Learning to recognize objects

Guy Wallis and Heinrich Bülthoff

Evidence from neurophysiological and psychological studies is coming together to shed light on how we represent and recognize objects. This review describes evidence supporting two major hypotheses: the first is that objects are represented in a mosaic-like form in which objects are encoded by combinations of complex, reusable features, rather than two-dimensional templates, or three-dimensional models. The second hypothesis is that transform-invariant representations of objects are learnt through experience, and that this learning is affected by the temporal sequence in which different views of the objects are seen, as well as by their physical appearance.

In the late 1950s Gregory and Wallace had the rare opportunity to investigate the case of a man who, thanks to an operation, was able to see for the first time in over fifty years[1]. Three months after the operation they made the following observation:

> Quite recently [S.B.] 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.
>
> Richard Gregory and Jean Wallace (in Ref. 1)

As the result of his postponed exposure to the visual world, S.B. may be one of the very few adults to have appreciated how strange and difficult a problem it is to recognize objects from different viewpoints. We are continually required to recognize objects in the process of our everyday life, but the apparent speed and ease with which we are able to solve the task makes it difficult to appreciate how remarkable this ability is. The image cast on our retina by an object changes markedly as a function of viewpoint, lighting, size or location, but we are nevertheless able to interpret these images correctly as indicating the presence of one or other object. Consider, for example, the scene depicted in Fig. 1, in which the many different views of a chair are all recognizable as such.

This review describes theories of how humans solve the recognition problem. One of the major issues that we consider is how the very perception of objects changes with experience. The role of experience in human perception has yet to be fully understood but there is now good evidence that much of our perceptual apparatus is affected by learning. In this article we describe both neurophysiological and human psychophysical evidence for an experience-based influence on object representation. The evidence from learning studies is then used to support the proposal that objects are stored as collections of views each represented by small collections of neurons – countering the many alternative approaches summarized in Box 1. In the final section we describe how disparate views of objects can be associated on the basis of the temporal as well as spatial regularities governing their appearance.

Neurophysiology

From lesion studies and cellular recording it has been proposed that the sequence of primate visual areas (V4→PIT→CIT→AIT) – often referred to as the ventral stream – solve 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 located[2–5] (Fig. 2). In particular, cells in the latter part of the ventral stream in the inferior temporal areas (IT) have been implicated in the task of object recognition, ostensibly because of their transformation-tolerant selectivity for views of faces[6–10]. Transformations that can be tolerated by IT cells include changes in the position, viewing angle, image contrast, size and spatial frequency content[6,7,11] – indeed all of the types of transformation invariance required for view-invariant object recognition. Although there are concentrations of face cells in IT, this area is not simply a face cell area, since clusters of face cells are interspersed with clusters of cells not selective to faces[12,13]. In fact, although face cells account for as much as 20% of neurons in some regions of IT and STS, they only account for around 5% of all cells present in inferior temporal cortex[14]. In the early 1990s, Tanaka and his colleagues[11] showed that many of the remaining neurons are selective for complex combinations of features, including a basic shape with bounded light, dark or colored regions etc. and that these neurons also demonstrated useful invariance properties.

Apart from their tolerance to stimulus transformations, inferotemporal neurons are of interest in that they provide a source of evidence of learning in the recognition system. Rolls *et al.*[15], for example, were able to demonstrate rapid adaptation of a neuron's selectivity for faces. In addition, both Miyashita[16] and Kobatake *et al.*[17] found cells responsive to

G. Wallis and H.H. Bülthoff are at the Max-Planck Institute for Biological Cybernetics, Spemannstraße 38, 72076 Tübingen, Germany.
G. Wallis is currently at the Perception and Motor Systems Laboratory, Connell Building, University of Queensland, St Lucia, QLD 4072, Australia.

tel: + 61 7 3365 6817
fax: + 61 7 3365 6877
e-mail: gwallis@hms.uq.edu.au; hhb@kyb.tuebingen.mpg.de

1364-6613/99/$ – see front matter PII: S1364-6613(98)01261-3

Fig. 1. The problems inherent in analysing complex scenes. A complex scene comprising many chairs seen with different sizes, viewpoints, lighting conditions, etc., demonstrating the range of problems faced in recognizing and categorizing objects. For example, we appear to find it trivial to distinguish cast shadows or wall paintings of chairs from the genuine article. It also 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 despite the fact that the images formed by the two chairs on our retinas are actually identical.

familiar, non-face stimuli used in previous training. In particular, Kobatake *et al.* demonstrated that the number of cells selective for the trained stimuli was significantly higher than in the cortex of naive monkeys. More recently, Logothetis and Pauls[18] trained monkeys to recognize particular aspects of the novel object class of paper clips which had been used in earlier recognition studies[19]. After training, many IT neurons were shown to have learned representations of particular paper clips – including some neurons that were selective to specific views.

Learning in IT cells can be built up over many months, but can also be almost instantaneous, reflecting behavioral changes measured in human responses to stimuli. Tovee *et al.*, for example, presented two-tone images of strongly lit faces to monkeys[20] (see Fig. 3a). Some IT neurons which did not respond to any of the two-tone faces did so if once exposed to the standard grey-level version of the face (Fig. 3b). This accords with findings in humans, who often struggle to interpret two-tone images for the first time, but then have no difficulty interpreting the same image, even weeks later[21].

Psychophysical studies

Apart from the accumulating evidence for the experience-dependent modification of neural responses, there are also ample examples in the field of human object recognition. One of the important recent developments has been the use of stimuli chosen from novel object classes. What emerged from this work was that if two views of a novel object were learned, recognition was better for new views oriented between the two training views, than for views lying outside them[19,22] (see Fig. 4). These view-dependent effects were at odds with most of the then current theories of object recognition and helped establish the view-based approach to object recognition[19,22]. This theory proposes that objects are represented as collections of views rather than explicitly related parts or 3-D models. For discussion of this and other approaches to the problem of object recognition see Box 1.

One important aspect of view-dependent recognition is that it is most noticeable for unfamiliar objects, or for objects usually seen from a particular viewpoint[23] – for which the familiar view is referred to as 'canonical'. For other familiar objects it has long been known that recognition is view-invariant[24,26]. This is, however, still consistent with a view-based model if one assumes that for familiar objects, enough views are stored to remove any view-specific effects. To demonstrate that view-dependence is purely a function of familiarity, Edelman and Bülthoff investigated the effects of extensive training, and showed that it can indeed counter initial view specificity[27] (Fig. 5).

