The State of the Art in Visual Object Recognition

The human brain has the impressive capacity to rapidly, spontaneously, and effortlessly recognize a large variety of visual objects even under unusual visual conditions. This remarkable ability has been studied by scientists from a variety of disciplines. In this thesis, I seek to describe the state of the art in our understanding of visual object recognition. To do so, I review influential recent and historical findings from the disciplines of psychophysics, neurophysiology, neuropsychology, and neuroimaging.

Hinze Hogendoorn

Neuroscience & Cognition Master Thesis
August 11, 2006

Presented to:
F.A.J. Verstraten
T.A. Carlson

Contents

Introduction

The study of visual object recognition concerns itself with the impressive capacity of the brain to recognize three-dimensional objects solely on the basis of two-dimensional patterns of light striking the retinas. The human brain is able to rapidly, spontaneously, and effortlessly recognize a large variety of visual objects to any degree of specificity, even when they are encountered in unusual orientations or illumination conditions. More than a century ago, Helmholtz first pointed out the immensity of the task to convert incomplete visual information into a meaningful interpretation. How does the brain solve this formidable challenge? Over the last decades, this question has been approached by scientists from a wide array of disciplines, including cognitive psychology and psychophysics, neuroimaging, and computation and engineering.

It is the purpose of this thesis to give an impression of the state of the art in visual object recognition. First, I outline some of the conceptual questions about recognition and categorization that shape the way object recognition is studied. Subsequently, I discuss what can be learned from the literature in four significant contributors to the topic: psychophysics, neurophysiology, neuropsychology, and neuroimaging. In each section, I describe important empirical findings and influential theoretical publications. I highlight some of the questions that remain unresolved and the evidence put forward by different researchers to attempt to answer them. Finally, I summarize the goals and challenges of current and future research in each field.

Categorization

The visual world has a seemingly infinite number of stimuli that can be discriminated from one another to an arbitrary level of detail. An important conceptual question is how humans classify objects for recognition, and how recognition information is organized. In 1958, Brown observed that all members of society tend to refer to everyday objects such as pineapples and chairs by the same names. His conclusion was that “the most common name for each of these categorizes them as they need to be categorized for the community’s nonlinguistic purposes. The most common name is at the level of usual utility”. For example, when shown a photograph of an apple, most people will report “apple” rather than “fruit” or “granny smith”.

Rather than being based on utility, Rosch et al (1976) argues that visual stimuli are classified according to different perceivable cues, which form natural categories. The level of abstraction at which category cues have maximum predictive validity is the most natural level at which to recognize objects. Rosch and colleagues call this the ‘basic level’. For example, members of the basic-level category “dog” share a large number of visual features that are predictive for that category. General (or superordinate) categories, such as “mammal”, are so inclusive that its members share few perceivable characteristics. Conversely, specific (or subordinate) categories such as “Doberman” are so specific that its members share many perceivable characteristics with members of other categories.

Clearly, human beings group objects into basic-level categories according to natural groupings of attributes. Furthermore, evidence from various lines of reason suggests that these categorizations are perceptual in nature rather than cognitive (Logothetis & Sheinberg, 1996). Subjects performing recognition tasks with novel objects, for example, can form categories for those stimuli without linguistic labels. Similarly, young prelinguistic infants are able to make categorical representations, which are later enriched by the use of language (Piaget, 1969). Finally, many animals, including not just nonhuman primates but even relatively primitive animals such as pigeons and goldfish, have been found to be able to discriminate between basic-level categories. There is therefore considerable evidence supporting the notion that there are universal perceptual principles underlying the formation of perceptual categories.

Humans can recognize objects more rapidly and more accurately at the basic level than at other levels of abstraction (eg. Jolicoeur et al, 1984). Interestingly, categorization at basic level takes no longer than merely detecting the presence of an object (Grill-Spector & Kanwisher, 2005) – “as soon as you know it is there, you know what it is”. However, atypical exemplars that have significant shape deviations from the prototype of a class form an exception. Such objects (such as penguins), tend to be recognized more rapidly as a member of the subordinate class (i.e. “penguin”) rather than the basic level (“bird”). Jolicoeur et al (1984) coined the term ‘entry-level’ to describe the specific level at which objects are first accessed.

Interestingly, for observers with expertise in discriminating the subtle differences between objects in a certain category, the entry level of object recognition can shift from the basic level to subordinate categories. For example, Tanaka & Taylor (1991) reported that dog experts and bird experts spontaneously recognize objects from their respective areas of expertise at the subordinate level as quickly and accurately as is usual for basic-level classification.

In summary, objects can be identified at various levels in a taxonomical hierarchy. Object categories are most likely organized according to universal perceptual principles, and the level at which objects are most perceptually similar is known as the basic or entry level. With the exception of atypical exemplars, most people spontaneously report the basic level when identifying an object, and recognition occurs most rapidly at basic level. Expertise with a category can influence the entry level at which objects are recognized.

Psychophysics

As in many fields of vision, the most accessible approach to study object recognition is psychophysics. The most frequently used psychophysical tool to study object recognition is recognition performance. By manipulating aspects of the task and inspecting the effect of the manipulation on recognition performance, scientists hope to better understand the steps involved in the process of recognition. The ability of subjects to recognize objects is generally measured using either recognition accuracy or recognition latency. In the past a considerable number of possible object features have been manipulated, which can be divided into linear and nonlinear transformations.

When an image undergoes a linear transformation, the visual information available to an observer remains unchanged. Examples of such transformations are changes in size (disregarding limits of visual acuity), position, or orientation in the picture plane. Conversely, nonlinear transformations do result in changes in the visual information available to the observer, and therefore present a particular challenge to object recognition systems. The most-studied example of such a transformation is rotation in depth.

Linear Transformations

The effect of stimulus size on the latency with which subjects made ‘same object’ vs ‘different object’ decisions (disregarding size) was studied by Jolicoeur (1987) and Ellis et al (1989). Both papers report increasing recognition latencies with increasing size discrepancies. In a later study, Milliken and Jolicoeur (1992) manipulated viewing distance and found that the perceived size of the object, rather than its retinal size, determined the effect of size on recognition latency.

However, in a picture priming task in which previously presented objects were perceived more quickly and accurately than unfamiliar objects, no effect was found when the size of the prime was manipulated (Biederman & Cooper, 1992). Conversely, an explicit memory task using the same objects showed a clear effect of size. This dissociation suggests that the influence of size on object recognition is likely to be due to processes involving memory-based comparisons rather than the perception of the object’s shape. As such, size does not seem to be involved in visual object recognition per se.

It seems intuitively evident that we can visually recognize objects irrespective of where they are. However, empirical data has been strangely equivocal with respect to position dependence. Foster & Kahn (1985) report finding a detrimental effect on recognition accuracy when successive presentations of objects were in different positions. Conversely, Biederman & Cooper (1991) report finding that recognition latency was independent of translation. To further complicate the picture, Nazir & O’Regan (1991) found a detrimental effect of translation only in the first few trials for each object and each subject.

Despite this apparent controversy, translation invariance is an accepted feature incorporated into virtually all current models of object recognition (Riesenhuber & Poggio, 2000). Furthermore, both physiological studies in monkeys and human neuroimaging studies have found that brain regions involved in the final stages of object perception are completely non-retinotopic, lending further support to the idea that the ability of the brain to recognize objects is independent of position (Grill-Spector, 2003; Tanaka, 1997).

