6 From the Highest Echelons of Visual Processing to Cognition

Supplementary content at http://bit.ly/364H8WR

The inferior temporal cortex (ITC) is the highest echelon within the visual stream concerned with processing visual shape information. The Felleman and Van Essen diagram (Chapter 1, Figure 1.5) places the hippocampus at the top. While visual responses can be elicited in the hippocampus, people with bilateral lesions to the hippocampus can still see very well. A famous example is a patient known as H. M., who had no known visual deficit but gave rise to the whole field of memory studies based on his inability to form new memories. The hippocampus is not a visual area and instead receives inputs from all sensory modalities (Chapter 4).

The history of how the inferior temporal cortex became accepted and described as a visual area is fascinating and follows the refinements in the ability to make more precise lesions and controlled behavioral experiments. In stark contrast to the hippocampus, bilateral lesions to the ITC are associated with impairment in visual object recognition in macaque monkeys (Section 4.7), and with several object agnosias in humans (Section 4.8). We are beginning to decipher the neural code that represents how visual scenes are interpreted.

6.1 A Well-Connected Area

The inferior temporal cortex (ITC) spans Brodmann's cytoarchitectonic areas 20 and 21 (Figure 5.1). The ITC is a vast expanse of cortex that is usually subdivided into a posterior area (PIT), a central area (CIT), and an anterior area (AIT). Biologists are fond of confusing people by using different names for the same thing, a phenomenon that can be partly explained by independent investigators working on related topics in parallel and coming up with new nomenclature to describe their findings. The ITC is also referred to in the literature as areas TEO and TE. The degree of functional specialization among different parts of the ITC remains poorly understood, and it is extremely likely that we will have to subdivide the ITC into many different subareas beyond the current coarse subregions, based on connectivity, neurophysiological, and computational properties.

Like most other parts of cortex, the connectivity patterns of the ITC are extensive and complex. When we describe computational models of vision in Chapters 7 and 8, it will be apparent that most models represent a major simplification of the actual connectivity diagram. The ITC receives feedforward topographically organized inputs from areas

V2, V3, and V4 along the ventral visual cortex. The ITC also receives fewer inputs from areas V3A, MT, and MST, highlighting the interconnections between the dorsal and ventral streams (Section 4.5). The ITC projects back to V2, V3, and V4. There are also interhemispheric connections between the ITC in the right and left hemispheres through the main set of fibers connecting the two hemispheres, the corpus callosum.

The ITC also has extensive projections to and receives signals from non-visual regions, including (i) areas that provide critical inputs to the medial temporal lobe memory system such as the perirhinal cortex, parahippocampal gyrus, and entorhinal cortex; (ii) areas involved in processing emotions such as the amygdala; and (iii) areas in prefrontal cortex that are relevant for decision making, planning, and working memory. Thus, from an anatomical standpoint, the ITC is ideally situated to interpret visual inputs in the context of current goals and previous history, and to convey this information to make behavioral decisions and create episodic memories.

6.2 ITC Neurons Show Shape Selectivity

Over the last five decades, a heroic school of investigators has studied ITC responses in monkeys due to the overall similarity between their visual system and that of humans. Most, if not all, ITC neurons show visually evoked responses, firing vigorously to color, orientation, texture, direction of movement, and shape. Posterior portions of the ITC show a coarse retinotopic organization and an almost complete representation of the contralateral visual field. The receptive field sizes of posterior ITC neurons are about 1.5–4 degrees; on average, the receptive fields are more extensive than those found in V4 neurons.

In more anterior locations along the ITC, there is a weaker retinotopic organization. The receptive field sizes in more anterior parts of ITC are often large. Estimates vary widely, ranging from ~2 degrees receptive fields to neurons with receptive fields that span several tens of degrees. Most receptive fields in anterior ITC include the foveal region.

Example responses from three ITC neurons in response to five pictures are shown in Figure 6.1. In this figure, each picture was repeated 10 times, and the stochasticity of the neuronal responses is evident in the heterogeneous patterns from one trial to the next. This trial-to-trial variability is not specific to ITC and is prevalent throughout the visual cortex. There is considerable discussion in the field about the origin of this variability – which does not seem to be intrinsic to neurons but may constitute a network phenomenon that reflects different levels of attention, expectations, eye positions, and other changes across trials.

Despite this trial-to-trial variability, there are several consistent features that are evident in the neuronal responses in Figure 6.1. All three neurons show an increased firing rate that commences approximately 100 milliseconds after stimulus onset (approximately near the end of the white horizontal line denoting the duration of stimulus presentation). This latency should not be interpreted as a response triggered by the stimulus offset; if the stimulus duration were longer, the neurons would still start to fire at around 100 milliseconds after stimulus onset. These 100 milliseconds reflect

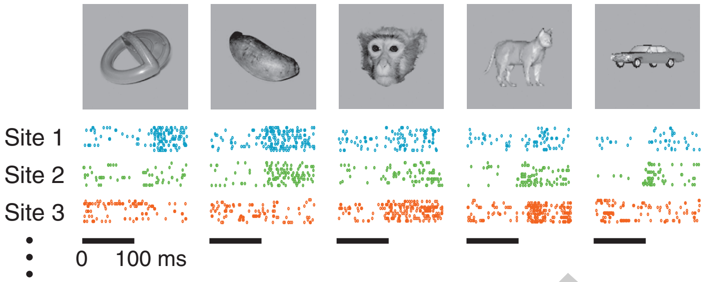

Figure 6.1 ITC neurons are picky. Example responses from three neurons in inferior temporal cortex (labeled “Site 1”, “Site 2”, “Site 3”) to five different grayscale images. Each dot represents a spike, each row represents a separate repetition (10 repetitions per object), and the horizontal white lines denote the duration of the image (100 milliseconds presentation time). Data from Hung et al. 2005

the latency for all the computations that transpire throughout the ventral visual cortex before reaching these ITC neurons (Section 5.12). The neurons are picky in their stimulus preferences. The “Site 1” neuron showed a stronger response to the first two pictures (toy, food) compared to the last two pictures (synthetic rendering of a cat, car). In contrast, the “Site 3” neuron showed an increased response to the third and fourth pictures (monkey face, cat).

Investigators have effectively tested the responses of ITC neurons to a wide range of visual stimuli. For example, some studies have used parametric descriptors of abstract shapes. Logothetis and colleagues trained monkeys to recognize paperclips forming different three-dimensional shapes and subsequently found neurons that were selective for specific three-dimensional object configurations. ITC neurons can be driven by pictures of cars, toys, faces, and fruits.

This wide range of response preferences might seem puzzling at first. Perhaps one would like to conjecture that an area that plays a vital role in object recognition would have neurons that respond specifically to objects in the real world. There could be banana neurons (i.e., a neuron that responds if and only if investigators show a picture of a banana to the monkey), peanuts neurons, chair neurons, face neurons, paperclip neurons, hand neurons, and spaghetti with meatball neurons. Indeed, if we ignore momentarily the “if and only if” part, it is possible to find neurons activated selectively by these types of images. As illustrated by the examples in Figure 6.1, the responses are *not* all-or-none. ITC neurons do not seem to be activated *only* upon presentation of one specific type of object in the real world with baseline level responses to everything else. Instead, ITC neurons show graded activations with stronger responses to some stimuli compared to others.

