Neurocomputational bases of object and face recognition

IRVING BIEDERMAN AND PETER KALOCSAI

University of Southern California, Department of Psychology and Neuroscience Program, Hedco Neuroscience Building, Los Angeles, CA 90089-2520, USA

SUMMARY

A number of behavioural phenomena distinguish the recognition of faces and objects, even when members of a set of objects are highly similar. Because faces have the same parts in approximately the same relations, individuation of faces typically requires specification of the metric variation in a holistic and integral representation of the facial surface. The direct mapping of a hypercolumn-like pattern of activation onto a representation layer that preserves relative spatial filter values in a two-dimensional (2D) coordinate space, as proposed by C. von der Malsburg and his associates, may account for many of the phenomena associated with face recognition. An additional refinement, in which each column of filters (termed a 'jet') is centred on a particular facial feature (or fiducial point), allows selectivity of the input into the holistic representation to avoid incorporation of occluding or nearby surfaces. The initial hypercolumn representation also characterizes the first stage of object perception, but the image variation for objects at a given location in a 2D coordinate space may be too great to yield sufficient predictability directly from the output of spatial kernels. Consequently, objects can be represented by a structural description specifying qualitative (typically, non-accidental) characterizations of an object's parts, the attributes of the parts, and the relations among the parts, largely based on orientation and depth discontinuities (as shown by Hummel & Biederman). A series of experiments on the name priming or physical matching of complementary images (in the Fourier domain) of objects and faces documents that whereas face recognition is strongly dependent on the original spatial filter values, evidence from object recognition indicates strong invariance to these values, even when distinguishing among objects that are as similar as faces.

1. INTRODUCTION

We propose a theoretical account of the neural, perceptual, and cognitive differences that are apparent in the individuation of faces and the entry- and subordinate-level classification of objects. After a general theoretical overview, we review some of the behavioural and neural phenomena by which face and object recognition can be contrasted, and then present a neurocomputational account of these differences, with particular attention to the perceptual representation of faces. Finally, original experiments testing a key assumption of this account are described.

2. A THEORETICAL OVERVIEW: FACE AND OBJECT RECOGNITION

The basic theoretical differences that we will propose are diagrammed in figure 1. The object model follows that of Hummel & Biederman (1992) and only a brief overview will be presented here. Specification of the edges at an object's orientation and depth discontinuities in terms of non-accidental properties (NAPs) is employed to activate units that represent simple, viewpoint invariant parts (or *geons*), such as bricks, cones, and wedges. Other units specify a geon's attributes, such as its approximate orientation (e.g. horizontal) and aspect ratio, and still other units specify the relative relations of pairs of geons to each other, such as *top-of, larger-than, end-to-middle-connected*. The separate units associated with a given geon, its attributes, and its relations, are bound (through correlated firing) to a unit termed a geon feature assembly (GFA). A unit representing a geon structural description (GSD) specifying the geons and their relations in a given view of the object can then self-organize to the activity from a small set of GFAs.

Differences in GFAs are usually sufficient to distinguish entry level classes and most subordinate level distinctions that people can make quickly and accurately in their everyday lives. Sometimes the GSDs required for subordinate level distinctions are available at a large-scale, as in distinguishing a square table from a round table. Sometimes they are at a small-scale, as when we use a logo to determine the manufacturer of a car.

Although there are some person individuation tasks that can be accomplished by the information specified by a GSD ('Steve is the guy wearing glasses'), generally we will focus on cases where such easy information as a distinctive GSD or texture field ('Steve is the guy with freckles') is insufficient. We will argue that the information required for general purpose face recognition is holistic, surface-based, and metric, rather than parts-based, discontinuous, and non-accidental (or qualitative), as it is with objects. A representation that preserves

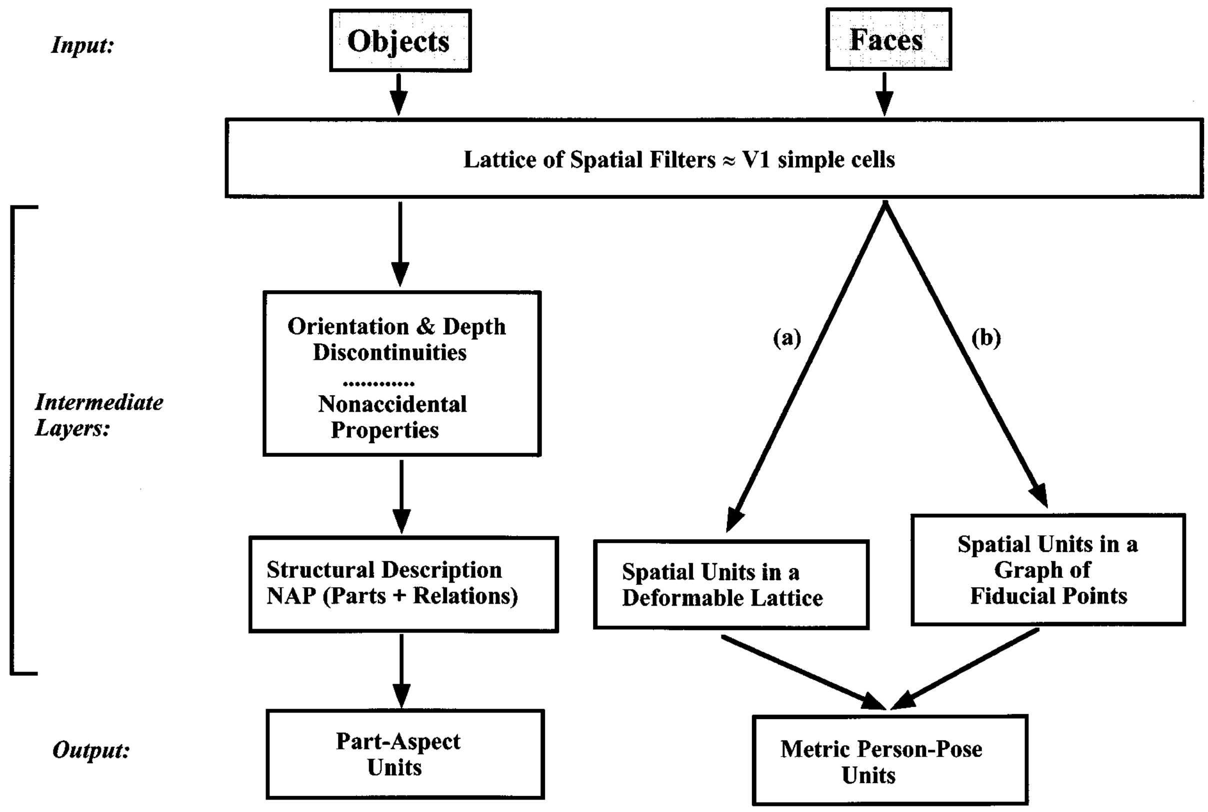

Figure 1. Relations between presumed models of object and face recognition. Both start with a lattice of columns of spatial filters characteristics of six hypercolumns. The object pathway is modelled after Biederman (1987) and Hummel & Biederman (1992) and computes a geon structural description (GSA) which represents the parts and their relations in a view of an object. Both face pathways retain aspects of the original spatial filter activation patterns. In the (*a*) pathway, modelled after Lades *et al.* (1993), the default position of the columns (termed 'jets') of filters is a lattice similar to that of the input layer, but which can be deformed to provide a best match to a probe image. In the (*b*) pathway, modelled after Wiskott *et al.* (1997), the jets are centred on a particular facial feature, termed a fiducial point.

the relative scale of the original spatial filter values in a coordinate space that normalizes scale and position may allow specification of the metric variation in that region for determining the surface properties of a face. The coordinate system is preserved because the locations of facial characteristics are highly predictable from a given pose of a face. For objects they are not. (What is in the upper, right hand part of an object?) Relative (cycles/face) rather than absolute (cycles/degree) allows invariance over size changes of the face.

