7
ACQUIRED FORM REPRESENTATION IN VISION

KUNIYOSHI SAKAI*

Department of Physiology, School of Medicine, University of Tokyo, Tokyo 113, Japan

Primate vision has a remarkable ability to recognize a variety of similar faces or objects. This ability suggests that form information is perceptually organized so that it enables fine discrimination of faces or objects. In this review I use the term *form* for the geometry of an object's overall structure. I believe that form is an indispensable concept in understanding mechanisms of object recognition, because form directly represents an object's entity and enables its recognition.

The inferotemporal (IT) cortex has been proposed to be the memory storehouse in object vision (*9, 15, 24, 26–28*). Along the visual pathway from the primary visual cortex (V1) to the anterior inferotemporal (AIT) cortex, both the receptive field size and the complexity of neuronal processing increase (*10, 33*). Consequently, IT neurons respond selectively to complex forms such as hands, faces, and computer-generated forms (Fig. 1). Whereas the orientation selectivity of V1 and V2 neurons has been well characterized, the form selectivity of IT cells has not been thoroughly studied, owing in

*Present address: MGH-NHR Center, Massachusetts General Hospital & Harvard Medical School, Charlestown, MA02129, U.S.A.

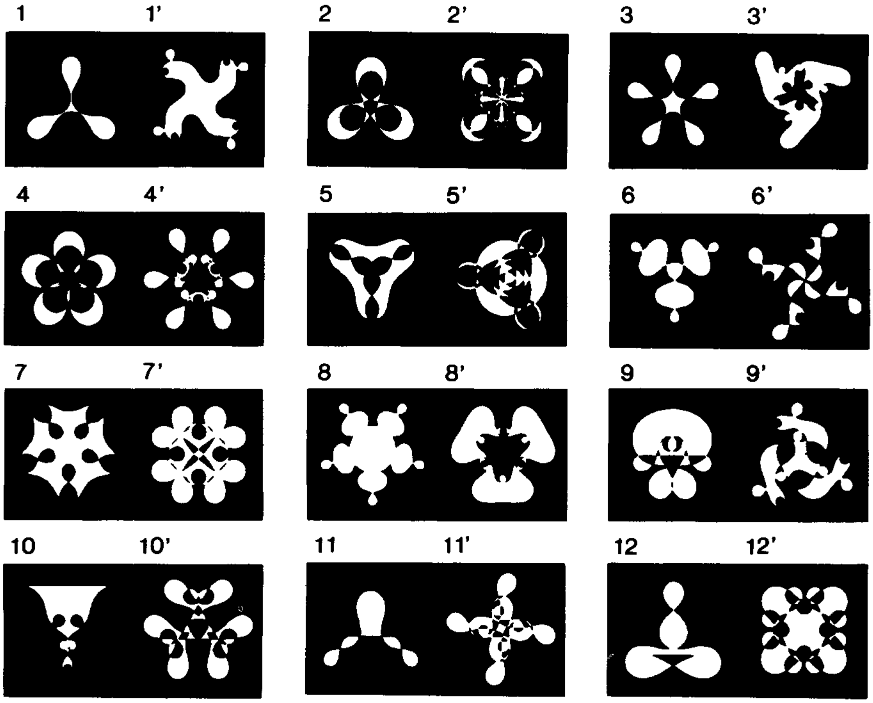

Fig. 1. Twelve pairs of Fourier descriptors (1 & 1′ to 12 & 12′) for stimuli in the pair-association task. When one member of each pair is shown, trained monkeys can retrieve and select the other member of the paired associates.

part to the complexity of form features. When recording a V1 cell, its orientation preference and selectivity can be systematically determined. Orientation is of one-dimensional parameter ranging from 0° to 180°. In contrast, the range of form variety is almost infinite. Moreover, no critical parameters are known for the specification of general forms.

There still remains a controversial question whether form perception occurs as overall identification or as a synthesis of structural components. Some theories of object recognition have proposed bottom-up determination of an object's components and subsequent matching of the arrangement of components with a memory representation (*2*, *13*, *20*). The properties of face-selective cells may be in agreement with these theories (*21*). However, there is psychological evidence that object features (not specifically related to parts) are matched directly with such overall features stored in long-term memory (*5*).

