Neuronal Mechanisms for Hierarchical Encoding in Inferior-Temporal Cortex

Narihisa Matsumoto
Graduate School of Science and Engineering, Saitama University, JAPAN:
RIKEN Brain Science Institute, JAPAN
xmatumo@brain.riken.go.jp

Masato Okada
RIKEN Brain Science Institute, JAPAN:
"Intelligent Cooperation and Control", PRESTO, JST, JAPAN
okada@brain.riken.go.jp

Abstract

Sugase et al. found that global information is processed at initial transient firings of face-responsive neurons in the inferior-temporal (IT) cortex, and that finer information is processed at the subsequent sustained firings. An attractor network is employed to elucidate neuronal mechanisms producing this dynamics. The results of computer simulations show that the behavior of both neuronal population and a single neuron qualitatively coincides with Sugase's data. Furthermore, we propose two physiological experiments and predict the results from our model. Therefore, if actual experimental results coincide with our predicted results, the attractor network might be the neuronal mechanisms of their data.

1 Introduction

In the inferior-temporal (IT) cortex, which is regarded as a final stage of the visual cortex, some neurons respond to faces or complex objects [1, 2]. Previous works on neuronal activity indicated that a pattern recognition is complete at initial transient firings of the neurons [3]. However, recent experimental results were shown, which suggest the possibility that not only the stationary state of the neuronal activity but its dynamics encodes some information [4, 5]. Sugase et al. recorded the activity of the face-responsive neurons in the IT cortex by using single-unit electrodes, while presenting visual stimuli [4]. The visual stimuli in their experiment consisted of 12 human faces (3 models with 4 expressions), 16 monkey faces (4 models with 4 expressions), and 10 simple objects (2 models with 5 colors). The relationship among the stimuli was hierarchical because the stimuli were categorized into global categories, e.g., human, monkey and object, and finer categories, e.g., individual and expression. The temporal change in neuron firing rates indicated that the single neuron encodes the information of the global categories at the initial transient firing and that of the finer categories at the subsequent sustained firing. This implies that the hierarchical relationship among the visual stimuli is extracted and coded in the transient and the sustained activity of the single neuron.

It is possible that a feed-forward model and an attractor network can explain Sugase's findings [3, 6, 7, 8]. Two different pathways that reach the IT cortex should be considered in the framework of the feed-forward model. Each pathway is assumed to have different resolution and different transmission time of the information. In the faster pathway, that is, the transmission

time is shorter, a visual stimulus is processed by a coarse resolution. In the slower pathway, it is processed by a fine resolution. The global information is processed on the faster pathway, while the finer information is processed on the slower pathway. Thus, the feed-forward model might be able to explain Sugase's data. The attractor network is assumed to involve recurrent connections, e.g., horizontal connections within the IT cortex, or feed-back connections from the higher visual area. The process time of the finer information gets longer since the recurrent process takes longer [9]. Thus, the attractor network might be able to explain their data.

In this paper, we employ the attractor network to explain Sugase's data. Amit et al. showed that the stationary state of the network might produce the sustained activity of the IT neurons by employing the associative memory model [8]. It is easily anticipated that the convergence process to the attractor, i.e., attractor dynamics, might produce the dynamics of the face-responsive neurons. Amari found that when the memory patterns with the hierarchical structure are stored in the associative memory model, not only a memory pattern but a mixed state, which is a nonlinear superposition of the memory patterns, becomes an attractor spontaneously [10]. Sugase's data indicates that the activity of the neuronal population is initially a superposition of patterns representing the different faces, and then it converges to a single pattern representing the specific face [11]. The mixed state represents the superposition of the different faces while each stored memory pattern represents each face. The results of computer simulations show that the macroscopic state of the network approaches the mixed state, and finally it converges to the memory pattern. The results qualitatively coincide with the population dynamics of the face-responsive neurons [11]. The behavior of a single neuron in our model also coincides with their data.

