A Self-Organization Model of Feature Columns and Face Responsive Neurons in the Temporal Cortex

Takashi TAKAHASHI† and Takio KURITA‡
† Postdoctoral Research Fellow of the Japan Society for the Promotion of Science,
‡ Neuroscience Research Institute,
The National Institute of Advanced Industrial Science and Technology,
Ibaraki 305–8568, Japan
Takashi@Takahashi.com

Abstract

We investigate a self-organizing network model to account for the computational property of the inferotemporal cortex. The network can learn sparse codes for given data with organizing their topographic mapping. Simulation experiments are performed using real face images composed of different individuals at different viewing directions, and the results show that the network evolves the information representation which is consistent with some physiological findings. By analyzing the characteristics of the neuron activities, it is also demonstrated that the present model self-organizes the efficient representation for coding both of the global structure and the finer information of the face images.

1 Introduction

The temporal cortex is known to be responsive for complex patterns such as human faces. Recent experimental studies revealed some properties of neurons in this cortex. For instance, Young and Yamane[1] investigated the response properties of face responsive neurons in the inferotemporal cortex and showed that facial features are represented as the ensemble of sparse neuron activities. They called such coding scheme "sparse-population coding." On the other hand, Wang et al.[2] reported that the inferotemporal cortex consists of columnar modules and the activated region systematically moves on the cortical surface with the change of object views. These findings indicate that, in the inferotemporal cortex, objects are encoded as ensembles of sparse neuron activities preserving the topological relationship among similar object views.

Aiming to reveal underlying relationships between the information coding scheme and the computational function of the visual cortex, many researchers have proposed the self-organization models. However, there seems to be no model which can explain both of the above findings, sparseness of neuron activities and topographic mapping of object views. For example, Olshausen and Field[3] introduced a learning algorithm for sparse coding. It was shown that, by seeking sparse code for natural images, the network could develop a set of receptive fields similar to those found in the striate cortex. Although such network successfully captures sparse nature of the input data, topographic structure is not incorporated. On the other hand, Suzuki and Ueda[4] proposed a model of feature columns based on modular architecture of the self-organizing map networks. They showed that the network self-organized the topographic maps of object views. However, such network cannot develop sparse-population coding since the learning algorithm is based on winner-takes-all mechanism favoring a representation in which only one neuron is activated for each input.

In this paper, we investigate a self-organization model which evolves the information representation being consistent with the above physiological findings. The network model is based on Olshausen and Field's efficient coding scheme[3], but an additional constraint, "topographic smoothness," is incorporated so as to emerge a topographic map. The network is trained using real face images composed of different individuals at different viewing directions. The simulation results show that the network evolves the information representation which is consistent with the above physiological findings. It is also confirmed by analyzing the characteristics of the neuron activities that the present model evolves the efficient information representation for coding both of the global structure and the finer information of the face images.

In the following section, we explain our self-organizing network model. The learning method for sparse coding is first described, then the topographic

smoothness constraint is introduced. Section 3 shows the simulation results obtained by the present network using real face images. The properties of the network is further examined in Section 4. Then Section 5 concludes the present work.

2 Self-organization model

2.1 Network

Following Olshausen and Field[3], we investigate a simple network model:

$$\boldsymbol{x} = \sum_{i=1}^{m} a_i \boldsymbol{w}_i \tag{1}$$

where $\boldsymbol{x}$ is an n-dimensional vector denoting the input to the network, a_i denotes the activity of the i-th neuron, and $\boldsymbol{w}_i$ is an n-dimensional vector composed of the connection weights between the i-th neuron and the input. This model is based on the assumption that the input data can be represented in terms of a linear combination of the basis vectors(weight vectors). Given an input, the network encodes it as the coefficients(activities of the neurons) of Equation (1). The objective of the network is, given a set of data, to find the basis vectors giving efficient codes which satisfies some given criteria. Unsupervised learning is adopted for seeking the optimal basis vectors and the coefficients.

2.2 Efficient coding and sparseness

One of the criteria for efficient coding is how well the code describes the input. It can be measured by the squared error between the input and its reconstruction by the network:

$$E = \frac{1}{2} \left\| \boldsymbol{x} - \sum_{i=1}^{m} a_i \boldsymbol{w}_i \right\|^2 \quad . \tag{2}$$

It is known that the minimization of E gives the results which is substantially equivalent to those obtained by applying principal component analysis to the input data.