It is unclear whether ITC neurons show any special treatment to naturally occurring objects like bananas or faces. ITC neurons may represent a sufficiently rich dictionary of complex features. These features can be used to represent any number of naturally occurring objects in an analogous way to forming words by combining different letters or sentences by combining words. Those features can be found in fractal patterns, in paperclips, in faces, and chairs. We will come back to a more quantitative description of these properties and the responses of ITC neurons when we describe current computational models of visual processing in Chapters 7 and 8.

As discussed in the case of neurons in earlier visual areas (Sections 2.7 and 5.6), there is a clear topography in the ITC response map. By advancing the electrode in a trajectory that is approximately tangential to cortex, investigators find neurons that have similar tuning. This level of organization can be represented by "columns" of neurons with similar preferences. Moving horizontally, neighboring neurons in the ITC also show similar, but not identical, preferences.

6.3 Selectivity in the Human Ventral Visual Cortex

Less is known about the internal machinery that processes visual information in the human brain. The primary source of information about the inner workings of human ventral visual cortex comes from invasive neurophysiological recordings in epilepsy patients, which were introduced through the work of Penfield in Section 4.9. A fraction of patients with epilepsy can be treated pharmacologically. In cases of focal epilepsies that do not respond to current drug treatments, an important approach has been to surgically remove the epileptogenic focus. In most cases, this surgical procedure requires first carefully mapping brain activity to discern where seizures are coming from and also to ensure that the brain excisions do not interfere with future cognitive function. For this purpose, neurosurgeons typically implant electrodes inside the human brain. Because current noninvasive techniques are too coarse to map the origin of seizures, the neurosurgeons typically implant many tens of electrodes in different brain areas with the hope of pinpointing the seizure onset. After implantation, the patients stay in the hospital for about one week for observation, granting investigators a rare and unique opportunity to scrutinize human brain function at high spatiotemporal resolution and with high signal-to-noise ratio compared to anything that can be done from outside the brain.

The location of the electrodes is strictly dictated by clinical needs. Sometimes, those electrodes are placed along ventral visual cortex. An example of visually selective responses in human ITC is shown in Figure 6.2. The human intracranial field potentials – that is, the voltage recorded at these electrodes – show many of the hallmarks of the macaque ITC responses. Signals along the human ventral visual cortex also show circumscribed receptive fields, which increase in size from the fovea to the periphery, and from one area to the next. Field potentials along the human ventral visual cortex are also selective and graded (Figure 6.2A). The intracranial field potential signals also show trial-to-trial variability, yet the visually evoked responses can be readily appreciated in single trials (Figure 6.2B). There have been many more neurophysiological studies scrutinizing responses in monkeys compared to humans. Many details about the response properties along the human ventral visual cortex remain unexplored. For example, to the best of our knowledge, nobody has investigated responses in the human ventral visual cortex to fractal patterns or paperclips, as shown in monkey studies.

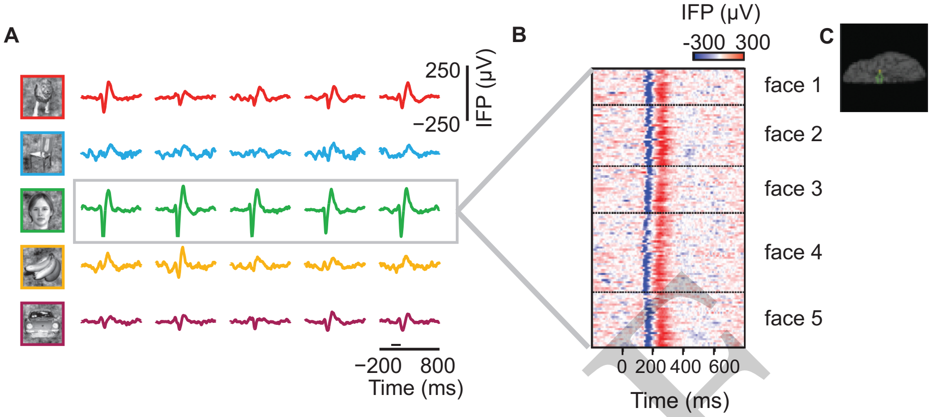

Figure 6.2 The human ITC also shows shape selectivity. Example electrode describing the physiological responses to 25 different exemplar objects belonging to five different categories. (**A**) Responses to each of 25 different exemplars (each color denotes a different category of images; each trace represents the response to a different exemplar). (**B**) Raster plot showing every single trial in the responses to the five face exemplars. Each row is a repetition; the horizontal lines separate the exemplars; the color shows voltage (see scale bar on the right). (**C**) Electrode location. Reproduced from Liu et al. 2009

However, the absence of evidence does not imply evidence of absence. As far as we can tell, responses along the human ventral visual cortex show selectivity to a wide variety of visual shapes like their macaque monkey counterparts.

Response latencies in the human brain seem to be slightly longer than those in macaque monkeys, perhaps because of the larger brain size, or perhaps because there might be more computational steps before the information reaches the human ITC. Within the scarce and preliminary neurophysiological evidence available today and, to a reasonable extent, many of the properties of the macaque ITC are recapitulated in the human ITC.

It should be noted that it is not entirely clear how to meaningfully compare brain areas and functional responses between humans and monkeys (or any other pair of species separated by long evolutionary timespans). First of all, we should be cautious about comparing spikes in monkeys to intracranial field potential signals in humans. It turns out that field potential signals show similar selectivity patterns to spiking signals in monkey ITC. The coarser field potential responses are somewhat less picky than spikes in terms of their ability to distinguish different stimuli, perhaps due to averaging over many neurons.

A more challenging consideration involves establishing rigorous homologies between species. It seems evident that the eyes in monkeys are homologous to the human eyes. Additionally, although the neuroanatomical connections in humans remain unclear, it is quite tempting to assume that the monkey primary visual cortex may be homologous to the human primary visual cortex. As we go deeper into the ventral visual cortex beyond V1, homologies become murkier. Regardless of whether we can establish a unique evolutionarily rigorous one-to-one map between specific structures in different species, it is nevertheless clear that human ventral visual cortex shows rapid and selective responses to complex shapes that are qualitatively similar to those observed in monkeys.

6.4 What Do ITC Neurons *Really* Want?

ITC neurons seem to respond to a wide variety of different shapes that investigators have used to probe their stimulus preferences. Recording time is limited, and investigators need to make choices about which stimuli to use in an experiment; we introduced this problem in Section 5.11. Typically, investigators choose stimuli based on a combination of inspiration from previous studies (if a particular type of stimulus worked before to drive neurons in a given area, it should work now too) or intuitions based on the prevalence of natural stimulus statistics (it seems logical to assume that neurons may represent the types of inputs that the animal experiences daily), or arguments about the presumed evolutionary importance of certain classes of stimuli. Additionally, important advances about neuronal tuning properties have been based on semi-serendipitous discoveries.

