A Computational Model of Early Development of Predictive Eye Movement

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

In this work, we propose a computational model employing a recurrent neural network for modeling the developmental process of action prediction ability. Our hypothesis is that temporal memory is a driving force mechanism for the development of this ability. While keeping consistency with the psychological findings, our experimental results confirmed the hypothesis, showing that memory capacity has a strong influence on the development of prediction abilities.

Key Words**:** Cognitive developmental Robotics, Recurrent Neural Network, Visual Attention

1. Introduction

A recent work by Kanakogi and Itakura [1] has indicated that in humans the ability to make predictions of action goals of others emerges in infancy as early as six months. In that work, trajectory of the gaze of infants 4, 6, 8 and 10 month-olds and adults was measured in order to assess their predictive ability. From the experimental results, adults was found to shift their gaze in a predictive manner, six month and older infants performed predictively but only for actions which were clearly goal-oriented, and four month-old infants were not able to make predictions. Additionally, it was found that prediction in adults was performed significantly earlier in time in comparison with infants 6, 8, and 10 months, which in turn performed earlier than 4 month old infants. This experiment demonstrated that infants undergo a process of development of their ability to make predictions of action goals, which involves perception of acting agents and improvement of cognitive capacities. Myowa et al [2] has also contrasted anticipation of action goals in humans and chimpanzees, revealing that chimpanzees are also able to anticipate goals as humans, although they scan goal-directed actions differently.

Regarded to play a crucial role in cognition development, working memory is an important concept in the field of cognitive developmental psychology. Working memory is a conceptual structure in charge of storing temporal information during execution of actions, and then, is related to several cognitive processes. In relation to the development of working memory, Pelphrey et al [3] has indicated that visuospatial short-term capacity, which can be considered a type of working memory, increases in infancy between the ages of 6 and 12 months following a linear pattern, and the age at which it becomes evident may vary from 6 to 8 months of age depending on the complexity of tasks.

In this work we address the problem of emergence and development of action prediction within the field of Cognitive Developmental Robotics [4]. Based on the findings of Kanakogi and Itakura [1], and Pelphrey et al [3], we found that the period of development of visuospatial short-term memory and the period of development of predictive eye movement are synchronized. We therefore hypothesize that working memory supports the development of predictive behavior. Then we propose a computational model based on the concept of working memory to explain the developmental mechanism of predictive eye movement. The experimental results demonstrated that temporal memory capacity is intrinsically related to the development of predictive behaviour.

2. Basic Ideas of our Approach

In order to explain the findings in Kanakogi and Itakura [1] regarding the development of eye predictive movement, we state the following hypotheses from a computational perspective:

1. It is possible to infer the goal of observed movements by learning to predict the attention target, which in turn originates from the visual saliency-based attention.
2. Development of working memory capacity supports the emergence and development of predictive eye movement.
3. The performance on prediction depends on the visual familiarity with an acting object. Regarding the experiment of Kanakogi and Itakura [1] in which infants were not able to perform in a predictive manner with a mechanical claw, we attribute that to the lack of visual experience with claws.

The computational model that we propose to carry out the validation of our hypotheses is shown in Fig. 1. In this model, first, the primitive features of the input images are extracted by using a bottom-up approach. Then, those primitive features are employed for training a Recurrent Neural Network. During the training, the temporal memory capacity is increased progressively. The output of the neural network is used to calculate the predicted location. And finally the gate module selects the next attention location between the saliency-based attention location and the prediction-based attention location. The components of this model are explained in the following sections.

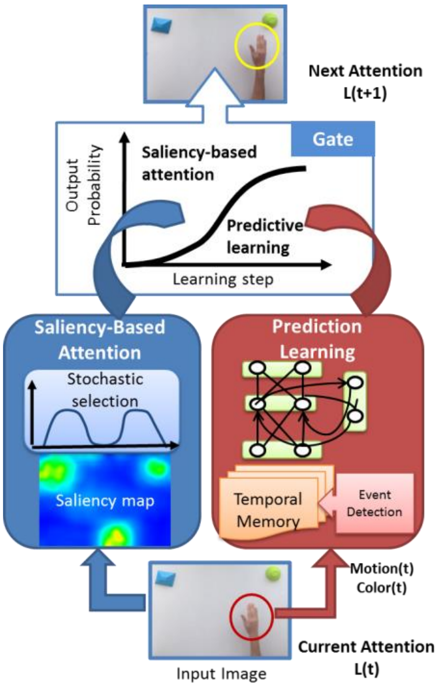

Fig. 1 Proposed computational model for development of predictive eye movement.