We consider two recent proposals by C. von der Malsburg and his associates for face representation. The first (see figure 1*a*), is described by Lades *et al.* (1993). This system maps columns (or 'jets') of V1-like spatial filter activation values to images of faces or objects. The jets are arranged in a hypercolumn-like lattice where they are stored. This stored lattice serves as a representation layer, and is then matched against probe faces or objects by correlating the filter values of the original lattice against a new lattice that has been allowed to deform to achieve its own best match. The second model (figure 1*b*), proposed by Wiskott *et al.* (1997), positions each of the jets not on the vertices of a rectangular lattice but to assigned 'fiducial points' on a face, such as the left corner of the mouth. These face models will be considered in more detail in a later section.

3. DISTINGUISHING FACE AND OBJECT RECOGNITION: EMPIRICAL RESULTS

One problem in distinguishing face and object recognition is that there are a large number of tasks that can be loosely described as 'recognition'. We will consider the identification of an image of a face to the criterion of individuation, and that of an object with its assignment to its basic level or common subordinate level class.

(*a*) *Behavioural differences*

Table 1 lists eight behavioural differences between face and object recognition. See Bruce (1988) and Bruce & Humphreys (1994) for more extensive reviews. These will be considered in turn with respect to the different properties that should be captured by a particular representation.

(i) *Configural effects*

Tanaka & Farah (1993) trained their subjects to recognize a set of Identikit faces, each of which had a different pair of eyes, nose, and mouth. In testing, they presented pairs of images that differed in the shape of a single face part, the eyes, nose, or mouth (figure 2). In one condition, only a pair of face parts was shown, for example, two slightly different noses. In the other, the

Table 1. *Some differences in the recognition of faces and objects*

	faces	objects
configural effects?	yes	no
basis of expertise	holistic representation	feature discovery
differences verbalizable?	no	yes
sensitive to contrast polarity?	yes	no
sensitive to illumination direction?	yes	no
sensitive to metric variation?	yes	slightly
sensitive to rotation in depth?	yes	no, within part aspects (*ca.* 60°)
sensitive to rotation in the plane?	yes	slightly

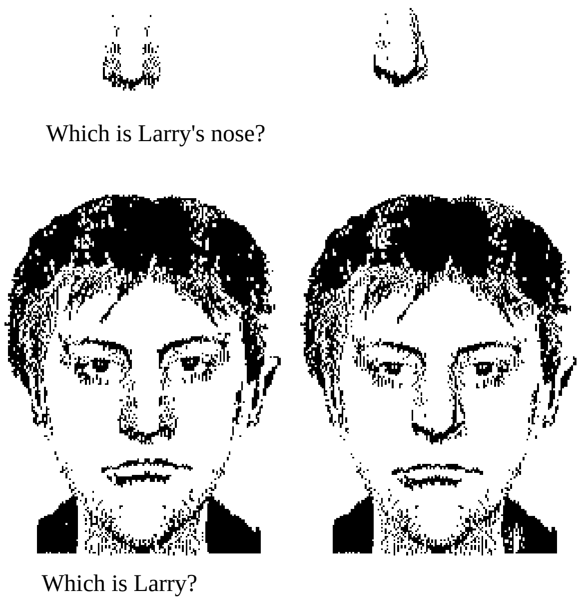

Figure 2. Sample stimuli from Tanaka & Farah's (1993) single feature and whole face conditions. In the single feature condition, subjects were presented with, for example, the upper pair of noses and were to judge, 'Which is Jim's nose?'. In the whole face condition, the subjects were presented with a pair of faces whose members were identical but for a single feature, the one shown in the feature condition, and they had to judge, 'Which is Jim?'. Used with permission.

stimuli were part of a context of a whole face; one with one of the noses, the other with the other nose. The subjects did not know which face part might differ when they viewed a complete face. Remarkably, the context of the face facilitated detection of the difference. The facilitation from the presence of the context was not found for non-face objects, such as a house, or when the faces were inverted.

(ii) *Expertise*

Good face-recognizers use the whole face, although with unfamiliar faces the overall external shape and hairline receive extremely high weight (Young *et al.* 1985). When asked to describe a picture of a person's face, these individuals will often refer to a famous person, perhaps with some modification in the descriptions (Cesa 1994). Poor recognizers tend to pick a single feature or small set of distinctive features. As people age, face recognition performance declines. This decline is marked by a qualitative shift in the representation such that older people, like poor face-recognizers in general, search for distinctive features. Prosopagnosics often report a distinctive feature strategy as well (Davidoff 1988).

In contrast to the holistic processing of faces, expertise in the identification of an object from a highly similar set of objects is most often a process of discovery or instruction as to the location and nature of small differences that reliably distinguish the classes (Gibson 1947; Biederman & Shiffrar 1988). If such features are not present, then performance is often slow and error prone (Biederman & Subramaniam 1997). Gibson (1947) described the consequences of attempting to teach aircraft identification during World War II by 'total form' versus distinctive features of the parts:

'Two principal observations made by the instructors who took part in the experiment are of some bearing on the question of the two methods under consideration. The impression was obtained by all three of the instructors, at about the time the course was two-thirds completed, that the group taught by emphasis on total form was definitely "slipping" in comparison with the other group. The second observation was that a single question was insistently and repeatedly asked by the cadets in the group taught by emphasis on total form. This question was "How can I distinguish between this plane and the one which resembles it closely (e.g. the C-46 and the C-47)?"' (Gibson 1947, p. 120.)

