Max–Planck–Institut
für biologische Kybernetik
Spemannstraße 38 • 72076 Tübingen • Germany

Technical Report No. 084

Learning to Recognize Objects

Guy Wallis[1] & Heinrich H. Bülthoff[2]

October 2000

Guy Wallis is now at:
Perception and Motor Systems Laboratory
Department of Human Movement Studies
University of Queensland, St. Lucia
QLD 4072, Australia

(Chapter submitted for inclusion in 'Perceptual Learning': Ed. Manfred Fahle and Tomaso Poggio. Please do not cite without authors' permission)

[1] AG Bülthoff, E–mail: gwallis@hms.uq.edu.au
[2] AG Bülthoff, E–mail: heinrich.buelthoff@tuebingen.mpg.de

Learning to Recognize Objects

Guy Wallis & Heinrich H. Bülthoff

Abstract. In this report we review a large body of literature describing how experience affects recognition. Both neurophysiology and psychophysics provide clear evidence for the development of recognition over time. In particular, we show how perceptual learning in recognition tasks can be directly linked to learning in feature tuned inferotemporal lobe neurons in the primate brain. The environment as we experience it, is so structured that potentially very different images appearing in close temporal succession are likely to be views of the same object. We argue that this temporal structure forms the basis of a tendency (a *prior* in the sense of Bayesian Statistics) of the human visual system to associate images of objects together over short periods of time.

---

1 Introduction

In the process of everyday life we are continually analysing and interpreting our visual environment. We effortlessly convert the flat retinal images supplied by our eyes, into a rich three-dimensional world, filled with licorice and ladybirds, shipyards and woods. The apparent speed and ease with which we do this is deceptive. The images cast on our retinas by objects change drastically as a function of viewpoint, lighting, size or location. Consider, for example, the scene depicted in figure 1 in which the same office chair appears several times. We seem to find it trivial to distinguish cast shadows or wall paintings from the genuine article, and it seems self-evident that the chair on the desk is small enough to hold in the hand whereas the chair in the adjacent office is large enough to sit on. We happily conclude this from various cues in the image, despite the fact that the images formed on our retinas by the two chairs are actually identical.

This chapter describes theories of how humans solve the recognition problem, and particularly, how our perception of objects changes with experience. The question of how we recognise objects is an active area of research, and part of this chapter is dedicated to a summary of the proposals that have been made. This is followed by a review of the evidence for perceptual learning in object recognition, ranging from the level of single neurons to that of human behaviour. The chapter concludes by considering how we might learn to associate very dissimilar views of an object, describing how temporal as well as spatial correlations present in our environment can be used to make the necessary associations.

2 Interpreting our visual world

2.1 Introduction

Before launching into the topic of object recognition in detail, this section provides a broader review of visual perception in general. In particular, it covers some of the philosophical arguments which have inspired research, as well as the experimental evidence for and against the influence of perceptual learning in the human visual system.

2.2 Philosophical beginnings

The question of how we construct an internal representation of our visual world has provided fruitful labour for philosophers, psychologists and engineers for many years. Among the first recorded ruminations on the topic appear in the writings of the ancient Greeks, including those of Pythagoras and Plato. Plato was particularly interested in how we recognise and categorise objects. He came to won-

Figure 1: A complex scene comprising many chairs seen with different sizes, viewpoints, lighting conditions etc., demonstrating the range of problems faced in recognising and categorising objects.

der what it was about a cat, for example, that let it be classified as a cat and yet remain distinct from all other cats. He proposed an innate fund of perfect, universal forms, to which all seen objects are likened. Plato's ideas find parallels in current theories of prototypes (Rosch 1973; Posner and Keele 1968; Edelman 1995), but the proposal that such categories are innate gifts has been replaced by the belief that much about visual perception is learnt, and is therefore shaped by our environment. This belief was first voiced in the early 17th century in the writings of René Descartes, and within 50 years it became the guiding principle of a new philosophical movement called empiricism.

The founding father of empiricism was John Locke. Locke rejected all theories of innate knowledge, and although he was prepared to accept a certain amount of prenatal knowledge, he attributed it mainly to sensory experience within the womb. Despite a core of truth to his theories, Locke's uncompromising stance left plenty of scope for criticism and counter argument. Nativist philosophers such as Porterfield and Kant argued that perception requires a framework, an assumed space and time, and a concept of categories to be able to begin to represent the real world. This argument was later championed by the Gestalt movement, which strongly influenced thinking in the first half of the 20th century. Gestalt psychologists such as Köhler and Koffka took the view that human perception is

littered with assumptions which are used to transform the retinal image into an object. They both believed the principles of organisation which they proposed to be fundamental, like laws of physics, enforcing unavoidable and universal constraints on perception.

Köhler condemned the empiricist view, stating:

> Now, when the concept of organisation was first introduced, we were at every step hampered by empiricist explanations... It has been shown, I hope, that ... [Gestalt laws] ... do not allow of explanations in terms of learning, and that therefore, organisation must be accepted as a primary phase of experience. At present we may go further and claim that, on the contrary, any effects which learning has on subsequent experience are likely to be after-effects of previous organisation.
> *(Gestalt Psychology, Köhler 1947)*

Nowadays, many of the Gestaltist laws seem rather vague and anecdotal, but their work did succeed in highlighting a large number of instances in which humans infer form and shape on the basis of a few inbuilt assumptions. Perhaps their only disservice to science was their success, which stifled progress on perceptual learning throughout the early part of the 20th century. It was not until the 1960s and 70s that interest in the seminal work of late 19th century empiricist writers like Helmholtz and James, enjoyed a resurgence of interest. One of the tasks for current researchers is to provide evidence for how much of visual perception is altered by experience and how much is innate.

2.3 The case of S.B.

Having reviewed the nativist vs empiricist debate it is time to consider some of the evidence for the two philosophies. The earliest arguments in favour of perceptual learning stem from the work of the early empiricists. Of course, being philosophers rather than scientists in the modern sense, most of their evidence is enshrined in intellectual argument rather than experimentation. Their preferred approach was to present the reader with a mental conundrum followed by an elegant explanation which furthered their cause. Nevertheless, through such mind games they did make at least some, seemingly testable predictions. One favourite concerned a man born blind, who is suddenly able to see. They speculated on how his tactile experience of the world would transfer to his interpretation of a visual world.

In early 1693, Molyneux was sending comments on a draft of Locke's Essay on Human Understanding, and concluded with:

> ...a jocose problem: Suppose a man born blind... and taught by his touch to distinguish between a cube and a sphere. Suppose then... the blind man made to see; query whether by his sight... he could now distinguish and tell which the globe and which the cube.
> *(From a collection of letters published by Locke 1708)*

Both Locke and Molyneux thought not.

So, did anyone ever live out Molyneux's Gedanken experiment? Several hundred years ago a man born blind was almost certain to remain so, and in more modern times operable cases are usually dealt with soon after birth. However, it turns out that there are a very few cases of people recovering their sight after years of blindness. In their *Experimental Psychology Society Monograph 1963,* Gregory and Wallace review several such cases making reference to the discussions of the empiricist philosophers. They then go on to describe in detail, experiments conducted with S.B., a man who lost his sight at the age of ten months and then had it restored some fifty-two years later. Finally, after two hundred years of waiting, it seemed that S.B. could provide scientists with the opportunity to supplant the empiricists' reasoning with facts.

The first remarkable thing about S.B. was that some tactile information clearly trans-

fered almost instantaneously to his seeing world. In other words, things which S.B. had felt with his hands could often be readily perceived with his eyes. He could, for example, read the time on a clock face across a room. He could also read printed capital letters, although lower case letters meant nothing to him as he had not been taught to read them as a blind schoolboy. Clearly, this type of transferral is at odds with the predictions of Locke and Molyneux. Far from being an exception, there is now good evidence that the transferral from touch to visual perception is quite normal. This has lead theorists to propose a much closer link between the two sense modalities than the early empiricists would have expected. Despite this setback, many other of their predictions did stand the test of time. For example, Locke proposed that depth cues such as shading and perspective are only useful after experience of their meaning:

> When we set before our eyes a round globe... it is certain that the idea thereby imprinted in our mind is of a flat circle variously shadowed... But we having by use been accustomed to perceive what kind of appearance convex bodies are wont to make in us... we have by these an idea of the thing as it is in itself.
> *(Essay on Human Understanding, Locke 1706)*

Gregory and Wallace tested S.B. on various illusions and found him abnormally insusceptible to those shown in figure 2, revealing an unusual insensitivity to depth cues. Indeed S.B. described looking from an upstairs window and feeling that he could step to the ground several floors below without harm. This misapprehension of depth accords both with Locke's words above as well as those of George Berkeley. In Berkeley's *Three dialogues between Hylas and Philonous, 1725*, Philonous argues:

> Now, I find from Experience that when an Object is removed still farther and farther off... its visible Appearance still grows lesser and fainter... [We] perceive Distance, not immediately, but by mediation of a Sign, which hath no Likeness to it, or necessary Connexion with it, but only suggests it from Experience as Words do Things.