2.1 Saliency-Based Attention Module

The saliency-based attention module is a bottom-up approach for visual attention selection based on the work of Nagai et al [5] which uses a computational model of saliency map proposed by Itti et al [6]. The importance of the concept of saliency map lies in the fact that it is regarded to be a model of primates' low-level visual attention.

The input of this module is a sequence of images. In the first stage, motion flow, color component and edge component of images are employed for calculating a saliency map, as shown in Fig. 2. Then, by using a probabilistic function, one location is selected among the most salient locations of the saliency map. Finally, primitive features are extracted from the region around the selected location: color, motion length and motion orientation. The output of this module is the stochastic-based attention location and the extracted primitive features.

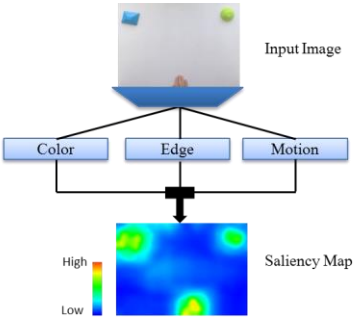

Fig. 2 Computation of Saliency Map.

2.2 Predictive Learning Module

The function of the predictive learning module is learning to predict color and motion features which are employed for calculating the prediction-based attention location. As shown in Fig. 1, the main components of this module are: a recurrent neural network, which is a special class of neural network whose main characteristic is exhibiting dynamic temporal behavior; a memory component for temporal storage; and a component called "Event Detection" whose functionality will be explained later in this section. The inputs of this module are the primitive features extracted in the saliency-based attention module: color, motion length and motion orientation. The outputs of the neural network are the predicted motion and color features.

The recurrent neural network receives the primitive features as inputs. Then, the output of the neural network is compared to the teaching data provided by the component "Event Detection", and the result is fed back as training error.

The component called "temporal memory" is in charge of providing temporal storage of feature data, and its capacity is relative to the amount of data stored simultaneously for a certain amount time (Hereafter the storage capacity of the temporal memory will be expressed in frames and will be referred as "Time Window").

The component "Event Detection" is in charge of providing the appropriate teaching data by accessing and analyzing the data stored in temporal memory. The teaching data is calculated as follow:

- If the component called Event Detection detects a significant change in the color feature within the color features of temporarily stored frames, then the teaching data will be the color of the detected frame and the motion trajectory required for reaching the location of the detected frame.
- If it is not detected a significant change among temporarily stored frames, then the teaching data will be the color of the next frame and the motion trajectory for reaching the next frame.

2.3 Gate Module

The function of the gate module is selecting the next attention location between the saliency-based attention location and the prediction-based attention location. The criterion for selection is based on the principle of gradual development employing stochastic selection, and therefore we employ the following equations:

$$L(t+1) = \cdot Dc(t)\cdot M(t)\cdot Lp(t) + \left(1 - Dc(t)\cdot M(t)\right)\cdot Ls(t) \qquad (1)$$

$$M(t) = P\left[Sig\cdot Sal[Lp(t)], (1-Sig)\cdot Sal[Ls(t)]\right] \qquad (2)$$

Where,

- L(t+1) represents the next location and is the output of the gate module.
- Dc(t) changes from 0 to 1 when the color predicted by the neural network differs from the color of the current attention point, which indicates that a future event has been predicted.
- Ls(t) is the location given by the saliency-based module.
- Lp(t) is the location predicted by the neural network.
- Sal(X(t)) is the value of saliency in the location X at time t.
- M(t) is the output of the function P, which is a probability function which uses a distribution of weights based on the saliency values of the candidate

locations (Ls and Lp) and a sigmoid function “Sig”. The most salient location has more probability to be selected. The output of P is 0 or 1.

The sigmoidal function, mentioned previously, is employed for shifting gradually the attention selection from the stochastic-based attention to the prediction-based attention while the learning time increases. The sigmoidal function can be formulated as follow:

$$Sig = 1 - \frac{1}{1 + e^{-\alpha_{gate} n}} \qquad (3)$$

Where Sig is the probability of shifting from the saliency-based attention to the prediction-based attention, α is a scale factor, and n is the learning step. The graphical representation of the sigmoidal function is shown in Fig. 3.

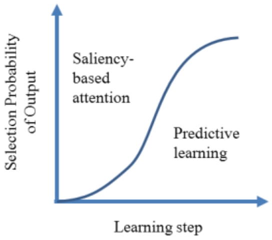

Fig. 3 Probabilistic function for attention selection in the gate module.

