Development of goal-directed gaze shift based on predictive learning

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

***Abstract*—Understanding others' actions as goal-directed is a key mechanism to develop social cognitive abilities such as imitation and cooperation. Recent findings in psychology have demonstrated that the ability to predict the goal of observed actions emerges as early as six months in infancy. However, what mechanisms are involved and how they trigger the development of this ability are still open questions. In this paper, we propose a computational model employing a recurrent neural network to reproduce the developmental process of goal-directed gaze shift. Our hypothesis is that it is possible to infer the goal of observed actions by learning to predict the attention target originated from bottom-up visual attention. While keeping consistency with psychological findings, our experimental results confirmed the hypothesis that learning to predict the attention targets leads to the development of predictive gaze shift to the action goal.**

Key Words: Cognitive developmental robotics, recurrent neural network, visual attention

I. Introduction

The ability to infer the intentions of others is crucial in humans to engage in social relations. It is argued that a prerequisite to infer the intentions of others is the ability to understand their actions as goal-directed, which means to evaluate actions based on causal relations. From a developmental point of view, it is suggested that understanding actions as goal-directed is a key mechanism to develop social cognitive abilities such as imitation and cooperation. However, what mechanisms are involved and how they trigger the development of this ability are still open questions.

Several studies have been carried out to understand when and how infants start to get involved in goal-directed actions. A remarkable work regarding this issue is the one conducted by Woodward [1]. In that study infants (5, 6 and 9 months old) were shown actors reaching for and grasping one of two objects. For the experiments four actors were employed: a human arm, a rod, a flat occluder and a mechanical grasping tool. The experimental results indicated that when infants were habituated to a goal-directed action (i.e., the human arm condition), they showed a stronger novelty response to test events that varied the goal of the action (e.g., the grasped object) than test events that varied the physical properties of the action (e.g., the motion path). On the other hand, if the actions were not goal-directed (i.e., the rod and the flat occluder conditions), or were goal-directed but difficult to infer the agency of the actor (i.e., the mechanical grasping tool condition), infants did not prefer one type of response to the other (i.e., the goal of the action versus the properties of the action). These results showed that infants differentiate between the actions of human beings and the motions of inanimate objects. Nonetheless, it is noteworthy to remark that their gaze analysis indicated that both the hand and the inanimate objects were equally effective as spotlights of attention.

Sommerville et al. [2] in a subsequent study focused on the impact of action experience on action perception and vice versa in relation to infants' ability to detect the goal of a grasping event. In this study infants participated in a visual habituation experiment, similar to the one in Woodward [1]. However, the experiment differed from the previous one in that a group of infants were allowed to interact with the objects prior to the visual habituation. The experimental results indicated that the experience of grasping objects enabled infants to detect the goal-directed structure of other persons' actions, that is, an impact of action execution on action perception.

In a recent work, Kanakogi and Itakura [3] addressed the issue of the emergence of prediction of goal-oriented actions in infancy. They showed that the ability to predict the action goals of others emerges as early as six months. Their experiment measured the visual attention of infants (4, 6, 8 and 10 months old) and adults in order to assess their prediction ability. The experimental settings consisted of three conditions: the first one was a hand reaching for and grasping one of two objects, the second one was the back of a hand reaching one of two objects but without grasping it, and the third one was a mechanical claw reaching for and grasping one of two objects. From the experimental results, adults were found to shift their gazes in a predictive manner under the three conditions, whereas 6-month-old and older infants performed predictively but only for actions which were clearly goal-oriented (i.e., hand grasping condition). In contrast, 4-month-olds were not able to make predictions. Additionally, adults were found to make predictions significantly earlier in comparison with 6-, 8-, and 10-month-olds. This study demonstrated that the ability to predict action goals undergoes a developmental process and that infants' prediction capability differs depending on the reaching agent.