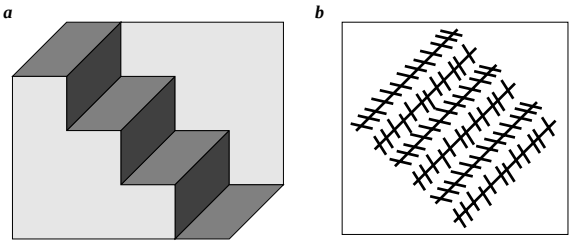

Figure 2: Examples of optical illusions to which S.B., a man 52 years blind and now able to see, was exposed during testing. ***a*** The staircase illusion in which S.B. failed to infer depth from the oblique lines, hence avoiding the usually bistable percept. ***b*** The Zöllner Illusion in which S.B. described seeing straight, parallel lines rather than the collection of bowed black lines described by normal observers.

Not surprisingly, Berkley concludes a similar conversation in his book *Alciphron: Or, The Minute Philosopher, 1732,* thus:

> Now, is it not plain, that if we suppose a Man born blind was on a sudden made to see... [he] could not at first have any Notion of Distance annexed to the things he saw...

Gregory and Wallace's findings support the idea that a great deal of S.B.'s tactile experience of essentially 2D views (of letters and clock faces) carried over to the visual world, but that his perception of depth and the use of cues to 3D form were missing. This is to some extent surprising because S.B. probably went blind at the age of 10 months. Although there is no record of his visual development before this time, had it been normal, developmental studies have shown that by this age he would have gained some ability to process depth information, including the rigidity assumption (Gibson et al. 1978), and the use of stereo disparity (Held 1987). However, none of this early experience seems to have been

retained, and instead S.B.'s experience with his tactile world appears to have formed the basis for what he later saw. Indeed, there is some evidence that shortly after the operation his previous tactile experiences caused curious misperceptions of objects. For example, he initially saw busses as having spoked wheels - as they had had at the turn of the century when he last had cause to touch them as an inquisitive boy. He only correctly perceived the modern solid wheels months after his operation. On the other hand, he had no difficulty in correctly perceiving the shape of a quarter moon, which as a blind man he had imagined would be sliced like quarter of a cake!

Over the next few months, S.B.'s conception of things around him continued to improve. Gregory and Wallace noted that since leaving the hospital, S.B. had become fascinated by the varying appearance of objects: 'Quite recently he had been struck by how objects changed their shape when he walked round them. He would look at a lamp post, walk around it, stand studying it from a different aspect, and wonder why it looked different and yet the same'.

It is clearly tempting to read a great deal into S.B.'s words, but as Gregory and Wallace point out in their paper, this temptation is dangerous, since we cannot see directly into the mind of a blind man. A blind man's vocabulary derives from that of the seeing, and one should therefore exercise caution in interpreting his descriptions of visual experience. Indeed, a recent brain imaging study by Sadato et al. (1996) has demonstrated that the visual cortex of blind people becomes recruited for tasks such as braille reading in a manner quite unlike that in normal subjects. This was confirmed by Cohen et al. (1997) who used magnetic stimulation to disrupt processing in particular brain areas to show that blind people require visual cortex to interpret braille, once again, unlike normal subjects. Bearing this in mind, we should shy away from extrapolating too much from S.B.'s experiences. However, the case does raise some intriguing questions about the influence of experience in interpreting our visual environment, and if the case succeeds only in whetting the reader's appetite for a more detailed treatment of the issues, then it has served a useful purpose.

2.4 Pragmatism in perception

To conclude this opening section we briefly review some examples from the psychophysical literature which argue both for and against perceptual learning. The purpose is to highlight the limitations of nativist and empiricist philosophy, and to draw a more modern perspective in which the goal is to isolate what is and what is not learnt.

For many years, researchers have been aware of perceptual differences arising directly from people's experience, even in adults (Fahle et al. 1995; Gregory 1972). One example to have gained renewed interest recently, concerns an illusion described by Pollock and Chapais (1952). This illusion causes subjects to over-estimate the length of vertical lines relative to horizontal ones, which Baddeley (1997) explains in terms of the level of image correlation occurring at different orientations within natural scenes. Evidence supporting this hypothesis comes from the reliable difference in the magnitude of the horizontal-vertical line length illusion between country-folk from the Norfolk Fens and towns-folk in the City of Glasgow (Ross 1990), environments containing very different amounts of image correlation at different orientations - see figure 3. This work requires further corroboration, but if correct, it provides remarkable evidence that the characteristics of our everyday visual environment directly affect basic perceptual judgements such as line length.

Despite the considerable evidence supporting perceptual learning, there are many counter examples in the literature. For example, an earlier study by Ross and Woodhouse (1979) on the same population of city and country dwellers, found no influence of environment for sensitivity to differences in line orientation. Also, in the field of depth percep-

Figure 3: Images of Glasgow and the Somerset Levels. Like the Norfolk Fens, the Somerset Levels were large, flat floodlands which have since been drained for farming. Glasgow, in contrast, is a large industrial city littered with tall, closely packed buildings. Daily exposure to vertical structures typical of cities may be responsible for the reduced size of the classical horizontal line length illusion in city dwellers compared with country folk, as reported by Ross (1990). Such results provide evidence that our visual diet affects fundamental perceptual judgements. Pictures ©1997 Martin Smith and ©1998 Pete Harlow, reproduced with permission.

tion, it has been shown that we perceive depth from cast shadows on an inbuilt assumption that the light source is situated above and to the left of the viewed object (Ramachandran 1988). Preliminary evidence that this assumption is innate was provided by Hershberger (1970), who showed that without prior experience of shadows, chicks perceive depth from cast shadows on the same assumption.

Apart from innate assumptions for interpreting our environment, some psychologists have also claimed that certain cues or sensory modalities are immutable in the presence of conflicting evidence, forming the framework within which other cues are adapted. For example, Harris (1963) showed that the relearning of hand eye coordination when wearing left-right reversing spectacles is purely motor rather than vision based. Spelke (1990) too, maintains that some systems do not adapt. She argues, for example, that the impression of distance derived from apparent object motion during head movements (motion parallax) is not adaptable, and that it forms an anchor for adapting other sources of information such as stereo vision. However, recent work has revealed a great deal of evidence that all cues to depth can be overridden in the presence of strongly competing cues (Landy et al. 1995), and that the cues interact in a non-linear manner (Bradshaw and Rogers 1996). Indeed Wallis and Bülthoff (1998) have shown that in the presence of additional depth cues, stereo information recalibrates perceived depth from motion, directly countering the argument made by Spelke.

The point in raising these examples is to make it clear that neither the nativist nor empiricist view is exclusively correct. It is important to bear in mind that the force which has shaped us, evolution, is both eclectic and pragmatic. If some advantage is to be had from hard wiring certain assumptions, whilst leaving others to be discovered, then there is nothing to say that evolution has not devised such a compromise. It is also possible that some of the basic assumptions shared with animals as distant from us as chickens, are useful left-overs from before the rapid expansion of neocortex. As Ramachandran (1985) puts it, we should not be surprised to find that evolution has supplied us with a 'bag of tricks' for interpreting the world. Ultimately, what is interesting for those investigating perceptual learning is how many of these tricks are inherited assumptions and how many extrapolated from our environment. For that reason, recent perceptual models have focused on the use of Bayesian mechanics, in which assumptions can be formally incorporated as statistical priors (Bülthoff and Yuille 1996).