These types of experiments carry the potential biases injected by the investigators in selecting the stimuli. Obviously, we can only find tuning for those stimuli that we probe. Even the title of this section has a strong anthropomorphic spin. Neurons do not really “want” anything. The question is meant to allude to what types of visual stimuli maximally activate a given neuron (in the sense of triggering more spikes). As emphasized in Section 5.11, the critical difficulty in elucidating the response preferences of neurons involves the *curse of dimensionality*: too many possible images and too little time.

A promising line of research to elucidate the feature preferences in ITC involves changing the stimuli in real time, dictated by the neuron’s preferences. Recent work based on this approach suggests that we may need to rethink the neural code for features in ITC (and perhaps earlier visual areas as well). One of the first applications of this approach was developed by Charles Connor’s group to let neurons themselves reveal what they like rather than impose a strong bias in the stimulus selection. Recent work by Will Xiao involved developing a computational algorithm that is capable of generating images guided by neuronal firing rates (Figure 6.3). The investigators combined an image generator and a genetic algorithm based on the neuron’s firing rate as a fitness function to guide the evolution of stimuli in real time. In a given generation, the investigators probe the responses to a set of images. Images that trigger high firing rates are kept, and the rest are modified and recombined by the genetic algorithm in combination with the image generation algorithm.

In Section 8.5, we will introduce deep hierarchical models of vision that start with pixels and yield a high-level feature representation of the image. Additionally, in Sections 9.8 and 9.9, we will introduce generative adversarial networks that create images by inverting a deep hierarchical model. The generative algorithm deployed by Xiao and colleagues, inspired by work in machine learning to build image generators, is

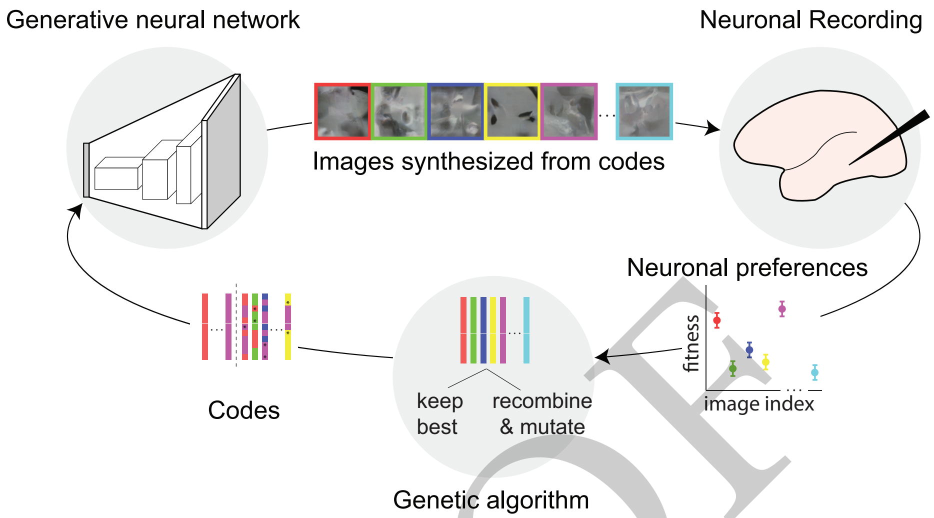

Figure 6.3 Letting neurons reveal their tuning preferences. An approach to investigate neuronal tuning in an unbiased manner. A generative neural network is used to create images by inverting a model of visual recognition (Section 9.9, Figure 9.10). The synthetic images are presented while recording neuronal activity. The neuronal responses are used as a fitness index to guide a genetic algorithm to select a new generation of improved images. Reproduced from Ponce et al. 2019

essentially an inverted version of deep hierarchical computational models, starting with high-level features and ending up with the generation of an image.

By running this generative computational algorithm while recording the activity of a neuron in ITC, Xiao and colleagues discovered images that elicited higher firing rates than natural images that had been used before to test the responses of the neurons. The investigators refer to these synthetic images as “super-stimuli.” These super-stimuli contain naturalistic combinations of textures and broad strokes, which have been likened to impressionist renderings of abstract art. The fundamental novel concept here is that neurons may be optimally activated by combinations of sophisticated features that cannot be easily described in words. In contrast to anthropomorphic descriptions of feature preferences in ITC (“this neuron likes faces,” “this neuron likes chairs,” “this neuron likes curved shapes”), the new line of work suggests that neurons might be activated by complex shapes that defy a language-based definition. A rich basis set of neurons tuned to such complex features is capable of allowing the organism to discriminate real-world objects, but the basis set does not have to be based on icons of real-world objects.

6.5 ITC Neurons Show Tolerance to Object Transformations

As emphasized in Sections 1.4 and 3.4, an essential property of visual recognition is the capacity to recognize objects despite the transformations of the images at the pixel level (Figure 3.6). It is therefore interesting to ask whether the visual selectivity at the neuronal level, as described in the previous sections, is maintained across image transformations. For example, would the neuron shown in the top row in Figure 6.1 continue to respond selectively to the first two objects if they are shown at a different scale, in a different position with respect to fixation, or in a different color?

ITC neurons show a significant degree of tolerance to certain object transformations. ITC neurons have larger receptive fields and therefore show more tolerance to object position changes compared to neurons in earlier parts of the ventral visual cortex. ITC neurons also show similar responses in spite of substantial changes in the retinal size of the stimuli. Tolerance does not necessarily imply that the firing rate in response to a given object should be *identical* across different transformations. Even if the absolute firing rates are affected by a transformation, like changing the stimulus size, the rank-order preferences among different objects – and, therefore, the relative stimulus preferences – are maintained. ITC neurons also show a certain degree of tolerance to depth rotation. Additionally, while luminance changes typically define most shapes, ITC neurons also respond to shapes defined by other cues. For example, shape can be defined by noise patterns that move in a coherent fashion or by texture changes without luminance edges.

An extreme example of tolerance to object transformations was provided by recordings of single-neuron responses from the medial temporal lobe (not the ITC) in human epilepsy patients. Recording from the hippocampus, entorhinal cortex, amygdala, and parahippocampal gyrus, investigators found neurons that show responses to multiple objects within a semantically defined object category. They also found some neurons that show a remarkable degree of selectivity to individual persons or landmarks. For example, one neuron showed a selective response to images where the ex-president Bill Clinton was present; another neuron preferred pictures of the famous actress Jennifer Aniston. Remarkably, the images that elicited a response in these neurons were quite distinct from each other in terms of their pixel content ranging from a black-and-white drawing to color photographs with different poses and views. Such an extreme combination of selectivity and tolerance has not been described in ITC areas but rather in areas of the medial temporal lobe. As noted at the beginning of this chapter, these medial temporal lobe structures receive visual inputs but are not strictly visual areas. In fact, damage to medial temporal lobe structures does not seem to be associated with any apparent visual impairment, or any other perceptual deficit, but rather with memory problems. Therefore, it is likely that this combination of selectivity and tolerance reflects a *readout* of activity from a population of ITC neurons to transform sensory inputs into episodic memories.

