A functional role of multiple spatial resolution maps in form perception along the ventral visual pathway

Yoshiki Kashimori[1], Nobuyuki Suzuki [2], Kazuhisa Fujita [2], Meihong Zheng [1], and Takeshi Kambara[1]

[1] Division of Bioinformatics
Department of Applied Physics and Chemistry,
[2] Department of Information Network Sciences,
Graduate School of Information Systems,
The University of Electro-Communications,
Chofu, Tokyo, 182-8585 Japan
e-mail:kashi@pc.uec.ac.jp

Key words: form perception, ventral visual pathway, spatial resolution map, prediction, neural model

Corresponding author:
Dr. Yoshiki Kashimori
Division of Bioinformatics
Department of Applied Physics and Chemistry,
The University of Electro-Communications,
Chofu, Tokyo, 182-8585 Japan
Tel: +81-424-43-5470
Fax: +81-424-89-9748
e-mail:kashi@pc.uec.ac.jp

Abstract:

We present a functional model of form pathway in visual cortex based on the idea of predictive recognition, in which the prediction for input image is compared with the feedforward signals from retina. Three kinds of spatial resolution maps of V1 and V4, broad, middle, and fine resolution maps, are effectively used to achieve the form perception. The prediction is generated by the feedforward signals of main neurons in broader resolution maps of V1 and V4, and then is compared with the feedback signals of main neurons in fine map of V4. We propose here the functional roles of the three kinds of spatial resolution maps in predictive recognition of object form.

1. Introduction

Object recognition is fundamental to the behavior of higher primates. The visual system rapidly and effortlessly recognizes a large number of diverse objects in cluttered, natural scenes, - a very difficult computational task. For the invariant object recognition task, the shape of one object projected on the retina is often quite different depending on the view direction, but our visual system can recognize that these distinct images belong to the same object. How does the brain recognize the form of object ?

Several models have been proposed to solve the problem of the form perception. Fukushima[2] has proposed a neural network model, referred to as 'neocognitron', in which a little deformation of visual input is converged through a hierarchical network. Rolls and Deco [8] have proposed VisNet model in which the translation invariance is achieved based on a learning rule with time delays. Poggio and Edelman[7] have proposed a standard view model in which the invariant recognition is achieved based on a particular standard view. These models have been made based on the feedforward signals from the retina. The form perception also crucially depends on previous visual experience, or visual memory. However, little is known about the mechanism by which the form of object is recognized using visual memory in high order visual system. In order to study the mechanism of how the visual memory interacts with the object information encoded in the lower system, we propose a functional model of visual system in which the main function of the model is generated based on the prediction of visual memory.

The basic idea of our model is as follows. When an object image is presented on the retina, a visual memory is chosen in high order visual system, based on the coarse image of the retina input. The memory is projected into the lower areas as a prediction for the perception of the input image. The prediction signal is compared with the feedforward signal form the retina. The form perception is accomplished by matching of both the signals. When both the signals do not represent the same object information, other prediction is chosen in the memory system.

Our model includes multiple resolution maps in V1 and V4. There exists the experimental evidence of the multiple resolution maps in V1 and V4 areas which are tuned to different spatial frequencies [1,4,6,12]. Although these maps process the information of visual input with different resolution, the role of these maps is poorly understood. The present model provides a clear answer for the functional role of multiple resolution maps; the broader resolution maps are used to retrieve the memories of temporal cortex (TE) which are categorized by broad shape of object, while the fine resolution map is used to compare the feedforward signals from lower system with the feedback signals from higher order memory system.

In the present study, we present a neural network model which may make an essential processing of visual information of objects in the form pathway from early visual area V1 to temporal cortex. Here we show the functional roles of prediction and multiple resolution maps in the form perception.

2. Neural model of visual areas from retina to TE

2.1. Information processing pathways for form perception

Our perception of objects in the field is derived from various low-level cues such as orientation, color, velocity, and binocular disparity. These cues are extracted from retinal images by relevant feature detection neurons in the first visual are V1. Each of these cues makes an essential contribution to the visual perception of relevant specific attribute such as form, color, motion, and depth. There exists the specific pathway in the visual system along which the neural computation of perception of specific attribute is made based on the relevant-low level cue [11].

