Visual Recognition Based on Temporal Cortex Cells: Viewer-Centred Processing of Pattern Configuration*

David I. Perrett[a] and Mike W. Oram[b]

[a] Psychological Laboratory, St. Andrews University, Scotland, KY16 9JU, UK
[b] Building 49, Room 1B80, NIMH, NIH, 9000 Rockville Pike, Bethesda, MD 20892, USA

Z. Naturforsch. **53c,** 518–541 (1998); received May 11, 1998

Visual Recognition, Temporal Cortex Cells, Pattern Configuration

A model of recognition is described based on cell properties in the ventral cortical stream of visual processing in the primate brain. At a critical intermediate stage in this system, 'Elaborate' feature sensitive cells respond selectively to visual features in a way that depends on size (±1 octave), orientation (±45°) but does not depend on position within central vision (±5°). These features are simple conjunctions of 2-D elements (e.g. a horizontal dark area above a dark smoothly convex area). They can arise either as elements of an object's surface pattern or as a 3-D component bounded by an object's external contour. By requiring a combination of several such features *without regard to their position* within the central region of the visual image, 'Pattern' sensitive cells at higher levels can exhibit selectivity for complex configurations that typify objects seen under particular viewing conditions. Given that input features to such Pattern sensitive cells are specified in approximate size and orientation, initial cellular 'representations' of the visual appearance of object type (or object example) are also selective for orientation and size. At this level, sensitivity to object view (±60°) arises because visual features disappear as objects are rotated in perspective. Processing is thus viewer-centred and the neurones only respond to objects seen from particular viewing conditions or 'object instances'. Combined sensitivity to multiple features (conjunctions of elements) independent of their position, establishes selectivity for the configurations of object parts (from one view) because rearranged configurations of the same parts yield images lacking some of the 2-D visual features present in the normal configuration.

Different neural populations appear to be selectively tuned to particular components of the same biological object (e.g. face, eyes, hands, legs), perhaps because the independent articulation of these components gives rise to correlated activity in different sets of input visual features. Generalisation over viewing conditions for a given object can be established by hierarchically pooling outputs of view-condition specific cells with pooling operations dependent on the continuity in experience across viewing conditions. Different object parts are seen together and different views are seen in succession when the observer walks around the object. The view specific coding that characterises the selectivity of cells in the temporal lobe can be seen as a natural consequence of selective experience of objects from particular vantage points. View specific coding for the face and body also has great utility in understanding complex social signals, a property that may not be feasible with object-centred processing.

1. Introduction

Few of the existent models of visual recognition incorporate the range of physiological data now available concerning the higher stages of visual processing. This is due to several reasons. First, particular models have been designed for different purposes, for example, to achieve recognition by any means independent of biological plausibility (e.g. Principle Component Analysis of raw pixel intensity values without further visual processing, Sirovitch and Kirby, 1987; Turk and Pentland, 1991), or models are designed to account for particular psychological and/or neuropsychological observations (Warrington, 1982; Warrington and James, 1986; Weiskrantz and Saunders, 1984) or to explain stages of processing subsequent to visual encoding (e.g. Bruce and Young, 1986). Second,

* This communication is a contribution to the workshop on "Natural Organisms, Artificial Organisms, and Their Brains" at the Zentrum für interdisziplinäre Forschung (ZiF) in Bielefeld (Germany) on March 8–12, 1998.

Reprint requests to Prof. M. W. Oram.
Fax: 301-4020046.
E-mail: mike@ln.nimh.nih.gov.

0939–5075/98/0700–0518 $ 06.00

physiological data on feature sensitivity in higher association cortex of the temporal lobe (Kobotake and Tanaka, 1994; Tanaka *et al.,* 1991) and the range of viewing conditions that temporal lobe cells tolerate while maintaining sensitivity to complex patterns (Perrett *et al.,* 1984; 1991; Wachsmuth *et al.,* 1994; Rolls and Baylis, 1986; Tovee *et al.,* 1994; Logothetis *et al.,* 1995) has become available only relatively recently. Finally, authors base their models in areas of their own expertise. Physiologists rarely discuss psychological phenomena and even less frequently do psychologists refer to physiological data.

An account of object recognition is usually characterised by a statement of the nature of the problems that any recognition system needs to overcome. The major problem is that the system must cope with the change in viewing circumstance over which an object may be encountered. Marr and Nishihara (1978) drew attention to the need to retrieve experiences of an object's properties across different viewing circumstances and pointed out the benefits of having information about experiences linked to a single description of the object which could be accessed despite changes in perspective view, ambient lighting, part occlusion, size and orientation. Such object descriptions are termed object-centred representations (Marr and Nishihara, 1978). Although some accounts still maintain that recognition is independent of view (Biederman, 1987) many recent psychological and computational accounts of recognition utilise multiple viewing condition specific descriptions of the same object (Jolicour, 1992; Poggio and Edelman, 1990; Ullman, 1989; Seibert and Waxman, 1992a, b; Verfraillie, 1992; Logothetis *et al.,* 1994a, b). Such descriptions are termed viewer-centred representations. Multiple descriptions of the same object allow invariance in recognition across different views only when these descriptions are inter-linked or converge hierarchically (Koenderink and van Doorn, 1979; Seibert and Waxman, 1992a, b; Perrett and Oram, 1993; Oram and Perrett, 1994).

The ability of humans and other animals to cope with changes in viewing conditions has been reviewed elsewhere (e.g. Wachsmuth and Perrett, 1997). This literature can be summarised as indicating that, after learning the appearance of an object under one set of conditions, normal animals show only a limited ability to generalise recognition of the same object when viewing conditions change. The limitations in ability to generalise are not restricted to change in perspective view as might be surmised from many computational models but extend to change in orientation and size. The degree of success in generalisation is of course dependent on task difficulty and the similarity of target and non-target objects.

It is argued here that the problems of object invariance (recognising equivalence across image transformations) are largely side-stepped by the visual system. View, orientation and size generalisation are achieved by the visual system performing a separate analysis of the different instances in which an object is experienced. These initially separate analyses generalise to only a limited extent across change in image orientation and size. More substantial generalisation is achieved through associative learning when the different instances are seen in temporal succession as a continuous transform (Foldiak, 1991). This allows hierarchical pooling of the outputs of the early view specific analyses. Generalisation across position, by contrast, appears to be distinct: the visual system seems to achieve position invariance for visual features prior to generalisation over other image transformations.

This article provides an overview of the increase in the complexity of stimulus features required to elicit cell responses along the ventral cortical stream (section 2). We then consider how cells at the higher levels of visual processing generalise across changes in viewing conditions. The response selectivity is interpreted in terms of viewer-centred or object-centred processing. We conclude that processing is predominantly viewer-centred (section 3). The viewer-centred processing allows interpretation of social and other interactions, an ability that would not be possible using object-centred processing (section 4). In section 5 we argue that response selectivity to conjunction of features can overcome the "binding problem" for recognition of familiar objects. We continue by examining the relationship between an object's intuitive features and feature conjunctions (section 6). Some of the arguments used are based, at least in part, on studies of response selectivity to faces and bodies. Examination of psychological and physiological data suggests that striking similarities exist

between processing of faces and processing of other familiar objects (section 7). We end by comparing the presently proposed model with previous models (section 8) and consideration of the role of neural populations in the representation of objects (section 9).

2. Ventral Cortical Stream of Visual Processing

Early vision is characterised by neural populations detecting visual elements at very localised regions of the image. Areas V4, posterior and anterior inferior temporal cortex (PIT and AIT respectively) form a chain of areas at more anterior locations and a greater number of synapses from the retinal input (for review see Oram and Perrett, 1994). As the ventral stream is sampled along this stream of processing, there are systematic changes in stimulus selectivity displayed by cell responses and in the size of the visual space over which cells respond (Tanaka and *et al.,* 1991; Kobatake and Tanaka, 1994). We briefly review the changes in cell response characteristics along the ventral stream of processing in the following section.

2.1. Receptive field size

Receptive field size increases from V1 to V2. This trend continues into V4, where the average receptive field areas = 4 degree^2, to PIT (16 degree^2) and AIT (150 degree^2). Many cells along the V4, PIT and AIT pathway also have receptive fields close to or including the fovea (75% of AIT cells' receptive fields include the fovea). The size of receptive fields presumably increases because cells at higher levels pool the outputs of a number of cells in the preceding area. Pooling that is restricted to inputs from cells with the same type of feature selectivity would allow the system to maintain stimulus selectivity while generalising over position. Pooling different types of input would develop sensitivity to more complex patterns while generalising across position. The selection of one type of feature sensitive input by higher level cells can be guided by associative learning mechanisms coupled with exposure to translation of the relevant feature across the local receptive fields of the lower level cells (Foldiak, 1991; Oram and Foldiak, 1996). Such translation of visual features would occur as the observer makes tracking eye movements or moves through the environment.

2.2. Increasing complexity of effective stimuli for ventral stream neurones

In areas V4 and PIT the majority (70%) of cells are sensitive to the primary qualities of the stimuli, such as orientation, size and colour (Kobotake and Tanaka, 1994). About 5% of cells in these areas code for the texture of stimuli (for example, requiring a repeating stripe pattern or an array of spots). The most important types of cells for understanding shape and object recognition are those labelled by Tanaka *et al.* (1991) as 'Elaborate'. These cells are selective for stimulus shape or for shape in addition to other primary qualities. Elaborate cells are more frequent in AIT (45%) but are also present in PIT (9%) and even V4 (2%). Selectivity for elaborate stimulus features may first be defined at local retinal regions within area PIT (Kobotake and Tanaka 1994). The simplest type of Elaborate cell is selective for a single shape with different cells being most responsive to different shapes. Examples of different cell selectivities include: long and rounded, star shaped, elongated with sharp taper at both ends and circular with protruding element. Many Elaborate cells have specific shape requirements coupled with specific texture or colour requirements (e.g. green star shape or triangular with a spotted texture, Tanaka *et al.* (1991)). A further level of complexity is observed for some IT cells that are tuned to the visual characteristics of two image regions. Some Elaborate cells are selective for the specific relationships between two shapes (e.g. an opaque circle above a clear circle; a light rounded area at the base of a dark egg shape; a dotted brown disk attached to a narrow bar) or two textures (e.g. a horizontal striped area above a separate horizontal striped area).

2.3. Summary of ventral stream processing

Cortical areas PIT and AIT appear to be coding the 2-D appearance of features in the image in terms of shape, relationships between shapes, texture, colour and intensity gradients. The stimulus selectivity of cells in anterior areas becomes progressively more complex compared to cells in posterior areas (PIT/V4) and show increasing recep-

tive field sizes. The IT cortex has a columnar or modular structure, cells within one module tending to respond to similar visual features. The visual stimuli activating different modules are not necessarily equally complex as some modules may contain cells selective for more complex features than neighbouring modules (Perrett *et al.,* 1984; Harries and Perrett, 1991; Fujita *et al.,* 1992).

It is not difficult to imagine how the more complex stimulus requirements for neural responses could derive from a combination of outputs from cells with simpler response selectivity. For example, a cell sensitive to a triangular area containing a vertically oriented texture can be seen as combining inputs specifying triangular shape selectivity with other inputs specifying texture sensitivity. Likewise cells displaying elaborate sensitivity for the shape of two areas (e.g. a disk shape and elongated horizontal bar) can be seen as combining two separate shape descriptions available within IT. The cells displaying Elaborate shape selectivity should not be considered selective for an object concept (e.g. pineapple); instead the cells code simple qualities (green and star shape) which may be present in some objects but not in many others. Selectivity of these IT cells appears high, but not extreme, and the cells will presumably respond to all objects containing the requisite shape features.