6.6 Neurons Can Complete Patterns

During natural vision, objects are often only partially visible due to poor illumination or because there are other objects in front of them (Section 3.5). In early visual areas with small receptive fields, occlusion may cover the entire part of the visual field that a given neuron is interested in. In contrast, in higher visual areas with larger receptive fields, occlusion may only obstruct part of the input to a given neuron. The degree of tolerance to object transformations described in the previous section suggests that neurons might potentially also tolerate inputs that only contain some of the preferred features.

Indeed, the ITC shows a large degree of robustness to occlusion. The neural responses in the ITC can complete patterns and maintain their selectivity even when more than half of the preferred object features are invisible. At both the behavioral level (Section 3.5) and the neurophysiological level, pattern completion requires additional computation time: the latencies of the visually selective evoked responses elicited by partially visible objects are about 50 milliseconds longer than those triggered by fully visible objects. These observations suggest the need for additional processing to make inferences from partial information. We will come back to this point in Sections 7.6 and 8.16 when we discuss the computational mechanisms of pattern completion.

In the previous section, we noted that tolerance to object transformations does not necessarily imply that the neural responses to transformed versions of an object should be identical. Scaling, rotation, color changes, and other transformations can alter a neuron's firing rate, and tolerance refers to the maintained neural selectivity. In the same fashion, completing patterns does not imply that neural responses to heavily occluded objects are identical to the responses to the fully visible counterparts; pattern completion at the neuronal level indicates that selectivity is maintained.

Whereas certain image transformations, such as scale or position changes, maintain the same object features visible (albeit in different places or sizes), other image transformations like three-dimensional rotation or heavy occlusion alter which features are visible and which ones are not. Therefore, it is perhaps unsurprising that the disappearance of certain object features and the appearance of new features during rotation may lead to different firing rates. What is remarkable is that some of the relative stimulus preferences are maintained under these conditions that carry substantial changes at the pixel level.

6.7 IT Takes a Village

The observation that individual neurons can show a high degree of selectivity and tolerance to image transformations should not be taken to imply that there is a one-to-one map between the activity of a single neuron and recognition of a specific object. The idea of a one-to-one map between neurons and specific objects is erroneously referred to as the "grandmother cell" theory. A one-to-one system would be extremely unwieldy and fragile. Losing that one neuron might lead to an inability to recognize that particular object. Additionally, in most cases, readout neurons depend on inputs from hundreds to thousands of other neurons and cannot be reliably or exclusively driven by a single input.

As noted in Section 6.2, nearby neurons in the visual cortex tend to show similar feature tuning properties. Even if we cannot currently monitor the activity of every neuron in a local area, finding a neuron with a specific tuning function is likely to imply the existence of a large number of other nearby neurons with similar tuning properties. In fact, the idea of a "grandmother cell," as coined by Jerry Letvin in 1969, referred to a whole population of cells with identical selectivity and tolerance properties (in the original description, he referred to a “mother cell” rather than a “grandmother cell”). Understood as in the original definition, the idea of a grandmother cell – that is, a population of probably nearby neurons that show selectivity and tolerance to related stimulus properties – is an adequate description of neuronal tuning throughout the visual cortex. Retinal ganglion cells are grandmother cells for changes in illumination at sparse and specific locations in the visual field, primary visual cortex neurons are grandmother cells for oriented lines, and ITC neurons are grandmother cells for complex shape features.

While each neuron shows a preference for some shapes over others, the amount of information conveyed by individual neurons about overall shape is limited. Additionally, there seems to be a significant amount of “noise” in the neuronal responses in any given trial. The term noise is somewhat of a misnomer, as it refers to the trial-to-trial variability in the spike timing and spike counts, as noted in Figure 6.1. Whether this is real noise or part of the signal and what the origin of this variability is remain topics of debate in the field. For suprathreshold stimuli, perception is quite robust: you can look at the shape of the letter A a thousand times, and it will always look like an A. Therefore, somewhere in the brain, a postsynaptic neuron receiving inputs from capricious presynaptic neurons that emit different responses to presumably identical inputs in each trial still needs to be able to discount the variability and decipher what is out there in the world.

Can animals use the neuronal representation of a population of somewhat capricious ITC neurons to discriminate among objects in single trials? The critical emphasis is on single trials. Unlike what many investigators do when they analyze neural recordings, the brain cannot average over trials (we do not need to look at the letter A 10 times to be able to recognize it). The brain is not constrained to making inferences from the activity of a single neuron. Any given neuron in cortex receives input from approximately 10,000 other neurons. Such a population could show interesting properties that ameliorate or eliminate the challenges associated with interpreting the output of a single neuron.

Chou Hung and colleagues addressed this question by recording activity (sequentially) from hundreds of ITC neurons and using machine learning classifiers to decode the activity of a pseudo-population of neurons in single trials. The term pseudo-population refers to the notion that these neurons were not simultaneously recorded. The machine learning decoding approach aims to learn a map between (i) the activity patterns of a population of neurons in response to a set of images and (ii) the labels of objects in those images (Figure 6.4). Consider an experiment where we present pictures of cats or pictures of fish. Let ${}_jx_i$ represent the activity of neuron i in response to image j. For example, x could represent the total number of spikes emitted by the neuron in a given time window. Due to the latency of ITC responses (Figure 6.1), we can consider a window between 100 and 300 milliseconds after stimulus onset. The population response of N neurons to image j is $\boldsymbol{jx} = \left[jx_1, \ldots, {}_jx_N\right]$.

If we imagine that all of these inputs might project to a given postsynaptic neuron, we can write the total aggregated input to the postsynaptic neuron as the weighted sum of

