2004 International Symposium on Nonlinear
Theory and its Applications (NOLTA2004)
Fukuoka, Japan, Nov. 29 - Dec. 3, 2004

Hierarchical encoding of faces and its neuronal mechanism in the inferior-temporal cortex

M. Okada[†‡], N. Matsumoto[‡*], K. Toya[**], Y. Sugase-Miyamoto[*], and S. Yamane[*]

†Department of Complexity Science and Engineering, University of Tokyo Kashiwa, Chiba, Japan
‡"Intelligent Cooperation and Control", PRESTO, JST, Japan
∗ Neuroscience Research Institute, AIST, Tsukuba, Japan
∗∗ Department of Systems and Human Science, Osaka University Toyonaka, Osaka, Japan
Email: okada@brain.riken.jp

Abstract—Sugase et al. analyzed the time-course of the information carried by the firing of face-responsive neurons in the IT cortex. They found that the initial transient firing correlated well with a rough categorization, and the subsequent sustained firing represented more detailed information. In order to investigate the mechanism of such temporal dynamics, we employ a simple correlation-type associative memory model with distributed, hierarchically correlated memory patterns. We found that the retrieval dynamics can qualitatively replicate the temporal dynamics of face-responsive neurons. Based on this model, we propose a new physiological experiment employing a noise-degraded image.

1. Introduction

Sugase et al. [1] analyzed the time-course of the information carried by the firing of face-responsive neurons in the IT cortex, while performing a fixation task of monkey and human faces with various expressions, and simple geometrical shapes. They found that the initial transient firing correlated well with a rough categorization (e.g., face vs. non-face stimuli), and the subsequent sustained firing represented more detailed information (e.g., specific person or expression). They divided neurons into three groups, called *dual*-, *fine*- and *rough*-neurons, based on the information the neurons coded. Their results suggest that the neuron firing pattern is initially a superposition of patterns representing different faces or expressions, but it then converges to a single pattern representing a specific face or expression. In order to investigate the mechanism of such temporal dynamics, we employ a simple correlation-type associative memory model with distributed, hierarchically correlated memory patterns [2, 3], which may mimic their input stimuli. The model dynamics is stable not only for stored memory patterns but also for mixed states, which are superpositions of memory patterns. Based on analytical and numerical studies, we found that the retrieval dynamics can qualitatively replicate the temporal dynamics of face-responsive neurons as follows. Initially, the network state approaches a mixed state that is a superposition of patterns representing different persons or expressions. After that it diverges from the mixed state, and finally converges to a single memory pattern representing a specific person or expression. The model neurons can be classified into rough-, fine and dual neuron groups according to whether they are active in the mixed state or the specific memory pattern(s). The time constant of groups having the finer or rougher information is slower or faster, respectively. The neurons in the slowest group tend to become the dual neurons. Based on this model, we propose a new physiological experiment, where a noise-degraded image is employed as in the work of the Shidara et al. [4] (see also [5]). We conjecture that there is a critical noise value above which temporal behaviors of fine and dual-neurons dramatically change because the network state converges not to the memory pattern but to the mixed state.

2. Summary of experimental findings

We summarize the results of Sugase et al. [1]. An information-theoretic analysis [7, 8] was applied to qualitatively measure neural responses according to two different categorical levels. One level consisted of rough categorizations such as face vs. non-face stimuli, the other level corresponded to fine (more detailed) categorizations (e.g., facial expression or a specific person). They measured the temporal change of the mutual information $I(S;R)$ represented by the entropy $H(\cdot)$,

$$\begin{aligned} I(S;R) &= H(S) - H(S|R) \\ &= \sum_{s\in S}(-p(s)\log p(s)) \\ &\quad - < \sum_{s\in S}(-p(s|r)\log p(s|r)) >_{p(r)}, \end{aligned} \tag{1}$$

where S and R are the set of the stimuli s and the neuronal responses (spike counts) r, respectively. The bracket in eq. (1) stands for the average of the probability $p(r)$ of the spike count r. $p(s)$ and $p(s|r)$ are the prior probability of stimulus s and the conditional probability of stimulus s given by the spike count r, respectively.