2.4. Implications for models of visual recognition

It is interesting to note that particular IT cells are selective for intersecting edges of specific orientations (e.g. T junctions but not Y junctions nor the right angled components of the T shape, Tanaka *et al.* (1991)). The angles of vertices at T and Y junctions have often been considered important in artificial intelligence for automated classification of geometrical faceted objects (Guzman, 1968). The sensitivity to the curvature of convexities in shapes that is exhibited by IT cortical neurones is more specific than might be expected from psychological models of recognition. In the model of Biederman (1987), a set of 36 3-D shape components or geons are defined only qualitatively; one geon characteristic would be thin at one end, fat in the middle and thin again at the other end. Such qualitative descriptions may be sufficient for distinguishing different classes of object but they are insufficient for distinguishing within a class of objects possessing the same basic components (Saund, 1992; Perrett and Oram, 1993). For example all birds have beaks which taper gradually, yet the exact rate of taper and degree of taper may help differentiate closely related bird species. Single cell studies in IT cortex present direct evidence that shape and curvature are coded within the nervous system more precisely than would be expected from Biederman's recognition by components model (Tanaka *et al.,* 1991).

3. Viewer-Centred and Object-Centred Representations

Within the cortex of AIT, temporal pole (TG), AMTS (Anterior Medial Temporal Sulcus) and STS (Superior Temporal Sulcus), cells have been found which respond selectively to complex stimuli such as faces, hands, bodies (Gross *et al.,* 1972; Perrett *et al.,* 1982; 1989; Wachsmuth *et al.,* 1994) and arbitrary visual patterns that have become meaningful as a result of extensive training (e.g. fractal patterns, Miyashita and Chang, 1988; Miyashita *et al.,* 1993; or wire frame objects, Logothetis *et al.,* 1995). Such cells can exhibit highly selective responses, discriminating between exemplars within one class of objects. For example, some cells may respond to one or a small subset of faces, fractal patterns or wire frame objects (Perrett *et al.,* 1984; 1989; Miyashita and Chang, 1988; Logothetis *et al.,* 1995). It is assumed here that Pattern sensitive cells tuned to specific object types are at a stage of processing beyond the Elaborate feature sensitive cells of IT. Cells selective for unique patterns of objects are sensitive to specific combinations or configurations of greater numbers of visual component features than Elaborate cells (Perrett *et al.,* 1982; Yamane *et al.,* 1988; Tanaka *et al.,* 1991; Miyashita *et al.,* 1993; Wachsmuth *et al.,* 1994). In this section we review the sensitivity of the Pattern selective cells to changes in viewing conditions.

3.1. View

Initial accounts of cells responsive to faces stressed the degree of generalisation across face colour, size, orientation and sub-part. Even during the early studies it was apparent that cells were selective for view and failed to respond to the sight of the face rotated away from the subject by 90

degrees in depth (Perrett *et al.,* 1982). Subsequent studies revealed that multiple views of the same object (face, left profile, right profile, back of head, face turned up and turned down) excited different populations of cells (Desimone *et al.,* 1984; Perrett *et al.,* 1984; 1985; 1991).

Within STS and AIT a small percentage of cells (~5%) generalise across view in a way that cannot be attributed to response to a single feature common to all views (Perrett *et al.,* 1984; 1985; 1989; 1991; 1992; Hasselmo *et al.,* 1989; Logothetis *et al.,* 1995). We note that the proportion of view-invariant or object-centred cells may increase in areas that have not yet been studied extensively. Several authors have suggested that generalisation across view could be established through associative learning mechanisms pooling the outputs of view specific descriptions (Poggio and Edelman, 1990; Foldiak, 1991; Seibert and Waxman, 1992a, b; Vetter *et al.,* 1995; Perrett *et al.,* 1984; 1985; 1992; Perrett and Oram, 1993; Oram and Perrett, 1994). There is some evidence for such a hierarchical scheme since view sensitive responses to the sight of the head are shorter in latency than view general responses (Perrett *et al.,* 1992).

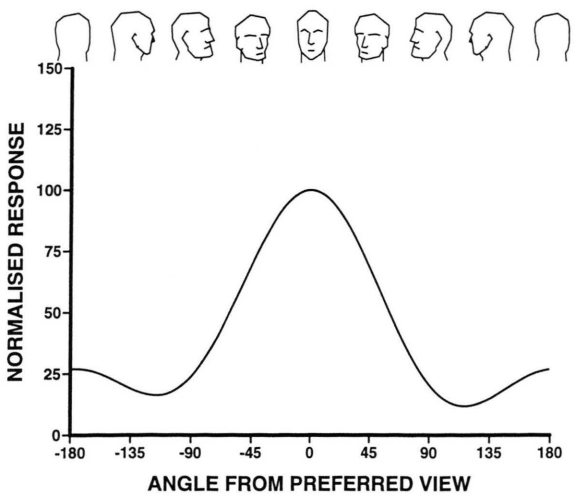

Fig. 1. Average tuning of STS neurones to change in perspective view. The average tuning curve was calculated for a population of 56 cells, each selective for one view of the head. Different cells were tuned to different head views but for computation of the average tuning curve, the best views were aligned before averaging. View is defined as angle of rotation from best view (arbitrarily defined as 0 degrees) and depicted schematically as the face view (data from Perrett *et al.* (1991)).

While a small percentage of cells show object-centred response selectivity, the majority of neurones show strong modulation of response with change in perspective view (Perrett *et al.,* 1991). The view-point selectivity may vary with object structure, complexity and defining features. For cells sensitive to the sight of the head and body in STS, responses typically decrease to ½ maximal rate with a perspective rotation of 60 degrees from the cells' preferred view (Perrett *et al.,* 1991). Figure 1 shows the average tuning curve for change in perspective view. Cells which have become tuned to complex wire frame objects as a result of extensive discrimination training, exhibit narrower view tuning with responses decreasing to ½ maximal for a 20–40 degree rotation of the stimulus object from the cells' preferred view (Logothetis *et al.,* 1995). This narrow tuning may reflect the difficulty in discrimination of target stimuli from non-targets and the use of features resolvable only with high sensitivity to orientation.

In conclusion, only a small percentage of Pattern sensitive cells respond to all views equivalently. The majority of response selectivity is viewer-centred in that the cells responded to only a limited number of views. Viewer-centred response selectivity is also seen for the appearance of learnt objects.

3.2. Lighting and contrast

The direction and the strength of ambient lighting have profound effects on the visual appearance of the face and body. Under very strong illumination the shadows cast by a single light source on a 3-D object such as a face alter dramatically with light position relative to the object (above, to the side or underneath). With less strong illumination, the features of the face may fall within one shadowed area, but features may still be visible as low contrast fluctuations within a low intensity (shadowed) facial region. Responses of STS neurones selective for faces are tolerant to changes in stimulus contrast (Rolls and Baylis, 1986). The cells also show extensive generalisation across different directions of strong illumination. Hietanen *et al.* (1992) compared responses to faces under normal (front) lighting with responses to faces viewed under unusual lighting (from the top, side and bottom). Hietanen *et al.* (1992) found that that many

(29%) tested STS cells selective for faces responded to all 4 tested directions of lighting. 57% of neurones failed to respond to the optimal view of the head under one or two of the unusual lighting directions but responded equally to the other two or three directions of lighting. The remaining 14% of cells displayed more selective responses and failed to respond to the face in the unusual lighting conditions. Figure 2 shows the population response to the different tested lighting conditions (adapted from Hietanen *et al.*, 1992).

The tolerance for direction of lighting (Hietanen *et al.*, 1992) and changes in input contrast (Rolls and Baylis, 1986) may depend on localised luminance and contrast normalisation occurring in the retina or lateral geniculate nucleus (e.g. Hubel and Livingstone, 1990). Luminance and contrast normalisation would facilitate simultaneous contour extraction across the whole image independent of the associated levels of light or contrast (Watt, 1996). Thus for lighting conditions and contrast level it appears that STS cells selective for the face and body show object-centred properties.

3.3. Object part visibility

A complication to recognition arises when part of an object is occluded from sight. Recognition of the entire pattern from a partial image is termed 'completion'. Cells responsive to faces vary in their generalisation to components of the face. Some cells are responsive to the mouth but not eyes and others show the converse selectivity (Perrett *et al.*, 1982). One might predict such component sensitivity given the selectivity of some Elaborate cells in IT and V4 to concentric circular features or features found in lips (Fujita *et al.*, 1992). Most cells respond to many face parts tested in isolation and some cells require all parts of the face to be visible before a response is generated (Perrett *et al.*, 1982; Oram and Perrett, 1994).

We have recently studied the responses of STS neurones to the presentation of single body parts with the rest of the body occluded from sight (Wachsmuth *et al.*, 1994). A large proportion of the cells responding to the whole body responded to either presentation of the head alone with the body occluded from sight or the body alone with the head occluded from sight (41.5%). A similar proportion (41.5%) responded to one but not a second body part (e.g. responding to the body presented alone, but not responding to the head when the body was occluded from sight). The remaining cells in the study (17%) responded only to the sight of the whole body and did not respond to

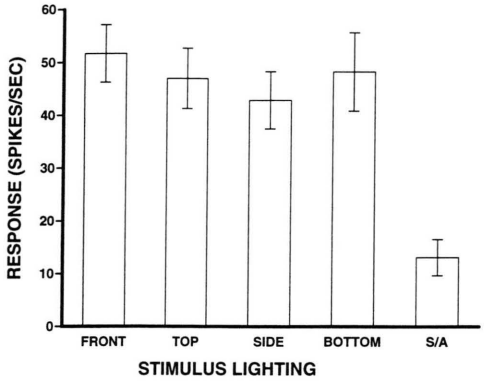

Fig. 2. Population response to changes in ambient lighting directions. The average response of 21 STS cells to the preferred view of the head and body (± SEM). The image of the head was lit with strong directional lighting from the front (normal viewing condition), from above the head (top), from one side (side) or from ground level (bottom). Responses to the four lighting conditions were statistically indistingushable ($p > 0.05$) and all were above the background or spontaneous activity (S/A) level ($p < 0.05$ each comparison). Data from Hietanen *et al.* (1992).

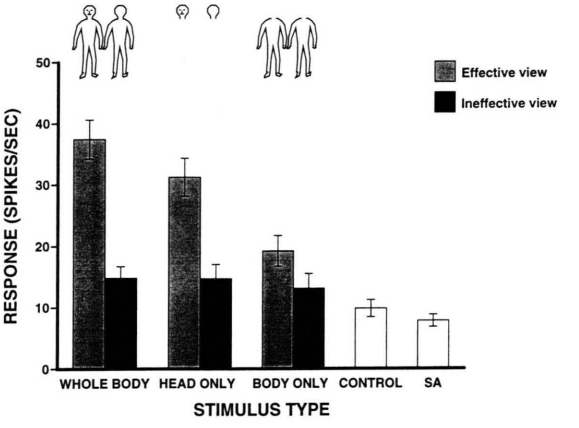

Fig. 3. Population response to body parts. Mean response (± SEM) of 53 STS cells to the whole body, the head alone and the body alone. Single body parts were presented by occlusion of the rest of the body with a matt black sheet of cardboard. Schematic figures representing the body stimuli are shown at the top of the figure. The most and least effective differed from cell to cell but are depicted as front and back views. SA = spontaneous activity (data from Wachsmuth *et al.* (1994)).

presentation of the body parts when presented in isolation.

The majority of cells (~60%) showed response selectivity that was not object-centred in the sense that the cells could not be excited from the sight of every particular object part (Wachsmuth *et al.,* 1994). This is reflected in the population response magnitudes to the different tested conditions (Figure 3). A greater population response was found when the whole body was presented than the head when presented in isolation, which in turn produced a greater population response than presentation of the body with the head occluded from sight.