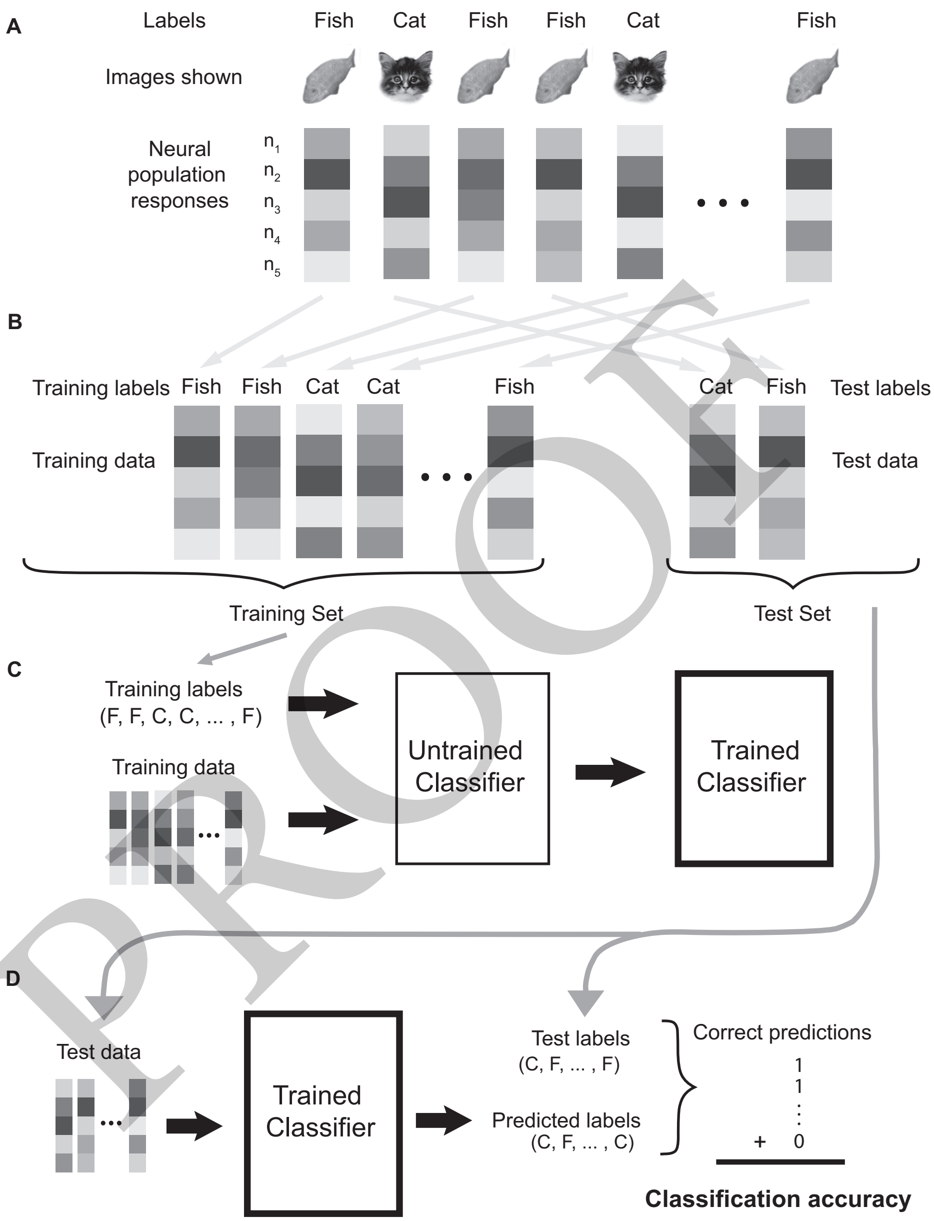

Figure 6.4 Decoding population responses. Basic steps involved in training and testing a classifier. (A) Illustration of an experiment where images of cats and fish were shown in random order to a subject while simultaneous recordings were made from five neurons/channels. The grayscale level denotes the activity of each neuron/channel. (B) Data points and the corresponding labels are randomly selected to be in either the training set or in the test set. (C) The training data points and the training labels are passed to an untrained classifier that "learns" which neural activity is useful at predicting which image was shown – thus becoming a "trained" classifier. (D) The test data are passed to the trained classifier, which produces predictions of which labels correspond to each unlabeled test data point. These predicted labels are then compared to the real test labels (i.e., the actual labels that were presented when the test data were recorded), and the percentage of correct predictions is calculated to give the total classification accuracy. Modified from Meyers and Kreiman 2011

all these inputs: $w_{1j}x_1 + \ldots + w_{Nj}x_N$. Those weights can be thought of as a measure of the synaptic strength, the impact that a given input will have on the postsynaptic neuron. Can such a downstream neuron detect the presence of a cat or a fish? We can build a detector that can read out from the population activity whether the image shown in a given trial contained a cat or a fish. We will set a threshold on the total combined inputs, $g(\boldsymbol{w} \bullet \boldsymbol{x})$ for short, where g indicates a nonlinear function like a sigmoid, $\boldsymbol{w}$ and $\boldsymbol{x}$ are the vectors defined above of dimension N, and the "•" represents a dot product. We can define that if $g >$ threshold, the image contains a cat, and if $g <$ threshold, then the image contains a fish. Machine learning algorithms offer several astute ways of choosing those weights $\boldsymbol{w}$ to minimize the number of classification errors that the algorithm makes. We will not go into the details here, but just to be concrete, we can imagine that we use a support vector machine (SVM) classifier with a linear kernel, which is a robust way of choosing those weights and which is the approach followed by Chou Hung and colleagues. This approach can be extended to many categories, not just binary classification. The key inference is that, if a reliable and simple (e.g., linear) classifier can be learned, then the pseudo-population of neurons contains sufficient information about the stimuli that can be readily extracted by biologically plausible computations (dot product followed by nonlinearity).

Using this approach, Hung et al. found that a relatively small group of ITC neurons (N ~ 200) could support object categorization quite accurately: up to ~90 percent accuracy in a task consisting of eight possible categories (where chance is one in eight). Furthermore, the pseudo-population response could extrapolate across changes in object scale and position. In other words, it is possible to fit the $\boldsymbol{w}$ values using the responses $\boldsymbol{x_1}$ to images at a particular scale, and then subsequently use the responses $\boldsymbol{x_2}$ to images at a different scale to accurately predict object labels. Thus, even if each neuron conveys only noisy information about shape differences, a small population of neurons can be powerful in discriminating among visual objects in individual trials, even extrapolating to transformed versions of the images used for training.

6.8 ITC Neurons Are More Concerned with Shape than Semantics

In the previous section, we considered whether it is possible to discriminate which object category was presented to the monkey by reading out neural activity from the ITC. Instead of decoding the object category, it is also possible to ask which specific exemplar was presented to the monkey. A population of ITC neurons excels at this question as well. Quantitatively comparing exemplar identification and categorization performance is tricky because the two tasks are not equated in terms of difficulty. First, in the experiment discussed in the previous section, there were eight categories and close to 80 exemplars. Therefore, even by chance, it is easier to get the object category right. Equating chance levels can be easily achieved by randomly subsampling and picking only eight exemplars. Yet this does not quite address a more challenging problem in this type of comparison: it is easier to distinguish a picture of a face from a picture of a house than to distinguish between two different houses.

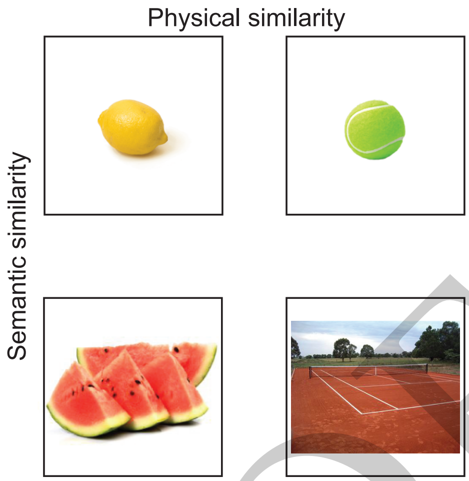

Figure 6.5 ITC neurons are more concerned with shape similarity than semantics. These images share more physical similarity along the horizontal dimension and more semantic similarity along the vertical dimension. Responses in the ITC more closely reflect the physical properties of the stimulus – including color, size, and shape.