3. Experimental Settings and Results

In our experiment, we reproduced similar experimental settings to those described in Kanakogi and Itakura [1]. We considered two experimental cases, the hand grasping condition and the claw grasping condition, which are shown in the left side and the right side of Fig. 4, respectively.

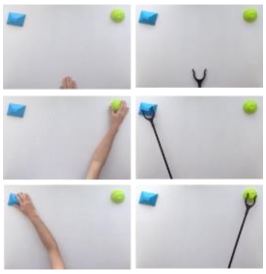

Fig. 4 Right: hand grasping condition. Left: mechanical claw condition.

In the learning phase six videos were employed for training the system: Three videos corresponded to a hand grasping an object located at the left of the image and the other three videos to a hand grasping an object located at the right image. The positions of the two target objects were fixed and the same for all the videos. And, in order to test our hypothesis regarding the effect of the lack of visual experience with claws, no video of mechanical claw was used for training. The recurrent neural network consisted of 26 input neurons, 26 output neurons and 25 hidden layer neurons. The size of the Time Window was modified gradually from 0 to 19 frames during training, the learning iterations for each window was of 5000 steps, and the scale factor α of the sigmoidal function was 10.

The results of the experiments explained in the following sections were classified into the following categories for analysis purposes:

- Correct predictions: The system predicted successfully color and location (angle and distance) of the target object.
- Non prediction: The target object was not predicted before the hand (or claw) arrival.
- Incorrect prediction: The neural network predicted an incorrect color regardless of the accuracy of the predicted location.

3.1 Hand Condition Test

In the testing phase twenty videos were employed for the hand grasping condition. Ten videos were right-side object grasping and the other ten videos were left-side object grasping. The average size of each video was about 70 frames, and the portion corresponding to motion was about 30 frames. The result of the test is shown in Fig. 5. The vertical axis represents the percentage rate for each category, and the horizontal axis represents both the learning step and the time window which increased gradually during training.

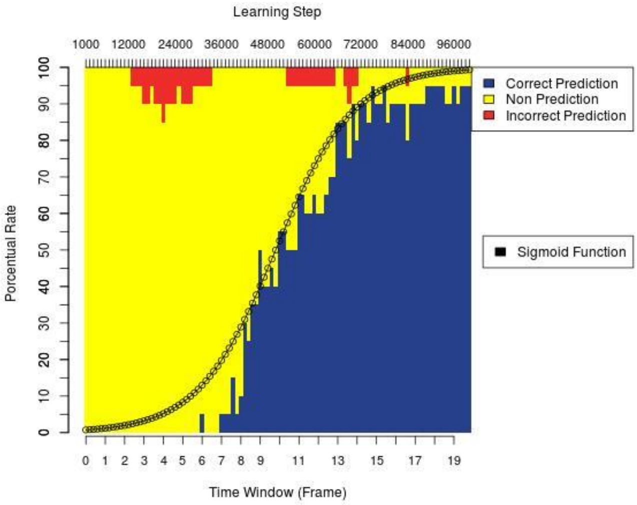

Fig. 5 Prediction Rate versus size of Time Window per each category for hand condition.

In hand condition, the success rate increased gradually when the size of the time windows was larger than 9 frames, reaching its maximum average value of 90% for time windows larger than 14 frames. However, when the size of the time windows was smaller than 9 frames, the “Non Prediction” category was high in relation to the possible effect of the sigmoidal function. This can be explained by the fact that, when the time window is still relatively narrow the learning module is not still able to detect significant changes from the visual information. Finally, the average rate of “Incorrect Prediction” category was significantly low. These results demonstrate the validity of our computational model for inferring goals by learning to predict the attention target, in accordance with our first hypothesis. Additionally, in relation to the emergence of prediction, the results also confirmed that the temporal memory played a crucial role, as we stated in our second hypothesis.

The relation between prediction time and the size of the time window is shown in Fig. 6, where the vertical axis represents how many frames earlier the prediction was achieved. In this graph we can see that the prediction was achieved in average 5 frames earlier than the arrival of the hand when the size of the time window was equal to 8. Then, the time at which prediction was achieved continued to become earlier in proportion to the size of the time window. And when the size of the time window was 19 frames, the prediction was achieved in average 14 frames earlier than the arrival of the hand. This result demonstrates that the prediction became earlier when the size of the time window was expanded, which is also in line with our second hypothesis.