In a later study, Cannon and Woordward [4], arguing that previous experiments lacked distinguishing goal prediction from movement anticipation, carried out a modified experiment. In their experiment infants were first familiarized with a reaching action directed to one of two objects. The location of the objects were then swapped, and infants' reactions were

assessed as the agent made an incomplete reach between the objects. The focus of their experiment was to assess the infants' predictions when the context had changed (i.e., the same movement would not realize the prior goal). Their experimental results showed that infants in the hand condition generated predictive gaze at the goal object at a different location, whereas infants in the claw condition generated predictive gaze at the location of the familiarized movement. Finally, their results showed, as in [1], that both the hand and the inanimate objects attracted infants' attention effectively. Among other studies on development of goal-directed gaze shift, Myowa-Yamakoshi et al. [5] have also contrasted the anticipation of action goals in humans with in chimpanzees, revealing that chimpanzees are also able to anticipate the goals as human do, but scan the actions differently.

In relation to the influence of perceptual features on goal-directed gaze shift, Henrich et al. [6] assessed the impact of the goal saliency on infants' ability to anticipate reaching actions. In their experiment, small and large goal objects were used for representing the low-salience and high-salience conditions, respectively. Their experimental results demonstrated that the goal saliency had an impact on goal anticipation in infants: The high-salient objects elicited earlier prediction of infants than the low-salient objects. They argued that action prediction might be influenced by the properties of the goal such as the size of the goal object.

In this work we address the issue of the development of goal-directed gaze shift from the perspective of cognitive developmental robotics [7]. Our hypothesis is that social cognitive abilities can develop through experiences based on simple innate mechanisms. A preceding study based on this hypothesis is the work done by Nagai et al. [8]. They proposed a computational model for the development of joint attention. Their model consisting of saliency-based visual attention and sensorimotor learning enabled a robot to acquire the ability to follow other person's gaze direction. Similarly, Balkenius and Johansson [9] proposed a developmental mechanism of smooth pursuit. Their model reproduced a transition from saccadic pursuit to smooth pursuit as a result of a gradually developing ability to predict the movement of a target.

Inspired by the above studies, we propose a computational model for the development of goal-directed gaze shift. We suggest that the ability of goal-directed gaze shift emerges through learning to predict the attention target based on the perceptual salience. Our model consists of three modules: a saliency-based attention module, which enables the system to gaze at a conspicuous target; a predictive learning module, which learns to predict the gaze shift generated by the saliency module; and a gate, which selects the output from the above two modules. Our model also integrates a working memory as a function of the predictive learning module in order to reproduce the gradual decrease in the prediction time as shown in [3]. Pelphrey et al. [10] have indicated that visuospatial short-term capacity, which is regarded as an early form of working memory, increases linearly from 6 to 12 months. The following sections, first, describe in detail our proposed model,

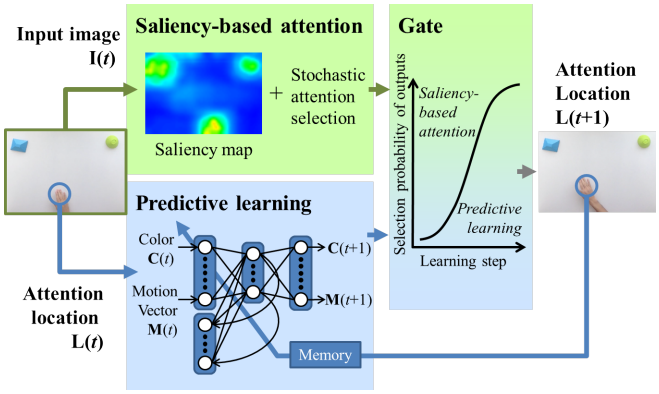

Fig. 1: Proposed computational model for development of predictive eye movement.

then explain the experiments settings employed for verifying our hypothesis, and finally introduce the experimental results, conclusions and future work.