The effect of training has again been raised by several recent articles investigating how continued exposure to an object class may affect the manner in which the objects

Box 1. Object recognition paradigms

Object recognition, shape constancy or shape perception – call it what you will – the question of how we identify and classify objects, has provided fruitful labor for philosophers, psychologists, engineers and scientists for many years. Here, we briefly review some of the more current and popular recognition theories.

Extracting 3-D information

The idea that we can extract 3-D information from a scene which we then use to access 3-D models of objects, is referred to as an 'object-based' representation. One approach for extracting 3-D information from 2-D projections is to use depth cues. Natural images usually contain depth information in properties such as texture gradients, shading, hue, binocular disparity and motion. Most of these cues have been used in shape extraction including shape from shading (Refs a,b) and shape from motion (Ref. c). These cues are certainly useful, but appear unnecessary for recognition because of the fact that we are able to infer shape from line drawings. Of course, even line drawings provide some depth information if only because of our predisposition to infer depth information from oblique lines (e.g. in a Necker Cube), or to make assumptions about object symmetry (Ref. d).

However one extracts the 3-D form, one needs to find some means of matching the viewed object to stored representations of familiar objects. Many authors proposed that matching occurred at different scales depending on the difficulty of the discrimination (Refs e–g). All of these theories center around the idea that objects are represented as specific configurations of basic parts, or building blocks. One could think of it as a LEGO-land representation. There have been numerous suggestions for the choice of LEGO bricks, including polyhedra, spheres, cylinders (Ref. f), superquadrics (Ref. a) and 'geons' (Ref. h).

Projective invariants

A quite different means of identifying objects to have received considerable attention is the notion of projective invariants, that is, the characterization of intrinsic elements of shape that are unaffected by the act of projection onto a 2-D surface such as the retina. Approaches include both projective invariants (Ref. i) and affine invariants (Ref. j), the latter being an approximation to the effects of projection in which foreshortening is characterized but the effects of line convergence are ignored. Affine invariance immediately suffers from its approximation: any associated models cannot distinguish simple shapes like rectangles because they are all linear (affine) transforms of each other. It is also possible to impose projective transforms on objects that leave them unrecognizable – suggesting that we do not use projective invariances for object recognition. In their stead, a number of authors have proposed that we make use of perspective invariance (Ref. k).

The types of projective invariances that authors refer to include: the fact that a triangle remains a triangle from all but the most pathological viewing directions; that parallel lines remain parallel; or that a circle always transforms to an ellipse. In order to recover true shape from a projected image one must decide how much foreshortening has taken place. Obviously, the problem has a free parameter, namely rotation in depth, which without depth information must be inferred via some constraint. One could, for example, assume that the true shape corresponds to the form in which the ratio of an object's area to the square of its perimeter is maximized (Ref. l) or that the true form has bilateral symmetry (Ref. m).

Geon structural description

One development of the part-based approach to recognition is the 'Geon Structural Description' of Biederman (Ref. h). The geon theory once again proposes that objects are represented by explicit relationships of a small set of LEGO-like blocks, which Biederman calls 'geons' (see Fig. IA). 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 volumetric primitives (36 in Biederman's opinion) can be used to describe any objects. This representation would then support invariant recognition so long as the parts remain in the same relative configuration, and are all visible.

Despite the similarity to other structural descriptions, geon theory also owes something to the idea of projective invariants. Biederman's approach specifically excludes any textural or similar depth cues, concentrating instead on the mapping of 2-D space relations to inferences about 3-D shape. Biederman describes a list of non-accidental 2-D properties in his discussion of how 2-D cues, such as collinearity, skew symmetry and coincident line termination, can be used to infer 3-D shape, and how they are critically important cues to recognition. The fact that features such as co-linearity and skew-symmetry can be extracted directly from a 2-D image without any recourse to 3-D modeling sets Biederman's approach apart from the fully 3-D approaches described earlier. The main weaknesses of this theory are that there remains little or no neurophysiological evidence for the explicit encoding of spatial relation or the representation of geon primitives, and that several psychophysical studies have revealed view-dependent recognition, even using basic geon shapes (Refs n–p).

Active shape matching

Another alternative to have received consideration is shape or template matching. The precise details of how to implement such a system vary considerably, but in practice all matching approaches fall into one of two conceptually important types. The first type argues that stored models should contain explicit

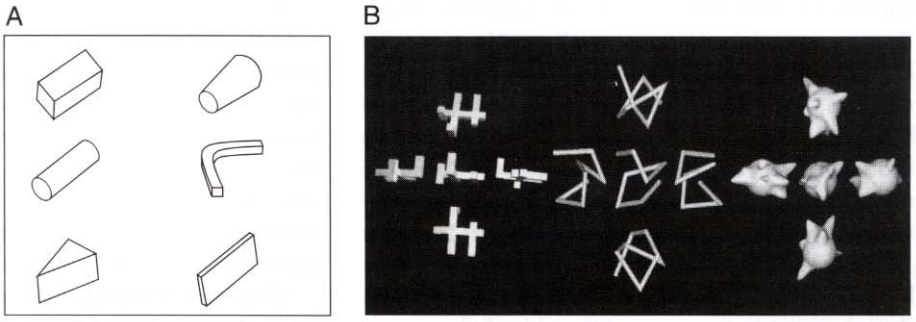

Fig. I. Volumetric primitives or views? Objects used in recognition paradigms. **(A)** 'Geons': volumetric building blocks used to describe everyday objects. **(B)** Aerials, Paperclips, and Amoebas: novel object categories used to test recognition performance for novel views.

within that class are represented. Indeed, Schyns argues that sufficient exposure to a particular stimulus type causes the representation of these stimuli to alter and be enhanced[28,29]. 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 an increase in the number of cells selective for those images or objects[16–18]. 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 his articles[28,29]. Indeed, in another recent article, Gauthier and Tarr have also made this point and shown that experience of an originally novel object-class heightens the subjects' awareness of small changes to objects within the same class[30].

Gauthier and Tarr go on to argue that this type of specialization underlies our highly sophisticated ability to

3-D shape information, whereas the second uses groups of 2-D views. One example of the 2-D approach is elastic pattern matching in which a non-linear image transformation is made of the stimulus in the image plane (or equivalently of the stored object views). 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 include Gabor-like patches or 'jets' (Ref. q) and edge-based facial features like ovals for eyes and a triangle for a nose (Ref. r).

The 3-D approach assumes that it is possible to extract the location of three (or more) anchor points in 3-D space, which are matched to those in the stored models. Matching the anchor points requires a 3-D 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 (Ref. s).