If the form analysis in vision is hierarchical and bottom-up, then some aspects of the form selectivity of IT cells reflect the constraint in the early visual processing. Besides, if this visual processing is based on computational logic, then there must be some principles that guide the generation of the form selectivity. In this review I propose the guiding principles of association and neuronal tuning in the AIT cortex. My working hypothesis is that learning of a form or repeated exposures to a form produces neurons that selectively respond to that particular form. This idea bears some relation to prior conjectures for the picture-selective responses that depend on temporal contiguity (*16*).

I. PAIR-ASSOCIATION TASK

Most of our long-term memories of episodes or objects are organized so that we can retrieve them by association. Anderson and Bower (*1*) proposed that human memory only stores "propositions", which are conceived as "structured bundles of associations between elementary ideas or concepts". This notion may be applied to object recognition: visual memory stores forms as structured bundles of associations between elementary views of objects. The neurophysiological evi-

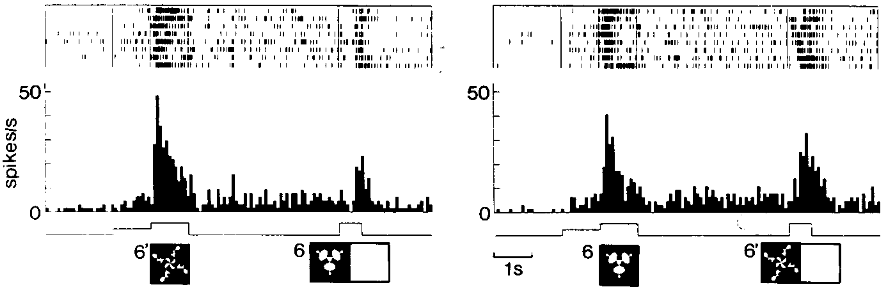

Fig. 2. Responses of a pair-coding neuron, which exhibited form-selective activity during the cue period. Left: trials for cue 6′ that elicited the strongest cue response. Right: trials for cue 6 that elicited the second-strongest cue response.

dence of associative mechanisms in visual memory has been found only recently (*23*, *24*).

Figure 1 shows a set of visual stimuli that I have used. These pictures were sorted randomly into pairs, numbered from 1 & 1′ to 12 & 12′. The pair combinations were fixed throughout the experiments. To impose the acquisition of this screening set (see the section IV) as long-term memory, I trained monkeys (*Macaca fuscata*) in the pair-association (PA) task (*18*, *23*). In each trial of the PA task, a cue stimulus (one of 24 pictures shown in Fig. 1) is presented at the center of the video monitor for 0.5 sec. After a delay period of 5 sec, two stimuli, the paired associate of the cue (correct choice) and a distractor (incorrect choice) are shown. The monkey obtains a fruit juice reward for touching the correct one within 1.2 sec. This task paradigm can reliably demand the learning of visual stimuli, because monkeys cannot select a paired associate correctly without memorizing and recalling pair combinations.

It should be noted that the PA task cannot be solved by employing short-term or working memory within a single trial. Instead, the PA task is essentially the memory *recall* task, which explicitly demands the memory retrieval and thus generation of images from the long-term memory. Therefore, memory components of the PA task present a sharp contrast to those of the delayed matching-to-sample (DMS) task. In the DMS task, the subject indicates whether a test stimulus matches a previously shown sample stimulus. When the stimulus set size is small and thus the same trial is repeated, working memory is mainly involved in the DMS task. When the stimulus set size is large and thus each trial is unique, recognition memory is also involved, because it is possible to indicate whether a test stimulus has appeared or not without employing working memory. We have to take these points into consideration in interpreting the results of behavioral and physiological studies.

II. PAIR-CODING NEURON

In the AIT cortex, I found one type of neuron (*pair-coding neuron*), which manifested selective responses to both paired associates (*23*). The properties of pair-coding neurons indicate that memory storage is organized such that single neurons can code both paired associates in the PA task (Fig. 2). This result provides new evidence that single neurons acquire form selectivity through associative learning. Here, this type of coding is termed *associative coding*, in which the involvement of associative learning is essential for memory storage.