In the present study, we consider the pathway of form perception, because the computation of form perception seems to give us a good case study for making clear a role of prediction in visual perception. The form pathway consists of the brain areas involved in ventral pathways, retina/LGN(lateral geniculate nucleus), V1, V2, V4, TEO(posterior temporal cortex) and TE(temporal cortex) [5]. Our model also includes PP(posterior parietal) of dorsal pathway. We consider explicitly the functional roles of these areas except for V2.

2.2. Basic concept of our model

2.2.1. Basic structure of our neural model

To investigate the neural mechanism of visual perception, we made a neural network model for a form perception pathway from retina to TE. The network structure of our model is illustrated in Fig. 1. The model consists of seven groups of layers corresponding to retina/LGN, V1, V4, TEO, PP, TE, and WM (working memory area), respectively, from bottom to top.

The retinal network is an input layer, on which object image is projected. The LGN network transforms the retina image into the firing rates of LGN neurons. The outputs of LGN neurons were calculated by 2D-Gabor functions with three different spatial frequencies.

The neurons of V1 network have the detection ability of the simple feature of object image, such as orientation and edge of a bar. The V1 network also consists of three different types of neurons with respect to the spatial resolution of feature detection, fine tuned neurons with high spatial resolution(V1F), middle tuned neurons with middle spatial resolution(V1M), and broad tuned neurons with low spatial resolution(V1B). The V1 network contains M x M hypercolumns, each of which contains L orientation columns. The three types of V1 networks have different network size, and receive the output of LNG relevant to the spatial frequency.

The V4 network consists of three different networks with high, middle, and low spatial resolution, which receive the outputs of V1F, V1M, and V1B, respectively. The convergence of outputs of V1 neurons enables V4 neurons to respond to a combination of features encoded by V1 layer.

The functional role of TEO is to detect an elemental figure that corresponds to particular combination of the features encoded by V4 layer. The existence of the elemental figure neurons has been reported by Tanaka [10]. The TEO neurons may respond to the relevant elemental figure even if the shape of the figure is a little transformed.

The PP network consists of N x N neurons, each of which corresponds to the spatial position of each pixel of the input image. The functional roles of PP network are to encode the spatial position of a whole object and the spatial arrangement of its parts in retinotopic coordinate and to mediate the spatial position of the attended part of object.

In TE network, the information of form of an object is memorized as attractors corresponding to some pairs of the part and its spatial arrangement. The memory of object is categorized by the elemental figures of TEO.

The WM(working memory) network makes the dynamical linkage of the visual memory within the group categorized by the elemental figures of TEO. The network model was made based on the dynamical map model [3].

The mathematical descriptions of our neural network models from V1 to TE are described in Ref. [9].

2.2.2. Roles of predictive signals and resolution maps in form perception

Feedforward process

The visual image is processed in parallel in three kinds of V1 maps, V1F, V1M, and V1B, responding strongly to specifically orientation lines or edges in their receptive fields. Because of high resolution of V1F, it can detect the detailed features of object image, but it takes a long time to process the object information because of a large number of object information. Whereas,V1M and V1B process object information with lower resolutions, but they do quickly the object feature. The outputs of V1F, V1M, and V1B are converged in V4F, V4M, and V4B, respectively, and then the further convergence of outputs of V4 neurons enables TEO neurons to detect the elemental figure of object. The elemental figure is mainly determined by the signals from broader maps, V4M and V4B, because they arrive at TEO network earlier than the signals from fine map V4F and suppress the late signals. The outputs of broader V4 maps are also sending to PP network, in which the spatial position of a whole object and the spatial arrangement of its parts are represented. Receiving the outputs of TEO, TE network retrieves the memory attractors corresponding to object form that are categorized by the elemental figures of TEO. These memory attractors are dynamically linked in WM network. Then one of these object memory is chosen in WM network as a predictive signal, and furthermore the attractor corresponding to the part and arrangement of a retina image is enhanced by attention.