Recognition based on 2-D image features

Although the recognition of familiar, everyday objects proceeds almost effortlessly, some views are generally easier to recognize than others, both in terms of reaction times and accuracy. Such views are referred to as 'canonical' in the recognition literature (Ref. t).

Many researchers have since studied view specificity using novel objects trained in particular views (Fig. IB). They all point to a decrease in recognition performance as a function of viewpoint's disparity from a previously learned view (Refs u–x). Similar drops in recognition performance with viewing angle, have also been reported for unfamiliar objects (Ref. y) and faces (Ref. z).

These results have led to a new alternative for how objects are represented and recognized, namely the feature-based, multiple-view approach (Ref. w). It bears some relation to earlier 2-D matching theories and similarly benefits from the result of Ullman and Basri (Ref. aa) that any 2-D projection of a 3-D object can be written as a *linear* combination of 2-D views. However, the multiple-views model differs from that of the classical 2-D 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. The approach also represents a significant departure from object based models, as it requires neither the extraction of depth information, nor the exhaustive matching of 3-D models.

References

a Pentland, A.P. (1986) Perceptual organization and the representation of natural form *Artif. Intell.* 28, 293–331

b Lehky, S.R. and Sejnowski, T.J. (1988) Network model of shape from shading: neural function arises from both receptive and projective fields *Nature* 333, 452–454

c Koenderink, J.J. and van Doorn, A.J. (1975) Invariant properties of the motion parallax field due to the movement of rigid bodies relative to an observer *Optica Acta* 22, 773–791

d Hochberg, J. and Brooks, V. (1960) The psychophysics of form: reversible perspective drawings of spatial objects *Am. J. Psychol.* 73, 337–354

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

f Marr, D. and Nishihara, H.K. (1978) Representation and recognition of the spatial organization of three dimensional structure *Proc. R. Soc. London Ser. B* 200, 269–294

g Tversky, B. and Hemenway, K. (1984) Objects, parts and categories *J. Exp. Psychol. Gen.* 113, 169–193

h Biederman, I. (1987) Recognition by components: a theory of human image understanding *Psychol. Rev.* 94, 115–147

i Cutting, J.L. (1986) *Perception with an Eye for Motion*, MIT Press

j Rivlin, E. and Weiss, I. (1997) Deformation invariants in invariant object recognition *Comput. Vis. Image Understand.* 65, 95–108

k Pizlo, Z. (1994) A theory of shape constancy based on perspective invariants *Vis. Res.* 34, 1637–1658

l Brady, M. and Yuille, A. (1984) An extremum principle for shape from contour *IEEE Trans. Pattern Anal. and Machine Intell.* 6, 288–301

m Vetter, T., Poggio, T. and Bülthoff, H.H. (1994) The importance of symmetry and virtual views in 3-dimensional object recognition *Curr. Biol.* 4, 18–23

n Tarr, M.J. and Bülthoff, H.H. (1995) Is human object recognition better described by geon-structural-descriptions or by multiple-views? *J. Exp. Psychol. Hum. Percept. Perform.* 21, 1494–1505

o Tarr, M.J. *et al.* (1998) Three-dimensional object recognition is viewpoint dependent *Nature Neurosci.* 1, 275–277

p Edelman, S. (1997) Computational theories of object recognition *Trends Cognit. Sci.* 1, 296–304

q Lades, M. *et al.* (1993) Distortion invariant recognition in the dynamic link architecture *IEEE Trans. Comput.* 42, 300–311

r Yuille, A.L. (1991) Deformable templates for face recognition *J. Cogn. Neurosci.* 3, 59–71

s Ullman, S. (1979) *The Interpretation of Visual Motion*, MIT Press

t Palmer, S.E., Rosch, E. and Chase, P. (1981) Canonical perspective and the perception of objects, in *Attention and Performance* (Vol. IX) (Long, J. and Baddeley, A., eds), pp. 131–151, Erlbaum

u Shepard, R.N. and Cooper, L.A. (1982) *Mental Images and their Transforms* (3rd edn), MIT Press

v Rock I. and DiVita, J. (1987) A case of viewer-centered object perception *Cognit. Psychol.* 19, 280–293

w Bülthoff, H.H. and Edelman, S. (1992) Psychophysical support for a two-dimensional view interpolation theory of object recognition *Proc. Natl. Acad. Sci. U. S. A.* 92, 60–64

x Tarr, M.J. and Pinker, S. (1989) Mental rotation and orientation-dependence in shape recognition *Cognit. Psychol.* 21, 233–282

y Jolicouer, P. (1990) Orientation congruency effects on the identification of disoriented shapes *J. Exp. Psychol. Hum. Percept. Perform.* 16, 351–364

z Troje, N.F. and Bülthoff, H.H. (1996) Face recognition under varying poses: the role of texture and shape *Vis. Res.* 36, 1761–1771

aa Ullman, S. and Basri, R. (1991) Recognition by linear combinations of models *IEEE Trans. Pattern Anal. and Machine Intell.* 13, 992–1005

recognize faces, which they regard as no more than a highly trained object class. Some researchers would argue that this is not the case, and that face recognition is special. One of the main motivations for this has been the existence of the neurological disorder prosopagnosia. Prosopagnosia is characterized by a normal ability to recognize common objects, contrasted with extreme difficulty in recognizing people's faces[31]. The fact that the locus of the brain damage in patients pointed to a part of the temporal lobe homologous to that of face cells in monkeys[6,8], made a strong case for the suggestion that prosopagnosia was caused by damage to face cells[2]. Psychological studies have revealed a dissociation between face and object recognition in the past[32,33], but the latest neurophysiological evidence is not as clean as some theorists had first hoped. Direct attempts to find the elusive area responsible for face recognition in monkeys has been controversial and until now unfruitful[7,34]. This in turn lends more weight to Tarr and Gauthier's proposal that

