprioceptive predictions generated in the closed-loop method. The result confirmed the model's mental simulation capability (See the supplementary video also). With the given initial states, the model was able to generate the visuo-proprioceptive patterns consisting of the different primitive sequences and transitions between them. Similar to the previous experiment, the visual prediction generated during the closed-loop method was in synchrony with the proprioceptive prediction, implying the coordinated vision and proprioception. Furthermore, it was observed that the model was able to generate the sequential data of the different human subject, suggesting that the low-level representation (appearance of the human subjects) were also preserved in the visual predictions.

In order to clarify the internal dynamics, we conducted a PCA on the neural activation in the highest-level vision layer (V_S) and the lowest-level layers (V_F and P_F). Fig. 6 illustrates the development of the internal representations for the exemplar cases. The result showed that the sequential training data was hierarchically represented within the model. The lower-level layers (V_F and P_F) were directly related to the current visuo-proprioceptive sequence being generated whereas the higher-level layer (V_S) showed the switch between the primitive sequences. In other words, the low-level representation of the visuo-proprioceptive patterns was encoded in the lower-level layers meanwhile the higher-level representation was encoded in the higher-level layer. This result implies the self-organized functional hierarchy within the model. It is assumed that the functional hierarchy could be self-organized by means of the spatio-temporal hierarchy achieved the different spatio-temporal constraints imposed on each level of the model. As a result, the model was capable of learning compositional visuo-proprioceptive sequences by means of the self-organized

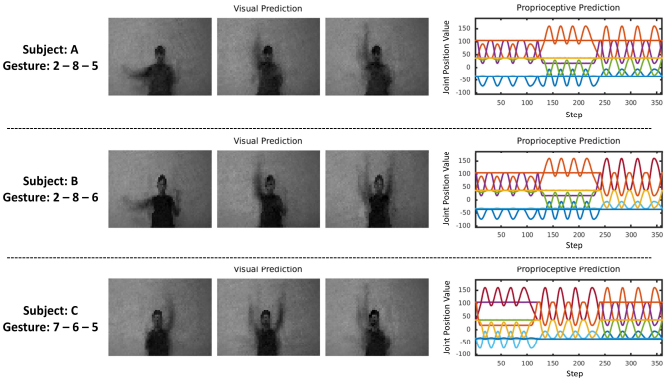


Fig. 5. The closed-loop generation of the sequential visuo-proprioceptive patterns.

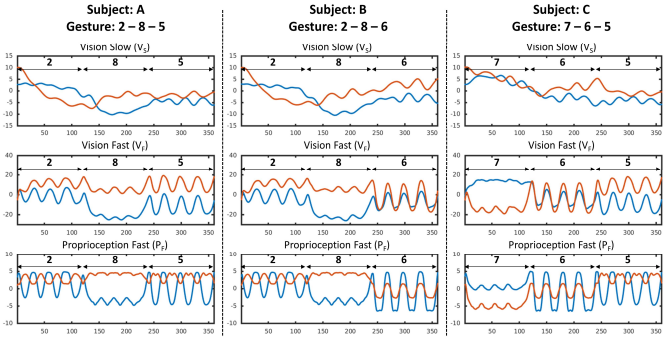


Fig. 6. The PCA plot showing the development of internal representations in V_S, V_F and P_F for the exemplar cases (sequential patterns). The x axis indicates the time step and the blue and the red colors denote the first and the second principal component respectively.

hierarchy as in [4]. This result is also in line with the previous studies [55, 58, 59] that showed the hierarchical representation of the action. In sum, the first experiment verified that the proposed model was capable of mentally simulating the perceptual consequences of the action in a coordinated manner by utilizing the self-organized functional hierarchy.

C. Experiment 2. Inferring Intention States by Prediction Error Minimization

In Experiment 2, we investigated the model's capability of inferring the underlying intention in the observed visuo-proprioceptive patterns by means of the prediction error (PE) minimization mechanism (i.e. error regression scheme, ERS. See Section II. F). Since the proposed model generated both visual and proprioceptive prediction, the two different conditions were examined: minimizing the visual PE and minimizing the proprioceptive PE.

1) Minimizing the Visual Prediction Error

In the minimizing the visual PE condition, it was assumed that the robot observed the human subject's gestures displayed on the screen. Then, the visual prediction error defined as the discrepancy between the perceived and the predicted visual images (gestures of the human subject) was minimized through updating the initial states of the error regression window. Note that the model generated both visual and proprioceptive predictions in the closed-loop manner, meaning that the visual observation was used as the target signal for computing the prediction error, not as the input to the model.

TABLE II
AVERAGE MEAN SQUARED ERROR (MSE) IN THE MINIMIZING VISUAL PREDICTION ERROR CONDITION

	Learned Subject		Unlearned Subject	
	Vision	Proprioception	Vision	Proprioception
Error Regression	0.0046	78.92	0.0063	269.91
Sensory Entrainment	0.0259	1067.58	0.0374	964.05

The learnable parameters except the initial states were initialized to the values obtained from the previous experiment. The initial states of the neurons at each layer were initialized with the neutral values. During the ERS, the size of the temporal window was set to 20 steps and the initial states of the temporal window were updated 50 times at each time step using the ADAM optimizer [71] with the learning rate of 0.1. Note that only the initial states of the window were updated in the direction of minimizing the prediction error during the ERS. Two visuo-proprioceptive sequences consisting of the five sequential primitive visuo-proprioceptive sequences were used during the ERS. One sequence consisted of the visuo-proprioceptive patterns used in the previous experiment (i.e. learned human subject data) whereas another sequence consisted of the novel visuo-proprioceptive patterns (i.e. unlearned human subject data).