The associative coding proposed here provides one organizing principle by which the special selectivity of neuronal responses is produced. The spatiotemporal patterns of neuronal discharges selective to object forms thus constitute the basis of ensemble coding. A possible molecular mechanism of the associative coding lies in the change of synaptic connections through repetitive learning, whereby two inputs are always paired with each other. The associative mechanism based on temporal contiguity is further discussed in section VIII.

III. PAIR-RECALL NEURON

I found another type of neuron (*pair-recall neuron*) that is presumably involved in the process of memory retrieval (*23*). In the PA task, the monkey is required to recognize the paired associate of a cue stimulus after a delay period. Pair-recall neurons exhibited form-selective delay activity (Fig. 3). This response is closely coupled with the paired

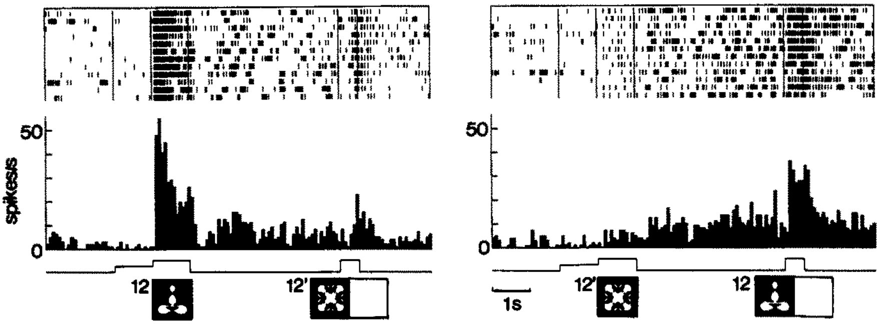

Fig. 3. Responses of a pair-recall neuron, which exhibited form-selective activity during the delay period. Left: trials for cue 12 that elicited the strongest cue response. Right: trials for cue 12′ that elicited the delay response, presumably reflecting retrieval of the paired associate 12. Note the tonic increasing activity during the delay period, which is much stronger than the cue response.

associate that is not actually seen, but retrieved by the cue stimulus. There are two possibilities for the critical process during the delay period. One is to hold a *retrospective code*, which is a cue stimulus, in working memory. The other is to generate a *prospective code* by converting a cue into its paired associate. The increasing delay activity of the pair-recall neurons is consistent with the claim that subjects can employ a prospective code. On the grounds that the AIT cortex serves as the memory storehouse, these neurons could serve as memory storage elements, also activated in the retrieval process. The finding of pair-recall neurons is the first neurophysiological demonstration that visual imagery is also implemented by the same neuronal mechanism that subserves memory retrieval.

IV. FINE-FORM SELECTIVITY

The conventional method for determining the response selectivity of a single IT neuron utilizes a *screening set* of various object forms. Because experimenters are not able to test every possible form, they must prepare a convenient screening set with many visual stimuli. Figure 1 shows an example of a screening set that can specify global form selectivity. If a recorded neuron responds to at least one of the forms in a screening set (responsive), but not equally to all of them (selective), the global form selectivity of this cell can be further characterized.

One may mistakenly conclude that the most effective form in a screening set is the optimum stimulus for a recorded cell. One way to overcome the limitation of using a screening set is the analysis of *fine-form selectivity*, which provides better resolution to discriminate among effective stimuli with similar features. In other words, the test of a cell's form selectivity should be performed in two steps. First, test the cell's responses with a broad screening set. Second, test them more finely by preparing forms similar to the effective stimuli selected in the first step. The analysis of fine-form selectivity becomes particularly important in searching for memory traces of specific forms, which could be acquired through learning experience.