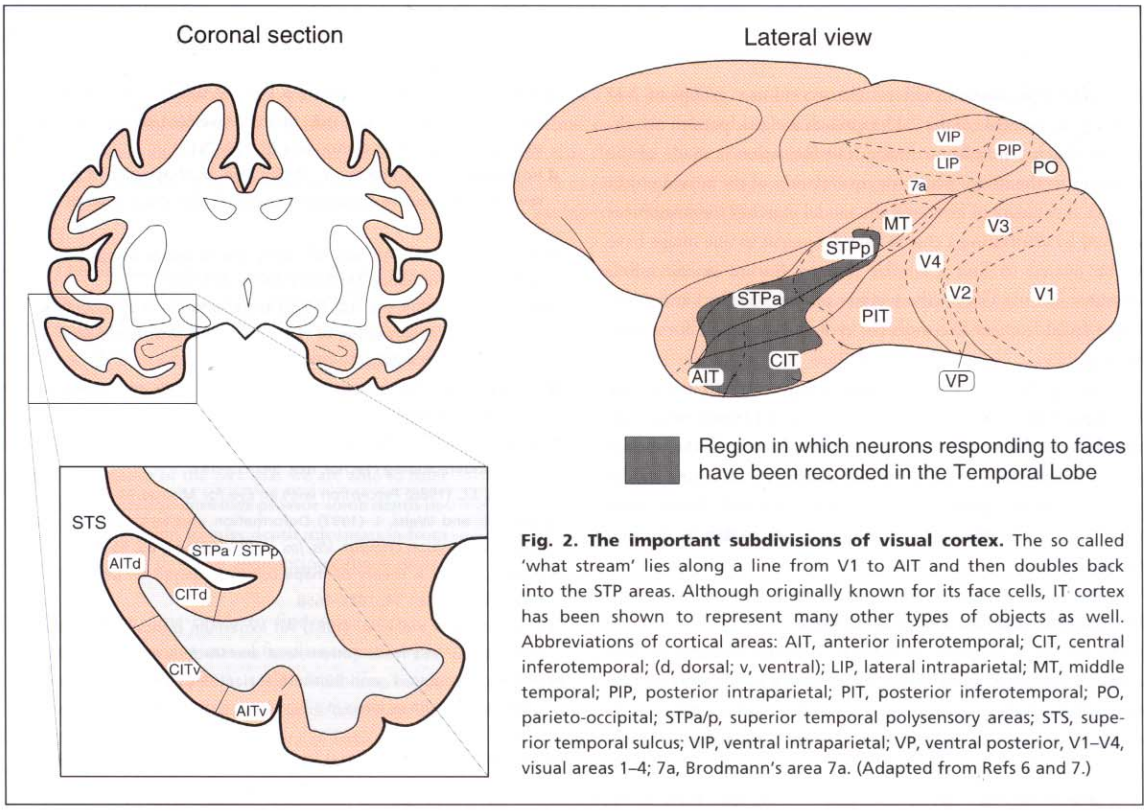

Fig. 2. The important subdivisions of visual cortex. The so called 'what stream' lies along a line from V1 to AIT and then doubles back into the STP areas. Although originally known for its face cells, IT cortex has been shown to represent many other types of objects as well. Abbreviations of cortical areas: AIT, anterior inferotemporal; CIT, central inferotemporal; (d, dorsal; v, ventral); LIP, lateral intraparietal; MT, middle temporal; PIP, posterior intraparietal; PIT, posterior inferotemporal; PO, parieto-occipital; STPa/p, superior temporal polysensory areas; STS, superior temporal sulcus; VIP, ventral intraparietal; VP, ventral posterior, V1–V4, visual areas 1–4; 7a, Brodmann's area 7a. (Adapted from Refs 6 and 7.)

prosopagnosia might reveal a general deficit in the area dedicated to fine-level discriminations of objects to which we are highly trained[30].

Representation through image features

The view-based approach to object recognition accords well with a large portion of the available neurophysiological data on face cells. However, the precise nature of this representation remains as yet unclear. Although there is good evidence that neurons represent faces through some form of distributed representation, there is neurophysiological evidence that this is sometimes at the level of complete views[35–37] and sometimes at the level of facial features[38–40]. Representation in the form of complete views is very similar to classical 2-D template models, whereas a feature-based system would be quite distinct. Different types of object encoding are summarized in Box 2, along with a description of what is meant by a 'feature' in this case.

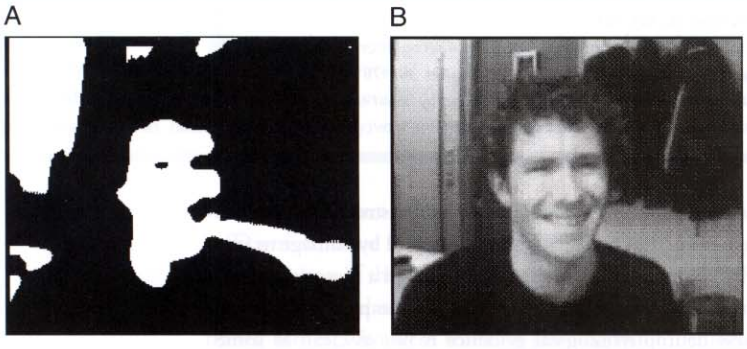

Fig. 3. A Mooney face. (A) If subjects or face-selective neurons are exposed to the two-tone image, they often fail to see a face. **(B)** Upon seeing the veridical image, both neurons and subjects now identify the face and will continue to do so in the future, providing evidence for rapid and lasting learning.

The results of Bülthoff and Edelman are actually best explained in terms of a feature-based neural code. For example, one can argue that through training, neurons become tuned to the features present in certain views of an object, and that in turn this pattern of firing becomes associated with the identity of that object. The interpolation result then follows if one assumes that two views have been learned, and that both are identified with the same object. Identification will clearly be easiest for views of the object nearest those trained, since this view is most likely to contain one or more of the features supporting the representation of the learnt views. It also follows that 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 falling outside that range, and is therefore more likely to activate one or more of the stored object features than a view outside this range. The rise in reaction times away from trained views follows the same logic.

It turns out that a view-based, distributed representation can also explain several other well known phenomena in object recognition. For example, evidence from neurophysiology suggests that populations of neurons are trained to recognize extreme views of faces[38] (e.g. frontal and profile), which seems to conflict with the good recognition performance reported for faces in 3/4 view[41,42]. However, using a distributed representation, recognition of the 3/4 view would be mediated by partial excitation of both frontal and profile neurons, thereby preventing any drop in reaction time or recognition accuracy[42].