To examine the importance of minimizing visual PE, we also examined the model's performance without minimizing the visual prediction error (sensory entrainment). In the sensory entrainment condition, the visual prediction was generated in the open-loop manner, meaning that the visual input (pixel-level image) was given from the external source (camera) at each time step. On the other hand, the proprioceptive prediction was generated in the closed-loop manner by feeding the proprioceptive output at the current time step to the proprioceptive input at the next time step.

The result showed that the model was able to predict the movements of the human subject successfully by minimizing the visual prediction error (Table II). In the case of the learned human subject gesture (Fig. 7 (a)), the model was able to reconstruct the gestures showing the shape of the specific human subject (MSE = 0.0046). Moreover, the model generated the proprioceptive prediction that corresponded to the visual prediction, resulting in successful imitation in both learned (MSE = 78.92) and unlearned (MSE = 269.91) subject cases (See the supplementary video).

In the sensory entrainment condition, however, the model was not able to generate neither the visual nor the proprioceptive predictions, leading to unsuccessful imitation. This result shows the importance of the prediction error minimization in communication and interaction between the two agents. Interestingly, in the sensory entrainment condition of the unlearned subject case (the bottom row in Fig. 7 (b)), the shape of the human subject that appeared in the visual

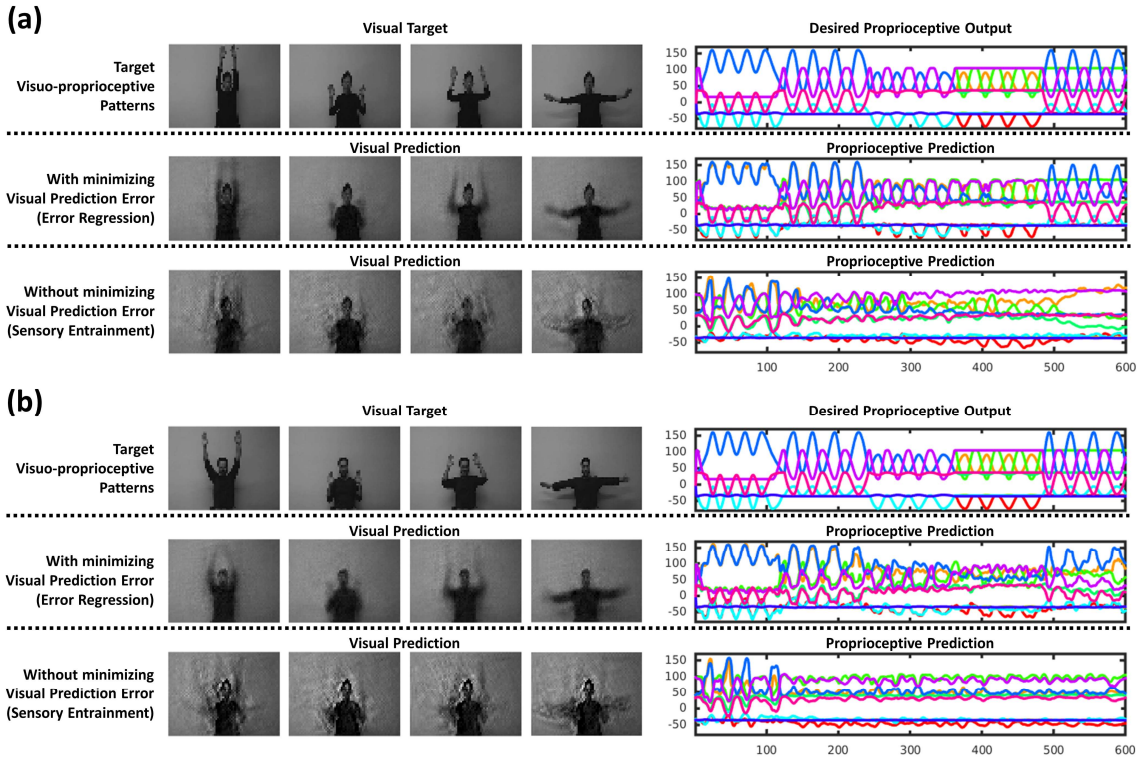


Fig. 7. The visuo-proprioceptive predictions generated in the minimizing visual prediction error condition, tested with (a) the learned human subject gestures and (b) the unlearned human subject gestures.

