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

Yoshiki Kashimori [a,b], Nobuyuki Suzuki [b], Kazuhisa Fujita [b], Meihong Zheng [c], and Takeshi Kambara [a,b]

[a] Division of Bioinformatics
Department of Applied Physics and Chemistry,
University of Electro-Communications, Chofu, Tokyo, 182-8585 Japan
[b] Department of Information Network Sciences,
Graduate School of Information Systems,
University of Electro-Communications, Chofu, Tokyo, 182-8585 Japan
[c] Department of Applied Physics, School of Science and Engineering,
Waseda University, Shinjuku, Tokyo 169-8555, Japan.

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

Corresponding author:
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 predictive recognition scheme, in which feedback image signals predicted from the memory system are compared with the feedforward image signals from retina. We propose the functional roles of three kinds of spatial resolution maps in predictive recognition of object form. The broad resolution map is used to categorize visual image based on the outline of object and to choose the prediction image in the memory system based on the categorization. The middle resolution map is used for more detailed categorizations and for focusing prediction range. The fine resolution map is used to compare the feedforward image of a part of the object with the image predicted.

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, but this is 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 independently of its images?

Several models have been proposed to solve the problem of the invariant form perception. Fukushima[3] has proposed a neural network model, referred to as 'Neocognitron', in which a little deformation of visual input image is converged through a hierarchical network. Rolls and Deco [10] have proposed VisNet model in which the translation invariance is achieved based on a learning rule with time delays. Poggio and Edelman[9] 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 hierarchical processing of image signals along the feedforward pathway from the retina to visual cortex. 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 the high order visual system.

In order to study the mechanism of how the visual memory interacts with the object information encoded by 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 to the lower system.
The basic idea of our model is as follows. When an object image is presented on the retina, a visual memory is chosen in the high order visual system, based on the coarse image on the retina. The memory is projected to the lower systems as a prediction for the perception of the input image. The image signal predicted is compared with the feedforward signal sent from the retina. The form perception is made based on matching between both the signals. When the image signal predicted does not match with the feedforward signal, other image in the memory system is chosen as a new prediction.

Our model includes multiple resolution maps in V1 and V4 areas of visual cortex. There exists the experimental evidence of the multiple resolution maps in V1 and V4 areas which are tuned to different spatial frequencies for image resolution [1,6,8,14]. Although these maps process the information of single visual input with different resolutions, the role of these maps is poorly understood. The present model provides a clear possible answer for the functional role of multiple resolution maps; the broad resolution map is used to retrieve the relevant image memorized in temporal cortex (TE), based on the categorization of the object made using the coarse image on the map. The middle resolution map is used for more detailed categorizations and for forcusing attention on a part of object form. The fine resolution map is used to compare the feedforward image of the part sent from the lower system with the image predicted by the higher order memory system.

In the present study, we present a neural network model which may make an essential processing of visual information of object in the form pathway from early visual area V1 to temporal cortex. Using the model, we show the functional roles of multiple resolution maps in the form perception based on the predictive recognition scheme.

2. Neural model of visual pathway from retina to TE

2.1. Information processing pathway for form perception

Our perception of objects is derived from various elemental 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 area V1. Each of these cues makes an essential contribution to the visual perception of relevant specific attribute of object 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-elemental cue [13].

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) [7]. Our model includes PP(posterior parietal) of dorsal pathway beside the visual areas mentioned. 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 neural networks corresponding to retina/LGN, V1, V4, TEO, PP, TE, and WM (working memory area).

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 transformed by 2D-Gabor functions with three different spatial frequencies.

The neurons of V1 network have the ability to detect the elemental features of object image, such as orientation and edge of a bar. The V1 network 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 specifically to a combination of elemental features represented on V1 network.

The functional role of TEO is to detect an elemental figure that corresponds to a particular combination of the features such as cross and triangle represented on V4 network. The existence of the neurons tuned to elemental figures has been reported by Tanaka [12]. The TEO neurons tuned to an elemental figure have the ability to respond actively to various figures whose shape is slightly different from the real shape.

