Published online in Cognitive Neurodynamics, 8 October, 2011

Neuro-Robotics Study on Integrative Learning of Proactive Visual Attention and Motor Behaviors

Sungmoon Jeong[a], Hiroaki Arie[b], Minho Lee [a,*] and Jun Tani[b]

[a]School of Electronics Engineering, Kyungpook National University,

1370 Sankyuk-Dong, Puk-Gu, Taegu 702-701, Korea

[b]Lab. for behavior and Dynamic Cognition, RIKEN Brain Science Institute,

2-1 Hirosawa, Wako-shi, Saitama, 3510198, Japan

[*]Corresponding Author: Tel: +82-53-950-6436, Fax: +82-53-950-5505, E-mail: mholee@knu.ac.kr

Abstract

The current paper proposes a novel model for integrative learning of proactive visual attention and sensory-motor control as inspired by the premotor theory of visual attention. The model is characterized by coupling a slow dynamics network with a fast dynamics network and by inheriting our prior proposed multiple timescales recurrent neural networks model (MTRNN) that may correspond to the fronto-parietal networks in the cortical brains. The neuro-robotics experiments in a task of manipulating multiple objects utilizing the proposed model demonstrated that some degrees of generalization in terms of position and object size variation can be achieved by organizing seamless integration of the proactive object-related visual attention and the related sensory-motor control into a set of action primitives in the distributed neural activities appearing in the fast dynamics network. It was also shown that such action primitives can be combined in compositional ways in acquiring novel actions in the slow dynamics network. The experimental results presented substantiate the premotor theory of visual attention.

Keywords: proactive visual attention, multiple time recurrent neural networks, premotor theory of visual attention, multiple objects manipulation task.

1. Introduction

Humans perform gaze shifts and fixations (visual attention) proactively to gather visual information for guiding movements (Johansson et al, 2001). This visual attention can improve the reach accuracy (of the hand) to a specific target in multiple objects through visual feedback of the hand position (Berkinblit et al., 1995; Bowman et al., 2009; Carlton 1981; Land et al., 1999; Paillard, 1996; Sarlegna et al., 2004; Saunders et al., 2004) and even when the hand is not visible (Prablanc et al., 1979, 1986, 1992, 2003). In addition, the visual attention can effortlessly detect (location) and help to recognize (identification) an interesting area or object within natural or cluttered scenes through the selective attention mechanism using various visual features such as color, orientation, scale, symmetry and top-down knowledge (Ballard, 1991; Berkinblit et al., 1995; Jeong et al., 2008; Johansson et al., 2001; Linda et al., 2009; Rao et al. 1995; Yuqiao et al., 2007). Due to its capacity limitations, the brain can process only a fraction of this information using visual attention (Lennie, 2003; Tsotsos, 1990; Tsubomi et al., 2009).

Especially in the cases of goal-directed action generation accompanying complex sensory-motor interactions with environments, proactive shifts of visual attention from one part of the environment to other should be indispensable to generate adequate motor behaviors. For example, let us consider a situation of multiple object manipulation where we are asked to place a red object on a blue object located on a table in front of us. Through visual search, we attempt to locate an object with red color as a feature. Once fixated, our hands reach out for the red object for grasping while visual attention is shifted to fixate the blue object. After fixating the blue object, the red object is moved to the location of the current fixation and is placed on the blue object. We see that this type of visually-guided actions dealing with multiple objects manipulation require proactive sequential visual attention shifts synchronized with accompanied adequate hand movements (Crawford et al., 2004; Furneaux et al., 1999; Hayhoe et al., 2003; Herst et al., 2001; Land et al., 2000; Pelz et al., 2001). A particular motivation in the current study is to investigate the possible neuronal mechanism underlying this type of function by conducting synthetic neuro-robotics experimental studies.

Tani's group proposed a dynamic neural network model, the so called multiple timescale recurrent neural network (MTRNN) (Nishimoto et al., 2009; Yamashita & Tani, 2008) to account for the underlying neuronal mechanisms for learning and generation of complex goal-directed actions. This work show that a particular functional hierarchy can develop through learning by utilizing timescale differences of neural activities set into the network model. More specifically, it was shown that a set of primitive behaviors is acquired in the fast dynamic network at the lower level, while sequencing of these primitive behaviors for particular intention or goals develops in the slow dynamic network at the higher level. It was interpreted that the fast dynamics network may correspond to the inferior parietal cortex in which visuo-proprioceptive states for on-going actions are anticipated (Mulliken et al., 2008) while the slow dynamics network may correspond to the premotor or supplementary motor area in which plans of connecting action primitives into sequences for the current intentions or goals are generated (Rizzolatti et al., 1996; Tanji & Shima, 1994). The idea of the model is formally related to "active inference" (Friston et al., 2011), which can be regarded as a form of predictive coding (Rao & Ballard, 1999). The proposed model was evaluated through a humanoid robotics experiment dealing with single object manipulation tasks.

The current study renovates the MTRNN model to facilitate functions of the proactive visual attention shifts in generating goal-directed actions. We expect the robots controlled by this renovated model to learn to manipulate multiple objects with generating adequate visual attention shifts in a goal-directed manner. More specifically, the MTRNN predicts the object to be attended in terms of its attribute, e.g., by its color. The visual guiding system (Carpenter et al., 1992, Choi et al., 2006; Jeong et al., 2008), as an external device, spots the region of a specific object among others in the retinal image that (which) accords with the predicted attribute, e.g., red color. Then, the visual guiding system fixates the object, e.g., the red color object. The direction of fixation, which is the input to the MTRNN, provides information regarding the movement of the hands to manipulate the fixated object. One important proposal in the renovated model is that object manipulative motor behavior and the accompanying visual attention shift would be processed inseparably in distributed neural activities in the MTRNN.

This idea accords with the recent thoughts on embodied cognition (Beer, 2000; Metta et al., 1999; Varela & Thompson, 1991) that argues that cognitive processes such as executive control of attention cannot be separated from sensory-motor processes of interacting with environments. Furthermore, the idea is analogous to the so-called “premotor theory for visual attention” (Craighero et al., 1999; Rizzolatti et al., 1987) in the cognitive neuroscience literatures. Conventionally, spatial attention was considered as a dedicated meta-control mechanism, which is anatomically distinct from local regions underlying sensory-motor processing (Posner & Dehaene, 1994). The new theory, however, argues that there is no need to segregate these two mechanisms, one for attention and one for sensory-motor processing, as various psychological studies (Rizzolatti et al., 1994; Sheliga et al., 1995), neuroimaging studies (Corbetta et al., 1998; Nobre et al., 2000) and neurophysiological studies (Kustov & Robinson, 1996; Moore & Fallah, 2001) suggested that the same fronto-parietal circuits involve both processes. In addition, it was surprisingly shown that the same neuronal activities were observed even for covert visual attention without gaze or eye saccadic motions in the neurophysiological studies of monkeys by Kustov and Robinson (1996).

Upon this thought, the current synthetic neuro-robotics study investigates how goal-directed skilled behaviors of multiple objects manipulation can be learned through seamless integration between proactive visual attention and bimanual movement controls. A particular expectation is that such seamless integration between the two functions can be acquired through iterative learning of related sensory-motor experiences, which could facilitate generalization and compositionality of the system in generating a diversity of multiple objects manipulative actions. More specifically, if some basic action primitives, such as approaching an object, grasping an object or placing an object on other objects can be acquired as an integration of visual attention and movement controls, such action primitives could be insensitive to changes in position and size of objects by means of generalization to some extent. It is also expected that novel actions can be easily developed by combining those primitives in compositional ways. The current paper evaluates these ideas by conducting neuro-robotics experiments utilizing our newly proposed model. The following sections present our model and experimental tasks.

1.1 Multiple objects manipulation task

A humanoid robot is trained for a set of actions concerning multiple objects manipulation. A part of our experimental scenario is similar to the Ballard group’s and Johansson et al.’s human experiment setting (Ballard et al., 1992, 1995; Johansson et al., 2001; Smeets et al., 1996). They presented an eye-/hand coordination task when subjects moved color blocks from a pickup area and placed them in a desired location. Subjects sequentially shift the attention from a block before picking it up to be placed at a desired location. In our study, the robot learns to perform similar visual attention shifts and behaviors followed by acquired bimanual movement patterns; the robot sequentially attends to an object before picking it up and then to a destination location to place the object, as with human subjects. Therefore, visual attention shifts and behavior sequences are simultaneously cooperated and anticipated by considering current behavior and environments.

1.2 System overview

As shown in Fig. 1, the neuro-robotic system consists of MTRNN, an extra device of visual guiding system and a humanoid robot. When the inputs of specifying current action in the slow dynamics part of MTRNN are given, the MTRNN generates prediction of the proprioceptive state ($m_{t+1:}$ arm joint angles) and visual attention command (v_{t+1}) of next time step. Here, the visual attention command represents four object categories (red, green, blue and default preference color) to be attended. The proactive visual attention is performed overtly by directing the camera head (there is no eye saccadic movements). The visual guiding system receives a visual attention command from the MTRNN and the retina image from the robot's vision camera. Then, the visual guiding system localizes position of an object with the color specified by the visual attention command in X-Y coordinate within a robot camera view. The center position of the attended object is converted into the two-dimensional head joint angles (s_{t+1}) by a pre-programmed map (visuo-head directional map) within the visual guiding system. By making the head move to the obtained target angle (s_{t+1}) by PID control, the camera head can fixate the object. The current head joint angles ($\bar{s}_t$) are fed back as inputs to the MTRNN, by which the MTRNN receives positional information of the attended object. The current proprioceptive state ($\overline{m_t}$) as well as the current visual attention command (v_t), are fed back as inputs to MTRNN. The prediction of the proprioceptive state (m_{t+1}) is sent to the arm PID controller, as target joint angles at the next time step and the arms physically move to the target.

1.3 Multiple Timescales Recurrent Neural Network (MTRNN) for Behavior Generation Related to (With) Top-down Visual Attention Command and Arm Movement

The MTRNN is a variation of the continuous time recurrent neural network (CTRNN) model (Doya et al., 1989; Williams et al., 1989) in which neural units at different levels in the network are assigned different time scales. It was observed that a certain functional hierarchy can be self-organized through learning of a set of sensory-motor sequences in which neural units with a fast time constant encode a set of behavioral primitives, and those with a slow time constant prepare for the compositional sequences of these primitives (Nishimoto et al., 2009; Yamashita & Tani, 2008).

1.3.1 Forward Dynamics in MTRNN for Behavior Generation

In the current model, corresponding to each action different behavioral trajectory is generated by utilizing the initial sensitivity characteristics of the MTRNN. More specifically, each different action is started by providing specific initial states for some neural units in the slow dynamics network while initial states for other neural units in slow and fast dynamics network are set with neutral values. Therefore, the initial state is considered to represent the top-down intention of the action to be generated (Nishimoto et al., 2009; Yamashita & Tani, 2008).

