Visual Object Recognition: Do We Know More Now Than We Did 20 Years Ago?

Article *in* Annual Review of Psychology · February 2007

DOI: 10.1146/annurev.psych.58.102904.190114 · Source: PubMed

2 authors:

Jessie J Peissig
California State University, Fullerton
57 PUBLICATIONS **568** CITATIONS

SEE PROFILE

Michael J. Tarr
Carnegie Mellon University
238 PUBLICATIONS **13,301** CITATIONS

SEE PROFILE

Visual Object Recognition: Do We Know More Now Than We Did 20 Years Ago?

Jessie J. Peissig and Michael J. Tarr

Department of Cognitive and Linguistic Sciences, Brown University, Providence, Rhode Island 02912; email: jpeissig@fullerton.edu, Michael_tarr@brown.edu

Annu. Rev. Psychol. 2007. 58:75–96

First published online as a Review in Advance on August 11, 2006

The *Annual Review of Psychology* is online at http://psych.annualreviews.org

This article's doi: 10.1146/annurev.psych.58.102904.190114

0066-4308/07/0110-0075$20.00

Key Words

structural descriptions, view-based, neural codes, visual features, category-selectivity

Abstract

We review the progress made in the field of object recognition over the past two decades. Structural-description models, making their appearance in the early 1980s, inspired a wealth of empirical research. Moving to the 1990s, psychophysical evidence for view-based accounts of recognition challenged some of the fundamental assumptions of structural-description theories. The 1990s also saw increased interest in the neurophysiological study of high-level visual cortex, the results of which provide some constraints on how objects may be represented. By 2000, neuroimaging arose as a viable means for connecting neurons to behavior. One of the most striking fMRI results has been category selectivity, which provided further constraints for models of object recognition. Despite this progress, the field is still faced with the challenge of developing a comprehensive theory that integrates this ever-increasing body of results and explains how we perceive and recognize objects.

Contents

INTRODUCTION

At a functional level, visual object recognition is at the center of understanding how we think about what we see. Object identification is a primary end state of visual processing and a critical precursor to interacting with and reasoning about the world. Thus, the question of how we recognize objects is both perceptual and cognitive, tying together what are often treated as separate disciplines. At the outset, we should state that in spite of the best efforts of many to understand this process, we believe that the field still has a long way to go toward a comprehensive account of visual object recognition. At the same time, we do believe that progress has been made over the past 20 years. Indeed, visual object recognition is a poster child for a multidisciplinary approach to the study of the mind and brain: Few domains have utilized such a wide range of methods, including neurophysiology, neuroimaging, psychophysics, and computational theory. To illustrate this progress, we review the state of the art circa 1985 and contrast this with the state of the art today (2006). We note that some problems have been solved, some have evolved, some have become extinct, and new ones have arisen.

Although there were clearly many neuroscientific and behavioral antecedents (e.g., Konorski 1967, Rock 1973, Selfridge 1959) to Marr & Nishihara's (1978; popularized in Marr's 1982 book) seminal paper, it, more than any other single publication, is arguably the spark for what we think of as the modern study of visual object recognition. Interestingly, although it was heavily motivated by neuropsychological data and behavioral intuition, Marr and Nishihara's theory was purely computational, with no attempt at empirical validation. Colleagues of Marr took a similar approach, identifying in principle problems with then state-of-the-art theories of recognition, but presenting little in the way of concrete data to validate, invalidate, or extend such theories (Pinker 1984).

One reason for this hesitancy to step into the fray may have been the enormous level of flexibility exhibited by the primate visual system—an issue that remains with us today and challenges all would-be accounts of recognition. If anything, the more we have learned about our recognition abilities, the more daunting the problem has become. For example, results regarding the incredible rapidity with which successful recognition is achieved have imposed significant new constraints on current theories. Consider the study by Thorpe et al. (1996), in which they allowed observers only 20 ms to determine whether an animal was present in a natural scene. Event-related potentials (ERPs) measured during performance of this task reveal, approximately 150 ms after stimulus onset, a significant difference between the neural responses for trials in which there is an animal and trials in which there is not. Such data indicate that the primate visual system processes complex natural scenes quite rapidly and with only the briefest of inputs. Interestingly, this result and many more from the past two decades have not been integrated into any extant theory. Thus, although we have made significant empirical progress, as discussed in the next two sections, theoretical models have lagged behind. In future model building, it

behooves the field to consider the breadth of psychophysical, neurophysiological, neuropsychological, and neuroimaging data that form the basis of this progress.

CIRCA 1985—STRUCTURAL DESCRIPTION MODELS

Marr & Nishihara (1978) introduced the idea of part-based structural representations based on three-dimensional volumes and their spatial relations. In particular, they proposed that object parts come to be mentally represented as generalized cones (or cylinders) and objects as hierarchically organized structural models relating the spatial positions of parts to one another. To some extent, this proposal was motivated by 1970s-era models from computer vision and computer graphics, but also by the desire of Marr and Nishihara to have their scheme satisfy several computational criteria. First, representations should be accessible. That is, the necessary information to recover a representation in a computationally efficient manner should be available in the visual image. Second, representations should be unique. That is, objects that seem psychologically different from one another should be representationally discriminable from one another. At the same time, representations should be generic, so that the same representational predicates are sufficient to capture the wide variability of objects we encounter. Third, representations should be both stable and sensitive. That is, the wide range of two-dimensional images generated by a single object seen under different combinations of object pose, configuration, and lighting should map to a common object representation (i.e., it is the same object), but the representation should also be sufficiently detailed to make discriminations between visually similar objects (i.e., those are two different objects).

One of the most challenging issues for Marr and Nishihara was the fact that, when rotated in depth, three-dimensional objects change their two-dimensional retinal projection (the problem of viewpoint invariance); stable object representations require addressing this problem. Thus, in their theory, object parts encoded as generalized cones are represented in an object-centered manner, that is, in a coordinate system that decouples the orientation of the object from the position of the viewer. The significance of this assumption is that the same generalized cones can be recovered from the image regardless of the orientation of the object generating that image. Consequently, object recognition performance should be independent of both observer position and object orientation. Thus, at least for changes in viewing position—the most daunting problem in the eyes of Marr and Nishihara (and much of the field, as we discuss below)—the many-to-one mapping called for by the stability constraint is satisfied. Conversely, the sensitivity constraint is satisfied by two properties of the proposed representation. First, generalized cones—a two-dimensional cross-section of any shape swept along an axis of that shape—can capture an infinite number of part shapes. Clearly, such powerful representational units have the potential to discriminate between objects having only subtle shape differences. Second, these object parts are related to one another by metrically precise spatial relations at multiple scales. That is, a given representation can be refined down to the shape and configural details necessary to distinguish it from other objects of similar coarse shape. For example, two different faces might have subtly different relations between the angles of their noses and eyes as well as subtly different generalized cones representing the shapes of the noses. However, as mentioned above, Marr and Nishihara offered no empirical support for this model.

RBC: recognition by components

By far the most well-received structural-description model is recognition by components (RBC; Biederman 1985). RBC is quite similar to Marr and Nishihara's model, but has been refined in important ways. First and foremost is the psychophysical support for the model presented by Biederman (1985; Biederman & Cooper 1991, 1992; Biederman

& Gerhardstein 1993).[1] Second, Biederman included two important properties to make the model more tractable: (*a*) a restricted set of three-dimensional volumes to represent part shape—Geons—defined by properties, including whether the edge is straight or curved, whether the object is symmetrical or not, if the cross-section is of constant size or expands or contracts, and whether the axis is straight or curved;[2] and (*b*) a single layer of qualitatively specified spatial relations between parts—for example, "above" or "beside." At the same time, Biederman retained the idea of view-invariant representations, but modified it to be based on three-dimensional shape properties that project to stable local contour configurations—so-called viewpoint-invariant properties (Lowe 1985). Critically, Geons are specified by the co-occurrence of multiple instances of these properties in the image. For example, a brick Geon might project to three arrow junctions, three L junctions, and a Y junction that remain visible over many different rotations in depth. RBC assumes that these sets of contour features are identified in the image and used as the basis for inferring the presence of one of the 30 or so Geons that constitute RBC's building blocks for representing part shapes. Because these features are themselves viewpoint invariant (up to occlusion), the recovery of parts is also viewpoint invariant. Object representations are simply assemblies of such parts—deemed "Geon-structural descriptions," and are constructed by inferring the qualitative spatial relations between recovered parts. Because these relations are viewpoint invariant across rotations in depth, the recognition process is likewise viewpoint invariant (but not for picture-plane rotations; e.g., the relation "above" is perturbed when an object is turned upside down).