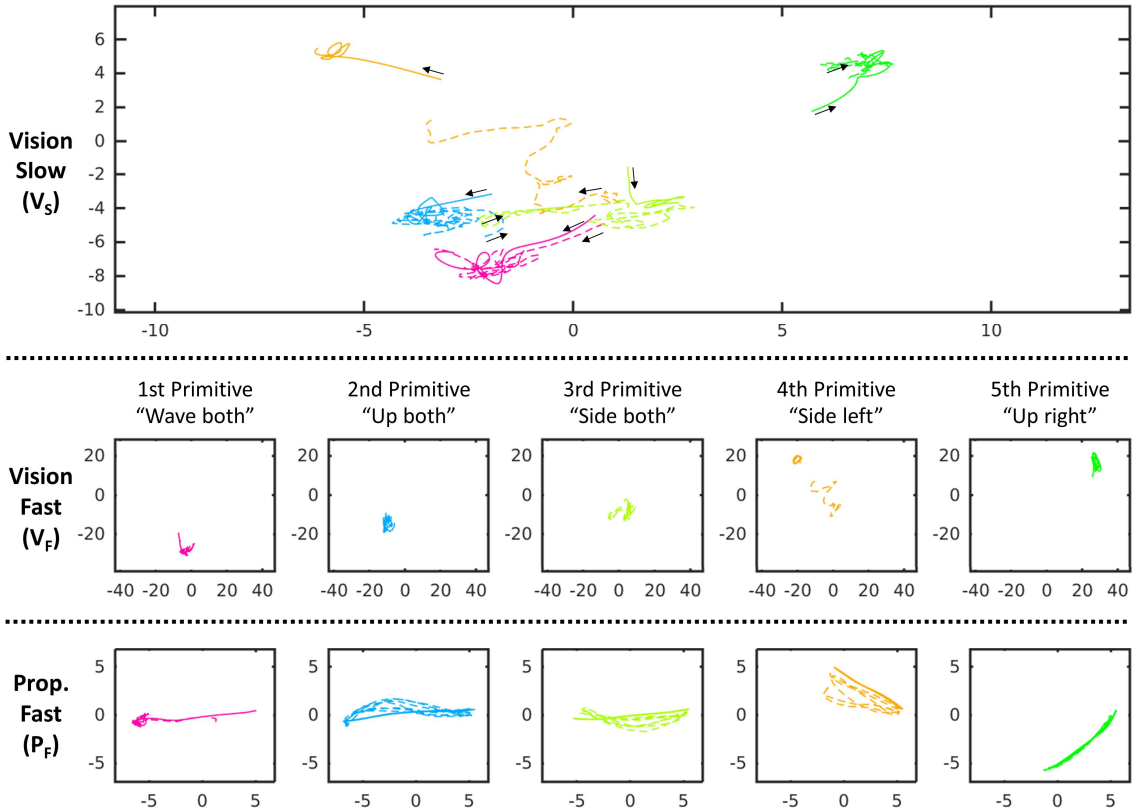


Fig. 8. PCA plot showing the internal representation emerged after the training (solid lines) and during the ERS in the visual PE minimization (dashed lines). The X and Y axes indicate the first and the second principal component respectively. The colors denote the type of the gesture. The black arrows in Vision Slow (V_S) indicate the direction of temporal evolution.

prediction was different from the shape of the human subject in the target sequence, but it was closer to one of the trained human subjects. This implies that the model predicts the visual image not by simply mapping from the visual input in a

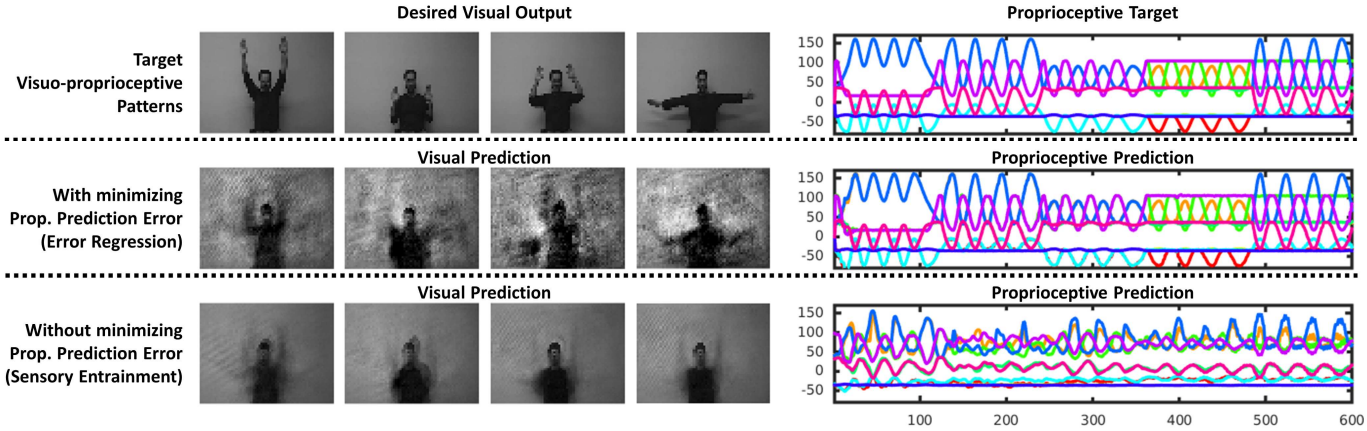


Fig. 9. Visuo-proprioceptive predictions generated in the proprioceptive PE minimization condition (the middle row) and the sensory entrainment condition (the bottom row). Note that the desired visual output was depicted to illustrate the desired type of gesture.

previous step, but by recalling the visual representation acquired from the training.

In order to clarify the internal dynamics during the ERS, we conducted a PCA on the neural activation at the highest-level of the visual pathway (V_S) and the lowest-level layers (V_F and P_F). Fig. 8 illustrates the internal representations emerged after the training (solid lines) and the ones emerged during the ERS (dashed lines). The horizontal and the vertical axes indicate the first and the second principal components respectively and the colors denote the type of gesture. It was observed the internal representations emerged during the ERS were close to the ones emerged after the training (i.e. overlapping between the solid and the dashed lines in the plots). This result suggests that the visual image showing the human subject's gesture was successfully recognized by recalling the corresponding internal representations in the model's repertoire acquired during the training. Consequently, the visual prediction of the other's action, as well as the proprioceptive prediction of its own action, could be generated, resulting in successful imitation.

2) Minimizing the Proprioceptive Prediction Error

In the minimizing the proprioceptive PE condition, it was assumed that the desired joint position values were provided from the environment and the discrepancy between the perceived and the predicted proprioceptive signals was minimized through updating the initial states of the error regression window. This condition emulates the situation in which the user grasps the robot's arms and moves as he/she wants while the robot is imagining the visual imagery of the human subject's gesture that corresponds to the given proprioceptive signal. During the ERS, the model generated both visual and proprioceptive predictions in the closed-loop manner, meaning that the desired proprioceptive pattern was used as the target signal for computing the prediction error, not as the input to the model. The same network examined in the visual PE minimization condition was used in the proprioceptive PE minimization condition. A visuo-proprioceptive sequence consisting of the five sequential primitive visuo-proprioceptive sequences was used during the ERS. Note that the sequence was different from the training data.