The MTRNN has 216 CTRNN units (indices i=1-216) that consist of two groups of neural units in the present study, namely input-output units and context units. The first 116 units are namely input-output units to receive the external input and their activation values correspond to the CTRNN output. Of these, the first 64 units (indices i = 1-*6*4) correspond to the proprioceptive input (arm joint angle), the next 36 units (indices i = 65-100) correspond to the visual input (head movement), and the last 16 units (indices i = 101-116) correspond to the visual attention command, respectively. The 14 dimensional inputs, which consist of 8 joint angles for two arms with four degrees of freedom in each arm, two real head directional joint angles, and four dimensional visual attention commands, were thus transformed into 116 dimensional sparsely encoded vectors by a topology preserving map (TPM) with $3x10^6$ training epochs (Kohonen, 2001; Saarinen et al., 1985). The TPM was used to separately cluster the various sensory, motor and visual attention sequences and preserve the topological information of the input vectors (proprioceptive: arm joint angles, vision sense: real head directional joint angles, visual attention command) by localized firing neurons. The TPM was trained by offline unsupervised learning with training signals. The samples to train the TPMs included all sequences of behavioral tasks for the CTRNN by manually controlling the robot through the task. The size of the TPMs is 64 (8×8) for proprioception, 36 (6×6) for the vision sense information and 16 (4×4) for the top-down visual attention command. The number of input-output units is determined by the sizes of the TPMs as 116 input and output neural units. This transformation reduces the redundancy of the input trajectories for units. The sizes of the TPMs were selected for the current experiment, such that they were the minimum value sufficiently large to allow the TPMs to reproduce in real time, sensory-motor sequences through the process of vector transformation to reduce time spent on computation. The remaining 100 units are context units that consist of fast context units and slow context units. The first 70 of these 100 context units (indices i = 117-186) correspond to the fast context units with a small time constant value and the next 30 units (indices i = 187-216) correspond to the slow context units with a large time constant value τ. The fast context units are connected to the input-output units and slow context units. However, the input-output units are not directly connected to the slow context units. The number of context units was also selected to be the minimum value sufficiently large to successfully allow the network to learn the task sequences. The synaptic weights of each unit are determined through learning by examples. The activation of these neurons is calculated by Eq. (1)

$$\tau_i (du_{i,t} / dt) = -u_{i,t} + \sum_j w_{ij} x_{j,t} \qquad (1)$$

where $u_{i,t}$ is the membrane potential of each i-th neural unit at time step t and $x_{j,t}$ is the neural state of the i-th unit, and w_{ij} is the synaptic weight from the j-th unit to the i-th unit. The time constant τ is defined as the decay rate of a unit's membrane potential. One might consider this decay rate to correspond to an integrating time window of the neurons, in the sense that the decay rate indicates the degree to which the earlier history of synaptic inputs affects the current state. If the τ value is large, the activation of the unit changes slowly, because the internal state potential is strongly affected by the history of the unit's potential. Conversely, if the τ value is small, the effect of the history of the unit's potential is also small, and thus it is possible for the activation of the unit to change quickly. Context units were divided into two units, fast and slow context units, based on the value of time constant τ. The activity of the fast context units with small time constant (τ = 4) change quickly, while the activity of the slow context unit with a large time constant (τ = 20) changed slowly. Among the input-output units, the units corresponding to the proprioception and visual attention command were not connected to each other. In addition, input units were also not directly connected to the slow context units.

Neurons in the CTRNN are modeled by a conventional firing rate model, in which the activity of each unit constitutes an average firing rate over units of neurons. Continuous time characteristics of the model neurons are described by Eq. (2). Actual updating of $u_{i,t}$ values is computed according to Eq. (2), which is the numerical approximation of Eq. (1)

$$u_{i,t+1} = (1-1/\tau_i)u_{i,t} + (1/\tau_i)(\sum_{j\in N} w_{ij}x_{j,t}) \tag{2}$$

The activation of the *i*-th unit at time *t* is determined by the following Eq. (3)

$$y_{i,t} = \begin{cases} \dfrac{\exp(u_{i,t})}{\sum_{j\in Z}\exp(u_{j,t})} & if\ i \in Z \\ f(u_{i,t}) & otherwise \end{cases} \tag{3}$$

where Z is a set of output units that correspond to proprioception or vision. The softmax activation function is applied only to the output units, and not to the context units. Activation values of the context units are calculated by the function *f* which is a conventional unipolar sigmoid function. The softmax activation function applied to the CTRNN enables (maintaining consistency with the output of TPMs that are calculated using the softmax function. The output vector of the MTRNN is sent to the TPM and subsequently transformed into the predictions of the proprioception M_{t+1} at time t+1, and the top-down visual attention command V_{t+1} at time t+1. After this process, the robot initiates movement using these predicted values.

1.3.2 Learning of synaptic weights

In the present study, the TPMs were trained, prior to the MTRNN training, in an unsupervised manner. The MTRNN training could find the optimal values of the connective weights that minimize the value of learning error *E.* The error function *E* was defined by the Kullback-Leibler divergence, as shown in Eq. (4)

$$E = \sum_{t}\sum_{i\in O} y^{*}_{i,t}\log(y^{*}_{i,t}/y_{i,t}) \tag{4}$$

where $y^{*}_{i,,t}$ is the desired activation value of the output neuron at time *t*, O is a set of output units, and $y_{i,t}$ is the activation value of the output neuron with the current connective weight. A conventional back propagation through time (BPTT) algorithm was used to train the model (Rumelhart et al., 1986). In the actual learning process, the update rule of a connective weight from the *i*-th neuron to the *j*-th neuron at the *n*-th learning iteration step considering the opposite direction of gradient $\partial E / \partial w_{ij}$ is follows;

$$w_{ij}(n+1) = w_{ij}(n) - \alpha(\partial E / \partial w_{ij}) \tag{5}$$

where α is the learning rate. The gradient $\partial E / \partial w_{ij}$ is given by Eq. (6)

$$\frac{\partial E}{\partial w_{ij}} = \sum_{t} (1/\tau_i) \frac{\partial E}{\partial u_{i,t}} x_{j,t-1} \tag{6}$$

and is recursively calculated from the following recurrence formula Eq. (7)

$$\frac{\partial E}{\partial u_{i,t}} = \begin{cases} y_{i,t} - y^{*}_{i,t} + (1-1/\tau_i)\dfrac{\partial E}{\partial u_{i,t+1}} & i \in O \\ \displaystyle\sum_{k \in N} \frac{\partial E}{\partial u_{k,t+1}} \left[\delta_{ik}(1-1/\tau_i) + (1/\tau_k) w_{ki} f^{'}(u_{i,t}) \right] & i \notin O \end{cases} \tag{7}$$

where $f^{'}$ is the derivative of the unipolar sigmoid function and $\pm\delta_{ik}$ is Kronecker's delta ($\pm\delta_{ik}$ = 1 if $i = k$, otherwise $\pm\delta_{ik} = 0$). Through iterative calculation of the BPTT, the values of the connective weights reach their optimal values in the sense that the errors between a teaching sequence and an output sequence are minimized. During the learning iterations, the learning rate α is fixed at 0.0003. The initial values of the connective weights were set with random values ranging from -0.1 to 0.1.

In training mode, predicted values of m_{t+1} serve as virtual sensory feedback for the next time step, instead of sensory feedback ($\overline{m}_{t+1}$) from actual robot movements. In the process of this closed-loop training, error between generated sequences and teaching signals sometimes grow too large to estimate the gradient of the error landscape. (The target sensory-motor state $\overline{m}_{t+1}$ is also incorporated into the predicted values of m_{t+1} as in Eq. (8) to avoid this problem in learning

$$m_{i,t+1} = 0.99 m_{i,t+1} + 0.01 \overline{m}_{i,t+1} \tag{8}$$

1.3.3 Additional novel training sequences

During the additional training after the basic training, which is to learn a set of action primitives, only the connections in the slow dynamic units and those from the slow dynamics units to the fast ones were allowed to change. The other connective weights were fixed at the same values acquired in the basic training. The initial states of the slow dynamics units were set to different values from those of the basic action patterns. The slow dynamic units remembered the sequences of their primitive behaviors that were recalled by the fast dynamics units. Therefore, we could show how the model could generate the combination of basic actions with novel object positions, sizes, and colors using the previously learned action skills.

2. Experiments and Results

2.1 Task design

A small humanoid robot, namely HOAP3, plays the role of a physical body interacting with the actual environment. A table was set in front of the robot where a fixed pedestal attached to a green (G) sheet was placed, as shown in Fig. 2 (a). The robot was supposed to displace two different size objects, 6x8x6 cm^3 red (R) object and 6x10x6 cm^3 blue (B) object between the top of the table and a pedestal on the table. The robot was given a demonstration on how the objects should be manipulated. For example, the red object in the basement was placed on top of the blue object on the pedestal. There are four different height levels in our experimental environment; (1) object in the basement, (2) one object on the other in the basement, (3) object on the pedestal box and (4) one object on the other in the pedestal as shown in Table I and Fig. 2 (a). These sorts of level differences were introduced in the robot workspace due to the limited manipulation capability of the robot with multiple objects on a flat workspace. The position of the object on the basement as well as that of the green sheet on the pedestal, can be varied within an 8 cm range from left [L] to right [R] in Fig. 2 (b). In the current convention, an object located in the first, second, third, and forth level denote that the object is in the basement, on an object placed in the basement, on the green sheet attached to the pedestal, and on an object placed on the pedestal as in Fig. 3 (a), respectively. It should also be noted that the exact robot arm posture for holding these two objects differs due to the size difference between these two objects.

In the current experiment, the robot was initially trained for three types of basic displacement actions, with all possible combinations for the object position variations of left [L], center [C], and right [R] in the source and the destination, as shown in Fig. 2 (b). The test was conducted to regenerate them, after training the basic action sets. Then, the experiment was further conducted for generalization tests. In the generalization tests, the robot was trained for two additional types of actions that share some action primitives that appeared in the basic actions, but needed to be combined in different ways to achieve the novel task. Not all possible combinations of the position variations were trained in this additional training. Moreover, only the synaptic weights for the slow dynamics network were trained in generating a new sequence utilizing previously trained behaviors stored in the fast dynamics network. We expect to achieve this by the generalization capability of the MTRNN. Only one teaching sequence was used to train each action type, and learning by the BPTT algorithm was iterated for $5\text{x}10^3$ training epochs.

Table I shows the experimental conditions of the robot tasks. All the basic actions are shown in Figs. 3 (a), (b), and (c) where basic action I is to displace a blue object placed on a green sheet, basic action II is to displace a blue object placed on the pedestal onto a red one in the basement, and basic action III is to displace a red object located on the basement onto a blue one located on the pedestal. The sequence patterns for the visual attention shifts to be learned were provided for each action by an experimenter. Training of each basic action, accompanied with the visual attention shifts, was repeated for nine different positions under the physical guidance of the arm movement trajectories by the experimenter.