V. EVIDENCE OF NEURONAL TUNING

The patterns shown in Fig. 1 are Fourier descriptors (FDs). FDs are appropriate in testing fine-form selectivity because they are specified by a set of parameters: harmonic amplitude (A_k) and phase angle (α_k), where $k = 1,2,...$ is each term's ordinal number in a Fourier series (*34*). A slight alteration of one FD parameter from its original value produces a very similar form. This manipulation of an original pattern is called ***parametric transformation***. The similarity of forms with slightly altered FD parameters is validated by the fact that the position of each point in the drawing plane is a continuous function of FD parameters (some examples are shown in Fig. 4).

Figure 4 shows data from one AIT neuron with form-selective

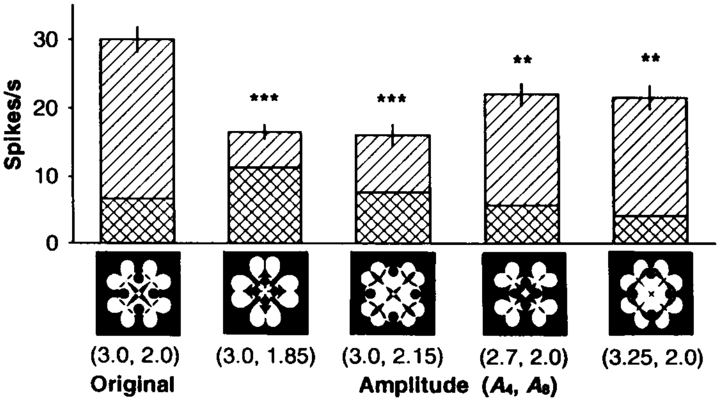

Fig. 4. Responses of an exemplary neuron for the parametric transformation. The original pattern shown is specified by two harmonic amplitudes and two phase angles ($A_4 = 3.0$, $\alpha_4 = \pi/2$, $A_8 = 2.0$, $\alpha_8 = \pi/2$). The set of two numbers attached to each pattern denote its amplitude parameters (A_4, A_8). In each transformed pattern, one of these parameters is slightly altered from the original value. Note the close similarity between these four transformed patterns and the original. Histograms show mean discharge rates (mean±S.E.M.) for each cue presentation of the original or transformed patterns. The cross-hatched area in each histogram bar corresponds to the spontaneous discharge rate. This neuron exhibited optimal responses to the original pattern rather than to the transformed patterns. These data were taken from the first five trials in which the original or transformed pattern was used as a cue stimulus. Two asterisks, $p < 0.01$; three asterisks, $p < 0.001$: according to t test with unequal variances (*29*).

responses. The critical comparison is the neuronal response to the learned form *versus* unlearned transforms during the cue period. The original FD pattern shown in Fig. 4 has two amplitude parameters. A_4 and A_8 that can take continuous values. It can be readily noted that the global features of patterns with a slightly altered value of A_4 or A_8 are strikingly similar. All of the transformed patterns elicited significantly weaker responses than the original pattern. This observation was further confirmed by responses of other neurons. The parametric transformation resulted in a significantly weaker neuronal response for most cases. Moreover, I did not observe any case in which responses to transformed patterns were stronger than those to the original patterns (*25, 28*). This result suggests that AIT neurons are subject to *tuning* mechanisms for particular forms in a learning process. This neuronal tuning cannot be explained by learning-independent innate selectivity, since the original patterns in the screening set were selected randomly.

For the following reasons, the possibility that the weaker response to unlearned transforms is attributable to purely sensory reasons can be excluded. First, every tested transform was derived from one of the original patterns, which elicited the strongest or the second-strongest response. Therefore, these transforms are likely to produce equally strong responses as the original pattern, obviously stronger than other ineffective original patterns. Furthermore, it is impossible to predict beforehand the relative effectiveness between the original pattern and transforms that derived from it, because original patterns have no special form attributes and their FD parameters were selected randomly. Second, a cell's preference within the screening set should not be confused with fine-form selectivity among transforms, because the range of form variety in the screening set is incomparably wider than that in transforms. Therefore, the learning-dependent *fine*-form selectivity cannot be explained by the *global* form selectivity for the screening set.

VI. INTERACTION BETWEEN MEMORY MECHANISMS

Pair-coding neurons and pair-recall neurons can participate, respectively, in the coding and recall processes of visual long-term memory.