We also examined the model's performance without minimizing the proprioceptive PE (sensory entrainment). In the sensory entrainment condition, the proprioceptive prediction was generated in the open-loop manner, meaning that the proprioceptive input (joint position values) was given from the external source (encoders) at each time step. On the other hand, the visual prediction was generated in the closed-loop manner by feeding the visual output at the current time step to the visual input at the next time step.

Fig. 9 illustrates the target visuo-proprioceptive patterns (top), the visuo-proprioceptive predictions generated under the minimizing proprioceptive PE condition (middle) and the sensory entrainment condition (bottom). In the proprioceptive PE minimization condition, the model successfully generated the proprioceptive predictions (MSE = 2.70), showing that the model was able to adapt to the incoming proprioceptive signals.

Interestingly, the model was also able to generate the visual prediction showing the human subject's gesture which corresponded to the proprioceptive prediction. Although the visual prediction was a bit noisy, the gestures appeared in the visual prediction were still identifiable (See the supplementary video also). As similar to the previous experiment, it is assumed that minimizing the proprioceptive PE induced the recall of the proprioceptive representation as well as the visual representation of the corresponding gesture and in turn, the prototypical shape of the corresponding human subject's gestures appeared in the visual prediction. Without the PE minimization condition (sensory entrainment), however, neither the proprioceptive (MSE = 801.42) nor the visual (MSE = 0.0314) predictions were generated successfully, highlighting the importance of the PE minimization.

IV. Discussion

In this study, we proposed a dynamic neural network model which could build a predictive internal model of the environment from consolidative learning of spatio-temporal patterns. The experimental findings illustrated several key characteristics of the proposed model.

In Experiment 1, it was verified that the proposed model was able to anticipate the possible incoming visuo-proprioceptive

patterns through mental simulation (closed-loop generation) in a top-down manner. With the given intention specified as the initial states, the proposed model was able to generate visuo-proprioceptive predictions for each primitive sequences as well as the transition between the sequences in the compositional sequences. The mental simulation experiments also revealed that vision and proprioception were tightly coupled within the model. It is assumed that the coordinated visuo-proprioceptive representations were acquired during the consolidative learning of the patterns on the tightly coupled structure.

In addition, the experimental results showed the self-organized functional hierarchy of the proposed model. The analysis on the neural activation revealed that the visuo-proprioceptive patterns were hierarchically represented at each level of the model. That is, the lower-level layers encoded the low-level details of the visuo-proprioceptive patterns (e.g., the initial states grouped by the human subject in Fig. 4) whereas the higher-level layers encoded the abstract information of the patterns (e.g., the initial states grouped by the type of the gesture in Fig. 4). This finding is in line with the previous studies [28, 29, 55, 58, 59] and supports the notion of a hierarchical representation of actions [34, 57, 72, 73].

The findings in Experiment 2 highlighted the importance of the prediction error minimization, supporting the predictive coding account of the MNS as proposed in [34]. First, the results showed the role of the prediction error minimization in recognition of the intention in the observed patterns. The underlying intention in the perceived visuo-proprioceptive pattern was recognized by minimizing prediction error between the perceived and the predicted patterns. By recognizing the intention in the observed patterns, the model was also able to predict the possible incoming patterns in the next time step. Second, it was observed that minimizing the prediction error in one modality induced the recall of the corresponding representation in another modality acquired during the training. In the visual PE minimization condition, corresponding proprioceptive prediction was generated while the visual prediction error was minimized. Similarly, in the proprioceptive PE minimization condition, the model was able to generate the visual imagery showing the human subject's gesture that corresponded to the proprioceptive signals. Previous studies have shown the similar findings such that the activation in the cortical motor region was modulated when the actions that existed in the motor repertoire were recognized [74-76].

The importance of the prediction error minimization mechanism was further highlighted by comparing the model's performance in the sensory entrainment condition (i.e. without minimizing prediction error). Although the model was equipped with the same perception-action link, the MNS-like activity was not developed in the sensory entrainment condition in our experiments. Therefore, the MNS-like activity of the proposed model can be considered as the consequence of the several key features of the model, including the cortical connectivity, consolidative learning of the visuo-proprioceptive patterns and the PE minimization mechanism as suggested in the previous studies [2, 34].

In Experiment 2, the model was also able to respond to the gestures of the unknown human subject, illustrating the generalization capability of the proposed model. In our preliminary experiment with only one human subject data, we observed that the performance of the model with unknown human subject's data degraded. The generalization performance of the proposed model is expected to be enhanced by incorporating more training data as demonstrated in [55].

Note that the proposed model exploited the same neural architecture for generating the visuo-proprioceptive patterns as well as for recognizing the intention in the perceived visuo-proprioceptive patterns. It has been argued that the same neural substrate is shared for both perception and production of actions [7, 8]. In sum, the findings highlight the importance of the prediction error minimization mechanism in terms of inferring higher-level intention as well as recalling the corresponding visuo-proprioceptive representations acquired during the training.