As an additional criterion for efficient coding, Olshausen and Field[3] proposed the "sparseness" cost for seeking sparse codes. The sparseness cost function, S, is given by

$$S = -\sum_{i} s(a_i) \tag{3}$$

where $s(x)$ is a nonlinear function such as $|x|$, $-\exp(-x^2)$, and $\log(1+x^2)$. The cost S favors the codes which consist of minimal number of non-zero coefficients. As a result, the network seeks the coefficients which are statistically independent each other over an ensemble of input data. In the case that the data contains some forms of higher-order statistical structure as found in natural images, it can be captured by using this sparseness cost function.

2.3 Topographic smoothness

In order to organize a topographic mapping in the above network, it is necessary to incorporate a constraint for preserving the topographic structure of data. One of the approach is to bring each basis vector close to those of its topological neighborhood. Although Self-Organizing Map(SOM) algorithm[5] is able to organize a topographic map by adopting this approach, it evolves winner-takes-all type representation in which only one neuron is activated at a time. When such algorithm is applied to the data composed of multiple views of several objects, each neuron becomes to respond to a specific view of a specific object[4]. Such coding scheme is inconsistent with physiological findings as mentioned above, therefore, we investigate a different approach. In this paper, we incorporate the "topographic smoothness" cost function which constrains the neuron activities to be as smooth as possible on a two-dimensional map.

Let us start with considering a continuous system in which activities of the neurons form a two-dimensional field, $a(x,y)$. In this case, the topographic smoothness can be measured by using the first order differentials with respect to x and y:

$$T = -\iint \left[\left(\frac{\partial a}{\partial x} \right)^2 + \left(\frac{\partial a}{\partial y} \right)^2 \right] dxdy \tag{4}$$

Hence, the smoothness constraint for a discrete model is given by approximating the differentials. Assuming the neurons are arranged on a two-dimensional square lattice with unit intervals, the topographic smoothness is derived:

$$T = -\frac{1}{8} \sum_{x,y} \left[(a_{x,y} - a_{x+1,y})^2 + (a_{x,y} - a_{x,y+1})^2 \right] \tag{5}$$

where $a_{x,y}$ denotes the activity of the neuron at position (x,y) on the lattice. The effect of maximizing this smoothness can be seen by taking the derivative of T with respect to $a_{x,y}$:

$$\frac{\partial T}{\partial a_{x,y}} = -\sum_{x',y'} h(x,y,x',y') a_{x',y'} \tag{6}$$

where $h(x,y,x',y')$ is a function which determines the

interaction between neighboring neurons:

$$h(x,y,x',y') = \begin{cases} 1 & x = x' \text{ and } y = y' \\ -\frac{1}{4} & x = x' \text{ and } |y-y'| = 1 \text{ or} \\ & \quad |x-x'| = 1 \text{ and } y = y' \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

Consequently, Equation (6) forces the activity level of each neuron to be equal with its neighborhood neurons. By combining this smoothness cost to learning of efficient codes, the coefficients, the neuron activities, are biased so that nearby neurons are activated simultaneously. As the result, the network organizes a topographic map of data in which similar patterns activate similar ensemble of neurons but distinct patterns activate different neurons at distant location in the map.

2.4 Learning

Learning is accomplished by minimizing the total cost function:

$$F = E - \lambda_S S - \lambda_T T \tag{8}$$

where λ_S and λ_T are positive constants. The learning process consists of two phases. First, for each input data, F is minimized with respect to a_i. Then $\boldsymbol{w}_i$ is modified by gradient descent on F averaged over the set of input data. In the first phase, the optimal value of a_i is sought on condition that

$$a_i = \exp\left(-\frac{1}{2}b_i^2\right) \tag{9}$$

in order to constrain a_i to be in $[0, 1)$. Accordingly, instead of a_i, the parameters b_i are evolved by the following differential equation:

$$\frac{db_i}{dt} = -\eta_b \frac{\partial F}{\partial b_i} \tag{10}$$

where η_a denotes the learning rate. Equations (9) and 10) can be regarded as the equations defining the temporal dynamics of the neuron activities when a static stimulus keeps being presented for the network[6]. On the other hand, the learning rule for the second phase is given by

$$\frac{d\boldsymbol{w}_i}{dt} = -\eta_w \left\langle \frac{\partial F}{\partial \boldsymbol{w}_i} \right\rangle \tag{11}$$

where η_w is the learning rate and $\langle\ \rangle$ denotes the ensemble average over the input data. The learning rate η_w is set to an appropriately small value for stability of learning. This process can be regarded as learning at longer time scale. Furthermore, in the following simulations, each basis vector was normalized to unit length after each learning step.