A final issue raised by Biederman in the RBC model is the default level of object recognition. That is, there is really no claim that all of object recognition is accomplished as outlined above. Rather, Biederman suggests that typical object recognition tasks occur at the basic level (Rosch et al. 1976) or entry level (Jolicoeur et al. 1984). More specifically, the first and fastest label applied to most objects is their category label (e.g., bird), the exception being visually idiosyncratic category exemplars (e.g., penguin). RBC only explains how observers recognize objects at this level, making no attempt to account for how we arrive at either superordinate labels (e.g., animal—probably more cognitive than visual) or subordinate labels (e.g., species labels such as fairy-wren or individual labels such as Tweety Bird). Thus, even given RBC as a plausible model of basic-level recognition circa 1985, there is no particular theory for how a wide variety of visual recognition tasks are accomplished.

So Where are We?

The late 1970s and early 1980s harbored significant changes in how the field thought about the mind and brain. In particular, accounts of recognition and categorization tasks shifted from purely cognitive problems to, at least in part, perceptual problems. Shepard & Cooper's (1982) and Kosslyn's (1980) empirical investigations into mental imagery brought home the idea that vision is more than an input system. That is, considerable mental work is accomplished before we ever invoke symbolic modes of processing (Fodor 1975) or semantics. Building on this transformation, theorists such as Marr and Biederman formulated theories of visual recognition that postulated high-level visual representations for recognition and categorization. Their models reflected the emerging bodies of both empirical data and work in computational vision. The elder author of this article remembers the excitement

[1] Although the bulk of these results were published during the early 1990s, similar empirical designs and results are outlined in the original 1985 paper, as well as in several technical reports.

[2] Geons constitute a highly restricted subset of generalized cones.

surrounding this time—solutions to problems as complex as visual object recognition were just around the corner (e.g., Pinker 1984).

At the same time, this excitement was tempered by clear gaps in knowledge, not the least of which was the relatively small amount of behavioral and neuroscientific data on how humans and other primates actually recognize objects. Although hindsight is 20/20, it appears that many of us underestimated both the extreme complexity of cortical areas past V4 (Van Essen 1985), the flexibility and complexity of the visual recognition tasks we routinely solve, and the computational intricacies of building an artificial vision system. In the sections below, we review some, but certainly not all, of the results that emerged from 1985 to present—not in small part due to the excitement generated during these seminal years.

CIRCA 1990—VIEW-BASED MODELS

As discussed above, one of the core characteristics (and appeals) of structural-description models is their viewpoint invariance. Such models predict that faced with a novel view of a familiar object, observers should be able to recognize it and should do so with no additional cost.[3] That is, response times and errors should be equivalent regardless of viewpoint. Interestingly, despite being one of the core tenets of these models, this assumption had not been tested. This omission may have been in part due to the strong intuition we have that object recognition is effortless even when faced with novel viewing conditions. Even Shepard & Cooper (1982), who explored extensively the nature of mental rotation in making handedness judgments, argued on logical grounds that this viewpoint-dependent process was not used for object recognition.

View-based models: models of visual object recognition in which objects are represented with respect to their original viewing conditions

At nearly the same time that structural-description models became popular, several groups undertook empirical studies of invariance in visual object recognition.[4] Jolicoeur (1985) simply asked observers to view picture-plane misoriented line drawings of common objects and then to name them as quickly as possible. He found that the time to name a given object was related to how far it was rotated from the upright, revealing a systematic response pattern similar to that found by Shepard and Cooper. However, Jolicoeur also found that this effect was relatively small and diminished quite rapidly with repeated presentations of the objects. Thus, it remained an open question as to whether the effect was due to viewpoint-dependent representations or to more transient viewpoint-dependent processes elicited by the task. Building on this ambiguity, Tarr & Pinker (1989) argued that the critical question was not how observers recognize familiar objects—which potentially had already been encountered in multiple viewpoints—but rather how observers recognize novel objects when viewpoint has been controlled during learning.

Tarr & Pinker (1989) studied this proposal by teaching observers to name several novel shapes appearing at select orientations. They found that observers exhibited a significant cost—in both response times and error rates—when recognizing trained shapes in new orientations and that these costs were systematically related to the distance from a trained view. Interestingly, the pattern looked a good deal like

[3] In Marr and Nishihara's model, this prediction is predicated on the successful recovery of the axes and cross sections describing the generalized cones representing parts. In Biederman's model, the same prediction is predicated on the successful recovery of the same Geons from different viewpoints—due to self-occlusions within objects, some rotations in depth will obscure Geons that are part of the representation; in such instances, recognition will become viewpoint dependent.

[4] This increased interest was of course due to the theoretical excitement discussed above. However, it also may have been spurred on in part by the new availability of desktop personal computers (PCs) that were sufficiently powerful to both display images and to record response times (the first IBM PC appeared in 1981 and the first Apple Macintosh™ in 1984).

that seen in mental rotation tasks. These and related results were taken as powerful evidence for viewpoint-dependent object representations—sometimes called views—and the use of a continuous mental transformation process to align images of objects in novel viewpoints with familiar views in visual memory (e.g., Shepard & Cooper 1982, Ullman 1989). Bolstering these claims, Tarr (1995) found similar results for three-dimensional objects rotated in depth. That is, the same relationship between familiar and unfamiliar viewpoints holds, even when depth rotation changes the visible surfaces in the image.

More recent results suggest that the viewpoint dependencies seen in object recognition tasks do not arise as a consequence of mental transformation processes. In a study that demonstrates how neuroimaging can inform functional models, Gauthier et al. (2002) explored whether viewpoint-dependent object recognition and viewpoint-dependent handedness judgments (i.e., mental rotation) recruit the same or overlapping neural substrates. Using fMRI, they found that localized regions of the dorsal pathway responded in a viewpoint-dependent manner during mental rotation tasks, while, in contrast, localized regions of the ventral pathway responded in a viewpoint-dependent manner during object recognition tasks. That is, although the behavioral data were nearly identical for the two tasks, the neural bases of the behaviors were qualitatively different, being subserved by entirely different brain areas. This finding strongly suggests that Tarr & Pinker's (1989, Tarr 1995) hypothesis regarding shared mechanisms for mental rotation and object recognition is incorrect. As such, alternative models as to how disparate views of the same object are matched must be considered.

The early 1990s not only saw several studies that supported the hypothesis that objects are represented in a viewpoint-dependent manner, but also offered new ideas as to how unfamiliar views are matched with familiar views. At the same time, these studies helped to reinforce the overall picture of a significant role for viewpoint-dependent recognition processes (e.g., Lawson et al. 1994). Most notably, Poggio & Edelman (1990) developed both a computational framework and a supporting collection of empirical results (Bülthoff & Edelman 1992, Edelman & Bülthoff 1992) that reinforced the idea of view-based recognition. Their model offers one possible mechanism for matching inputs and representations without mental rotation, proposing that the similarity between input and memory is computed using a collection of radial basis functions, each centered on a meaningful feature in the image. Object representations are view-based in that they are encoded with respect to a particular set of viewing parameters, and matching such representations to novel views of known objects produces errors proportional to the dissimilarity between the two. Thus, larger rotations are likely to produce larger errors, but no mental transformation is used. Another important implication of their model is that similarity is computed with reference to all known views. Therefore, a novel view centered between two known views will be better recognized (interpolation in the model's representational space) as compared to a novel view an equivalent distance away from only one known view (extrapolation in the model's representational space). This model was tested by Bülthoff & Edelman (1992; also Edelman & Bülthoff 1992). Consistent with the specific predictions of the model, Bülthoff and Edelman found the best recognition performance for unfamiliar viewpoints between trained views; poorer performance for viewpoints outside of trained views, but along the same axis; and the poorest performance for viewpoints along the orthogonal axis.

Models based on principles similar to those originally proposed by Poggio and Edelman are still popular today (Riesenhuber & Poggio 1999). Lending empirical support to this framework, Jiang et al. (2006) used an fMRI adaptation paradigm to examine the neural responses to target faces and face morphs

between the target face and a nontarget face. Consistent with similarity-based recognition models, they found that adaptation increased as the similarity between target faces and morphed faces increased, indicating a larger degree of overlap in the neural populations coding for the two stimuli.