There are several directions suggested for future research. First, the speed of the ERS should be improved to apply the proposed model in a real robot setting. Minimizing prediction error in our method requires iterative optimization at each time step. A different optimization technique can be examined to enhance the speed of the ERS so that it can be applied in real-time interaction. Second, the scalability of the proposed model in the different settings can be also examined. For instance, self-other distinction based on prediction error as suggested in [1] can be examined where visual input contains not only the gesture of other's but also robot's own. In addition, the proposed model can be examined under the circumstances where interaction between robots and humans goes through multiple developmental stages.

V. Conclusion

In this study, we investigated how the cognitive-like functions, such as mental simulation and intention recognition could be developed from consolidative learning of the low-level sensorimotor information under the predictive coding framework. We proposed a dynamic neural network model called P-VMDNN (Predictive Visuo-Motor Deep Dynamic Neural Network) which could perceive and predict the dynamic visuo-proprioceptive patterns. The experimental results validated several core features of the proposed model. First, the proposed model was able to develop the predictive internal model of the environment by directly learning the visuo-proprioceptive patterns acquired from the interaction with the environment. Due to the spatio-temporal hierarchy of the proposed model, the functional hierarchy was self-organized in a way that the visuo-proprioceptive patterns were encoded at the different level of the representation within the model. Second, the experimental results verified the mental simulation capability of the proposed model. With a given intention represented as the initial states, the model generated visuo-proprioceptive predictions proactively through the top-down process. By feeding its own output to input in the next time step (closed-loop generation), the model was capable of mentally simulating its own action (proprioceptive prediction) as well as

other's action (visual prediction) without inputs from the external world. Third, the experimental results highlighted the importance of minimizing prediction error in terms of inferring higher-level intention from the observed patterns as well as recalling the corresponding visuo-proprioceptive representations acquired during the training. The higher-level intention in the observed patterns was recognized in the process of minimizing prediction error through updating the internal states. It was also observed that updating the internal states to minimize the prediction error in one modality induced the recall of the corresponding representation of another modality, resulting in the generation of the corresponding perceptual sequences. To conclude, the current study suggests how artificial agents can develop higher-level cognitive functions from learning to perceive and predict the dynamic sensorimotor information. In addition, the findings of the current study support the predictive coding account of the mirror neuron system as proposed in [34]. The future study should involve with scaling of the proposed model experimented with real robots allocated with a much longer time period for developmental tutoring as well as interaction with human subjects.

REFERENCES

[1] Y. Nagai and M. Asada, "Predictive learning of sensorimotor information as a key for cognitive development," presented at the *Int. Conf. Intell. Robots Syst. (IROS) Workshop on Sensorimotor Contingencies for Robotics*, Hamburg, Germany, 2015.

[2] F. Pulvermüller, R. L. Moseley, N. Egorova, Z. Shebani, and V. Boulenger, "Motor cognition–motor semantics: Action perception theory of cognition and communication," *Neuropsychologia,* vol. 55, pp. 71-84, 2014.

[3] D. M. Wolpert, Z. Ghahramani, and M. I. Jordan, "An internal model for sensorimotor integration," *Science,* vol. 269, no. 5232, pp. 1880-1882, 1995.

[4] J. Tani, *Exploring robotic minds: actions, symbols, and consciousness as self-organizing dynamic phenomena*. New York, NY, USA: Oxford University Press, 2016.

[5] G. Rizzolatti and L. Fogassi, "The mirror mechanism: recent findings and perspectives," *Philos. Trans. R. Soc. Lond., B, Biol. Sci.,* vol. 369, no. 20130420, 2014.

[6] G. di Pellegrino, L. Fadiga, L. Fogassi, V. Gallese, and G. Rizzolatti, "Understanding motor events: a neurophysiological study," *Exp. Brain Res.,* vol. 91, no. 1, pp. 176-180, 1992.

[7] V. Gallese, L. Fadiga, L. Fogassi, and G. Rizzolatti, "Action recognition in the premotor cortex," *Brain,* vol. 119, no. 2, pp. 593-609, 1996.

[8] A. Aly and A. Tapus, "An Online Fuzzy-Based Approach for Human Emotions Detection: An Overview on the Human Cognitive Model of Understanding and Generating Multimodal Actions," in *Intelligent Assistive Robots: Recent Advances in Assistive Robotics for Everyday Activities*, S. Mohammed, J. C. Moreno, K. Kong, and Y. Amirat, Eds.: Springer International Publishing, 2015, pp. 185-212.

[9] E. Kohler, C. Keysers, M. A. Umiltà, L. Fogassi, V. Gallese, and G. Rizzolatti, "Hearing Sounds, Understanding Actions: Action Representation in Mirror Neurons," *Science,* vol. 297, no. 5582, pp. 846-848, 2002.

[10] E. Oztop, M. Kawato, and M. Arbib, "Mirror neurons and imitation: A computationally guided review," *Neural Netw.,* vol. 19, no. 3, pp. 254-271, 2006.

[11] E. Oztop, M. Kawato, and M. A. Arbib, "Mirror neurons: Functions, mechanisms and models," *Neurosci. Lett.,* vol. 540, no. Supplement C, pp. 43-55, 2013.

[12] J. Fuster, "The Prefrontal Cortex—An Update: Time Is of the Essence," *Neuron,* vol. 30, no. 2, pp. 319-333, 2001.

[13] M. Jeannerod, *Motor cognition: What actions tell the self* (Oxford Psychology Series 42). New York, NY, USA: Oxford University Press, 2006.

[14] P. Andry, P. Gaussier, S. Moga, J.-P. Banquet, and J. Nadel, "Learning and communication via imitation: An autonomous robot perspective," *IEEE Trans. Syst., Man, Cybern. A, Syst.,Humans,* vol. 31, no. 5, pp. 431-442, 2001.