Similarly, studies investigating the impact of rotating stimuli in the picture plane away from the canonical orientation initially found only minimal effects of misorientation on either recognition latency or accuracy (Jolicoeur, 1985). Furthermore, even these small effects disappeared after practice. Experiments in monkeys show a similar pattern of results. Monkeys trained to recognize exemplars of various stimuli initially show orientation dependence, but over the course of a few test sessions quickly develop orientation-independent patterns of responses (Logothetis & Pauls, 1995).

However, Jolicoeur (1985) was criticized for using familiar stimuli, and using a set of stimuli deliberately designed to be unfamiliar and possess no diagnostic features, Tarr & Pinker (1989) showed that there were both latency and accuracy costs associated with recognizing stimuli rotated away from the training view. Furthermore, although continued practice resulted in an increase in recognition performance for the newly familiarized, misoriented view, this did not generalize to other, novel, misoriented views. Perhaps certain diagnostic features are necessary for rotation-invariant recognition to be possible, in the absence of which some alternative, rotation-dependent processing mechanism mediates recognition.

The special case of rotations where the stimulus is up side down is particularly interesting because, in humans, it seems to adversely affect recognition performance of only a few classes of stimuli, especially faces (Valentine, 1988). However, as pointed out Logothetis & Sheinberg (1996) it is debatable whether faces are special in and of themselves, or just the most common class of objects that are distinguished on the basis of subtle shape variations. Indeed, a study by Diamond & Carey (1986) using different dog breeds, which are also highly similar, familiar objects, also reported an effect of inversion. Since distinguishing members of familiar subclasses is a process that relies more on configurational cues than on featural cues, they argue that such within-class identification is differentially disrupted by inversion. It is interesting to note that inversion studies in primates (Bruce, 1982), who regularly encounter faces in non-upright positions, and human infants (Carey & Diamond, 1977), which presumably have not developed orientation-specific sensitivity, do not find any effects of inversion.

In summary, with a few exceptions, human object recognition performance is accepted to be largely invariant to linear transformations such as changes in size, position, or orientation.

Nonlinear Transformations: Image-Based vs Viewpoint-Invariant Models

The effect of nonlinear transformations on recognition performance is one of the central debates in the field of visual object recognition. The most hotly debated nonlinear transformation is rotation in depth. The ability to recognize three-dimensional objects that are rotated in depth is remarkable if one considers the fact that such rotations result in different sets of visible features, contours, geometry, and often illumination. Over the past decades, a number of theories have been developed to explain the stability of object recognition across such nonlinear transformations. These can be broadly grouped into one of two groups: image-based models and structural description models.

Image-based, or viewpoint-dependent, models hold that objects are represented in a viewer-centered frame of reference determined by the relative positions and orientations of the object relative to the observer. Theories of viewpoint-dependent recognition are motivated by the frequent finding that there is a monotonic decrease in recognition performance with increasing misorientation from a preferred view of an object when it is rotated in depth (eg. Jolicoeur, 1995; Tarr & Pinker, 1989). Objects have been proposed to be represented as single, canonical views, or as a collection of distinct views to which the observer has been exposed (Tarr & Pinker, 1989)

Recognizing an object from a viewpoint that differs from the stored representation or representations therefore requires some sort of transformation, either towards the single canonical view (in the former case) or towards the nearest stored view (in the latter). Initially, mental rotation was proposed as a possible mechanism by which misoriented stimuli were matched to canonical representations (Tarr & Pinker, 1989). Mental rotation tasks where subjects make same-different judgments about object rotated in depth show effects on accuracy and latency similar to recognition tasks with rotated stimuli. However, recent human neuroimaging studies (Gauthier et al, 2002) have provided compelling evidence that mental rotation and object recognition do not share a common substrate, and as such that mental rotation is unlikely to be the process mediating the recognition of objects rotated in depth.

The existence of viewpoint-tuned neurons in regions of the monkey cortex implicated in object recognition has been a mainstay of the argument for image-based models for many years. Although it seems likely that such neurons exist in humans, the resolution of current noninvasive neuroimaging techniques means that at present this question is best addressed indirectly. For example, in an innovative psychophysics study, Fang & He (2005) find support for the existence of viewpoint-dependent neurons in humans using viewpoint aftereffects in an adaptation paradigm. After visual adaptation to an object viewed from a small angle, the perceived viewing angle of a subsequent frontal view is biased towards the other direction. Fang and He found that this viewpoint aftereffect was category–specific; viewing a car obliquely caused a subsequent frontal view of another car to be biased, but not a frontal view of a face. This provides strong support for the existence of object-selective neurons tuned to certain specific viewing angles in the human brain.

The image-based approach is not without its weaknesses, however. A categorization strategy should clearly be sensitive to differences in appearance, in order to distinguish between visually similar objects. However, if representations are too specific, then separate representations will need to be stored for each small variation in the image if image-based information is not shared across viewing conditions. Considering the huge number of objects we are able to recognize across a range of viewing conditions, this would result in a combinatorial explosion of separate representations that would require unfeasible amounts of memory to store. This has been one of the main points of critique of the image-based approach.

An extension to the several-stored-views model has been proposed in which a new stimulus, rather than being transformed towards the nearest stored view, is recognized by an interpolation of stored views (Bülthoff & Edelman, 1992). Furthermore, such interpolation models have been shown to be computationally viable (Ullman, 1998) and neurally plausible (Perret et al, 1998). Such an approach seems especially attractive if one considers another type of nonlinear transformation: recognizing different members of the same object class. While objects within an object class (chairs, for example) share many defining features, they can nonetheless differ from one another in a multitude of ways without losing their “chair” identity. Interpolation models have been argued to be able to explain recognition of both objects rotated in depth and different exemplars within an object class (Tarr & Gauthier, 1998).

Conversely, structural description or viewpoint-invariant models hold that objects are represented in an object-centered frame of reference on the basis of the relations between their distinctive features, which remain constant under changes of viewpoint (Marr & Nishihara, 1978; Biederman, 1987; Biederman & Gerhardstein, 1993). Marr and Nishihara proposed that local image features such as edges and surfaces are combined in a bottom-up way into a viewer-centered, 2½-D ‘sketch’. This viewpoint-dependent representation is then converted to a 3D object representation, where the object’s features are defined in a reference frame centered on the object.

The notion that object representations should be object centered and might be represented as configurations of 3D parts or simple volumes (‘geons’) was further developed in Biederman’s (1987) influential Recognition-By-Components model. RBC specified a lexicon of allowable 3D volumes, the kind of qualitative spatial prepositions that could connect such volumes, and a list of 2D cues, that uniquely identified the volumes. The model was later refined by Biederman and Gerhardstein (1993), who added a number of criteria for viewpoint invariance, on the basis of which they attempted to dismiss a number of psychophysics studies that had shown viewpoint dependence. For example, they argue that if a change in viewpoint results in the occlusion of a part of the visible part structure, it is only logical that recognition will be impaired, since the viewpoint-invariant processes do not have access to the same information.

The 2D visual cues that were proposed to underlie the detection of these geons were termed non-accidental properties. Non-accidental properties are 2-D retinal patterns that are only rarely caused by the accidental alignment of the eye and the 3D object. For example, a straight line on the retina is assumed by the visual system to be a straight line in 3D, rather than a coincidental alignment of a curved line with the eye.

Such features tend to be especially informative of an object's shape, and are the underlying principle of many powerful optical illusions such as the Ames Room (*Figure 1*).