II. Basic Ideas of our Approach

From a computational perspective, Kanakogi and Itakura's [3] work on the development of predictive eye movement could be interpreted as the following hypotheses:

1) It is possible to infer the goal of observed movements by learning to predict the attention target originated from saliency-based visual attention.
2) The performance on action prediction depends on the visual familiarity with an agent. The difficulty to predict the goal of an inanimate movement (i.e., a mechanical claw in [3]) can be attributed to a lack of visual experience with it.
3) The development of working memory capacity supports the acceleration of goal-directed gaze shift.

A computational model to verify our hypotheses is shown in Fig. 1. In this model, first, the saliency-based attention module extracts the primitive features of the input images by using a bottom-up approach. Then, the predictive learning module employs those primitive features for training a recurrent neural network. During the training, the temporal memory capacity increases progressively. The output of the neural network is used to calculate the predicted attention location. Finally the gate module selects the next location to attend to between the saliency-based and the prediction-based attention. The following sections explain in detail these modules.

A. Saliency-Based Attention Module

The saliency-based attention module employs a bottom-up approach for visual attention selection [11], which extends Itti et al.'s [12] model with a stochastic attention selection. An important concept of the saliency model is to rely only on perceptual features of the environment and thus to reproduce visual attention similar to younger infants [13]. Additionally,

visual saliency of action goals is suggested to influence on infants ability to anticipate reaching actions [6].

The input of this module is a sequence of images. The module first computes color, edge and optical flow, and calculates a saliency map. Then, one location $\mathbf{L}_s(t+1)$ at time t+1 is selected among the most salient locations of the saliency map by using a probabilistic function. Color and motion data are extracted as primitive features from the surrounding region of $\mathbf{L}_s(t+1)$. These features, representing the static and dynamic information of the attention, are used by the predictive learning module to train a recurrent neural network. Finally, the module outputs $\mathbf{L}_s(t+1)$ and the features $\mathbf{F}(t+1)$, which is represented with a population coding [14].

B. Predictive Learning Module

This module learns to predict $\mathbf{F}(t+1)$ (i.e., the color and motion vector) and then calculate the future attention location $\mathbf{L}_p(t+1)$ based on $\mathbf{F}(t+1)$. The main components of this module are: a working memory for temporal storage; a recurrent neural network capable of encoding the dynamic structure of the visual features; and an event detector.

The working memory is in charge of storing the past feature data $\mathbf{F}(t-w+1), \mathbf{F}(t-w+2), ..., \mathbf{F}(t)$, where w indicates the length of a time window. The data are used to train a recurrent neural network, which predicts the above features sequentially from $\mathbf{F}(t-w+1)$. The training is conducted by back propagation through time, where the teaching data is provided by the event detector. The function of the event detector is to detect a change in the attention target using the color information and calculate the training data for the recurrent neural network. The importance of this mechanism is that events involving a perceptual feature change are used to detect the goal of the observed action: when the module detects a color change in $\mathbf{F}(t)$, the training data is calculated so as to the neural network can learn to associate the current location $\mathbf{F}(t)$ as a target location to be predicted from the past locations $\mathbf{F}(t-w+1)$,...,$\mathbf{F}(t-1)$. The required steps to calculate the teaching data are as follows:

- if no significant color changes are detected among the frames stored in temporal memory, the training data for input $\mathbf{F}(t-w+1)$ will be $\mathbf{F}(t-w+2)$;
- instead, if a color change is detected in $\mathbf{F}(t-w+n)$, the training data for input $\mathbf{F}(t-w+1)$ will be the color component $\mathbf{F}_{color}(t-w+n)$ and the vectorial sum of motion vectors $\mathbf{F}_{motion}(t-w+2)$+...+$\mathbf{F}_{motion}(t-w+n)$.