3.4. Orientation

The vast majority of cells responsive to simple features and cells responsive to more elaborate features in IT cortex show orientation tuning (Gross *et al.,* 1972; Tanaka *et al.,* 1991). Some cells show preferences for radially symmetrical patterns (star shapes or Fourier descriptors, Tanaka *et al.,* 1991; Schwartz *et al.,* 1983; Gross, 1992; Miyashita and Chang, 1988; Miyashita *et al.,* 1993). Such cells show little orientation sensitivity because the shapes are often equivalent after rotation through relatively small angles (45 degrees). The earliest reports suggested the orientation tuning of cells responsive to the pattern of complex objects such as faces and hands appeared to differ between IT and STS; cells in the IT being selective for face and hand orientation (Gross *et al.,* 1972; Tanaka *et al.,* 1991) while cells in the STS generalised across orientation (Perrett *et al.,* 1982; 1988).

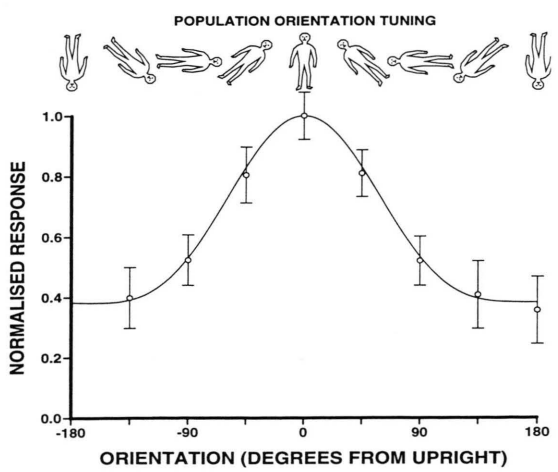

Fig. 4. Population response tuning curve of STS to change in stimulus orientation. The curve was calculated from the normalised response magnitudes of 18 cells to eight different orientations (shown schematically across the top). 0 = spontaneous activity (data from Wachsmuth and Perrett (1997)).

Re-examination of the cells in the STS responsive to the face and body indicates that most cells (82%) are tuned for orientation (Wachsmuth and Perrett, 1995; 1997). The rate of fall of response as the stimulus orientation is changed from optimal varies from cell to cell but on average declines to ½ maximal with a 60 degree orientation change. Different cells are optimally tuned to different orientations. In STS more cells appear tuned for the upright orientation (21/25, 84%) though some are selective for inverted and horizontal orientations (Wachsmuth and Perrett, 1997). The same preference for upright orientations of the face is also evident in IT (Tanaka *et al.,* 1991). Figure 4 shows the population response in STS to different orientations of the body. Note that while there is clear modulation of the response with orientation, the response to bodies is significantly higher than the response to controls for all tested orientations. Cells in AIT sensitive to other complex objects with trained significance such as wire frame objects also exhibit orientation selectivity (Logothetis *et al.,* 1995). Thus, the majority of cells in the temporal cortex selective for meaningful patterns (hand, face and body or learned patterns) exhibit sensitivity to orientation in the image plane. Interestingly the prevalence of orientation sensitivity in STS is slightly lower than that reported for the IT cells tuned to simple and Elaborate features. In IT cortex all cells show orientation selectivity (Tanaka *et al.,* 1991). In the STS, 16% of cells responsive to faces generalise response to all orientations in the image plane yet do not respond to other types of object. Cells both tuned to complex patterns such as faces and responsive to all orientations have not been reported in IT. As IT cortex provides input to the STS, this suggests that there may be a gradual increase in the extent of the object-centred coding in brain areas increasingly distal from the retina (Wachsmuth and Perrett, 1997).

Orientation affects cell responses in the temporal cortex in a similar way to perspective view. Most cells are selective for one particular orientation or view but there is evidence, at least in STS,

that a minority of cells display orientation or view invariant responses. Invariance across orientation could be achieved in an analogous way to that suggested for view processing (Perrett *et al.*, 1984, 1985; 1991; 1992; Logothetis, 1994a, b). Cells responsive to all views or all orientations may be established by pooling outputs from viewing condition (view/orientation) specific cells. To establish this generalisation it may be necessary to experience the transition across a range of some orientations. This hierarchical account predicts cells responsive to all orientations should respond at an increased latency relative to cells that are orientation specific.

3.5. Size

Cells responsive to Elaborate visual features in IT cortex are sensitive to stimulus size (Ito *et al.*, 1995). Tuning for size varies across cells, 43% were found responsive over a narrow range of stimulus sizes (½ width < 2 octaves), though some 21% of cells tolerated a 16-fold size change (>4 octaves for response to decrease by half maximal; a change in linear dimension of the stimulus from 6.25% to 100%).

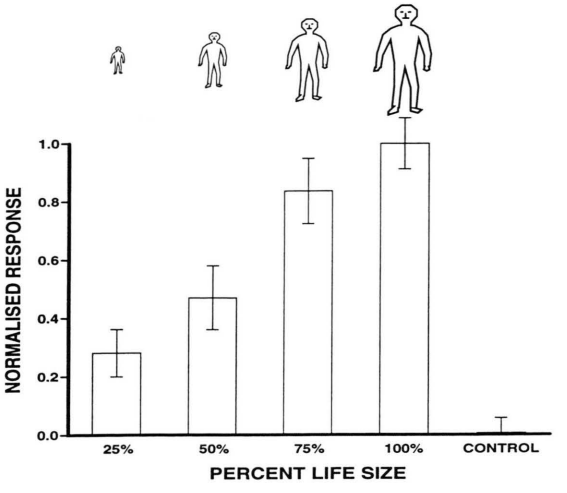

Fig. 5. Normalised population response of STS to change in stimulus size. Stimuli of 4 different sizes were presented to 16 neurones. The normalised response (± SEM) to each stimulus is shown. Life size images were projected to be the appropriate height on a viewing screen 4 m from the subject. Control objects were projected to have the approximately the same size as the largest (life size) images of the body (data from Wachsmuth and Perrett (1997)).

Studies of cells responsive to faces, hands and whole bodies indicate a range in generalisation over size with some cells tolerating a 10-fold increase in size (Perrett *et al.*, 1982; 1984; 1989; Rolls and Baylis, 1986). Other cells respond selectively to a narrower range of sizes (Wachsmuth and Perrett, 1997; Rolls and Baylis, 1986). Our recent experiments suggest that size tuning for faces and bodies reflects the degree of experience with different retinal sizes. Few cells were found responsive to small image sizes of bodies that would not be experienced often (see Fig. 5). It is suggested here that the formation of size general descriptions of objects (with cells tolerating size change > 2 octaves) may depend on inputs from several size specific descriptions (tolerating size change < 2 octaves) of the same objects.

3.6. Tolerance of stimulus position

Cells within IT show very large receptive fields, invariably incorporating the fovea and extending more than 10–20 degrees into the periphery (Gross *et al.*, 1972; 1992). It should be realised that the extent of translation invariance exhibited by cells will depend on the image size of features or patterns used to test receptive fields and the size of the elements/patterns for which cells are tuned. One would not expect to find extensive receptive fields when cells are tested with a tiny stimulus since it may only be possible to differentiate these from other stimuli within the high acuity foveal region. Cells with responses selective for more complex patterns, such as faces, also maintain stimulus selectivity across position over the central 5–15 degrees (Perrett *et al.*, 1989; Bruce *et al.*, 1981; Desimone *et al.*, 1984; Tovee *et al.*, 1994). The contention here is that the generalisation over position for these Pattern selective cells follows naturally from the convergence of Elaborate feature selective inputs that are already position tolerant. That is, the position invariance of Pattern sensitive cell responses to one object type is inherited from the input of Elaborate feature cells which are already position invariant for the object features.

Cells selective for arbitrary patterns with trained significance in anterior and ventral temporal areas also show position invariance over central vision (±5 degrees, Miyashita and Chang,

1988; Logothetis *et al.*, 1995). Studies of cells in AIT selective for particular views of wire frame objects showed a range of tolerance of position, some cells responding to stimuli reduced in size to ½ of that seen in training and moved to positions ±7 degrees (Logothetis *et al.*, 1995). Other cells exhibited more limited response zones restricted to ±2 degrees from the fixation point. Such limited receptive fields may reflect several factors. First, all training of the objects took place with presentations restricted to the fixation point. Second, the target stimuli are extremely difficult to differentiate from non-targets and discrimination may require features resolvable only with the acuity of the fovea.

3.7. Sensitivity to multiple changes in viewing conditions

In the above sections we reviewed the available data concerning tolerance of cell responses to single image transformations. Neural responses showing invariance were found for each transformation. It is important to realise that a simple count of cells need not reflect the nature of the processing of visual information. The responses of a single cell to a visual stimulus could, in principle, lead to perception.

We have tested cell selectivity to examine whether invariance to one image transformation is related to the degree of tolerance to other viewing conditions. As expected, in the majority of cases view sensitivity was found. Cells responsive to multiple parts of the same object were additionally view selective responding selectively to component parts from one view (Wachsmuth *et al.*, 1994). The view sensitivity was equivalent for the whole body and component parts (e.g. the left profile view of the head or the left profile view of the body, but not the right profile view of any body part). View sensitivity was present for cells selective for different lighting conditions (Hietanen *et al.*, 1992), size and orientation (Wachsmuth and Perrett, 1995; 1997). There was no clear relationship between the tolerance shown by a cell's responses to one image transformation and a second transformation. Some view specific cells generalised across orientation, others did not.

3.8. Discussion: Viewer-centred response selectivity

We have reviewed the response selectivity of neurones in IT and STS cortices when stimuli were changed to simulate different viewing conditions. It is clear that the majority of cells show strong viewer-centred response characteristics for at least one image transformation. An object-centred representation has by definition no information about the object's disposition relative to the viewer. It does not seem plausible therefore for the viewer-centred response selectivity of temporal cortex cells to be a product of initial object-centred pro-

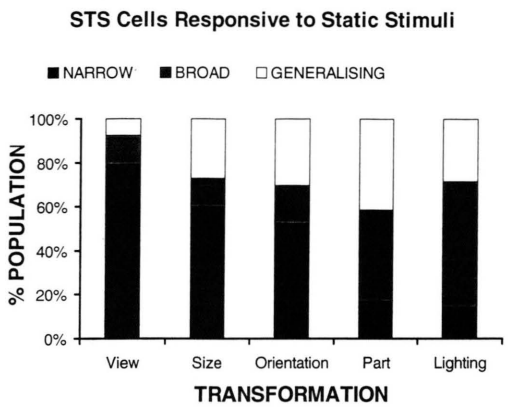

Fig. 6. Proportions of STS cells showing tolerance to a selection of image transformations. The proportions of cells showing narrow, broad, or generalising response selectivity to image transformations. View and Orientation. Cells whose response half-height half-width measure was less than 90 degrees were defined as narrow tuned. Broad tuning was defined as 90° < half-height half-width < 180°. Cells which responded to all views or orientations of the head and body at rates more than to control objects were defined as generalising. Image part visibility. Cells that responded to only the whole body and not to the tested parts (head alone and body alone) were defined as narrow tuned. Broad selectivity was assigned to cells which responded to the whole body and one body part but not a second body part presented in isolation. Cells responding to the whole body and both the head and the body when presented in isolation were defined as generalising. Size. Narrow tuning was assigned to cells whose tolerance was less than 1 octave, broad tuning was assigned to cells responsive to changes in size of up to 1.3 octaves. Cells with responses tolerant to size changes of at least 2 octaves were defined as generalising. Lighting. Cells which responded to the head in all four tested lighting conditions (normal, from side, above and below) above response levels to control objects were defined as generalising. Cells with responses above control levels to 2 or 3 lighting conditions were defined as broadly tuned. Cells that responded to only one lighting condition were labelled as narrow tuned.

cessing. Indeed the sensitivity to perspective view in temporal cortex is present from within 5 ms of response onset and hence is not an emergent property of the response, but reflects direct processing of inputs (Oram and Perrett, 1992; 1994).