Sugase et al. found that the initial transient firing correlated well with a rough categorization (e.g. face vs. non-face stimuli), and the subsequent sustained firing repre-

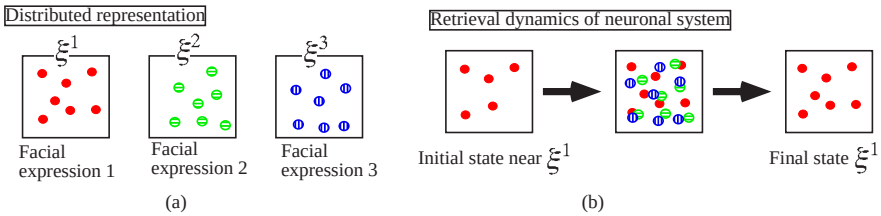

Figure 1: Schematic diagram of neural dynamics of face-responsive cells in IT cortex.

sented finer information regarding more detailed categorizations. Their results suggest that the neuron firing pattern is initially a superposition of patterns representing different faces or expressions, but it then converges to a single pattern representing a specific face or expression. Matsumoto et al. applied principal component analysbis and mixture of Gaussian analysis. They found that the population vectors of neural firing split into three clusters corresponding to human faces, monkey faces and other simple shapes in the early phase (90-140msec), and that these three clusters split into sub-clusters corresponding to finer categorizations in the later phase (140-190msec) [9].

Let us reinterpret this transient phenomenon in the manner presented below. The individual facial expressions (or faces) are encoded with the distributed representation as shown in Fig.1(a). Each circle represents a excited neuron in the neuron pools denoted by the rectangles. We might be able to say that a mixed state appears in the intermediate part of dynamics of neurons, and the network state finally converges to the memory pattern schematically shown in Fig. 1(b). The main purpose of the present work is to investigate the mechanism of such temporal dynamics.

3. Model

A recurrent neural network consisting of N neurons with an output function $\mathrm{sgn}(\cdot)$ is discussed, where $\mathrm{sgn}(u) = +1$ for $u \geq 0$, otherwise $\mathrm{sgn}(u) = -1$. We employ a case of the thermodynamic limit ($N \to \infty$) and the synchronous dynamics,

$$x_i^{t+1} = \mathrm{sgn}(\sum_{j \neq i}^{N} J_{ij} x_j^t), \tag{2}$$

where x_i^t represents a state of the i-th neuron at discrete time t, and J_{ij} denotes a synaptic weight from the j-th neuron to the i-th neuron. In order to mimic the stimuli of Sugase et al., we utilize a set of ultrametric patterns [6]. For simplicity, we treat two stages of hierarchy which is the most simple case. This can be easily extended to more complex situations of multistage and/or inhomogeneous hierarchies, and the results do *not* qualitatively change. One can use many procedures to make the set of ultrametric patterns, but we employ the following method. Each element $\xi_i^{\mu\nu}$ of memory pattern $\boldsymbol{\xi}^{\mu\nu}$ is independently generated using eqs. (3) and (4),

$$\mathrm{Prob}[\xi_i^{\mu} = \pm 1] = 1/2, \tag{3}$$

$$\mathrm{Prob}[\xi_i^{\mu\nu} = \pm 1] = (1 \pm b\xi_i^{\mu})/2. \tag{4}$$

The distance between memory patters $\boldsymbol{\xi}^{\mu\nu}$ with the same parent $\boldsymbol{\xi}^{\mu}$ is represented by a correlation matrix $\mathbf{B}$,

$$(\mathbf{B})_{\nu\nu'} = \mathrm{E}[\xi_i^{\mu\nu}\xi_i^{\mu\nu'}] = \delta_{\nu\nu'} + b^2(1-\delta_{\nu\nu}). \tag{5}$$

If their parents are different, they are completely orthogonal to each other. Thus, we have the two stage ultrametric structure in the set of $\boldsymbol{\xi}^{\mu\nu}$'s. The synaptic coupling is determined by a simple correlation learning,

$$J_{ij} = \frac{1}{N}\sum_{\mu=1}^{\alpha N}\sum_{\nu=1}^{s}\xi_i^{\mu\nu}\xi_j^{\mu\nu}. \tag{6}$$

The number of clusters is αN, where α is called the loading rate and each cluster has s memory patterns. Many researchers have analyzed this model (for example, [2]) or our analyses in [3]. Other more complex and sophisticated learning methods storing the hierarchically correlated patterns have been proposed compared with eq. (6) (for example, [10]). However, most of them have two kinds of essential drawbacks. One is from an information-theoretic view point. These method are not local with respect to the pattern suffixes, μ, ν; for example, they need an information about which cluster the memory pattern belongs to. The other one is strictly essential in treating the target experiment of Sugase et al. The ultrametric structure of the memory pattern is completely destroyed in these learning methods. As a result, we can definitely not mimic the transient phenomena observed by Sugase et al. by using these methods.

