A study on visual object recognition system

Visual object recognition is crucial for our ability to interact with the environment and for our survival. It is not surprising then, that a large percentage of the cortex, extending from the occipital lobe to the parietal and temporal lobes, is devoted to visual processing. The ‘ ventral stream’ pathway (also known as the ‘ what’ pathway) is responsible for object identification (and is the focus on this article) and the ‘ dorsal stream’ pathway (or the ‘ where’ pathway) is concerned with spatial location. The ‘ ventral stream’ pathway runs from the occipital lobe anteriority to the inferior temporal lobe. 1 Activity in the occipito-inferior temporal lobe is necessary for object perception, in other words, it is region that ‘ performs’ the task of object recognition. 2 This article will focus on theories of object recognition and the mechanisms by which objects are encoded and represented within the ventral stream.

Task of the visual recognition system

The task of the visual object recognition system is to be able to accurately and quickly (within hundreds of milliseconds) recognize and categorize objects that are perceived by the retina. The system must be able to discriminate among similar, but unique, shapes. And simultaneously be able to generalize across a variety of stimuli to recognize patterns and categorize objects appropriately. Finally, the recognition system must also be robust – that is, it needs to be able to successfully recognize objects viewed under different lighting, view-point and size. 3, 4

Theories of object recognition

Theories attempting to explain how the visual object recognition system achieves these tasks can be categorized into view-dependent and view-independent models. View-independent models were first to be proposed and attempt to explain the mechanisms by which the visual system is able to recognize objects viewed from different angles without (seemingly) any temporal delay. In recent years, however, evidence has been building up for the view-dependent model, which suggests that there is a temporal delay (at least initially) in recognizing objects viewed from a non-familiar perspective. 5

View-independent object recognition

The critical aspect of view-independent theories of object recognition is that our ability to recognize an object is not influenced (at least not greatly) by viewpoint. The theory was originally developed in order to explain what is called viewpoint invariance: the relativity stability in our ability to recognize a particular object regardless of the angle or perspective (and lighting, colour, background, etc.) that it is used to view it. The first model was proposed by Marr and Nishihara (1978) and although they did not offer any empirical evidence, they did provide a testable hypothesis: object recognition performance should be independent of observer position and object orientation. 3, 6 In 1985, Biederman expanded upon the Marr and Nishihara's model by developing what's called the Recognition by Components (RBC) theory of object recognition. 7 Just as the previous model, this theory attempts to explain how the visual system achieves robust recognition of 3D objects despite the wide range of possible and perceived 2D representations of the object (on the retina). In the RBC theory, every object is composed of a finite number of geons (geometric ions) – basic geometrical shapes – the geons are such that they are equally recognizable from almost any viewpoint (objects such as cylinders, cones, wedges). Every object is composed of a combination of geons in a structural arrangement that is characteristic of that object. Experimental evidence for the RBC theory comes from an experiment of a line drawing in which the object is obscured (for example, parts of it have been whitened out), if there is sufficient information for geons to be identified, the object can be recognized easier than if the geons are obscured. However, the theory also predicts that geons should be recognized at the same speed regardless of view-point and thus far, experiments have not been able to show that. 3

View-dependent object recognition

The defining aspect of view-independent models of object recognition and the key prediction these models make is that object recognition should be equally fast and accurate from any (or almost any) perspective. If the RBC theory were true, the subjects' ability to decide that two images represent an identical geon should remain constant regardless of the viewpoint. Tarr et al. set out to test this prediction and found that time and accuracy is actually dependent upon the degree of object rotation/viewpoint. They found a decrease in subject performance (the ability to recognize the object) from 0° (trained viewpoint) to 45° to 90° rotation conditions. The findings suggest that, because there is a view-point dependent effect even for simple 3D objects tested in the study, it is unlikely that more complex stimuli will utilize a viewpoint-invariant strategy of recognition. Further support for view-dependent object recognition comes from a study in macaque monkeys which reported evidence of human face responsive neurons in the superior temporal sulcus that exhibit a preference (greater firing) for a specific face angle. 8 The authors suggest that features, surfaces, parts or entire images of objects are encoded in a viewpoint-specific manner, which means that object recognition relies on the similarity between the encoded object and what is perceived by the retina. This is consistent with the findings that objects which are viewed from points that are increasingly different from view learned during training result in both an increased time-to-recognition and increase in errors of accuracy. 5

Neural coding and representation of objects in cortex

Coding schemes

More recently, researchers have begun asking how neuronal activity in the cortex encodes information about visual objects. It appears that while some specific subject categories (faces, body parts and places) are represented by a ‘ sparse code’ (activity in a subset of highly tuned neurons), the majority of objects are encoded by a population of neurons. 9