Use of a view-based representation might also be able to explain one important ability of human subjects, which

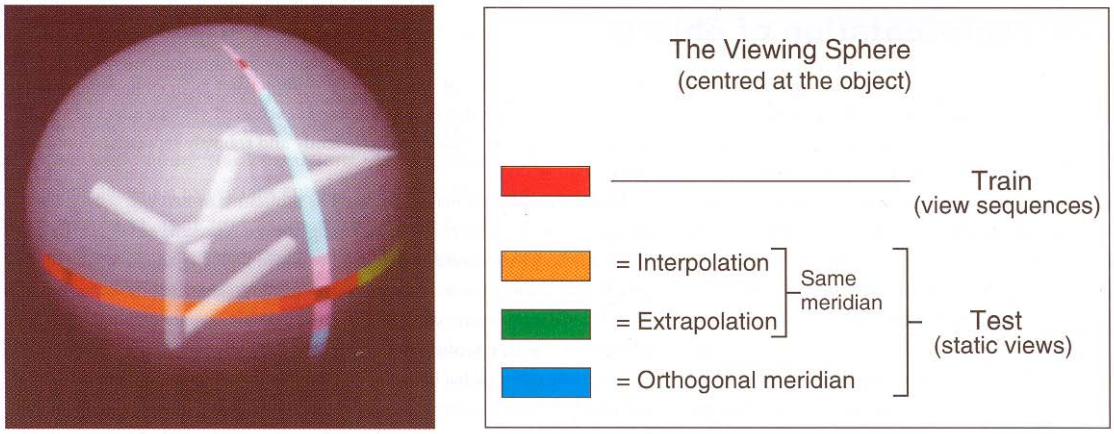

Fig. 4. Generalization to novel views. If two views of a novel object are learned, recognition is better for new orientations located between the two training views (Inter-polation) than outside them (Extra-). Recognition of the latter is better than for orientations away from the axis linking the trained views (Orthogonal) (see Refs 19 and 22).

until now was held as proof for some internal 3-D model. This is our ability to mentally rotate objects, i.e. to imagine how they would appear from a different viewpoint. Shepard and Cooper showed that the time required to recognize an object from a new viewpoint correlates with the view's disparity from a previously learned view[43]. Many interpreted this as evidence for the presence of a rotatable, internal 3-D model. However, it turns out that even here a distributed, view-based representation can provide an explanation for the results. Perrett *et al.*[44] discuss how the number of features supporting recognition drops with the difference in viewing angle between the two views. Since a disparate view will only activate a few neurons it will take time for this activity to raise following neurons to their threshold of firing. Response times would hence be correlated to the difference in angle between the trained and test views[44].

Despite the advantages of a feature-based system, there is some psychophysical evidence that we represent faces holistically[45,46] which supports those neurophysiological findings that report neurons responsive to complete views. On the other hand, there is also evidence that the feature-based approach can better explain the results of other recognition tasks. For example, Solso and McCarthy conducted an experiment in which subjects were presented with photo-fit pictures of people and then tested on a familiarity task[47]. 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 the salience of the latter group not only relative to the totally novel faces, but to the familiar faces as well. In other words, completely unfamiliar faces were regarded as very familiar, simply because features within these faces had been seen before.

It is possible that the mixed type of holistic and feature-based representations are common to all objects and not just faces. It may be the case, for example, that familiar objects are afforded larger neural reserves in the manner described in the previous section, and that the level to which an object is represented as features or as a more integrated representation closer to a holistic template, is governed by the level of exposure to that object class.

Temporal continuity as a cue to invariance learning

A broadly tuned feature-based system of the type under consideration in this review, would be sufficient to perform recognition over small transformations[48]. However, associating images over larger shape transformations either requires separate pre-normalization for size and translation of the image, or the use of separate view-specific feature detectors that would then feed into a view-invariant detector. The use of pre-normalization is at odds with the available neurophysiological evidence, which instead points to a gradual development of invariance over many processing stages, culminating in the types of cell responses found in inferior temporal cortex[6,36]. The feed-forward model also fits with the results of Perrett *et al.*[7] that response latencies for view-dependent cells are shorter than for view-invariant cells, which also accords with the earlier suggestion that view-invariant cells in STPa might pool the outputs of view-selective AIT cells[38].

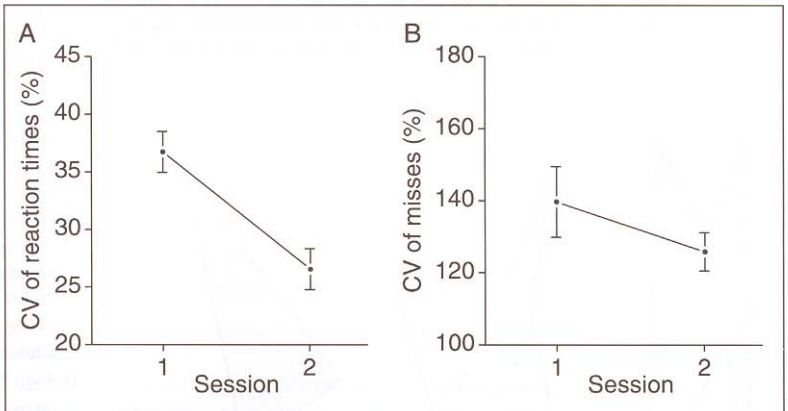

Fig. 5. Loss of canonical view effects. After subjects train on large numbers of views of novel objects, the shape of their recognition curves changes. Not only do reaction times decrease and accuracy increase, but view-specific effects, such as canonicality, gradually disappear. In this figure, the effect is quantified in terms of the coefficient of variation (CV), a measure that relates the amount of variability or range of a particular variable (be it reaction time or hit rate) to that recorded across all views. **(A)** CV of reaction times for paper clips seen from many viewpoints. The decrease in variation from one session to the next indicates a drop in view-specific effects with increased exposure to the object, providing evidence of object-specific learning. **(B)** CV of error rate (misses) for the same paper clips. The slight decrease in variation from one session to the next indicates that the decrease in reaction time is not due to a speed–accuracy trade-off.

Box 2. The neural representation of objects

Early theories of object recognition built on the assumption that the goal of the visual system was to represent a single object with the response of a single neuron. Such cells were given names such as 'Gnostic units' (Ref. a) or 'cardinal cells' (Ref. b) but are more often referred to as 'grandmother neurons', on the basis of a fundamental criticism of such encoding schemes. Simply put, the scheme in practice would be highly susceptible to cell damage. If you lose your grandmother-recognizing cell, then she will appear unfamiliar the next time you see her. The scheme is also highly inefficient, requiring impractically large numbers of uniquely stimulated neurons to represent the totality of real world objects. Alternative distributed codes have therefore been proposed, although a fully distributed encoding is wasteful in energy terms (Ref. c) and even more disastrous sensitivity to cell damage (Ref. d) (see Fig. I).