3 Simulation

In this section, we show some simulation results using real face images.

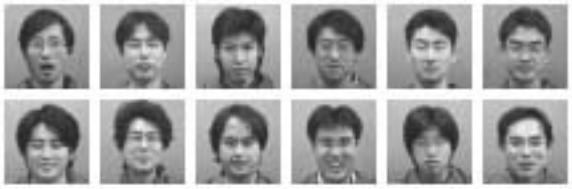
3.1 Experimental conditions

The data set was composed of 32×32 pixel face images of 12 individuals with different view directions(Figure 1). The face images were taken at 25 viewing directions(at every 5 degrees ranging from -60 to 60 degrees), and their position and size were normalized. The total number of the data was $1,500$ (12individuals$\times$25views$\times$5sets). Out of these images, nine views(every 15 degrees for each individuals) were chosen as learning data. Hence the learning data set consisted of 108 images. The other images were used for examining the properties of the network(see Section 4). To reduce the dimensionality of the learning data, each data was converted into a 100-dimensional vector by principal component analysis. Then the input vector, $\boldsymbol{x}$, was obtained by normalizing it so that the mean of $\boldsymbol{x}$ has unit length. In the results shown below, the inverse of these preprocessing steps were performed for visualization.

The network is trained so as to acquire efficient codes for the above data by minimizing the cost function (8). The parameters λ_S and λ_T were set to 0.01 and 0.4, respectively, and $s(x) = |x|$ was chosen as the function for Equation (3). The dimensionality of the input, n, was 100, and the number of neurons, m, were set to 36. As the topology of the neurons we chose a two-dimensional torus lattice to avoid boundary effects. For each of the input data, the parameters b_i were evolved by gradient descent method with momentum term. The learning rate η_b and the momentum coefficients were set to 0.1 and 0.8, respectively. The initial values of b_i were randomly chosen so that each a_i became a small random value, and the updating step was iterated 100 times. Then the basis vectors $\boldsymbol{w}_i$ were also modified by gradient descent method with momentum term. The learning rate η_w and the momentum coefficients were set to 0.01 and 0.8, respectively. The learning step was repeated $1,000$ times starting from random initial values.

3.2 Results

Figure 2(a) shows the basis vectors of 6×6 neurons acquired by the network. None of them correspond to a specific view of a specific person, however, we can find rough structure in their arrangement. The basis vectors around the upper left corner in the figure, around the bottom, and around the upper right seem to correspond to the faces viewing right, front, and



(a) Frontal views of 12 individuals.

(b) Images at different viewpoints. Top: mean faces of 12 people. Center and Bottom: two sets of examples.

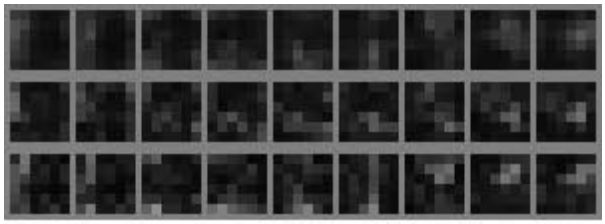
Figure 1: Samples of the face images used in the simulation.

left, respectively. To confirm this observation and to examine the response property of the network, activities of the neurons were computed. Figure 2(b) shows the resulting activity patterns. The figure shows that neurons in a spot region are activated for each face image and the activation spot systematically shifts with rotation of the faces. Furthermore, it is also shown that the images of different individuals at the same direction activate the neurons in similar area but evoke different activation patterns. These results mean that each face image is encoded by the ensemble of several neuron activities and the network possesses the topographic map of the facial views, which seem to be qualitatively consistent with the physiological findings mentioned in Section 1.