In the rest of this chapter we shall be con-

centrating on the mechanisms underlying the representation and recognition of objects, attempting to get to the core of why, as S.B. put it, an object can 'look different and yet be the same'. In particular, we shall explain to what extent the beliefs of the early empiricists are still to be taken seriously today, by describing evidence that our recognition and representation of objects is largely learnt from experience.

3 Object recognition paradigms

3.1 Introduction

This section briefly reviews some of the more current and popular theories of how we represent and recognise objects. By studying the various theories we hope to demonstrate why recognising objects is a difficult task, despite the relative ease with which humans seem able to do it.

3.2 Extracting 3D information

One of the most influential writers in the field of object recognition was David Marr. Marr believed that recognition of an object requires the matching of elemental parts of that object to the parts of 3D models which we have memorised (Marr 1982). The correct apprehension of those 3D features was, for Marr, achieved in three consecutive stages. Firstly, the primal sketch, a wholly two dimensional representation which contains information about lines and edges visible in the scene. Secondly, a $2\frac{1}{2}$D sketch derived from these edges and depth information, describing closed surfaces in space. And finally, a full, three dimensional representation of our environment built from the identified surfaces.

The third stage, Marr argued, provided all of the information required to recognise objects. Recognition itself, involved taking the shapes drawn from the environment and matching them to stored 3D models. These models were themselves built up of constituent parts, or building blocks, which defined an object's shape at various levels of abstraction and detail. Within such a scheme the form of a standing human can, for example, be said to fill the volume of an upright cylinder. A tree or a tower block would be abstracted into the same group, but not a car or a bed whose major axis is horizontal. Beyond this primary level, Marr proposed that we would recognise the six major bodily divisions of the head, trunk and four limbs, which afford discrimination from most non-animal object categories including trees and tower blocks. This recursive analysis then proceeds to the level required to solve a particular discrimination task or to make a specific categorisation judgement.

Although strongly associated with this type of hierarchical approach, Marr was neither the only nor the first person to propose using it in categorisation and recognition. In fact the idea underpins a whole series of theories (Guzman 1971; Marr and Nishihara 1978; Brooks 1981; Tversky and Hemenway 1984) which can be traced back to early attempts to build artificially intelligent systems in the 1970s. Irrespective of the detailed implementation, the approaches are united in the assumption that we use 3D information from our environment to extract 3D parts, and that objects are represented as configurations of these parts. One can think of it as a LEGOland representation. The only major difference in each case is the precise shape and range of LEGO bricks used. Examples include: polyhedra (Waltz 1975), spheres (Badler and Bajcsy 1978), cylinders (Nevatia and Binford 1977; Marr and Nishihara 1978), and 'superquadrics' (Pentland 1986).

3.3 Projective invariants

A quite different approach to identifying objects is the use of projective invariance. Projective invariance refers to the fact that projection of a 3D shape onto a flat surface (like our retina) produces certain characteristic patterns irrespective of the angle at which that 3D feature is being viewed. For example, a triangle remains a triangle from all but the most contrived viewing directions, and so if we detect any surface with three sides, we

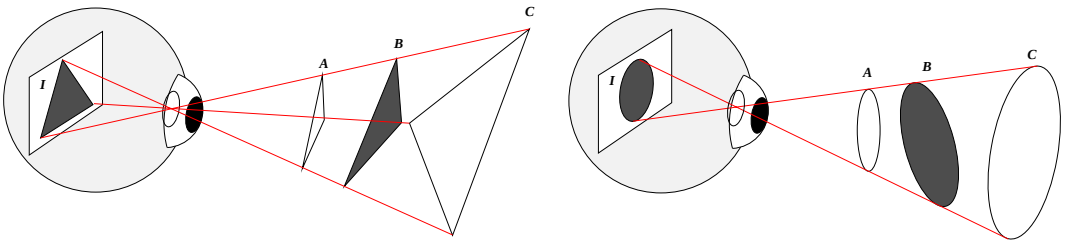

Figure 4: When we view a surface at some random orientation in space, it often results in a characteristic 2D pattern on our retina from which we can infer its true shape. For example: triangles remain triangles and ellipses remain ellipses irrespective of viewpoint. However, the viewer must then decide which of the many possible shapes is responsible for the image seen. In the above examples the image ***I*** is due to the object ***B***, but it could equally well have been due to ***A*** or ***C***.

can label it as a triangle - see figure 4. We can then use the presence of the triangle as the basis for working out what the object is that includes this feature. Other useful invariances include the fact that parallel lines suggest parallel lines in the object, and that ellipses are views of ellipses of equal or smaller aspect ratio.

There are several different levels of tolerance to projective distortion which the human visual system might exhibit. At one extreme there is full projective invariance (Duda and Hart 1973; Cutting 1986; Weiss 1988), which assumes that full 3D information can be recovered from the 2D image on our retina. However, it is possible to subject objects to projective transforms which leave them unrecognisable - suggesting that humans cannot achieve this. Alternatively, humans might simply ignore the affect linear perspective has on the appearance of objects, namely the narrowing of straight lines with distance (Hoffman 1966; Lamdan et al. 1988; Koenderink and van Doorn 1991). Unfortunately, this type of 'affine' approximation cannot distinguish simple shapes like rectangles since they are all affine transforms of each other.

The third, and most promising type of invariance to be investigated is perspective invariance (Grimson et al. 1992; Mundy and Zisserman 1992; Pizlo 1994). Perspective invariance relies upon the types of predictable mappings of triangles and circles etc. mentioned above, and authors have argued that such invariances can form a basis for recognition. There is, however, a problem. There are infinitely many triangles oriented in 3D space which map to any one triangle on our retina, and infinitely many ellipses which map to a single ellipse. Therefore, some heuristic has to be used to decide which of one of the family of possibilities to select - see figure 4. Possible constraints include assuming that the true shape corresponds to the form in which the ratio of an object's area to its perimeter is maximised (Brady and Yuille 1984), or that the true form has bilateral symmetry (Vetter and Poggio 1994). By applying these constraints, variability in appearance due to viewing angle can be eliminated in many cases.

3.4 Geon structural description

One development of the part based, Marrian approach to recognition is the 'Geon Structural Description' due to Biederman (1987). Biederman once again suggests that objects are represented by explicit relationships of a small set of LEGO-like blocks, which he calls 'geons' - see figure 5a. For example, a house might be represented as the base of a pyramid on top of a cube, and a US mail box as a cube on top of a narrow cylinder. In this manner a few (24 in Biederman's opinion) volumetric primitives can be used to describe any object.

Despite the similarity to Marrian ideas, geon theory also owes something to the idea of perspective invariants. Biederman's approach specifically excludes any textural or similar depth cues, concentrating instead on the mapping of 2D space relations to inferences about 3D shape. Biederman cites Lowe's list of non-

accidental 2D properties (Lowe 1984), in his discussion of how 2D cues such as collinearity, skew symmetry and coincident line termination can be used to infer 3D shape.

3.5 Active shape matching

Another alternative to have received consideration - especially from researchers hoping to build working recognition systems - is template matching. The precise details of how to implement such a system vary considerably, but in practice all matching approaches are one of two conceptually important types. The first utilises stored models containing explicit 3D shape information. It therefore assumes that it is possible to extract the location of three (or more) anchor points in 3D space, which are matched to those in the stored models. Matching the anchor points requires a 3D rotation and scaling of the stored model until the anchor points are most closely aligned. Recognition then proceeds by measuring the amount of overlap in the two views (for example, Ullman 1979).

The second approach relies on representations based upon groups of 2D views. For example, in elastic pattern matching a non-linear image transformation is made to the incoming image of the object being viewed. A measure of how well the model matches the stimulus is derived by attributing a cost to how far points in one image have to be moved to find a similar looking feature in the other. Features which have been tried include: Gabor like patches or 'jets' (Buhmann et al. 1990), specific features like end-stopped lines or junctions (Hinton et al. 1992), and edge based facial features like ovals for eyes and a triangle for a nose (Yuille 1991).

3.6 Recognition based on 2D image features

Although the recognition of familiar, everyday objects proceeds almost effortlessly, some views are generally easier to recognise than others, both in terms of reaction times and accuracy. Such views are referred to as 'canonical' in the recognition literature (Palmer et al. 1981).