Feedback process

The attended attractor of the part and arrangement in WM network excites the relevant memory of TE by the feedback connections. Through the descending connections from the neurons in TE to the neurons in V4 and PP, the information of object form and position encoded by a pair of attractors are sending back to V4 and PP, respectively, as the prediction for retinal input signals. Then the feedback image of object is reproduced in the V4 network by binding the information of spatial position from PP and the information of form from TE. The feedback signals from PP also enhance the activities of V1M and V1F neurons corresponding to the attended location. The V4F neurons fire depending on the difference between the feedforward image filtered in V1F network and feedback image given by the prediction. The V4F network has not necessarily the same input pattern for the feedforward and feedback signals, because the feedback signal from TE is merely a prediction. When both signals in V4F network represent the same information of object's part, the firing pattern is more stabilized, and then attention is directed to the other pairs of parts and arrangement in the same object. When the feedforward signals match with the feedback signals for all the pairs, form perception is accomplished. When both signals represent different information of object's part, the firing pattern of V4F generated by both signals are cancelled out each other due to the lateral inhibition of V4F neurons. Then the V4F neurons do not fire. After the silence of V4F neurons, the WM network chooses a new pair of different object memory as prediction. The matching process is performed in V4F again. It is continued until the feedforward and feedback signals to V4F network are matching for all the pairs of object.

3. Results

3.1. A role of lower resolution maps in generating prediction

To investigate the neural mechanism for generating a prediction in form perception, we calculated the responses of early visual modules, V1, V4, and TEO, for different two retina images shown in Fig. 2a. Three types of outputs of LGN neurons were calculated based on the visual images filtered by Gabor functions with different spatial frequencies. These outputs for the retina image A are shown in Fig. 2b. After the detection of orientation lines in V1, the V4 maps extract the image of a whole object, depending on the spatial resolutions of three V4 maps, as shown in Fig. 2c. The TEO neuron encodes the broad image of the retina input, because the processing of the lower resolution maps are faster than the processing of the fine resolution map. Receiving the outputs of TEO neuron, TE network retrieves the memory corresponding to retina images, A and B, which are in the same group categorized by TEO neuron. Then one of the memory, which corresponds to object A or B, is chosen as a prediction signal for form perception. The visual systems always work in parallel, but the different latencies associated with different spatial resolutions allow the generation of prediction signals based on coarse features in the retina image.

3.2. A matching process using fine resolution maps

When the object image A shown in Fig. 2a is presented on the retina, the TE network retrieves the memories of objects, A and B, which are represented by the attractors corresponding to three parts and their arrangement (up, middle, down) as shown in Fig. 3a. These attractors are dynamically linked in WM network as shown in Fig. 3b. When the memory of object B is chosen as a prediction and then the attractors b1 is enhanced by attention, the information of the part and arrangement of b1 are sending to V4F and PP via TE network. The feedback signals from PP enhance the activity of V1 neurons encoding the relevant location of b1. In V4F, the feedforward and feedback signals are compared only for the attended pair of object image. Then the firings of V4F neurons are suppressed, because the patterns of feedforward and feedback signals do not represent the same object information. After the silence of V4F, the WM network generates another prediction that corresponds to the memory of object A, under the switching bias signal, as shown in Fig. 3b. The feedback signals are sending back to V4F and PP again. The matching of both signals makes V4F more stable. The pairs of part and arrangement about object A are sequentially compared in V4F by switching the attention signal in WM network, as shown in Fig. 3b. The form perception is accomplished when, for all the pairs of the object image A , the feedforward signals match with feedback signals.

4. Concluding Remarks

In the present study, we have proposed a functional model of visual system in which the form perception is achieved based on the prediction signal from WM area via TE area. The prediction signal is one of dynamical memory categorized by the broad shape of object in TEO. We have also proposed the functional roles of broad and fine resolution maps in the form perception. The prediction is mainly determined by the feedforward signals from the broader resolution maps, V4M and V4B through V1M and V1B, while the prediction signals are compared with the feedforward signals on the fine resolution map V4F through V1F.

The prediction signals may also play an important role in reconstructing from the two-dimensional images the physical properties of three dimensional surface and the boundary between patches that may correspond to the outlines of physical objects in the scene. Even if the information is lacking where part of the scene is hidden from view, the visual system must be able to separate structures of a partly occluded object from those of the occluding objects. Because the prediction provides the information about the depth and boundary between patches of objects, it seems quite useful to use the predictive signals in the perception of the more complex object scene.