Whether still more extensive training on non-face stimuli can lead to face-like processing is an open issue. Gauthier & Tarr (1997) provided extensive training to some of their subjects in distinguishing among a family of 'greebles' a set of stimuli composed of three rounded parts—a base, body, and head—one on top of the other, with protrusions that are readily labelled penis, nose, and ears. Unfortunately, these rounded, bilaterally symmetrical creatures closely resemble humanoid characters, such as the Yoda (in *Return of the Jedi*). This characteristic of the stimuli is termed unfortunate because even if face- or body-like results were obtained from the training, it would be unclear whether the stimuli engaged face or body processing because of their physical resemblance to people. In the other direction, some of the differences in body parts distinguishing groups or sex of the greebles appeared to be non-accidental, the presence of a brick at the base of a body part, for example, a characteristic that would not distinguish faces in general. Despite Gauthier & Tarr's conclusion that they were able to mimic face processing with their training, their results were clearly inconsistent with face-like processing. For example, there were no effects of inverting the greebles or testing for the identity of a part outside of its greeble context. Gauthier & Tarr's results are, perhaps, more consistent with the viewpoint-invariant recognition of objects by geons than they are to face recognition.

(iii) *Differences verbalizable?*

People find it exceedingly difficult to express verbally the differences between two similar faces. This fact is well known to the chagrin of police investigators interviewing witnesses. When asked to describe an object, however, people readily name its parts and provide a characterization of the shape of these parts in terms of NAPs (Tversky & Hemenway 1984; Biederman 1987). Within highly similar shape classes, such as Western American male Quail, people will spontaneously employ local shape features that closely correspond to those specified—verbally—by the bird guides (Biederman *et al*. 1997). Gibson (1947) concluded that the problem of training aircraft spotters was best solved by informing them of the non-accidental differences in the shapes of parts. It was a simple matter for Gibson to construct an outline—in words—providing this information.

(iv) *Sensitivity to contrast polarity and illumination direction?*

Whereas people have great difficulty in identifying a face from a photographic negative or when illuminated from below (Johnston *et al*. 1992), there is little, if any, effect of reversing the polarity of contrast of a picture of an object (Subramaniam & Biederman 1997). Viewing an object at one polarity provides essentially the same information about the structure of the object as does the other polarity. A major reason for this difference between faces and objects is that, as noted previously, object recognition is largely based on distinguishable parts based on differences in NAPs of edges marking orientation and depth discontinuities. The position of these edges and their non-accidental values (e.g. straight or curved) are unaffected by contrast reversal. Individuating faces typically requires metric differences that may be specified in terms of the convexities and concavities that characterize a facial structure. A change in contrast polarity would reverse the interpretation of the luminance and shadow gradients that are employed to determine the convexity or concavity of a smooth surface. A similar explanation may account for some of the increased difficulty in identifying faces when they are illuminated from below as this would violate the strong assumption that illumination is from above.

(v) *Metric variation?*

Metric properties are those such as aspect ratio or degree of curvature that vary with the orientation of the object in depth. Such properties are to be contrasted with NAPs, such as whether an edge is straight or curved, which are only rarely affected by slight changes in the viewpoint of an object. Other NAPs are the vertices that are formed by coterminating lines and whether pairs of edges are approximately parallel or not, given that the edges that are not greatly extended in depth.

Before looking at figure 3 (from Cooper & Wojan 1996), please cover the left and center columns. In looking at the right column, the reader can assess for himself or herself how modest variation in the metrics of a face can result in marked interference in the recognition of that face (see also Hosie *et al*. 1988). In these images of celebrities, the eyes have been raised. A similar variation in the length (and, hence, aspect ratio) of an object part, as illustrated in figure 4, has little or no effect in the assignment of objects to classes. As long as the relative relations, such as *larger-then* or *above*, between parts are not changed by altering a part's length, the effects of the variation appear to be confined to that part, rather than affecting the object as a whole. Unlike what occurs with the holistic effects with faces, there is little effect of the variation on a metric attribute of a part in the recognition of objects. Cooper & Biederman (1993) presented two images of simple, two-part objects (illustrated in figure 4) sequentially. Subjects had to judge whether the two objects had the same name. When the objects differed in the aspect ratio of a part, reaction times (RTs) and error rates were only slightly elevated compared to when the images were identical. A change in a NAP produced a much larger interference effect on the matching.

(vi) *Rotation in depth*

If objects differ in NAPs, then little or no cost is apparent when they are rotated in depth, as long as the same surfaces are in view (Biederman & Gerhardstein 1993). In contrast, when the differences are in metric properties, such as aspect ratio or degree of curvature, then marked rotation costs are observed (e.g. Edelman 1995). The robustness of the detection of non-accidental differences under depth rotation is not simply a function of greater discriminability of NAPs compared to metric properties. Biederman & Bar (1995) equated the detectability of metric and non-accidental part differences in a sequential same–different matching task with novel objects. Presenting the objects at different orientations in depth had no effect on the detectability of non-accidental differences. When easy non-accidental cues are eliminated, such as glasses, facial hair, and the hairline, even modest rotations of faces, from 20° left to 40° right, as illustrated in figure 7 (middle row), can result in marked increases in RTs and error rates in their matching (Kalocsai *et al*. 1994).

(vii) *Rotation in the plane*

Recognizing an upside-down face is extremely difficult relative to identifying an upside-down object, such as a chair (e.g. Yin 1969; Johnston *et al*. 1992; Jolicoeur 1985). According to the Hummel & Biederman (1992) network, turning an object upside down would leave most of the units coding the structural description intact, affecting only the relations *top-of* and *below*. Consequently, only a small effect for objects would be expected. Some of the large effect of inversion with face photographs lies in the misinterpretation of luminance gradients where the light source is typically assumed to be coming from above. But when the light source is controlled, there still remains a large cost to viewing a face upside down (Johnston *et al*. 1992; Enns & Shore 1997), as expected from their representation in a 2D coordinate space.