Many researchers have since studied view specificity using novel objects trained in particular views - see figure 5b. Results consistently point to a decrease in recognition performance as a function of the viewpoint's disparity from a previously learned view (Shepard and Cooper 1982; Rock and DiVita 1987; Bülthoff and Edelman 1992; Tarr and Pinker 1989; Jolicoeur 1990). Similar drops in recognition performance with viewing angle have also been reported for unfamiliar faces (Troje and Bülthoff 1996).

These results have lead to a new alternative for how objects are represented and recognised, namely the feature based, multiple view approach (Bülthoff and Edelman 1992). It bears some relation to earlier 2D matching theories and similarly benefits from the result of Ullman and Basri (1991) that any 2D projection of a 3D object can be written as a *linear* combination of 2D views. However, the multiple views model differs from that of the classical 2D models in two important respects. Firstly, the views are not deformed to match each incoming image, and secondly, each view is represented as a collection of small picture elements, each tolerant to small view changes - i.e. not as single templates. In the feature based scheme, individual neurons are selective to features which occur frequently in the environment. These features may be selective for identifiable things such as noses, or eyes, but most will be responsive to more abstract combinations of edges and surface textures. An ensemble of many hundreds of cells would then be required to act in unison to uniquely identify any one object. The emergent properties of robustness to small variations in the input image (due to view, size or location changes) as well as cell damage, have long been realised by the neural network community (Hinton et al. 1986).

As well as its distinction from other 2D representation schemes, the approach represents a significant departure from object based models, since it neither requires the extraction of depth information, nor the exhaustive

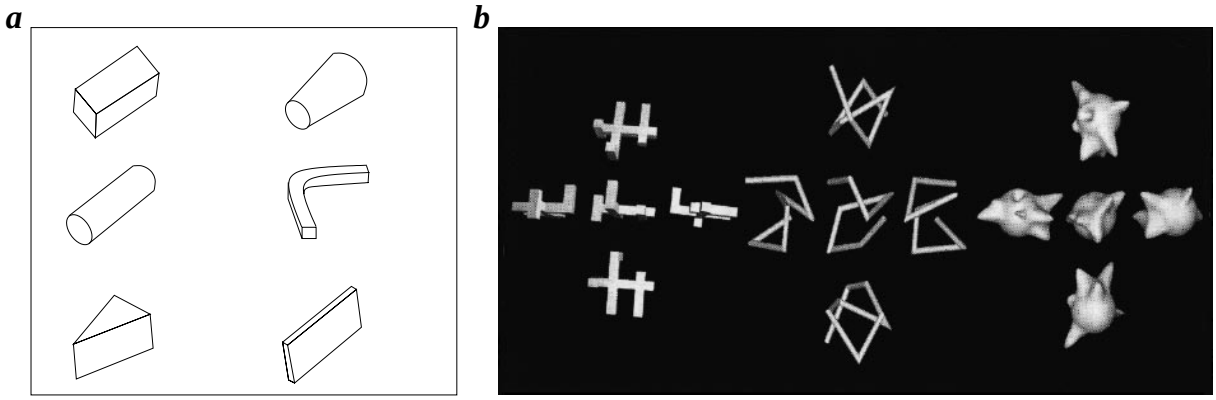

Figure 5: ***a*** Examples of geons, volumetric shape primitives for categorising and recognising natural shapes, adapted from Biederman (1987). In his paper, Biederman describes how a set of 36 geons can exhaustively describe all everyday objects. ***b*** Examples of three data sets used by Bülthoff, Edelman, Tarr and colleagues, to investigate within category view generalisation in novel objects. From left to right: Aerials, Paper clips and Amoebas

matching of 3D models. It is also consistent with a great deal of neurophysiological evidence, as we shall describe in the next section. For those readers interested in a more detailed discussion of the pros and cons of this and the other representation schemes we can recommend several papers which deal directly with this issue (Wallis and Bülthoff 1999; Pizlo 1994; Tarr and Bülthoff 1995; Biederman and Gerhardstein 1993).

4 Learning from examples

4.1 Introduction

Having described various proposals of how objects are represented and recognised, the question arises as to how such representations are learned, and in what sense such learning affects what is perceived. In this section we describe how object representations are established both at the cellular and cognitive level, and how recognition performance alters with experience.

4.2 Neurophysiology

From lesion studies and cellular recording it has been proposed that a series of cortical regions starting in V1 and running ventrally through the occipital into the temporal lobe (V1-V2-V4-Intraparietal areas), solves the problem of *what* we are looking at. In contrast, a second stream leading dorsally and into the parietal lobe (V1-V2-V3-Intraparietal areas), has been implicated in the role of deciding *where* that object is (Farah 1990; Ungerleider and Mishkin 1982; Goodale and Milner 1992; Young 1992) - see figure 6.

Cells in the latter part of the ventral stream, in the inferior temporal areas (IT), are of particular relevance to object recognition because of their tolerance to changes in the precise appearance of their preferred stimuli. Transformations which may be tolerated by IT cells include changes in an object's position, viewing angle or size, as well as overall image contrast or spatial frequency content (Rolls 1992; Desimone 1991; Tanaka et al. 1991) - indeed all of the types of transformation invariance required for view-invariant object recognition. These neurons are also of interest in that they provide a source of evidence of learning in the recognition system. Evidence for experience dependent learning in IT has now been reported by many researchers (Rolls et al. 1989; Miyashita 1993; Logothetis and Pauls 1995; Kobatake et al. 1998).

The link from view based recognition to representations in IT was strengthened through recording work by Logothetis and colleagues (Logothetis and Pauls 1995; Logothetis et al. 1995), in which monkeys were trained to recognise particular aspects of the paper clip stimuli originally used by Bülthoff

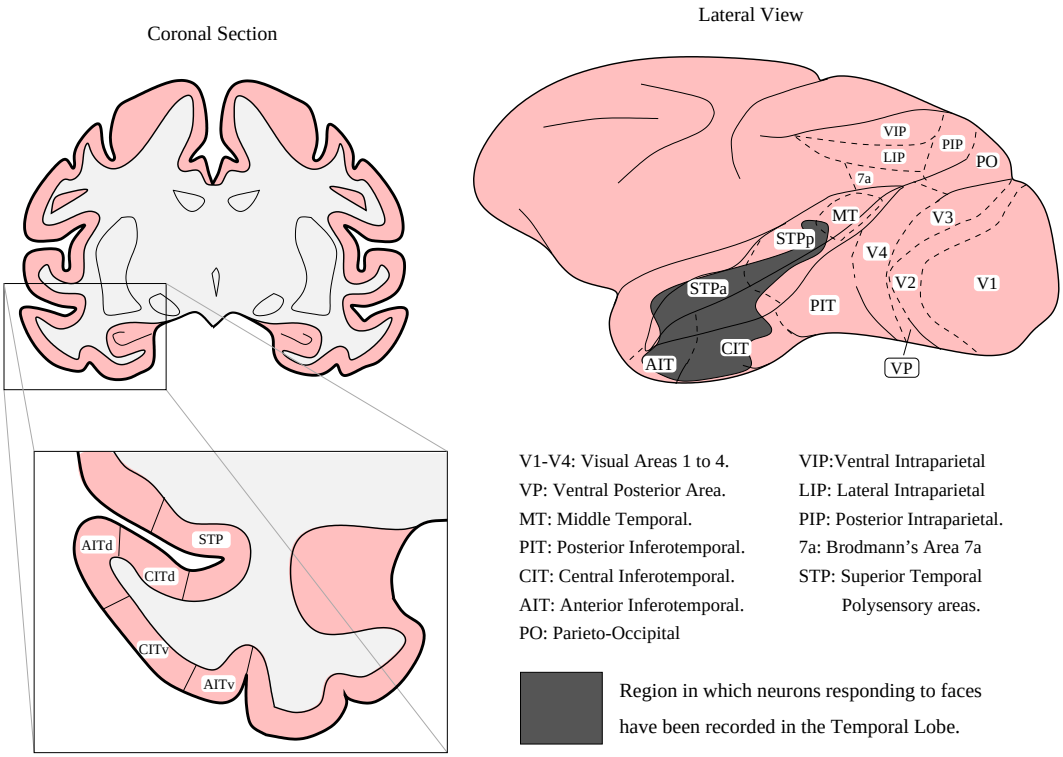

Figure 6: Lateral view and coronal section of the primate cortex showing some of the significant visual processing areas. The expanded coronal section portrays some of the important subdivisions of the temporal lobe. Adapted from Rolls (1992); Perrett et al. (1992).

and Edelman (1992) - see figure 5b. After training, many neurons were shown to have learned representations of particular paper clips, including some selective to specific views.