4. Equilibrium properties

In the present model, not only the stored pattern $\boldsymbol{\xi}^{\mu\nu}$ but also mixed state, $\boldsymbol{\eta}^{\mu} = \mathrm{sgn}(\sum_{\nu=1}^{s}\boldsymbol{\xi}^{\mu\nu})$, which is a nonlinear superposition of the memory patterns, $\boldsymbol{\xi}^{\mu\nu}$, $\nu = 1, \cdots, s$, is stable. In order to investigate the stabilities of the memory patterns $\boldsymbol{\xi}^{\mu\nu}$ and the mixed states $\boldsymbol{\eta}^{\mu}$, we have derived order parameter equations that describe the the equilibrium state of eqs. (2) through (4) using the SCSNA [11]. Details of

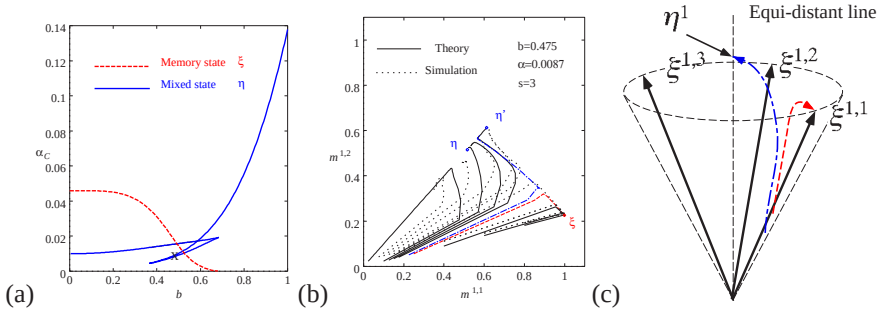

Figure 2: (a): Phase diagram of equilibrium states. (b): Retrieval dynamics when $s = 3$, $b = 0.45$ and $\alpha = 0.0087$. (c): Schematical description of trajectories denoted by dashed and dash-dotted lines in (b).

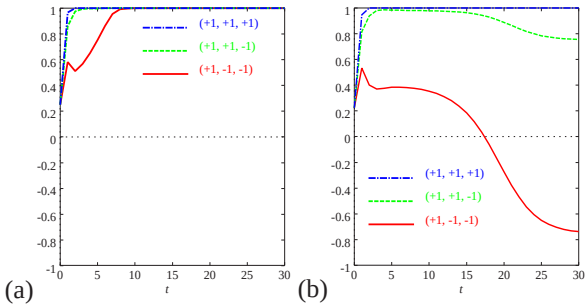

Figure 3: Retrieval dynamics of each sub-lattice. (a): Memory retrieval denoted by dashed line in Fig. 2(b). (b): Mixed state retrieval denoted by dash-dotted line in Fig. 2(b).

this section are described in [3]. We solve the order parameter equations, and show in Fig. 2(a) the resultant phase diagram of $s = 3$ on a parameter space of the loading α and b, representing the correlation among the memory patterns in the same cluster. The dashed and solid lines represent the storage capacities for the memory pattern and the mixed state, respectively. Each equilirium state becomes unstable above its value of storage capacity. When $b \leq b_C = 1/\sqrt{s-1} = 1/\sqrt{2}$, the memory patterns and the mixed states coexist as equilibria.