Do ITC neurons carry any type of categorical information, or is shape the main variable that is represented in ITC? To answer this question, we need to better define what we mean by “category.” The word category is typically associated with semantic labels. One way to dissociate semantic information from pure shape information is to consider objects that are physically similar but semantically distinct and vice versa (Figure 6.5). For example, a lemon is similar to a tennis ball in terms of its color, size, and approximate shape. However, a lemon is *semantically* closer to a watermelon or a tree, and a tennis ball is semantically closer to a tennis court or a tennis racquet. There is no evidence to date that ITC neurons can link a tennis ball to a tennis court, or link a lemon to a watermelon. Instead, there is evidence that ITC neuronal responses to physically similar images are closer than responses to semantically similar but physically distinct objects.

An elegant series of experiments that tackled the question of categorization was conducted by Earl Miller’s group. They created synthetic images of cats and dogs and morphed between them in such a way that they could continuously change shape similarity without affecting categorical ownership or change category ownership with small changes in shape similarity. They found that ITC neuronal responses correlated with shape similarity better than with categorical ownership. They also recorded responses from neurons in the prefrontal cortex, which is one of the targets of ITC neurons. In contrast with the ITC neurons, the responses of those prefrontal cortex neurons did reflect the task-dependent categorical boundaries.

Another intriguing case where neuronal responses seemed to be dissociated from pure shape information is the case of those neurons recorded from the human medial temporal lobe discussed earlier (Section 6.5). Those neurons do seem to carry semantic information that transcends physical shape similarity, and those neurons receive either direct or indirect information from the anterior ITC, but they are not part of the ITC.

As repeatedly stated, the absence of evidence should not be interpreted as evidence of absence. It is conceivable that there may be semantic information that can be dissociated from pure shape information in ITC, but there is no clear evidence for this yet. Semantic information is a critical component of how we use language. In addition to the medial temporal lobe and prefrontal cortex, structures responsible for language are likely to contain neurons that represent semantic information. Furthermore, it is plausible that such semantic neurons may project back to the ventral visual cortex and modulate or sharpen visually evoked responses.

6.9 Neuronal Responses Adapt

Neurons throughout visual cortex are particularly sensitive to change. Neuronal responses dynamically depend on the temporal context. Temporal context can dramatically alter visual experience (Section 3.8), as in the illusory perception of upward motion after fixating on a waterfall, due to adaptation. As a consequence of adaptation, the responses of ITC neurons, as those in earlier parts of visual cortex, are transient (Sections 2.9, 5.7, and 5.12). If a constant stimulus is shown for many seconds, the neuronal responses only last a few hundred milliseconds.

Adaptation is an evolutionarily conserved property of visual processing that is also prevalent in other sensory systems. One function of adaptation is probably to save energy by reducing the number of spikes triggered by an unchanging stimulus. At least partly, the biophysical mechanisms underlying such suppression may be due to intrinsic changes in a neuron through transient modulation of its membrane conductance. However, adaptation is also evident at much longer time scales than the presentation of a single stimulus. For example, exposure to an adapter stimulus leads to a reduction in the neural response to subsequent presentations of the same or similar stimuli, a phenomenon known as repetition suppression. The repetitions need not be adjacent in time. Suppression is also evident even when there are other intervening stimuli, though the strength decreases with the time interval between repetitions.

Adaptation is evident at multiple time scales. As discussed in Section 2.9 and Section 5.7 (Figure 5.6), neuronal responses are typically transient and are quickly attenuated during a single trial over scales of hundreds of milliseconds, even if the stimulus remains on the screen. Repetition suppression is a manifestation of adaptation at a scale of multiple trials, typically occurring over several seconds. Figure 6.6 shows an example paradigm where the effects of adaptation can take place over minutes. In the so-called oddball paradigm, a given stimulus is repeated multiple times (high-probability stimulus shown in blue), whereas another stimulus is shown only rarely (low-probability stimulus shown in orange). Figure 6.6B–C shows average population responses from multiple neurons in the rat primary visual cortex (V1) and in a higher

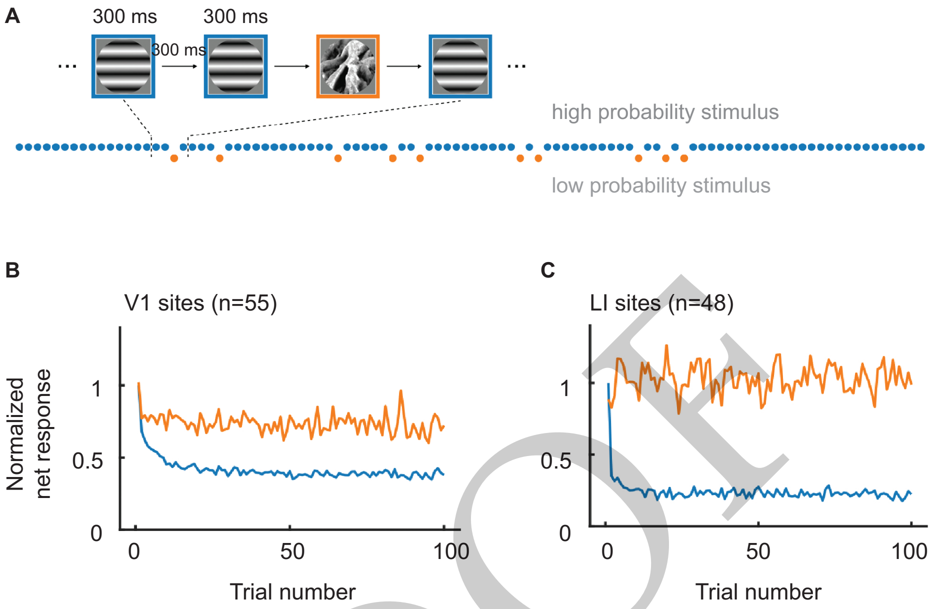

Figure 6.6 Neural adaptation increases the salience of novel stimuli. (**A**) Oddball paradigm where one stimulus is presented with high probability (blue) and another stimulus is presented with low probability (orange). (B)–(C) Normalized average population responses from neurons in rat primary visual cortex (B) and the latero-intermediate area (C) as a function of trial number for low- and high-probability stimuli. Modified from Vinken et al. 2020

visual area called the latero-intermediate area (LI). Whereas there is general agreement about what constitutes the primary visual cortex across species, it is less evident how to establish homologies between higher visual areas across species, and therefore, the nomenclature diverges across species. Repeated presentation of the high-probability stimulus leads to a sharp reduction in the neural responses over trials (blue). In contrast, the low-probability stimulus evokes a larger response, especially in area LI. This effect can help detect novel stimuli or changes in the environment.