In order to decide a potential future attention location $\mathbf{L}_p(t+1)$ the prediction learning module first calculates the locations corresponding to the predicted color $\mathbf{L}_{color}(t+1)$ and to the predicted motion $\mathbf{L}_{motion}(t+1)$. $\mathbf{L}_{color}(t+1)$ is determined by finding the image location which contains color similar to the predicted one, whereas $\mathbf{L}_{motion}(t+1)$ is calculated by adding the predicted motion vector to the current attention location. The module then determines the output $\mathbf{L}_p(t+1)$ using $\mathbf{L}_{color}(t+1)$ and $\mathbf{L}_{motion}(t+1)$ as follows:

$$\mathbf{L}_p(t+1) = \begin{cases} \mathbf{L}_{color}(t+1), & \text{if } \theta_1 \geq r(t+1) \\ \mathbf{L}_{motion}(t+1), & \text{otherwise} \end{cases} \quad (1)$$

$$\theta_1 = \frac{Sal[\mathbf{L}_{color}(t+1)]}{Sal[\mathbf{L}_{color}(t+1)] + Sal[\mathbf{L}_{motion}(t+1)]}, \quad (2)$$

where $Sal[\mathbf{X}(t)]$ is the saliency value at the location $\mathbf{X}(t)$; and $r(t+1)$ is a random value ranging from 0 to 1. That is, the module selects stochastically the attention location based on the salience at $\mathbf{L}_{color}(t+1)$ and $\mathbf{L}_{motion}(t+1)$.

C. Gate Module

The function of the gate module is to select the next attention location $\mathbf{L}(t+1)$ from $\mathbf{L}_s(t+1)$ and $\mathbf{L}_p(t+1)$. The module employs the following equations:

$$\mathbf{L}(t+1) = \begin{cases} \mathbf{L}_p(t+1), & \text{if } \theta_2 \geq r(t+1) \\ \mathbf{L}_s(t+1), & \text{otherwise} \end{cases} \quad (3)$$

$$\theta_2 = \frac{H_{sig} \cdot Sal[\mathbf{L}_p(t+1)]}{H_{sig} \cdot Sal[\mathbf{L}_p(t+1)] + (1 - H_{sig}) \cdot Sal[\mathbf{L}_s(t+1)]}, \quad (4)$$

where H_{sig} is a sigmoidal function formulated by:

$$H_{sig} = 1 - \frac{1}{1 + e^{-\alpha \cdot n}}. \quad (5)$$

α is a scale factor, and n is the learning step. Using this function the system gradually shifts the attention location from the saliency-based gaze location $\mathbf{L}_s(t+1)$ to the prediction-based location $\mathbf{L}_p(t+1)$, and as a consequence the system develops the ability of predictive gaze shift.

III. Experimental Settings and Results

In our experiment we reproduced similar experimental settings to those described in [3]. We designed two experimental conditions: a hand grasping condition and a claw grasping condition. In the video stimuli a hand and a claw moved diagonally from the bottom center of the image towards a top left/right corner until reaching one of two target objects, as shown in the top and the bottom of Fig. 2, respectively.

In the learning phase, six videos were employed for training the proposed model: Three videos corresponded to a hand grasping the top-left object, while the other three corresponded to a hand grasping the top-right object. No video of mechanical claw was used for training in order to verify our hypothesis regarding the effect of the lack of visual experiences with claws. The positions of the two objects were fixed and the same for all the videos. The six videos slightly varied in terms of the motion trajectory of the hand in order to make the network acquire generalization capability. In addition, the videos were created so as to have the same length of time: In the hand condition, the average size of each video was about 64 frames, and the portion corresponding to motion was about 30 frames; In the claw condition, the average size of each

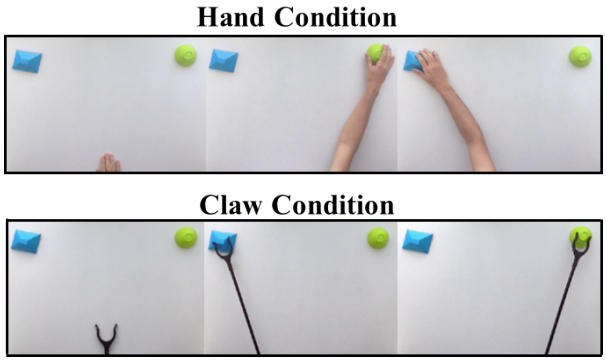

Fig. 2: Hand grasping condition (top) and mechanical claw condition (bottom). The leftmost image shows the initial state, and the middle and rightmost images show the final states under each condition.