Jiang et al. (2006) also computationally tested whether their exclusively feature-based model can account for recognition behaviors that appear to be configurally based. Surprisingly, they found that their model was sensitive to configural manipulations (e.g., Rotshtein et al. 2005). Thus, a simple view-dependent, featural model is able to account for a significant number of behavioral and neural findings. At the same time, the large majority of studies supporting this approach speak only to the issue of the nature of the features used in object representations, not to whether structural information, independent of the particular features, is used (Barenholtz & Tarr 2006). Put another way, demonstrations of viewpoint dependence only implicate viewpoint-dependent features; not how those features are related to one another. At the same time, the fact that feature-based models are able to produce configural effects does not rule out, in and of itself, the possibility that such effects are a consequence of structural information. In particular, although configural coding may be realized through larger view-based features (Zhang & Cottrell 2005), the same coding may take the form of explicit relations between features. In the end, it is unclear whether the large body of work focused on view-based models is compatible with, incompatible with, or just orthogonal to structural models of object representation such as RBC (Biederman 1985).

CIRCA 1995—WHAT'S HAPPENING IN THE BRAIN?

Well before most cognitive scientists were thinking about the problem, neurophysiologists were studying visual object recognition by mapping the responses of single neurons in primate visual cortex. In two landmark studies, Gross and colleagues (Gross & Bender 1969, Gross et al. 1972) reported that neurons in the inferotemporal (IT) cortex of macaques responded most strongly to complex visual stimuli, such as hands and faces. As a measure of the times, this result was met with great skepticism and it was years before the field accepted the idea of such strong stimulus selectivity for single neurons. Much of this conservatism stemmed from the fact that prior to this study, recordings from single units in the visual system typically employed simple stimuli, such as light spots and bars. Moreover, most recordings were made in early visual areas such as V1 and V2, not the higher-level regions investigated by Gross and Bender. Interestingly, Gross and Bender found that cells in these higher-level areas of IT showed very little response to simple stimuli, but great sensitivity to complex stimuli. The logical conclusion is that this and related areas of IT are critical to complex visual processing and, presumably, visual object recognition. As reviewed below, this study was the first of many to come, with neurophysiological results coming to play an important role in how we understand the object recognition process.

IT: inferotemporal cortex

Before we discuss more recent results from neurophysiology, it is worth stepping back and considering some of the assumptions that make single-unit recording a viable tool for the study of visual object recognition. Because most neurophysiological results are based on the responses of single neurons, it is quite possible that years of neuron-by-neuron probing will reveal almost nothing. Consider a plausible neural architecture in which the representation of objects is wholly distributed: A good proportion of the neural population available for object representation participates in the encoding of each known object. Moreover, this particular population code is distributed across all of IT cortex, and active neurons are interleaved with inactive neurons. Such a coding scheme would render single-unit responses all but useless: The neural firing patterns produced by the perception and

STS/AMTS: superior temporal sulcus/anterior medial temporal sulcus

recognition of different objects or object classes would mostly look alike. And when different patterns were found within the small populations recorded in most studies, they would not be particularly diagnostic relative to the overall code. Conversely, it is also possible that only a tiny proportion of the neural population available for object representation participates for any one object. In this case, a neuroscientist might die, still never having managed to find even one neuron selective for any of the particular subset of objects used as stimuli. Happily, neither extreme appears to hold. That is, neuroscientists find object- and class-selective visual neurons in short order—usually in about 10% to 20% of the neurons from which they record. At the same time, they do not find uniform selectivity; different objects produce different and consistent patterns of firing across the measured neural populations. These basic facts suggest a sparse, distributed code that, when uncovered by neurophysiology, is capable of providing constraint on how objects come to be both represented and recognized.

Gross and Bender's discovery of neuronal selectivity for complex stimuli eventually came to be the de facto model of how objects are processed in IT. For example, Perrett et al. (1984) proposed that IT is organized into anatomical groupings of neurons labeled as columns and minicolumns (one cell wide!) that encode for visually similar high-level features. A host of similar findings by Perrett and many other groups (e.g., Desimone et al. 1980; nice reviews are provided by Gross 1994 and Rodman 1994) helped paint a picture of IT as a highly organized structure in which single units appear to code for individual objects or object classes. Of course, such data were rightly taken only as evidence for an orderly cortex, not one in which single units uniquely code for specific objects (e.g., "grandmother" cells). At the same time, the actual neural code for objects seemed (and seems!) elusive. Familiar objects must be represented in the brain somewhere; thus, it is not surprising to find evidence for this. The problem is that simply enumerating the selectivity of hundreds or even thousands of single units does not tell us much about how such exquisite selectivity arises in the first place or what role it plays in object recognition.

Studies exploring neural learning take us beyond such simplistic cortical mapping and give us some insight into one of the most salient characteristics of object recognition, its flexibility. In a study motivated by Bülthoff & Edelman's (1992) psychophysical results and Poggio & Edelman's (1990) model, Logothetis & Pauls (1995) trained monkeys to recognize paper clip and blob objects at a small number of specific viewpoints. Over the course of training, the experimenters added new views of the objects that were generated by rotations in depth. As with humans, behaviorally the monkeys showed better generalization to views that were closer to the trained, familiar views. Again similar to humans, further experience with these novel views eventually led the monkeys to exhibit invariant performance across a wide range of viewpoints. The questions at the neural level are, how did they accomplish generalization in the first place, and what did they learn to achieve invariance?

To address this question, Logothetis & Pauls (1995) recorded from IT neurons in superior temporal sulcus (STS) and anterior medial temporal sulcus (AMTS) while the monkeys were presented with the familiar objects in familiar viewpoints. About 10% of the neurons recorded from (71 out of 773) responded selectively to particular wire-frame objects at specific trained views. In contrast, they found only eight neurons (~1%) that responded to a specific object in a view-invariant manner. Critically, Logothetis and Pauls found no neurons that were preferentially selective for unfamiliar object views. Invariant recognition performance was apparently achieved by pooling across the generalization gradients of the neurons found to be selective for familiar views. The finding of view-tuned neurons is highly consistent with and provides a neural mechanism for the behaviorally based argument that observers represent objects at

multiple viewpoints (Bülthoff & Edelman 1992, Tarr & Pinker 1989). That is, neural selectivity emerges as a consequence of specific visual experiences, whether these experiences be new objects or new views of familiar objects. Moreover, it does not appear that the primate visual system is "recovering" an invariant object representation. Rather, it is relying on the specific representation of many different examples to efficiently cover image space in a manner that supports robust generalization.

Although the results of Logothetis & Pauls (1995) are compelling, some groups voiced concern that overtraining with static, specific viewpoints of objects leads to idiosyncratic object representations that are more view-selective than one might expect to arise from real-world experiences where both observers and objects interact dynamically (the same critique may be applied to the majority of the human psychophysical findings on view sensitivity). To address this concern, Booth & Rolls (1998) recorded from single neurons in the STS of monkeys that had experienced novel objects by playing with them in their cages. However, actual neurophysiological recordings were done with a fixed observer and with static images depicting photographs of the objects at different viewpoints. Of the neurons that were recorded from, Booth and Rolls found that 49% (75 out of 153) responded to specific views and only 14% (21 out of 153) showed view-invariant responses. However, using an information theoretic analysis, they found that the combined firing rates of the view-invariant cells alone were sufficient to discriminate one object from another. That is, the pattern of firing rates across the population of 21 view-invariant neurons differed more for different objects than it did for different views of the same object. This discriminatory power was not feature dependent in that it was observed for different views of the same object that shared no visible features and for grayscale versions of the object images (e.g., discrimination was not based on color features). Critically, Booth and Rolls also demonstrated that the view-invariant neurons were selective for the particular objects with which the monkeys had played and not for similar objects they had not previously seen. Thus, Booth and Rolls were able to both reinforce and elaborate on the results of Logothetis and Pauls. Specifically, view-invariant neurons are intermingled with view-tuned neurons, and it is theorized that the pooling of the responses of collections of the view-tuned neurons in the population enables invariance in the view-invariant neurons. Finally, such neurons appear capable of supporting recognition in the sense that their responses, considered together, were sufficient to discriminate one familiar object from another.