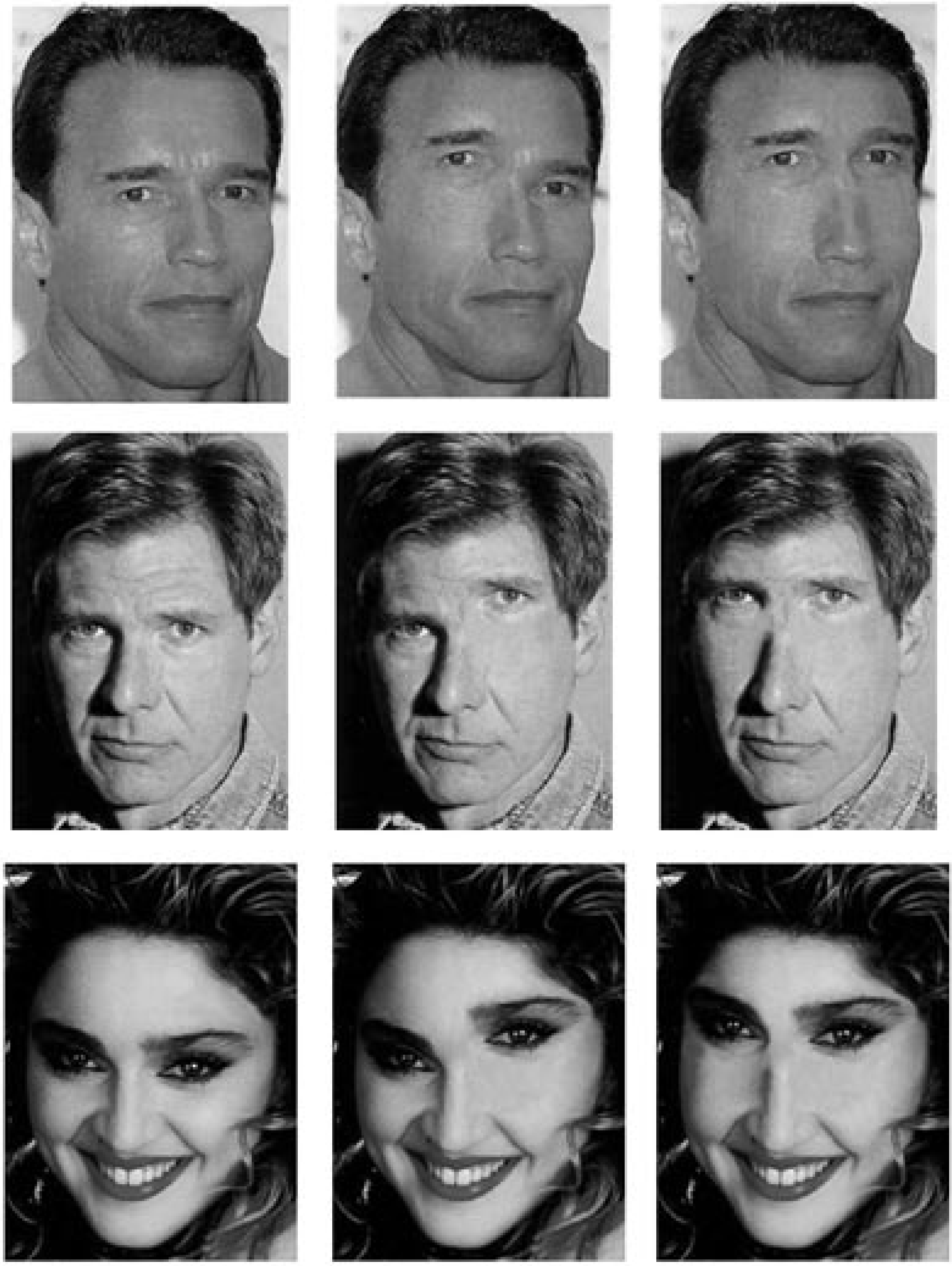

Figure 3. Sample stimuli from Cooper & Wojan (1996). Subjects were much worse at identifying the celebrities in the third column, where both eyes were raised, compared to those in the second column where only one eye was raised, despite the greater difficulty in judging the later as a face. Copyright Eric E. Cooper. Used with permission.

(*b*) *Neural differences between faces and objects*

There are several neural differences distinguishing the representation of faces and objects. Only a brief summary will be presented here. (See Grüsser & Landis (1991) for a comprehensive treatment of this general area.)

(i) *Selective impairment: prosopagnosia and object agnosias*

Prosopagnosia, the inability to recognize familiar faces but with a normal or near normal capacity for object recognition, is a well-documented phenomenon, generally associated with lesions to the right, inferior mesial hemispheric (Grüsser & Landis 1991), although

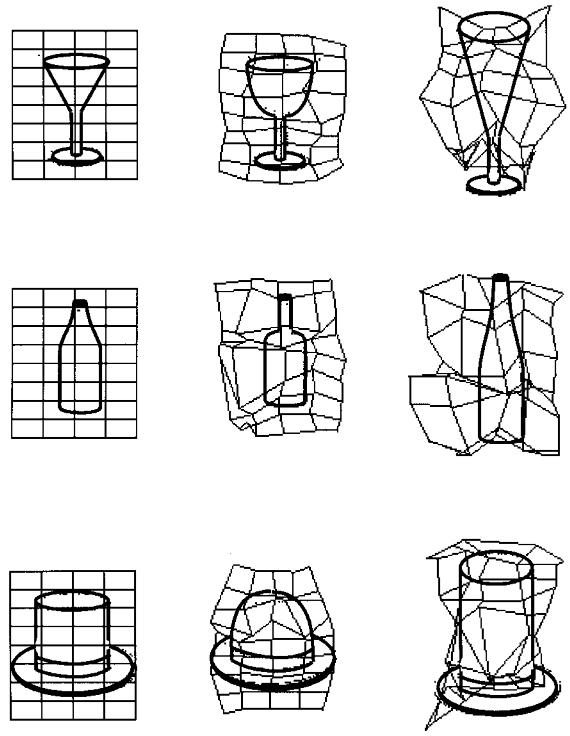

Figure 4. Sample object stimuli from Cooper & Biederman (1993). (*a*) standard; (*b*) NAP change; and (*c*) metric change. Given the standard object on the left, a NAP of only a single part was changed in the objects in the middle column (NAP condition), and that same part was lengthened in the metric condition, illustrated by the objects in the third column. The magnitude of the metric changes were slightly larger than the NAP changes, according to the model of Lades *et al*. (1993). Whereas the differences between metric and standard images were more readily detected when performing a simultaneous physical identity matching task ('Are the objects identical?'), in a sequential object matching task ('Do the objects have the same name?'), a change in a NAP resulted in far more disruption than a change in a metric property.

some (e.g. Damasio *et al*. 1985) have argued that the lesions must be bilateral. Farah (1990) theorized that the underlying continuum in visual recognition extended from holistic processing, which would be required for faces, to the capacity to represent multiple shapes (or parts), which would be typified by the integration of letters into words in reading. She surmised that bilateral parietal and superior occipital lesions affected holistic processing, whereas lesions to the left inferior temporal-occipital region (including the fusiform) resulted in a condition, ventral simultagnosia, in which the patient could not simultaneously process multiple parts of an object or letters of a word (alexia). Object recognition, according to Farah, employs both types of processing, so object agnosia should be accompanied by either prosopagnosia or alexia. Two recent cases have confirmed that a loss of the capacity for parts-based representation need not interfere with face recognition (Rumiati *et al*. 1994; Behrmann *et al*. 1992). We interpret these findings (and those described in the section on expertise) as evidence that object recognition does not generally entail holistic processing.