Figure 6 shows the degree of tolerance shown by cell responses to changes in viewing conditions for cells in the STS selectively responsive to static images of the head and body. View specific codes at the cellular level have been evident for some time (Desimone *et al.,* 1984; Perrett *et al.,* 1982; 1984) and such view specificity is now gaining widespread acceptance (Poggio and Edelman, 1990; Logothetis *et al.,* 1994a ,b). Orientation specific codes are only recently emerging and are potentially more unexpected given earlier reports (Wachsmuth *et al.,* 1994). The finding of orientation specific coding for upright faces is perhaps not surprising since the face is often seen in this orientation (Gross *et al.,* 1972; Kendrick and Baldwin, 1987). For view there is a slight but significant bias towards front views compared with back views (Perrett *et al.,* 1992) whereas for orientation, there appears to be a much stronger bias towards coding of upright compared to inverted orientations. The bias in tuning for the upright orientation and frontal views may reflect the increased importance and familiarity of these viewing conditions. The difference in orientation and view coding may again reflect experience as inverted orientations may be less frequently encountered than back views.

The width of orientation and view tuning in STS is approximately equivalent (±60 degrees, Perrett *et al.,* 1991; Wachsmuth and Perrett, 1995). A similar bandwidth was found for direction of motion tuning (Oram *et al.,* 1993). Cells in visual cortex show tuning for bar orientation (±30 degrees); in all cases these tuning widths are approximately ⅙ of the available range (180 degree range of bar orientation, 360 degree of perspective view and orientation). The constancy of the tuning width value may arise as an optimal solution over which there is competition. The tuning of cell response selectivity in the later stages of processing (e.g. AIT, TG, AMTS, STS) may simply be a product of input characteristics. If cells responsive to faces receive input from Ealborate IT cells the inputs will be orientation tuned with a 60 degree bandwidth. Rotation of face images by >60 degrees in the picture plane would eliminate activity in the feature sensitive cells providing the inputs to cells responsive to faces.

To summarise, under Marr and Nishihara's model (1978), an object-centred representation should be accessible from the sight of isolated parts of an object, and from different views, orientations and lighting conditions. Many STS neurones showed object-centred properties with independent responses to separate object parts and responses that were tolerant to changes in lighting conditions. The extent of generalisation for other visual dimensions was, however, more limited than that expected of object-centred models since the majority of cells showed selectivity for view, orientation and size. Thus the majority of the cells show viewer-centred properties.

4. Functions of View and Orientation Specific Coding

From the studies described it is evident that there is a wide variety of cells responsive to the face and body. Information relevant to the visual analysis of an individual is thus distributed across a large array of cells each coding particular parts or features and combinations thereof. There appears to be a hierarchy within this analysis. At lower levels in this hierarchy, cells may be sensitive to only one facial component, whereas cells at a higher level may combine the outputs of many separate analyses of the appearance of the eyes, head and body each from specific views and orientations (Perrett *et al.,* 1992). Indeed to code social signals such as "attention down", information from widely different views (head and body postures with specific orientations) needs to converge on individual cells so that they respond to a range of visual signals that have the same conceptual meaning (Perrett *et al.,* 1992).

Pooling the outputs of multiple view and orientation specific descriptions of the same object establishes one description of an object's identity independent of viewing condition but simultaneously throws away information about the specific 3-D orientation of the object in question. This act of pooling may characterise ventral cortical processing where identification of object meaning is one of the chief goals of processing (Ungerleider and Mishkin, 1982). To guide motor interaction with the object one can speculate that outputs of

orientation and view specific cell descriptions in temporal cortex would need to reach the dorsal parietal cortical system (Walsh and Perrett, 1994; Carey *et al.,* 1997). Alternatively orientation and size specific descriptions need to be computed independently in parietal cortex (Sakata *et al.,* 1995).

For the head and body, orientation specific coding may have additional functional roles in guiding comprehension in social interactions. Understanding where someone else is attending is essential to the life of social primates. Analysis of attention direction requires specification not only of head view but orientation in the picture plane. For example, the sight of a left profile view of an animal's head tilted up can signify that the animal's attention is directed to the viewer's left and somewhere high up. The object of attention in this circumstance could be an aerial predator, or ripe fruit in a tree. By contrast, a view of the left profile with head down could indicate the attention to a good spot for foraging or a snake. Note that by specifying only the head view (left profile) without specification of the orientation (pointing up, level or down), it would be impossible to understand the focus of another's attention or actions. More generally, in any social activity (be the activity affiliative play or an agonistic encounter), recognition of the precise posture, orientation and movement of the social partner's face, hands and body is important if the participants are to adjust their own body movements appropriately (Perrett *et al.,* 1992; Perrett *et al.,* 1995a; Perrett and Emery, 1996; Oram and Perrett, 1996; Carey *et al.,* 1997).

5. The Binding Problem

'Illusory conjunctions' occur when features belonging to different objects are falsely attributed to the same object. Consider an array of red squares and blue circles. If an observer claims to see a red circle in such an array, such a claim would be false. The conjunction of redness and circular shape would be illusory as the colour and shape features belong to different items in the array.

Under normal viewing circumstances the visual system does not suffer from illusory conjunctions. The brain thus solves the 'binding problem' and is able to keep track of which visual attributes (e.g. redness and shape) arise from the same item. How the brain manages to solve the binding problem and avoid illusory conjunctions is a matter conjecture.

We have reviewed how receptive field size increases along the ventral stream. In IT and STS, cells respond to stimuli independent of their position within large receptive fields. We argued that such selectivity could be achieved by the pooling of inputs having smaller receptive fields and showing response selectivity for the appropriate component parts. This change in two aspects of response selectivity immediately raises the problem of false conjunctions. For example, a cell receiving inputs selective for a horizontal bar and inputs from a second population of cells selective for vertical bars would respond to T shapes. The same cell would also respond to separate horizontal and vertical bars and to + shapes. One aspect of the binding problem is to determine where one feature is relative to another feature (see Tables I, II). Solution of this problem allows T and + shapes to be interpreted as different shapes even though they have the same component parts. As previously suggested, the binding problem may not be as complex for familiar or frequently seen stimuli as it first appears (Oram and Perrett, 1994).

5.1. Coding conjunctions of features to overcome the binding problem

Consider a visual stimulus consisting of 6 arbitrary elements, El 1–6. If cells exist which are selective for the elements El 1 through El 6 and at 7 different retinal positions, then these 42 cells can be represented as in Table I. Table I represents position by font and elements of features by letters. The third column in Table II shows the resultant activity of an imaginary IT Elaborate cell receiving input from the 42 cells of Table I a to a variety of stimuli. As can be seen, if the input units have sensitivity to isolated features, the binding problem is severe, with the putative IT cell responding to many false conjunctions of the individual elements. While the goal of processing was to detect the stimuli rows 1–3 of column 1 Table II, the processing arrangement gave responses to the non-targets on rows 4–10.

IT Elaborate feature sensitive cells can be selective for the presence and relative position of more than one visual item in the image. When the selec-

Table I. Possible processing elements selective for single or conjunction of visual features. Two sets of feature detectors are shown. Table Ia. Grid of 42 cells selective for visual elements El 1 to El 6. The elements are incicated by the letters A–F, different retinal positions are indicated by the different font types and sizes. Table Ib. Grid of 42 cells selective for conjunctions of visual features El 7 to El 12. The conjunction of elements are indicated by the pairs of letters (ab, bc and so on).

	Table Ia						Table Ib					
	El 1	El 2	El 3	El 4	El 5	El 6	El 7	El 8	El 9	El 10	El 11	El 12
Position 1	A	B	C	D	E	F	AB	BC	CD	DE	EF	F
Position 2	a	b	c	d	e	f	ab	bc	cc	de	ef	f
Position 3	**A**	**B**	**C**	**D**	**E**	**F**	**AB**	**BC**	**CD**	**DE**	**EF**	**F**
Position 4	**a**	**b**	**c**	**d**	**e**	**f**	**ab**	**bc**	**cd**	**de**	**ef**	**f**
Position 5	**A**	**B**	**C**	**D**	**E**	**F**	**AB**	**BC**	**CD**	**DE**	**EF**	**F**
Position 6	A	B	C	D	E	F	AB	BC	CD	DE	EF	F
Position 7	a	b	c	d	e	f	ab	bc	cd	de	ef	f

Table IIa. Responses of a cell receiving input from the 42 cells in Ia. The presence or absence of a response to 10 stimuli is shown. Incorrect responses are shown in bold and underlined. Table IIb. Responses of a cell receiving input from the 42 cells in Ib. The presence or absence of a response to 10 stimuli is shown. Note there are no possible incorrect responses.

Stimulus	Target stimulus	Response to set Ia	Response to set Ib
ABCDEF	YES	YES	YES
abcdef	YES	YES	YES
ABCDEF	YES	YES	YES
ab**c**D**e**f	NO	**YES**	NO
A**B**cdeF	NO	**YES**	NO
ab**CdE**F	NO	**YES**	NO
ABCD**EF**	NO	**YES**	NO
ABcdEF	NO	**YES**	NO
a**b**Cdef	NO	**YES**	NO
aB**C**D**e**F	NO	**YES**	NO

tivity of several such Elaborate feature cells is combined the positional interdependence of many pattern elements is established. This interdependence circumvents the binding problem as illustrated in Table Ib and Table II. In Table Ib, each of the cells is sensitive to the relationship between two elements. As can be seen, a cell with inputs from the 42 units no longer responds to the distracter stimuli (Table II, right column). Furthermore, only 35 input cells are needed – the cells sensitive to element 12 (El 12, the Fs) are redundant.

The principle of using feature conjunctions to overcome issues of binding is of course easily extended to transformations other than retinal position. In terms of information coding, Elaborate cells code for multiple dimensions of the stimulus. When sensitivity to more than one aspect of the

stimulus exists (e.g. colour, shape and texture), each attribute modulates the response separately (Komatsu and Ideura, 1992). The responses of such cells therefore contain information about each attribute that is not dependent on other attributes. It is the overlap of feature selectivity between cells sensitive to conjoint features that allows "binding" to occur (Richmond and Optican, 1992; Oram and Perrett, 1994). The utilisation of feature conjunction is not restricted to the Elaborate cells of IT cortex. There is ample evidence that cells in V1 are sensitive to elements in the area surrounding the classical receptive field (Knierem and Van Essen, 1992; see Allman (1985) for review). These cells in V1 show response selectivity to conjunctions of visual elements (e.g. vertical bar surrounded by horizontal bars). The cells in subsequent areas of the ventral stream (V2, V4, PIT, AIT, STS) can utilise the principles outlined above to circumvent the binding problem.

5.2. Feature conjunction for word and face recognition

There is a simple demonstration that processing may utilise combinations of features and that these features might not be those that first spring to mind. ThIs SeNtEnCe Is NoT tHe EaSiEsT tO rEaD, eVeN tHoUgH iT cOnSiStS Of ThE uSuAl LeTtErS aNd SpAcEs. When written normally all letters are the same case or may have a capital at the beginning. Under these normal conditions, we have no difficulty in reading. "This Sentence Is Not So Difficult To Read, Even Though It Consists Of Upper And Lower Case Letters". The relative ease in reading the second case alternated sentence is presumably because we are more familiar with the change in case (initial upper case to lower case). The confusion in the first example is due to the unusual combinations of letters of different cases. This indicates that visual processing of words uses "between letter" features and is not restricted to the analysis of isolated letters.