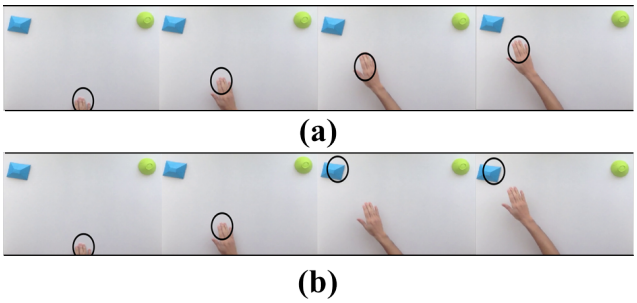

Fig. 3: Example of gaze pattern in hand condition: (a) before training and (b) at later stage of training (learning steps =75000, w=15).

video was about 66 frames, and the portion corresponding to motion was about 29 frames.

The recurrent neural network consisted of 27 input neurons (7 for the motion length, 12 for the motion angle and 8 for the color), 27 output neurons (the same distribution as the input neurons) and 25 hidden layer neurons. The time window of the working memory w was gradually expanded from 1 to 19 frames while the network was trained for 5000 steps with each window size. The scale factor α in Eq. (5) was 10. For statistical analysis purposes, we set ten different random initial weights for training the neural network.

Results of the model's prediction were categorizes as follows:

- Correct predictions: The model predicted successfully the target object using color and/or a motion vector (i.e., location) before the hand or the claw reached the object.
- Non prediction: The target object was not predicted before the hand or the claw arrival.
- Incorrect prediction: The predicted object did not correspond to the target object.

Fig. 3 shows examples of the attention location (indicated by a circle in the images) before (a) and after learning (b). In Fig. 3 (a), the system always gazes at the reaching hand (i.e., no prediction) because it mainly uses the output of the saliency-based attention module. In Fig. 3 (b), in contrast, the system predicts the target object before the hand reaches it. This prediction ability is acquired by learning the temporal sequences of saliency-based attention.

A. Hand Condition Test

In the testing phase twenty videos were employed for the hand grasping condition. Ten videos were reaching for right-side object and the other ten videos were for left-side object. The results of the test are shown in Fig. 4. For both Figs. 4 (a) and (b) the horizontal axis represents the learning step and the time window indicated at the bottom and the top of the graph, respectively. The vertical axis in Fig. 4 (a) represents the percentage rate for correct, incorrect, and non prediction, and that in Fig. 4 (b) represents how many frames earlier the prediction was achieved than the hand arrival at the target.

In the hand condition, the correct prediction rate increased gradually when the size of the time windows was larger than 9 frames as shown in Fig. 4 (a), and reached about 70 % with the time window of 14 frames. The average standard deviation was about 15 % which is relatively small. After training, the correct prediction rate was significantly higher than the non-prediction rate, which indicates that the system performed predictively. In contrast, the average rate of incorrect prediction category was maintained significantly low over learning. These results demonstrate the validity of our computational model for predicting a reaching action.

With respect to the relation between the anticipation time and the size of the time window, Fig. 4 (b) shows that the gaze shift was first achieved when the time window expanded to 8 frames. Then, the prediction time became earlier in proportion to the size of the time window. The earliest prediction (i.e., 14 frames earlier than the goal achievement) was produced when the time window became 19 frames. This result demonstrates that the development of the working memory facilitates the action prediction, which is in accordance with our third hypothesis.

B. Claw Condition Test

Similar to the previous experiment, twenty videos were employed for testing under the mechanical claw condition. Ten videos were reaching for the right-side object, and the other ten videos were for left-side object. The results are shown in Fig. 5.