Further demonstrating the explanatory power of using neural population codes to represent objects, Perrett et al. (1998) demonstrated that collections of feature-selective neurons coding for different parts of an object produce, when their responses are considered together, patterns of viewpoint dependency in object recognition tasks. This finding is based on two facts: (*a*) each feature-sensitive neuron is highly viewpoint dependent, and (*b*) the overall rate at which the population will reach a response threshold sufficient for recognition is dependent on its cumulative responses. That is, recognition is a function of the similarity between known local features and their appearance in the input image. As such, the overall neural response pattern is consistent with that observed behaviorally in many different object recognition studies (e.g., Bülthoff & Edelman 1992, Tarr 1995), suggesting a neural basis for the similarity computations proposed by Poggio & Edelman (1990; also Riesenhuber & Poggio 1999).

Further evidence for population codes in the neural representation of objects comes from the work of Kobatake et al. (1998). As in the studies discussed above, a learning paradigm was used to examine how the selectivity of IT neurons in TE change with experience with novel objects. Monkeys were overtrained in a match-to-sample task with

28 simple two-dimensional shapes. Following this training, Kobatake et al. (1998) recorded the responses of IT neurons in five anesthetized monkeys: two trained with the shapes and three controls that had not seen the shapes. A total of 131 neurons were recorded from in the trained animals, while a total of 130 neurons were recorded from in the controls. Not surprisingly, Kobatake et al. (1998) found significantly more neurons selective for the familiar test shapes in the trained group as compared to the control group (although there were some neurons selective for the test shapes in this group as well). More interestingly, they found that the similarity distances between the shape-selective neurons—computed by placing the neural pattern best selective for each trained shape in a location in a high-dimensional response space equal to the total number of recorded neurons—were larger for the trained group as compared to the control. Such results provide additional constraints on how objects come to be coded in IT cortex: With experience, a larger population of neurons is likely to be selective for the familiar objects (a point we return to in the discussion of neuroimaging below), and the responses of these neurons are likely to become more selective (corresponding to larger distances in neural response-defined feature space).

A somewhat different approach to how neurons come to represent complex stimuli was developed by Sigala (2004, Sigala & Logothetis 2002). They trained monkeys to recognize diagrammatic drawings of faces or simple fish composed of separable features that could be systematically altered. For example, it was possible to change the eye height, eye separation, nose length, and mouth height of the faces. Likewise, the fish also had four changeable parameters: the shape of the dorsal fin, the tail, the ventral fins, and the mouth. For both sets, each parameter could independently have three different values. For this discrimination, two of the four parameters were task relevant (the remaining two parameters varied randomly), and maximal performance required attending to the values of both relevant parameters simultaneously. Monkeys were trained to perform this classification task with 10 objects per stimulus set (fish or faces) and then were given the task of classifying 24 new objects, which were also instances of the two trained classes. Following this training regimen, Sigala recorded from anterior IT neurons while the monkeys performed the same categorization task. At issue is whether neural selectivity occurs at the level of the class, the object, or diagnostic features. Sigala (2004) found that a large number of the class-selective neurons responded differentially to at least one parameter value for both faces (45.8%) and fish (43.1%). Of those cells, a large percentage were selective only for those diagnostic parameters (i.e., not to class-irrelevant features; 72.7% for faces, 75% for the fish). Thus, although it is tempting to associate highly selective neural responses with the representation of complete objects, these data suggest that object-selective neurons in IT may be coding for particular task-relevant diagnostic features rather than for objects as a whole.

Some of the most striking evidence with regard to how object features are represented in IT comes from the work of Tanaka (1996). Recording within the TE region of IT, Tanaka uncovered a highly structured pattern of feature selectivity. These results were built on a novel method for determining the best stimulus for a given neuron. Tanaka used an image-reduction technique on a neuron-by-neuron basis to determine which combination of minimal image features maintains the same firing rate as that measured in response to a complex object image for which a given neuron is preferential. These single-neuron object preferences were established by rapidly presenting a large number of common objects to anesthetized monkeys. Once an object was found that produced a strong response in the neuron being recorded, features in the image of this object were systematically manipulated and reduced in a direction that continued to drive the neuron to the same degree as the

original image. However, the particular features identified by these methods were not what one might have expected: They tended to be moderately complex two-dimensional shapes depicting lollipops, stars, etc. That is, despite a high degree of organization in area TE, the particular neural code being used to represent objects is not at all obvious. Indeed, Tanaka's (1996) results have proven to be something of a Rorschach test for the field—almost everyone can find something here that seems consistent with their pet theory.

Tanaka also found that TE neurons are neuroanatomically organized according to this feature selectivity. That is, adjacent neurons tended to be selective for the same minimal features. In addition, recordings along electrode penetrations perpendicular to the surface of the cortex revealed neurons that tended to exhibit a maximal response to visually similar objects. For example, within a perpendicular penetration, neurons might respond to small variations in the frontal view of a face, suggesting that area TE is also organized into minicolumns. Tanaka (1996) also used a technique known as optical imaging, which measures the presence of deoxidized hemoglobin, to examine these findings at a more macro scale. Again using faces as the example, he found that adjacent areas (corresponding to several minicolumns) all responded to faces, but at slightly different views. That is, one region might have its highest activation for a frontal view, while adjacent areas might have their highest activations to rotations varying from the frontal view in a systematic manner. Such findings also help to reinforce the hypotheses of Logothetis & Pauls (1995) and Booth & Rolls (1998) regarding the use of view-tuned neurons to achieve viewpoint invariance.

Interestingly, Tanaka himself has argued that these feature columns provide the visual object recognition system with two of the desiderata identified by Marr & Nishihara (1978): sufficient sensitivity to support discrimination among very similar objects, but sufficient stability to generalize across changes in viewpoint, lighting, and size. Tanaka suggests that objects represented by a large population of neuroanatomically adjacent neurons with overlapping selectivities are best able to realize these properties.

Finally, a recent study has gained notoriety for revealing "Jennifer Aniston" cells in the human brain (Quiroga et al. 2005). The nominal question addressed here is how neuron activity is organized into visual representations. The remarkable specificity of the neurons recorded in humans with epilepsy—some neurons responded to Jennifer Aniston alone, but not to Jennifer Aniston with Brad Pitt—seems to hark back to grandmother cells. Yet, it seems clear that if we know something, it must be encoded somewhere in the brain. Thus, it should not be surprising that knowing about Jennifer Aniston (apparently the subjects read *People*) is reflected in neural responses. Moreover, the actual locations of these responses were all in the medial temporal lobe, in regions typically considered to be nonvisual and often implicated in generic memorial processes. Similarly, strong neural responses were prompted by both visual and verbal stimuli. Thus, although it is nice to know that semantic knowledge about things is indeed represented independent of modality, this does not speak to how invariance is achieved within a modality—that is, how the visual system compensates for the infinite variations in size, viewpoint, lighting, and configuration that we encounter every day.

CIRCA 1890/1990—NEUROPSYCHOLOGY REDUX

Neuropsychology—the study of human subjects with brain lesions—is one of the oldest sources of evidence regarding the mechanisms underlying visual object recognition (e.g., Lissauer 1890). Although there has been a steady stream of neuropsychological research over the past century (e.g., Warrington & James 1967), arguably, it was the publication of Farah's (1990) monograph, *Visual*

Visual agnosia: impaired visual recognition abilities, with a preservation of low-level visual functions

Prosopagnosia: impaired visual face recognition, with the relative preservation of visual recognition of other object categories

Double dissociation: a neuropsychological framework in which one type of brain injury impairs Function 1 while sparing Function 2, while a second type of brain injury impairs Function 2 while sparing Function 1

Agnosia, that rekindled interest in the use of neuropsychological case studies. Specifically, she reviewed almost every published case study of visual agnosia—disorders of visual recognition—to gain a clearer picture of what they tell us, in toto, about the process of object recognition in unimpaired subjects. Broadly speaking, Farah's argument is that there exist two independent recognition systems: one that is part-based and one that is holistic. Put another way, one route to recognition is to parse an object into its constituent features or parts and then to use these parts—possibly in a configural manner (which would make such part-based representations truly structural; see Barenholtz & Tarr 2006)—as the match to similarly parsed input images. The other route assumes no such parsing: Objects are represented as undifferentiated images, and recognition proceeds by matching these holistic representations to input images.