As well as longer-term changes to cell selectivity, there is also good evidence of almost instantaneous learning in IT cells. Tovee et al. (1996), for example, presented images of strongly lit, two-tone (black and white) faces, referred to as Mooney faces in the literature - see figure 7a. Some IT neurons which did not respond to any of the Mooney faces did so if once exposed to the standard grey-level version of the face - see figure 7b. This accords with findings in humans, who often struggle to interpret Mooney face images the first time, but then have no difficulty in seeing the face a second time, even weeks later (Ramachandran 1994).

4.3 Psychophysics

There is considerable psychophysical evidence that our perception of objects is affected by experience. Even a few days or hours of training can affect the speed and accuracy with which objects are recognised. Bülthoff and Edelman (1992), for example, were able to show that if subjects learn to recognise two views of a novel object, recognition performance is better for new orientations located between the two training views (INTER) than outside them (EXTRA), itself better than for orientations away from the axis linking the trained views (ORTHO) - see figure 8. A first step to explaining these results, within a feature based representation scheme, is to understand why recognition performance drops with distance from a learnt view. We can start by first imagining what happens when leaning a single view. When a view of a novel object is presented, many feature selective neurons respond, and the associated pattern of responses

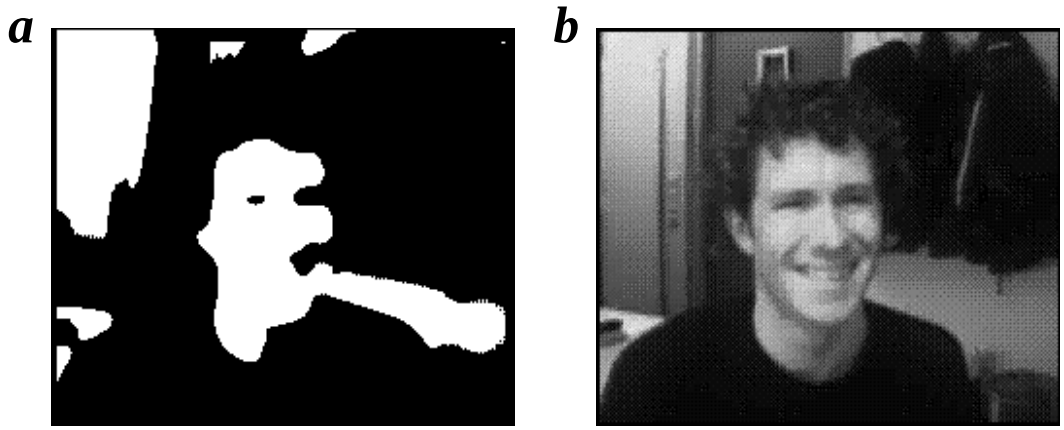

Figure 7: Example of a Mooney face, similar to those used by Tovee et al. (1996). ***a*** If subjects or face selective neurons are exposed to the two-tone image, they often fail to see a face. ***b*** Upon seeing the veridical image, both neurons and subjects can now identify the face and will continue to do so in the future, providing evidence for rapid and lasting learning.

they produce comes to represent the presence of that object. Now, identification of a novel view of the object will clearly be easiest for views nearest to the one trained, since these views are most likely to contain one or more of the features supporting the representation of the learnt view.

Using a similar line of argument, if two views of the object have been learnt, and both are identified as being the same object, then the presence of any of the features seen in either learnt view will tend to evoke recognition of that same object. This second step is important because the INTER and EXTRA results follow as a natural consequence. Clearly, any view falling within the range of the two trained views is more likely to have features in common with either or both of the trained views, than a view of the object from outside that range. Hence INTER views are more likely to be easily recognised than EXTRA views.

The ORTHO effect stems from the type of training used. The training views used were not stationary but rather rocked back and forth through a few degrees. This had the effect that only cells tolerant to changes in the object's appearance along the training meridian (see figure 8) were strongly activated during learning, and hence only they strongly support recognition of the object. Views of the object lying along the training meridian (i.e. INTER and EXTRA views) are much more likely to contain the features for which these cells are selective than views lying on an orthogonal meridian (ORTHO views). Hence, INTER and EXTRA views are more readily recognised.

In a further study, Edelman and Bülthoff (1992) investigated the effects of extensive training, to see if it can override view specificity. After training large numbers of views they were able to change the shape of the recognition curves. Not only did reaction times decrease and accuracy increase, but view specific effects, such as canonicality, gradually disappeared. This issue has been raised again recently in several articles investigating how continued exposure to an object class may affect the manner in which the objects within the class are represented. In a pair of articles, Schyns (Schyns et al. 1998; Schyns 1998) argues that sufficient exposure to a particular stimulus type causes the representation of these stimuli to alter and be enhanced. This in turn relates to the findings of researchers mentioned earlier, who were studying learning in IT neurons. Their work showed that extensive experience of a class of images or objects, causes a rise in the amount of cells selective for those stimuli (Miyashita 1988; Logothetis and Pauls 1995; Kobatake

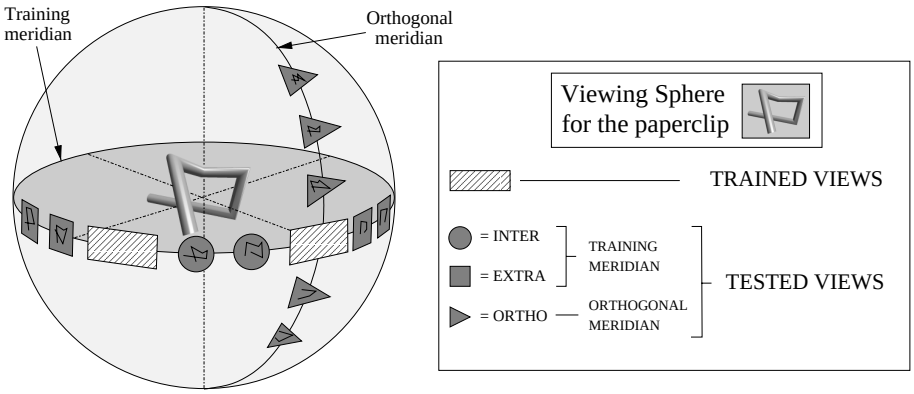

Figure 8: If two views of a novel object are learned, recognition is better for new viewing angles located between the two training views (INTER) than outside them (EXTRA), itself better than for orientations away from the axis linking the trained views (ORTHO).

et al. 1998). By devoting more neural hardware to the representation of the features present in an object class, one would presumably be better able to discriminate subtleties in their form, of relevance to the types of visual expertise raised by Schyns. In another recent article, Gauthier and Tarr (1998) have also made this point, showing that experience of an originally novel object class heightens the subjects' awareness of small changes to objects within the same class.

Gauthier and Tarr go on to argue that our highly sophisticated ability to recognise faces is simply due to a natural concentration of neural resources, resulting from our lengthy exposure to the particular object class we call faces. Some researchers would argue that this is wrong, and that face recognition is special. One of the main motivations for this has been the neurological disorder prosopagnosia. Prosopagnosia is characterised by a normal ability to recognise common objects, contrasted with extreme difficulty in recognising people's faces (De Renzi 1997). The fact that the locus of the brain damage in patients pointed to a part of the temporal lobe homologous to that of cells selective for faces in monkeys (Rolls 1992; Desimone 1991), made a strong case for the suggestion that prosopagnosia was caused by damage to these cells (Farah 1990). Psychological studies have revealed a dissociation between face and object recognition in the past (Tanaka and Farah 1993; Fiser et al. 1996), but the latest picture from the neurophysiological evidence is not as clean as some theorists had first hoped. Direct attempts to find the illusive area responsible for face recognition in monkeys has been controversial and till now unfruitful (Perrett et al. 1992; Cowey 1992). This in turns lends more weight to Tarr and Gauthier's proposal that prosopagnosia may reveal a general deficit in the area dedicated to fine level discriminations of highly trained objects, rather than to a specialist face area *per se*.