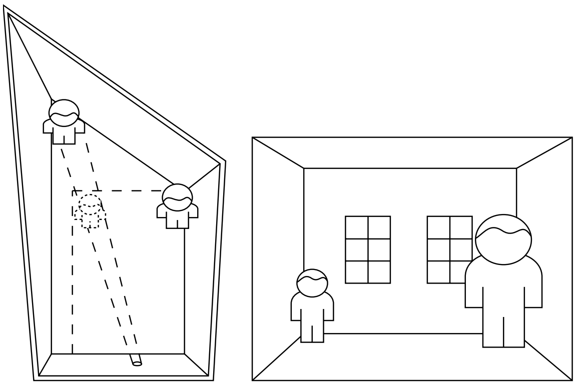

Figure 1: The Ames Room is a trapezoidal room specifically designed to appear rectangular from one perspective. As a result, a person standing in the far corner appears much smaller than a person in the near corner.

Biederman and Bar (1999) directly compared variations in such non-accidental properties to variations in metric properties, such as length and width. After appropriate calibration, they found that when objects were rotated in depth, objects that differed on non-accidental properties were much easier to distinguish than objects that differed on metric properties. Combined with the fact that non-accidental image properties are highly diagnostic 3D shape cues, it seems clear that non-accidental properties are especially important in the object recognition process.

A parsimonious explanation of object recognition will likely combine elements from both image-based and structural description models. Such an account would need to be able to explain the continuous spectrum of viewpoint dependence reported in the literature, which ranges from complete viewpoint-invariance (eg. Biederman & Gerhardstein, 1993) to extreme viewpoint-dependence (eg. Bülthoff & Edelmann, 1992), depending on experimental conditions and task demands. Another challenge facing the development of such a model is that it would need to consider that we are able to recognize objects at many levels of categorization, meaning that our stored object information has multiple levels of access. It would need to account for the effects of expertise on both recognition performance and the default level of access. Finally, it is instructive to realize that performance on any given recognition task is the result of a complex interaction of all these factors, and that a complete model should take into account the nature of these interactions. Indeed, more recent research has already started considering combinations of these factors rather than simply making dichotomous contrasts (Tarr & Bülthoff, 1998).

In short, empirical research using psychophysical paradigms has contributed significantly to our understanding of object recognition in the human brain. Recognition performance has been found to be independent of most linear transformations, indicating that object representations are stored in such a way that they are independent of position, size, and to a degree, rotation in the image plane. Changes in the retinal image associated with non-linear transformations such as rotations in depth are a great deal more complicated, and empirical studies have reported results ranging from minimal to extreme viewpoint-dependence. Two influential models have been proposed. The image-based approach holds that object representations are stored as one or a small number of discrete views of an object, and that a stimulus is recognized by transforming it and matching it to one of these stored views. Conversely, the structural description approach proposes that object representations are stored as object-centered 3D representations, and that a stimulus is broken down into its component parts, or geons, and converted into an object-centered frame of reference before it is compared to stored representations. More recently, researchers have begun to combine elements from these two approaches.

Neurophysiology

The anatomy of the visual system has been studied for decades, even centuries. The physiology of the early parts of the visual system, such as the eyes and sub cortical nuclei, has been studied using simple dissection and in vivo techniques in primitive animals such as frogs. However, studying the neurophysiology of higher visual functions is much more difficult in such primitive animals because it is difficult to conceptualize and measure their mental state. For this reason, nonhuman primates such as macaque monkeys have been extensively used to study aspects of visual cognition such as object recognition.

More than 2 decades ago Ungerleider and Mishkin (1982) first reported the existence of separate processing streams in the macaque visual pathway. They identified a ventral stream leading from primary visual areas to inferotemporal (IT) cortex and a dorsal stream leading from primary visual areas to dorsal parietal areas. The former, which they dubbed the ‘what’-pathway, was involved in object vision, enabling the monkey to identify objects, whereas the latter, the ‘where’-pathway, was implicated in spatial vision and was important for localizing objects in space.

More recently, the two streams have been suggested to differ on the output systems that each stream serves, rather than on the nature of the information available to them (Goodale & Milner, 1992). According to Goodale and Milner, a perceptual representation of the world and the objects in it is formed by the ventral processing stream, while the dorsal processing stream mediates visual control of motor actions towards those objects. Essentially, the ventral stream identifies an object, and the dorsal stream provides information critical to interact with it, such as shape, size, and location. Although the two systems have access to the same kind of visual information, the output of each stream is used for different purposes.

The Dorsal Stream

The dorsal visual pathway in monkeys projects from V1 through V3 and V3A to areas MT and MST and finally to the posterior parietal lobe. The activity of cells in the dorsal pathway is dependent both on certain properties of the stimulus and on the concurrent behavior of the animal (Andersen, 1987). Many cells in the posterior parietal cortex are gaze-dependent, their activity being determined by where the animal is looking. Neurons in the dorsal stream, particularly areas MT and MST, are highly sensitive to motion, including rotation and motion in depth, but are insensitive to features such as color and texture. Additionally, dorsal stream neurons seem to be tuned for form and shape properties only insofar as these are necessary for motor interaction (Sakata & Taira, 1994). Intestingly, and in contrast to many motion-sensitive neurons in temporal visual areas, motion-sensitive neurons in the dorsal stream respond to self-produced motion, and as such are well-suited to the visual monitoring of limb position (Perret et al, 1990).

The dorsal parietal lobe has projections onto pre-motor areas of the frontal cortex implicated in ocular control, reaching motions of the limb, and grasping actions of the hand and fingers (Gentilucci & Rizzolatti, 1990). As such, the posterior parietal area is strategically situated to mediate the visual control of motor actions. Indeed, monkeys with damage to the posterior parietal lobe show impairments in the visual guidance of their motor behavior, such as deficits in reaching for objects or failing to shape or orient their hands when attempting to retrieve food (Goodale & Milner, 1992).

It is outside the scope of this essay to review in-depth the literature on dorsal stream visual processing in monkeys, especially considering the general consensus that the object recognition process seems largely the domain of temporal areas. I refer the interested reader to Milner & Goodale (1995) for more information on the monkey literature.

The Ventral Stream

The ventral visual pathway in monkeys goes from V1 through V2, V4, and posterior inferotemporal (PIT) cortex to IT. Although projections exist that skip steps in the pathway, the step-by-step projections predominate. Lesions along the pathway to IT result in the same deficits in object recognitions as lesions to IT itself. The clear retinotopic organization of the primary visual cortex gets gradually coarser further downstream, and no retinotopic organization has been found in area IT (Tanaka, 1997).

Since the discovery of cells in the macaque IT selective to specific features (particularly faces), the representation of features in the ventral visual pathway has seen a substantial amount of interest. According to Tanaka (1996), cells in the monkey IT are selective to features that are considerably more complex than features such as orientation, size, color, and texture, some of which are processed as early as V1. However, they are not sufficiently complex to uniquely represent entire natural objects, and in many cases a stimulus that results in maximum activation for a particular cell can be simplified to some degree without any decrease in activation. The only reported exceptions are faces: for many face-selective cells, the image of the face cannot be simplified very much without losing activation.

Electrophysiology experiments have shown that the monkey IT cortex is organized into columns of cells selective for similar stimulus features. Recordings from electrodes in long vertical penetrations into the cortex demonstrate very similar stimulus selectivity. Conversely, measurements from oblique penetrations crossing into nearby columns indicate that cells with similar stimulus selectivity are only evident within a mean range of about 400 microns (Fujita et al, 1992).