References

[1] T. R. Born, and R.B.H. Tootell, Spatial frequency tuning of single units in macaqu supragranular striate cortex, Proc. Nat. Acad. Sci.88(1991)7066-7070.

[2] K. Fukushima, Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, Biol. Cybern. 36(1980) 193-202.

[3] O. Hoshino, Y. Kashimori, and T. Kambara, An olfactory recognition model of spatio-temporal coding of odor quality in olfactory bulb, Biol. Cybern. 79(1998) 109-120.

[4] L. Maffei, and A. Fiorentini, The visual cortex as a spatial frequency analyzer, Vision Res. 13(1973) 1255-1267.

[5] S .Marcelija, mathematical description of the responses of simple cortical cells, J. Opt. Soc.Am.(1980)1297-1300.

[6] G.F. Poggio, R.W. Doty, and W. H. Talbot, Foveal striate cortex of behaving monkey: single-neuron responses to square-wave grating during fixation of gaze, J. Neurophysilol. 40(1977)1369-1391.

[7] T. Poggio, S. Edelman, A network that learns to recognize three dimensional objects. Nature 343(1990) 263-266.

[8] E. Rolls, and G. Deco, Computational neuroscience of vision(Oxford University Press, 2002).

[9] N. Suzuki, N. Hashimoto, Y. Kashimori, and T.Kambara, A neural model of predictive recognition in form pathway of visual cortex, in: Proc. IPCAT'03(Lausanne, Switzerland, 2003)105-122..

[10] K. Tanaka, Mechanism of visual object recognition:monkey and human studies, Curr. Oppnin. Neurobiol. 7(1997)523-529.

[11] D.C. Van Essen, and E.A.Yoe, Concurrent processing in the primate visual cortex, in:M.S.Gazzania, ed., The Cognitive Neuroscience(MIT Press, Cambridge MA,1995)383-400.

[12] H.R. Wilson, D.K. MacFarlane, and G.C. Phillips, Spatial frequency tuning of orientation selective units estimated by oblique masking, Vision Res. 23(1983)873-882.

Figure Legends

Fig.1. The structure of our model. The model consists of seven groups of layers; retina/LGN(lateral geniculate nucleus), V1, V4, PP(posterior parietal), TEO (posterior temporal cortex), TE (temporal cortex), and WM (working memory area). F, M, and B mean the fine, middle, and broad spatial resolution maps, respectively. a ~c in TE are the attractors of visual memory. The solid line and dotted one indicate the feedforward and feedback signals, respectively.

Fig. 2. (a) Two kinds of retina images used. (b) The outputs of LGN filtered by Gabor functions with three kinds of spatial frequencies. B, M and F mean broad, middle, and fine resolution, respectively. (c) Responses of V4 and TEO to the retina image A. The filled rectangles in each V4 map indicate the firing neurons encoding the object features at the each resolution level. The filled rectangle of TEO means the firing neuron encoding coarse image of a whole object.

Fig. 3. (a) Three pairs of attractors constructing the object memory. Each pair consists of the part (X, |, ^, … etc) and the arrangement (top, middle, and bottom). The object memory for A and B consist of the attractors a1 ~ a3 and b1 ~b3, respectively. (b) Dynamic state of WM network during the form perception. A horizontal bar on row corresponding to a1 ~ b3 indicates that the network activity stays in the attractors. The switching bias means the uniform impulse stimulus to each WM network in order to change the dynamical state of WM network.