Note that it may seem logical to associate the former with structural-description models (e.g., Biederman 1985) and the latter with view-based models (e.g., Poggio & Edelman 1990, Tarr & Pinker 1989). This would be wrong in that the question of view specificity is orthogonal to whether object representations are structural or not (Barenholtz & Tarr 2006). What Farah proposes speaks only to the structural nature of object recognition, her conclusion being that structure is used for some, but not all, recognition tasks. Farah (1990, 1992) proposed this division of labor to account for three major types of neuropsychological deficits that emerged from her review: (*a*) prosopagnosia, a profound deficit in recognizing faces; (*b*) object agnosia, a deficit in recognizing at least some classes of objects; and (*c*) alexia, a deficit in reading written text. She suggests that all three deficits can be explained in the context of damage to these two basic systems. For example, damage to the holistic system is likely to severely impact face recognition (to the extent we believe face recognition is holistic, a hypothesis supported by results such as Tanaka & Farah 1993). Conversely, damage to the part-based system is likely to severely impact text reading (to the extent we believe that reading requires structural representations; a hypothesis supported by results such as those of Pelli et al. 2003). Somewhere in between, damage to either system is also likely to impair the recognition of some objects, depending on the degree to which the recognition of a particular class relies more on structural or holistic mechanisms. And perhaps obviously, damage to both systems is likely to impair the recognition of most objects, including faces and words.

The critical neuropsychological prediction of Farah's model concerns the pattern of sparing and loss seen across different cases of agnosia. Specifically, she predicts that both prosopagnosia and alexia will occur with only limited object agnosia. However, there should never be cases of prosopagnosia and alexia with no object agnosia. Across her exhaustive review of the literature, she finds cases of impaired face recognition, impaired reading, and impaired object recognition, the latter often in combination with one of the former, but she finds not a single clear case in which face recognition and reading are impaired but object recognition is spared. Farah's meta-analysis of neuropsychological data provides powerful evidence for a two-system account of object recognition.

Although Farah did not explicitly make the distinction, it is tempting to associate the structural mechanisms with normal object recognition and holistic mechanisms with face recognition. Indeed, much of the field has adopted this particular dichotomy. Although there are certainly apparent dissociations in face and object recognition in both psychophysics (but see Gauthier & Tarr 1997) and neuroimaging (the focus of the next section), perhaps the most compelling piece of evidence for separable mechanisms comes from a single neuropsychological case. The Holy Grail in neuropsychology is a double dissociation, in which two different cases show opposite patterns of sparing and loss across two abilities (Plaut 1995, however,

presents a cogent computational argument as to why double dissociations do not necessarily entail separable mechanisms). For face and object recognition, this would manifest as cases in which face recognition abilities are lost, but object recognition abilities are intact, and cases in which face recognition abilities are intact, but object recognition abilities are severely impaired. Cases of the former—prosopagnosia—are quite prevalent in the literature. At the same time, cases of the latter—intact face recognition with severe agnosia—are rare. In fact, there is only one such published case, that of CK (Moscovitch et al. 1997).

Strikingly, CK appears to have acute object agnosia and alexia, but no deficit in face recognition. Moscovitch et al. found that CK's face recognition abilities for upright faces were comparable to that of unimpaired control subjects. This ability extended to Arcimbaldo's highly schematic faces composed of vegetables or two-tone Mooney faces. At the same time, CK was functionally blind at object recognition in that he was unable to perform even basic-level recognition tasks (although he was able to use context to infer the identity of some objects). Thus, CK appears to be the other end of the double dissociation with prosopagnosia. Moscovitch et al. (1997) suggest that this pattern strongly supports separable mechanisms for face and object recognition. It should be noted, however, that CK's face recognition performance is hardly normal. For example, his recognition of inverted faces was significantly worse as compared to control subjects. One interpretation of this result is that CK's recognition of inverted faces defaults to the standard object recognition process (e.g., Farah et al. 1995), for which he is severely impaired. An alternative is that all normal routes to object recognition are impaired, but that CK has developed an idiosyncratic template-like strategy for recognizing upright faces only (other idiosyncratic but successful recognition strategies have been observed in prosopagnosia; e.g., Bukach et al. 2006).

Moscovitch et al. (1997) also investigated whether CK was capable of recognizing overlapping line drawings of faces or objects, as well as two-tone Mooney faces or objects (Moore & Cavanagh 1998), all tasks that presumably require part segmentation for successful object recognition. Not surprisingly, given his severe object recognition deficit, CK was impaired at recognizing either overlapping objects or Mooney objects. However, CK performed at the same level as control subjects when recognizing overlapping faces or Mooney faces. Such findings indicate that CK's impairment is not in the preprocessing components of object recognition, for instance, segmentation of objects into parts, but rather is located downstream, perhaps at the point at which parts are related to one another to form structural representations. This claim is something of a paradox in that most holistic models assume object features are assembled into higher-order configurations (e.g., Maurer et al. 2002); that is, they are not template models in the strict sense. Yet, configural processing is often cited as a core property of face recognition mechanisms. Thus, it is difficult to see how CK could be impaired at "putting things together" to form configural representations, yet could be good at face recognition. As mentioned above, one wonders whether CK has actually lost most of his visual object recognition abilities, but has somehow retained a much more literal template-like process in which undifferentiated representations are matched to face images (possibly as a residual effect of premorbid face expertise).

fMRI: functional magnetic resonance imaging

CIRCA 2000—THE RISE OF A NEW MACHINE

If any single device since the tachistoscope could be said to have revolutionized the study of the mind and brain, it is the functional magnetic resonance imaging (fMRI) scanner.[5]

[5] As experimental devices, computers have mostly been used as fancy tachistoscopes. On the other hand, the idea of a

BOLD: blood oxygen level-dependent

FFA: fusiform face area

Neuroimaging studies have the potential to bridge the gap between human psychophysics and single neuron physiology by allowing us to measure the neural activity at the scale of hundreds of thousands of neurons concurrently with real-time task performance.

As with neurophysiology, before we discuss specific results from fMRI, it is worth stepping back and considering some of the assumptions that make it a viable tool for the study of visual object recognition (and the mind and brain more generally). The dependent measure in fMRI—the blood oxygen level-dependent (BOLD) effect—provides a neuroanatomical location where neural activity (or its signature) occurs. What can location tell us about the functional properties of the brain? Perhaps nothing. Again consider a wholly distributed neural code. Such a representational scheme would render fMRI useless: The neural patterns produced by the perception and recognition of different objects or object classes, at the scale of fMRI, would be indistinguishable from one another. Again, luckily for us, this sort of encoding does not seem to be the norm. As hinted at by single-unit neurophysiology studies, regions of IT appear to be organized into columns based on visual similarity—the defining characteristic of object classes. That is, hundreds of thousands of adjacent visual neurons tend to be selective for the same object or variations on that object. Thus, there is every reason to expect that fMRI should be able to resolve differential object processing across stimuli and tasks.

One of the most robust and oft-studied results regarding object processing to come out of the neuroimaging is that of category-selectivity. This phenomenon is roughly analogous to the object selectivity seen at the single-unit level: as revealed by fMRI, clusters of hundreds of thousands to millions of adjacent neurons show a selectively higher level of activity in response to objects within a visually similar class. As with single-unit selectivity, the most well-known finding of this sort is with reference to faces. That is, if subjects view a series of faces and the resultant pattern of activity is compared to the pattern measured when viewing a series of nonface objects, one finds higher activity in a small region of the fusiform gyrus located within the ventral-temporal pathway (actually first revealed using PET; Sergent et al. 1992; replicated by Puce et al. 1995, and later Kanwisher et al. 1997). This face-selective region—dubbed the fusiform face area or FFA—has attained a level of notoriety that is arguably far out of proportion relative to its potential to inform us about the nature of object recognition. That being said, there has been a great deal of interest in both uncovering the computational principles underlying the FFA and exploring whether it is truly exclusive to faces or can be recruited by nonface stimuli.