Wang et al (1996) used optical imaging to further investigate the spatial distribution of the columnar organization. In this technique, the cortical surface of the brain is exposed and illuminated with red light. Because the optical properties of hemoglobin vary with its oxygenation state, it is possible to measure the metabolic activity at high resolution. After determining the critical feature of a single cell using a microelectrode, they found activation spots surrounding that cell using optical imaging when presenting the monkey with that feature. Since the activity of a single cell is unlikely to cause measurable metabolic fluctuations, this localized activation confirms the regional clustering of cells tuned to similar stimulus features.

Another striking finding reported in the same paper was the overlap of IT cells tuned to different but related features. After finding both cells tuned to frontal and profile views of faces in the same microelectrode penetration, Wang and colleagues presented the monkey with pictures of a doll's face from five different angles. They found that the activation spots caused by all the stimuli overlapped, but their centers moved around as the face rotated from left profile to frontal to right profile. Interestingly, they reported no similar effect when rotating the doll's head around a different axis or when changing its facial expression. Apparently, either only horizontal rotations (arguably the most commonly experienced) are explicitly mapped on the cortical surface, or rotations around other axes are represented elsewhere in the cortex. This is a particularly notable finding in light of the previously discussed adaptation study by Fang & He (2005), which reported psychophysical evidence in humans for object-selective cells tuned to different viewpoints on the horizontal plane. They did not examine rotations around other axes; it would interesting to see whether object-selective neurons exist in the human visual system that are tuned to unconventional orientations.

The existence of such viewpoint-dependent cells is one of the important arguments supporting the existence of viewpoint-specific representations. However, the fact that the activation spots overlap might indicate that viewpoint-dependent cells together form a viewpoint-independent representation. Furthermore, to the best of my knowledge, no similar continuous maps of non-face features have been reported, indicating that perhaps the processing of faces enjoys special status in the cortex, or that perhaps the viewpoints of other objects are mapped as overlapping cell populations rather than as the principal component in the columnar organization.

Although it is powerful evidence that the presentation of certain stimuli can drive the activation of particular cells, it is important to realize that this does not in itself say anything about the actual percept experienced by the monkey. Fortunately, there is good agreement between the activity of IT cells and the monkey's conscious percept. A number of studies have shown that the activity of IT cells tuned to specific shapes is invariant over changes in size, location, and even the visual cue (luminance, texture, relative motion) used to define the shape, all of which drastically change the low-level image properties while leaving the shape percept intact (Tanaka, 1997).

When stimuli are partially occluded by a foreground figure, which in humans in many cases does not impair shape recognition, IT cells similarly were found to maintain their selectivity for shapes (Kovacs et al, 1997). A study using two-tone images found similar correlations between the activity of IT cells and the conscious percept. Two-tone image depict objects illuminated to cast strong shadows, and are difficult to recognize if shadow contours are continuous with the outer contours of the object. After being shown a grayscale version of the same image, however, it becomes easy to ‘see’ the same figure in the previously ambiguous two-tone image. Similarly, IT cells that do not respond during the initial presentation of a two-tone image do respond to a second presentation after the image has been disambiguated by the presentation of a grayscale image (Tovee et al, 1996).

Another technique to disambiguate conscious perception from neural activity as a result of visual stimulation is binocular rivalry. When different stimuli are presented to the two eyes, the conscious percept is alternatingly dominated by one of the two stimuli. By presenting a trained monkey with two stimuli simultaneously, one of which is chosen to evoke a strong response from a measured cell, the activity of inferotemporal cells can be compared to the conscious percept as reported by the monkey (Leopold & Logothetis, 1996). Whereas nearly all inferotemporal cells responded consistently with the reported percept, cells further up the processing stream were less consistent, with 25% of V4 cells and only 18% of V1 cells correlating with the monkey’s report (Sheinberg & Logothetis, 1997).

While compelling, the results from all these studies is correlational in nature. Until recently, no study has been able to report a causal link between monkey IT cortex and conscious perception. In a ground-breaking study, Afraz et al (2006) report the first causal link between activity in face-selective neuron populations in monkey IT and the conscious perception of a face. Using two monkeys trained to report seeing faces, Afraz et al used microelectrodes to stimulate neural populations that responded selectively to faces. They found that, within a certain time window, stimulation of these cells made the monkey more likely to report seeing a face in a noise stimulus.

Convergent evidence from these different paradigms therefore makes a good case for a strong link between the activity of single cells in the monkey inferotemporal cortex and the overall percept, and accordingly, the involvement of monkey IT in conscious object recognition.

In summary, the visual system in monkey cortex is functionally divided into a dorsal stream, which projects from V1 to dorsal parietal areas and is implicated in the guidance of motor actions, and a ventral stream, which projects from V1 to temporal areas and is involved in representing and recognizing objects. Evidence from different experimental paradigms provides strong evidence that ventral stream neurons are correlated to the conscious percept. [needs more about object recog]

Neuropsychology

The study of patients with brain damage and the ensuing cognitive deficits has been an important contributor to our understanding of the human brain for several decades. The neuropsychological study of such patients has facilitated the integration of various neuroscientific disciplines. For example, patients such as DF whose impairments correspond to the deficits caused by experimental lesions in monkeys are informative about the degree of evolutionary conservation from the monkey brain to the human brain. Psychophysical experiments with brain-damaged patients can bring to light deficits that might otherwise not be evident, and neuroimaging studies can elucidate the functions associated with the damaged brain regions in normal subjects.

Care must always be taken generalizing findings from studies with monkeys to humans. However, the functional map of brain areas involved in vision is relatively conserved from the monkey to the human. Furthermore, a number of human neuropsychological cases in the past years have provided support for a functional distinction between dorsal and ventral stream processing in humans comparable to the one reported in monkeys. Damage to the dorsal parietal lobes can cause a deficit called optic ataxia, in which patients are impaired on their ability to visually guide their motor actions. Such patients are usually able to guide their actions normally using other sensory cues such as proprioception, and are similarly unimpaired on their ability to describe objects in the visual world. Since both the visual and motor systems are otherwise intact, the deficit is visuomotor rather than uniquely visual or uniquely motor.

A complementary neuropsychological deficit is visual form agnosia, the most famous of which is probably patient DF (Goodale & Humphreys, 1998). Carbon monoxide poisoning left this patient with a lesion centered on the ventrolateral area of the occipital cortex and a profound impairment of her ability to recognize objects. Although unable to visually distinguish even simple geometric shapes such as circles and triangles, she is completely unimpaired on her ability to visually guide reaching and grasping. When presented with blocks of various sizes, she is unable to report the size of the blocks by indicating the width between thumb and forefinger, yet when instructed to pick one up, she changes the aperture between thumb and forefinger as appropriate for the size of the block. It seems that the information about object properties necessary for motor guidance is still available to the appropriate motor regions in the cortex, but no longer available to mediate recognition or verbal report.

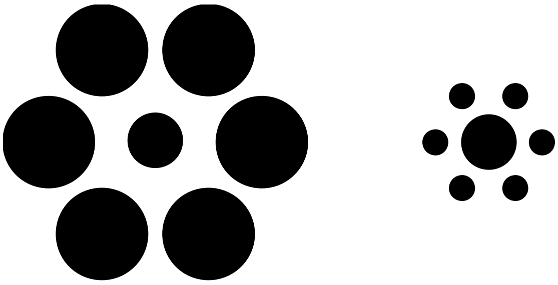

Figure 2: In the Ebbinghaus illusion, the presence of large and small flankers makes identical discs seem to differ in size.