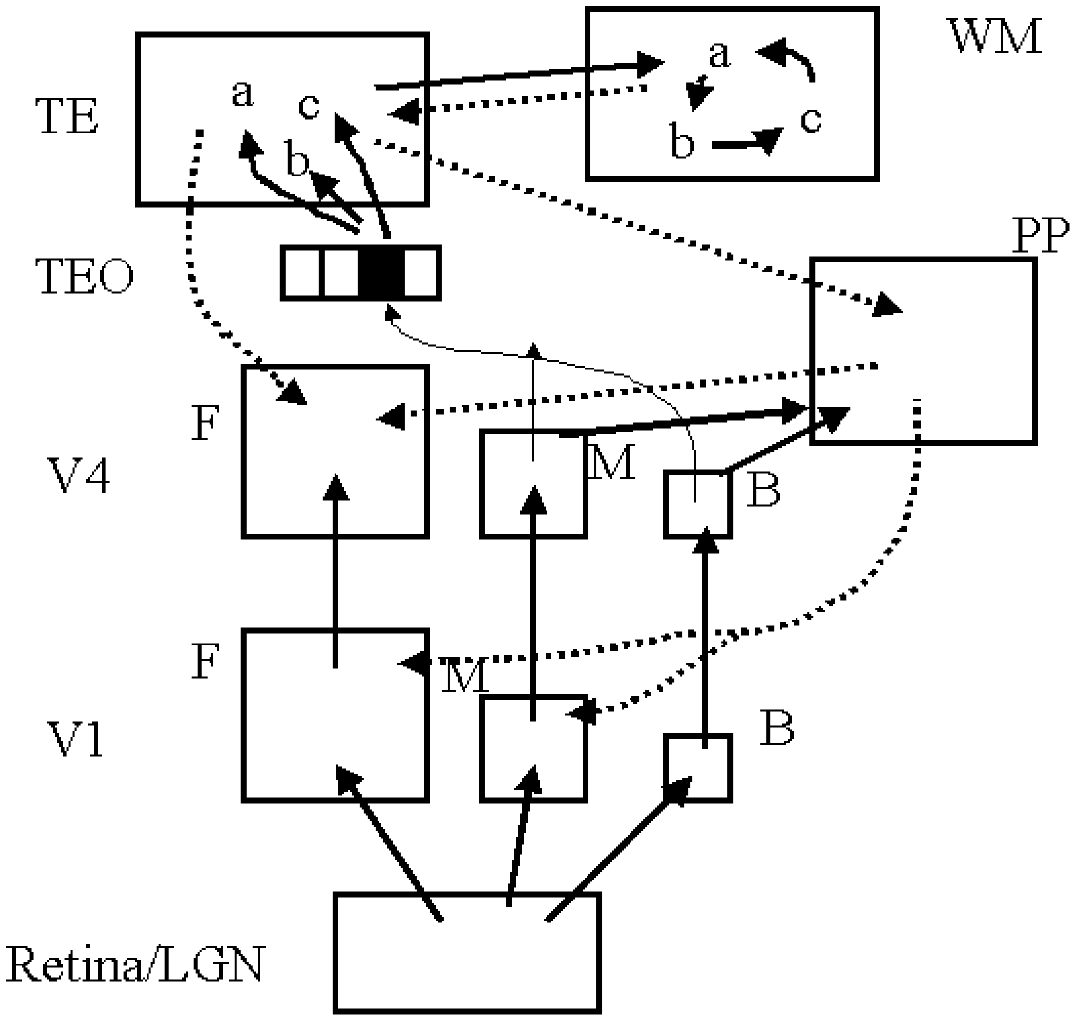

Fig. 1 Y. Kashimori et al.