The additional action IV, for the generalization test, was to displace a red object located on a green sheet onto a blue object located on the basement, as shown in Fig. 3 (d). Although this action seems to share a similar motor profile with basic action II, there is a slight difference in the arm posture to grasp different objects in the same position. This action was trained for six out of the nine possible object position variations; the remaining three unlearned position variations were used for the generalization test. The additional action V was a sequential combination of part of basic action II and action I, with the bold rectangular in Fig. 4, in which the blue object located on the pedestal is moved onto the red object located on the basement, and then moved back on the green sheet located on the pedestal, as shown in Fig. 4. The dashed rectangular in Fig. 4 is not used to generate additional action V. This action was trained for eight out of 27 possible object position variations and the remaining 19 unlearned position variations were used for the generalization test. The number of time sequences of basic actions I, II, III and the additional action IV is the same, but the additional action V has 1.4 times the number of sequences for basic actions by experiment trials.

2.2 Results

Each action generation was tested with every possible combination of object positions (left [L], center [C], and right [R]) in an origin and in a destination. If the goal-directed task for the robot is to place the source object located in the left basement to the place on the right destination located on the right pedestal, we called it the "action of [LR]". Performance was scored in terms of a success rate across all trials for each task. A trial was considered as "success" if the object was moved to the desired destination within the range of 2 cm. We applied principal component analysis (PCA) to visualize the different neural activities of context units in the network during execution of behavioral tasks. We used different data sets for every behavior and at every position to reconstruct the PCs. 70 dimensional vectors, made up of fast context units, and 30 dimensional vectors, made up of slow context units, at each time were separately used to construct the transformation vectors PCs. After the calculation of the PC conversion vectors, basic behavior sequences and additional novel behavior sequences were separately transformed using calculated PCs.

2.2.1 Basic actions I, II and III

Fig. 5 shows the result of basic task II displacing the blue object located on the left of the pedestal onto the red object located on the right of the basement (basic action II of [LR]). The teaching signals and the actual signals generated by the robot trial are shown as paired for the proprioception (m^*, m) as arm joint angles, the vision sensation (s^*, s) as camera head position and the visual attention command (v^*, v) of the top six rows. Proprioceptive signals were plotted using four values consisting of two left and right arm joint angles out of the entire eight arm motor joint angles. In the vision sensation, the bold dash-dot line represents real head x-directional joint angle and the solid line represents real head y-directional joint angle. If the head x-directional joint angle value exceeds 0.5, the robot attended to the left or otherwise the robot attended to the right from initial postures. If the head y-directional joint angle value exceeds 0.5, the robot attends upward or otherwise the robot attends downward from initial postures. In the visual attention command, the bold line is for the blue color effect, the solid line is for the green color effect, the bold dash-dot line is for the red color effect, and the dashed line is for the default color effect. The seventh and eighth rows show the activation profiles for the fast dynamics units and the slow dynamics units by PCA with a 1st principal component (PC) to a 4th PC, respectively. The detailed robot behavior is as follows:

Initially, the robot was set to home position with a neutral visual attention command (no color to attend). The MTRNN simultaneously predicts a visual attention command (which color to attend) and arm proprioceptive value for the next time step t+1 while receiving the input vectors including visual attention command, arm joint angles, and head directional angles of the current time step t. The MTRNN predicts the next visual attention command that is input to the visual guiding system to control the robot head direction. The target arm joint angles, which are the predicted proprioceptive state, are used as an input to the robot's PID controller to generate the arm movement. Let us look at the behaviors generated by the robot as shown in Fig. 5, precisely. (1) From the initial step to 25th step: according to the visual attention command, the robot camera detects and attends to a target object between two objects. (2) From the 26th step to 30th step: after encoding the location of the target object, the target object is grasped by robot hands, in which the robot camera is still attending to a target object. (3) From the 31st step to 50th step: to encode the destination of the target object, the robot camera detects and attends to the destination for the target object with suitable arm behavior. (4) From the 51st step to 90th step: the robot arms move the target object to the destination place, in which the robot camera is still attending to the destination place. Preserving the visual attention during the robot arm movement can accurately guide the arm behaviors to achieve a specific action by considering both the object coordinates and the robot postures. (5) From the 91st step to the final step: finally, the robot goes back to the home position. Here it can be observed that all dimensions of teaching signals are reconstructed in the generation process, only with minimal errors. In addition, it can be seen that the profiles of the fast dynamic units after the PCA contain more complex patterns compared to those in the slow units. This might occur, because the activities in the fast dynamics units are self-organized, such that they are responsible for reconstructing details of proprioceptive trajectories, and the visual attention command sequences, by utilizing their fast timescale dynamics, as have been shown in the prior studies (Yamashita & Tani, 2008).

A two-dimensional vector using the 2nd and 4th PCs of the PCs was plotted in Fig. 6 for basic actions and additional action IV at every position in order to visualize the state changes in the network during execution of behavioral tasks effectively. Figs. 6 (a) and (b) represent the fast dynamics and slow dynamics activities with PCA using the 2nd PC and 4th PC axes, because this vector space most efficiently visualizes the neural dynamics. Nine different types of trajectories represent different object position cases. As shown in Fig. 6 (a), it is reasonable to consider that the difference of the slow dynamics between I and III was caused by the difference of the visual attention sequences, because motor behavior is almost the same but visual attention shift sequences are significantly different. However, it is difficult to dissociate the effects of visual attention and motor behaviors for representing context units due to the distributed representation characteristics of MTRNN. This result implies that the internal representations for visual attention and motor behaviors are seamlessly integrated in the neural activities in the model which accords with the aforementioned idea of premotor theory for visual attention. Finally, it can be observed that there is a smaller variance in the slow dynamics trajectories, depending on the object positions for each action, whereas the variance becomes larger in the fast dynamics trajectories. This means that the slow dynamics network in the higher level is successful in categorizing action primitives regardless of differences in object positions, while the fast dynamics network in the lower level turns out to be sensitive to the position differences in exact manipulation of the objects. This is a signature of the top-down control by the slow dynamics network on the fast dynamics network by means of parametric bifurcation and modification, as will be detailed in the later section. Although similar observations were made in prior studies (Yamashita & Tani, 2008; Nishimoto et al., 2009), this result confirms that functional hierarchy of segregating the sensory-motor processing level and the action primitive manipulation level can be self-organized even in tasks of complex skilled action learning that involve the integration of visual attention control and sensory-motor control, such as shown in the current example.

Table II summarizes learning errors and performance for the basic robot actions. The robot could efficiently reproduce the entire collection of learned basic behaviors by interacting with the real environment. All the basic actions are simultaneously generated by one network which has a learning error of 0.003631 between the teaching and output sequences, as calculated by the Kull-back-Leibler divergence (Yamashita & Tani, 2008). Additionally, we examined several trials for each action by placing the target object at arbitrary points between the left and the right location of the trained positions. It turned out that the robot can perform the tasks successfully with more than a 95% success rate. This indicates that the robot achieved the position generalization for each object to be manipulated via learning.

2.2.2 Additional action IV

As shown in Fig. 6 (a), the fast dynamics unit activation of basic action II and the additional action IV are similar patterns, and basic action I and III are similar shape of trajectories because they represent the same upward movement of the object. However, fast dynamics unit activation of basic action II and additional action IV have less similar patterns compared to basic action I and III. This is caused by the untrained behavior patterns in action IV. As shown in Fig. 6 (b), slow context units have small within class variance of each action but quite different characteristics in each action that are caused by differing arm behaviors, together with dissimilar visual attention commands, between the two actions.

It was expected that additional action IV could be learned with positional generalization, even with a partial set of training examples, for position variances because the action is similar to basic action II of the previously learned action, but with a target object of different size and color. Table III summarizes the two experimental cases with a different number of training behavior patterns. In the case of the first experiment with four trained positions out of all nine possible positional combinations, it successfully generated half of the behavior patterns in trained sessions and one in the behavioral patterns of the untrained session. The lower success rates for behavioral generation are mainly caused by an insufficient amount of training data. When the amount of training data is increased to six behavior patterns, all the behavior patterns including both the six trained cases and the three untrained cases can be successfully generated. It was discovered that the number of training examples in additional action IV was an important factor to effectively design the dynamic neural networks. Fig. 7 shows a comparison between basic action II of [CR] case and that of the untrained additional action IV of [CR] case. From the teaching signal, as shown in Fig. 7 (a), the proposed model generated exact proprioceptive, visual sensory information, and visual attention command as shown in Figs. 7 (b) and (c). As shown in Figs. 7 (b) and (c), those two actions have different visual attention shifts, which are blue to red attention shifts for basic action II and red to blue attention shifts for additional behavior IV, and slightly different arm behaviors caused by different object sizes. The model network can generate exact arm behavior sequences with attention shifts on time even through untrained training samples, as additional behavior IV of [CR], as shown in Fig. 7 (c).

The test for additional action IV shows that the novel behaviors with different object features such as sizes, colors and positions compared to those in basic actions are successfully re-generated by the MTRNN utilizing previous experiences of learning the basic behaviors.

2.2.3 Additional action V

In experiment V, we tested whether the proposed model can generalize a combination of two different basic actions using a previously learned primitive behavior. It was expected that additional action V could be learned with a combination of basic actions with novel object positions, composed of basic action II and I, of the previously acquired actions. The basic actions II and I are to grasp the blue object after attending to the blue one and then move it onto a red object and a green sheet, respectively. When those two actions are combined into a new task, which means that the robot action is to find a blue one and move it onto red one and then move it on the green one successively, the robot does not need to find the blue object to achieve basic action I (move a blue object on a green sheet) as shown in Fig. 4, because it has already attended to the blue object from the first basic action II (move a blue object on a red target).

We considered two experimental conditions with a different number of trained behavioral patterns, as shown in Table IV. In the first experiment with four trained behavior patterns out of all 27 possible positional combinations, almost half the behavioral patterns for both the trained and untrained cases could be successfully generated. Using more trained patterns, such as eight behavior patterns; it was found that 15 untrained object position cases could be successfully generated, in additional training sessions, along with eight trained cases with a 0.003678 value of learning error. Even though the visual attention sequences skip the attention shift to the blue object in the middle stage between basic action II and I, the additional action V is successfully achieved by utilizing the primitive behaviors acquired in the basic actions. It was observed that four unsuccessful cases resulted from just a slight position error in placing an object at the destination. In the same manner, with additional action IV experiment, it was found that the number of training examples is also an important factor to successfully generate the desired behaviors.

Fig. 8 shows the results of reproducing the performance with a novel combination of actions, as with the example of the additional untrained action V. Fig. 8 shows the actual sensory feedback in the physical environment of (a) basic action II of [RC], (b) basic action I of [CL], and (c) additional action V of [RCL], in which this action displaces the blue object located to the right [R] of the pedestal onto the red object located at the center [C] in the basement and then place the blue object onto the green sheet located to the left [L] on the pedestal; this is a sequentially combined basic action II of [RC] and a basic I of [CL]. As shown in Fig. 8, the additional action V of [RCL] is successfully generated by efficiently combining the two basic actions. Actually, it can be seen that the profiles of the proprioception, as well as fast and slow dynamics of additional action V of [RCL], are mostly similar to the two basic actions; (1) between the initial step to the 80th step, with basic action

II, in Fig. 8 (a), and from the initial step to the 80th step with an additional action V, in Fig. 8 (c), when the object was grasped and then placed onto the red object, (2) between the 40th step to the final step with basic action I, in Fig. 8 (b), and from the 81st step to the final step with the additional action V, when the object was placed at the destination; then, the robot returned to its initial postures. The visual attention command and hand postures were successfully generated by the network, considering a top-down visual attention shift sequence, to achieve the task. The experiment of additional learning of action V shows that the novel behaviors are efficiently re-generated by adopting the previously learned actions. In this task, the MTRNN was able to make a plan to achieve untrained behavior patterns of additional task V using a part of the basic actions II and I without introducing specific supervisions.