This latter question is central to the issue of modularity in visual processing (and in the brain more generally). That is, one model of how the brain works is that it is divided into functional modules, each specialized for a particular task (Fodor 1983). Although this is certainly true at some level of analysis—for instance, there are clearly separable areas for vision and language—the jury is still out on whether more specific processing mechanisms within such systems are domain-general or domain-specific. To some extent, fMRI plays into domain-specific arguments. A subject performs a particular task with particular stimuli, and a small area of the brain selectively "lights up" in response. It is certainly tempting to label each such spot as a module and give it a good acronym. Hence we hear about modules for all sorts of things, including, recently, love (Aron et al. 2005) (see **http://www.jsmf.org/badneuro/**). Thus, interpreting neuroimaging results requires some degree of caution—a putative module may be nothing more than an artifactual peak that is potentially neither necessary nor sufficient for the associated task.

computer certainly has had dramatic impact on how we think about the mind and brain (e.g., Turing 1950).

Returning to the question of category selectivity in IT, two chief issues have emerged. First, do category-selective regions occur for stimuli other than faces? And, if so, do they rely on the same neural substrates? Second, what are the origins of category selectivity; that is, why does it occur in the first place? Kanwisher (2000) has been the strongest proponent of the view that the FFA is a "face module" and, as such, is face exclusive. At one level, her arguments are weakened by the ever-growing list of putatively category-selective regions in IT: We have the parahippocampal place area (PPA; Epstein & Kanwisher 1998), the extrastriate body area (EBA; Downing et al. 2001), as well as areas for cars (Xu 2005), birds (Gauthier et al. 2000), chess boards (Righi & Tarr 2004), cats, bottles, scissors, shoes, chairs (Haxby et al. 2001), and novel objects called "Greebles" (Gauthier et al. 1999). Category selectivity ceases to have any explanatory power if a new functional module is postulated for each new object class that results in category selective activity (**Figure 1**).

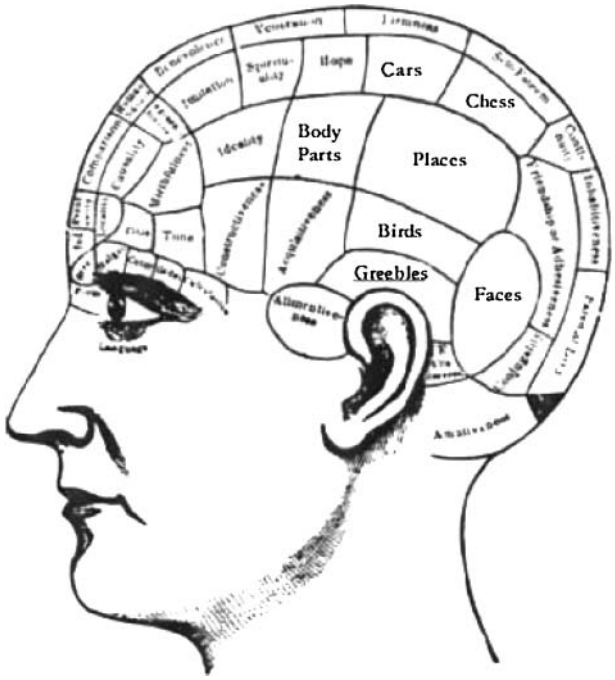

Figure 1

The new phrenology. fMRI reveals numerous category-selective regions in the brain. Simply labeling each as exclusively "for" a particular category provides no explanatory power regarding the functional properties of high-level vision. Please note that the depicted neuroanatomical locations in this figure are arbitrary, but it would not add much more information if they were correct! Adapted from the original drawing by Franz Joseph Gall.

Clearly, the answer to the first question raised above is yes—category selectivity occurs for many nonface objects. As to whether the same neural substrates used for faces are recruited in these cases, the answer is that it depends on whom you ask. There is certainly strong evidence that, up to the resolution of fMRI, the selective responses seen for Greebles (Gauthier et al. 1999), birds (Gauthier et al. 2000), and cars (Xu 2005) appear to colocate with the responses seen for faces. At the same time, it has been argued that better resolution might reveal separable neural populations for each object class, for example, columns similar to those found by Perrett and Tanaka. On the other hand, one might argue that colocalization at this scale is irrelevant. The functional architecture of the visual system is often thinly sliced with regard to selectivity—for example, the tuning to different orientations in adjacent regions of V1—yet the overall picture is one of a common computational basis (no one would reasonably argue that the different orientation columns in V1 were distinct modules). What is more telling is that the large majority of category-selective areas in IT seem to occur nearly on top of one another or quite close to one another. Thus, the best current evidence suggests that they form a single functional subsystem within object processing.

A more informative question is why this functional architecture occurs at all. Efforts to simply map out category selectivity entirely miss this point. In contrast, there has been significant effort by Gauthier and her colleagues to better understand the specific factors that contribute to the origins of this phenomenon. In particular, they have found that the concatenation of a visually homogeneous object class with the learned ability to quickly

Category selectivity: a phenomenon in neurophysiology and neuroimaging where a single neuron or small brain region shows higher responses for a particular object category relative to all other object categories

and accurately discriminate between members within this class—perceptual expertise—leads to a shift in the default level of processing and category-selective regions as measured by fMRI (Gauthier et al. 1999, 2000; see also Xu 2005). Critically, these same properties are true for face recognition in almost all humans: we are experts at recognizing individual faces and do so automatically when we encounter a face (Tanaka 2001). Thus, the same explanation applies to why we see category selectivity for faces (we all become face experts during development) and nonface objects (subjects were birdwatchers, car buffs, or were trained in the laboratory to individuate visually similar objects). As such, this account undermines any need to postulate category-exclusive modules.

Different computational principles have been used to account for other divisions of category selectivity. In particular, the explanation offered above would not seem to apply to the place area. Instead, Levy et al. (2001) have suggested that these separable functional regions arise due to the way in which we usually see faces (and objects) as compared to places/buildings. Faces are typically perceived centrally, that is, using one's fovea, whereas places are typically perceived peripherally. Such differences in the eccentricity of viewing are preserved throughout the visual system, so that object representations in IT come to be organized with respect to retinal eccentricity. Again, a putative modular difference based on category (faces versus places) can be accounted for by functional properties arising from the way in which a given object category is typically processed.

As alluded to at the beginning of this section, one issue in interpreting fMRI data is understanding how the measured activity of millions of neurons within a single voxel relates to single neuron responses. To explore this question, Sawamura et al. (2005) used an fMRI adaptation paradigm in both monkeys and humans to measure localized neural activity in the ventral pathway in response to changing object size. At issue was how the primate brain achieves invariance over image variation—in this case, over changes in the size of an image on the retina. In contrast to an earlier study that suggested minimal size dependency (Grill-Spector et al. 1999), Sawamura et al. found the greatest decrease in neural responses for repetitions of the same object at the same size, intermediate levels of response for repetitions of the same object at different sizes, and lowest response levels for repetition of different objects. Critically, such size dependency at the voxel level is consistent with single-neuron physiology that suggests changes in size result in systematic decreases in the response rate (Ashbridge et al. 2000). A similar correspondence across levels may be seen in the viewpoint-dependent patterns revealed by behavioral, neuroimaging, and neurophysiology studies reviewed above. Overall, such data suggest that it is possible to make connections between methodologies operating at different scales so long as similar methods and stimuli are used at all levels. Unfortunately, corresponding studies are still relatively rare, leaving open many questions regarding the nature of the neural code as understood at different levels of analysis.

Finally, perhaps mirroring errors of the field itself, in this section we have emphasized a particular subset of results at the expense of others. Although the interpretation of fMRI results remains problematic, the past decade has seen a wealth of new findings in this area, many of them speaking to issues that have been with us for far longer. For example, Bar et al. (2006) used fMRI (as well as magnetoencephalography) to study top-down connections within the visual recognition system—clearly an understudied problem relative to its probable importance in visual processing. Bar et al. propose that quite early during the recognition process the orbitofrontal cortex uses low spatial frequencies to determine the coarse shape of objects. This information is then fed back to the temporal cortex as a means for narrowing the candidate object search space. The fMRI results of Bar et al.

(2006) support the existence of this feedback pathway, and their magnetoencephalography results support the claim that these prefrontal areas actually respond more quickly than do early ventral areas during object processing. These results provide one possible mechanism for how the visual system achieves the incredible recognition speeds demonstrated by Thorpe et al. (1996).