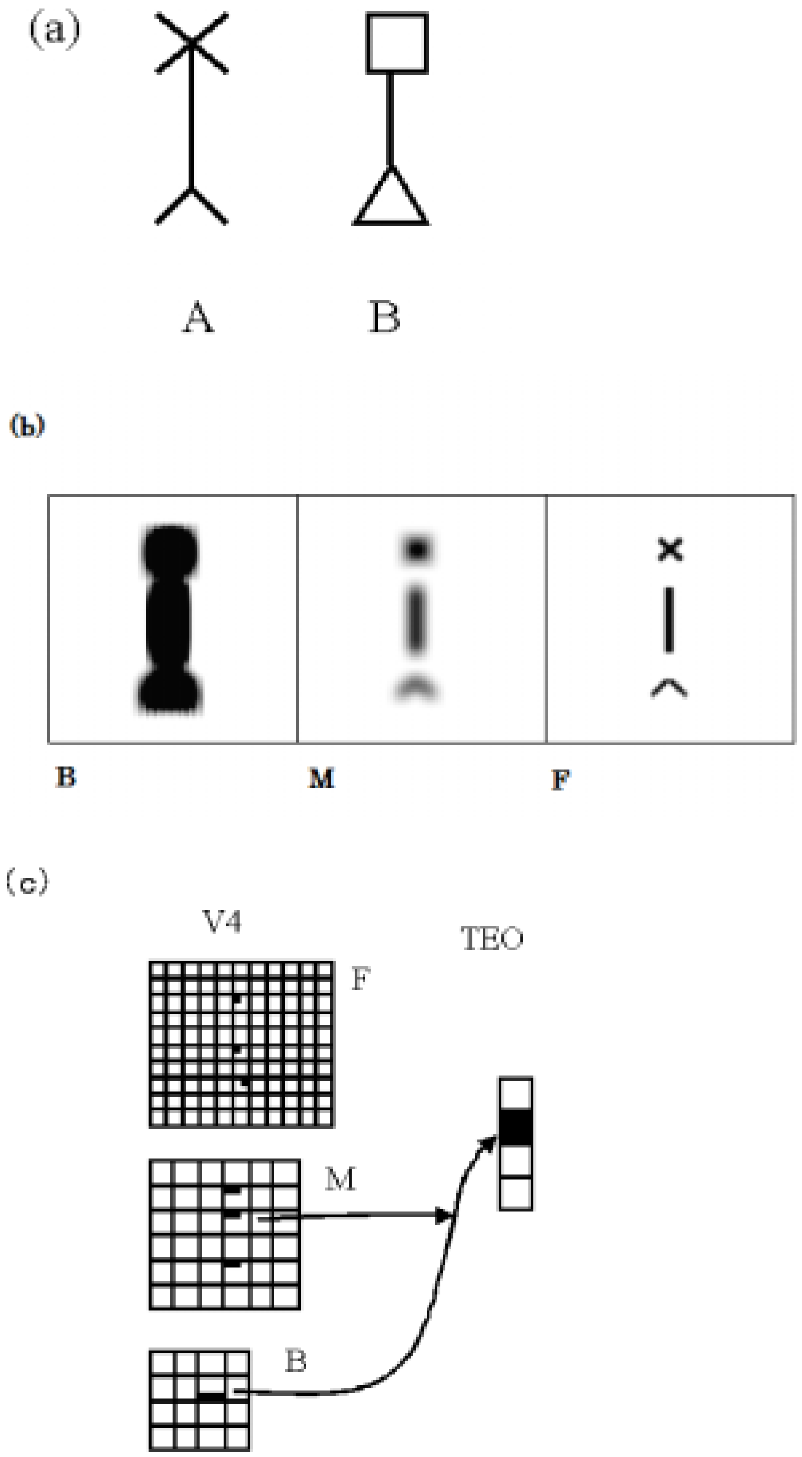

Fig. 2 Y. Kashimori et al.