Adaptation occurs throughout the visual system. The consequences of adaptation are stronger in higher areas like the ITC, or area LI in the rat, compared to earlier neurons like those in V1, probably due to the cumulative effects through a hierarchical cascade of neurons, each showing increasingly larger effects of adaptation that impact the next stage. In other words, adaptation leads to a reduction in response in RGCs and in the LGN, which in turn implies a weaker input to V1, and this is compounded with the intrinsic effects of adaptation in V1. The weaker V1 signals lead to a reduced input to V2, which is compounded with the intrinsic adaptation effects in V2, and so on. Another effect that could contribute to the increased adaptation in higher stages is that earlier areas are more sensitive to small eye movements, hence reducing the similarity in the inputs for prolonged stimulus durations or repetitions of the same stimulus.

6.10 Representing Visual Information in the Absence of a Visual Stimulus

Perceptually, prolonged exposure to a stimulus often leads to a temporarily reduced sensitivity to its features. The lingering effects after removal of the stimulus are called aftereffects, which have been described for a wide range of low- to high-level visual stimulus properties, and they are considered to be related to adaptation.

In addition to aftereffects, exposure to a stimulus leaves a memory trace that allows subjects to remember what they have just seen. A classical experiment used to study memory effects at short time scales is the delayed match-to-sample task. Subjects are presented with an image, the image disappears, and there is a delay of several seconds. After this delay, a second image is shown, and subjects have to indicate whether the second stimulus matches the first one (because it is identical, or because it is a scaled or rotated version of the same object, or because they match in color or any other property). Typically, the delay period consists of a blank screen. For subjects to be able to execute this task, neurons somewhere in the brain need to be able to maintain information about the preceding stimulus, even during the blank screen. Such information stored for a few seconds is typically referred to as *working memory*.

It turns out that, although the responses of neurons in the ITC are drastically reduced in the absence of visual stimulation, the activity does not fully return to baseline (Figure 6.7). Instead, ITC neurons maintain a small activation above baseline during the delay. Furthermore, this delay activity is stimulus selective: a neuron will maintain higher delay activity if its response to the preceding stimulus was higher.

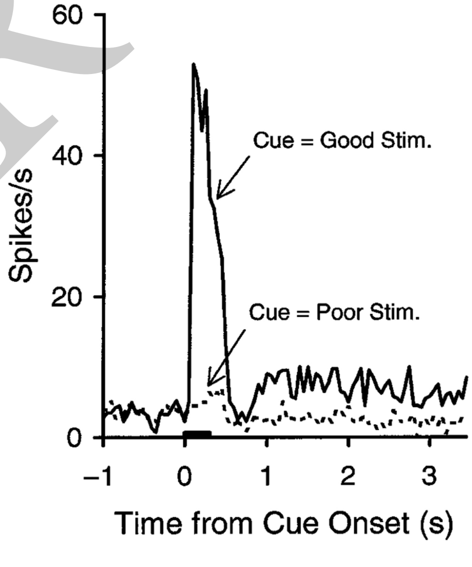

Figure 6.7 Selective neuronal response during working memory. Responses of a neuron during a delayed match-to-sample task when the cue was a good stimulus (solid) or a poor stimulus (dashed). The horizontal black bar denotes the cue duration (300 milliseconds). Reproduced from Chelazzi et al. 1998

Some investigators have interpreted those neuronal responses during the delay period in the absence of visual stimulation as an example of *visual imagination*. They argue that the subjects are imagining the sample stimulus to hold it in memory during the delay. To the extent that this is the case, it would seem that ITC neurons may show a selective response that matches the animal's internally generated percept irrespective of sensory inputs. It is difficult to directly test this idea due to the challenge in eliciting volitional visual imagery in animals. In humans, several investigators have measured neuronal correlates of volitional imagery, but those responses have been investigated in the medial temporal lobe rather than in the ITC.

Taking these ideas a step further, another situation where visual percepts can be generated in the absence of concomitant visual inputs is during dreams. Humans often report vivid visual imagery during dreams. Whether the visual cortex is involved in the representation of those visual percepts remains to be determined. We will come back to this discussion in Section 10.4.

6.11 Task Goals Modulate Neuronal Responses

We have described properties of ITC neuron responses as if they were static and immutable, but this is far from the case. For example, the reduced response to repeated presentation of the same stimulus (Section 6.9, Figure 6.6) shows that temporal aspects of the task can modulate neuronal responses. Beyond the temporal reduction in the response, other aspects of the current task goals can also modulate responses throughout ventral visual cortex.

One of the most studied forms of task-dependent modulation of neural responses is the effect of attention introduced in Section 5.17. A typical paradigm to study spatial attention is to train a monkey to fixate in the middle of the screen while devoting covert attention to either the left or right hemifields. Under these conditions, monkeys show enhanced performance and faster reaction times during visual discrimination tasks when a stimulus is presented within their locus of attention. Furthermore, the same visual stimulus, presented at the same location in the receptive field of the neuron under study, evokes a stronger response when the monkey is paying attention to the location encompassing the stimulus (Figure 6.8). Such attentional modulation is evident throughout the entire range of stimulus preferences.

Other aspects of the task goals can also modulate neuronal responses along the ventral visual cortex. During visual search experiments, the subject is looking for a particular object or a particular feature (e.g., looking for Waldo). For example, Robert Desimone's laboratory trained monkeys to look for red oriented bars. Under these conditions, neuronal responses to red objects were enhanced throughout the visual field. Other typical tasks involve flashing images while subjects have to indicate in a forced-choice yes/no fashion whether a particular target object is present or not. Here again, trials containing the target object or object category trigger enhanced neural responses. In Section 3.7, we described two forms of temporal contextual modulation: priming and backward masking. Both of these manipulations also impact the responses

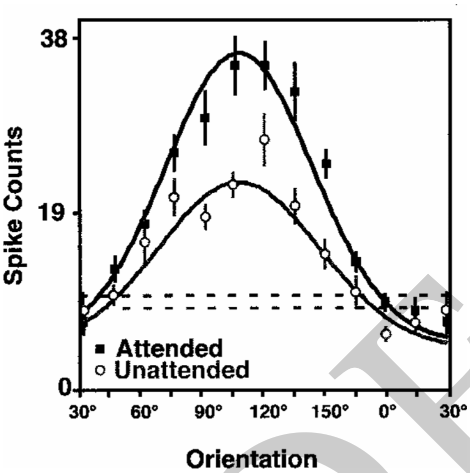

Figure 6.8 Spatial attention modulates responses in area V4. Modulation of the tuning curve of a V4 neuron in response to gratings of varying orientations when the animal attends to the receptive field location (solid squares) or when attention is diverted away from the receptive field location (empty circles). Reproduced from McAdams and Maunsell 1999

in the ITC. In sum, while the contents of what is on the screen at a particular moment are the main determinants of the responses of ITC neurons, current goals, spatial context, temporal context, and other task demands can modulate the responses throughout the ventral visual cortex.