(ii) *Imaging studies*

Recent facial magnetic resonance imaging (fMRI) studies in humans have given clear evidence for object and shape specific regions in the occipital cortex. Tootell *et al*. (1996) have documented an area just anterior to V4v and partly overlapping with regions of the fusiform, termed the lateral occipital (LO), that gives vigorous responses to interpretable faces and objects even when they are unfamiliar, such as an abstract sculpture, but not to these stimuli when they have been rendered into textures as, for example, quantized blocks characteristic of the 'Lincoln' illusion or in gratings, texture patterns, or highly jumbled object images. In contrast to LO, V4 does not show this specificity to objects as compared to textures. The LO is therefore sensitive to shapes—faces or objects—that have an interpretable structure rather than being characterizable as a texture pattern. More anterior regions in the ventral pathway such as IT are sensitive to the familiarity of the objects, as described in the next section. That the LO's responsivity is unaffected by familiarity suggests that it may be a region where shape descriptions—even novel ones—are created. A number of fMRI and positron emission tomography (PET) studies have demonstrated that the processing of faces and objects activate different loci in or near the LO. These areas are generally consistent with the results of the lesion work, showing greater posterior right hemisphere activity, particularly in the fusiform gyrus, for face processing and greater left hemisphere activity for object processing (Kanwisher *et al*. 1996; Sergent *et al*. 1992, 1994). The two PET studies by Sergent *et al*. are noteworthy in showing virtually identical loci for the differential activity of judging whether a face was that of an actor. The control task was one of judging whether the orientation of a gratings was horizontal or vertical.

(iii) *Single unit recording*

It is well established that individual IT cells can be found that are differentially tuned either to faces or to complex object features, but not both (e.g. Baylis *et al*. 1987; Kobatake & Tanaka 1994; Young & Yamane 1992). However, as recently argued by Biederman *et al*. (1997), it is likely that these IT cells are not involved in the initial perceptual description of an image—rather that they suggest that this is accomplished by the LO or in the area immediate anterior to it—but, instead, in coding episodic memories *following* perception. Because these experiences include contribution of the dorsal system in which position, size, and orientation of the stimulus is specified, it is not surprising to find cells that are tuned to the specific orientations and characteristics of the trained stimuli (e.g. Logothetis *et al*. 1994). That IT may not be involved in the perceptual recognition of a face or object is suggested by the requirement of an interval between stimulus presentation and testing in order to show any deficits in object processing of macaques who have undergone bilateral ablation of IT (Desimone & Ungerleider 1989). However, the differential tuning of IT cells to faces and complex object features indicates that these two classes of stimuli are distinguished neurally. A given IT face cell does not fire in an all-or-nothing fashion to a given face, but participates in a population code to that face by which the firing of the cell is modulated by the specific characteristics of the face (Young & Yamane 1992; Rolls 1992). Young & Yamane showed that the code for macaques looking at pictures of men could be summarized by two dimensions, one coding the width of the face and one the distance of the pupil of the eye to the hairline. Somewhat remarkably, as noted earlier, these same two dimensions characterize human performance with unfamiliar faces.

(iv) *Universal classes of facial attributes*

All cultures appear to process faces in highly similar ways. Faces are not only processed for identity, but for the information they provide about emotion, age, sex, direction of gaze, and attractiveness. Different areas mediate at least some of these attributes. Cells tuned to differences in emotional expression and direction of gaze are found in the superior temporal sulcus in the macaque, an area different from the IT locus of the units that contribute to a population code that can distinguish identity. Prosopagnosics can often readily judge these other attributes, e.g. sex, age, etc., as we have recently witnessed in our laboratory. To the extent that these areas are segregated from those for object recognition, we have additional evidence supporting the faceobject distinction. However, it is not clear to what extent, if any, these attributes contribute to face individuation.

4. A THEORY OF THE PERCEPTUAL RECOGNITION OF FACES

A biologically inspired face recognition system developed by Christoph von der Malsburg and his associates (Lades *et al*. 1993; Wiskott *et al*. 1997) suggests a theoretical perspective from which many of the phenomena associated with face perception described in the previous section might be understood. The fundamental representation element is a column of multiscale, multiorientation spatial (Gabor) kernels with local receptive fields centered on a particular point in the image. Each column of filters is termed a 'Gabor jet', and each jet is presumed to model aspects of the wavelet-type of filtering performed by a V1 hypercolumn. We will first consider the initial version of the model (Lades *et al*. 1993), which will be referred to as the lattice version. This model can be applied to the recognition of faces and objects, so it has the potential to serve as a device for the scaling of both kinds of stimuli. A more recent version (Wiskott *et al*. 1997), the 'fiducial point' model, incorporates general face knowledge. We will ignore preprocessing stages by which a probe image is translated and scaled to achieve a normalized position and size. Overall illumination levels and contrast are similarly normalized.

(*a*) *The lattice model*

As illustrated in figure 5, Lades *et al*. (1993) posited a two-layer network. The input layer is a rectangular lattice of Gabor jets. The pattern of activation of the 80 kernels (five scales × eight orientations × two phases, sine and cosine) in each of the jets is mapped

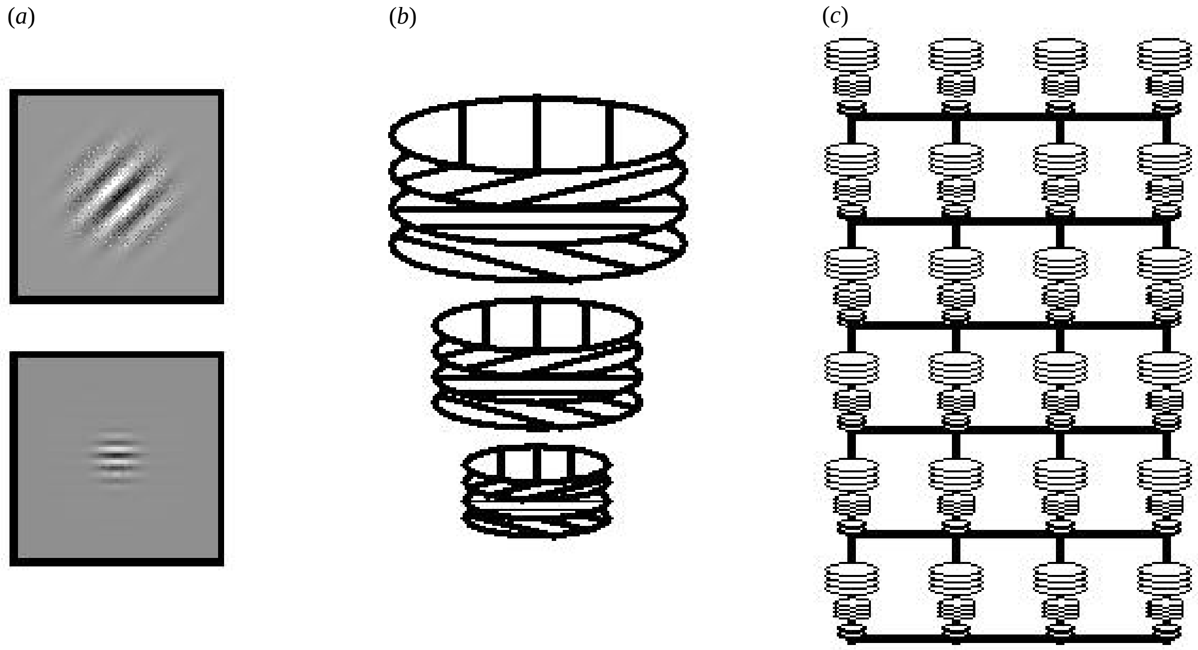

Figure 5. Illustration of the input layer to the Lades *et al*. (1993) network. (*a*) The basic kernels are Gabor filters at different scales and orientations, two of which are shown here (eight orientations, five scales). (*b*) The centre figure illustrates the composition of a jet, with the larger disks representing lower spatial frequencies. There are a set of 8 × 5 filter responses at one image location (only 4 × 3 are represented here). (*c*) This is a model showing a grid of 4 × 6 connected jets. The number of jets, scales, and orientations can be varied.