In the claw condition, the average rate of the correct prediction was about 45 % at the end of learning, and did not surpass the non-prediction rate as shown in Fig. 5 (a). It is, however, surprising that our model could achieve the goal prediction to some extent despite no training with the claw stimuli. The large variance of the result indicates that our model could correctly predict the goal of claw reaching under certain conditions. To understand this, we should remember that our model utilizes both the static and dynamic image features for predictive learning. The former represents what

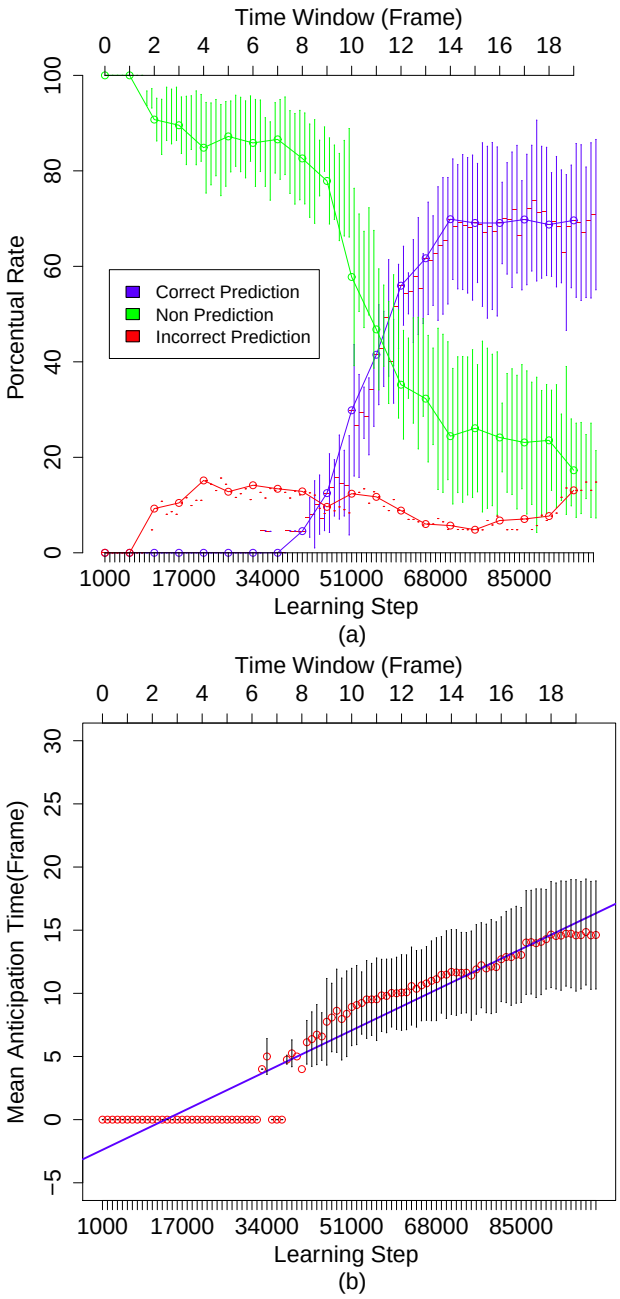

Fig. 4: Experimental results in hand condition. (a) prediction rate versus size of time window (b) prediction time versus size of time window for correct prediction.