Further support for the existence of separate recognition and visuomotor pathways comes from psychophysical studies using optical illusions such as the Ebbinghaus Illusion (*Figure 2).* In this illusion, two circles of identical size are surrounded by a ring of either small or large circles. The circle surrounded by small circles appears larger than the circle surrounded by large circles, even though they are of equal size. Interestingly, Aglioti et al (1995), using a 3D version of this illusion with discs in the place of circles, showed that although the illusion affects the judgment of which central disc is larger, it does not affect the aperture of the thumb and forefinger when subjects are instructed to pick it up. Apparently, although the metric information in the perceptual experience is distorted by the proximity of other objects, the metric information available to guide motor action remains intact.

Taken together, a double dissociation emerges from neuropsychological studies between the ability to recognize objects and the ability to guide actions towards them, which parallels findings from electrophysiological studies in monkeys.

The study of patients with brain damage has also provided insight into the neural basis of the object categorization taxonomy. A number of patients with very explicit category-specific deficits in visual recognition have been reported in the literature. Prosopagnosia is one such deficit, in which the patient is impaired in the ability to recognize the faces of familiar or famous people, but is able to recognize other objects normally. Prosopagnosics can sometimes recognize people by their voice or their gait, and can recognize faces as being faces as well as being able to point to facial features, but are incapable of identifying individual faces.

Although prosopagnosia at first seemed an impairment specific only to faces, more subtle recognition deficits have since been reported, indicating a more general impairment in object recognition than was initially surmised. A pattern emerged showing impairments in recognizing individuals within object categories. Patients have been reported that are unable to distinguish housefronts or cars, a birdwatcher unable to visually distinguish birds, and another patient unable to discriminate between his cows (Damasio et al, 1982). After reviewing such cases, Damasio and colleagues proposed that rather than being a specific impairment of the visual processing of faces, prosopagnosia involved a deficit in the ability to make within-category discriminations without affecting their ability to recognize to which class a stimulus belongs.

The already difficult task of recognizing human faces is complicated further by the fact that human faces are expressive and have to be recognized despite the nonrigid deformations that characterize such changes in expression. Facial expressions and the emotional content evident from such configurational changes are clearly informative and can be extremely important in social interactions. Neuropsychological evidence suggests that the processing of facial identity is at least partly distinct from the processing of emotional content. A number of prosopagnosics, while impaired at identifying individual faces, can recognize emotions from facial expressions (Tranel et al, 1974). Conversely, Adolphs and colleagues (1994) report a patient who suffered a near-complete bilateral destruction of the amygdala as a result of Urbach-Wiethe disease. Although this patient’s ability to recognize faces of individuals was completely preserved, she was much worse than age-matched controls when tested on rating facial expressions. Together, these two studies show a double dissociation of the processing of facial expression and facial identity.

In short, the neuropsychological study of human patients with brain damage has contributed significantly to our understanding of the processes involved in visual object recognition in humans. It has provided strong support for the monkey model of the human visual system, thereby allowing findings from the monkey literature to be generalized to humans. Furthermore, the existence of impairments specific to object categories or depth of processing, such as visual form agnosia and prosopagnosia, imposes constraints on neural models of object recognition.

Human Neuroimaging

With the advent of non-invasive neuroimaging techniques, it has become possible to more directly study the neural basis of object recognition in humans. Various groups have contributed to understanding which areas of the human ventral visual processing stream are involved in object recognition, what their functional organization is, and what characterizes their respective roles.

A large constellation of areas just anterior and lateral to primary visual cortex have been reported to respond selectively to objects, compared to textures, noise, or scrambled objects (Grill-Spector, 2003). Different objects have been found to activate slightly different areas, with inanimate objects, houses, and scenes activating areas in both dorsal and ventral regions, and faces and animals activating primarily ventrolateral areas. Patterns of activation associated with these stimuli are replicable both within subjects and across subjects (Haxby et al, 2001; Levy et al, 2001). Frequently studied areas include LOC (lateral occipital complex), the fusiform gyrus (divided into posterior, mid, anterior fusiform areas and the fusiform face area, FFA), PPA (parahippocampal place area), and the inferior temporal gyrus (Grill-Spector, 2003).

Models of Object Recognition

A number of possible models have been proposed to describe the functional organization of the human ventral stream. Kanwisher’s group (Downing et al, 2001) has suggested that the temporal cortex is characterized by a small number of category-specific modules in combination with a general purpose recognition region. Regions have been found that respond selectively to biologically salient stimuli such as faces (the fusiform face area, FFA), places (the parahippocampal place area, PPA), and even body parts (extrastriate bodypart area, EBA). The remaining cortex is hypothesized to be a general mechanism capable of perceiving any kind of visually presented object.

An alternative model has been presented by Tarr & Gauthier (2000), who argue that rather than being selective for particular object categories, object representations are clustered according to the required level of processing. They argue that different levels of processing may require dedicated computations, and as such, rather than being selective for faces in particular, areas such as FFA carry out subordinate identification of any object that is automated by expertise. Indeed, Gauthier et al (2000) report finding activity in the supposedly face-specific fusiform face area when expert subjects are required to identify cars and birds at a subordinate level.

A third model, proposed by Malach et al (2002), holds that the organization of object areas in both ventral and dorsal object areas is driven by the resolution needs and eccentricity biases of different objects. For example, areas with peripheral visual representations tend to respond to houses, whereas areas biased towards central representations tend to respond to faces (Levy et al, 2001). Objects recognized on the basis of fine details will therefore differentially activate areas with centrally biased representations, whereas objects recognized on the basis of large-scale features will be associated with peripherally biased representations.

Finally, motivated by replicable patterns of activation in response to particular stimuli, rather than isolated, localized foci, Haxby et al (2001) have proposed that the occipito-temporal cortex is organized according to form attributes. The representation of an object is reflected as a distinct pattern of activation across the whole cortical area. In support of this model, they report finding that the patterns of activation elicited by 8 different categories of objects were replicable within subjects, even without taking the region showing maximal activation to an object into account. Additional support for this theory was reported by Spiridon & Kanwisher (2002), who showed that the patterns of activation associated with different object categories were stable across changes in stimulus format and viewpoint. They do, however, point out that areas that prefer faces and places do not contribute to discriminating between non-preferred categories, indicating that the cortex might not be as homogenous as Haxby and colleagues originally proposed.

It is interesting to note that the model advocated by the Kanwisher group and the model proposed by Haxby et al are not irreconcilable. If the ventral temporal cortex is divided into a small number of specialized modules combined with a general region, as the Kanwisher model posits, then it is unsurprising that objects elicit predictable patterns of activation. Even the finding that the correspondence holds when disregarding the area of maximum activation can be explained if the general region proposed by Kanwisher is inhomogeneous and if we assume that even the processing of objects for which a specific module exists requires some amount of activity outside of that module. In such an interpretation, the general module might be organized according to form attributes as the Haxby model claims. Alternatively, perhaps the entire ventral stream cortex is organized according to form, with particular areas overspecialized for certain form configurations to such a degree that they become selective to certain biologically salient objects, such as faces, body parts, or places.