There is a common misconception that the discovery of face-selective cells in the cortex is linked to the earlier single cell theories of object representation. However, the truth is that in general, face cells respond not to one specific face, but to a subset of all faces (Ref. e). It is also not suggested that the cortex is full of face-selective neurons. There are perhaps 20 times as many neurons in temporal lobe cortex that do not respond to faces. These neurons appear to be selective for more abstract, complex visual stimuli (Ref. f) or for elements of other familiar objects (Refs g,h). As a consequence, it has been suggested that object encoding is achieved via ensembles of simultaneously firing cells, which both efficiently and robustly represent all objects (Refs i–l).

Under this scheme, many hundreds or thousands of neurons – each selective for its specific feature – would act together to represent an object. In this instance, the term 'feature' is intended to mean any diagnostic combination of light and shade, colour, form, etc., such as those described by Tanaka *et al.* (Ref. f). Although many of these features appear to represent only small regions of an object, it is also conceivable that they respond to the object's outline, or some other global but general property. In addition, the neural representation of these features is more sophisticated than a simple template, as they can exhibit invariance to scale or size, etc., as is typical of neurons in IT cortex.

New combinations of these features could then be recruited to uniquely represent a completely new object. In so doing, each neuron would bring knowledge of how its feature changes in appearance with changes in viewpoint. For example: Although a face may be new, experience with that type of nose or configuration of mouth and eyes, would provide some level of generalization for the face across view change. Indeed, the numerous beneficial, emergent properties of a distributed representation have long been realized by neural network theorists (Refs m,n).

This type of representation also provides a counter to the criticism of theorists to the view-based approach to object recognition. For example, Pizlo dismisses the idea of a multiple view representation on the grounds that novel shapes could not be correctly associated together, without prior exposure to multiple views of that shape (Ref. o). This he refutes with examples of the type shown in Fig. II. If views are represented as many features rather than single images, then this criticism no longer applies.

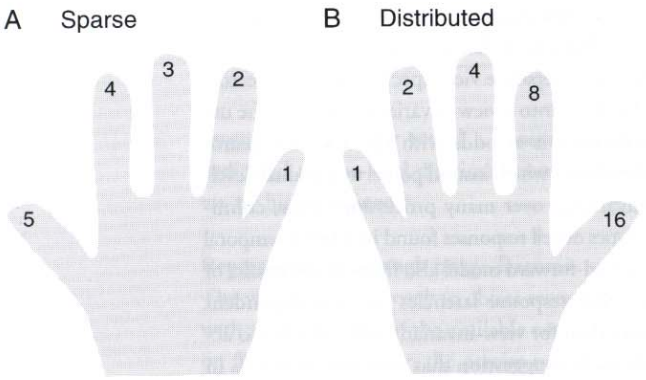

Fig. I. What you can count on the fingers of one hand. Two means of encoding information. **(A)** Sparse coding: each finger represents an object yielding five in total. **(B)** Distributed coding: using combinations of fingers a binary code can represent up to 31 objects. Two hands could represent 1023. In practice, both forms of encoding are susceptible to cell damage and the brain appears to compromise between the two.

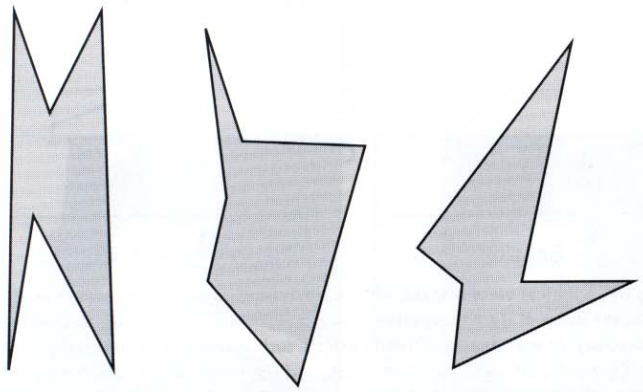

Fig. II. Recognition at first sight. The figure on the left is probably novel to all readers; despite this fact we have no difficulty in deciding which of the two other images show the same object from a second viewpoint. Hence, some generalization to novel views is immediate, even for novel objects. This type of immediate generalization is possible using a view-based representation, if each view comprises a collection of many subfeatures.

References

a Konorski, J. (1967) *Integrative Activity of the Brain: An Interdisciplinary Approach*, University of Chicago Press

b Barlow, H.B. (1972) Single units and sensation: a neuron doctrine for perceptual psychology? *Perception* 1, 371–394

c Baddeley, R. (1996) Visual perception: an efficient code in V1? *Nature* 381, 560–561

d Barlow, H. (1995) The neuron doctrine in perception, in *The Cognitive Neurosciences* (Gazzaniga, M.S., ed.), pp. 415–435, MIT Press

e Baylis, G.C., Rolls, E.T. and Leonard, C.M. (1985) Selectivity between faces in the responses of a population of neurons in the cortex in the superior temporal sulcus of the monkey *Brain Res.* 342, 91–102

f Tanaka, K. *et al.* (1991) Coding visual images of objects in the inferotemporal cortex of the macaque monkey *J. Neurophysiol.* 66, 170–189

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

h Logothetis, N.K., Pauls, J. and Poggio, T. (1995) Shape representation in the inferior temporal cortex of monkeys *Curr. Biol.* 5, 552–563

i Perrett, D.I., Mistlin, A.J. and Chitty, A.J. (1987) Visual cells responsive to faces *Trends Neurosci.* 10, 358–364

j Young, M.P. and Yamane, S. (1992) Sparse population coding of faces in the inferotemporal cortex *Science* 256, 1327–1331

k Rolls, E.T. (1992) Neurophysiological mechanisms underlying face processing within and beyond the temporal cortical areas *Philos. Trans. R. Soc. London Ser. B* 335, 11–21

l Abbott, L.F., Rolls, E.T. and Tovee, M.J. (1996) Representational capacity of face coding in monkeys *Cereb. Cortex* 6, 498–505

m Hinton, G.E., McClelland, J.L. and Rumelhart, D.E. (1986) *Distributed representations, Parallel Distributed Processing* (Vol. 1) (Rumelhart, D.E. and McClelland, J.L., eds), pp. 77–109, MIT Press

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

o Pizlo, Z. (1994) A theory of shape constancy based on perspective invariants *Vis. Res.* 34, 1637–1658

The only problem with such an approach is how disparate views can be associated by a neuron seeking to build an invariant representation of an object feature.

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 might act as an important cue for predicting the identity of an object as it undergoes transformations due to change of viewing-position relative to the object. This change in viewing-position can be the result of our approaching the object, watching it move, rotating it in our hand, and so on. The prediction that temporal correlations in image appearance affect the representation of objects, is clearly testable. For example, 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[6,36,49] and has been successfully used in various neural network models of recognition[49–51]. In particular, Wallis and Baddeley demonstrated how the temporal statistics of the real world can be used optimally, to establish transform-invariant representations of objects, using a biologically realizable learning rule[52].