[15] A. Cangelosi, M. Schlesinger, and L. B. Smith, *Developmental robotics: From babies to robots*. Cambridge, MA, USA: MIT Press, 2015.

[16] P. F. Dominey and F. Warneken, "The basis of shared intentions in human and robot cognition," *New Ideas Psychol.,* vol. 29, no. 3, pp. 260-274, 2011.

[17] G. Novembre, L. F. Ticini, S. Schütz-Bosbach, and P. E. Keller, "Motor simulation and the coordination of self and other in real-time joint action," *Soc. Cogn. Affect. Neurosci.,* vol. 9, no. 8, pp. 1062-1068, 2014.

[18] J. Su, "Representation and inference of user intention for Internet robot," *IEEE Trans. Syst., Man, Cybern., Syst,* vol. 44, no. 8, pp. 995-1002, 2014.

[19] J. Tani, "Self-organization and compositionality in cognitive brains: a neurorobotics study," *Proc. IEEE,* vol. 102, no. 4, pp. 586-605, 2014.

[20] M. Tomasello, M. Carpenter, J. Call, T. Behne, and H. Moll, "Understanding and sharing intentions: the origins of cultural cognition," *Behav. Brain Sci.,* vol. 28, no. 5, pp. 675-91, 2005.

[21] D. Vernon, S. Thill, and T. Ziemke, "The role of intention in cognitive robotics," in *Toward Robotic Socially Believable Behaving Systems-Volume I*: Springer International Publishing, 2016, pp. 15-27.

[22] A. Wohlschläger, M. Gattis, and H. Bekkering, "Action generation and action perception in imitation: an instance of the ideomotor principle," *Philos. Trans. R. Soc. Lond., B, Biol. Sci.,* vol. 358, no. 1431, pp. 501-515, 2003.

[23] D. M. Wolpert, K. Doya, and M. Kawato, "A unifying computational framework for motor control and social interaction," *Philos. Trans. R. Soc. Lond., B, Biol. Sci.,* vol. 358, no. 1431, pp. 593-602, 2003.

[24] Y. Kuniyoshi, "Learning from Examples: Imitation Learning and Emerging Cognition," in *Humanoid Robotics and Neuroscience: Science, Engineering and Society*, G. Cheng, Ed. Boca Raton, FL, USA: CRC Press, 2015.

[25] E. Ugur, Y. Nagai, H. Celikkanat, and E. Oztop, "Parental scaffolding as a bootstrapping mechanism for learning grasp affordances and imitation skills," *Robotica,* vol. 33, no. 5, pp. 1163-1180, 2015.

[26] M. Lopes, F. S. Melo, B. Kenward, and J. Santos-Victor, "A Computational Model of Social-Learning Mechanisms," *Adapt. Behav.,* vol. 17, no. 6, pp. 467-483, 2009.

[27] G. Saponaro, G. Salvi, and A. Bernardino, "Robot anticipation of human intentions through continuous gesture recognition," in *Proc. Int. Conf. Collaboration Tech. Syst. (CTS)*, San Diego, CA, USA, 2013, pp. 218-225.

[28] J. Hwang, M. Jung, J. Kim, and J. Tani, "A deep learning approach for seamless integration of cognitive skills for humanoid robots," in *Proc. IEEE Int. Conf. Dev. Learn. Epigenetic Robot. (ICDL-EpiRob)*, Cergy-Pontoise, France, 2016, pp. 59-65.

[29] J. Hwang, M. Jung, N. Madapana, J. Kim, M. Choi, and J. Tani, "Achieving "synergy" in cognitive behavior of humanoids via deep learning of dynamic visuo-motor-attentional coordination," in *Proc. IEEE-RAS Int. Conf. Humanoid Robots (Humanoids)*, Seoul, Korea, 2015, pp. 817-824.

[30] 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.

[31] J. Hwang, J. Kim, A. Ahmadi, M. Choi, and J. Tani, "Predictive Coding-based Deep Dynamic Neural Network for Visuomotor Learning," presented at the *IEEE Int. Conf. Dev. Learn. Epigenetic Robot. (ICDL-EpiRob)*, Lisbon, Portugal, 2017.

[32] K. Friston, "The free-energy principle: a unified brain theory?," *Nat. Rev. Neurosci.,* vol. 11, no. 2, pp. 127-138, 2010.

[33] K. Friston, J. Kilner, and L. Harrison, "A free energy principle for the brain," *J. Physiol. Paris,* vol. 100, no. 1, pp. 70-87, 2006.

[34] 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.

[35] R. P. Rao and D. H. Ballard, "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects," *Nat. Neurosci.,* vol. 2, no. 1, pp. 79-87, 1999.

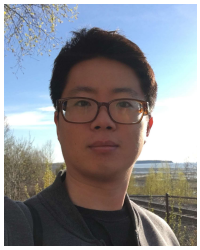
[36] J. L. Copete, Y. Nagai, and M. Asada, "Motor development facilitates the prediction of others' actions through sensorimotor predictive learning," in