4 Discussion

The above simulation results imply a hierarchical coding scheme of the network model: global changes of the data are represented as the systematic shifts of the activation spot, while finer information is coded as the ensemble activity of the neurons in each spot region. In this section, we further discuss the characteristics of the information representation by calculating some quantities which reflect the organization of the neuron activities.

(a) The learned basis vectors. They are arranged according to the topology of 6×6 torus lattice.



(b) Activity patterns. Rows correspond to the three sorts of the data shown in Figure 1(b). Panels in each row show the activity patterns of 6×6 neurons. The gray level indicates the magnitude of the activity. Black means zero, while white is one.

Figure 2: Simulation results.

4.1 Reconstruction error and discriminability

We examine the information representation of the network in terms of the reconstruction error and the "discriminability." Suppose that a set of vectors is given with class labels and each vector can be classified among K classes. Then the discriminability measures the distance among clusters of the vectors corresponding to each class. By calculating the discriminability in several conditions, it is able to characterize the information representation of the network.

For a set of m-dimensional vectors composed of m neuron activities, the discriminability, d, is defined by

$$d = \frac{\sum_{k=1}^{K} \|\bar{\boldsymbol{a}}_k - \bar{\boldsymbol{a}}\|^2}{\sum_{k=1}^{K} \frac{1}{N_k} \sum_{l=1}^{N_k} \|\boldsymbol{a}_{k,l} - \bar{\boldsymbol{a}}_k\|^2} \qquad (12)$$

where $\boldsymbol{a}_{k,l}$ corresponds to the l-th data in class k, and N_k denotes the number of data in the class. The vectors $\bar{\boldsymbol{a}}_k$ and $\bar{\boldsymbol{a}}$ denote the mean of $\boldsymbol{a}_{k,l}$ and the mean of the all activity vectors, respectively. We calculated the discriminability values for the face data in two different conditions, "view direction" condition and "identity" condition. In the view direction condition, the face images were divided into five classes according to their viewing directions. On the other hand, in the identity condition, they were divided into 12 classes according to their identities. Since the variation of face images over different viewing directions is larger than the difference between individuals facing the same direction, only coarse information is required for discriminating the viewing directions, while finer information is required for identifying the individuals.

Table 1 shows the reconstruction errors and the discriminability values of the three different networks, the present model(with T), the same network without the topographic smoothness(without T), and the conventional SOM. These values were computed for the test data described in Section 3.1. In the case of SOM, after finding the basis vectors, $\boldsymbol{w}_i$, the coefficients, a_i, were calculated by using Gaussian softmax function:

$$a_i = \frac{\exp\left(-\|\boldsymbol{x} - \boldsymbol{w}_i\|^2/2\sigma^2\right)}{\sum_{i'=1}^{m} \exp\left(-\|\boldsymbol{x} - \boldsymbol{w}_{i'}\|^2/2\sigma^2\right)} \qquad (13)$$

provided that $\sigma = 0.25$, which gave the best reconstruction error. Both of the networks with and without the topographic smoothness cost achieve considerably smaller errors and higher discriminability than SOM. This might reflect the difference of their coding schemes. Although a kind of topographic map emerges in the present network, the network can encode the images more precisely using the ensemble of several neuron activities unlike SOM. It is also show that the discriminability for global information is enhanced without degrading that for finer one by incorporating the topographic smoothness. These results demonstrate that the present model can develop the representation which is useful for describing both of the global structure and the finer information of the face images.

Table 1: Reconstruction errors and discriminability values of the three different networks. The discriminability values are shown as the relative values.

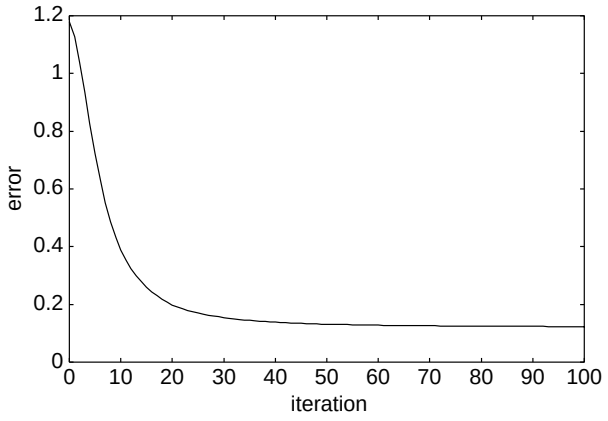
	error	view	identity
without T	0.085	0.621	1.05
with T	0.123	1	1
SOM	0.271	0.177	0.511