The pair-recall neuron represents a typical example of interaction between sustained activation and association among the neuronal mechanisms that subserve the formation or expression of memory traces (*6*). Evidence for form-selective or color-selective sustained activation has been documented in the IT cortex (*7*, *8*, *17*). Furthermore, there is another type of interaction among neuronal mechanisms. The fact that each pair-coding neuron responds to both paired associates suggests an important role of interaction between neuronal tuning and associative mechanisms.

VII. IS FORM REPRESENTATION 2D VIEWER-CENTERED OR 3D OBJECT-CENTERED?

One of the most fundamental problems in object recognition is how a 3D object form in the real world is represented in long-term memory. Marr and Nishihara (*14*) proposed that a form representation for recognition should use a 3D *object-centered* coordinate system, which is transformed from a 2D *viewer-centered* coordinate system. In a viewer-centered coordinate system an object's geometry is specified relative to the viewer. In an object-centered coordinate system an object's geometry is specified relative to the object itself. The latter emphasizes the computation of a form-specific description that is independent of the vantage point (*2*). In contrast to this 3D object-centered model, recent studies propose that multiple 2D views are directly stored for each object representation (*3*, *4*, *12*, *22*, *31*). In this scheme, a viewer-centered recognition can be achieved by interpolating between a small number of stored views. To establish that memory representations of object forms are multiple 2D views rather than internal 3D models, we need more supporting data that can elucidate the underlying processes for object recognition.

VIII. THE IMPLICIT ASSOCIATIVE PROCESS AND OBJECT RECOGNITION

I further propose a possibility that associative mechanisms are shared by two neuronal processes: one is the explicit process that is critical in the PA task, whereas the other is the implicit or automatic process. The latter implicit associative process may subserve object recognition, in which distinct views of an object can be treated as common views of the same object because these views are overwritten in the common neuronal network that consists of many pair-coding neurons. Using this scheme, one standard view can be naturally associated with any observed views of the same object, thus enabling correct recognition.

In our visual world, multiple views of an object are nearly always presented in succession, resulting from relative movement between the observer's eyes and the object. This situation is an example of the implicit associative process in that two different views of an object are associated and memorized by AIT neurons. Such an associative mechanism based on temporal contiguity agrees well with the principle of generic image of sampling proposed by Nakayama and Shimojo (*19*). Incorporating this significant conceptual framework, the implicit associative process can be elucidated further: the relationship of generic 2D views is automatically acquired by the neuronal mechanism of association.

IX. A MODEL OF THE COGNITIVE MEMORY SYSTEM

A model of the cognitive memory system that I have proposed (*24*, *26*, *28*) is shown in Fig. 5A. This scheme is based on structures and functions of the visual memory system for object recognition (Fig. 5B), and it contains some ideas that have been put forward by several researchers (*9*, *15*, *30*, *32*, *33*).

In the memory acquisition process, sensory information is transformed into a memory code of neuronal responses (encoding) with the bottom-up information flow from feature analyzers to a memory storehouse. The primary visual area and the prestriate area serve as feature analyzers in vision. A possible candidate for the memory storehouse is the temporal association area. In visual perception, prominent features in a visual field are selected and located by a focal-attention controller. One candidate for the focal-attention controller is the pulvinar nucleus. I hypothesize that perception is implemented by the interaction between memory acquisition (encoding) and focal-attention mechanisms (Fig. 5A).