3. Discussion

3.1 Correspondence to empirical studies

We consider that the current model corresponds to the fronto-parietal network in the cortical brains that which has been considered essential in integration of sensory-motor processing and visual attention (Corbetta, 1998; Rizzolatti & Craighero, 1998). The visual attention discussed in the current robotics task can be categorized as object-related visual attention (Craighero et al., 1999; Schubotz & Cramon, 2002) in which some properties of encountering objects, such as shape, size or color, are anticipated. Schubotz and Cramon (2001) found a fronto-parietal network comprising the pre-supplementary motor area, the ventral premotor cortex, and the left anterior intraparietal sulcus to be activated independently of the attended stimulus property, but most intensively during object-related attention. Particularly, it was observed that left superior ventrolateral premotor cortex was activated in an object related visual attention task in contrast to the observation that dorsolateral premotor cortex was activated in a spatial visual attention task and frontal opercular cortex in a timing task.

Related evidence was demonstrated by Jellema and colleagues (1994) in the recording of Superior Temporal Sulcus (STS) of monkeys while the monkeys observe human experimenters to reach objects. They described the unique response characteristics of two distinct populations of the STS cells. The first population of cells responded to particular head views and gaze directions of the experimenters. The second population of cells responded to reaching movements of the experimenters' hands. An interesting finding was that the activities of the second population for encoding perception of reaching behavior were enhanced when the gaze of experimenters was directed to the object. This finding suggests that, in observation of other's behaviors, some STS cells combine visual attention of others to objects and their particular behaviors toward the objects into meaningful concepts. Furthermore, it is reasonable to assume that mirror neurons in the premotor cortex (F5 in monkeys) could have similar properties, as it has been known that the STS provides large inputs to F5 in the premotor cortex in monkeys through the inferior parietal lobe (Seltzer & Pandya, 1994). Of particular relevance was the observation of an F5 cell that responded to the observation of an experimenter grasping an object when the experimenter was looking at the object but not when the experimenter looked away from the object (L. Fogassi, personal communication to Jellema et al.). This result implies a possibility that some premotor cells encode meaningful concepts for object manipulative actions both for own generation and observation of the same actions by others by combining visual attention (of self or other) and corresponding motor behaviors. It would be intuitive to think that an action primitive, such as approaching an object, should be one package made of tight coupling of visual attention control to the object and behavior control for arm reaching to the object. What we propose here is that such packaging into action primitives might be undertaken in the premotor cortex, in terms of premotor theory for “object-related” visual attention.

Based on the above discussions, we examine how the proposed model can correspond to findings in cognitive neuroscience and neurophysiology. An essential assumption is that MTRNN might correspond to the premotor-parietal network in a broad sense in which the slow dynamic network might correspond to the premotor cortex and the fast dynamics network to the parietal cortex. First, it is considered that an intention to initiate an action might originate in the prefrontal cortex, as the prefrontal cortex has been known as a distinguished source to generate goal-directed actions (Fuster, 2009). In the current model, an intention for action is represented by the initial state in the slow dynamic network that is assumed to be provided from the prefrontal cortex.

By receiving the intention for action, the premotor cortex dispatches action primitives, related to visually-guided object manipulation, with organizing their adequate sequences to accomplish the intended action. By taking account of the observation that the mirror neurons in the premotor cortex tend to exhibit relatively stable tonic firing during performance of the corresponding action reparatory such as reaching towards an object or grasping an object (Rizzolatti et al., 1996), it is presumed that their rate coding level activities may correspond to the dynamic property in the neural units in the slow dynamics network. One central assumption here is that the premotor cortex should deal with both sensory-motor processing and object-related visual attention inseparably, as discussed in the premotor theory for visual attention. Results of our preliminary experiments suggest that this inseparability can be observed in the model. As shown in the Fig. 1, the MTRNN receives the current visual attention command as the inputs from the output feedback. In the experiment, when this channel of the inputs was deleted, the prediction outputs for the next step proprioceptive state were significantly disturbed. This implies that the sensory-motor processing (prediction of the proprioceptive state in the current context) and proactive visual attention are tightly coupled in the distributed representation of the neural activities in the model.

The premotor cortex may interact mainly with two regions, one for the frontal eye field (FEF) and the other for the parietal cortex. The FEF may implement both overt and covert visual attention by receiving the visual attention command from the premotor cortex. This part was implemented in the current model by connecting the output of the visual attention command to the visual guiding system of an external device. However, in the current implementation, the visual attention command is output not from the slow dynamics network but from the in-out network. This treatment was necessary, because the visual attention shift requires exact synchrony with on-going manipulative behavior (with stepwise sharp shifts) by receiving real time information of the visuo-proprioceptive state. Therefore, the visual attention command should be output from the in-out network having direct interaction with the sensory inputs, even though the cause of visual attention shift mostly originated in the slow dynamic network of mimicking the premotor cortex. This may account for why visual attention is generated cooperatively or in complementary ways between endogenous controls in the frontal cortex (Connolly et al., 2002) and exogenous control in the parietal cortex (Corbetta & Shulman, 2002).

Nishimoto et al. (2009) assumed that the visuo-proprioceptive sequences are anticipated in the inferior parietal lobe (IPL) by receiving abstract information about the current context of the intended action or more specifically, the currently dispatched action primitive from the premotor cortex. This assumption corresponds to the recent observations suggesting that the IPL may play the role of forward prediction for multi-modal perceptual sequences (Ehrsson et al., 2003; Eskandar et al., 1999; Mulliken et al., 2008). This idea has been implemented in

MTRNN by allowing reciprocal interactions between the slow dynamics network (corresponding to the premotor cortex) and the fast dynamics network (corresponding to the IPL). The fast dynamics network can predict change of the visuo-proprioceptive state with precise timing due to its given fast dynamic property. The proprioceptive state predicted for the next time step is assumed to be sent to the primary motor cortex and to the cerebellum as the target proprioceptive state to be achieved. The inverse model assumed in the cerebellum (Wolpert & Kawato, 1998) may compute the corresponding motor commands, i.e., required motor torque to achieve a target joint angle in the robot implementation. A modification in the current renovated model is about the treatment of the visual perception and prediction in the fast dynamics network. The current model receives the head direction that represents the position of the currently attended object in the egocentric view as visual inputs, and it predicts the feature of the next attended objects, i.e., color of the object.

3.2 Generalization and compositionality

The current model which is inspired by the idea of the fronto-parietal network (Corbetta, 1998; Rizzolatti & Craighero, 1998), demonstrated both characteristics of generalization and compositionality in learning complex goal-directed skilled actions that require integration of object-related visual attention and sensory-motor control. A distinct characteristic of the current model which has been found through the synthetic neuro-robotics experiment is that proactive visual attention and sensory-motor control are seamlessly integrated in the distributed neural activities appearing in the reciprocal interactions between the slow dynamics network (corresponding to the premotor cortex) and the fast dynamics network (corresponding to the parietal cortex). This characteristic, corresponding to the premotor theory for visual attention (Craighero et al., 1999; Rizzolatti et al., 1987), can allow precise adjustments and coordination between the two processes including timing control between visual attention shifts and the starting of arm movements and the spatial coordination between the direction of visual fixation and target position for arm movement. Those details should be acquired through generalization in learning from various experiences and practices in object manipulation, because the details of those can vary and be complex, depending on the size and position of objects and exact sequences of action primitives. Our robotic experiment results showed that distributed neural representation of coupling these two processes emerged in the results of such generalization through iterative learning. It was also shown that the robot can achieve object manipulative actions for untrained position cases by generalization of learned skills.

Another distinct feature observed in the presented model is that behavioral compositionality is achieved by self-organizing functional hierarchy, which is inherit from our prior studies (Nishimoto et al., 2009; Yamashita & Tani, 2008). The current result showed that integration of object-related attention (Craighero et al., 1999) and sensory-motor control for manipulating objects appeared in the fast dynamics network as reusable action primitives. Conversely, dynamic functions for sequential manipulation of those action primitives appeared in the slow dynamics network. However, it is noted that the hierarchical manipulation of the action primitives from the higher level to the lower level is not mechanized via symbolical manipulation and on-off type dispatching of the primitives but by parametric interactions between two dynamic networks of different timescales. More specifically, bifurcations in the fast dynamics network caused by parametric interactions from the slow dynamics network (in terms of neural activation inputs) enable a shift from one primitive to another. This parametric control from the slow dynamics network to the fast dynamics network also enables precise adjustments of on-going behaviors. For example, motor control for approaching and grasping an object is adjusted by means of proactive attention to object features, either a smaller object (the red object in the current example) or a larger one (the blue one). One interesting observation in Figs. 7 and 8 is that the activities of slow dynamics units change continuously and smoothly, even in the discrete events of visual attention switches. It is highly speculated that this smoothness enables fluent shifts from one action primitive to another seamlessly. It is considered that a successive smooth connection between one primitive and another require fine adjustments between the “tail” of the foregoing primitive and the “head” of the subsequent one. Such fine adjustment could be done again by the parametric modulation of those successive action primitives from the higher level control. The experiment about additional learning of action V by smoothly connecting two of the previously learned action primitives would demonstrate this characteristic. This task was successfully achieved by allowing synaptic changes only in the slow dynamics network; this means that control by the slow dynamics network on the fast one, via parametric bifurcation and modulation achieved this.

Therefore, highly sophisticated actions of manipulating multiple objects with accompanying proactive visual attention shifts are considered to require compositionality on one side, which is more like a symbolical process, and generalization on the other side, which is more like an analogical process. The presented model of the renovated MTRNN can satisfy these two seemingly conflicting requirements, utilizing two essential characteristics of distributed representation and multiple timescales dynamics in its neural activities.

3.3 Related studies

McCallum proposed rule based reinforcement learning using selective attention and short-term memory, with similar experiments to teach the visual attention shift sequences for goal-directed behaviors (McCallum, 1996). They defined the states of environments (world state space and sensory state space) taught then by policy, which consists of if-then-else tree statements, in the model, to avoid the obstacle in the path. In contrast to this rule based reinforcement model, our model can automatically generate the actions, associating visual attention and motor behaviors, except providing adequate policy to manipulate the object using supervised teaching.

Suzuki and Floreano showed a similar neural architecture to the current study that agents can learn in a scheme of active vision by switching attention with a small image size retina for goal-directed behaviors using a genetic algorithm (Suzuki et al., 2008). The scheme was successfully evaluated by navigation experiments using real mobile robots as well as a humanoid robot. A limitation in Suzuki's approach comes from the neural architecture, as it must be carefully designed for each task (Suzuki et al., 2008). In contrast, the proposed model can share the previously learned primitive behaviors and generate combined actions, based on the primitive behaviors.