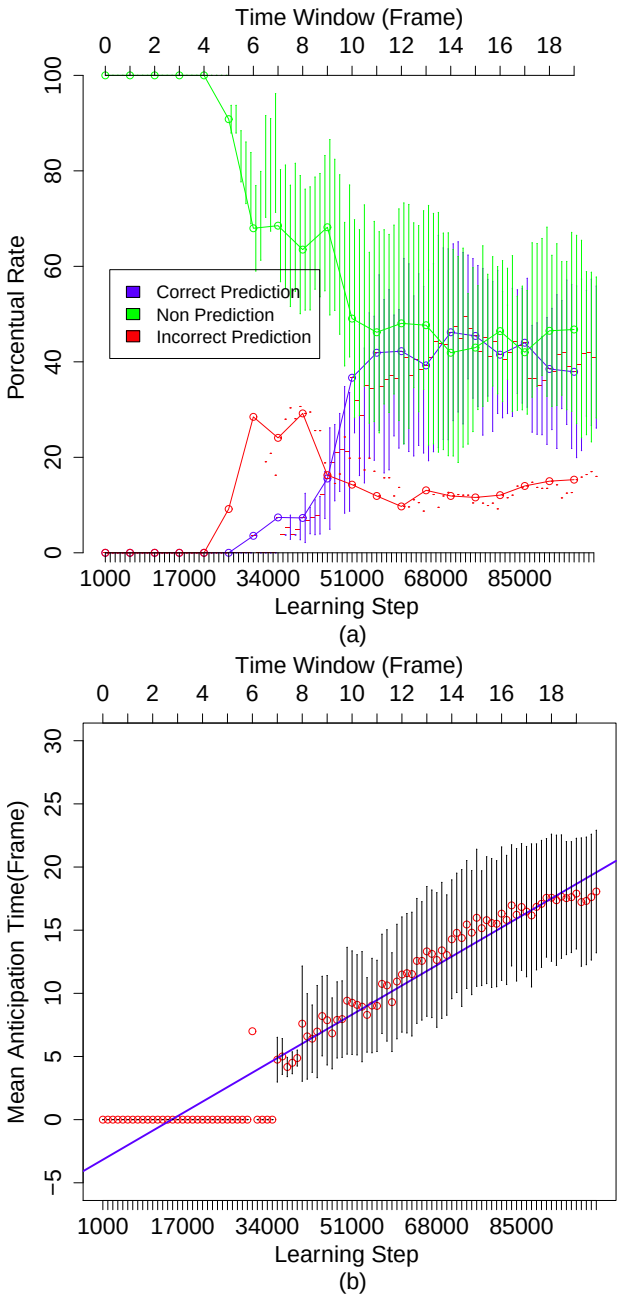

Fig. 5: Experimental results in claw condition. (a) prediction rate versus size of time window (b) prediction time versus size of time window for correct prediction.

the attention target is (i.e., a reaching agent or an object to be reached for), whereas the latter indicates in which direction the attention moves. As a consequence, even though the static information of the claw differed from the one used for training (i.e., the hand), the dynamic feature enabled the system to predict the direction of the target object.

The relation between the anticipation time and the size of the time window is plotted in Fig. 5 (b). In this graph we observe a similar result to the one obtained in the hand condition: The time at which prediction was achieved became earlier in proportion to the size of the time window.

Note that one of the objectives of our experiment was to test whether the neural network did make any prediction if the moving agent was a mechanical claw but not a hand. The results for the claw condition show that the performance of the prediction ability decreased in comparison to the results for the hand condition. This supports our hypothesis that visual experiences of actions lead to the development of the ability

to predict the same actions.

IV. Conclusion

In this paper, we have proposed a computational model to explain the development of goal-directed gaze shift. The experimental results showed that, in accordance with our first hypothesis, learning to predict the attention target originated from visual saliency could lead to the emergence of goal-directed gaze shift. Regarding the second hypothesis, the prediction performance of our system in the claw condition was significantly lower than in the hand condition. That is, our results verified that the visual experience of actions influences on the ability to predict the actions. For the third hypothesis, the experimental results demonstrated that temporal memory facilitates the development of prediction. To conclude, our computational model reproduced the multiple aspects of infant development of predictive gaze shift.

Our current model employed a sigmoid function for the gate to generate a transition from saliency-based attention to prediction-based gaze shift. This idea is based on the assumption that the neural network becomes able to predict more accurate locations by decreasing the prediction error as learning steps increase. Further work is required to model an automatic transition using a value of the prediction error.