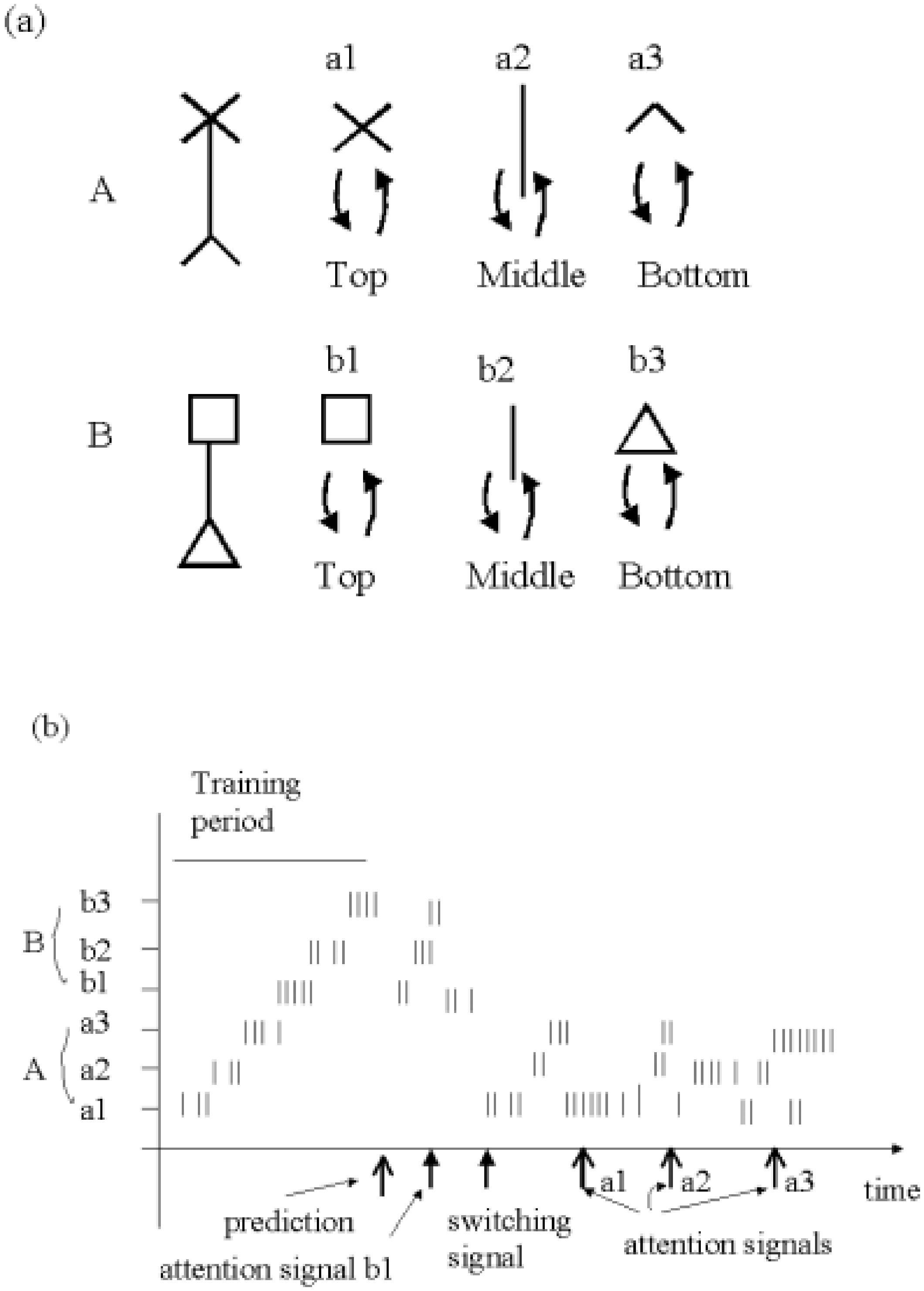

Fig. 3 Y. Kashimori et al.

Biosketches

Yoshiki Kashimori received his Ph. D. degree from Osaka City University in 1985. He is an associate professor in the Department of Applied Physics and Chemistry at University of Electro-Communications. His research interest is to clarify the neural mechanism of information processing in the electrosensory, auditory, and visual systems, based on modeling of neurons and their network. He also investigates the emergence of dynamical orders in various biological systems, based on the nonlinear dynamics.

Nobuyuki Suzuki is presently a student in the Graduate school of Information Systems at the University of Electro-Communications. His research interest is to clarify the neural mechanism of visual perception.

Kazuhisa Fujita is presently a student in the Graduate school of Information Systems at the University of Electro-Communications. His research interest is to clarify the neural mechanism of electrolocation.

Meihong Zheng received her Ph.D. degree from University of Electro-Communication in 2002. She has been engaging in the research area of biological complex system using computer simulation.

Takeshi Kambara received his Ph.D. degree from Tokyo Institute of Technology in 1970. He is a professor of biophysics in the Department of Applied Physics and Chemistry and professor of Biological Information Science in the Graduate School of Information Systems at University of Electro-Communications. His scientific interests cover the neural mechanism of information processing in the olfactory, auditory, visual, gustatory, and electro-sensory systems, and emergence of dynamical orders in various biological complex systems. His research work uses the "in silico" method. neural mechanism of visual perception.