The influence of image order presentation has also found support from neurophysiological studies[53,54]. Miyashita tested monkeys on a sequential match to sample task, in which he repeatedly displayed a set sequence of random fractal images[16]. He later found that if an IT cell responded to one stimulus in the series very strongly, it also responded to neighboring patterns in the sequence as well. Hence he was able to show that the efficacy of a stimulus dropped purely as a function of temporal and not spatial disparity between the stimuli. Of course, the match to sample task is somewhat removed from normal object viewing, but a link to normal object invariance learning has since been established[55].

Until recently, there was little or no psychophysical evidence to support the theoretical and neurophysiological findings. However, Sinha and Poggio recently described the use of sequences to establish the perception of the form of ambiguous wire-frame objects[56], and Wallis addressed the question directly, by considering the effect of temporal sequences for natural objects such as faces[57]. Wallis hypothesized that by exposing observers to sequences of different faces, they could confuse the identity of faces since the views would be falsely associated to form a new composite face. The results did indeed reveal poorer discrimination for faces associated in sequences, and that this effect increased with each session of training[57].

In another recent article, Stone presented temporal sequences of amoeba-like shapes[58], similar to those of Edelman and Bülthoff[27]. The experiment was divided into two phases. In the first phase subjects had to distinguish objects shown rotating anti-clockwise. In the second phase some of the familiar objects were presented rotating clockwise, which reduced the subjects' discrimination performance for these objects. This is similar to the earlier two results, but also distinct, in that it suggests that temporal information forms part of the representation of the object. This may have bearing on the results from the biological motion literature, that the motion of abstract dots can evoke recognition of a moving creature[56,59].

Whichever approach to object recognition one prefers, the data discussed here require that they be modified to include the use of temporal sequences either in setting up the representations, or as an integral part of the representation. It is also important to realise that 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 justly criticised alternatives, such as supervised training 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 understanding object recognition learning within the scope of a view-based representation.

Outstanding questions

- Assumptions (i.e. priors) are known to play an important role in perception[60] But what specific role do they play in object recognition? The temporal-association hypothesis is one example, but there might be others. To what extent are these priors learnt, and to what extent innate?
- This review has drawn much of its evidence from work on neurons responsive to faces. But to what extent are faces and objects related?
- Why do some cells represent faces holistically[35–37], and others as features[38–40]? Does this underlie a process in recognition by which very familiar objects are stored in a more holistic manner in later parts of IT cortex[28,30]? Or have we evolved a special processing systems for faces because they are difficult to discriminate and are socially important[8].
- If temporal associations affect object recognition, what element of the temporal nature is encoded? Is there a temporal signature stored, as Stone[58] suggests. Or is the temporal signal merely used to help establish spatial representations[56,57]? Indeed, if both spatial and temporal cues are used in setting up representations of objects, under what circumstances would one prefer to use one than the other? How might these cues be used in combination?
- Although there is some evidence that neurons in IT cortex associate images on the basis of temporal continuity, the time period over which this association takes place is too long to be of relevance to normal recognition learning[16,55]. Do neurons in IT cortex associate views of objects when they appear in short smooth sequences, as the temporal association hypothesis would predict?
- Most of the work done on recognition has been restricted to isolated shapes displayed before a blank background. How does the representation of a scene differ from that of an object, if at all? To what extent can the view-based approach be extended to the representation of places, for use in tasks like navigation[61].

Conclusion

Our intention in this paper has been to draw together much of the research currently underway in the field of object recognition, and to highlight the encouraging parallels between neurophysiological and psychophysical evidence in this field. In the main body of the article we have concentrated on the questions of whether and how representations of objects are learnt, reviewing studies ranging from adaptation to Mooney faces, to the fall in canonical view effects with experience. We have also sought to describe how a feature-based approach to recognition would work and the manner in which the choice of stored features develops with experience.

The final section described how we may draw on the temporal structure of our environment to learn more difficult object transformations. The simple but effective heuristic makes use of the fact that potentially very different images appearing in close temporal succession are likely to be views of the same object. This piece of information about environmental structure then takes the form of a tendency (a 'prior' in the sense used in Bayesian statistics) of the human visual system to associate images of objects together over short periods of time. Evidence for this hypothesis stems both from neurophysiological and human psychophysical studies.

Taken as a whole, the results described in this paper strongly support the empiricist view that object recognition and categorization is largely an ongoing process, affected by experience of our environment. By using novel stimuli it has been possible for researchers to describe how object representation and recognition develops with experience in normal subjects. In this manner the results serve to support the proposal that much of perception is mediated via a dynamic learning system, the modification of which continues throughout our lives.

References

1 Gregory, R.L. and Wallace, J.G. (1963) *Recovery from Early Blindness: a Case Study* (Exp. Psychol. Soc. Monogr. No. 2), W. Heffer & Sons

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

3 Ungerleider, L.G. and Mishkin, M. (1982) Two cortical visual systems, in *Analysis of Visual Behavior* (Ingle, D.J., Goodale, M.A. and Mansfield, R.J.W., eds), pp. 549–586, MIT press

4 Goodale, M.A. and Milner, A.D. (1992) Separate visual pathways for perception and action *Trends Neurosci.* 15, 20–25

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

6 Rolls, E.T. (1992) Neurophysiological mechanisms underlying face processing within and beyond the temporal cortical areas *Philos. Trans. R. Soc. London Ser. B* 335, 11–21

7 Perrett, D.I. *et al.* (1992) Organisation and functions of cells responsive to faces in the temporal cortex *Philos. Trans. R. Soc. London Ser. B* 335, 23–30

8 Desimone, R. (1991) Face-selective cells in the temporal cortex of monkeys *J. Cogn. Neurosci.* 3, 1–8

9 Gross, C.G., Rocha-Miranda, C.E. and Bender, D.B. (1972) Visual properties of neurons in inferotemporal cortex of the macaque *J. Neurophysiol.* 35, 96–111

10 Logothetis, N.K. and Sheinberg, D.L. (1996) Visual object recognition *Annu. Rev. Neurosci.* 19, 577–621

11 Tanaka, K. *et al.* (1991) Coding visual images of objects in the inferotemporal cortex of the macaque monkey *J. Neurophysiol.* 66, 170–189

12 Fuster, J.M. and Jervey, J.P. (1982) Neuronal firing in the inferotemporal cortex of the monkey in a visual memory task *J. Neurosci.* 2, 361–375