Object Representations

Any computationally viable object recognition system should be relatively insensitive to exactly what physical cues define an object. The nature of object representation in the cortex has been a much-studied topic over the last few years, and convergent evidence indicates that object-selective regions in the dorsal and especially in the ventral visual stream do indeed show a high degree of perceptual constancy under many manipulations. Such regions have been found to respond consistently when objects are defined by luminance, texture, motion, or stereo cues. Furthermore, object-selective areas are activated both for gray-scale and line drawings, and even when shapes are defined by illusory contours (Grill-Spector, 2003).

Furthermore, object-selective areas have been shown to represent shapes rather than contours. Using an adaptation paradigm, Kourti & Kanwisher (2001) demonstrated that objects with the same shape but different contours share a common neural representation, whereas objects with different shapes but the same contours do not. Similarly, Lerner et al (2002) showed that activation was similar for occluded and non-occluded objects, which differ in their local contours but not in their percept. However, scrambling the occluded stimuli, which left the local contours intact but created a different percept, resulted in different levels of activation.

It is clear that object representations in the ventral occipotemporal cortex are independent of the precise physical cues that visually define the object. However, a second question is whether these representations are also robust across changes in external viewing condition that affect an object's appearance, such as changes in viewpoint, size, and position.

Grill-Spector et al (1999) report that ventral occipotemporal areas, in contrast to lateral occipital areas, show considerable invariance to objects' size and location. This finding is consistent with the functional distinction between ventral and dorsal streams originally made by Ungerleider & Mishkin (1982) and later reinterpreted by Goodale & Milner (1992), since the size and position of an object on the retina are important for guiding motor interaction with that object, but are relatively unimportant for the purposes of object recognition.

As in the psychophysical literature, the neuroimaging evidence on the topic of viewpoint dependence is much less consistent. Grill-Spector et al (1999) report finding viewpoint-dependent activation in object-selective areas in ventral occipitotemporal areas, whereas for example James et al (2002) instead showed some degree of invariance. On the basis of recent indirect evidence such as the adaptation study by Fang & He (2005) and the existence of viewpoint-dependent cells in monkeys, it seems likely that similar viewpoint-dependent cells exist in humans. However, as noted before, the existence of such cells does not necessarily mean that representations as a whole are viewpoint-dependent.

Just like in the monkey literature, the fact that certain areas in the occipito-temporal cortex that are activated when subjects are presented with objects does not prove that these areas are actually involved in object recognition. However, evidence is accumulating that shows that the ventral occipitotemporal area is indeed involved in conscious object perception, rather than simply responding to low-level features (see Grill-Spector, 2003, for a review). By manipulating parameters that impair the ability of subjects to recognize objects, such as the display duration, the amount of occlusion, contrast, and the coherence of the stimulus in noise, it is possible to correlate recognition performance with cortical activity. The consensus from these studies is that object-selective regions in the ventral occipitotemporal cortex are more active when objects are recognized than when they are present but not recognized.

Because the visual stimuli used in these studies differ in the two conditions, it remains impossible to rule out that changes in brain activation were correlated to changes in the visual stimulus rather than to recognition performance. One solution to this problem is to use bistable or ambiguous stimuli. The advantage of these approaches is that the percept changes in the absence of physical changes in the system. Even without changes in the evoking stimuli, strong correlations are found between activation in ventral occipitotemporal object-selective areas and conscious object perception. During binocular rivalry between houses and faces, activation in the parahippocampal place area (PPA) increased, and activation in the fusiform face area (FFA) decreased, when the percept shifted from a face to a house, and vice versa (Tong et al, 1998). Furthermore, the changes in activation were comparable to the changes evoked during real, nonrivalrous alternations. Andrews et al (2002) report that activation in face-selective areas in the fusiform gyrus is higher during periods when a subject viewing the Rubin face-vase illusion (*Figure 3*) perceives two faces compared to the activation when a vase is perceived.

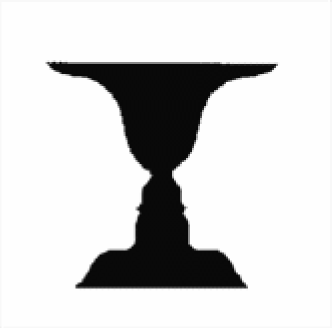

Figure 3: The Rubin face-vase illusion. It is possible to see either two faces (in white) or a vase (in black).

Another solution is to present images at recognition threshold, and then to compare trials in which the subject successfully recognized an object to trials in which the subject failed to recognize an object. Using this approach, Bar et al (2001) found that lateral occipital and ventral occipitotemporal areas show higher activation during trials where the object was recognized than in trials where it was not. Furthermore, as subjects improved in their ability to recognize the brief 40 ms masked stimulus over the course of a training period, the fMRI signal from object-selective areas increased.

Building on this finding, Grill-Spector and colleagues (2004) used the same approach to attempt to separate areas involved in different recognition tasks: detection of an object category and subordinate identification within an object category. Regions that were correlated with successful identification within a category were also correlated with successful detection of that category. Interestingly, they found that only a subset of regions in ventral and lateral occipitotemporal cortex were correlated with successful detection and identification, and that different areas were correlated with different object categories. Activity in the fusiform face area (FFA) was correlated with the successful detection and identification of faces and birds, for example, but not of guitars, whereas the activity in PPA was independent of all three. Similarly, earlier retinotopic visual areas such as V1 and V4 were also not correlated with recognition performance.

The finding that only a small number of subregions were correlated with the detection and identification of particular object categories supports the sparse model proposed by the Kanwisher group (Downing et al, 2001), rather than the model proposed by Haxby et al (2001), which predicts activation across the entire ventral cortex. However, the finding that activity in the fusiform face area (FFA) was correlated to the both detection and identification of both birds and faces contradicts the Kanwisher model, which held that FFA was specifically involved in processing faces. The model by Tarr and Gauthier (2000), which proposes that areas are organized according to level of processing, predicts that areas such as FFA would be correlated to successful identification within a category rather than detection, and as such is also not supported by the data.

Taken together, it is clear that the object-selective regions in the human occipito-temporal cortex are involved in the perception and representation of objects. In contrast to earlier visual areas, these areas are selective for shapes irrespective of the low-level physical cues defining them. Furthermore, the finding that these areas are more active when an object is recognized than when it is not recognized strongly suggests they are involved in the actual conscious recognition process. Although a number of theoretical models of object recognition in the human occipito-temporal cortex have been developed, none of these has yet been able to explain all the variations in the data. It seems likely that a parsimonious model would incorporate features from several of the influential models presented to date.

Conclusion

Vision is arguably the most important of the five sensory systems possessed by human beings. Although the study of the eye and the visual system dates back centuries, the study of the cortical areas of the brain involved in visual perception is a much more recent development. One of the most critical functions of visual perception is to recognize objects, a task that the human brain performs spontaneously, rapidly, and with impressive accuracy across enormous variations in visual conditions. In this thesis I have illustrated how scientists from different disciplines have contributed to our understanding of visual object recognition. I have discussed influential findings in the fields of psychophysics, neurophysiology, neuropsychology, and neuroimaging, as well as the conceptual questions that remain unresolved.

Research in the field of psychophysics has attempted to identity the nature of object representations and the mechanisms used to match visual stimuli to such representations. The consensus seems to be that our mental representations of objects are independent of many linear translations, such as changes in size and position and in some cases rotations. One of the central debates in the psychophysics literature is between image-based and structural-description models. Image-based models propose that internal object representations are viewer-centered such that a visual stimulus is somehow transformed to match the internal representation. Conversely, structural-description models propose that internal object representations are object-centered, and that visual stimuli are broken down into components such that a representation of the object is constructed in its own reference frame, which is then compared to the stored representation. Neither model is without its weaknesses, and it seems likely that the actual mechanism shares features with both approaches.