IT Elaborate feature sensitive cells are selective for the relative position of elements over large receptive fields. In the example of Tables I and II, it was assumed that input cells were position sensitive. IT cells, however, show tolerance of position. The scheme for binding features together works even if the feature detectors are tolerant of position. For example, a Pattern sensitive cell might receive input from 4 hypothetical Elaborate feature sensitive cells with large receptive fields. Of these hypothetical cells, Elaborate feature cell 1 requires visual elements A + B in given spatial arrangement; cell 2 requires elements B + C + D; cell 3 requires elements D+E and cell 4 requires elements E + F. If the Pattern sensitive cell can only be activated by the combined input of these 4 input cells, it will require all 6 visual elements (A, B, C, D, E and F) to be present in the image. More importantly it will only respond to one unique set of spatial relationships between those elements. The cell will be tuned to the arrangement of element F and E relative to elements A and B, even though none of the input cells are sensitive to the arrangements between these particular elements. Note that the higher level cell will inherit the positional tolerance of the input cells. Thus while the higher level Pattern sensitive cell shows increased specificity to pattern configuration, the cell generalises in its response to the optimal configuration of letters over a range of retinal positions.

Figure 7 illustrates the advantages of using conjunction of features for analysing facial patterns.

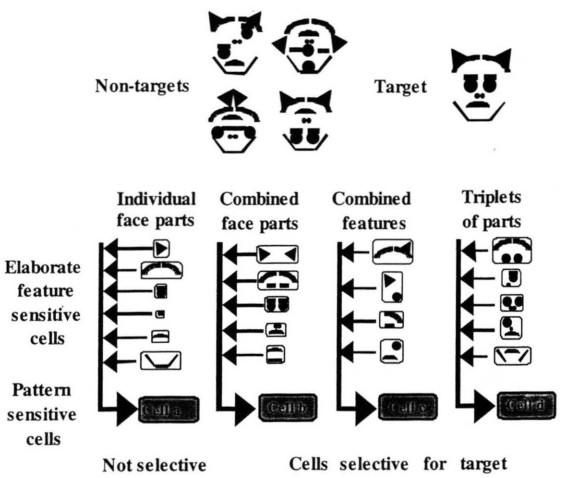

Fig. 7. Establishing sensitivity to configuration independent of position. Top: target (face) and non-target (jumbled face patterns). Lower: columns of hypothetical Elaborate feature sensitive cells, sensitive to the visual element displayed within one box but generalising over the feature position within central vision. Hypothetical pattern sensitive cells a–d pool the outputs of 4–6 elaborate feature sensitive cells. Cells b–d but not cell a are sensitive to the pattern configuration and would respond more to the face than to the jumbled faces.

In the centre of the figure are 4 columns of simple geometrical shapes which occur in facial patterns. Individual hypothetical Elaborate cells in IT cortex could be tuned to the shapes drawn in any one of the separate boxes. Note that an Elaborate feature sensitive cell need not be tuned to a face part *per se* and may respond to any object containing the relevant features. The Elaborate cell second from the top of the left column would respond to the upward convex edge of a face or a melon. The left column gives a collection of geometrical shapes that occur often in individual face parts (ear, top of the head, eye, nose, mouth and jaw line). These parts can be arranged into a face structure (target) and various jumbled arrays (non-targets). Many Elaborate cells in IT are selective for local features involving more than one shape component. The three columns to the right illustrate the fact that the whole facial pattern contains many features that involve more than one facial part. The column second from the left illustrates several intuitive facial regions defined by the combination of pairs of face parts: for example, the forehead (framed by the eyebrows and top of the head); the upper lip (formed by the relationship between the nose and mouth) and the chin (formed by the mouth and jawline). The third column illustrates less intuitive features formed by pairs of face parts and the right column features formed from triplets of face parts.

One can consider the selectivity of Pattern sensitive cells (a–e) formed by pooling the outputs of Elaborate cells tuned to each of the features present in one vertical column. Cell (a), pooling outputs from cells tuned to isolated face parts, would respond well to the target face pattern. Indeed by setting a high threshold for firing, simultaneous input from several face part features would be necessary for any activity. Thresholds for firing would depend on the strength of synaptic inputs, electrotonic distance of the synapses from the site of spike generations and the timing between input spikes: simultaneous inputs from several weak synapses on the distal parts of the dendrite would be needed to cause a cell to spike (Gochin *et al.,* 1991). In this way the Pattern sensitive cell (a) of Fig. 7 would respond well when the entire face was visible but only weakly or not at all when one or two parts of the face were presented in isolation with the rest occluded from sight. Such selectivity has been documented (Oram and Perrett, 1994; Perrett *et al.,* 1982). Cell (a) would be, however, insensitive to configuration and would respond to the non-target jumbled face patterns. This is because each input tolerates the presence of one face part in any position. Some cells do indeed respond to both normal and jumbled arrays of facial features. Most cells, however, show greater responses to the normal pattern than to jumbled patterns (Perrett *et al.,* 1982).

Pattern sensitive cells (b), (c) and (d) pool outputs of Elaborate cells tuned to features formed by conjunctions of face parts or geometrical elements that occur within regions of faces. Cells (b)–(d) would also respond well to the entire face pattern but in contrast to cell (a), these cells would be additionally selective for configuration; they would not respond well to the jumbled faces. Cells (b)–(d) would respond to face patterns over a range of retinal positions because each of the input Elaborate cells tolerates its preferred feature over a range of positions. In this way it is possible to generate pattern selectivity which is both capable of generalising across retinal position and which is capable of discriminating between different configurations.

Of course, different Pattern sensitive cells could receive inputs from different combinations of Elaborate feature sensitive cells and individual Pattern sensitive cells need not specify all aspects of the configuration arising from a familiar object. Increasingly exacting configuration sensitivity can be established by further processing in which the outputs of several Pattern sensitive cells each sensitive to different aspects the same object (e.g. face) are combined. Moreover, the process can be repeated with successively higher cells pooling the outputs of cells sensitive to different parts of the same complex object (e.g. body). The resulting sensitivity to multiple components and their spatial configuration can be seen to offer a progressively improved selectivity amongst patterns. Exactly this type of selectivity would be needed for discriminating members of the same category of objects whether they are words, faces, fractal patterns or wire frame objects.

Figure 8 provides an analogous example for individual word recognition and text reading. High level 'features' such as those typical of letters themselves or more importantly those typical of

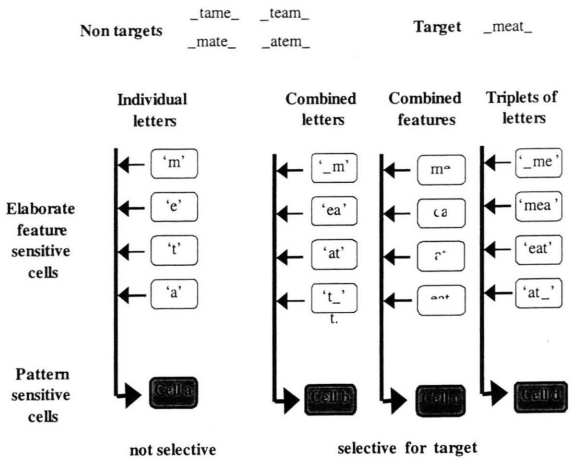

Fig. 8. Establishing sensitivity to letter order. Top: target and non-target words. Middle: columns of hypothetical Elaborate feature cells, sensitive to the visual element displayed within one box but alising over feature position within central vision. Lower: hypothetical Pattern sensitive cells (a)–(d) pool the outputs of 4 Elaborate feature sensitive cells. Cells (b)–(d) but not cell (a) are sensitive to the letter configuration and would respond more to the word 'meat' than to the other words with the same letters ('tame, mate, team') or non-words (atem).

letter pairs and commonly occurring letter triplets would be appropriate for recognising words. In Fig. 8, hypothetical Elaborate feature sensitive cells in the left column are tuned to the approximate shape of individual letters (lower case, perhaps in a small range of visually similar fonts). Such hypothetical Elaborate cells are tuned to more complex features than those described by Tanaka *et al.* (1991). The scheme suggested here can be reiterated to generate sensitivity to progressively more complex patterns. Thus it is possible to imagine that Elaborate features of the type described by Tanaka are used to create sensitivity to letters and letter pairs, and these are used to create sensitivity to individual words or longer strings of letters.

The choice of Elaborate features as inputs in Fig. 8 is again critical. Pooling the outputs of feature sensitive cells selective for individual letters by Pattern sensitive cell (a) reiterates the problem of detecting letter configuration. Cell (a) would respond equally to all words and non-words with letters m, e, a and t (e.g. tame, atem, mate, team, maet, etc.). By contrast, Pattern sensitive cells (b) and (c) which pool features that are conjunctions of letters (e.g. _m, me, ea, at and t_, column 2) or conjunctions of pattern elements typical of such letter pairs (column 3) would be able to differentiate between words with the same letters. The Pattern sensitive cell (d) that pools features that are triplets of letters in column 4 would also be able to differentiate the target word from non-targets with the same letters. Indeed there are many visual components of words that can be used for generating selectivity between words, whether the components are conjunctions of letter features or parts of letters; the components do need not need to be explicit suffixes and prefixes of words.

5.3. Summary of the binding problem and feature conjunction

The thesis advanced here is that in order to detect complex patterns such as those approximating the face or the eyes, Pattern sensitive cells pool the outputs of a number of Elaborate feature sensitive cells. The pooling operations might superficially appear unable to generate configuration sensitivity (i.e. selectivity for the relative position of features within the image). As illustrated, however, the pooling features which are themselves pairs of visual elements can increase sensitivity to pattern configuration.

In essence, the properties of higher level cells selective for complex pattern configurations and capable of generalising over position are produced by combining the outputs of lower level cells tuned to simple configurations. Pattern sensitive cells established in this way will have several response properties; as described they will be selective for object type and capable of generalising across stimulus position in the central region of vision. The cells will also display selectivity for image size, orientation, view and object part. Selectivity for orientation and size will be inherited directly from the Elaborate feature cells which are themselves orientation and size selective. Selectivity for object view and part will arise because the features required by input Elaborate cells will be visible only for particular views and parts of the object.

A key aspect of the account presented here is the realisation that a relatively limited list of feature requirements automatically produces sensitivity for configuration of an object's components.

This is because changing the configuration of components automatically removes or diminishes the visibility of some of the required features. In Figure 7, the Pattern sensitive cells (c)–(d) fail to respond to the jumbled faces because displacing the eyes, nose and mouth has the effect of destroying various features that would activate the input Elaborate feature sensitive cells. Therefore, the Pattern cells will fire sub-optimally to rearranged faces because they will no longer receive input from a full set of Elaborate cells. Reduced response to jumbled faces has been described previously (Perrett *et al.,* 1982; 1988). Selectivity for one type of object is also conferred by the list of required Elaborate feature inputs; most other types of object will not contain sufficient of these features to generate false positive activation and spurious recognition.

The configuration of face parts is radically different between the target and non-target patterns illustrated. Cell sensitivity to more subtle changes in configuration within the same class of object could be achieved in an exactly analogous way. Elaborate feature sensitive cells are sensitive to the size, orientation and the relative position of components (Tanaka *et al.,* 1991). Pattern sensitive cells combining outputs of multiple Elaborate feature sensitive cells could therefore be sensitive to subtle differences in face configurations as facial components change in size and position relative to one another. In this way Pattern sensitive cells could be selective for the differences between familiar faces as has been observed in several investigations (Baylis *et al.,* 1985; Hasselmo *et al.,* 1989; Perrett *et al.,* 1984; 1989; 1991; Yamane *et al.,* 1988; Young and Yamane, 1992).