Regarding the extension of our model to more natural scenarios, some considerations must be done. First, in relation to the experience of infants, the works of Cannon and Woodward [4] and Kanakogi and Itakura [3] employed two different approaches for their experiments: the former relied on specific learning of infants (i.e., infants acquired experience during the familiarization stage), whereas the later relied on the natural experience of infants (i.e., there was no familiarization stage). Regarding this point, we adopted the specific learning approach as an starting point for our current model and limited the experimental conditions. However, it is also required that our model learns from more natural experiences, and therefore further work is required to include changing conditions which can be considered more natural like multiplicity of objects, locations of objects, visual appearance, and other features. Next, it is important to note that the hypotheses of our work slightly differ from those in [3]. In their work, the main objective was finding correspondence between the development of action prediction of others and the development of infants' motor abilities (i.e., mirror neuron system). Therefore, integrating the motor development into the prediction learning model will be part of our future work. Finally, our model implements a mechanism to predict the goal of the action which relies on the visual information (i.e., color and motion). However, this information is not sufficient to account for the general phenomenon of goal detection. For example, in Cannon and Woodward's [4] study, where infants were familiarized with two reaching objects (a hand and a mechanical claw) and then exposed to the them moving in the same direction, the infants produced totally different patterns of prediction for the two object: in the hand condition infants gazed at the goal object at a different location, whereas infants in the claw condition gazed at the location of the familiarized movement. This might indicate that infants make predictions based on their experience with humans (i.e., experience of goal-directed actions) that could not be explained only in terms of visual perception. Then, an additional mechanism to account for this phenomenon is required. Regarding this issue, we intend to extend our model to learn from own experiences acquired during the generation of goal-directed actions. Specifically, we propose to include tactile information (e.g., when grasping an object) in order to allow the system to get feedback from own goal-directed actions. This approach would be also in line with the findings in [3], which reported the development of motor abilities to be synchronized with the development of prediction abilities. Then, the relation between tactile perception and goal detection can be regarded as natural consequence of motor development.

Acknowledgment

This work is partially supported by MEXT/JSPS KAKENHI (Research Project Numbers: 24119003, 24000012, 25700027) and JSPS Core-to-Core Program, A. Advanced Research Networks.

References

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

[2] J. A. Sommerville, A. L. Woodward, and A. Needham, "Action experience alters 3-month-old infants' perception of others' actions," *Cognition*, vol. 96, no. 1, pp. B1–B11, 2005.

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

[4] E. N. Cannon and A. L. Woodward, "Infants generate goal-based action predictions," *Developmental science*, vol. 15, no. 2, pp. 292–298, 2012.

[5] M. Myowa-Yamakoshi, C. Scola, and S. Hirata, "Humans and chimpanzees attend differently to goal-directed actions," *Nature communications*, vol. 3, p. 693, 2012.

[6] I. Henrichs, C. Elsner, B. Elsner, and G. Gredebäck, "Goal salience affects infants goal-directed gaze shifts," *Frontiers in psychology*, vol. 3, 2012.

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

[8] Y. Nagai, K. Hosoda, A. Morita, and M. Asada, "A constructive model for the development of joint attention," *Connection Science*, vol. 15, no. 4, pp. 211–229, 2003.

[9] C. Balkenius and B. Johansson, "Anticipatory models in gaze control: A developmental model," *Cognitive processing*, vol. 8, no. 3, pp. 167–174, 2007.

[10] K. A. Pelphrey, J. S. Reznick, B. Davis Goldman, N. Sasson, J. Morrow, A. Donahoe, and K. Hodgson, "Development of visuospatial short-term memory in the second half of the 1st year.," *Developmental psychology*, vol. 40, no. 5, p. 836, 2004.

[11] Y. Nagai, "From bottom-up visual attention to robot action learning," in *Proceeding of the International Conference on Development and Learning*, 2009.

[12] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[13] M. C. Frank, E. Vul, and S. P. Johnson, "Development of infants attention to faces during the first year," *Cognition*, vol. 110, no. 2, pp. 160–170, 2009.

[14] Y. Nagai, "The role of motion information in learning human-robot joint attention," in *Proceeding of the International Conference on Robotics and Automation*, pp. 2081–2086, 2005.