Proc. IEEE Int. Conf. Dev. Learn. Epigenetic Robot. (ICDL-EpiRob), Cergy-Pontoise, France, 2016, pp. 223-229.
[37] A. Hanuschkin, S. Ganguli, and R. Hahnloser, "A Hebbian learning rule gives rise to mirror neurons and links them to control theoretic inverse models," *Front. Neural Circuits,* vol. 7, pp. 1-15, 2013.
[38] K. Rebrova, M. Pechac, and I. Farkas, "Towards a robotic model of the mirror neuron system," in *Proc. IEEE Int. Conf. Dev. Learn. Epigenetic Robot. (ICDL-EpiRob)*, Osaka, Japan, 2013, pp. 1-6.
[39] J. Tani, M. Ito, and Y. Sugita, "Self-organization of distributedly represented multiple behavior schemata in a mirror system: reviews of robot experiments using RNNPB," *Neural Netw.,* vol. 17, no. 8, pp. 1273-1289, 2004.
[40] G. Tessitore, R. Prevete, E. Catanzariti, and G. Tamburrini, "From motor to sensory processing in mirror neuron computational modelling," *Biol. Cybern.,* vol. 103, no. 6, pp. 471-485, 2010.
[41] K. Noda, H. Arie, Y. Suga, and T. Ogata, "Multimodal integration learning of robot behavior using deep neural networks," *Rob. Auton. Syst.,* vol. 62, no. 6, pp. 721-736, 2014.
[42] A. Ahmadi and J. Tani, "How can a recurrent neurodynamic predictive coding model cope with fluctuation in temporal patterns? Robotic experiments on imitative interaction," *Neural Netw.,* vol. 92, pp. 3-16, 2017.
[43] G. Hesslow, "Conscious thought as simulation of behaviour and perception," *Trends Cogn. Sci.,* vol. 6, no. 6, pp. 242-247, 2002.
[44] M. Ito and J. Tani, "On-line imitative interaction with a humanoid robot using a dynamic neural network model of a mirror system," *Adapt. Behav.,* vol. 12, no. 2, pp. 93-115, 2004.
[45] G. Pezzulo, M. A. van der Meer, C. S. Lansink, and C. M. Pennartz, "Internally generated sequences in learning and executing goal-directed behavior," *Trends Cogn. Sci.,* vol. 18, no. 12, pp. 647-657, 2014.
[46] J. Tani, "Model-based learning for mobile robot navigation from the dynamical systems perspective," *IEEE Trans. Syst. Man. Cybern. B Cybern.,* vol. 26, no. 3, pp. 421-436, 1996.
[47] T. Ziemke, D.-A. Jirenhed, and G. Hesslow, "Internal simulation of perception: a minimal neuro-robotic model," *Neurocomputing,* vol. 68, pp. 85-104, 2005.
[48] A. Di Nuovo, D. Marocco, S. Di Nuovo, and A. Cangelosi, "Embodied Mental Imagery in Cognitive Robots," in *Springer Handbook of Model-Based Science*: Springer International Publishing, 2017, pp. 619-637.
[49] R. J. Gentili, H. Oh, D.-W. Huang, G. E. Katz, R. H. Miller, and J. A. Reggia, "A Neural architecture for performing actual and mentally simulated movements during self-intended and observed bimanual arm reaching movements," *Int. J. Soc. Robot.,* vol. 7, no. 3, pp. 371-392, 2015.
[50] L. Lukic, J. Santos-Victor, and A. Billard, "Learning robotic eye-arm-hand coordination from human demonstration: a coupled dynamical systems approach," *Biol. Cybern.,* vol. 108, no. 2, pp. 223-48, 2014.
[51] E. Şahin, M. Çakmak, M. R. Doğar, E. Uğur, and G. Üçoluk, "To Afford or Not to Afford: A New Formalization of Affordances Toward Affordance-Based Robot Control," *Adapt. Behav.,* vol. 15, no. 4, pp. 447-472, 2007.
[52] L. Jamone *et al.*, "Affordances in psychology, neuroscience and robotics: a survey," *IEEE Trans. Cogn. Develop. Syst.,* vol. PP, no. 99, 2016.
[53] L. Montesano, M. Lopes, A. Bernardino, and J. Santos-Victor, "Learning object affordances: from sensory-motor coordination to imitation," *IEEE Trans. Robot.,* vol. 24, no. 1, pp. 15-26, 2008.
[54] P. Zech, S. Haller, S. R. Lakani, B. Ridge, E. Ugur, and J. Piater, "Computational models of affordance in robotics: a taxonomy and systematic classification," *Adapt. Behav.,* vol. 25, no. 5, pp. 235-271, 2017.
[55] M. Choi and J. Tani, "Predictive coding for dynamic vision: Development of functional hierarchy in a multiple spatio-temporal scales RNN model," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Anchorage, AK, USA, 2017, pp. 657-664: IEEE.
[56] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient backprop," in *Neural Networks: Tricks of the Trade*: Springer-Verlag Berlin Heidelberg, 2012, pp. 9-48.
[57] A. F. d. C. Hamilton and S. T. Grafton, "Goal representation in human anterior intraparietal sulcus," *J. Neurosci.,* vol. 26, no. 4, pp. 1133-1137, 2006.
[58] 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.
[59] R. Nishimoto and J. Tani, "Development of hierarchical structures for actions and motor imagery: a constructivist view from synthetic neuro-robotics study," *Psychol. Res. PRPF,* vol. 73, no. 4, pp. 545-558, 2009.
[60] S. Cho and S. Jo, "Incremental online learning of robot behaviors from selected multiple kinesthetic teaching trials," *IEEE Trans. Syst., Man, Cybern., Syst,* vol. 43, no. 3, pp. 730-740, 2013.
[61] D. E. Rumelhart, J. L. McClelland, and PDP Rsearch Group, *Parallel distributed processing*. Cambridge, MA, USA: MIT Press, 1987.
[62] S. Murata, Y. Yamashita, H. Arie, T. Ogata, S. Sugano, and J. Tani, "Learning to perceive the world as probabilistic or deterministic via interaction with others: a neuro-robotics experiment," *IEEE Trans. Neural Netw. Learn. Syst,* vol. 28, no. 4, pp. 830-848, 2015.
[63] E. Oztop, D. M. Wolpert, and M. Kawato, "Mental state inference using visual control parameters," *Cogn. Brain Res.,* vol. 22, no. 2, pp. 129-151, 2005.
[64] V. Tikhanoff, A. Cangelosi, P. Fitzpatrick, G. Metta, L. Natale, and F. Nori, "An open-source simulator for cognitive robotics research: the prototype of the iCub humanoid robot simulator," in *Proc. Perform. Metrics Intell. Syst. Workshop (PerMIS)*, Gaithersburg, MD, USA, 2008, pp. 57-61.
[65] G. Metta *et al.*, "The iCub humanoid robot: an open-systems platform for research in cognitive development," *Neural Netw.,* vol. 23, no. 8-9, pp. 1125-1134, 2010.
[66] A. Di Nuovo, M. Vivian, and A. Cangelosi, "A Deep Learning Neural Network for Number Cognition: A bi-cultural study with the iCub," in *Proc. IEEE Int. Conf. Dev. Learn. Epigenetic Robot. (ICDL-EpiRob)*, Providence, RI, USA, 2015, pp. 320-325.
[67] A. Di Nuovo, D. Marocco, S. Di Nuovo, and A. Cangelosi, "Autonomous learning in humanoid robotics through mental imagery," *Neural Netw.,* vol. 41, pp. 147-155, 2013.
[68] V. Tikhanoff, A. Cangelosi, and G. Metta, "Integration of speech and action in humanoid robots: iCub simulation experiments," *IEEE Trans. Auton. Mental Develop.,* vol. 3, no. 1, pp. 17-29, 2011.
[69] M. Peniak, D. Marocco, J. Tani, Y. Yamashita, K. Fischer, and A. Cangelosi, "Multiple time scales recurrent neural network for complex action acquisition," presented at the *IEEE Int. Conf. Dev. Learn. Epigenetic Robot. (ICDL-EpiRob)*, Frankfurt, Germany, 2011.
[70] M. Abadi *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467,* 2016.
[71] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," presented at the *Int. Conf. Learn. Represent. (ICLR)*, Banff, Canada, 2015.
[72] A. N. Meltzoff, "Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children," *Dev. Psychol.,* vol. 31, no. 5, p. 838, 1995.
[73] T. R. Savarimuthu *et al.*, "Teaching a robot the semantics of assembly tasks," *IEEE Trans. Syst., Man, Cybern., Syst,* vol. PP, no. 99, pp. 1-23, 2017.
[74] G. Caetano, V. Jousmäki, and R. Hari, "Actor's and observer's primary motor cortices stabilize similarly after seen or heard motor actions," *Proc. Natl. Acad. Sci. U.S.A.,* vol. 104, no. 21, pp. 9058-9062, 2007.
[75] M. A. Giese and G. Rizzolatti, "Neural and computational mechanisms of action processing: interaction between visual and motor representations," *Neuron,* vol. 88, no. 1, pp. 167-180, 2015.
[76] J. Kilner and C. Frith, "A possible role for primary motor cortex during action observation," *Proc. Natl. Acad. Sci. U.S.A.,* vol. 104, no. 21, pp. 8683-8684, 2007.