We propose two decisive physiological experiments to differentiate the feed-forward model and the attractor network. The first experiment uses noise-degraded images [12, 13]. The other experiment is a cooling experiment [14, 15, 16]. The results from the attractor network are different from those of the feed-forward model. Therefore, the results obtained by future physiological experiments will be able to judge which model might be the neuronal mechanisms of Sugase's data.

2 Model

Our model consists of excitatory and inhibitory neurons [8]. The excitatory connections from the excitatory neurons to the inhibitory neurons are uniform. The inhibitory connections from the inhibitory neurons to the excitatory neurons are also uniform. There are no connections between each inhibitory neuron. Therefore, the population of the inhibitory neurons can be regarded as a single inhibitory neuron. The equations that describe the model are as follows:

$$\tau_{ex}\dot{I}_i^{ex} = -I_i^{ex} + J\sum_{j\neq i}^{N} J_{ij}V_j^{ex} - T + H_i^{ext}, \tag{1}$$

$$V_i^{ex} = \phi(I_i^{ex}), \tag{2}$$

$$\tau_{in}\dot{I}^{in} = -I^{in} + \frac{1}{fN}\sum_{i=1}^{N} V_i^{ex}, \tag{3}$$

$$T = k \times V^{in} = k \times \psi(I^{in}), \tag{4}$$

where N is the number of excitatory neurons and J_{ij} is the synaptic weight from the j-th neuron to the i-th neuron. J adjusts the gain of all synaptic weights among the excitatory neurons. τ_{ex} and τ_{in} are time constants of the excitatory neurons and the inhibitory neuron, respectively. I_i^{ex} is the membrane potential of the i-th excitatory neuron while I^{in} is the membrane potential of the inhibitory neuron. V_i^{ex} is the output of the i-th excitatory neuron while V^{in} is the output of

the inhibitory neuron. $\phi(I)$ is a saturation output function of the excitatory neurons while $\psi(I)$ is a threshold-linear output function of the inhibitory neuron. The outputs of these functions are mean firing rates of the neurons. f denotes the mean firing rate of the memory pattern.

The visual stimuli that Sugase et al. used had the hierarchical structure. Ultrametric patterns $\boldsymbol{\xi}^{\mu,\nu}$ are used as the memory patterns. The elements of the memory pattern are binary, $\{0, 1\}$. If $\xi_i^{\mu,\nu} = 1$, the mean firing rate takes the maximal value of the i-th neuron, otherwise it takes 0. The procedure for generating the ultrametric patterns is shown as follows. At first, each element ξ_i^{μ} of a parent pattern $\boldsymbol{\xi}^{\mu}$ is generated independently by $\mathrm{Prob}[\xi_i^{\mu} = 1] = 1 - \mathrm{Prob}[\xi_i^{\mu} = 0] = f$. Next, s memory patterns $\boldsymbol{\xi}^{\mu,\nu}$ are generated, preserving a correlation with $\boldsymbol{\xi}^{\mu}$. μ denotes a group number while ν denotes an element number. The memory patterns belonging to the same group are correlated, while the memory patterns belonging to the different groups are uncorrelated. Consequently, the memory patterns are structured hierarchically. The parent pattern $\boldsymbol{\xi}^{\mu}$ is regarded as the representative pattern of the group μ since $\boldsymbol{\xi}^{\mu}$ is considered to be average of the memory patterns $\boldsymbol{\xi}^{\mu,\nu}$ belonging to the same group μ. The synaptic weight J_{ij} from the j-th excitatory neuron to the i-th excitatory neuron is determined by the Hebb rule [17]: $J_{ij} = \frac{1}{fN}\sum_{\mu=1}^{p}\sum_{\nu=1}^{s}\xi_i^{\mu,\nu}\xi_j^{\mu,\nu}$. Note that the representative patterns $\boldsymbol{\xi}^{\mu}$ are not stored in the synaptic connections. External input $\boldsymbol{H}^{ext}$ is injected into only the excitatory neurons. When the stimulus $\boldsymbol{\xi}^{\mu,\nu}$ is input, the external input $\boldsymbol{H}^{ext}$ follows the memory pattern $\boldsymbol{\xi}^{\mu,\nu}$, i.e., $H_i^{ext} = H\xi_i^{\mu,\nu}$, where H denotes the coefficient of the external input.