Apart from questions of recognition speed and accuracy there is also the question of familiarity. Familiarity is by definition, experienced based, but one interesting prediction to come out of the feature based approach to recognition is that a face made up of previously experienced features, although itself novel to the observer, should appear familiar. This hypothesis has been tested by Solso and McCarthy (1981). In their experiment, subjects were presented with photo-fit pictures of people and then tested on a familiarity task. The test set of faces contained either familiar faces, wholly novel faces, or novel faces containing combinations of features present in the familiar ones. The most intriguing result was that subjects chose the composite faces as more familiar, not only than the unfamiliar faces but than the familiar faces as well. This result not only provides further support for the distributed feature based approach, but

also demonstrates that perceived familiarity need not correlate with true familiarity.

4.4 Temporal continuity as a cue to invariance learning

A broadly tuned feature based system of the type being advocated in this chapter, would be sufficient to perform recognition over small transformations (Poggio and Edelman 1990). However, associating images over larger shape transformations either requires separate pre-normalisation for size and translation of the image, or separate feature detectors which would then be fed into a final decision unit.

As it turns out, the use of pre-normalisation is contrary to the evidence we have from the responses of real neurons implicated in object recognition. Invariance seems to be established over a series of processing stages, starting from neurons with restricted receptive fields and culminating in the types of cell responses found in inferior temporal (IT) cortex mentioned earlier. With this in mind it remains to be explained how one might learn to associate very different views of an object.

One possible solution to this problem is that in the real world, we tend to see discrete sequences of images of an object, often undergoing transformations. This regularity in time may act as an important cue for predicting the identity of an object as it undergoes transformations, due to a change of position relative to the object. This change in viewing position may be simply due to our approaching the object, watching it move, rotating it in our hand and so on. If the time domain is truly influential in setting up representations of objects then there should presumably be some evidence for this in the learning of inferotemporal neurons. In effect one should expect to see quite different views of an object being associated to the same neuron in preference to other very similar images, simply on the basis of the sequence in which they are presented. This last section discusses evidence that temporal relations in the appearance of object views do indeed affect learning.

The temporal association hypothesis has been discussed in the past, and has been successfully used in various neural network models of recognition (Edelman and Weinshall 1991; Földiák 1991; Wallis and Rolls 1997). In particular, Wallis and Baddeley (1997) demonstrated how the temporal statistics of the real world can be optimally used to establish transform invariant representations of objects. The hypothesis has also found direct experimental support from neurophysiological recordings (Stryker 1991; Miyashita 1993). Miyashita (1988), for example, was able to show that repeating a temporal sequence of randomly selected fractal images establishes cells in IT which respond to one stimulus in the series very strongly, but also to those patterns appearing in close succession. He was also able to show that the efficacy of a stimulus dropped purely as a function of temporal and not spatial disparity between stimuli.

Until recently, there was little or no psychophysical evidence to support the theoretical and neurophysiological findings. However, Sinha and Poggio (1996) recently described the use of sequences to establish the perception of the form of ambiguous wire-frame objects, and Wallis (1998) addressed the question directly, by considering the effect of temporal sequences for natural objects such as faces.

In a series of papers, Wallis (Wallis 1998; Wallis and Bülthoff 1997) hypothesised that by exposing observers to sequences of different faces, he could confuse the identity of faces which were seen together in a sequence. This should then become apparent by the increased number of discrimination errors for those faces which were seen in sequences, in comparison with those faces which were not. Figure 9 puts this hypothesis in a more graphical light by displaying two possible sequences each containing two different people's faces. The temporal association hypothesis predicts a higher confusion rate for pairs of faces associated in this way, than for pairs of faces coming from two different sequences. During the experiment, subjects were exposed to 36 such pair-

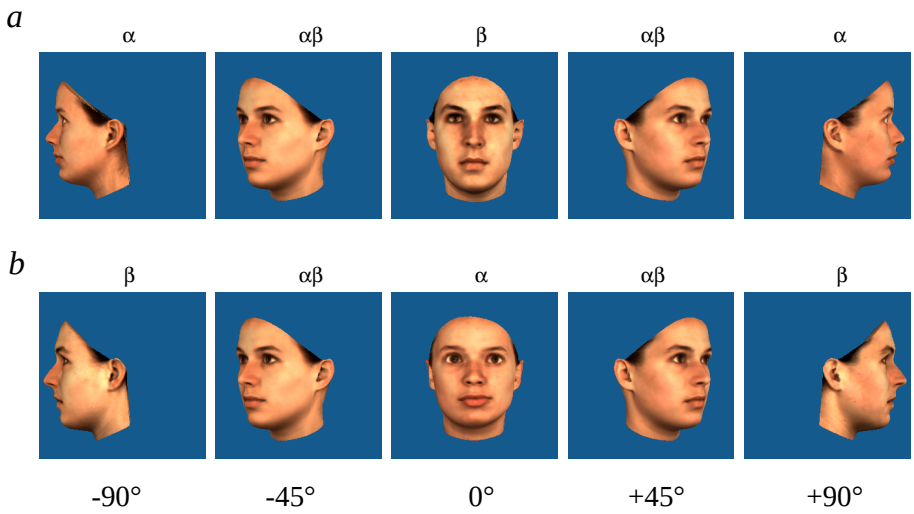

Figure 9: Example of a pair of faces used in the experiments of Wallis and Bülthoff (1997). Each association sequence consisted of the two faces (α and β) in profile and frontal view, and a morphed face ($\alpha\beta$) shown at $\pm 45°$. Subjects saw both sequences ***a*** and ***b*** during training.

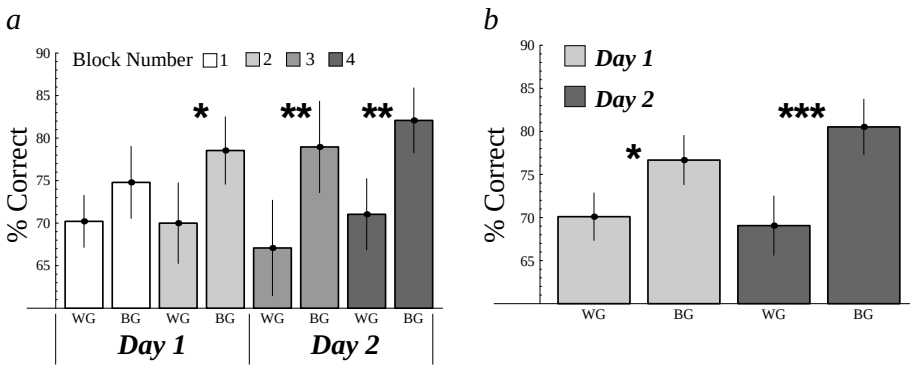

Figure 10: Results from an experiment by Wallis and Bülthoff (1997) in which subjects were asked to discriminate faces previously seen in morphed sequences. ***a*** Discrimination performance across training blocks, measured as percent correct in the mismatch trials. Figures show results for within group (WG) and between group (BG) comparisons. ***b*** The same results broken down across the two training days. ($P < 0.1$) =*, ($P < 0.01$) =**, ($P < 0.001$) =***.

ings of heads and then tested on their ability to discriminate them. The results of their experiment are displayed in figure 10. The symbol 'WG' indicates that the tested faces were from within a group i.e. appeared in a training sequence together. The symbol 'BG' indicates that the faces tested were once again familiar, but came from separate groups, and had thus not been seen in the same training sequence. The results clearly demonstrate that discrimination performance was indeed worse for faces associated in sequences. The difference between the WG and BG condition is also seen to increase with each session of training.

In another recent article, Stone (1998) too looked at the influence of temporal order on the representation of objects. Rather than wire-frame or familiar, facial objects, he used amoeba-like shapes, similar to those of Edelman and Bülthoff (1992). During a learning phase, subjects had to discriminate four objects from numerous distractors. In this first phase, all stimuli rotated in one particular direction. During testing, he allowed certain of

the trained objects to be rotated in the opposite direction which caused a drop in discrimination performance, and an increase in reaction times. It is worth noting that although similar to the results described above, Stone's results propose something new, since they suggest that temporal information forms part of the representation of the object.

The ability of a time-based association mechanism to correctly associate arbitrary views of objects, without an explicit external training signal, means that it could overcome many of the weaknesses of using supervised training schemes or associating views simply on the basis of physical appearance. For this reason, the three experiments described above may well represent a significant new step in establishing the 2D multiple view approach to object recognition.