6. Relationship Between Features and Feature Configurations

From the discussion in section 5, it is clear that the Elaborate cells described in IT cortices may not be responsive to only the intuitively obvious features of objects. Indeed, since cells are selective for the relationship between features, aspects of the binding problem can be circumvented and the processing in temporal cortex can resolve different configurations of patterns made up of the same elements (Oram and Perrett, 1994). In this section we examine the role of intuitive features of images and their configuration on object recognition. Accounts of recognition frequently differentiate between the role of key features and the role of feature configuration. Classes of object where all exemplars have the same parts of features (such as faces) are supposed to be differentiated on the basis of the configuration of component features (e.g. Farah, 1996). From the outset it should be noted that transformations that affect configuration necessarily affect the shape of multiple component features and *vice versa.*

During the rearrangement of the face (see Fig. 7) the appearance of some features of the face (nose, mouth and eye) need not be disrupted. There is a tendency to conceive of an object's features as being equivalent to the list of the minimum number of parts that when combined generate the overall appearance of the object. This assumption is explicit in the model of Biederman (1987) where objects are recognised by decomposition into a set of 3-D geon parts. It should be realised that the visible features (in the sense specified here) that are generated by a particular combination of object parts, greatly outnumbers the parts themselves. The eyes, nose, mouth, eyebrows, hair and head/jaw outline are the most obvious component parts of the face and are sometimes considered the only 'features'. The relationships between these intuitive parts of the face create numerous additional visual features. For example, the light space between the mouth and the nose parts form another feature, the upper lip. Consider an object composed of $n = 5$ discernible differently shaped dark blob parts. These blobs might be considered the only component features. In which case there would be five 1st order features ($n = 5$) but any view with all blob components visible will include further 2nd order visual features created by pairs of blobs [$n = 5!/((5 - 2)!2!) = 120/(6*2) = 10$] and 3rd order features as triplets of blobs [$n = 5!/((5 - 3)!3!) = 120/(2*6) = 10$]. It is not clear which type of feature (simple or configurational) is more important for object recognition. For any given object class both 1st and higher order features could form the basis of discrimination between different exemplars of that class.

Tanaka *et al.* (1993) found that learning about the parts of a face in the context of an entire face generates context sensitivity to the parts. After subjects learned the appearance of two whole

faces differing only in nose shape, subsequent testing showed that the noses were discriminated more accurately in the context of the trained face pattern than when tested in isolation (as noses alone). This contextual sensitivity was present only when the faces were learned in the upright orientation and was not found for other types of object (e.g. buildings). Such contextual sensitivity (and its orientation specificity) naturally arises when one considers the training pattern as consisting of more visual features than the intuitive parts. The combination of the nose with the eyebrows and the mouth creates additional sets of (2^{nd} order) visual features. For example, the bridge of the nose may make a T or Y feature with the eye brows; the base of a long nose may abut the mouth creating an inverted T feature. For a shorter nose, the upper lip may noticeably separate the base of the nose and the mouth. To reiterate a point above, it is impossible to change the appearance of one face part without changing the configuration of the features. Reciprocally, changing the configuration of a face by moving the intuitive parts relative to one another also changes the shape of component features: moving the nose up from the mouth creates a bigger upper lip feature.

When subjects learn new face shapes or configurations, they appear to be sensitive to component facial features in the broader sense of the term and do not restrict their visual analysis to the intuitive macroscopic face parts (Tanaka *et al.,* 1993). The orientation specificity of the training is expected given the orientation sensitive Elaborate and Pattern sensitive cells in the temporal cortex of the observers. This explanation does not account for the lack of context sensitivity for features trained in inverted faces. For these unusual inverted patterns subjects may have not had sufficient perceptual experience to develop appropriate Elaborate feature and Pattern sensitive cells tuned to relevant visual features and feature combinations to enable discrimination. Sensitivity to pattern configuration and disproportionate problems with inversion have frequently been noted to depend on expertise or familiarity with the pattern type (e.g. Carey, 1992; Gauthier and Tarr, 1997).

7. Are Faces Special?

One might argue that faces are a special case and recognition of other object classes utilise fundamentally different types of neural processing. The issue of whether faces are special has been raised in the contexts of neuropsychology and psychology (Yin, 1969; 1970; Farah, 1996). From the psychological perspective there are many indications of equivalent processing mechanisms for faces and objects. Factors affecting the efficiency of face recognition also affect recognition of other types of object in a qualitatively equivalent manner. Face processing is affected adversely by change in perspective view, lighting and stimulus orientation with such effects visible both at the behavioural and at the single cell level (Benton and Van Allen, 1968; De Renzie *et al.,* 1968; Yin, 1969; 1970; Perrett *et al.,* 1988; 1991; 1995b; Hietanen *et al.,* 1992). The same parameters also affect recognition of objects (Warrington, 1982; Weiskrantz and Saunders, 1984; Flin and Dziurawiec, 1989; Carey, 1992; Gauthier and Tarr, 1997). Conversely, automated caricaturing can improve recognition of both faces (Rhodes *et al.,* 1987; Benson and Perrett, 1991) and objects (Rhodes and McLean, 1990; Dodd and Perrett, 1997).

To find comparable effects of image manipulations on face and object recognition it may be necessary to tap the ability of subjects to use the subtle differences in the configuration of features (Bruce *et al.,* 1991). This skill is only acquired after a substantial amount of experience in differentiating members within the same object class (Carey, 1992; Rhodes and McLean, 1990; Dodd and Perrett, 1997; Logothetis *et al.,* 1994a, b).

Brain damage can selectively impair recognition of faces but leave recognition of other object classes intact (e.g. McNeil and Warrington, 1994; DeRenzie, 1986). Conversely different cases show intact recognition of faces but defective recognition of objects (Assal *et al.,* 1984; Moscovitch *et al.,* 1997). Such dissociation need not imply that faces and objects are processed differently, for example objects and words being processed by reference to single features and faces by reference to a configuration of features (Farah, 1996). The neuropsychological dissociation may imply only that faces and objects are processed in different cortical modules, even in the same architectonic area of the cortex. Cortical neurones processing the same type of stimulus tend to cluster together anatomically. Their proximity implies similar input selectivity, not necessarily differences in style of processing

(Fujita *et al.,* 1992; Perrett *et al.,* 1984; Harries and Perrett, 1991). Different columns in primary visual cortex process edges of horizontal and vertical orientations, yet we do not imply from this that the two orientations are processed in fundamentally different ways. Indeed, there is now evidence that different patches (1–4 mm) within the lingual and fusiform gyri (which could well be the human equivalent of the macaque inferior temporal cortex) are selectively activated by the sight of faces and words (Allison *et al.,* 1994a, b; Puce *et al.,* 1996a, b; Nobre *et al.,* 1994; c.f. Gauthier *et al.,* 1997).

In humans, cortical modules processing specific object classes may be located within same general region (e.g. the lingual and fusiform gyri). If such modules are located in slightly different places for different individuals (Puce *et al.,* 1996a, b; Nobre *et al.,* 1994), differently positioned lesions would be expected to correlate with different symptoms. If modules analysing one class of object are interspersed with other modules analysing other object classes, then most lesions would be unlikely to produce highly selective deficits. This is because a given lesion would be likely to disrupt several modules processing different classes of stimuli. Only occasionally would one expect a bilateral lesion in the fusiform region to selectively disrupt one type of object processing (Assal, 1984; McNeil and Warrington, 1994; Moscovitch *et al.,* 1997). Cellular specificity in processing a variety of different classes of object has now been established. Cells have been found in the anterior temporal cortex that are tuned to arbitrary visual patterns (colour fractal patterns; wire frame shapes) that have acquired significance through behavioural training (Sakai and Miyashita, 1991; Miyashita *et al.,* 1993; Logothetis *et al.,* 1995). Cells in the amygdala have been noted with selective responses to specific desirable foods and to animals such as centipedes with learned aversive associations (Ono *et al.,* 1992). Hence faces are not the only class of object for which cell populations can be selectively tuned.

8. Similarities to Other Models

Many accounts of recognition acknowledge the importance of component configuration (e.g. geon one to the left of geon 2 and joined at the mid point; Biederman, 1987). Few models, however, address the mechanisms for determining configuration. Several accounts generate sensitivity to pattern configuration using principal component analysis but these accounts are not designed to be biologically plausible (e.g. Sirovitch and Kirby, 1987) since such approaches subject the entire image patterns to a global analysis without the initial analysis of localised sub-regions of the image that is apparent in the nervous system.

The process of generating sensitivity to one view of an object from a list of 2-D features independent of their position is superficially similar to artificial intelligence models employing the 'viewpoint consistency constraint' (Lowe, 1987; Porrill *et al.,* 1988). In these schemes of recognition, a single all-encompassing object description is held in memory. This description specifies all the object's edges that will be visible under idealised lighting from any view. To determine whether the target object is present in the image, the procedure is first to hypothesise a given vantage point of the observer. From this view, the exact appearance of the object's edges including their size and orientation is predicted. The ensuing process checks for correspondence between triplets of edges or features that should be visible from the hypothetical view and the features present in the image. If a given set of three features exist in the image and match the prediction, then the match constitutes a limited amount of evidence that the object is present in the image at the hypothesised viewpoint. By performing checks on other triplets of simple features, confidence can accumulate that the hypothesis is correct. If there is insufficient evidence at the end of the checking process for one view then a new hypothetical view can be chosen and the process repeated.

The process of Lowe (1987) is similar to that suggested above, in that multiple low level features are checked to generate complex pattern selectivity for one object type from one view. Under the scheme of Lowe (1987), however, the only representation of the object that is stored is a single 3-D object-centred description of the object's edges relative to one another. From this, the precise appearance of the object's edges from all possible views can be generated.

Note that to establish the configuration of edges to enable "checking" of an image against a object-

centred representation, Lowe's scheme of processing would seem to require that a temporary viewpoint-dependent representation must be generated. We have shown that responses of STS neurones discriminate between input images within 5 ms of response onset and that the response latencies suggest that only feed-forward processing occurs to generate the initial response (Oram and Perrett, 1992; 1994). Therefore the view selectivity does not seem to arise from some time consuming interplay between object-centred representations and the image using the viewpoint consistency constraint of Lowe (1987).

The processing within temporal cortex can check for all familiar objects in all the familiar viewing conditions in parallel and in a single processing step because of the separate neural representations for different views of the same object and separate representations for particular image sizes and orientations. The parallel process depends only on the existence of neural populations tuned to the pattern of different views, orientations and sizes of familiar object classes. These populations might seem too numerous for such a scheme to be feasible but economy is achieved because each cell population is broadly tuned to view, orientation and size and populations are only dedicated to object classes that are important to the observer. Parallel processing is conducted in the temporal cortex with an enormous speed advantage even if the computing elements are several orders of magnitude slower than the computing elements in digital computers running programs with a view-point consistency constraint.

Fukushima's account of recognition is also somewhat similar to the account suggested here in that both utilise a gradual increase in receptive field size coupled with a gradual increase in complexity of feature sensitivity to generate selectivity for pattern configuration independent of position (Fukushima, 1980). The configurational selectivity for complex patterns of high level units in Fukushima's model is a product of sensitivity of low level units with preference for particular features *at particular locations*. Detecting an A is a product of detecting specific features at specific locations (e.g. a / feature at one retinal region to the left of another region containing a \ feature and below a region containing a $\wedge$ feature). While the receptive fields of lower level units detecting the features $\wedge$, / and \ are relatively localised, the receptive field of the higher level unit detecting A is broader. In Fukushima's scheme an "A" detector that is even more position invariant is made at a higher level by combining the outputs of many "A" detectors with limited fields.

The current proposal can be seen as an extension of Fukushima's model in that it uses a greater number of features and more elaborate features than those described for detecting an "A". Moreover the present scheme uses cells with large receptive fields extending over the entire foveal region. Fukushima's model generates sensitivity to pattern configuration over local regions of the image, and then establishes translation invariance. The processing within temporal cortex described here generates sensitivity to increased feature complexity by combining detectors for different features with large receptive fields.

The example given for detecting the letters is perhaps misleading as letter discrimination can be based on relatively simple features. It would not be advantageous for pattern recognition to utilise cells with large receptive fields and selective for simple features such as an oriented edge. It would not be possible to generate pattern specificity by pooling the outputs of such cells because too many different patterns would possess such simple features. In the case of individual word recognition and text reading, higher level 'features' such as those typical of letters themselves or more importantly those typical of letter pairs and commonly occurring letter triplets would be appropriate for recognising words (see Fig. 8).