The study of neurophysiology in monkeys has contributed immensely to our understanding of the cortical processes involved in visual perception. The monkey cortex is functionally divided into dorsal and ventral processing streams, the latter of which is most involved with object recognition. Neurons in the temporal cortex exist that are tuned to specific stimuli or stimulus properties. Neurons with similar selectivity are grouped together in columns, which in turn are organized such that columns selective to related stimuli are close to each other. A strong link has been established between the activity of these cells and the conscious percept, indicating that they are involved in the process of object recognition.

Evidence from human patients with brain damage suggests that the visual object recognition system in the human brain shares organizational characteristics with the macaque brain. Furthermore, considerable evidence suggests that distinct neural substrates underlie the recognition of different object categories. Patients have been reported that show deficits in visual object recognition specific to certain categories, such as faces or facial expressions, or for certain levels of processing.

The development of noninvasive neuroimaging techniques, in particular functional magnetic resonance imaging (fMRI), has made it possible to study in humans a lot of research questions which could previously be studied only in nonhuman primates. Neuroimaging work has implicated areas in the lateral occipital and temporal cortex in object recognition, which are areas largely homologous to the ventral stream areas in monkeys. Activation in these areas is similarly selective to objects or object features and correlated to the conscious percept rather than simply being driven by raw stimulus properties. A number of models have been proposed to explain the functional organization of these object-selective areas and how object representations are stored in the occipito-temporal cortex. Although these models are as yet unable to explain the entire body of reported findings, neuroimaging experiments are currently very popular, and our understanding of many of the commonly implicated areas, such as the lateral occipital area (LO) and the fusiform face area (FFA), is continually improving.

Our understanding of visual object recognition has come a long way over the last decades, but prominent scientists agree that it is not even close to a solved problem. Object recognition is a truly difficult problem. The ability of the human brain, and indeed the brains of even animals as primitive as goldfish, to learn from just a few exemplars, to be highly adaptable to task demands, to be robust across image variation and degradation, and to be effective at different levels of access, is quite simply impressive.

Although an all-encompassing model seems a long way off, scientists from the different disciplines will continue to study and learn to understand small pieces of the puzzle, in the knowledge that one day, we will be able to put together the pieces and see the big picture in the same way that our brain does almost instantly every time you open your eyes.

References

Adolphs, R., Tranel, D., Damasio, H., & Damasio, A. (1994). Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala. *Nature* (372), p. 669-672.

Afraz, S.R., Kiani, R., & Esteky, H. (2006) Microstimulation of inferotemporal cortex influences face categorization. *Nature* (doi:10.1038/nature04982).

Aglioti, S., DeSouza, J.F.X., & Goodale, M.A. (1995). Size-contrast illusions deceive the eye but not the hand. *Current Biology* (5), p. 679–685.

Andersen, R.A., 1987. Inferior parietal lobule function in spatial perception and visuomotor integration. In: V.B. Mountcastle, F. Plum, S.R. Geiger (Eds.), *Handbook of Physiology, Section 1: The Nervous System, Vol. V, Higher Functions of the Brain, Part 2*. American Physiological Association, Bethesda, MD, p. 483–518.

Andrews, T.J., Schluppeck, D., Homfray, D., Matthews, P., & Blakemore, C. (2002). Activity in the fusiform gyrus predicts conscious perception of Rubin’s vase–face illusion. *NeuroImage* (17), p. 890-901.

Bar, M., Tootell, R.B., Schacter, D.L., Greve, D.N., Fischl, B., Mendola, J.D., Rosen, B.R., & Dale, A.M. (2001). Cortical mechanisms specific to explicit visual object recognition. *Neuron (*29), p. 529-535.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review* (94)2, p.115-147.

Biederman, I., & Cooper, E.E. (1991). Evidence for complete translational and reflectional invariance in visual object priming. *Perception* (20)5, p. 585-593.

Biederman, I., & Cooper, E.E. (1992). Size invariance in visual object priming. *Journal of Experimental Psychology: Human Perception and Performance* (18)11, p. 121-133.

Biederman, I., & Gerhardstein, P.C. (1993). Recognizing depth-rotated objects: evidence and conditions for three-dimensional viewpoint invariance. *Journal of Experimental Psychology* (19)6, p. 1162-1182.

Brown, R. (1958). How shall a thing be called? *Psychological Review* (65), p. 14-21.

Bruce, C.J. (1982). Face recognition by monkeys: absence of an inversion effect. *Neuropsychologia* (20), p. 515-521.

Bruce, C., Desimone, R., & Gross, C.G. (1981). Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. *Journal of Neurophysiology* (46), p. 369-384.

Bülthoff, H.H., & Edelman, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Academy of Science USA* (89), p. 60–64.

Carey, S., & Diamond, R. (1977). From piecemeal to configuration representation of faces. *Science* (195), p. 312-313.

Damasio, A.R., Damasio, H., & Van Hoesen, G.W. (1982). Prosopagnosia: anatomic basis and behavioral mechanisms. *Neurology* (32), p.331-341.

Diamond, R., & Carey, S. (1986). Why faces are and are not special: an effect of expertise. *Journal of Experimental Psychology* (115), p. 107-117.

Downing, P.E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A cortical area selective of visual processing of the human body. *Science* (293), p. 2470-2473.

Ellis R., Allport, D.A., Humphreys, G.W., & Collis, J. (1989). Varieties of object constancy. *Quarterly Journal of Experimental Psychology A: Human Experimental Psychology* (41), p. 775-796.

Fang, F., & He, S. (2005). Viewer-centered object representation in the human visual system revealed by viewpoint aftereffects. Neuron (45), p. 793-800.

Foster, D.H., & Kahn, J.I. (1985). Internal representations and operations in the visual comparison of transformed patterns: Effects of pattern point-inversion, positional symmetry, and separation. *Biological Cybernetics* (51)5, p. 305-312.

Fujita, I., Tanaka, K., Ito, M., & Cheng, K. (1992) Columns for visual features of objects in monkey inferotemporal cortex. *Nature* (360), p. 343-346.

Gauthier, I., Hayward, W.G., Tarr, M.J., Anderson, A.W., Skudlarski, P., & Gore, J.C. (2002). BOLD activity during mental rotation and viewpoint dependent object recognition. *Neuron* (34), p. 161-171.

Gauthier, I., Skudlarski, P, Gore, J.C., & Anderson, A.W. (2000). Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience* (3), p. 191-197.

Gentilucci, M. & Rizzolatti, G. (1990). Cortical motor control of arm and hand movements. In: Goodale, M.A. (Ed.), *Vision and Action: The Control of Grasping*, Ablex, p. 147-162.

Goodale, M.A., & Humphrey, G.K. (1998). The objects of action and perception. *Cognition* (67), p. 181-207.

Goodale, M.A., &Milner, A.D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences* (15), p. 20–25.

Grill-Spector, K. (2003). The neural basis of object perception. *Current Opinion in Neurobiology* (13), p. 1-8.

Grill-Spector, K. (2004). The functional organization of the ventral visual pathway and its relationship to object recognition. In: Kanwisher, N., & Duncan, J. (Eds.) *Functional Neuroimaging of Visual Cognition: Attention and Performance XX*. Oxford University Press, Oxford, p. 169-194.