5 Conclusion

In this chapter we have reviewed a large body of literature describing how experience affects recognition. Both neurophysiology and psychophysics provide clear evidence for the development of recognition over time - such as the adaptation to Mooney faces, or the fall in canonical view effects with prolonged exposure to many views of the object. In particular, we have explained how perceptual learning in recognition tasks can be directly linked to learning in feature tuned inferotemporal lobe neurons in the primate brain.

We have also described that the environment as we experience it, is so structured that potentially very different images appearing in close temporal succession are likely to be views of the same object. We then argued that this temporal structure forms the basis of a tendency (a *prior* in the sense of Bayesian Statistics) of the human visual system to associate images of objects together over short periods of time.

The results described in this chapter strongly support the empiricist view that object recognition and categorisation is largely an ongoing process, affected by experience of our environment. S.B.'s insusceptibility to visual illusions and his failed perception of depth, all point to the fact that much of our ability to interpret the form of objects and scenes is learnt. By using novel stimuli it has been possible for researchers to study more precisely how object representation and recognition develops with everyday visual experience. Taken as a whole the results serve to underpin the main tenet of this book, namely that perception is mediated via a dynamic learning system, the modification of which continues throughout our lives.

References

Baddeley, R. (1997). The correlational structure of natural images and the calibration of spatial representations. *Cognitive Science*, 21(3):351–372.

Badler, N. and Bajcsy, R. (1978). Three dimensional representations for computer graphics and computer vision. *Computer Graphics*, 12:153–160.

Berkeley, G. (1725). *Three dialogues between Hylas and Philonous : the design of which is plainly to demonstrate the reality and perfection of human knowledge, the incorporeal nature of the soul, and the immediate providence of a deity: in opposition to sceptics and atheists.* Innys, London.

Berkeley, G. (1732). *Alciphron: Or, The Minute Philosopher : In Seven Dialogues; Containing an Apology for the Christian Religion, against those who are called Free-thinkers.* Tonson, London, 2 edition.

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

Biederman, I. and Gerhardstein, P. (1993). Recognizing depth–rotated objects: Evidence and conditions for 3D viewpoint invariance. *Journal of Experimental Psychology: Human Perception and Performance*, 20(1):80.

Bradshaw, M. and Rogers, B. (1996). The interaction of binocular disparity and motion parallax in the computation of depth. *Vision Research*, 36(21):3457–3468.

Brady, M. and Yuille, A. (1984). An extremumprinciple for shape from contour. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:288–301.

Brooks, R. (1981). Symbolic reasoning among 3–d and 2–d images. *Artificial Intelligence*, 17:205–244.

Buhmann, J., Lades, M., and von der Malsburg, C. (1990). Size and distortion invariant object recognition by hierarchical graph matching. In *International Joint Conference on Neural Networks*, pages 411–416. New York: IEEE.

Bülthoff, H. and Edelman, S. (1992). Psychophysical support for a two–dimensional view interpolation

theory of object recognition. *Proceedings of the National Academy of Sciences, USA*, 92:60–64.

Bülthoff, H. and Yuille, A. (1996). A Bayesian framework for the integration of visual modules. In McClelland, J. and Inui, T., editors, *Attention Performance XVI: Information Integration in Perception and Communication*, pages 49–70. MIT Press.

Cohen, L., Celnik, P., Pascual-Leone, A., Corwell, B., Falz, L., Dambrosia, J., Honda, M., Sadato, N., Gerloff, C., Catala, M., and Hallett, M. (1997). Functional relevance of cross-modal plasticity in blind humans. *Nature*, 389:180–183.

Cowey, A. (1992). The role of the face-cell area in the discrimination and recognition of faces by monkeys. *Philosophical Transactions of the Royal Society, London [B]*, 335:31–38.

Cutting, J. (1986). *Perception with an Eye for Motion*. Cambridge, Massachusetts: MIT Press.

De Renzi, E. (1997). Prosopagnosia. In Feinberg, T. and Farah, M., editors, *Behavioural neurology and neuropsychology*, pages 245–255. McGraw-Hill, New York.

Desimone, R. (1991). Face–selective cells in the temporal cortex of monkeys. *Journal of Cognitive Neuroscience*, 3:1–8.

Duda, R. and Hart, P. (1973). *Pattern Classification and Scene Analysis*. Wiley, New York.

Edelman, S. (1995). Representation of similarity in three-dimensional object discrimination. *Neural Computation*, 7:408–423.

Edelman, S. and Bülthoff, H. (1992). Orientation dependence in the recognition of familiar and novel views of 3D objects. *Vision Research*, 32:2385–2400.

Edelman, S. and Weinshall, D. (1991). A self–organising multiple–view representation of 3D objects. *Biological Cybernetics*, 64:209–219.

Fahle, M., Edelman, S., and Poggio, T. (1995). Fast perceptual learning in hyperacuity. *Vision Research*, 35(21):3003–3013.

Farah, M. (1990). *Visual Agnosia: Disorders of Object Recognition and What They Can Tell Us About Normal Vision*. Cambridge, Massachusetts: MIT Press.

Fiser, J., Biederman, I., and Cooper, E. (1996). To what extent can matching algorithms based on direct outputs of spatial filters account for human object recognition? *Spatial Vision*, 10(3):237–271.

Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Computation*, 3:194–200.

Gauthier, I. and Tarr, M. (1998). Becoming a "Greeble" expert: Exploring mechanisms for face recognition. *Vision Research*, 37:1673–1682.

Gibson, E., Owsley, C., and Johnston, J. (1978). Perception of invariants by five month old infants. *Developmental Psychology*, 14:407–415.

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

Gregory, R. (1972). *Eye and Brain : The Psychology of Seeing*. Weidenfeld and Nicolson, London, 2 edition.

Gregory, R. and Wallace, J. (1963). Recovery from early blindness: A case study. In *Experimental Psychology Society Monograph No. 2*. W. Heffer & Sons, Cambridge, UK.

Grimson, W., Horn, B., Poggio, T., and staff (1992). Progress in image understanding at mit. In *Proceedings of the Image Understanding Workshop*, pages 69–82.

Guzman, A. (1971). Analysis of curved line drawings using context and global information. *Machine Intelligence*, 6:325–375.

Harris, C. (1963). Adaptation to displaced vision: Visual, motor, or proprioceptive change? *Sciene*, 140:812–813.

Held, R. (1987). Visual development in infants. In Adelman, G., editor, *The Encyclopedia of Neuroscience*, volume 2. Birkhäuser, Boston.

Hershberger, W. (1970). Attached-shadow orientation perceived as depth by chickens reared in an environment illuminated from below. *Journal of Comparative and Physiological Psychology*, 73(3):407–411.

Hinton, G., McClelland, J., and Rumelhart, D. (1986). Distributed representations. In Rumelhart, D. and McClelland, J., editors, *Parallel Distributed Processing*, volume 1: Foundations, chapter 3. Cambridge, Massachusetts: MIT Press.

Hinton, G., Williams, C., and Revow, M. (1992). Adaptive elastic models for hand–printed character recognition. In Moody, J., Hanson, S., and Lippman, R., editors, *Advances in Neural Information Processing Systems*, volume 4, pages 512–519. San Meteo, California: Morgan Kaufmann.

Hoffman, W. (1966). The lie algebra of visual perception. *Journal of Mathematical Psychology*, 3:65–98.

Jolicoeur, P. (1990). Orientation congruency effects on the identification of disoriented shapes. *Journal of Experimental Psychology: Human Perception and Performance*, 16:351–364.

Kobatake, E., Tanaka, K., and Wang, G. (1998). Effects of shape discrimination learning on the stimulus selectivity of inferotemporal cells in adult monkeys. *Journal of Neurophysiology*, 80:324–330.

Koenderink, J. and van Doorn, A. (1991). Affine structure from motion. *Journal of the Optical Society of America, A*, 8:377–385.

Köhler, W. (1947). *Gestalt Psychology*. Liveright, New York.

Lamdan, Y., Schwartz, J., and Wolfson, H. (1988). Object recognition by affine invariant matching. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 335–344.

Landy, M., Maloney, L., Johnston, E., and Young, M. (1995). Measurement and modeling of depth cue combination: In defense of weak fusion. *Vision Research*, 35(3):389–412.