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 functions of PP network are to represent the spatial position of a whole object and the spatial arrangement of its parts in the retinotopic coordinate system and to mediate the location of the part of object to which is paid attention.

In the TE network, the information about form of an object is memorized as hierarchical dynamical attractors [4, 5] representing both the form of the part and its spatial position. The memory of object is categorized based on the elemental figures of TEO.

The WM(working memory) network mediates the projection of a complete form of the object memorized in TE on the V4 and PP networks. The network model was made based on the dynamical map model [4, 5].

The mathematical description of our neural network model of visual pathway from V1 to TE is given in Ref. [11].

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

Feedforward process shown in Fig.1 by solid arrow lines

The visual image is encoded in parallel by three kinds of V1 maps, V1F, V1M, and V1B, responding specifically to 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 encode the object image because V1F map needs to represent a large number of elemental features and their positions. Whereas, V1M and V1B encode object image with lower resolutions, and as a result they do quickly the mapping. The image signals decomposed into elemental features in V1F, V1M, and V1B are partially combined in V4F, V4M, and V4B, respectively. Then the further binding between three kinds of V4 maps in TEO enables TEO neurons to detect the elemental figure of object. Then, the elemental figure is represented by the signals mainly from broad and middle maps, V4M and V4B, because they arrive at TEO network earlier than the signals from V4F and suppress the late signals. The outputs of V4M and V4B maps are also sent to PP network, in which the spatial position of a whole object and the spatial arrangement of its parts are represented, respectively. Receiving the outputs of TEO, TE network retrieves the memory attractors

representing object forms which belong to the category corresponding to the elemental figure encoded by TEO. These memory attractors correspond to the different images of the same object and are dynamically linked in WM network. Then one of these object image memories is chosen in the WM network as a predictive signal.

Feedback process shown in Fig.1 by dotted arrows lines

The predicted attractor representing a pair of the part form and its position excites the relevant memory of TE by the feedback connections. Through the descending connections from the neurons in TE to the neurons in V4F and PP, the information of object form and position encoded by a pair of attractors are sent back to V4F and PP, respectively, as the prediction to compare with the retinal input signals. Then the feedback image of object is reproduced in the V4F network by binding the information of spatial position from PP with the information of form from TE. The feedback signals from PP also enhance the activities of V1M and V1F neurons so that only the image of the part attracting attention remains in the networks. The V4F neurons are inhibited depending on the magnitude of difference between the feedforward image filtered in the V1F network and feedback image given by the prediction. When the feedback signal in the V4F network matches with the feedforward signal, the firing pattern is stabilized, and then attention is paid to the pair of other part and its position in the memory of the same object. When the feedforward signals match with the feedback signals for all the parts, the form perception is accomplished. When the feedback signal is different from the feedforward one, the firing patterns of V4F generated by both signals are cancelled out each other due to the lateral inhibition between V4F neurons. Then the V4F neurons do not fire. After that, 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 match with each other for all the parts 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 areas, V1, V4, and TEO, to two different retinal images shown in Fig. 2a. Three types of outputs of LGN netwwork were calculated based on the visual images filtered by Gabor functions with different spatial frequencies. These outputs for the retinal image A are shown in Fig. 2b. The detection of line orientations is made in V1, as shown in Fig. 2c. Then the V4B and V4M maps represent the image of a whole object, depending on the spatial resolutions of their maps, as shown in Fig. 2c. However, the V4F map can not rapidly encode the object features. The TEO network encodes only the broad image of the retinal input. Receiving the outputs of TEO network, TE network retrieves the memory index corresponding to the broad resolution images shown in Fig 2b, which corresponds to the category including images A and B,