3.4 Future studies

We envision four directions for future studies to construct autonomously operating robots considering active environments; (1) we will add the object recognition function, utilizing texture, depth, and appearance of the object, to generate the complex top-down attention using high-level visual cognition to achieve the high-level object manipulation task. (2) We will introduce more diversity and complexity of actions, rather than current simple ones such as just placing objects. For example, we will consider how robots can acquire skills for tool usage actions that would require more complex spatio-temporal associations between the visual attention and behavior generations. (3) We will introduce a reinforcement learning paradigm to acquire the attention shift skills, as inspired by McCallum's model.

4. Conclusion

The main contribution of the paper is to present a cortical model of the fronto-parietal network that accounts for integrative learning of proactive visual attention shifts and sensory-motor control by extending our prior proposal of MTRNN. The model was evaluated by neuro-robotics experiments in tasks of multiple object manipulation. The experimental results show that some extents of generalization, in terms of position and object size variances, can be achieved by organizing seamless integration of the visual attention and the sensory-motor control in the distributed neural activities in the model network. Furthermore, it was shown that additional learning of combining prior learned actions can be efficiently achieved, because a functional hierarchy was developed by acquiring a set of action primitives with seamless integration of visual attention and sensory-motor control via multiple timescales properties employed in the network model. These accounts correspond to an idea of the premotor theory for object-related visual attention discussed in cognitive neuroscience literature.

Acknowledgments

This research was supported by the Ministry of Education, Culture, Sports, Science and Technology of Japan, Grants-in-Aid for Scientific Research on Innovative Areas (No. 22120523, 2010) (50%), the Korea-Japan Collaboration Core Project on Neuroinformatics (10%) and the Converging Research Center Program (40%) through the National Research Foundation(NRF) of Korea funded by the Ministry of Education, Science and Technology (2009-0082262).

References

Ballard, D. H. (1991). Animate vision, Artificial Intelligence Journal, 48, 57-86, 1991.

Ballard, D. H., Hayhoe, M. M., & Pelz, J. B. (1995). Memory representations in natural tasks. Journal of Cognitive Neuroscience, 7, 66–80.

Ballard, D. H., Hayhoe, M. M., Li, F., Whitehead, S. D., Frisby, J. P., Taylor, J. G., & Fisher, R. B. (1992). Hand-Eye Coordination during Sequential Tasks [and Discussion], Philosophical Transactions: Biological Sciences Royal Society, 337, 331-339.

Beer, R. D. (2000). Dynamical approaches to cognitive science. Trends in Cognitive Sciences, 4(3), 91-99.

Berkinblit, M. B., Fookson, O. I., Smetanin, B., Adamovich, S. V., & Poizner, H. (1995). The interaction of visual and proprioceptive inputs in pointing to actual and remembered targets. Exp. Brain Res., 107, 326–330.

Binkofski, F., Buccino, G., Posse, S., Seitz, R. J., Rizzolatti, G., & Freund, H. –J. (1999). A fronto-parietal circuit for object manipulation in man: evidence from an fMRI-study. European Journal of Neuroscience, 11(9), 3276-3286.

Bowman, M. C., & Johannson, R. S. (2009). Eye-hand coordination in a sequential target contract task, Exp. Brain Res, 195, 273-283.

Carlton, L. G. (1981). Processing visual feedback information for movement control. J. Exp. Psychol. Hum. Percept Perform, 7, 1019–1030.

Carpenter, G. A., Grossberg, S., Markuzon, N., Reynolds, J. H., & Rosen, D. B. (1992). Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps. IEEE Transaction on Neural Networks, 3(5), 698-713.

Choi, S. B., Jung, B. S., Ban, S. W., Niitsuma, H., & Lee, M. (2006). Biologically motivated vergence control system using human-like selective attention model. Neurocomputing, 69, 537-558.

Connolly, J. D., Goodale, M. A., Menon, R. S., & Munoz, D. P. (2002). Human fMRI evidence for the neural correlates of preparatory set. Nat. Neurosci., 5, 1345-1352.

Corbetta, M. (1998). Frontoparietal cortical networks for directing attention and the eye to visual locations: Identical, independent, or overlapping neural systems? Proc. Natl. Acad. Sci., 95, 831-838.

Corbetta, M., Akbudak, E., Conturo, T. E., Snyder, A. Z., Ollinger, J. M., Drury, H. A., Linenweber, M. R., Petersen, S. E., Raichle, M. E., Van Essen, D. C., & Shulman, G. L. (1998). A common network of functional areas for attention and eye movements. Neuron, 21, 761-773.

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain, Nature reviews: neuroscience, 3, 201-215.

Craighero, L., Fadiga, L., Rizzolatti, G., & Umiltà, C. (1999). Action for perception: a motor-visual attentional effect. J. Exp. Psychol. Hum. Percept. Perform., 25, 1673-1692.

Crawford, J. D., Medendorp, W. P., & Marotta, J. J. (2004). Spatial Transformations for Eye-Hand Coordination, J. Neurophysiol, 92, 10-19.

Doya, K., & Yoshizawa, S. (1989). Memorizing oscillatory patterns in the analog neuron network, in proceedings of 1989 international joint conference on neural networks, I:27–32, Washington, D.C..

Ehrsson, H., Fagergren, A., Johansson, R., & Forssberg, H. (2003). Evidence for the involvement of the posterior parietal cortex in coordination of fingertip forces for grasp stability in manipulation. Journal of Neurophysiology, 90, 2978-2986.

Eskandar, E., & Assad, J. (1999). Dissociation of visual, motor and predictive signals in parietal cortex during visual guidance. Nature Neuroscience, 2, 88-93.

Friston, K., Mattout, J., & Kilner, J. (2011). Action understanding and active inference. Biological Cybernetics, 104, 137-160.

Furneaux, S., & Land, M. F. (1999). The effects of skill on the eye-hand span during musical sight-reading, in proceeding of Royal Society, 266, 2435-2440.

Fuster, J. M. (2008). The Prefrontal Cortex (Fourth Edition) Academic Press, London.

Hayhoe, M. M., Shrivastava, A., Mruczek, R., & Pelz, J. B. (2003). Visual memory and motor planning in a natural task. J Vis, 3, 49–63.

Herst, A. N., Epelboim, J., & Steinman, R. M. (2001). Temporal coordination of the human head and eye during a natural sequential tapping task. Vision Res, 41, 3307–3319.

Iacoboni, M. (2009). Imitation, Empathy, and Mirror Neurons, Annu. Rev. Psychol. 60, 653-670

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. IEEE Transaction on Pattern Analysis and Machine Intelligence, 20(11), 1254-1259.

Jellema, T., Baker, C. I., Wicker, B., & Perrett, D. I. (2000). Neural representation for the perception of the intentionality of actions. Brain and Cognition, 44, 280-302.

Jeong, S., Ban, S.-W. and Lee, M. (2008). Stereo saliency map considering affective factors and selective motion analysis in a dynamic environment. Neural Networks, 21, 1420-1430.

Johansson, R. S., Westling, G., Ba¨ ckstro¨m , A., & Flanagan, J. R. (2001). Eye-Hand Coordination in Object Manipulation. The journal of neuroscience, 21, 6917-6932.

Kohonen, T. (2001). Self-Organizing Maps, Springer Series in Information Sciences, 30, Springer.

Kustov, A. A., & Robinson, D. L. (1996). Shared neural control of attentional shifts and eye movements. Nature, 384, 74-77.

Land, M., Mennie, N., & Rusted, J. (1999). The roles of vision and eye movements in the control of activities of daily living. Perception, 28, 1311–1328.

Land, M. F., & McLeod, P. (2000). From eye movements to actions: how batsmen hit the ball. Nat Neurosci, 3, 1340–1345.

Lennie, P. (2003). The cost of cortical computation. Curr. Biol., 13, 493–497.

Linda J. L., & Susan L. D. (2009). Modelling attention in individual cells leads to a system with realistic saccade behaviours. Cogn Neurodyn, 3, 223-242.

McCallum, A. K. (1996). Learning to Use Selective Attention and Short-Term Memory in Sequential Tasks. Proceedings of the fourth International Conference on Simulation of Adaptive Behavior, (SAB'96), 315-324.

Metta, G., Sandini, G., & Konczak, J. (1999). A developmental approach to visually-guided reaching in artificial systems. Neural networks, 12(10), 1413-1427.

Moore, T.,& Fallah, M. (2001). Control of eye movements and spatial attention. Proc. Natl. Acad. Sci., 98, 1273-1276.

Mulliken, G. H., Musallam, S., & Andersen, R. A. (2008). Forward estimation of movement state in posterior parietal cortex. Proc. Natl Acad Sci USA. 105(24), 8170-8177.

Nishimoto, R., Tani, J. (2009). Development of hierarchical structures for actions and motor imagery: a constructivist view from synthetic neurorobotics study. Psychological Research, 73, 545-558.

Nobre, A. C., Gitelman, D. R., Dias, E. C., & Mesulam, M. M. (2000). Covert visual spatial orienting and saccades: overlapping neural systems. Neuroimage, 11, 210-216.

Paillard, J. (1996). Fast and slow feedback loops for the visual correction of spatial errors in a pointing task: a reappraisal. Can J Physiol Pharmacol, 74, 401–417.

Pelz, J., Hayhoe, M., & Loeber, R. (2001). The coordination of eye, head, and hand movements in a natural task. Exp Brain Res, 139, 266–277.

Posner, M. I., & Dehaene, S. (1994). Attentional networks. Trends neurosci., 17, 75-79.

Prablanc, C., Desmurget, M., & Grea, H. (2003). Neural control of on-line guidance of hand-reaching movements. Prog. Brain Res., 142, 155–170.

Prablanc, C., Echallier, J. F., Komilis, E., & Jeannerod, M. (1979). Optimal response of eye and hand motor systems in pointing at a visual target. I. Spatio-temporal characteristics of eye and hand movements and their relationships when varying the amount of visual information. Biol. Cybern. 35, 113–124.

Prablanc, C., & Martin, O. (1992). Automatic control during hand reaching at undetected two-dimensional target displacements. J. Neurophysiol, 67, 455–469.

Prablanc, C., Pelisson, D., & Goodale, M. A. (1986). Visual control of reaching movements without vision of the limb.1. Role of retinal feedback of target position in guiding the hand. Exp Brain Res, 62, 293–302.

Rao, R. P. N. (1995). An Active Vision Architecture based on Iconic Representations. Artificial Intelligence, 78, 461-505.

Rao R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nature neuroscience, 2, 79-87.

Rizzolatti, G., & Craighero, L. (1998). Spatial attention: Mechanisms and theories. In M. Sabourin, F. Craik, & M. Robert (Eds.), Advances in psychological science, 2, Biological and cognitive aspects (171-198). East Sussex, England: Psychology Press.

Rizzolatti, G., Fadiga, L., Galless, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. Cognitive Brain Research, 3, 131-141.