Sparse coding

Visual objects can be represented by a ‘ sparse code’ – activity in a small number of highly tuned neurons. This means that visual object perception relies on a small number of neurons. In the most extreme example, a single neuron could signal meaningful and complex stimuli (e. g. the ‘ grandmother cell’ or the ‘ Jennifer Aniston cell’). A good example of sparse coding in the nervous system is in the songbird. The forebrain of the bird contains neurons that are selective for a temporally precise sequence of notes. 10 Sparse coding strategy offers two main advantages: (1) it is theoretically easy to understand by other brain regions because activity of that one cell, which might synapse onto hundreds of thousands of other cells, ‘ tells’ those postsynaptic cells something relatively complex and meaningful (and little integration may be required on part of the postsynaptic cell) and (2) metabolic efficiency. 11, 12 The disadvantage is obvious: loss of that cell or the inability of that cell to process presynaptic inputs or generate adequate electrical activity in response to the stimulus might prohibit the organism from recognizing that object entirely.

Population coding

Alternatively, objects can be represented by a large number of broadly tuned neurons – termed ‘ population coding’. In this strategy, the coordinated activity in a great number of these neurons represents the stimulus. In this case, relevant information about the stimulus cannot be extracted from a single or a small number of neurons. Examples of population-based coding in the nervous system are evident in the motor cortex. Movement direction (for example your arm) is coded by combining information across a population of neurons such that recording activity from a single individual neuron within that coding population is insufficient to predict the direction of an movement (or any relevant information about the motion). The advantage of coding using a large number of neurons is robustness against biological noise (such as cell death) or inherent variability that exists in neuronal circuits. 12 The disadvantage is that there is an inherent ambiguity that arises when two similar stimuli have to be encoded at the same time. Since the stimuli may share a large proportion of the neurons – that is, neuron A would be activated to stimulus 1 and stimulus 2 – it may be difficult for the decoder to accurately resolve that both, or indeed, either, of the stimuli are being perceived (think of a very overlapping Venn diagram). 13, 14

Visual object representation in cortex

The second aspect of visual object recognition is the distribution of the encoding neurons within the cortex. Besides the electrical activity and the number of neurons that are used to code for a particular object, it is important to know the spatial distribution of these neurons in the cortex and if there is an underlying principle that dictates the organization of these neurons (and if it is the same for all sub-regions). 2

Category-specific mapping

The cortex contains regions, notably the fusiform face area (FFA), the extrastirate body part area (EBA) and the parahippocampal place area (PPA), that are highly-selective for particular objects (faces, body parts and places, respectively). By contrast, the rest of the cortex is relatively unselective and appears to participate in object recognition more generally. 15-17 The most famous of the category-specific modules is the FFA, initially reported by Kanwisher et al. who used fMRI to show that the fusiform gyrus in twelve of the fifteen subjects was preferentially activated when faces were presented than when they viewed other common-place objects. They concluded that this region was selectively involved in face perception, and by extension, purposed that while the majority of the cortex is utilized for general object recognition, there are (at least one at the time) cortical regions which are specific for particular object categories. 15

Distributed and overlapping mapping

Conversely, all objects may be encoded by a distributed pattern of neuronal activity across the entire cortical visual coding areas. 18 In a study by Haxby et al., the authors test eight different object categories and show that within the same subject, the activation areas for each particular object is replicable, but perhaps most importantly, they show that the particular object category can be predicted even when the region most selective for the category is excluded from the analysis. This led the authors to propose that cortical coding of objects and faces ise widely distributed and overlapping. 18 Further support for this hypothesis comes from another study by this group which showed that although several regions in the ventral temporal cortex respond preferentially to specific categories, they found that each category also activates a significant response in regions whose maximal response occurs to stimuli from other object categories. Consequently, each object-category was not only associated with activation in a particular cortical region but also a differential pattern response across the ventral temporal cortex. This suggests that object representation is not restricted to a particular region, but rather distributed more broadly across the cortex (and that this pattern across the cortex is key for object coding). 19

Process-based mapping

The cortex can also be mapped based on the kind of processing that is necessary to identify the object. 20, 21 Different processing may be required to recognize visual objects that are identified on an individual basis (ie, faces of people you know) versus those that, unless we have expertise in the area, are identified by the group they belong to (ie, chairs). Gautheir et al. suggests that expertise in an area (such as faces, cars, birds, etc.) and the level of categorization required (drink/coffee/Starbucks Serena Organic Blend), not superficial properties of the object, that is important and determines the activation of category-selective regions. In other words, the FFA is not activated because the stimulus is a face per se, but because face perception (or identification of who the individual is) requires a high level of expertise on the part of the subject. 21

Resolution-based mapping

Finally, the cortex may be organized based on resolution requirements that are necessary to identify the visual stimulus. 22 Visual stimuli, such as letters, words, and faces, require a more central visual-field bias than stimuli such as buildings. Buildings, for example, can be discriminated using a more peripheral bias that requires incorporation of the entire visual scene. In particular, Malach et al. suggest that visual objects where fine-detail analysis is necessary for recognition will require different coding and different areas than objects that require integration of the entire visual-field (or a large aspect of it).