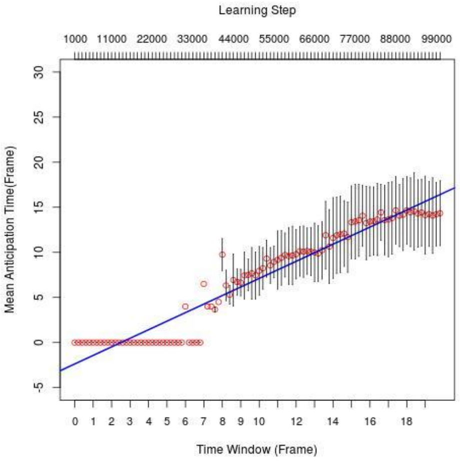

Fig. 6 Prediction Time versus size of Time Window for "Correct Prediction".

3.2 Claw Condition Test

In the testing phase twenty videos were employed for the mechanical claw condition. Ten videos were right-side object grasping and the other ten were left-side object grasping. The average size of each video was about 80 frames, and the portion corresponding to motion was about 40 frames. The result of the test is shown in Fig. 7. The vertical axis represents the percentage rate for each category, and the horizontal axis represents both the learning step and the time window which increased gradually during training.

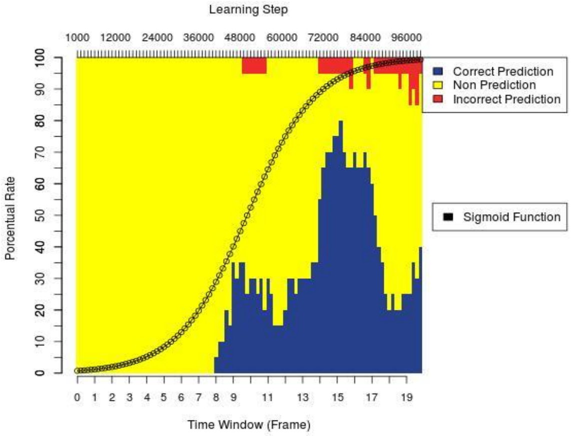

Fig. 7 Prediction Rate versus size of Time Window per each category for claw condition.

One of the objectives of this experiment was testing whether the neural network did make any prediction if the moving object was a mechanical claw but not a hand. The result shown in Fig. 7 revealed that the rate of "correct prediction" increased significantly during some period of the learning process, which means that the neural network tried to make predictions when claw grasping videos were introduced, in spite of the fact that the neural network was trained using only the physical description of the hand. However, the global average success rate was low, and then the prediction results were not stable enough. We attribute that dynamical and unstable change to the limited amount of types of objects employed for training the neural network, and then there was still some dependence on the physical attributes of the acting object.

These experimental results indicate to a certain degree that the motion information of the hand had a stronger influence on prediction training than its physical information. This result can be explained by the fact that during training the prediction of the target depends more on motion and less on the appearance of the moving object.

4. Conclusion and Discussion

In this paper, we have proposed a computational model to explain the developmental mechanism of predictive eye movement. The experimental results showed that, in accordance with our first hypothesis, learning to predict the attention target originated from visual saliency was proven to be valid for implementing a computational model of predictive eye movement. For our second hypothesis, the experimental results demonstrated that temporal memory facilitates the emergence of predictive behavior. Furthermore, when the size of the temporal memory was increased, the results showed that the prediction time became significantly earlier. This indicates that the development of the temporal memory had significant influence on the development of action prediction. Regarding the third hypothesis, which attributed the difference in behaviour under claw condition and hand condition to the lack of visual experience, the experimental results were not conclusive. However, some results suggested that our current model might not be sufficient to explain this phenomenon and additional cognitive mechanisms should be also considered. To conclude, our computational model was confirmed to follow a developmental process for acquiring prediction ability.

As a next step, visual mechanisms are required to distinguish between hand condition and mechanical claw condition. Also, considerations regarding development of motor skills in connection with the mirror neuron system will be part of a future work.

Acknowledgment

This work is partially supported by MEXT/JSPS KAKENHI (Research Project Number: 24119003, 24000012, 25700027).

References

[1] Kanakogi, Y. & Itakura, S. Developmental correspondence between action prediction and motor ability in early infancy. Nat. Communications. 2, 341 (2011).

[2] Yamakoshi M, Scola C, Hirata S, Humans and chimpanzees attend differently to goal-directed actions. Nature Comm 3: 693 (2012)

[3] 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, pp. 836–851, 2004.

[4] Asada, Minoru et al., Cognitive developmental robotics as a new paradigm for the design of humanoid robots. Robotics and Autonomous Systems, 2001, vol. 37, no 2, p. 185-193.

[5] Yukie Nagai. From bottom-up visual attention to robot action learning. In Proceedings of the 8th IEEE International Conference on Development and Learning, pp. 1–6, 2009.

[6] Laurant Itti, Christof Koch, and Ernst 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, Nov 1998.