Then one of the image belonging to the category, A or B, is chosen as a prediction signal for form perception. The visual systems always work in parallel, but the different latencies required for different spatial resolutions allow for the system to make the prediction based on coarse features of the retinal 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 both images, A and B, which are represented by the attractors X_n (X = a, b; n=1 ~3) corresponding to three parts and their positions (top, middle, down) as shown in Fig. 3a. These attractors are dynamically linked in the WM network, that is, activated cyclically as shown in Fig. 3b. When the memory of image B is chosen as a prediction and then the attractor b_1 is enhanced by attention, only attractor b_1 is activated and as a result the information of the part and its position represented by b_1 are sent to V4F and PP via TE network. The feedback signals from PP enhance the activity of V1 network corresponding to the real image encoded in a_1 whose position is the same as that in b_1. In V4F, the feedforward signal is compared with the feedback one. Then the firing of V4F neurons are suppressed, because the feedforward signal does not match with the feedback signal. After the silence of V4F, the WM network generates another prediction that corresponds to the memory of object A, under application of the switching bias signal, as shown in Fig. 3b. The feedback signals are sent back to V4F and PP again. The matching between both signals makes V4F more stable. The pairs of part and its position included in object A represented by a_1, a_2, and a_3, are sequentially compared with the relevant feedforward signals in V4F by switching the attention signal in WM network as shown in Fig. 3b. The form perception is accomplished when, for all the parts of the object image A , the feedforward signals match with the 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 chosen among the memories belonging to the category corresponding to the elemental figure in TEO activated by the outline of object. We have also proposed the functional roles of broad, middle, and fine resolution maps in the form perception. The prediction is determined based on the feedforward signals from the broad and middle resolution maps, while the prediction signals are compared with the feedforward signals on the fine resolution map V4F via V1F.

Our model has some resemblances to ART model proposed by Carpenter and Grossberg [2]. Both the models were made based on the idea that perception is achieved by the comparison between the input signal and top-down expectation signal. The ART model consists of two types of networks, a subnetwork with short-term memory, F1, and a higher order memory system, F2. F1 detects features of input signal, while F2 receives the signal from F1 and then gives a feedback signal to F1

in order to compare the input signal with the feedback signal corresponding to expectation. The functional roles of F1 and F2 networks correspond to those of V4 and TE network in our model, respectively.

However, there exists an important difference between ART model and our model. In ART model, it is not clear how the top-down expectation signal is generated in the memory system. Our model proposed a neural mechanism for generating the prediction signal (top-down expectation signal) and the functional roles of multiple resolution maps in the form perception. The broad resolution map is used to detect a category of visual memory whose members are chosen for prediction. The middle resolution map is used to select the part of object to which we pay attention. The fine resolution map plays an important role in comparing the input signal with the prediction.

The prediction signals may also play an important role in reconstructing the physical properties of three-dimensional surface from the two-dimensional images and detecting the boundary between patches that may correspond to the outlines of physical objects in the scene. Even if a part of the scene is hidden from view, the visual system can 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. Natl. Acad. Sci. U.S.A. 88(1991)7066-7070.

[2] G.A. Carpenter and S. Grossberg, A massively parallel architecture for a self-organizing neural pattern recognition machine, Computer Vision Graphics Image Process. 37(1987)54-115.

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

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

[5] O. Hoshino, S. Inoue, Y. Kashimori, and T. Kambara, A hierarchical dynamical map as a basic frame for cortical mapping and its application to priming, Neural Comput. 13(2001)1781-1810.

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

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

[8] T. 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.

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

[10] E.T. Rolls, and G. Deco, Computational neuroscience of vision, Oxford University Press, 2002.

[11] N. Suzuki, N. Hashimoto, Y. Kashimori, M. Zheng, and T. Kambara, A neural model of predictive recognition in form pathway of visual cortex, in: Proc. of Information Processing of Cells and Tissues 2003, Lausanne, Switzerland, 2003, pp.105-122..

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

[13] 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, pp. 383-400.