5. Dynamical properties

We employ the full-order version of the statistical neurodynamical method [12, 13] to investigate the retrieval process. The static and dynamic properties obtained by this theory were confirmed fairy well in extensive numerical simulations. Fig. 2(b) shows the trajectories of temporal evolutions of overlaps of the state $\boldsymbol{x}^t$ with $\boldsymbol{\xi}^{1,1}$ or $\boldsymbol{\xi}^{1,2}$, respectively,

$$m_t^{1,\nu} = \frac{1}{N}\sum_{i=1}^{N} \xi_i^{1,\nu} x_i^t. \tag{7}$$

The parameters are set to $\alpha = 0.087$, $b = 0.475$ and $s = 3$, respectively. Since the $\times$ in the bi-stable region of Fig. 2(a) shows this parameter set, the memory pattern (ξ in Fig. 2(b)) and two kinds of the mixed states (η and η') coexist. We have set the initial state $\boldsymbol{x}^0$ as $\mathrm{Prob}[x_i = \pm 1] = (1 \pm m_0^{1,1}\xi_i^{1,1})/2$. Since we obtain $m_0^{1,2} = m_0^{1,3} = b^2 m_0^{1,1}$ in this case, only $m_t^{1,1}$ and $m_t^{1,2}$ have been plotted, and the initial states are arranged in the line of $m^{1,2} = b^2 m^{1,1}$. The solid and dotted lines denote the theoretical and computer simulation result (N=30,000), respectively. All trajectories regarding $\boldsymbol{\xi}^{1,1}$ retrieval cases are inside the line $m^{1,2} = b^2 m^{1,1}$ as shown in Fig. 2(b). This fact means that the network state state approaches the mixed state at the first stage of the retrieval process. After that the network state parts from the mixed state, and finally converges to $\boldsymbol{\xi}^{1,1}$. These trajectories are schematically shown as the red dashed line in Fig. 2(c). The memory patterns and the mixed states are embedded in the surface of the cone and the equi-distant line, respectively.

Sugase et al. divided neurons into three groups, called dual-, fine- and rough-neurons, based on the information they code [1]. On the other hand, we divide the (model-)neurons into 2^s groups using the sub-lattice method. For example, a model neuron x_i is such that $(\xi_i^{1,1}, \xi_i^{1,2}, \xi_i^{1,3}) = (+1, +1, +1)$. Thus, we have eight groups for $s = 3$ according to $(\xi_i^{1,1}, \xi_i^{1,2}, \xi_i^{1,3}) = (\pm 1, \pm 1, \pm 1)$. These eight groups can be effectively classified into three groups,

$$(\xi_i^{1,1}, \xi_i^{1,2}, \xi_i^{1,3}) = (+1, +1, +1),\ (+1, +1, -1),\ (+1, -1, -1), \tag{8}$$

because we discuss the 50% firing rate case, and have the symmetry between $m_t^{1,2}$ and $m_t^{1,3}$. Hereafter, we call the neuron x_i with $(\xi_i^{1,1}, \xi_i^{1,2}, \xi_i^{1,3}) = (+1, +1, +1)$ the "$(+1, +1, +1)$ neuron" and so on. The $(+1, -1, -1)$ neuron codes the largest information about a classification of $\boldsymbol{\xi}^{1,1}$ among the three groups, while the $(+1, +1, +1)$ one does not have this information. On the other hand, the $(+1, +1, +1)$ neuron has much information about the cluster $\mu = 1$ for the memory patterns $\boldsymbol{\xi}^{1,\nu}$, because of their finite correlation, i.e., $b \neq 0$. Thus, the $(+1, -1, -1)$ neuron codes the fine information on the detailed categorization among

$\boldsymbol{\xi}^{1,\nu}$, while the $(+1,+1,+1)$ codes the rough categorization. From these considerations, we find correspondence between the fine-neuron of Sugase et al. and the $(+1,-1,-1)$ neuron as well as that between the rough-neuron and the $(+1,+1,+1)$ neuron. Each line in Fig. 3(a) shows the temporal evolution of the averaged output of each neuron group for the red dashed line in Fig. 2(b). The time constant of the relaxation computation for $(+1,+1,+1)$ and $(+1,-1,-1)$ is the fastest and slowest, respectively. This theoretical finding coincides with the results of Sugase et al. as described in §2. The nonmontonic evolution of the average output on the $(+1,-1,-1)$ neuron group around $t = 2$ implies that some of the $(+1,-1,-1)$ neurons, which correspond to the fine-neurons, tend to become the dual-neurons, because the number of the $(+1,-1,-1)$ neurons with the state -1, which corresponds to the mixed state, is locally maximized around $t = 2$. The dual-neuron dually codes the different categorization classes (rough and fine) using the different time domain. Their initial transient firing correlates with the rough categorization, and the sustained firing codes the finer information.