Rizzolatti, G., Riggio, L., Dascola, I., & Umiltá C. (1987). Reorienting attention across the horizontal and vertical meridians: evidence in favor of a premotor theory of attention. Neuropsychologia, 25, 31-40.

Rizzolatti, G., Riggio, L., & Sheliga, B.M. (1994). Space and selective attention. In C. Umiltà & M. Moscovitch (Eds.), Attention and Performance XV (231-265). Cambridge, MA: MIT Press.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. Parallel distributed processing: explorations in the microstructure of cognition, 1, MIT Press, Cambridge.

Saarinen, J., & Kohonen, T. (1985). Self-organized formation of colour maps in a model cortex. Perception, 14(6), 711–719.

Sarlegna, F., Blouin, J., Vercher, J. L., Bresciani, J. P., Bourdin, C., & Gauthier G. M. (2004). Online control of the direction of rapid reaching movements. Exp. Brain Res. 157, 468–471.

Saunders, J. A., & Knill, D. C. (2004). Visual feedback control of hand movements. J. Neurosci. 24, 3223–3234.

Schubotz, R. I., & von Cramon, D. Y. (2001). Functional organization of the lateral premotor cortex: fMRI reveals different regions activated by anticipation of object properties, location and speed. Cognitive Brain Research, 11(1), 97-112.

Seltzer, B., & Pandya, D. N. (1994). Parietal, temporal, and occipita projections to cortex of the superior temporal sulcus in the rhesus monkey: A retrograde tracer study. The Journal of Comparative Neurology, 343(3), 445-463.

Sheliga, B.M., Riggio, L., & Rizzolatti, G. (1995). Spatial attention and eye movements. Exp. Brain Res., 105, 261-275.

Smeets, J. B., Hayhoe, M. M., & Ballard, D. H. (1996). Goal-directed arm movements change eye-head coordination. Exp Brain Res, 109, 434–440.

Suzuki, M. & Floreano, D. (2008). Enactive Robot Vision. Adaptive Behavior, 16(2-3), 122-128.

Tanji, J., & Shima, K. (1994). Role for supplementary motor area cells in planning several movements ahead. Nature, 371, 413-416.

Toth, L. J. & Assad, J. A (2002). Dynamic coding of behaviourally relevant stimuli in parietal cortex. Nature, 415, 165–168.

Treisman, A.M., & Gelde, G. (1980). A feature-integration theory of attention. Cognitive Psychology, 12(1), 97-136.

Tsotsos, J.K. (1990). Analyzing vision at the complexity level. Behav. Brain Sci., 13, 423–469.

Tsubomi, H., Ikeda, T., Hanakawa, T., Hirose, N. Fukuyama, H., & Osaka, N. (2009). Connectivity and signal intensity in the parieto-occipital cortex predicts top-down attentional effect in visual masking: An fMRI study based on individual differences, NeuroImage, 45, 587-597.

Varela, F. J., Thompson, E., & Rosch, E. (1991) The embodied mind, Cambridge, Mass: MIT Press.

Williams, R. J., & Zipser, D. (1989). A learning algorithm for continually running fully recurrent neural networks. Neural Computation, 1(2), 270–280.

Wolpert, D., & Kawato, M. (1998). Multiple paired forward and inverse models for motor control. Neural Networks, 11, 1317-1329.

Yamashita, Y., & Tani, J. (2008). Emergence of functional hierarchy in a multiple timescale neural network model: A humanoid robot experiment. PLoS Computational Biology, 4(11), 1-18.

Yuqiao G., & Hans L. (2007). A neural network model of attention-modulated neurodynamics, Cogn Neurodyn, 1, 275-285.

Table captions

Table I. Experimental conditions for robot tasks

Table II. Error and performance of robot basic behaviors

Table III. Performance of robot behavior with additional behavior IV

Table IV. Performance of robot behavior with additional behavior V

Figure captions

Figure 1. The overall architecture of the proposed model. MTRNN: Multiple Timescales RNN, In-Out: input-output context unit, FAST: fast context unit, SLOW: slow context unit, v_t: top-down visual attention command at time step t (four object categories; red, green, blue and default preference color), s_t: vision sense at time step t through the environment (directing the camera head; there is no eye saccadic movements), m_t: proprioception values at time step t through the environments (arm joint angles vector with eight dimensions), m_{t+1}: predicted proprioception value at time step t, v_{t+1}: predicted top-down visual attention command at time step t, $\bar{s}_{t+1}$: vision sense at time step t+1 through the environment, $\bar{m}_{t+1}$: proprioception values at time step t +1 through the environments.

Figure 2. Workbench for robot experiments. (a) Real environment of workbench, (b) the specification of a workbench

Figure 3. Robot tasks employed. (a) Basic action I, (b) basic action II, (c) basic action III, and (d) additional action IV.

Figure 4. Additional robot task V by integrating basic actions II and I.

Figure 5. Example of robot trials in basic action II of the [LR] case. Proprioception (first and second row), vision sensation information (third and fourth row), visual attention command (fifth and sixth row) of teaching signals (odd row) and actual sensory feedback in physical environment (even row). The activation profiles for the fast dynamics units and the slow dynamics units using PCA with 1st PC axis to 4th PC axis, respectively (seventh and eighth row). In proprioception, four values consist two left and right arm joint angles out of eight motor joint angles were plotted. In vision sense, the bold dash-dot line represents the real head x-directional joint angle and the solid line represents the real head y-directional joint angle. In visual attention command, the bold line is the blue color effect, the solid line is the green color effect, the bold dash-dot line is the red color effects, and the dashed line is a default color effect.

Figure 6. Trajectories of basic actions I, II, III and additional action IV by the fast and slow dynamics units with PCA. (a) Projection results for fast dynamics unit activation with 2nd PC and 4th PC axes. (b) Projection results for slow dynamics unit activation on the 2nd PC and 4th PC axes. Red, green, and blue lines represent the basement object located at the center, left, and right, respectively. Solid, bold, and dashed lines represent the destination area located at the center, left, and right, respectively.

Figure 7. Comparison of the basic action II of the [CR] case and the additional action IV of the [CR] case. (a) Teaching trajectory of basic action II of the [CR] case, (b) Actual sensory feedback in physical environment of basic action II of the [CR] case, (c) Actual sensory feedback in the physical environment of the untrained additional action IV of the [CR] case. The figure legends are the same as for Fig. 5.

Figure 8. Comparison of the basic actions II, I, and the additional action V that is sequentially combined with basic action II and I. (a) Actual sensory feedback in the physical environment of basic action II of the [RC] case. (b) Actual sensory feedback in the physical environment of basic action I of the [CL] case. (c) Actual sensory feedback in the physical environment of an additional action V of the [RCL] case. The figure legends are the same as for Fig. 5.

Table I. Experimental conditions for robot tasks.

	Origin: object color	Destination: object or sheet color	Movement direction	Number of behavior patterns
Basic action I	Blue	Green	From red object onto green sheet	9
Basic action II	Blue	Red	From pedestal onto red object	9
Basic action III	Red	Blue	From basement onto blue object	9
Additional action IV	Red	Blue	From pedestal onto blue object	9
Additional action V	Blue	Red and then green	From pedestal onto red object and then green sheet	27

Table II. Error and performance of robot basic behaviors

# of total behavior patterns: 9	# of trained behavior patterns	Learning error	Success rate (# of success behavior patterns)
Basic action I	9	0.003631	100% (9)
Basic action II			100% (9)
Basic action III			100% (9)

Table III. Performance of robot behavior with additional behavior IV

# of total behavior patterns: 9	# of trained/ # of all possible positional combinations	Learning error	Success rate (# of success behavior patterns)	
			Trained behavior patterns	Untrained behavior patterns
Additional action IV	4/9	0.003672	50% (2)	20% (1)
	6/9	0.003689	100% (6)	100% (3)

Table IV. Performance of robot with additional behavior V

# of total behavior patterns: 27	# of trained/ # of all possible positional combinations	Learning error	Success rate (# of success behavior patterns)	
			Trained behavior patterns	Untrained behavior patterns
Additional action V	4/27	0.003683	50% (2)	49% (11)
	8/27	0.003678	100% (8)	79% (15)

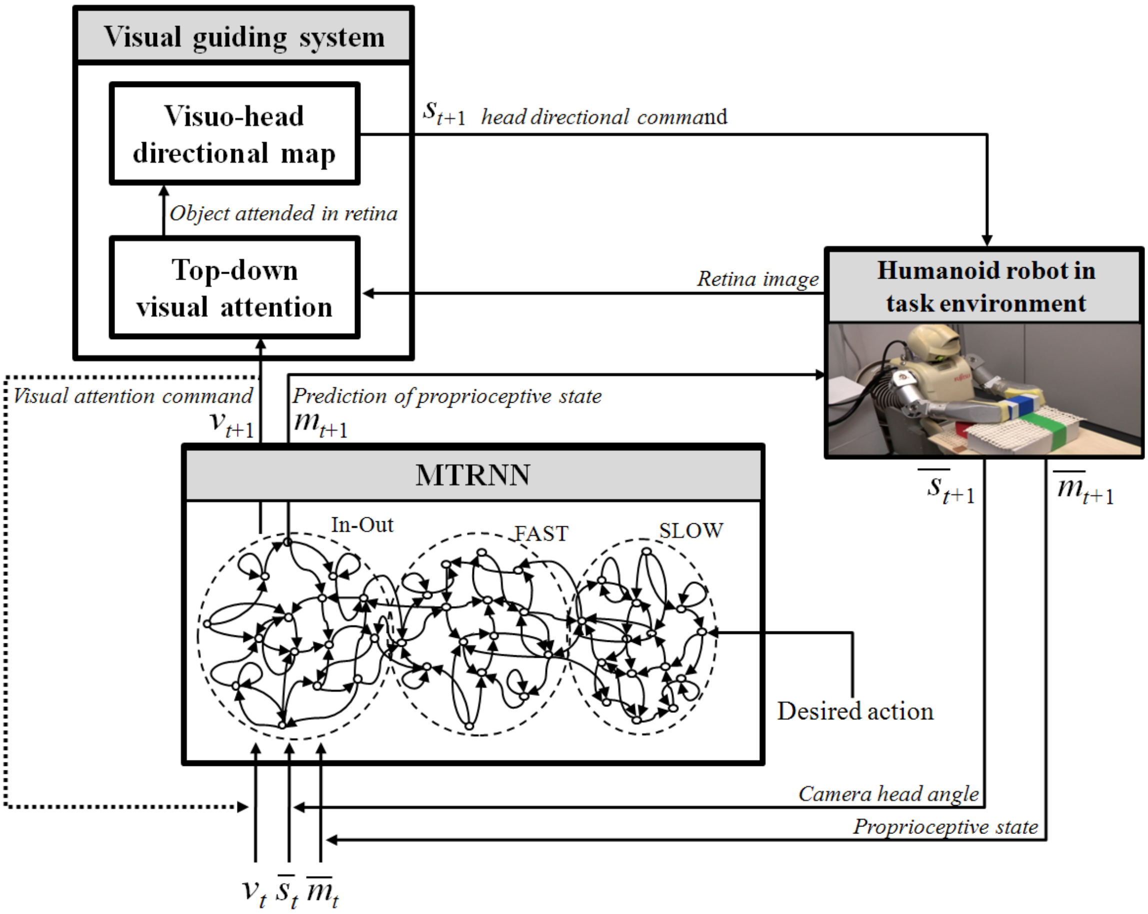
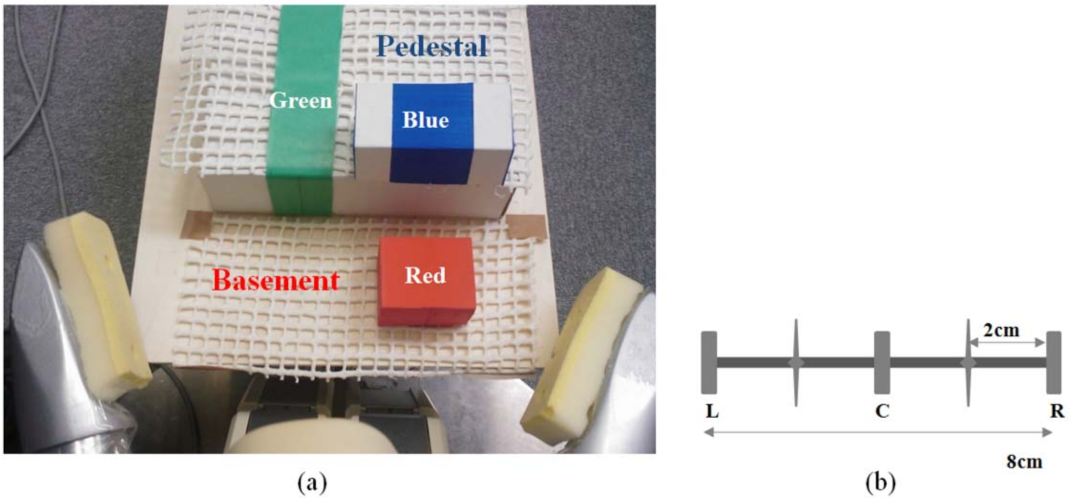