The scheme of using local groupings of features to generate patterns of greater complexity while maintaining the specificity in pattern configuration has a long history (Wicklegren, 1969; McClelland and Rumelhart, 1988). Wicklegren (1969) noted that words which were to be spoken could be represented as sequences of context sensitive phoneme production units, "Wicklephones". The word /kat/ could be represented uniquely by three wicklephones /_ka/, /kat/ and /at_/. Different words may contain the same letters or the same phonemes but they do not contain the same Wicklephones. Given the overlap (local configuration) in features, such a scheme ensures the production of the correct word and not some spurious word containing the same letters or phonemes in a

different combination. Recent neurophysiological studies provide direct support for the notion that complex motor output sequences are generated from the sequential coding of transitions between component movements (Tanji, 1996).

While Wicklephones are appropriate for word pronunciation, McClelland and Rumelhart (1988) suggest a similar scheme of "Wicklefeatures" for encoding the form of words from sensory input. One of the critical visual features utilised in the model of McClelland and Rumelhart (1988) and in Fig. 8 is the absence of visual elements in a retinal region (a space). Spaces are important for defining the beginning and end of words and segmenting separate words within text. Text without spaces would still be analysable with this approach but ambiguity is likely to arise. During normal reading, the processing of letter combinations would be restricted by the field of view in which letters around the fovea would be visible and under attention. The region in which letters may be resolved for a given point of fixation probably does not extend more than 15–25 letters wide by 6–8 lines of single spaced text. Other schemes of simulated word reading (e.g. interactive activation model of McClelland and Rumelhart, 1988) represent the letter position within a word explicitly. Such position information would be lost by neurones in temporal cortex. Coding relative position of letter pairs (or elements within letter pairs), independent of their absolute position within the foveal region, is one way that the visual processing can establish the letter order that is the basis of word construction.

9. Representation of Objects by Patterns of Neuronal Activity

One objection to the scheme of coding described above is that it depends on neurones in temporal cortex that are selective for the features or overall appearance of each and every familiar object class. To some this may seem an unacceptable return to notions of 'Grandmother' cell coding where individual cells in an observer's brain are hypothesised to respond selectively to the appearance of particular highly familiar faces (such as the observer's Grandmother). The relative merits of population coding in which each visual neurone contributes to the representation of many objects or sparse coding where single cells code more information about particular objects or examples have been debated elsewhere (Perrett *et al.*, 1987; 1989; Barlow, 1994; Foldiak and Young, 1995; Rolls and Tovee, 1995; Rolls *et al.*, 1977).

The visual pattern that characterises any object is specified at a given moment by a population of light sensitive retinal cones, or by a population of cells in primary visual cortex sensitive to local edges or spatial frequencies. Thus at early stages the pattern of an object is encoded by vast populations of cells. The processing which underlies object recognition, however, progresses beyond both retinal cones and edge detectors in primary visual cortex. In the same way, while cells in IT can provide an elaborate and extensive description of the features of objects, visual processing and the analysis of feature combinations, shapes and configurations can progress beyond this level. Cells in IT that are sensitive to feature configuration can easily participate in signalling the presence of many objects.

As each object class or exemplar becomes familiar and important to an observer then more and more neurones should become tuned to the specific features of the object class/exemplar. Feature tuning would be size and orientation specific reflecting the image conditions experienced. Since any novel object will share visual features with other already familiar objects, these features can be utilised initially to differentiate the new object. Biederman (1987) also noted that novel objects can be described by reference to a set of 3-D features or 'geons'. What is suggested here is a set of features thousands of times more numerous than those postulated by Biederman (1987). Moreover, unlike Biederman's account, the features would be universal (across observers) only to the extent that they would be frequently experienced and would be useful for differentiating common object classes. Feature sensitivity would be established through experience and each person's repertoire would be different. Thus the features would be adaptive (Miyashita and Chang, 1988; Miyashita *et al.*, 1993; Logothetis *et al.*, 1994a, b; 1995) and for any individual would reflect their lifestyle and interests. Someone with a professional interest in flowers would develop neurones sensitive to the visual differences between different types of leaf and petals while a motor mechanic would develop

sensitivity to features more appropriate for differentiating engine components.

For each class of object and visual pattern we know well, we may possess cells which are selectively activated by that object class. To recognise objects such as the daffodil flower, pineapple and schematic drawings of the sun, we may indeed utilise high level Pattern sensitive cells that are selective amongst these alternatives. It is not being suggested here that there will be only one cell selectively activated by each type of familiar object. Indeed, former proponents of single cell codes (Konorski, 1957; Barlow, 1972; 1985) all argued that there would be multiple cells selective for the same object, with the cell numbers depending on the object's importance. There may be thousands or millions of cells selective for the same object and any one cell in this population is unlikely to be as efficient as the psychophysical observer at detecting the object's presence (Barlow, 1985).

10. Summary

This paper has reviewed the development of pattern sensitivity of cells along the ventral cortical stream of processing. It is evident that the complexity of trigger features that are required to make cells respond increases progressively along the cortical pathway. At the highest levels of processing, in the temporal cortex, cell populations appear to encode the pattern of familiar classes of object from specific viewing conditions. The majority of cells show limited generalisation over transformations that modify the visibility of object parts, or the view, orientation and image size of the object. This view selectivity appears to be matched to the experience of the observer. For objects seen predominantly at one view, orientation and size, the majority of cells selective for that object respond more to the object when presented at the familiar view, orientation and size. The review has considered a plausible mechanism for generating selectivity for a specific pattern configuration at one orientation and size while maintaining the ability to generalise across position. Experience of the same object, from different views and orientations, at different retinal sizes, under different directions of strong illumination and with different parts occluded, can establish a range of cell types each tuned to the same object from specific viewing conditions. Generalisation of recognition of the same object across different viewing conditions can be achieved by linking the outputs of such cell types through associative learning as the transition between different views is experienced over time. In general, view specific coding is a natural consequence of selective experience and, in the case of biologically important objects such as bodies, view specific coding has great utility for understanding social signals.

Allison T., Ginter H., McCarthy G., Nobre A. C., Puce A., Luby M. and Spencer D. D. (1994a), Face recognition in human extrastriate cortex. J. Neurophysiol. **71,** 821–835.

Allison T., McCarthy G., Nobre A., Puce A. and Belger A. (1994b), Human extrastriate visual-cortex and the perception of faces, words, numbers and colours. Cereb. Cortex **4,** 544–554.

Allman J., Miezin F. and McGuinness E. (1985), Stimulus specific responses from beyond the classical receptive field: neurophysiological mechanisms for local-global comparisons in visual neurons. Ann. Rev. Neurosci. **8,** 407–430.

Assal G., Favre C. and Anderes, J. P. (1984), Non-reconnaissance d'animaux familiers chez un paysan: zoo-agnosie ou prosopagnosie pour les animaux. Revue Neurologique **140,** 580–584.

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

Barlow H. B. (1985), The twelfth Bartlett Memorial Lecture: The role of single neurons in the psychology of perception. Quart. J. Exp.Psychol. **37,** 121–145.

Barlow H. B. (1994), The Cognitive Neurosciences (M. Gazzaniga, ed.). Place Publishers, pp. 415–435.

Baylis C. G., 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.

Benson P. J. and Perrett D. I. (1991), Perception and recognition of photographic quality caricatures: implications for natural image processing. Eur. J. Cognit. Psychol. **3,** 105–135.

Benton A. L. and Van Allen M. W. (1968), Impairment in facial recognition in patients with cerebral disease. Cortex **4,** 344–358.

Biederman I. (1987), Recognition by components: a theory of human image understanding. Psycholog. Rev. **94,** 115–145.

Bruce C. J., Desimone R. and Gross C. G. (1981), Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. J. Neurophysiol. **46,** 369–384.

Bruce V., Doyle T., Dench N. and Burton M. (1991), Remembering facial configurations. Cognition **38,** 109–144.

Bruce V. and Young A. (1986), Understanding face recognition. Brit. J. Psychol. **77,** 103–327.

Carey S. (1992), Becoming a face expert. Phil. Trans. R. Soc. London B **335,** 95–103.

Carey D. P., Perrett D. I. and Oram M. W. (1997), Recognizing, understanding and reproducing action. In: Handbook of Neuropsychology, Vol. **11** (F. Boller and J. Grafman, Series eds.). Elsevier Science, pp. 111–129.

De Renzi E., Faglioni P. and Spinnler H. (1968), The performance of patients with unilateral brain damage on face recognition tasks. Cortex **4,** 17–34.

De Renzi E. (1986), Current issues on prosopagnosia. In: Aspects of Face Processing (H. D. Ellis, M. A. Jeeves, F. Newcombe and A. Young, eds.). NATO Advanced Science Institute. Series. Martinus Nijhoff Publishers, Dordrecht, pp. 243–252.

Desimone R., Albright T. D., Gross C. G. and Bruce C. (1984), Stimulus-selective properties of inferior temporal neurons in the macaque. J. Neurosci. **8,** 2051–2062.

Dodd J. and Perrett D. I. (1997), The effect of caricaturing on learning and recognition of car shapes (in submission).

Farah M. J. (1996), Is face recognition special – evidence from neuropsychology. Behav.Brain Res. **76,** 181–189.

Flin R. and Dziurawiec S. (1989), Development factors in face processing. In: Handbook of Research on Face Processing. North Holland, Amsterdam, pp. 335–350.

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

Foldiak P. and Young,M. P. (1995), Sparse coding in the primate cortex. In: The Handbook of Brain Theory and Neural Networks (M. A. Arbib, ed.). Massachusetts Institute of Technology Press, pp. 895–898.

Fujita I., Tanaka K., Ito,M. and Cheng K. (1992), Columns for visual features of objects in monkey inferotemporal cortex. Nature **36,** 343–346.

Fukushima K. (1980), Neocognitron: a self-organising neural network model for a mechanism of pattern recognition unaffected by shift in position. Biolog. Cybern. **36,** 193–202.

Gauthier I., Anderson A. W., Tarr M. J., Skudlarski P. and Gore J. C. (1997), Levels of categorization in visual recognition studied using functional magnetic resonance imaging. Curr. Biol. **7,** 645–651.

Gauthier I. and Tarr M. J. (1997), Becoming a "greeble" expert: exploring mechanisms for face recognition. Vision Res. **37,** 1673–1682.

Gochin P. M., Miller E. K., Gross C. G. and Gerstein G. L. (1991), Functional interactions among neurons in inferior temporal cortex of the awake macaque. Exp. Brain Res. **84,** 505–516.

Gross C. G., (1992), Representation of visual stimuli in inferior temporal cortex. Phil. Trans. R. Soc. London B **335,** 3–10.

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.

Guzman A. (1968), Decomposition of a scene into three-dimensional bodies. AFIRS Fall Joint Conferences **33,** 291–304.

Harries M. H. and Perrett D. I. (1991), Visual processing of faces in the temporal cortex: physiological evidence for a modular organization and possible anatomical correlates. J. Cognit. Neurosci. **3,** 9–24.

Hasselmo M. E., Rolls E. T., Baylis G. C. and Nalwa V. (1989), Object centred encoding by face-selective neurons in the cortex of the superior temporal sulcus of the monkey. Exp. Brain Res. **75,** 417–429.

Hietanen J. K., Perrett D. I., Oram M. W., Benson P. J. and Dittrich W. H. (1992), The effects of lighting conditions on responses of cells selective for face views in the temporal cortex. Exp. Brain Res. **89,** 157–71.

Hubel D. H. and Livingstone M. S. (1990), Color and contrast sensitivity in the lateral geniculate body and primary visual cortex of the macaque monkey. J. Neurosci. **10,** 2223–2237..