Grill-Spector, K, & Kanwisher, N. (2005). Visual Recognition: as soon as you know it is there, you know what it is. *Psychological Science,* 16(2), p. 152-160.

Grill-Spector, K., Kushnir, T., Edelman, S., Avidan, G., Itzchak, Y., & Malach, R. (1999). Differential processing of objects under various viewing conditions in the human lateral occipital cortex. *Neuron* (24), p. 187-203.

Harris, I.M., & Dux, P.E. (2005). Orientation-invariant object recognition: evidence from repetition blindness. *Cognition* (95), p. 73-93.

Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L, & Pietrini, P (2001). Distributed and overlapping representations of faces and objects in the ventral temporal cortex. *Science* (293), p. 2425-2430.

James, T.W., Humphrey, G.K., Gati, J.S., Menon, R.S., & Goodale, M.A. (2002). Differential effects of viewpoint on object-driven activation in dorsal and ventral streams. *Neuron* (35), p. 793-801.

Jolicoeur, P. (1985). The time to name disoriented natural objects. *Memory and Cognition* (13), p. 289–303.

Jolicoeur, P. (1987). A size-congruency effect in memory for visual shape. *Memory and Cognition* (15), p. 531-543.

Jolicoeur, P., Gluck, M. A. & Kosslyn, S. M. (1984). Pictures and names: Making the connection. *Cognitive Psychology* (16), 243-275.

Kourtzi, Z., & Kanwisher, N. (2001) Representation of perceived object shape by the human lateral occipital complex. *Science* (293), p. 1506-1509.

Kovacs, G., Vogels, R., & Orban, G.A. (1995). Selectivity of macaque inferior temporal neurons for partially occluded shapes. *Journal of Neuroscience* (15), p. 1984-1997.

Kraebel, K.S., & Gerhardstein, P.C. (2005). Three-month-old infant’ object recognition across changes in viewpoint using an operant learning procedure. *Infant Behavior and Development* (29) 1, p. 11-23.

Leopold, D.A., & Logothetis, N.K. (1996). Activity changes in early visual cortex reflect monkeys’ percepts during binocular rivalry. *Nature* (379), 549-553.

Lerner, Y., Hendler, T., Malach, R. (2002). Object completion effects in the human lateral occipital complex. *Cerebral Cortex* (12), p. 163-177.

Levy, I., Hasson, U., Avidan, G., Handler, T., & Malach, R. (2001). Center-periphery organization of human object areas. *Nature Neuroscience* (4), p. 533-539.

Logothetis, N.K., & Pauls, J. (1995). Psychophysical and physiological evidence for viewer-centered representations in the primate. *Cerebral Cortex* (52), p. 70-88.

Logothetis, N.K., Pauls, J., & Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology* (5), p. 552–563.

Logothetis, N.K., & Sheinberg, D.L. (1996). Visual Object Recognition. *Annual Reviews in Neuroscience* (19), p. 577-621.

Malach, R., Levy, I., & Hasson, U. (2002). The topography of high-order human object areas. *Trends in Cognitive Science* (6), p.176-184.

Marr, D., Nishihara, H.K. (1978). Representation and recognition of the spatial organization of three dimensional shapes. *Proceedings of the Royal Society of London B* (200), p. 269–294.

Milliken, B., & Jolicoeur P. (1992). Size effects in visual recognition memory are determined by perceived size. *Memory and Cognition* (20), p. 83-95.

Milner, A.D., Goodale, M.A. (1995). *The Visual Brain in Action*. Oxford University Press, Oxford.

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

Nazir, T.A., O'Regan, J.K. (1990). Some results on translation invariance in the human visual system. *Spatial Vision* (5), p. 81-100.

Perrett, D. I., Mistlin, A. J., Harries, M. H. and Chitty, A. J. (1990). Understanding the visual appearance and consequence of hand actions. In: Goodale, M.A. (Ed.), *Vision and Action: The Control of Grasping*, Ablex, p. 163-180.

Perret, D.I., Rolls, E.T., & Caan, W. (1982). Visual neurons responsive to faces in the monkey temporal cortex. *Experimental Brain Research* (47), p. 329-342.

Perrett, D.I., Oram, M.W., Ashbridge, E. (1998). Evidence accumulation in cell populations responsive to faces: an account of generalisation of recognition without mental transformations. *Cognition* (67). P. 111-145.

Piaget, J. (1969). *The Mechanisms of Perception*. (Translated from French by M. Cook). Basic Books, New York.

Rosch, E., Mervis, C.B., Gray, W.D., Johnson D.M. &Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology* (8), p. 382-439.

Riesenhuber, M., & Poggio, T. (2000). Models of object recognition. *Nature Neuroscience* (3), p. 1199 – 1204.

Sakata, H., Taira, M. (1994). Parietal control of hand action. *Current Opinion in Neurobiology* (4), p. 847–856.

Schyns, P.G. (1998).Diagnostic recognition: task constraints, object information, and their interactions. *Cognition* (67), p.147-179.

Sheinberg, D.L., & Logothetis, N.K. (1997). The role of temporal cortical areas in perceptual organization. *Proceedings of the National Academy of Sciences USA* (94), p. 3408-3413.

Spiridon, M., & Kanwisher, N. (2002). How distributed is visual category information in human occipito-temporal cortex? An fMRI study. *Neuron* (35), p. 1157-1165.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Reviews in Neuroscience* (19), p. 109-139.

Tanaka, K. (1997). Mechanism of visual object recognition: monkey and human studies. *Current Opinion in Neurobiology* (7), p. 523-529.

Tarr, M.J., & Bülthoff, H.H. (1998) Image-based object recognition in man, monkey, and machine. *Cognition* (67), p. 1-20.

Tarr, M.J., & Gauthier, I. (1998). Do viewpoint-dependent mechanisms generalize across members of a class? *Cognition* (67), p. 73-110.

Tarr, M.J., & Gauthier, I. (2000). FFA: a flexible fusiform area for subordinate-level visual processing automatized by expertise. *Nature Neuroscience* (3), p. 764-769.

Tarr, M. J., & Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology* (21), p. 233–282.

Tong, F., Nakayama, K., Vaughan, J.T., & Kanwisher, N. (1998). Binocular rivalry and visual awareness in human extrastriate cortex. *Neuron* (21), p. 753-759.

Tovee, M.J., Rolls, E.T., & Ramachandran, V.S. (1996). Rapid visual learning in neurones of the primate temporal visual cortex. *Neuroreport* (7), p. 2757-2760.

Tranel, D., Damasio, A.R., & Damasio, H. (1988). Intact recognition of facial expression, gender, and age in patients with impaired recognition of face identity. *Neurology* (38), p. 690-696.

Ullman, S. (1998). Three-dimensional object recognition based on the combination of views. *Cognition* (67), p. 21-44.

Ungerleider, L.G., Mishkin, M. (1982). Two cortical visual systems. In: Ingle, D.J., Goodale, M.A., & Mansfield, R.J.W. (Eds.), *Analysis of Visual Behavior*. MIT Press, Cambridge, Massachusetts, p. 549-586.

Valentine, T. (1988). Upside-down faces: a review of the effect of inversion upon face recognition. *British Journal of Psychology* (79) 4, p.71-92.

Wang, G., Tanaka, K., Tanifuji, M. (1996) Optical imaging of functional organization in the monkey inferotemporal cortex. *Science* (272), p. 1665-1668.