6.12 The Role of Experience in Shaping Neuronal Tuning Preferences

Neuronal responses can be altered during a task as a consequence of adaptation (Section 6.9), memory (Section 6.10), or task-oriented goals (Section 6.11). Neuronal responses can also be altered over longer time scales. Neuronal tuning preferences are malleable and depend strongly on the diet of visual experience that the animal is subject to. A perennial debate focuses on the relative role that nature and nurture play in shaping the architecture of visual cortex and neuronal response tuning functions.

Genetics largely dictates the basic architecture of the visual system. Animals are born with visual structures like the eyes, LGN, and different cortical areas. While there are differences between species, the six cortical layers, as well as their canonical connectivity with each other and between cortical areas, seem to be either already present at birth or formed shortly thereafter. Furthermore, there is a small but clear degree of orientation selectivity that can be measured in primary visual cortex right at the time of eye opening in ferrets, cats, and monkeys.

Mature tuning properties are a consequence of experience. Several experiments have shown that visual inputs shape the mature neural tuning in primary visual cortex. For example, monocular deprivation (i.e., eliminating inputs from one eye) leads to an expansion of neuronal preferences for the active eye at the detriment of neurons responding to inputs from the deprived eye. Dark rearing leads to impaired orientation tuning throughout the primary visual cortex. Furthermore, experiments in which cats are reared in environments where they are predominantly exposed to vertical lines rather than horizontal lines lead to a preponderance of V1 neurons preferring vertically oriented bars rather than horizontal ones.

Given that even the early stages in cortical processing depend on visual experience, it is perhaps less surprising that subsequent stages can also be modified by changing the statistics of the visual inputs. As mentioned earlier, neurons in macaque ITC can respond selectively to shapes like paperclips after the monkey is exposed to those images. It is clear that monkeys are not born with neurons tuned to arbitrary paperclip shapes. Furthermore, monkeys can be trained to recognize symbols, like numbers or letters. After training, ITC neurons can also respond selectively to those novel shapes, and, again, such tuning is not present at birth or without training. The presumed ethological relevance of natural stimuli like faces has led some investigators to suggest that tuning for those shapes could be innate. However, careful experiments have refuted this hypothesis. If monkeys are reared in an environment without any exposure to faces, then investigators do not find clusters of neurons tuned to faces. In sum, current evidence suggests that in the development of visual response functions, genetics provides the underlying architecture and the plasticity rules while environmental statistics guide the learning of tuning functions for neurons throughout visual cortex.

The shaping of ITC neuron tuning happens not only during development but also in adults. The paperclip and numeric symbol experiments were both conducted in adult monkeys who were exposed to those novel images over periods of several months.

Neuronal tuning can also even be changed much more rapidly. For example, it is likely that if we learn to recognize characters in a new language, or if we learn to recognize a new person, we would find changes in neuronal tuning in the ITC. Indeed, elegant experiments in monkeys have shown that it is possible to alter the tuning properties of ITC neurons over the course of a recording session lasting less than an hour.

6.13 The Bridge between Vision and Cognition

The studies discussed here constitute a non-exhaustive list of examples of the type of responses that investigators describe in the highest parts of the inferior temporal cortex. While the field has acquired a considerable number of such examples, there is an urgent need to put together these empirical observations into a coherent theory of visual recognition, which will be the focus of the next chapters.

It is critical to develop more quantitative and systematic approaches to examine feature preferences in extrastriate visual cortex (and other sensory modalities). The methodology described in Section 6.4 provides initial steps toward unbiased ways of interrogating neuronal tuning functions in visual cortex. At the same time, we should aim to describe a neuron's preferences in quantitative terms, starting from pixels.

What types of shapes would a neuron respond to? This quantitative formulation should allow us to make predictions and extrapolations to novel shapes. It is not sufficient to show stimulus A and A" and then interpolate to predict the responses to A'. If we could truly characterize the responses of the neuron, we should be able to predict the responses to any different shape B. Similarly, as emphasized multiple times, feature preferences are intricately linked to tolerance of object transformations. Therefore, we should be able to predict the neuronal response to different types of transformations of the objects. Much more work is needed to understand the computations and transformations along ventral visual cortex. How do we go from oriented bars to complex shapes such as faces? A big step would be to take a single neuron in, say, the ITC, be able to examine the properties and responses of its afferent V4 units to characterize the transformations from V4 to the ITC.

This formulation presupposes that a large fraction of the ITC responses is governed by their V4 inputs. However, we should keep in mind the complex connectivity in cortex and the fact that ITC neurons receive multiple other inputs as well (recurrent connections, bypass inputs from earlier visual areas, back projections from the medial temporal lobe and prefrontal cortex, and connections from the dorsal visual pathway). There is clearly plenty of unexplored territory for the courageous investigators who dare explore the vast land of the extrastriate ventral visual cortex and the computations involved in processing shapes. Another incipient area of active research that is still in its infancy and will require serious scrutiny in the near future is to further our understanding of how high-level visual information interfaces with the rest of cognition.

6.14 Summary

- The inferior temporal cortex (ITC) sits at the pinnacle of the visual cortical hierarchy, receiving strong inputs from both ventral and dorsal cortical areas and projecting widely to areas involved in episodic memory formation, decision making, and cognitive control.
- Monkey and human ITC neural responses are selective for a wide range of shapes, including abstract patterns, bananas, chairs or faces.
- ITC neurons represent an extensive overcomplete dictionary of features, are more concerned with shape rather than semantics, and show invariance to image transformations.
- ITC neurons can complete patterns from partially visible stimuli.
- The activity of neural populations in the ITC in single trials can be used to decode object information with linear classifiers.
- Neural responses continue representing selective visual information even in the absence of a visual stimulus.
- Neuronal tuning properties are the result of experience with the statistics of the visual world.

Further Reading

See more references at http://bit.ly/364H8WR

- Arcaro, M. J.; Schade, P. F.; Vincent, J. L.; Ponce, C. .R.; and Livingstone, M. S. (2017). Seeing faces is necessary for face-domain formation. *Nature Neuroscience* 20: 1404–1412.
- Freedman, D.; Riesenhuber, M.; Poggio, T.; and Miller, E. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science* 291: 312–316.
- Hung, C. P.; Kreiman, G.; Poggio, T.; and DiCarlo, J. J. (2005). Fast read-out of object identity from macaque inferior temporal cortex. *Science* 310: 863–866.
- Liu, H.; Agam, Y.; Madsen, J. R.; and Kreiman, G. (2009). Timing, timing, timing: fast decoding of object information from intracranial field potentials in human visual cortex. *Neuron* 62: 281–290.
- Logothetis, N. K., and Sheinberg, D. L. (1996). Visual object recognition. *Annual Review of Neuroscience* 19: 577–621.