CIRCA 2006—TOO MUCH DATA/TOO FEW MODELS

Our review touches upon only a small portion of the huge body of data that has been collected over the past two decades. This explosion of new data is due in large part to innovative methods—for example, computer graphics, psychophysics, ERP, and fMRI—that have invigorated the field, attracting new generations of scientists. We hope we have conveyed at least some of the significant theoretical and empirical progress that has been achieved. The huge successes of the Workshop on Object Perception and Memory (in its thirteenth year) and the Annual Meeting of the VisionSciences Society (in its sixth year) serve as barometers for how far the field has come.

At the same time, we should not get too self-congratulatory. One of the touchstones of the early 1980s was the integration of computational theory with empirical data (e.g., Marr 1982). For the most part, this promise has not come to pass. The wealth of data that has been accrued is almost overwhelming in its complexity, and there are few, if any, overarching models of the complete recognition process. Conversely, of the many computational theories that have been developed over the past two decades, few, if any, are strongly grounded in what we currently know about the functional and neural underpinnings of the primate visual system. Yet, there are reasons for optimism. In particular, within the empirical domain, methods such as fMRI have begun to forge tighter connections between single neurons and functional models of the mind. Similarly, computational theories have matured to the point where simple task completion (e.g., recognizing an object) is not a goal in itself; models have become more tied to neural architectures and to behaviorally based metrics of performance. Thus, we look forward to the next two decades and anticipate the new findings and integrative models that will bring us closer to the goal of explaining how we perceive and recognize objects.

SUMMARY POINTS

1. In the 1980s, structural-description models dominated the thinking about visual object recognition. These models proposed that visual recognition was based on "object-centered" three-dimensional parts. Consequently, recognition performance is predicted to be more or less independent of viewpoint.
2. In the 1990s, view-based models arose as an alternative to structural-description models. Such models propose that visual recognition relies on representations tied to an object's original viewing conditions. Critically, several behavioral studies found that visual recognition was strongly viewpoint dependent.
3. During this same period, neurophysiological researchers found that visual neurons are organized into columns that encode similar visual features. More recently, several studies have found that, with experience, neurons become progressively more selective for trained objects or their component features. However, individual neurons are not sufficient for discriminating between objects; rather, populations of neurons are used.

4. Neuropsychological research has found many cases of visual agnosia that appear to be face-specific. At the same time, there is one reported case of a general object agnosia with spared face recognition abilities. This double dissociation is intriguing regarding the independence of face and object recognition; however, more-detailed analyses suggest that such patterns may occur without separation of these processes.
5. fMRI has dramatically altered the landscape of how we study visual object recognition. Perhaps the most enduring finding from this new methodology has been that of category selectivity, that is, localized neural activity in response to a particular object category, for example, faces. Recent evidence suggests that this phenomenon is a consequence of the task requirements associated with face processing: expertise in individuating members of a homogeneous object category. Thus, similar category selectivity may be found for Greebles, cars, birds, etc.
6. For all of the recent progress in this area, we still have a long way to go. Behavioral and neuroscientific research must become better grounded in well-specified models of object recognition (as opposed to diagrams of boxes with arrows), and computational models must become better grounded in the rich set of constantly growing empirical data.

ACKNOWLEDGMENTS

The authors were supported by a grant from the James S. McDonnell Foundation to the Perceptual Expertise Network, and by National Science Foundation Award #0339122

LITERATURE CITED

Aron A, Fisher H, Mashek DJ, Strong G, Li H, Brown LL. 2005. Reward, motivation, and emotion systems associated with early-stage intense romantic love. *J. Neurophysiol.* 94(1):327–37

Ashbridge E, Perrett DI, Oram MW, Jellema T. 2000. Effect of image orientation and size on object recognition: responses of single units in the macaque monkey temporal cortex. *Cogn. Neuropsychol.* 17:13–34

Bar M, Kassam KS, Ghuman AS, Boshyan J, Schmid AM, et al. 2006. Top-down facilitation of visual recognition. *Proc. Nat. Acad. Sci. USA* 103:449–54

Barenholtz E, Tarr MJ. 2006. Reconsidering the role of structure in vision. *The Psychology of Learning and Motivation*, Vol. 47, ed. A Markman, B Ross B. In press

Biederman I. 1985. Human image understanding: recent research and a theory. *Comp. Vis. Graph. Imag. Process.* 32:29–73

A widely cited paper proposing a specific structural-description theory of human object recognition.

Biederman I, Cooper EE. 1991. Priming contour-deleted images: evidence for intermediate representations in visual object recognition. *Cogn. Psychol.* 23(3):393–419

Biederman I, Cooper EE. 1992. Size invariance in visual object priming. *J. Exp. Psychol.: Hum. Percept. Perform.* 18(1):121–33

Biederman I, Gerhardstein PC. 1993. Recognizing depth-rotated objects: evidence and conditions for three-dimensional viewpoint invariance. *J. Exp. Psychol.: Hum. Percept. Perform.* 19(6):1162–82

Booth MCA, Rolls ET. 1998. View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cereb. Cortex* 8(6):510–23

Bukach CM, Bub DN, Gauthier I, Tarr MJ. 2006. Perceptual expertise effects are NOT all or none: spacially limited perceptual expertise for faces in a case of prosopagnosia. *J. Cogn. Neurosci.* 18:48–63

Bülthoff HH, Edelman S. 1992. Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proc. Natl. Acad. Sci. USA* 89:60–64

Desimone R, Albright TD, Gross CG, Bruce C. 1984. Stimulus-selective properties of inferior temporal neurons in the macaque. *J. Neurosci.* 4:2051–62

Downing PE, Jiang Y, Shuman M, Kanwisher N. 2001. A cortical area selective for visual processing of the human body. *Science* 293:2470–73

Edelman S, Bülthoff HH. 1992. Orientation dependence in the recognition of familiar and novel views of three-dimensional objects. *Vis. Res.* 32(12):2385–400

Epstein R, Kanwisher N. 1998. A cortical representation of the local visual environment. *Nature* 392:598–601

Farah MJ. 1990. *Visual Agnosia: Disorders of Object Recognition and What They Tell Us About Normal Vision*. Cambridge, MA: MIT Press

Farah MJ. 1992. Is an object an object an object? Cognitive and neuropsychological investigations of domain-specificity in visual object recognition. *Curr. Dir. Psychol. Sci.* 1(5):164–69

Farah MJ, Tanaka JW, Drain HM. 1995. What causes the face inversion effect? *J. Exp. Psychol.: Hum. Percept. Perform.* 21(3):628–34

Fodor JA. 1975. *The Language of Thought*. Cambridge, MA: Harvard Univ. Press

Fodor JA. 1983. *The Modularity of Mind*. Cambridge, MA: MIT Press

Gauthier I, Hayward WG, Tarr MJ, Anderson AW, Skudlarski P, Gore JC. 2002. BOLD activity during mental rotation and viewpoint-dependent object recognition. *Neuron* 34(1):161–71

Gauthier I, Skudlarski P, Gore JC, Anderson AW. 2000. Expertise for cars and birds recruits brain areas involved in face recognition. *Nat. Neurosci.* 3(2):191–97

Gauthier I, Tarr MJ. 1997. Becoming a "Greeble" expert: exploring mechanisms for face recognition. *Vis. Res.* 37(12):1673–82

Gauthier I, Tarr MJ, Anderson AW, Skudlarski P, Gore JC. 1999. Activation of the middle fusiform "face area" increases with expertise in recognizing novel objects. *Nat. Neurosci.* 2(6):568–73

An innovative study demonstrating that category selectivity, as revealed by neuroimaging, emerges with expertise with nonface homogeneous object categories.

Grill-Spector K, Kushnir T, Edelman S, Avidan G, Itachak Y, Malach R. 1999. Differential processing of objects under various viewing conditions in the human lateral occipital complex. *Neuron* 24:187–203

Gross CG. 1994. How inferior temporal cortex became a visual area. *Cereb. Cortex* 4(5):455–69

Gross CG, Bender DB, Rocha-Miranda CE. 1969. Visual receptive fields of neurons in inferotemporal cortex of the monkey. *Science* 166:1303–6

The first paper to demonstrate that visual neurons can be selective for complex stimuli, such as hands.