To establish a long-lasting representation in the memory store-

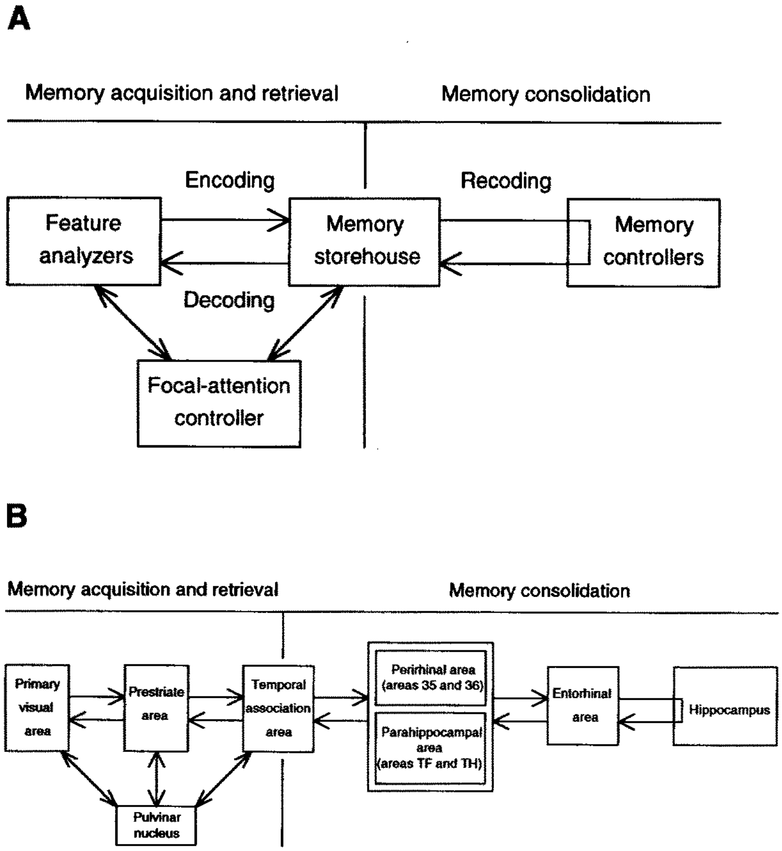

Fig. 5. A: a model of the cognitive memory system that unifies perception and imagery. Perception is implemented by the interaction between memory acquisition (encoding) and focal attention mechanisms. Imagery is implemented by the interaction between memory retrieval (decoding) and focal attention mechanisms. B: structures of the visual memory system for object recognition. Information flows between these structures correspond to the information flows indicated in A. Brodmann's area numbers 35 and 36, and von Economo's area symbols TF and TH are indicated. Note that most of the pathways between cortical visual areas are anatomically reciprocal, supporting their functional roles in encoding and decoding processes.

house, memory controllers are responsible for consolidation. During this memory consolidation process, the memory code is reorganized dynamically (recoding) by the interaction between the memory storehouse and memory controllers. Medial temporal structures, including the hippocampus, are regarded as memory controllers, based on their plasticity (*e.g.*, long-term potentiation and depression), neural networks (*e.g.*, autoassociation), clinical observations (*e.g.*, medial temporal lobe amnesia), and lesion studies.

In the memory retrieval process, memory codes stored in the memory storehouse are decomposed into more elementary attributes (decoding) by feature analyzers through a back-projection pathway. I hypothesize that imagery is implemented by the interaction between memory retrieval (decoding) and focal-attention mechanisms. As Kosslyn (*11*) pointed out, "imagery consists of brain states like those that arise during perception but occurs in the absence of the appropriate immediate sensory input". According to my scheme, visual imagery is generated by top-down activation of perceptual representations that are selected and located by the focal-attention controller. It is also possible that a signal from the focal-attention controller gates the top-down information flow toward early visual areas that function in the decoding process. This model predicts that the extent of visual areas devoted to decoding mental images is controlled dynamically by focal attention. The supporting evidence for this proposal on visual imagery has been discussed elsewhere (*26*).

SUMMARY

I examine the hypothesis that the form representation in the AIT cortex is acquired through learning. According to this hypothesis, perceptual aspects of the temporal association area are closely related to its visual representation, in that the response selectivity of AIT neurons can be influenced by visual experience. On the basis of neurophysiological evidence, I summarize two neuronal mechanisms that subserve the acquisition of form selectivity in AIT neurons. The first mechanism is association, with which relevant pieces of visual information are stored together and retrieved from each other. The second mechanism is neuronal tuning to particular stimuli that were learned in a cognitive task. The acquired form selectivity is a key feature in the capacity of temporal cortical neurons to establish the form representation with multiple 2D views. On the grounds that long-term memory of objects is acquired and organized by at least these two neuronal mechanisms in the temporal association area, I further present a model of the cognitive memory system that unifies perception and imagery.