All initial values of $\boldsymbol{I}$ are set at 0. The values of the parameters are set as follows: $N = 3000$, $p = 20$, $s = 3$, $C = 0.3$, $\tau_{ex} = 10$ ms, $\tau_{in} = 2$ ms, $J = 1.0$, $k = 8.0$, $f = 0.05$, $H = 0.055$.

3 Results

3.1 Population Behavior

We introduce a measure to describe the macroscopic state of the network, i.e., behavior of the neuronal population. This is called "overlap" and represents a distance between the memory pattern $\boldsymbol{\xi}^{\mu,\nu}$ and the vector of the outputs $\boldsymbol{V}^{ex}$. $m^{\mu,\nu}$ represents the distance between $\boldsymbol{\xi}^{\mu,\nu}$ and $\boldsymbol{V}^{ex}$ when the stimulus $\boldsymbol{\xi}^{\mu,\nu}$ is presented, and is calculated as follows:

$$m^{\mu,\nu}(t) = \frac{1}{fN}\sum_{i=1}^{N}\xi_i^{\mu,\nu}V_i^{ex}(t). \tag{5}$$

As $m^{\mu,\nu}$ approaches 1, $\boldsymbol{V}^{ex}$ converges to $\boldsymbol{\xi}^{\mu,\nu}$. When $\boldsymbol{\xi}^{1,2}$ is input, the temporal change of the neuronal population obtained by computer simulations is drawn in Figure 1(a). The solid line

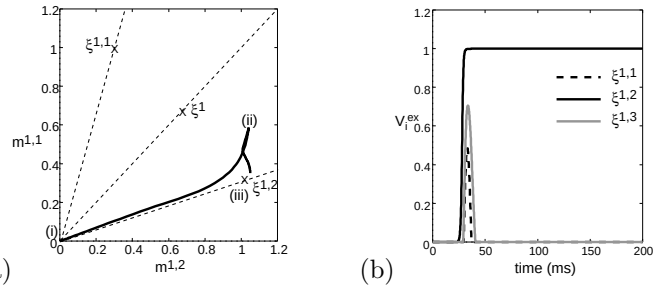

Figure 1: (a): temporal change of the neuronal population when $\boldsymbol{\xi}^{1,2}$ is input. The solid line shows a vector trajectory of $m^{1,2}$ between $\boldsymbol{\xi}^{1,2}$ and $\boldsymbol{V}^{ex}$. (b): temporal change of $(0, 1, 0)$ neuron.

shows a vector trajectory of $m^{1,2}$ between $\boldsymbol{\xi}^{1,2}$ and $\boldsymbol{V}^{ex}$. The abscissa axis is $m^{1,2}$, while the vertical axis is $m^{1,1}$ between $\boldsymbol{\xi}^{1,1}$ and $\boldsymbol{V}^{ex}$. (i) shows the initial state of the network. Then,

no neurons fire and $m^{1,1} = m^{1,2} = 0$. (ii) shows that the macroscopic state approaches the representative pattern $\boldsymbol{\xi}^1$ (t $\approx$ 33ms). (iii) shows that the macroscopic state converges to the memory pattern $\boldsymbol{\xi}^{1,2}$. This temporal behavior of the neuronal population is consistent with Sugase's data [11].