Visual guiding system
Visuo-head directional map
Object attended in retina
Top-down visual attention
s_{t+1} head directional command
Retina image
Humanoid robot in task environment
Visual attention command
v_{t+1}
Prediction of proprioceptive state
m_{t+1}
MTRNN
In-Out
FAST
SLOW
$\overline{s}_{t+1}$
$\overline{m}_{t+1}$
Desired action
Camera head angle
Proprioceptive state
v_t $\overline{s}_t$ $\overline{m}_t$

Fig. 1



(a) (b)

Fig. 2

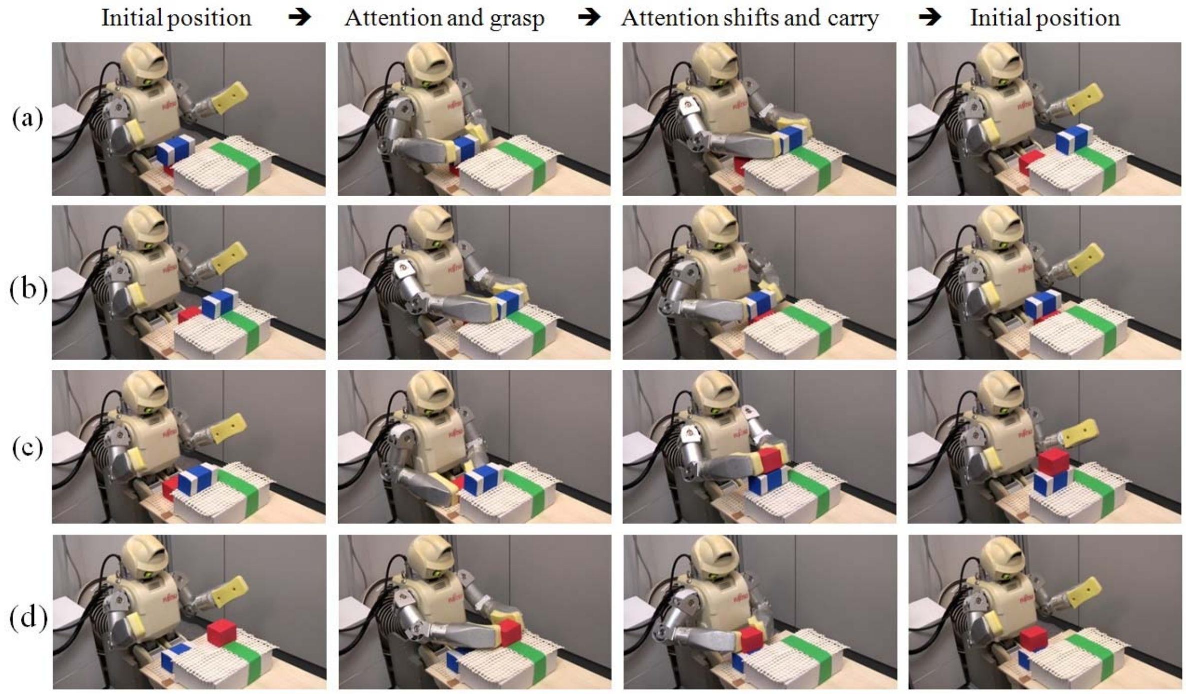

Initial position
➔
Attention and grasp
➔
Attention shifts and carry
➔
Initial position
(a)
(b)
(c)
(d)

Fig. 3

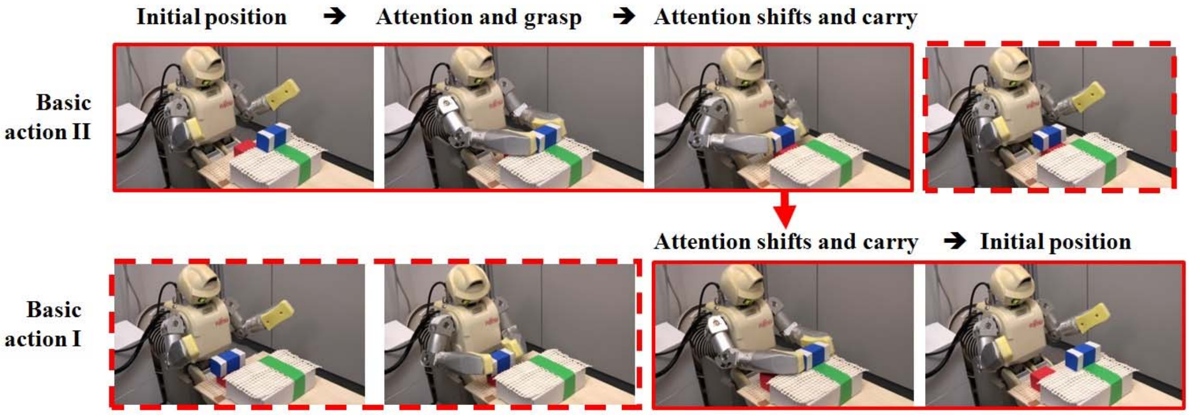

Initial position
Attention and grasp
Attention shifts and carry
Basic action II
Attention shifts and carry
Initial position
Basic action I

Fig. 4

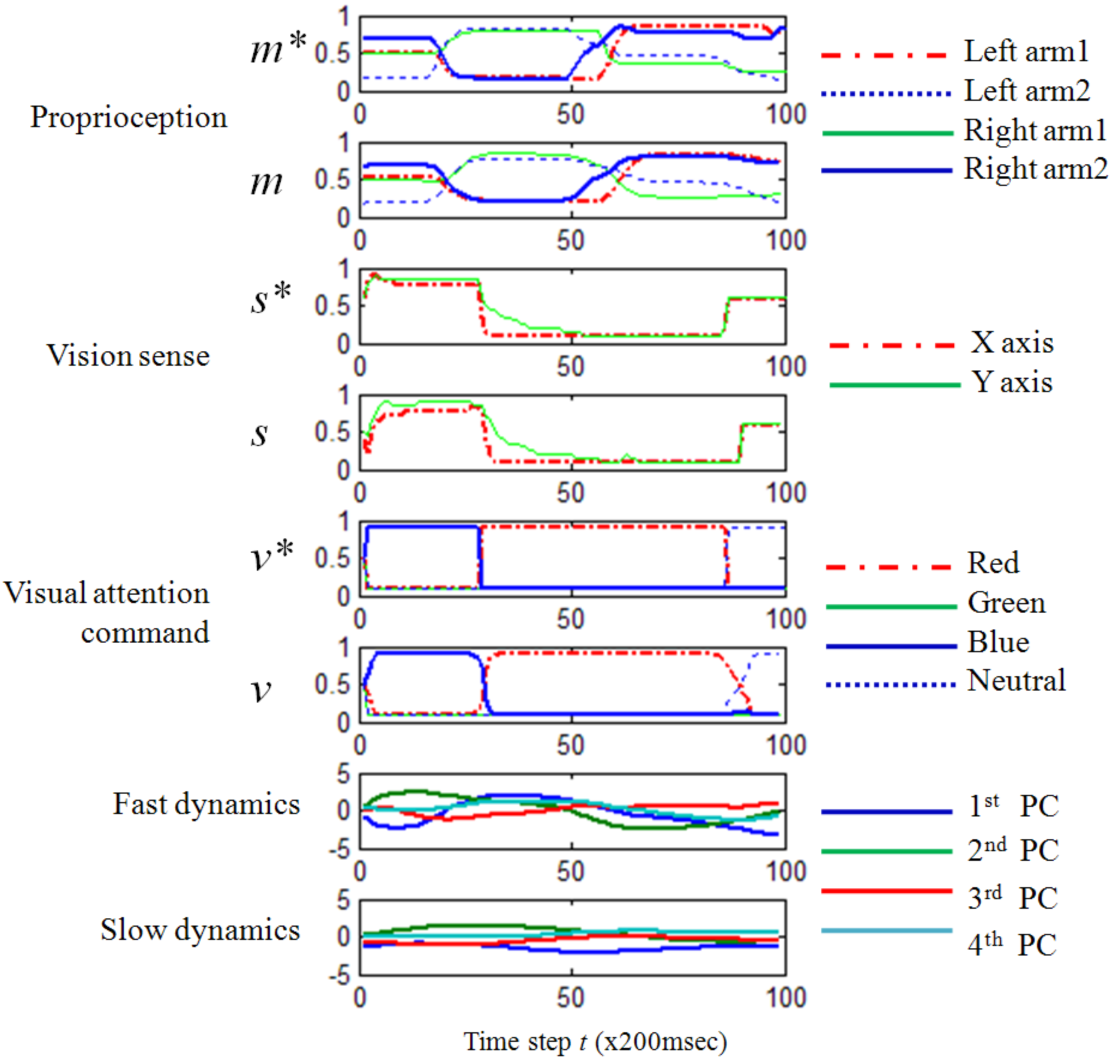

m^*
Proprioception
m
s^*
Vision sense
s
v^*
Visual attention command
v
Fast dynamics
Slow dynamics
1
0.5
0
0
50
100
5
-5
Left arm1
Left arm2
Right arm1
Right arm2
X axis
Y axis
Red
Green
Blue
Neutral
1st PC
2nd PC
3rd PC
4th PC
Time step t (x200msec)

