Inferotemporal cortex and visual object recognition

Keiji Tanaka
RIKEN Brain Science Institute

A key function of the primate brain is recognition of objects from their visual images. The recognition process is flexible, tolerating marked changes in images due to changes in illumination, viewing angle, and pose of the object. Moreover, the primate visual system processes images of novel objects based on previous visual experience of similar objects. Generalization or categorization may be an intrinsic property of the primate visual system. This lecture discusses the mechanisms of visual object recognition in the primate in reference to the response properties of single cells in the monkey inferotemporal cortex.

Area TE of the inferotemporal cortex represents the final stage of the ventral visual cortical pathway, which is thought to be essential for visual object-recognition. The ventral visual cortical pathway starts at the primary visual cortex (V1) and leads to TE after relays at V2, V4, and TEO (V1-V2-V4-TEO-TE). TE projects to various polymodal brain sites, including the perirhinal cortex (areas 35 and 36), the prefrontal cortex, the amygdala, and the striatum of the basal ganglia. Monkeys that underwent bilateral TE ablation showed severe and selective deficits in learning tasks that required the visual recognition of objects. Based on these behavioral results together with the above-described important anatomical position of TE, TE is assumed to be the site of neural organization essential for the flexible properties of visual object-recognition in the primate (Tanaka, 1996).

Most cells in TE respond only to complicated object images with strong selectivity. It sometimes appears that they are responding to the object itself. But, by simplifying the image of the most effective object stimulus while the activity of the single cell was recorded, we demonstrated that most TE cells respond to moderately complex features (Tanaka et al., 1991; Fujita et al., 1992; Kobatake et al., 1994; Itô et al., 1994, 1995; Wang et al., 1996, 1998). Single TE cells don’t represent objects themselves, but moderately complex features contained in object images. Combinations of several to several tens of TE cells representing different features contained in the image of the object are necessary to specify particular objects.

The selectivity to complex critical features is subject to changes according to changes in the visual environment in the adult. When TE cells were recorded after the monkeys had been trained to discriminate among 28 moderately complex shapes, about a quarter of the recorded TE cells responded more strongly to some of the learned shape stimuli than to the most effective object stimulus selected from a set of 75 objects (Kobatake et al., 1998). The proportion of TE cells that happened to maximally respond to some of these stimuli was only 5% in untrained control monkeys. These results suggest that the number of TE cells responsive to a particular class of stimuli increases through long-term discrimination training of the stimuli. In our more recent study, responses of cells in TE and adjoining perirhinal cortex obtained dependency on the associated reward condition after long-term experience of stimulus-reward associations (Mogami and Tanaka, 2006). Studies in other laboratories also show that the selectivity of TE cells is changeable in the adult according to long-term learning (Sakai and Miyashita, 1994; Sigala and Logothetis, 2002; Baker et al., 2002).

In monkeys trained to categorize stimuli into a few arbitrary groups, some single cells in the prefrontal cortex show responses covering most stimuli in one of the learned categories (Freedman et al., 2001, 2002). In humans, some cells in medial temporal lobe structures such as the hippocampus respond categorically (Kreiter et al., 2000; Quiroga et al., 2005). Although the stimulus selectivity of
TE cells is affected by training for visual categorization, responses of single TE cells appear to represent individual stimuli rather than learned categories (Freedman et al., 2003). It might be considered that object categories would be represented in the polymodal cortical areas, which receive visual inputs from TE. Although some types of categorization may rely on the prefrontal cortex and medial temporal structures, it is conceivable that certain classes of visual categories would be represented in TE, given our rapid and seemingly effortless categorization ability.

We therefore asked, for a large (>1,000) set of natural object images, whether responses of a population of TE cells (rather than single cells) represent the categorical structure of objects (Kiani et al., 2005). A large stimulus set is advantageous because it contains many categories of objects, thus allowing us to examine the representation of objects without setting an artificial category structure. We recorded 674 TE cells and examined each of them with all the stimuli in the object set. Each stimulus elicited a particular pattern of responses over the population of cells. Stimuli that were close to each other in our hierarchical category structure tended to elicit more similar response patterns in the population of IT cells. A hierarchical clustering analysis of the similarity of response patterns reconstructed the category structure, especially for animate object categories. We thus found that response patterns distributed over the TE cell population represented a large class of animate categories, as well as the structure among them. Information about categories was largely distributed over the cell population.

Pairs of objects belonging to a category may tend to share features more frequently than pairs of objects belonging to different categories. The probability of sharing a feature decreases as the categories of two objects get more distant. This tendency is statistical in that a single feature does not exist in the images of all objects belonging to a category and different features cover different, but overlapping, partial members of a category. On the other hand, features appear not only in images of objects of one category but also in images of objects belonging to other categories in a graded frequency. The frequency with which a feature appears in images of objects may be significantly different, even between pairs of categories other than its most affiliated category.

The important question that arises is whether responding to a set of moderately complex features by TE cells would automatically result in the representation of the category structure? To further examine this possibility, we tuned a population of shape-tuned model units, which simulate monkey TE cells, to a set of randomly selected images from the stimulus set used in the above-described study. Categories and their structure did not emerge from outputs of the shape-tuned model units. This result suggests that the monkey TE does something more than just respond to moderately complex features. There are a huge number of such features that TE cells could potentially be tuned for. However, TE cells do not randomly select their favorite features. They may select the features that are useful for the purposes of the monkey’s behavior.
This framework provides an explanation for the relationship between the selectivity for physical features and the representation of semantic knowledge. The purposes of behavior, which define the relevant semantic knowledge, also determine which features need to be represented by the TE cells. The features represented by TE cells do not directly reflect semantic knowledge, but the selection of features may reflect semantic knowledge. The images of objects belonging to a category may look similar to one another, but the similarity is found only when we define it based on particular features. The visual system has found these features, through post-natal experience and possibly through evolutionary processes, and implemented them in the selectivity of cells in TE and earlier stages along the ventral visual pathway.

References