3.2 Single Neuron Behavior

Next, we compare the temporal behavior of a single neuron in our model with in Sugase et at.'s data. The neurons in our model are classified into $8(= 2^3)$ classes according to the stored patterns $(\xi_i^{1,1}, \xi_i^{1,2}, \xi_i^{1,3})$. These 8 classes can be effectively classified into three groups: $(0, 1, 0)$, $(1, 1, 1)$ and $(1, 0, 1)$. Hereafter, we call the i-th neuron with $(\xi_i^{1,1}, \xi_i^{1,2}, \xi_i^{1,3}) = (0, 1, 0)$ the "(0,1,0) neuron" and so on. If the first group ($\mu = 1$) is regarded as the human category, $\boldsymbol{\xi}^{1,1}$, $\boldsymbol{\xi}^{1,2}$, and $\boldsymbol{\xi}^{1,3}$ correspond to the first individual, the second one and the third one, respectively. The $(0, 1, 0)$ neuron has the finer information to discriminate the second human individual from other individuals since it fires when $\boldsymbol{\xi}^{1,2}$ is input and does not fire when either $\boldsymbol{\xi}^{1,1}$ or $\boldsymbol{\xi}^{1,3}$ is input. Figure 1 (b) shows the temporal change of the $(0, 1, 0)$ neuron obtained by the computer simulations. The lines of $\boldsymbol{\xi}^{1,1}$, $\boldsymbol{\xi}^{1,2}$ and $\boldsymbol{\xi}^{1,3}$ are the firing patterns for the first individual, the second one, or the third one, respectively. The abscissa axis is time (ms), while the vertical axis is firing rate normalized by the maximal firing rate of the neuron. The neuron responds to $\boldsymbol{\xi}^{1,2}$ persistently, while it responds to $\boldsymbol{\xi}^{1,1}$ and $\boldsymbol{\xi}^{1,3}$ transiently. This behavior of the $(0, 1, 0)$ neuron coincides with that of a single neuron in Sugase et al's data [18].

The $(1, 1, 1)$ neuron codes all of $\boldsymbol{\xi}^{1,1}$, $\boldsymbol{\xi}^{1,2}$ and $\boldsymbol{\xi}^{1,3}$ as "1" while it codes all of the memory patterns belonging to the different groups as "0" under a sparse coding scheme, i.e., $f \to 0$. This neuron has the global information to discriminate the memory patterns in the first group from in the other groups. On the other hand, it does not have the finer information to discriminate the memory pattern in the first group from the other memory patterns in the first group. Computer simulations show that the neuron responds to $\boldsymbol{\xi}^{1,1}$, $\boldsymbol{\xi}^{1,2}$ and $\boldsymbol{\xi}^{1,3}$ persistently.

The $(1, 0, 1)$ neuron codes $\boldsymbol{\xi}^{1,1}$ and $\boldsymbol{\xi}^{1,3}$ as "1", while it codes $\boldsymbol{\xi}^{1,2}$ as "0". This neuron has smaller amount of the finer information to discriminate $\boldsymbol{\xi}^{1,2}$ from $\boldsymbol{\xi}^{1,1}$ and $\boldsymbol{\xi}^{1,3}$ than the $(0, 1, 0)$ neuron since the probability of "0" is higher than that of "1" under the sparse coding scheme. Computer simulations show that the neuron responds to $\boldsymbol{\xi}^{1,2}$ transiently, while it responds to $\boldsymbol{\xi}^{1,1}$ and $\boldsymbol{\xi}^{1,3}$ persistently.

3.3 Prediction of Results of Proposed Experiments

Here, we propose two decisive physiological experiments to defferentiate the feed-forward model and the attractor network. One is the experiment where the noise-degraded images are used. To mimic the noise-degraded images, the external input $\boldsymbol{H}^{ext}$ obeys the following relationship. If $\xi_i^{1,\nu} = 1$, $H_i^{ext} = H$ with the probability of $1 - a$ and $H_i^{ext} = 0$ with the probability of a. If $\xi_i^{1,\nu} = 0$, $H_i^{ext} = 0$. The increase of a corresponds to that of noise. Figure 2 shows the predicted results from the computer simulations at $a = 0.7$. Figure 2(a) shows the temporal change of the neuronal population, that is, the vector trajectory of the overlap when the stimulus $\boldsymbol{\xi}^{1,2}$ is input. This figure is drawn by the same procedure as in Figure 1(a). Figure 2(b) shows the firing patterns of the $(0, 1, 0)$ neuron when $\boldsymbol{\xi}^{1,1}$, $\boldsymbol{\xi}^{1,2}$, $\boldsymbol{\xi}^{1,3}$ are input, respectively. This figure is drawn by the same procedure as in Figure 1(b).