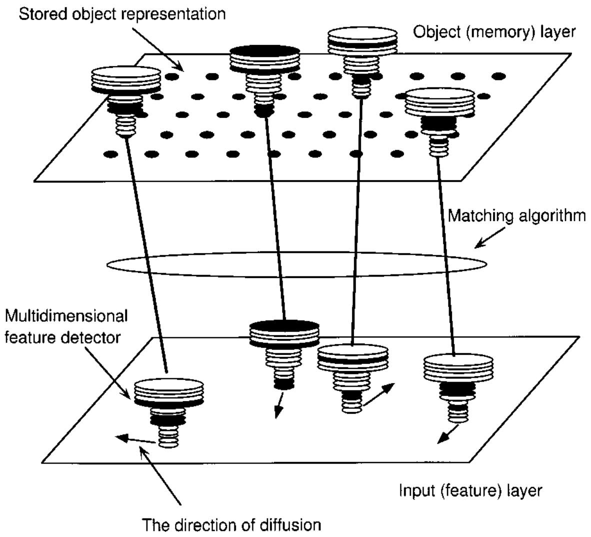

Figure 6. Schematic representation of Lades *et al*.'s (1993) two-layer spatial filter model. The model first convolves each input image with a set of Gabor kernels at five scales and eight orientations and sine and cosine kernels arranged in a 5 × 9 lattice. These values can be varied. The set of kernels at each node in the lattice is termed a 'Gabor jet'. The activation values of the kernels in each jet, along with their positions, are stored for each of the images to form a 'gallery'. The figure shows the diameters of the receptive fields to be much smaller than actual size in that the largest kernels had receptive fields that were almost as large as the whole face.

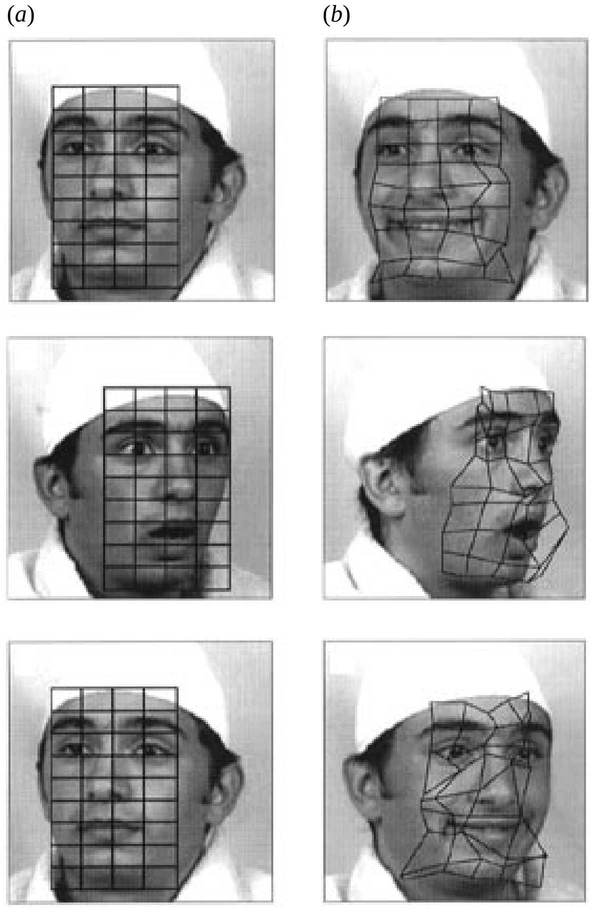

Figure 7. Sample images from the Kalocsai *et al*. (1994) experiment with the Lades *et al*. (1993) lattice deformations superimposed over different pairs of images of the same person. The positioning of the lattice over an original image is shown in the left-hand column (*a*) and the deformed lattice is shown in the right-hand column (*b*). Top, middle, and bottom rows show changes in expression, orientation (60°), and both expression and orientation, respectfully. The similarities as determined by the Lades *et al*. (1993) model correlated highly with performance in matching a pair of images when they were at different orientations and expressions (Kalocsai *et al*. 1994).

onto a representation layer, identical to the input layer, that simply stores the pattern of activation over the kernels from a given image. An arbitrary large number of facial images can be stored in this way to form a gallery.

Matching of a new image against those in the gallery is performed by allowing the jets (in either the probe or a gallery image) to independently diffuse (gradually change their positions) to determine their own best fit, as illustrated by the arrows on the jets in the input layer. The diffusion typically results in distortion of the rectangular lattice, as illustrated in figures 6 and 7. The similarity of two images is taken to be the sum of the correlation in corresponding jets of the magnitudes of activation values of the 80 corresponding kernels. The correlations (range 0 to 1) for each pair of jets is the cosine of the angular difference between the vectors of the kernels in an 80-dimensional space. (If the values are identical, the angular difference will be 0° and the cosine will be 1. A 90° (orthogonal) difference in angles will be 0.00.) The correlations over the jets are summed to get a total similarity score, expressed as a proportion of the maximum score. Figure 7 illustrates distortion of the lattice as a person changes expression, orientation, and both expression and orientation. Typically, the greater the deformation of the lattice, the lower the similarity of the match.

Given a test image matched against a number of stored images, the most similar image is taken to be the recognition choice if it exceeds some threshold value. In a set of over 1000 images of different people, test images differing moderately in pose and expression were recognized by a version of the model (incorporating fiducial points) at a 95% accuracy level (Phillips & Rauss 1997). Most important with respect to the model's relevance to biological vision, the model's similarity values for such images were strongly correlated with human performance in judging whether a pair of images depicted the same person's face (P. Kalocsai & I. Biederman, unpublished data).

How well does the model reflect the phenomena associated with faces listed in table 1?

(i) *Rotation effects*

We will first consider the model's handling of rotation effects, particularly rotation in depth, as that is an extremely common source of image variation and we have assessed its effects under well controlled conditions.

Kalocsai *et al*. (1994) had subjects judge whether two sequentially presented faces were of the same or different person in a task resembling that shown in figure 16. The faces could be at different orientations in depth and/or with a different expression, as shown in figure 7. Easy cues, such as facial hair, clothing and the hairline (all stimulus models wore a white bathing cap) were eliminated. A change in the depth orientation of the two poses, such as that shown in the middle row of figure 7, increased RTs and error rates for 'same' trials. The magnitude of this cost was strongly and linearly correlated with the lattice model's similarity values for the pair of pictures, 0.90 for RTs and 0.82 for error rates. That is, the more dissimilar the two figures according to the model, the longer the RTs and error rates for judging them to be the same person. We can consider the effects of depth rotation as a yardstick for determining the model's adequacy for handling other effects.