[14] 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 neural networks; 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 denote the attractors of visual memory. The solid line and dotted one indicate the feedforward and feedback signals, respectively.

Fig. 2. (a) Two kinds of retinal 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 resolutions, respectively. (c) Responses of V1, V4, and TEO networks to the retinal image A. The filled rectangles in each V1 and V4 maps indicate the firing neurons tuned to the object features at the each resolution level. The filled rectangle of TEO means the firing neuron tuned to coarse image (elemental figure) of a whole object.

Fig. 3. (a) Three pairs of attractors constructing the object memory. Each pair consists of the part form (X, |, ^ etc.) and the position (top, middle, and bottom). The object memory for A and B consist of the attractors a_1 ~ a_3 and b_1 ~b_3, respectively. (b) Variation of dynamic state of WM network during the form perception. A short-vertical bar on row corresponding to a_1 ~ b_3 indicates that the network activity stays in the attractor. The attention signal x means the application of input signal corresponding to the same firing pattern as that of attractor x. The switching bias means the uniform impulse stimulus to each WM neuron in order to change the dynamical state of WM network.

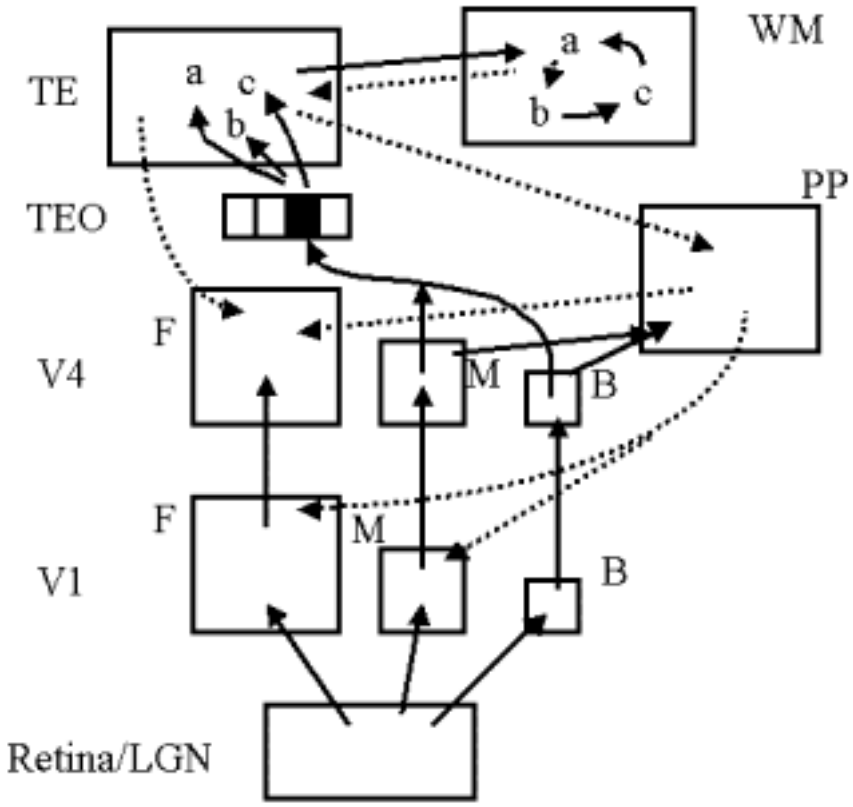

Fig. 1 Y. Kashimori et al.