4.2 Transient properties of the neuron activities

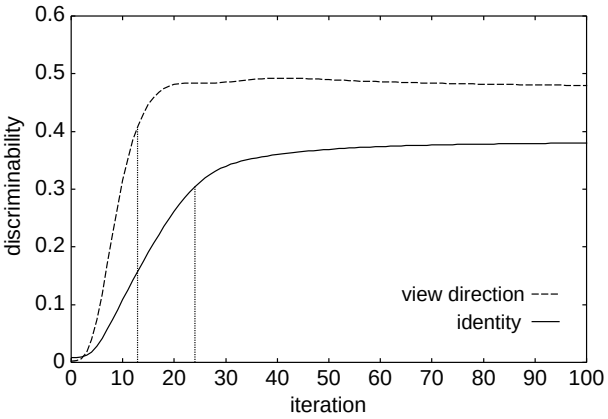
Recently, Sugase et al.[7] analyzed the transient response properties of face responsive neurons in the inferotemporal cortex. Their results suggest hierarchical information coding in time domain: the initial firing of the neurons encode coarse information while the subsequent firings encode finer information. Okada et al.[8] reported that the retrieval dynamics of an associative network model can explain such behavior. Because the present model also defines the temporal dynamics of neuron activity as described in Section 2.4, it is able to depict the change of the information representation scheme by computing the reconstruction error and the discriminability values at each time step. Figure 3 shows the results. It is seen that the discriminability for the global information rises and converges at earlier time step while that for the finer information requires longer time steps. Such characteristics resemble the response properties of the neurons analyzed by Sugase et al. though their stimulus patterns were different from our face images.

5 Conclusion

We investigated a self-organizing network model to account for several characteristics of neurons in the temporal cortex in terms of their information coding scheme. The simulation studies confirmed that the present model could reproduce the similar results to those obtained by experimental studies. The main results concern the emergent properties of the neuron activities evolved by the proposed learning algo-



(a) Mean reconstruction error.



(b) Discriminability values. The vertical dashed lines indicate the time when the discriminability values reached the 80% of their maxima.

Figure 3: Temporal changes of information representation.

rithm. It was shown that the neurons represented the global structure(viewing direction) and finer information(each view of each individual) in a hierarchical way. The global changes were represented as the systematic shifts of activation spot in the topographic map, the finer information were encoded as the ensemble of neuron activities. Since both of the learning processes for sparse coding and for topographic smoothness can be realized by biologically plausible implementation, it might be possible to interpret the above results connecting to the computational functions of the temporal cortex.

Acknowledgment

The authors are grateful to Mr. Hiroyuki Shimai of Saitama University for his invaluable aid for collecting the data. Takashi Takahashi was supported by Research Fellowships of the Japan Society for the Promotion of Science for Young Scientists.

References

[1] M. P. Young and S. Yamane. Sparse population coding of faces in the inferotemporal cortex. *Science*, 256:1327–1331, 1992.

[2] G. Wang, K. Tanaka, and M. Tanifuji. Optical imaging of functional organization in the monkey inferotemporal cortex. *Science*, 272:1665–1668, 1996.

[3] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

[4] S. Suzuki and N. Ueda. Self-organization of feature columns and its application to object classification. In *Proc. International Conference on Neural Information Processing(ICONIP)*, pages 1166–1169, 1997.

[5] T. Kohonen. *Self-Organizing Maps*. Springer, third edition, 2001.

[6] R. P. N. Rao and D. H. Ballard. Dynamic model of visual recognition predicts neural response properties in the visual cortex. *Neural Computation*, 9(4):721–763, 1997.

[7] Y. Sugase, S. Yamane, S. Ueno, and K. Kawano. Global and fine information coded by single neurons in the temporal visual cortex. *Nature*, 401:869–873, 1999.

[8] M. Okada, K. Toya, T. Kimoto, and K. Doya. Retrieval dynamics of associative memory model can explain temporal dynamics of face responsive neurons in the IT cortex. *Society for Neuroscience Abstract*, 25:part1–917, 1999.