Jungsik Hwang received the B.S. in Electronic and Electrical Engineering (2011) and the M.S. in Interaction Science (2013) from Sungkyunkwan University, Korea. He is currently a Ph.D. student in the school of electrical engineering at Korea Advanced Institute of Science and Technology (KAIST), Korea and a Special Research Student at Okinawa Institute of Science and Technology (OIST), Japan. His research interests include cognitive neuro-robotics, human-robot interaction, sociable robots, human-computer interaction and multimodal interaction.

Jinhyung Kim received his B.S. in electrical engineering (2014) from Sogang Univ., Korea and M.S. in electrical engineering (2016) from Korea Advanced Institute of Science and Technology (KAIST). He is currently a Ph.D. student at KAIST. His research interests include deep learning, reinforcement learning, and neurorobotics.

Ahmadreza Ahmadi received the B.S. in control engineering (2008) from Shiraz University of Technology and the M.E. in mechatronics and automatic control (2012) from Universiti Teknologi Malaysia (UTM). He is currently a Ph.D. student in the school of electrical engineering at Korea Institute of Science and Technology (KAIST). His research interests include neurorobotics, stochastic neural models, and deep learning.



Minkyu Choi received the B.S. in Electrical and Electronic Engineering (2015) in Yonsei University and the M.S. in Electrical engineering (2017) in Korea Advanced Institute of Science and Technology. His research interests include cognitive robotics, deep learning and machine learning.

Jun Tani received the B.S. degree in mechanical engineering from Waseda University, Tokyo, Japan in 1981, dual M.S. degree in electrical engineering and mechanical engineering from the University of Michigan, Ann Arbor, MI, USA in 1988, and the D.Eng. degree from Sophia University, Tokyo in 1995.

He started his research career with Sony Laboratory, Tokyo, in 1990. He had been a Team Leader of the Laboratory for Behavior and Dynamic Cognition, RIKEN Brain Science Institute, Saitama, Japan, for 12 years until 2012. He was a Visiting Associate Professor with the University of Tokyo, Tokyo, from 1997 to 2002. He was a Full Professor with the Electrical Engineering Department, Korea Advanced Institute of Science and Technology, Daejeon, South Korea, from 2012 to 2017. He is currently a Full Professor with the Okinawa Institute of Science and Technology, Okinawa, Japan. His current research interests include neuroscience, psychology, phenomenology, complex adaptive systems, and robotics. He is an author of “Exploring Robotic Minds: Actions, Symbols, and Consciousness as Self-Organizing Dynamic Phenomena." published from Oxford Univ. Press in 2016.