Locke, J. (1706). *An Essay Concerning Human Understanding*. London, 5 edition.

Locke, J. (1708). *Some familiar letters between Mr. Locke and several of his friends*. London.

Logothetis, N. and Pauls, J. (1995). Viewer-centered object representations in the primate. *Cerebral Cortex*, 3:270–288.

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

Lowe, D. (1984). *Perceptual Organization and Visual Recognition*. PhD thesis, Stanford University, Stanford, CA.

Marr, D. (1982). *Vision*. San Francisco: W.H. Freeman and Co.

Marr, D. and Nishihara, H. (1978). Representation and recognition of the spatial organization of three dimensional structure. *The Proceedings of the Royal Society, London [B]*, 200:269–294.

Miyashita, Y. (1988). Neural correlate of visual associative long–term memory in the primate temporal cortex. *Nature*, 335:817–820.

Miyashita, Y. (1993). Inferior temporal cortex: Where visual perception meets memory. *Annual Review of Neuroscience*, 16:245–263.

Mundy, J. and Zisserman, A. (1992). Introduction–towards a new framework for vision. In Mundy, J. and Zisserman, A., editors, *Geometric Invariance in Computer Vision*, pages 1–39. Cambridge, Massachusetts: MIT Press.

Nevatia, R. and Binford, T. (1977). Description and recognition of curved objects. *Artificial Intelligence*, 8:77–98.

Palmer, S., Rosch, E., and Chase, P. (1981). Canonical perspective and the perseption of objects. In Long, J. and Baddeley, A., editors, *Attention and Performance IX*, pages 131–151. Hillsdale, N.J.: Erlbaum.

Pentland, A. (1986). Perceptual organization and the representation of natural form. *Artificial Intelligence*, 28:293–331.

Perrett, D., Hietanen, J., Oram, M., and Benson, P. (1992). Organisation and functions of cells responsive to faces in the temporal cortex. *Philosophical Transactions of the Royal Society, London [B]*, 335:23–30.

Pizlo, Z. (1994). A theory of shape constancy based on perspective invariants. *Vision Research*, 34(12):1637–1658.

Poggio, T. and Edelman, S. (1990). A network that learns to recognize three-dimensional objects. *Nature*, 343:263–266.

Pollock, W. and Chapais, A. (1952). The apparent length of a line as a function of its inclination. *Quarterly Journal of Experimental Psychology*, 4:170–178.

Posner, M. and Keele, S. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, 77:353–363.

Ramachandran (1994). 2D or not 2D – that is the question. In Gregory, R. and Harris, J., editors, *The Artful Eye*. Oxford: Oxford University Press.

Ramachandran, V. (1985). The neurobiology of perception. *Perception*, 14:97–103.

Ramachandran, V. (1988). Perception of shape from shading. *Nature*, 331:163–166.

Rock, I. and DiVita, J. (1987). A case of viewer–centered object perception. *Cognitive Psychology*, 19:280–293.

Rolls, E. (1992). Neurophysiological mechanisms underlying face processing within and beyond the temporal cortical areas. *Philosophical Transactions of the Royal Society, London [B]*, 335:11–21.

Rolls, E., Baylis, G., Hasselmo, M., and Nalwa, V. (1989). The effect of learning on the face selective responses of neurons in the cortex in the superior temporal sulcus of the monkey. *Experimental Brain Research*, 76:153–164.

Rosch, E. (1973). On the internal structure of perceptual and semantic categories. In Moore, T., editor, *Cognitive Development and the Aquisition of Language*, pages 111–144. Academic Press, New York.

Ross, H. (1990). Environmental influences on geometrical illusions. In Muller, F., editor, *Frechner Day 90: Proceeding of the 6th Annual Meeting of the International Society of Psychophysicists*, pages 216–221.

Ross, H. and Woodhouse, J. (1979). Genetic and environmental factors in orientation anisotropy: a field study in the british isles. *Perception*, 8:507–521.

Sadato, N., Pascual-Leone, A., Grafman, J., Ibanez, V., Deiber, M., Dold, G., and Hallett, M. (1996). Activation of the primary visual cortex by Braille reading in blind subjects. *Nature*, 380:526–528.

Schyns, P. (1998). Categories and percepts: A bidirectional framework for categorization. *Trends in Cognitive Sciences*, 1:183–189.

Schyns, P., Goldstone, R., and Thibaut, J.-P. (1998). The development of features in object concepts. *Behavioural and Brain Sciences*, 21:1–54.

Shepard, R. and Cooper, L. (1982). *Mental Images and their Transforms*. Cambridge, Massachusetts: MIT Press, 3 edition.

Sinha, P. and Poggio, T. (1996). Role of learning in three-dimensional form perception. *Nature*, 384:460–463.

Solso, R. and McCarthy, J. (1981). Prototype formation of faces: A case of pseudo-memory. *British Journal of Psychology*, 72:499–503.

Spelke, E. (1990). Origins of visual knowledge. In Osherson, D. and Kosslyn, editors, *Visual Cognition and Action: An Invitation to Cognitive Science*, volume 2, chapter 10, pages 99–127. MIT Press, Cambridge, MA, USA.

Stone, J. (1998). Object recognition using spatiotemporal signatures. *Vision Research*, 38(7):947–951.

Stryker, M. (1991). Temporal associations. *Nature*, 354:108–109.

Tanaka, J. and Farah, M. (1993). Parts and wholes in face recognition. *Quarterly Journal of Experimental Psychology [A]*, 46:225–245.

Tanaka, K., Saito, H., Fukada, Y., and Moriya, M. (1991). Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *Journal of Neurophysiology*, 66:170–189.

Tarr, M. and Bülthoff, H. H. (1995). Is human object recognition better described by geon-structural-descriptions or by multiple-views? *Journal of Experimental Psychology: Human Perception and Performance*, 21:1494–1505.

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

Tovee, M., Rolls, E., and Ramachandran, V. (1996). Rapid visual learning in neurones of the primate temporal visual cortex. *Neuroreport*, 7:2757–2760.

Troje, N. and Bülthoff, H. (1996). Face recognition under varying poses: The role of texture and shape. *Vision Research*, 36:1761–1771.

Tversky, B. and Hemenway, K. (1984). Objects, parts and categories. *Journal of Experimental Psychology: General*, 113:169–193.

Ullman, S. (1979). *The Interpretation of Visual Motion*. Cambridge, Massachusetts: MIT press.

Ullman, S. and Basri, R. (1991). Recognition by linear combinations of models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13:992–1005.

Ungerleider, L. and Mishkin, M. (1982). Two cortical visual systems. In Ingle, D., Goodale, M., and Mansfield, R., editors, *Analysis of Visual Behavior*, pages 549–586. Cambridge, Massachusetts, USA: MIT press.

Vetter, T. and Poggio, T. (1994). Symmetric 3D objects are an easy case for 2D object recognition. *Spatial Vision*, 8(4):443–453.

Wallis, G. (1998). Temporal order in human object recognition learning. *Journal of Biological Systems*, 6(3):299–313.

Wallis, G. and Baddeley, R. (1997). Optimal unsupervised learning in invariant object recognition. *Neural Computation*, 9(4):883–894.

Wallis, G. and Bülthoff, H. (1997). Temporal correlations in presentation order during learning, affects human object recognition. *Perception*, 26(ECVP suppl.):32.

Wallis, G. and Bülthoff, H. (1998). Using a 'virtual illusion' to put parallax in its place. *Perception*, 27(ECVP suppl.):To appear.

Wallis, G. and Bülthoff, H. (1999). Learning to recognize objects. *Trends in Cognitive Sciences*, 3:22–31.

Wallis, G. and Rolls, E. (1997). A model of invariant object recognition in the visual system. *Progress in Neurobiology*, 51:167–194.

Waltz, D. (1975). Generating semantic descriptions from drawings of scenes with shadows. In Winston, P., editor, *The Psychology of Computer Vision*. New York: McGraw-Hill.

Weiss, I. (1988). Projective invariants of shapes. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 291–297.

Young, M. (1992). Objective analysis of the topological organization of the primate cortical visual-system. *Nature*, 358:152–155.

Yuille, A. (1991). Deformable templates for face recognition. *Journal of Cognitive Neuroscience*, 3(1):59–71.