For the smaller value of the initial overlap $m_0^{1,1}$, that is, the noisy input, the network state converges to the mixed state rather than the memory state (for example, the blue dash-dotted line in Fig. 2(b)). This means that a critical initial overlap $m_C^{1,1}$ exists under which the retrieval dynamics dramatically change. These trajectories are schematically shown as the blue dash-dotted line in Fig. 2(c). Each line in Fig. 3(b) shows the temporal evolution of the averaged output of each neuron group for the blue dash-dotted line in Fig. 2(b). The red solid line on the $(+1,-1,-1)$ neuron suggests that the temporal evolutions of $(+1,-1,-1)$ neurons dramatically change when $m_0^{1,1} < m_C^{1,1}$. It would be very interesting if this dramatic change could be obtained experimentally. One candidate out of many is to employ a noise-degraded image as the Shidara et al. have done [4] (see also [5]). We conjecture that there is a critical noise value above which the temporal behaviors of fine and dual-neurons dramatically change as denoted by the red solid line in Fig. 3(b). This is because the network state converges not to the memory pattern but to the mixed state.

We have already presented a more realistic model based on the excitatory-inhibitory network to properly treat the regulator mechanism of the mean firing rate [14]. The results of the realistic model more accurately agreeed with the behavior of IT cortex neurons suggested by Sugase et al. than the previously discussed 50% firing model.

6. Discussion

Recently, Parga and Rolls have used the mixed state as a mechanism of invariant recognition under a coordinate transformation [15]. Their model has the following *duality* with the present model. We have employed the *structured* memory patterns and the *unstructured* learning method of eqs. (3) through (6), while they utilized the *unstructured* memory patterns and the *structured* learning method. These two variants of the Hopfield model share the same theoretical structure as shown in [15] and the present paper. However, the two variants have a completely different dynamical property from each other. We conjecture that Parga and Rolls' model will be unable to observe the initial transient effect described in the present paper. Thus, by comparing these variants with physiological data of the higher visual areas, we may discuss how complex ultrametric structures in the external visual world are encoded in these visual areas.

References

[1] Sugase, Y., Yamane, S., Ueno, S., and Kawano, K. (1999). *Nature*, **400**, 869-873.

[2] Amari, S. (1977). *Biological Cybernetics*, **26**, 175-185.

[3] Toya, K., Fukushima, K., Kabashima, Y., & Okada, M. (2000) *Journal of Physics A*, **33**, 2725-2738.

[4] Shidara, M., Liu, Z., & Richmond, B. J. (1996). Stimulus degradation has a similar effect on stimulus-related information in inferior temporal (IT) neurons and on the monkey's behavior. *Soc. Neurosci. Abstr.*, **22**, 1615.

[5] Amit, D. J., Fusi, S., & Yakovlev, V. (1997). *Neural computation*, **9**, 1071-1092.

[6] Rammal, R., Toulous, G., & Virasoro, M.A. (1986). *Reviews of Modern Physics*, **58**, 765-788.

[7] Kjaer, T. W., Hertz, J. A., & Richmond, B. J. (1994). *Journal of Computational Neuroscience*, **2**, 175-193.

[8] Kitazawa, S., Kimura, T., & Yin, PB. (1998). *Nature*, **392**, 494-497.

[9] Matsumoto, N., Okada, M., Sugase-Miyamoto, Y., Yamane, S., and Kawano, K. Population dynamics of face-responsive neurons in the inferior temporal cortex. to be published in *Cerabral Cortex*.

[10] Gutfreund, H. (1990). Physical Review A, **37**, 570-577.

[11] Shiino, M. and Fukai, T. (1992). *Journal of Physics A: Mathematical and General*, **25**, L375-L381.

[12] Okada, M. (1995). *Neural Networks*, **8**, 833-838.

[13] Okada, M. (1996). *Neural Network*, **9**, 1429-1458.

[14] Matsumoto, N., & Okada, M. Neuronal Mechanisms for hierarchical encoding in inferior-temporal cortex. to be published in *Neurocomputing*.

[15] Parga, N., & Rolls, E. (1998). *Neural Computation*, **10**, 1507-1525.