Gross CG, Rocha-Miranda CE, Bender DB. 1972. Visual properties of neurons in inferotemporal cortex of the macaque. *J. Neurophysiol.* 35:96–111

Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P. 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* 293:2425–30

Jiang X, Rosen E, Zeffiro T, VanMeter J, Blanz V, Riesenhuber M. 2006. Evaluation of a shape-based model of human face discrimination using fMRI and behavioral techniques. *Neuron* 50:159–72

Jolicoeur P. 1985. The time to name disoriented natural objects. *Mem. Cogn.* 13:289–303

Jolicoeur P, Gluck M, Kosslyn SM. 1984. Pictures and names: making the connection. *Cogn. Psychol.* 16:243–75

Kanwisher N. 2000. Domain specificity in face perception. *Nat. Neurosci.* 3(8):759–63

An early neuroimaging paper that demonstrates category-selective neural responses for face stimuli.

Kanwisher N, McDermott J, Chun MM. 1997. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J. Neurosci.* 17(11):4302–11

Kobatake E, Wang G, Tanaka K. 1998. Effects of shape-discrimination training on the selectivity of inferotemporal cells in adult monkeys. *J. Neurophysiol.* 80(1):324–30

Konorski J. 1967. *Integrative Activity of the Brain: An Interdisciplinary Approach*. Chicago, IL: Univ. Chicago Press

Kosslyn SM. 1980. *Image and Mind*. Cambridge, MA: Harvard Univ. Press

Lawson R, Humphreys GW, Watson DG. 1994. Object recognition under sequential viewing conditions: evidence for viewpoint-specific recognition procedures. *Perception* 23(5):595–614

Levy I, Hasson U, Avidan G, Hendler T, Malach R. 2001. Center-periphery organization of human object areas. *Nat. Neurosci.* 4(5):533–39

Lissauer H. 1890. Ein fall von seelenblindheit nebst einem Beitrage zur Theori derselben. *Arch. Psychiatr. Nervenkr.* 21:222–70

Logothetis NK, Pauls J. 1995. Psychophysical and physiological evidence for viewer-centered object representations in the primate. *Cereb. Cortex* 5(3):270–88

Lowe DG. 1985. *Perceptual Organization and Visual Recognition*. Boston, MA: Kluwer Acad.

Marr D. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco: Freeman

A seminal computational paper on how visual objects might be represented using object-centered structural descriptions.

Marr D, Nishihara HK. 1978. Representation and recognition of the spatial organization of three-dimensional shapes. *Proc. R. Soc. Lond. B Biol. Sci.* 200:269–94

Maurer D, Le Grand R, Mondloch CJ. 2002. The many faces of configural processing. *Trends Cogn. Sci.* 6(6):255–60

Moore C, Cavanagh P. 1998. Recovery of 3D volume from 2-tone images of novel objects. *Cognition* 67(1–2):45–71

Moscovitch M, Winocur G, Behrmann M. 1997. What is special about face recognition? Nineteen experiments on a person with visual object agnosia and dyslexia but normal face recognition. *J. Cogn. Neurosci.* 9:555–604

Pelli DG, Farell B, Moore DC. 2003. The remarkable inefficiency of word recognition. *Nature* 423(6941):752–56

Perrett DI, Oram MW, Ashbridge E. 1998. Evidence accumulation in cell populations responsive to faces: an account of generalisation of recognition without mental transformations. *Cognition* 67(1–2):111–45

Perrett DI, Smith PAJ, Potter DD, Mistlin AJ, Head AS, et al. 1984. Neurones responsive to faces in the temporal cortex: studies of functional organization, sensitivity to identity and relation to perception. *Hum. Neurobiol.* 3:197–208

Pinker S. 1984. Visual cognition: an introduction. *Cognition* 18:1–63

Plaut DC. 1995. Double dissociation without modularity: evidence from connectionist neuropsychology. *J. Clin. Exp. Neuropsychol.* 17:291–321

Poggio T, Edelman S. 1990. A network that learns to recognize three-dimensional objects. *Nature* 343:263–66

Puce A, Allison T, Gore JC, McCarthy G. 1995. Face-sensitive regions in human extrastriate cortex studied by functional MRI. *J. Neurophyschol.* 74:1192–99

Quiroga RQ, Reddy L, Kreiman G, Koch C, Fried I. 2005. Invariant visual representation by single neurons in the human brain. *Nature* 435:1102–7

Riesenhuber M, Poggio T. 1999. Hierarchical models of object recognition in cortex. *Nat. Neurosci.* 2(11):1019–25

Righi G, Tarr MJ. 2004. Are chess experts any different from face, bird, or Greeble experts? *J. Vis.* 4(8):504 (Abstr.)
Rock I. 1973. *Orientation and Form*. New York: Academic
Rodman HR. 1994. Development of inferior temporal cortex in the monkey. *Cereb. Cortex* 4(5):484–98
Rosch E, Mervis CB, Gray WD, Johnson DM, Boyes-Braem P. 1976. Basic objects in natural categories. *Cogn. Psychol.* 8:382–439
Rotshtein P, Henson RNA, Treves A, Driver J, Dolan RJ. 2005. Morphing Marilyn into Maggie dissociates physical and identity face representations in the brain. *Nat. Neurosci.* 8:107–13
Sawamura H, Georgieva S, Vogels R, Vanduffel W, Orban GA. 2005. Using functional magnetic resonance imaging to assess adaptation and size invariance of shape processing by humans and monkeys. *J. Neurosci.* 25:4294–306
Selfridge OG. 1959. *Pandemonium: a paradigm for learning*. Presented at Symp. Mechanisation Thought Proc., London
Sergent J, Ohta S, MacDonald B. 1992. Functional neuroanatomy of face and object processing: a positron emission tomography study. *Brain* 115:15–36
Shepard RN, Cooper LA. 1982. *Mental Images and Their Transformations*. Cambridge, MA: MIT Press
Sigala N. 2004. Visual categorization and the inferior temporal cortex. *Behav. Brain Res.* 149(1):1–7
Sigala N, Logothetis NK. 2002. Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature* 415:318–20
Tanaka JW. 2001. The entry point of face recognition: evidence for face expertise. *J. Exp. Psychol.: Gen.* 130(3):534–43
Tanaka JW, Farah MJ. 1993. Parts and wholes in face recognition. *Q. J. Exp. Psychol.* 46A:225–45
Tanaka K. 1996. Inferotemporal cortex and object vision. *Annu. Rev. Neurosci.* 19:109–39
Tarr MJ. 1995. Rotating objects to recognize them: a case study of the role of viewpoint dependency in the recognition of three-dimensional objects. *Psychon. Bull. Rev.* 2(1):55–82
Tarr MJ, Pinker S. 1989. Mental rotation and orientation-dependence in shape recognition. *Cogn. Psychol.* 21:233–82
Thorpe S, Fize D, Marlot C. 1996. Speed of processing in the human visual system. *Nature* 381:520–22
Turing AM. 1950. Computing machinery and intelligence. *Mind* 49:433–60
Ullman S. 1989. Aligning pictorial descriptions: an approach to object recognition. *Cognition* 32:193–254
Van Essen DC. 1985. Functional organization of primate visual cortex. *Cereb. Cortex* 3:259–329
Warrington EK, James M. 1967. An experimental investigation of facial recognition in patients with unilateral cerebral lesion. *Cortex* 3:317–26
Xu Y. 2005. Revisiting the role of the fusiform face area in visual expertise. *Cereb. Cortex* 15:1234–42
Zhang L, Cottrell GW. 2005. Holistic processing develops because it is good. In *Proc. 27th Annu. Cogn. Sci. Conf.*, ed. BG Bara, L Barsalou, M Bucciarelli, pp. 2428–33. Mahwah, NJ: Erlbaum

The first neuroimaging paper to demonstrate category-selective neural responses for face stimuli.

An empirical paper that provides evidence in support of view-based object representations.

RELATED RESOURCES

Web sites with relevant information include:
http://www.psy.vanderbilt.edu/faculty/gauthier/PEN/
http://www.tarrlab.org/
http://web.mit.edu/bcs/sinha/
http://web.mit.edu/bcs/nklab/expertise.shtml/
http://www.jsmf.org/badneuro/

Annual Review of Psychology

Volume 58, 2007

Contents

Attraction and Close Relationships

Small Groups

Personality Processes

Cross-Country or Regional Comparisons

Organizational Groups and Teams

Career Development and Counseling

Adjustment to Chronic Diseases and Terminal Illness

Research Methodology

Indexes

Errata

An online log of corrections to *Annual Review of Psychology* chapters (if any, 1997 to the present) may be found at http://psych.annualreviews.org/errata.shtml