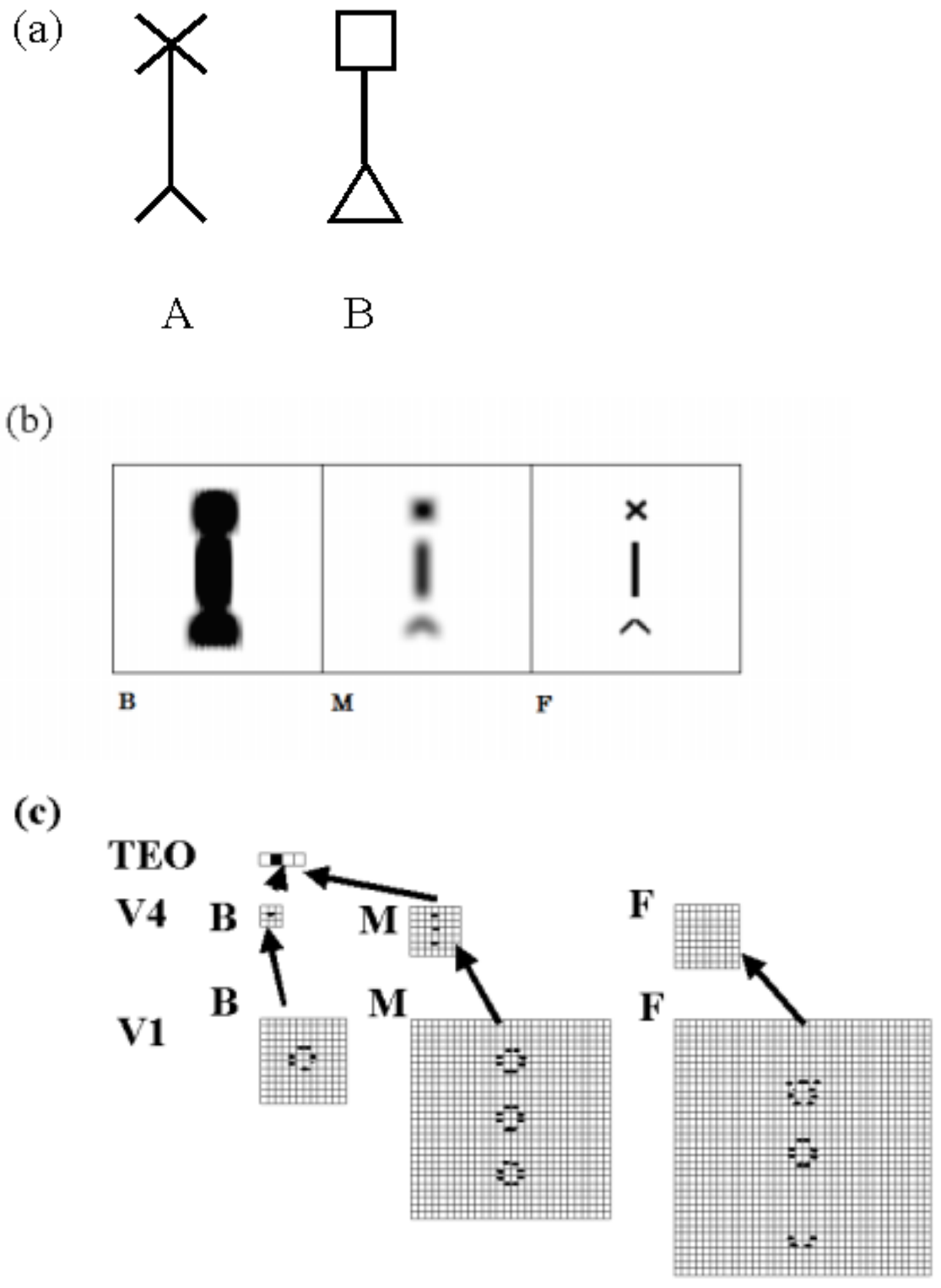

Fig. 2 Y. Kashimori et al.