Ito M., Tamura H., Fujita I. and Tanaka K. (1995), Size and position invariance of neuronal responses in monkey inferotemporal cortex. J. Neurophysiol. **73,** 218–226.

Jolicoeur P. (1992), Identification of Disoriented Objects: A Dual-System Theory. Blackwell, Cambridge, Mass.

Kendrick K. M. and Baldwin B. A. (1987), Cells in temporal cortex of conscious sheep can respond preferentially to the sight of faces. Science **236,** 448–450.

Knierim J. J. and Van-Essen D. C. (1992), Neuronal responses to static texture patterns in area V1 of the alert macaque monkey. J. Neurophysiol. **67,** 961–980.

Kobatake E. and Tanaka K. (1994), Neural selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. J. Neurophysiol. **71,** 856–867.

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

Komatsu H. and Ideura Y. (1993), Relationships between color, shape and pattern selectivities of neurons in the inferior temporal cortex of the Monkey. J. Neurophysiol. **70,** 677–694.

Konorski J. (1967), Integrative Activity of the Brain. University of Chicago Press, Chicago Illinois.

Logothetis N. K., Pauls J., Bulthoff H. H. and Poggio T. (1994a), View-dependent object recognition by monkeys. Curr. Biol. **4,** 401–414.

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

Lowe D. G. (1987), Three-dimensional object recognition form single two-dimensional images. Artificial Intelligence **31,** 355–395.

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

McClelland J. L. and Rummelhart and the PDP Research Group (1988), Parallel Distributed Processing Explorations in the Microstructure of Cognition, Vol. **2,** Psychological and Biological Models. MIT Press, Massachusetts.

McNeil J. E. and Warrington E. (1994), Prosopagnosia: a face specific deficit? Exp. Psycholog. Soc., Sussex, 3–5 July 1991. Meeting Abstracts, 47.

Miyashita Y. and Chang H. S. (1988), Neuronal correlate of pictorial short-term memory in the primate temporal cortex. Nature **331,** 68–70.

Miyashita Y., Date A. and Okuno H. (1993), Configurational encoding of complex visual forms by single neurons of monkey temporal cortex. Neuropsychologia **31,** 1119–1131.

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

Nobre A. C., Allison T. and McCarthy G. (1994), Word recognition in the human inferior temporal lobe. Nature **372,** 260–263.

Ono *et al.* (1992), Amygdala and learned aversive.

Oram M.W and Perrett D. I. (1992), The time course of responses of cells selective for faces in the temporal cortex. J. Neurophysiol. **68,** 70–84.

Oram M. W., Perrett D. I. and Hietanen J. K. (1993), Directional tuning of motion-sensitive cells in the anterior superior temporal polysensory area of the macaque. Exp. Brain Res. **97,** 274–294.

Oram M. W. and Perrett D. I. (1994), Modeling visual recognition from neurobiological constraints. Neural Networks **7,** 945–972.

Oram M. W. and Foldiak P. (1996), Learning generalization and localization: competition for stimulus type and receptive field. NeuroComputing **11,** 297–321.

Oram M. W. and Perrett D. I. (1996), Integration of form and motion in the anterior superior temporal polysensory area (STPa) of the macaque monkey. J. Neurophysiol. **76,**, 109–129.

Perrett D. I., Mistlin A. J. and Chitty A. J. (1987), Visual neurons responsive to faces. TINS **10**, 358–364.

Perrett D. I., Harries M. H., Bevan R., Thomas S., Benson P. J., Mistlin A. J., Chitty A. J., Hietanen J. K. and Ortega J. E. (1989a), Frameworks of analysis for the neural representation of animate objects and actions. J. Exp. Biol. **146,** 87–114.

Perrett D. I., Hietanen J. K., Oram M. W. and Benson P. J. (1992), Organization and functions of cells responsive to faces in the temporal cortex. Phil. Trans. R. Soc. London B **335,** 23–30.

Perrett D. I., Mistlin A. J., Chitty A. J., Smith P. A. J., Potter D. D., Broennimann R. and Harries M. H. (1988), Specialised face processing and hemispheric asymmetry in man and monkey: evidence from single unit and reaction time studies. Behav. Brain Res. **29,** 245–258.

Perrett D. I. and Oram M. W. (1993), The neurophysiology of shape processing. Image and Visual Computing **11,** 317–333.

Perrett D. I. and Emery N. J. (1994), Understanding the intentions of others from visual signals: neurophysiological evidence. Curr. Psych. Cognit. **13,** 683–694.

Perrett D. I., Oram M. W., Harries M. H., Bevan R., Hietanen J. K., Benson P. J. and Thomas S. (1991), Viewer-centred and object-centred coding of heads in the macaque temporal cortex. Exp. Brain Res. **86,** 159–173.

Perrett D. I., Rolls E. T. and Caan W. (1982), Visual neurons responsive to faces in the monkey temporal cortex. Exp. Brain Res. **47,** 329–342.

Perrett D. I., Smith P. A. J., Potter D. D., Mistlin A. J., Head A. S., Milner A. D. and Jeeves M. A. (1984), Neurons responsive to faces in the temporal cortex: Studies of functional organization, sensitivity to identity and relation to perception. Human Neurobiol. **3,** 197–208.

Perrett D. I., Smith P. A. J., Potter D. D., Mistlin A. J., Head A. S., Milner A. D. and Jeeves M. A. (1985), Visual cells in the temporal cortex sensitive to face view and gaze direction. Proc. R. Soc. London B **223,** 293–317.

Perrett D. I., Oram M. W., Wachsmuth E. and Emery N. J. (1995a), Understanding the behaviour and 'minds' of others from their facial and body signals: Studies of visual processing within the temporal cortex. In: Emotion, Memory and Behaviour: Studies on Human and Non-Human Primates (T. Nakajima and T. Ono, eds.). Taniguchi Symposia on Brain Sciences **18,** Tokyo. Japan Scientific Societies Press, Japan, pp. 155–167.

Perrett D. I. Benson P. J. Hietanen J. K. Oram M. W. and Dittrich W. H. (1995b), When is a face not a face? Correlations between perception and single neurones. In: The Artful Eye (R. Gregory, J. Harris, P. Heard and D. Rose, eds.). Oxford University Press, Oxford.

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

Porrill J., Pollard S. B., Pridmore T. P., Bowen J. B., Mayhew J. E. W. and Frisby J. P. (1988), TINA: a 3D vision system for pick and place. Image and Vision Computing **6,** 91–99.

Puce A. Allison T., Gore J. C. and McCarthy G. (1996a), Face-sensitive regions in human extrastriate cortex studied by functional MRI. J. Neurophysiol. **74,** 1192–1199.

Puce A., Allison T., Asgari M., Gore J. C. and McCarthy G. (1996b), Differential sensitivity of human visual-cortex to faces letter strings and textures – a functional magnetic resonance imaging study. J. Neurosci. **16,** 5205–5215.

Rhodes G., Brennan S. E. and Carey S. (1987), Identification and ratings of caricatures: implications for mental representations of faces. Cognit. Psych. **19,** 473–497.

Rhodes G. and McLean I. (1990), Distinctiveness and expertise effects with homogeneous stimuli: towards a model of configural coding. Perception **19,** 773–794.

Richmond B. J. and Optican L. M. (1992), The structure and interpretation of neuronal codes in the visual system. In: Neural Networks for Perception (H. Wechsler, ed.). Academic Press Ltd., San Diego, CA, USA.

Rolls E. T. and Baylis G. C. (1986), Size and contrast have only small effects on the responses to faces of neurons in the cortex of the superior temporal sulcus of the monkey. Exp. Brain Res. **65,** 38–48.

Rolls E. T. and Tovee M. J. (1995), Sparseness of neuronal represntation of stimuli in the primate temporal visual-cortex. J. Neurophysiol. **73,** 713–726.

Rolls E. T., Treves A. and Tovee M. J. (1997), The representational capacity of the distributed encoding of information provided by populations of neurons in primate temporal visual cortex. Exp. Brain Res. **114,** 149–162.

Sakai K. and Miyashita Y. (1991), Neural organization for the long-term memory of paired associates. Nature **354,** 152–155.

Sakata H., Taira M., Murata A. and Mine S. (1995), Neural mechanisms of visual guidance of hand action in the parietal cortex of the monkey. Cereb. Cortex **5,** 429–438.

Saund E. (1992), Putting knowledge into visual shape representation. Artificial Intelligence **54,** 71–119.

Schwartz E. L., Desimone R., Albright T. D. and Gross C. G. (1983), Shape recognition and inferior temporal neurons. Proc. Natl. Acad. Sci. USA **80,** 5776–5778.

Seibert M. and Waxman A. (1992a), Adaptive 3-D object recognition from multiple views. IEEE Transactions on Pattern Analysis and Machine Intelligence **14,** 107–124.

Seibert M. and Waxman A. (1992b), Learning and recognizing 3D objects from multiple views in a neural system. In: Neural Networks for Perception. Academic Press, Inc., 426–444.

Sirovitch L. and Kirby M. (1987), Low-dimensional proceedure for the characterization of human faces. J. Opt. Soc. America A **4,** 519–524.

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

Tovee M. J., Rolls E. T. and Azzopardi A. (1994), Translation invariance in the responses to faces of single neurones in the primate temporal cortical areas of the alert macaque. J. Neurophysiol. **72,** 1049–1060.

Turk M. and Pentland A. (1991), Eigenfaces for recognition. J. Cognit. Neurosci. **3,** 71–86.

Ullman S. (1989), Aligning pictorial descriptions: an approach to object recognition. Cognition **32,** 193–254.

Ungerleider L. G. and Mishkin M. (1982), Two cortical visual systems. In: Analysis of Visual Behaviour. MIT Press, Cambridge Massachusetts, 549–585.

Verfaillie K. (1992), Variant points of view on viewpoint invariance. Can. J. Psychol. **46,** 215–236.

Vetter T., Hurlbert A. and Poggio T. (1995), Models of 3D object recognition: invariance to image transformations. Cerebral Cortex **5**, 261–269.

Wachsmuth E., Oram M. W. and Perrett D. I. (1994), Recognition of objects and their component parts: responses of single units in the temporal cortex of the macaque. Cereb. Cortex **4,** 509–522.

Wachsmuth E. and Perrett D. I. (1995), Generalising across object orientation and size. In: Perceptual Constancies (V. Walsh and S. Butler, eds.). Oxford University Press, Oxford.

Wachsmuth E. and Perrett D. I. (1997), The physiology of shape generalisation (size and orientation). In: Perceptual Constancies: Why things Look as they do (V. Walsh and J. Kulikowski, eds.). Cambridge University Press, Cambridge, UK. (in press).

Warrington E. K. (1982), Neuropsychological studies of object recognition. Phil. Trans. R. Soc. London B **298,** 15–33.

WarringtonE. K. and James, M. (1986), Visual object recognition in patients with right hemisphere lesions: axes or features? Perception **15,** 355–366.

Watt R. J. (1996), Critical operations in low-level human vision International. J. Image Systems Technol. **7,** 65–77.

Weiskrantz L. and Saunders R. C. (1984), Impairments of visual object transforms in monkeys. Brain **107,** 1033–1072.

Wicklegren W. A. (1969), Context-sensitive coding, associative memory and serial order in (speech) behaviour. Psycholog. Rev. **76,** 1–15.

Yamane S., Kaji S. and Kawano K. (1988), What facial features activate face neurons in the inferotemporal cortex of the monkey. Exp. Brain Res. **73,** 209–214.

Yin R. K. (1969), Looking at upside-down faces. J. Exp. Psychol. **81,** 141–145.

Yin R. K. (1970), Face recognition by brain-injured patients, a dissociable ability? Neuropsychologia **8,** 395–402.

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