Turning a face upside down would greatly reduce its similarity to that of the original image. Although it would be a simple matter, computationally, to rotate the coordinate space of the jets to eliminate the effects of planar rotation, the large cost to human recognition performance from inversion suggests that such a transformation is not available to human vision. Given a yardstick of depth rotation, it is an open question whether the same similarity function would also account for the cost of 2D inversion or other variables. That is, would a 60° rotation in depth (around the y-axis) result in as much cost as a 60° rotation in the plane? What would human subjects evidence?

Given that we have a scaling device (namely, the model of Lades *et al*. model), the analysis that could be undertaken to compare rotation in depth to rotation in the plane can be illustrated by Kalocsai *et al*.'s (1994) comparison of the effects of differences in depth orientation to the effects of differences in expression. Kalocsai *et al*. (1994) showed that when the degree of image dissimilarity of two images of the same person produced by differences in depth orientation (holding expression constant) and expression differences (holding depth orientation constant) were equated, the increase in RTs and error rates in responding 'same' were three times greater when the dissimilarity was produced by expression differences than when produced by depth rotation. They modelled this effect by assuming that a classifier for expression, which was also highly correlated with Gabor similarity, would signal a mismatch to a decision stage (same versus different person?) between two face images that differed in expression, even though the images were of the same person. That mismatch signal resulted in the increased cost for faces differing in expression.

(ii) *Configural and verbalization effects*

Contrast variation within any small region of the face would affect all those kernels whose receptive fields included that region. The pattern of activation of the kernels implicitly contains a holistic or configural representation in that the shape of all facial features and their positions with respect to each other are implicitly coded by the activation of the kernels. Indeed, the representation if run with sufficient jets would be equivalent to a picture of a face, and so it does not distinguish contrast variation arising from the shape of facial features from contrast variation arising from translation of those features. It would be impossible to move a region or a feature, or to change a feature, without affecting the coding of a number of kernels from a number of jets. The representation thus becomes integral (Shepard 1964) or non-analytical (Garner 1966) in that it is not decomposed into readily perceivable independent attributes. This spatially distributed population code of activation values of many kernels of varying scales and orientations in a number of different jets thus captures many of the characteristics of what is generally meant by 'holistic representations'. Consistent with human performance, this spatially distributed code would be extraordinarily difficult to verbalize.

(iii) *Lighting, and contrast reversal effects*

Although the model's normalization routines allows its performance to be invariant to overall lighting and contrast levels, a change in the direction of lighting would result in a cost in similarity for the lattice model. It is not clear whether changing the light source vertically, from top to bottom, would result in a greater reduction in similarity, than a change from right to left, nor would the cost of contrast reversal necessarily be as severe as that evidenced in human performance when compared to, say, rotation in depth. There is nothing in the model, at present, that would identify regions on the surface as convex or concave.

(iv) *Metric sensitivity*

Metric variation such as that performed by Cooper & Wojan (1996) in raising the eyes in the forehead would alter the pattern of activation values in the lattice. Although the distortion of the lattice might be sufficient to account for the effects on recognition performance of such an operation, it is not obvious how lattice distortion would handle the much smaller effect of moving only one eye. The need to handle this result and others in this section on metric sensitivity (along with the benefits of improved recognition performance) provide motivation for the model's incorporation of fiducial points.

Another result that is not obviously derived from the lattice model is the extraordinary difficulty in recognizing the components of a face where the upper half is of one famous person and the lower half another, with the upper and lower halves smoothly aligned to constitute a single face (Young *et al*. 1987). When the upper and lower halves are offset it is much easier to identify the component individuals.

A third result is that we experience little distortion of other regions when a face is partially occluded as, for example, when a person holds his chin with his hand. The hand is not seen as part of the face but instead is regarded as another object, with the occluded regions contributing little, if anything, to the perception of the face.

Figure 8. Illustration of the mapping of jets onto fiducial points (the vertices of the triangles) on three images of the same person at different orientations and expressions.

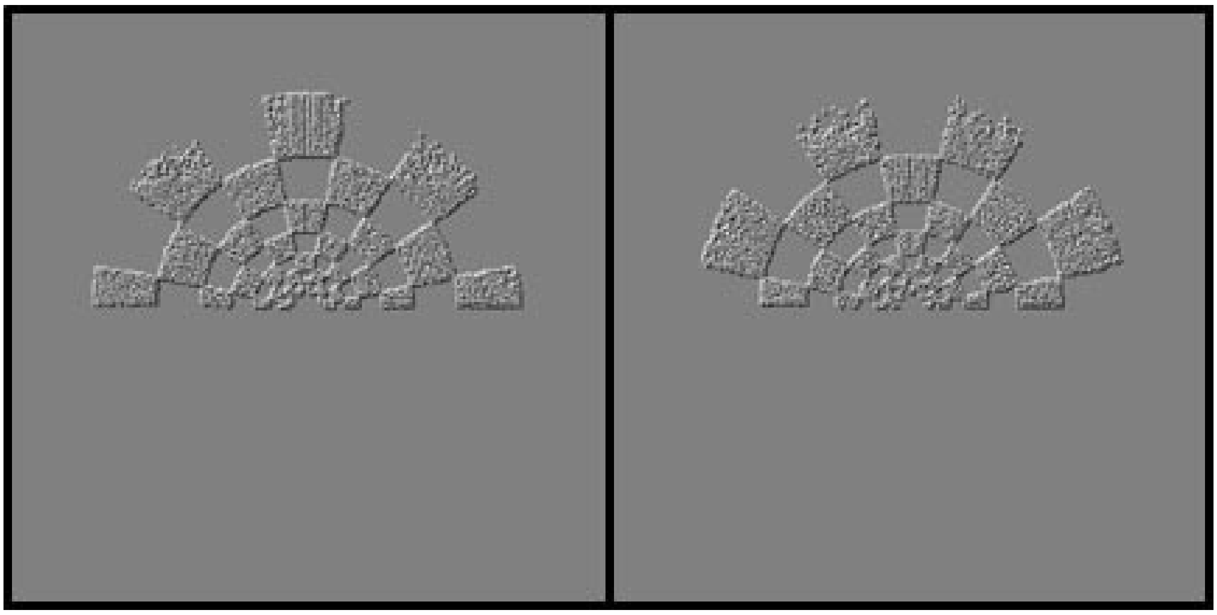

Figure 9. Illustration of how the eight scales × eight orientations were distributed to the members of a complementary pair. If arranged as a checkerboard with rows representing the spatial frequencies and columns the orientations, one image would have the specific scale-orientation values of the red squares, the other member the values of the black squares. Here the checkerboard is shown as two half radial grids, with scale varying with distance from the origin (low to high SF) and orientation varying as shown. (The lower half would continue the upper half.)