REFERENCES

1 Anderson, J.A. and Bower, G.H. (1980). *Human Associative Memory: A Brief Edition,* Hillsdale: Lawrence Erlbaum.

2 Biederman, I. (1987). *Psychol. Rev.* **94**, 115-147.

3 Bülthoff, H.H. and Edelman, S. (1992). *Proc. Natl. Acad. Sci. U.S.A.* **89**, 60-64.

4 Cavanagh, P. (1991). In *Representations of Vision: Trends and Tacit Assumptions in Vision Research,* ed. Gorea, A., pp. 295-304. Cambridge, U.K.: Cambridge University Press.

5 Cave, C.B. and Kosslyn, S.M. (1993). *Perception* **22**, 229-248.

6 Desimone, R. (1992). *Science* **258**, 245-246.

7 Fuster, J.M. (1990). *J. Neurophysiol.* **64**, 681-697.

8 Fuster, J.M. and Jervey, J.P. (1982). *J. Neurosci.* **2**, 361-375.

9 Gross, C.G. (1973). In *Handbook of Sensory Physiology,* ed. Jung, R., vol. VII/3, pp. 451-482. Berlin: Springer-Verlag.

10 Gross, C.G. (1992). *Philos. Trans. R. Soc. Lond. B* **335**, 3-10.

11 Kosslyn, S.M. (1988). *Science* **240**, 1621-1626.

12 Logothetis, N.K., Pauls, J., Bülthoff, H.H., and Poggio, T. (1994). *Curr. Biol.* **4**, 401-414.

13 Marr, D. (1982). *Vision,* New York: Freeman.

14 Marr, D. and Nishihara, H.K. (1978). *Proc. R. Soc. Lond. B* **200**, 269-294.

15 Mishkin, M. (1982). *Philos. Trans. R. Soc. Lond. B* **298**, 85-95.

16 Miyashita, Y. (1988). *Nature* **335**, 817-820.

17 Miyashita, Y. and Chang, H.S. (1988). *Nature* **331**, 68-70.

18 Murray, E.A., Gaffan, D., and Mishkin, M. (1993). *J. Neurosci.* **13**, 4549-4561.

19 Nakayama, K. and Shimojo, S. (1992). *Science* **257**, 1357-1363.

20 Pentland, A.P. (1986). *Artif. Intell.* **28**, 293-331.

21 Perrett, D.I., Mistlin, A.J., and Chitty, A.J. (1987). *Trends Neurosci.* **10**, 358-364.

22 Poggio, T. and Edelman, S. (1990). *Nature* **343**, 263-266.

23 Sakai, K. and Miyashita, Y. (1991). *Nature* **354**, 152-155.

24 Sakai, K. and Miyashita, Y. (1993). *Curr. Opin. Neurobiol.* **3**, 166-170.

25 Sakai, K. and Miyashita, Y. (1994a). *NeuroReport* **5**, 829-832.

26 Sakai, K. and Miyashita, Y. (1994b). *Trends Neurosci.* **17**, 287-289.

27 Sakai, K. and Miyashita, Y. (1994c). *Trends Neurosci.* **17**, 513-514.

28 Sakai, K., Naya, Y., and Miyashita, Y. (1994). *Learning and Memory* **1**, 83-105.

29 Snedecor, G.W. and Cochran, W.G. (1989). *Statistical Methods,* Ames: Iowa State University Press.

30 Squire, L.R. and Zola-Morgan, S. (1991). *Science* **253**, 1380-1386.

31 Ullman, S. and Basri, R. (1989). *Artif. Intell. Lab. Memo No. 1152*, Cambridge, MA.: MIT Artificial Intelligence Laboratory.

32 Ungerleider, L.G. and Haxby, J.V. (1994). *Curr. Opin. Neurobiol.* **4**, 157-165.

33 Van Essen, D.C., Anderson, C.H., and Felleman, D.J. (1992). *Science* **255**, 419-423.

34 Zahn, C.T. and Roskies, R.Z. (1972). *IEEE Trans. Comput.* **c-21**, 269-281.