As the noise increases, the behaviors of both the neuronal population and the $(0, 1, 0)$ neuron change drastically. The final state of the neuronal population approaches not the stimulus $\boldsymbol{\xi}^{1,2}$ but the representative pattern $\boldsymbol{\xi}^1$. The increase of the noise divides the firing pattern of the $(0, 1, 0)$ neuron into three parts. The first part represents the global information, the second one does the finer information, and the third one does the "global" information. The global information of the first part is the same as that of the third part. The first and the second

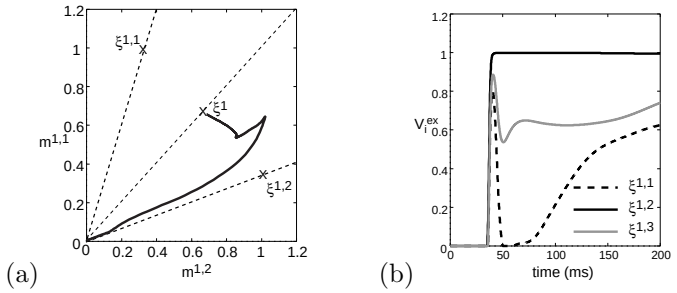

Figure 2: The predicted results of the experiment that uses noise-degraded images at $a = 0.7$. (a): temporal change of the neuronal population. (b): temporal change of $(0, 1, 0)$ neuron.

parts are the same as in the normal experiment. Moreover, the convergence time to the final state gets longer. This is consistent with the result obtained by Shidara et al. [12]. As for the $(1, 1, 1)$ neuron, the behavior does not change qualitatively. As for the $(1, 0, 1)$ neuron, each firing pattern for $\boldsymbol{\xi}^{1,1}$ or $\boldsymbol{\xi}^{1,3}$ diverges to three parts. Moreover the convergence time to the final state gets longer.

The other proposed experiment is a cooling experiment. The IT cortex is cooled down and becomes inactive [14, 15, 16]. The horizontal connections within the IT cortex may become ineffective. To mimic the cooling experiment in our model, the connections between the excitatory neurons are weakened. The value of J in Equation (2) denotes the ratio of the weakened excitatory connections and the normal ones. As J decreases, the excitatory connections weaken. Consequently, the macroscopic state of the network directly approaches the memory pattern $\boldsymbol{\xi}^{1,2}$ in the retrieval process. The $(0, 1, 0)$ neuron responds to $\boldsymbol{\xi}^{1,2}$, while it does not respond to $\boldsymbol{\xi}^{1,1}$ and $\boldsymbol{\xi}^{1,3}$. This implies that the global information is not represented on the firing pattern, while the finer information is. In other words, the hierarchical representation vanishes. As for the $(1, 1, 1)$ neuron, the behavior does not change qualitatively since this neuron does not originally have the finer information. As for the $(1, 0, 1)$ neuron, it responds to $\boldsymbol{\xi}^{1,1}$ and $\boldsymbol{\xi}^{1,3}$, while it does not respond to $\boldsymbol{\xi}^{1,2}$.

