Development of Visual Attention

- Random scan paths in young infants during the observation of a person’s face become simpler and more similar with age [Kato & Konishi, 2013].
- Attention shift among the eyes and mouth is driven by both featural processing and configural processing of input images.

Our Hypotheses about Underlying Mechanism
(a) The visual information obtained from the previous time steps needs to be integrated in order to predict the subsequent attention location.
(b) Attention shift among the eyes and mouth is driven by both featural processing and configural processing.

A Computational Model for Visual Attention

**Key idea:** A recurrent neural network (hypothesis (a)) learns to optimize attention locations through learning of featural and configural classification tasks (hypothesis (b)) of input images.

- **Input:** a face image \( x_t \), the attention location at previous time step \( l_{t-1} \)
- **Output:** the next attention location \( l_t \), classifications \( c^f_t \) for a featural task and \( c^c_t \) for a configural task
- **Four network modules:**
  - Glimpse network \( f_1(\theta^g) \): To extract a glimpse representation \( g_t \) corresponding to the fovea
  - Internal network \( f_2(\theta^h) \): To integrate \( g_t \) with the internal representation \( h_{t-1} \) with a recurrent connection
  - Classification networks \( f^f_1(\theta^f) \) and \( f^c_2(\theta^c) \): To estimate the categories \( c^f_t \) and \( c^c_t \) of \( x_t \)
  - Location network \( f_3(\theta^l) \): To determine the next attention location \( l_t \)
- **Training through reinforcement learning:**
  - Reward: 1 for a correct classification, 0 otherwise

Experimental Setting
- **Two classification tasks:**
  - Featural task: Emotion estimation
  - Configural task: Estimation of head orientation
- **Training conditions:**
  - KDEF dataset (600 pictures with seven types of emotion and three head orientations) [Lundqvist et al., 1998]
  - Image size: 128 x 128 [pixels]
  - Glimpse size: 26 x 26 [pixels]
  - Max. number of attention shift: 6
  - Learning period: 1,000 epochs

Exp. 1: Accuracy of Featural and Configural Tasks
- Emotion estimation (i.e., featural task): 39%
- Estimation of head orientation (i.e., configural task): 97%

Exp. 2: Development of Visual Attention
- The neural network reproduced the developmental change similar to that observed in infants.
- Glimpse locations (diameter of circles = frequency of attention)
- Scan paths (thickness of lines = frequency of attention shift)

Exp. 3: Complementary Roles of Featural and Configural Tasks
- The emotion estimation induced closer attention to salient facial features (i.e., featural processing).
- The estimation of head orientation required attention shift among facial regions (i.e., configural processing).

Conclusion
- The recurrent neural network reproduced infant-like development of visual attention.
- Featural and configural processing played complementary roles in the development of visual attention.

Future Issue
- Integration of bottom-up saliency-based attention

References

This work was supported by JST CREST “Cognitive Mirroring” (Grant Number: JPMJCR16E2), Japan.