Fig. 5

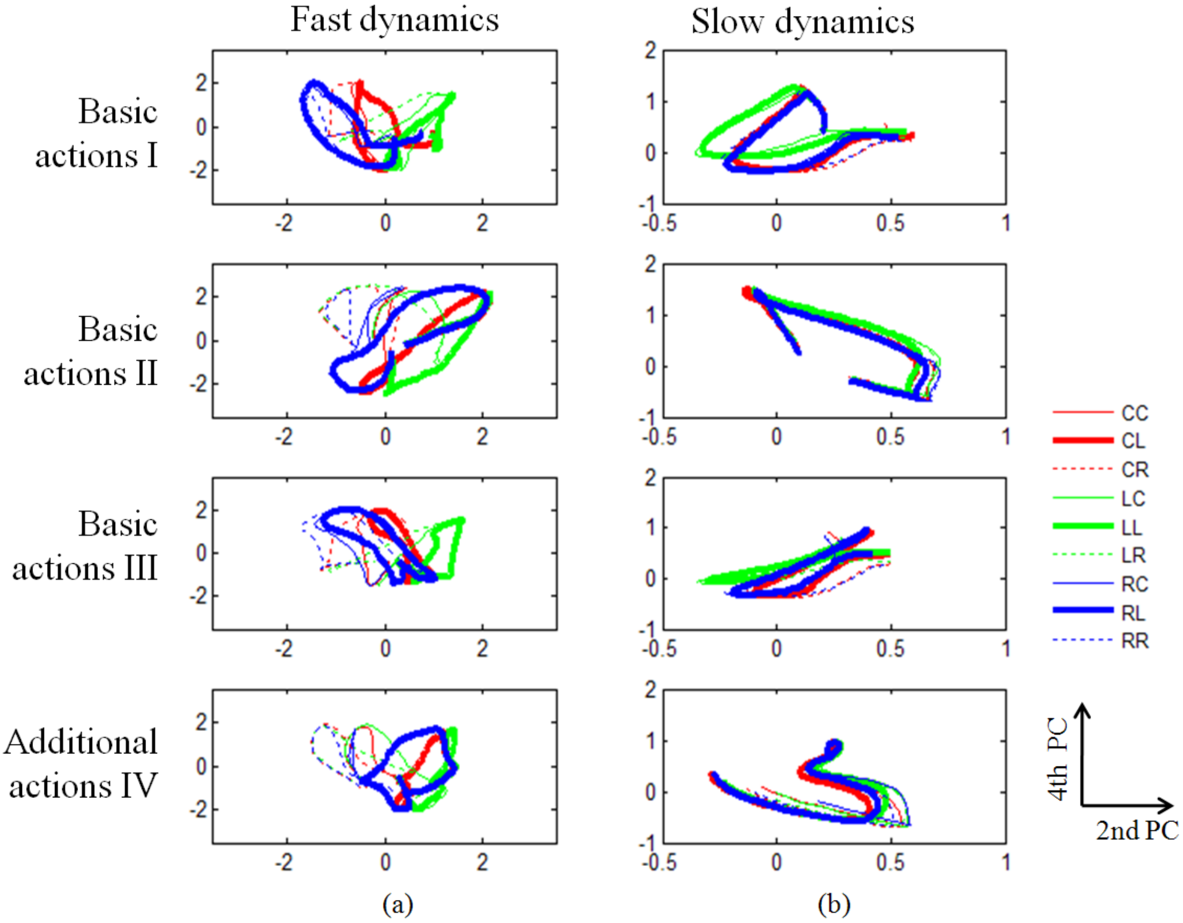

Fast dynamics
Slow dynamics
Basic actions I
Basic actions II
Basic actions III
Additional actions IV
2
0
-2
-2
0
2
1
-1
-0.5
0.5
CC
CL
CR
LC
LL
LR
RC
RL
RR
4th PC
2nd PC
(a)
(b)

Fig. 6

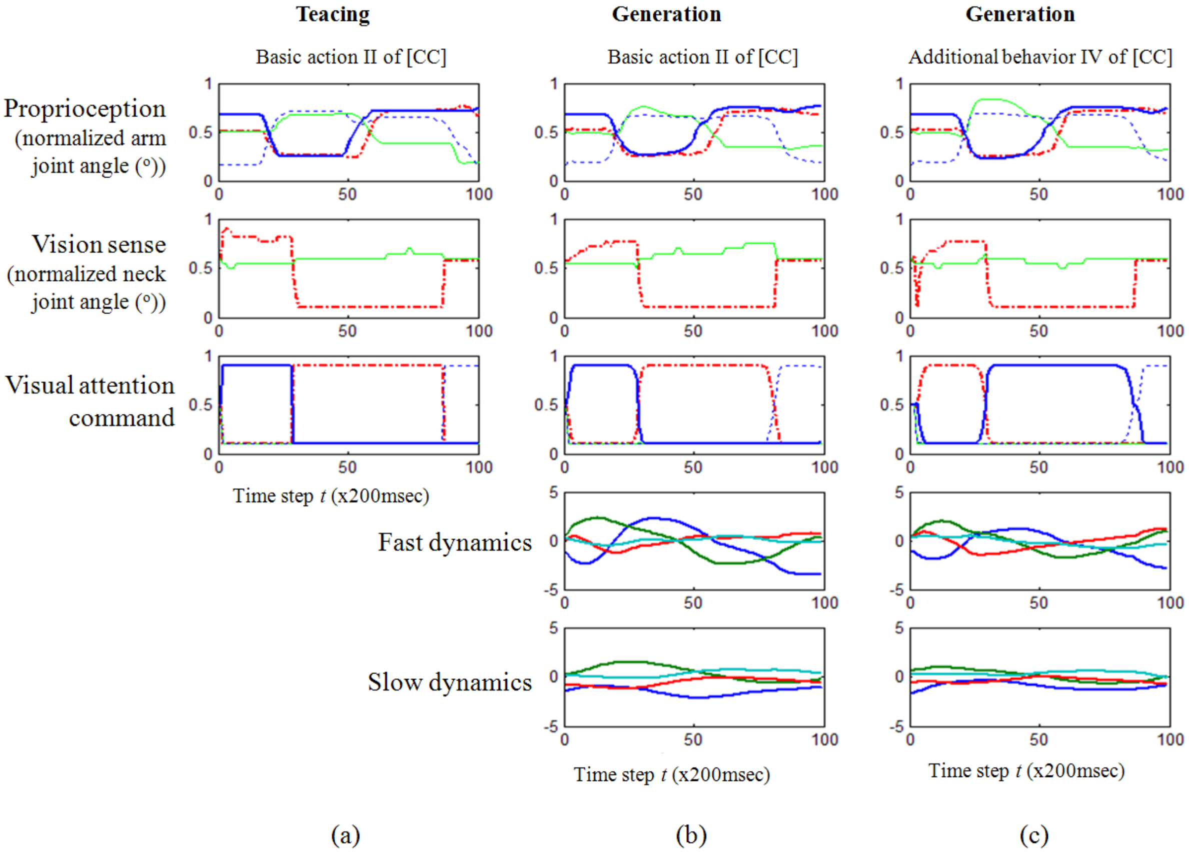

Teacing
Generation
Generation
Basic action II of [CC]
Basic action II of [CC]
Additional behavior IV of [CC]
Proprioception
(normalized arm
joint angle (°))
Vision sense
(normalized neck
joint angle (°))
Visual attention
command
Fast dynamics
Slow dynamics
Time step t (x200msec)
Time step t (x200msec)
Time step t (x200msec)
0
0.5
1
0
50
100
5
-5
(a)
(b)
(c)

Fig. 7

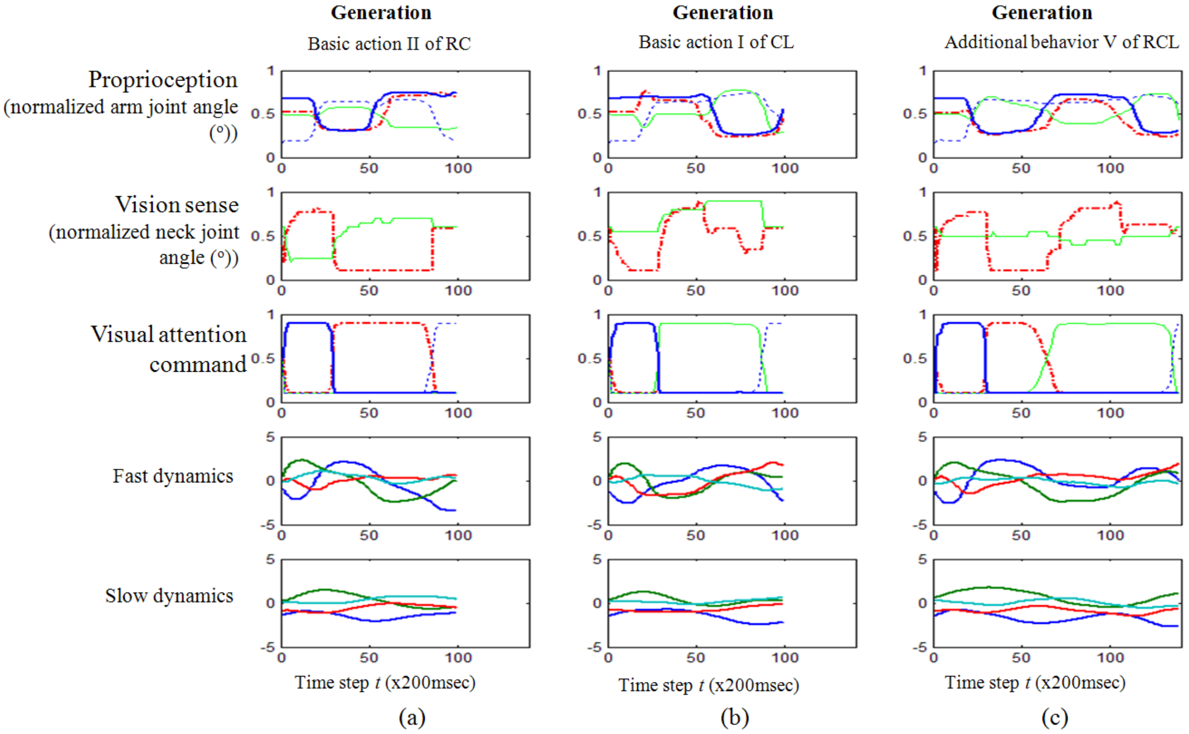

Generation
Basic action II of RC
Generation
Basic action I of CL
Generation
Additional behavior V of RCL
Proprioception
(normalized arm joint angle
(°))
Vision sense
(normalized neck joint
angle (°))
Visual attention
command
Fast dynamics
Slow dynamics
1
0.5
0
5
0
-5
0
50
100
Time step *t* (x200msec)
(a)
(b)
(c)

Fig. 8