13 Perrett, D.I. *et al.* (1991) Viewer-centered and object-centered coding of heads in the macaque temporal cortex *Exp. Brain Res.* 86, 159–173

14 Baylis, G.C., Rolls, E.T. and Leonard, C.M. (1985) Selectivity between faces in the responses of a population of neurons in the cortex in the superior temporal sulcus of the monkey *Brain Res.* 342, 91–102

15 Rolls, E.T. *et al.* (1989) The effect of learning on the face-selective responses of neurons in the cortex in the superior temporal sulcus of the monkey *Exp. Brain Res* 76, 153–164

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

17 Kobatake, E., Wang, G. and Tanaka, K. (1998) Effects of shape-discrimination training on the selectivity of inferotemporal cells in adult monkeys *J. Neurophysiol.* 80, 324–330

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

19 Bülthoff, H.H. and Edelman, S. (1992) Psychophysical support for a two-dimensional view interpolation theory of object recognition *Proc. Natl. Acad. Sci. U. S. A.* 92, 60–64

20 Tovee, M.J., Rolls, E.T. and Ramachandran, V.S. (1996) Rapid visual learning in neurones of the primate temporal visual cortex *NeuroReport* 7, 2757–2760

21 Ramachandran, V.S. (1994) 2D or not 2D – that is the question, in *The Artful Eye* (Gregory, R. and Harris, J., eds), Oxford University Press

22 Tarr, M.J. and Pinker, S. (1989) Mental rotation and orientation-dependence in shape recognition *Cognit. Psychol.* 21, 233–282

23 Palmer, S.E., Rosch, E. and Chase, P. (1981) Canonical perspective and the perception of objects, in *Attention and Performance* (Vol. IX) (Long, J. and Baddeley, A., eds), pp. 131–151, Erlbaum

24 Koenderink, J.J. and van Doorn, A.J. (1979) The internal representation of solid shape with respect to vision *Biol. Cybern.* 32, 211–216

25 Biederman, I. (1987) Recognition-by-components: a theory of human image understanding *Psychol. Rev.* 94, 115–147

26 Jolicoeur, P. (1985) The time to name disoriented natural objects *Mem. Cognit.* 13, 289–303

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

28 Schyns, P.G., Goldstone, R.L. and Thibaut, J-P. (1998) The development of features in object concepts *Behav. Brain Sci.* 21, 1–54

29 Schyns, P.G. (1997) Categories and percepts: a bi-directional framework for categorization *Trends Cognit. Sci.* 1, 183–189

30 Gauthier, I. and Tarr, M. Becoming a Greeble expert: exploring mechanisms for face recognition *Vis. Res.* 37, 1673–1682

31 De Renzi, E. (1997) Prosopagnosia, in *Behavioural Neurology and Neuropsychology* (Feinberg, T.E. and Farah, M.J., eds), pp. 245–255, McGraw-Hill

32 Tanaka, J.W. and Farah, M.J. (1993) Parts and wholes in face recognition *Q. J. Exp. Psychol. Ser. A* 46, 225–245

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

34 Cowey, A. (1992) The role of the face-cell area in the discrimination and recognition of faces by monkeys *Philos. Trans. R. Soc. London Ser. B* 335, 31–38

35 Rolls, E.T. *et al.* (1994) The responses of neurons in the temporal cortex of primates, and face identification and detection *Behav. Process.* 101, 473–484

36 Perrett, D. and Oram, M.W. (1993) Neurophysiology of shape processing *Image Vis. Comput.* 11, 317–333

37 Perrett, D.I. *et al.* (1984) Visual analysis of body movements by neurones in the temporal cortex of the macaque monkey: a preliminary report *Behav. Brain Res.* 16, 153–170

38 Perrett, D.I., Mistlin, A.J. and Chitty, A.J. (1987) Visual cells responsive to faces *Trends Neurosci.* 10, 358–364

39 Young, M.P. and Yamane, S. (1992) Sparse population coding of faces in the inferotemporal cortex *Science* 256, 1327–1331

40 Abbott, L.F., Rolls, E.T. and Tovee, M.J. (1996) Representational capacity of face coding in monkeys *Cereb. Cortex* 6, 498–505

41 Troje, N.F. and Bülthoff, H.H. (1996) Face recognition under varying poses: the role of texture and shape *Vis. Res.* 36, 1761–1771

42 Valentin, D., Abdi, H. and Edelman, B. (1997) What represents a face? a computational approach for the integration of physiological and psychological data *Perception* 26, 1271–1288

43 Shepard, R.N. and Cooper, L.A. (1982) *Mental Images and their Transforms* (3rd edn), MIT Press

44 Perrett, D., Oram, M.W. and Wachsmuth, E. Evidence accumulation in cell populations responsive to faces: an account of generalisation of recognition without mental transformations *Cognition* (in press)

45 Schwarzer, G. (1997) Development of face categorization: the role of conceptual knowledge *Space Cognit.* 16, 14–30

46 Carey, S. and Diamond, R. (1994) Are faces perceived as configurations

more by adults than by children? *Visual Cognit.* 1, 253–274

47 Solso, R.L. and McCarthy, J.E. (1981) Prototype formation of faces: a case of pseudo-memory *Br. J. Psychol.* 72, 499–503

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

49 Wallis, G. and Rolls, E.T. (1997) A model of invariant object recognition in the visual system *Prog. Neurobiol.* 51, 167–194

50 Edelman, S. and Weinshall, D. (1991) A self-organising multiple-view representation of 3D objects *Biol. Cybern.* 64, 209–219

51 Foldiak, P. (1991) Learning invariance from transformation sequences *Neural Comput.* 3, 194–200

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

53 Miyashita, Y. (1993) Inferior temporal cortex: where visual perception meets memory *Annu. Rev. Neurosci.* 16, 245–263

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

55 Wallis, G. (1998) Spatio-temporal influences at the neural level of object recognition *Netw. Comput. Neural Syst.* 9, 265–278

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

57 Wallis, G. (1998) Temporal order in object recognition learning *J. Biol. Syst.* 6, 299–313

58 Stone, J.V. (1998) Object recognition using spatio-temporal signatures *Vis. Res.* 38, 947–951

59 Johansson, G. (1973) Visual perception of biological motion and a model for its analysis *Percept. Psychophys.* 14, 201–211

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

61 Gillner, S. and Mallot, H.A. (1998) Navigation and acquisition of spatial knowledge in a virtual maze *J. Cogn. Neurosci.* 10, 445–463