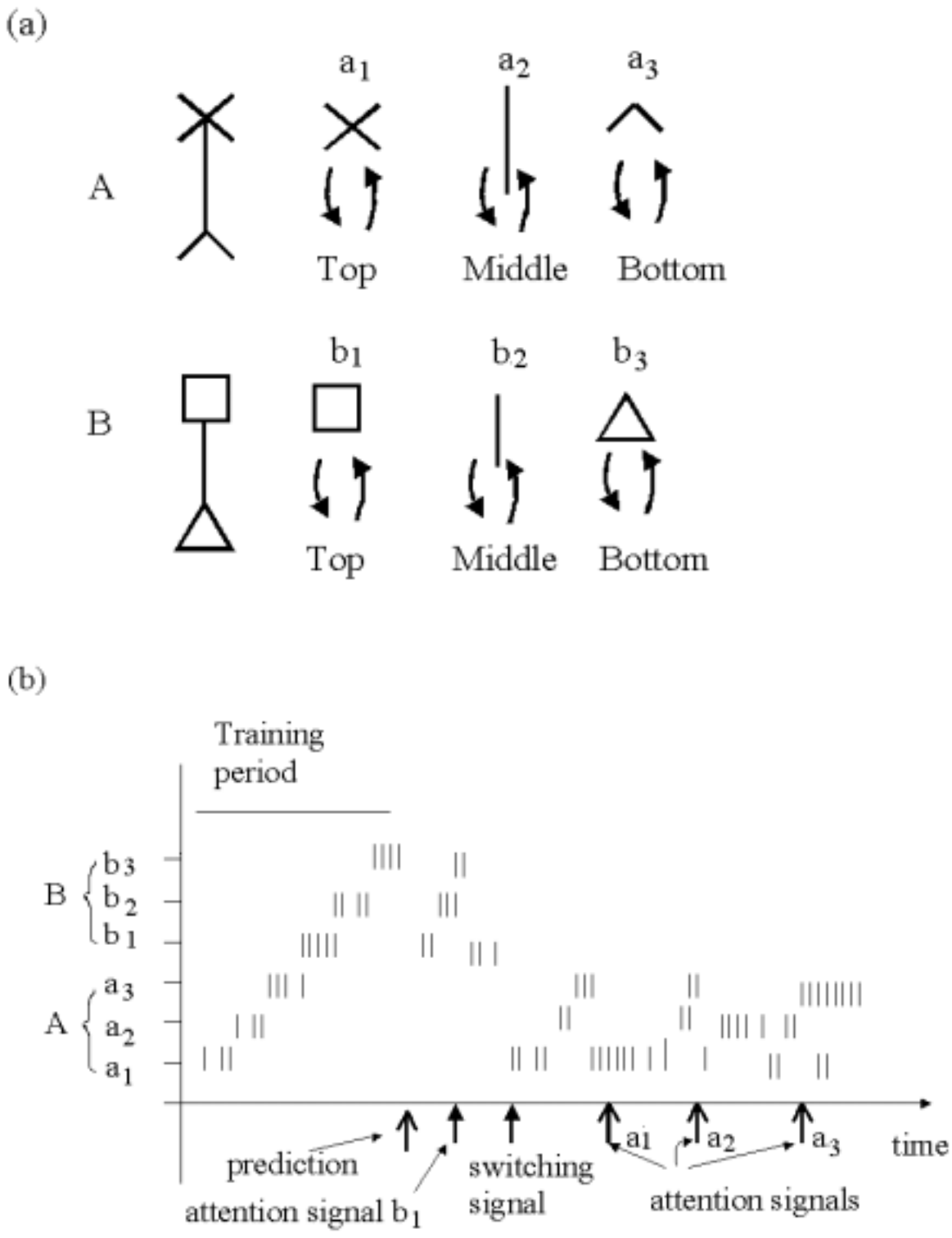

Fig. 3 Y. Kashimori et al.

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 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 a student in the Graduate School of Information Systems at the University of Electro-Communications. His research interest is to clarify the neural mechanisms of electrolocation, sound localization, and echolocation and visual recognition mechanism using the in silico method.

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 and of visual information processing using psychophysical approach.

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 auditory, visual, and electro-sensory systems, and emergence of dynamical orders in various biological complex systems. His research work is made using the "in silico" method.