4 Summary and Discussion

We employed the attractor network to elucidate the neuronal mechanisms producing the dynamics of the face-responsive neurons in the IT cortex [4]. Our model stored the memory patterns with the hierarchical structure corresponding to the hierarchical relationship among the visual stimuli. The results of the computer simulations showed that the macroscopic state of the network approaches the mean state of similar patterns transiently, and finally it converges to the memory pattern. This implies that the hierarchical structure of the memory patterns is extracted, which qualitatively coincides with the population dynamics of the face-responsive neurons [11] The behavior of the $(0, 1, 0)$ neuron also coincides with the dynamics of the single neuron [18]. We proposed two decisive physiological experiments to differentiate the feed-forward model and the attractor network. The first experiment was to present the noise-degraded images. The predicted results of the $(0, 1, 0)$ neuron from the attractor network showed that the increase of the noise divides the firing pattern into three parts and that the convergence time to the final state gets longer. On the other hand, the predicted results of the $(0, 1, 0)$ neuron from the feed-forward model show that The finer information processed in the slower pathway disappears, since the fine features disappear in the noise-degraded images, Thus, the firing patterns represent only the global information. Moreover, the convergence time to the final state gets shorter. The results of the feed-forward model contradict with the physiological data [12] and our predicted results. The other experiment was to cool down the IT cortex and made it inactive [14, 15, 16]. Then, the horizontal connections within the IT cortex might become

ineffective. The predicted results of the $(0, 1, 0)$ neuron from the attractor network showed that the global information is not represented on the firing pattern, while the finer information is. In other words, the hierarchical representation vanishes since the recurrent process vanishes. On the other hand, the predicted results of the $(0, 1, 0)$ neuron from the feed-forward model show that since the cooling has little influence on the feed-forward connections, the firing pattern remains unchanged by the cooling. In other words, the cooling experiment gives no impact on the results of the feed-forward model. The results of the feed-forward model contradict with our predicted results. In both proposed experiments, the results of the attractor network contradict the results of the feed-forward model. Therefore, the results obtained by the future physiological experiments will be able to differentiate the attractor network and the feed-forward model and to judge which model might be the neuronal mechanisms producing the dynamics of the face-responsive neurons in the IT cortex.

Acknowledgements

We would like to thank Dr. Sugase and Dr. Yamane, who gave us their data and a lot of useful comments about our research. We also would like to thank Dr. Kawato and Dr. Doya, who gave us a lot of valuable comments. This work was partially suported by Grant-in-Aid for the scientific research No.1458043 and 14084212.

References

[1] C. Bruce, R. Desimone, and C. G. Gross, Journal of Neurophysiology **46**, 369 (1981).

[2] I. Fujita, K. Tanaka, M. Ito, and K. Cheng, Nature **360**, 343 (1992).

[3] E. T. Rolls and M. J. Tovee, Proceedings of the Royal Society of London Series B; Biological Science **257**, 9 (1994).

[4] Y. Sugase, S. Yamane, S. Ueno, and K. Kawano, Nature **400**, 869 (1999).

[5] R. W. Friedrich and G. Laurent, Science **291**, 889 (2001).

[6] G. Wallis, Network: Computation in Neural Systems **9**, 265 (1998).

[7] D. J. Amit, *Modeling brain function: the world of attractor neural networks* (Cambridge university press, 1989).

[8] D. J. Amit, N. Brunel, and M. V. Tsodyks, Journal of Neuroscience **14**, 6435 (1994).

[9] S. Panzeri, E. T. Rolls, F. Battaglia, and R. Lavis, Network: Computation in Neural Systems **12**, 423 (2001).

[10] S. Amari, Biological Cybernetics **26**, 175 (1977).

[11] N. Matsumoto, M. Okada, K. Doya, Y. Sugase, and S. Yamane, in *Society for Neuroscience abstracts* (2001), vol. 27, p. 398.5.

[12] M. Shidara, S. Liu, and B. J. Richmond, in *Society for Neuroscience Abstract* (1996), vol. 22, p. 1615.

[13] D. J. Amit and N. Brunel, Cerebral Cortex **7**, 237 (1997).

[14] P. Girard, P. A. Salin, and J. Bullier, Journal of Neurophysiology **67**, 1437 (1992).

[15] D. Ferster, S. Chung, and H. Wheat, Nature **380**, 249 (1996).

[16] J. A. Horel, Behavioural Brain Research **76**, 199 (1996).

[17] D. O. Hebb, *The Organization of Behavior: A Neuropsychological Theory* (Wiley, New York, 1949).

[18] Y. Sugase, Ph.D. thesis, University of Tokyo (1999).