(*b*) *Beyond a lattice of spatial features: fiducial points*

We now consider the fiducial point version of the face recognition system so that we can appreciate the potential gains in making facial features explicit by centring designated jets onto salient feature points. We will also consider two other possible extensions of the model: the explicit use of (i) spatial distances and (ii) normative coding by which a face is represented in terms of its deviations from a population norm.

In the fiducial point model (Wiskott *et al*. 1997), the jets are not initially arranged in a rectangular lattice but, instead, each jet is centred on a particular landmark feature of the face, termed a *fiducial point,* such as the corner of the right eye. This step has been implemented and was achieved by centring each of 45 jets (by hand) on a particular fiducial feature, e.g. the outside corner of the right eye, for a 'learning set' of 70 faces, which differed in age, sex, expression, depth orientation, etc. Figure 8 shows some of the fiducial points on a face at different orientations and expressions. The 70 jets for each of the 45 points are stored as a 'bunch graph'. When a new face is presented to the system, not the mean *but the closest fitting* of the 70 jets for each feature is taken as a basis for refining the position by undergoing local diffusion. For example, if the right eye in the probe image is blinking, then a best match might be an eye that is blinking, rather than the mean. A jet on the centre of the chin might come from another face. Once a sufficiently large set of faces is

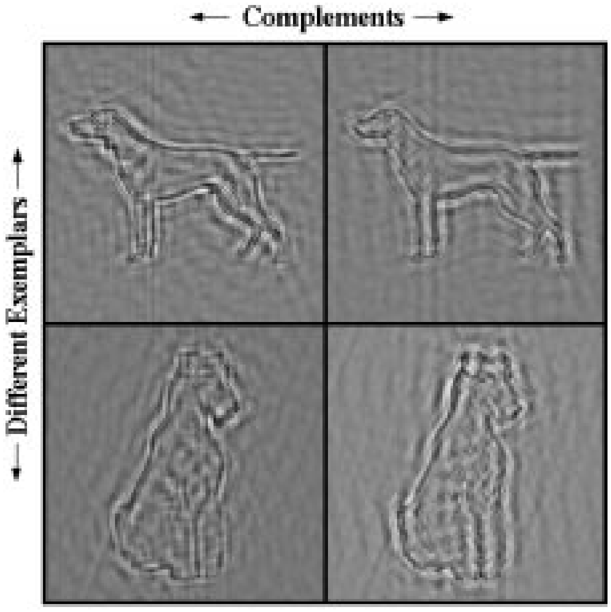

Figure 10. Example images for the object naming task of experiment 1. Shown are the four images (two exemplars × two complements) created for the entry level object 'dog'. In the priming paradigm one of the four images was displayed on the first block of trials and either the identical image, its complementary pair, or a different exemplar image was displayed on the second block of trials.

included in the bunch graphs (*ca*. 50), it is possible to automatically add new fiducial points. After the matching jet from the bunch graph finds its optimal position, the actual pattern of activation for a jet at that fiducial point is taken to be one of the jets representing that particular face.

The fiducial points, in additional to potentially allowing better resolution in matching, can readily be employed to reject inappropriate image information, such as would occur if the face was partially occluded by a hand. When none of the jets for a given fiducial point in the bunch graph can match their feature to some confidence level in a circumscribed region (constrained in part by the neighbouring jets), that jet is simply not employed in the matching phase. In this way partial occlusion can be made to exact a much smaller cost on recognition than it would if the occluder was incorporated into the representation of a face. Although not implemented, it may be possible to suppress the activity from parts of the receptive fields of jets that lie outside of the bounding contours of the face so they do not contribute to the representation as well. Young *et al*.'s (1987) finding that offsetting the upper and lower halves of a composite face resulted in much better performance in recognizing that the component individuals might be handled by a similar application of a fiducial point model. In this case the fiducial points in the upper and lower halves of the face were not in their expected locations, so their activation pattern would not be included in matching one half of the face to the other half. It is possible, of course, that beyond the offset of the fiducial points, the matched cusps provide strong evidence of separate parts, and this evidence could also aid easier retrieval of the offset face.

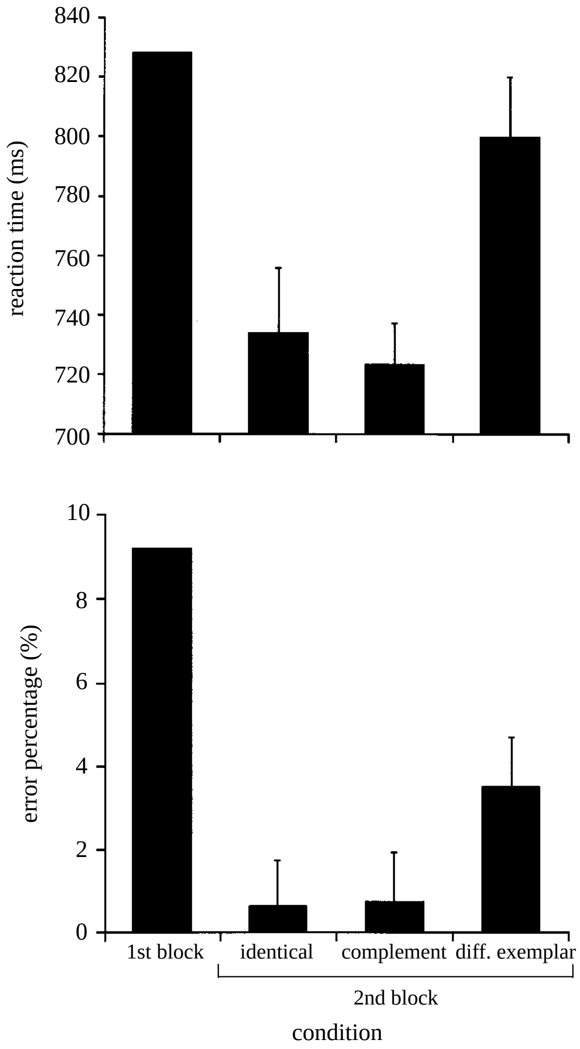

Figure 11. Mean correct naming RTs and mean error rates for the object naming task of experiment 1. The second block data are for those trials where the object was correctly named on the first block. (Because this was a within-subject design, error bars show the standard error of the distribution of individual subjects' difference scores, computed by subtracting each subject's mean score from that subject's score for a particular condition; and so they do not include between-subjects variability.)

It will be recalled that in the Cooper & Wojan (1996) experiment, better recognition was obtained for faces in which one eye was raised, rather than both of them, despite the former stimulus looking less like a face. If the expected locations of the fiducial points for the eye on the opposite side of the head differed for the left and right halves of the face, then each face half might not have been integrated into the fiducial points of the eye in the opposite half. Consequently, the original half could vote for the correct face, without incorporation of the distorted region.

In summary, in addition to greater accuracy in recognizing faces over a wider range of conditions, the

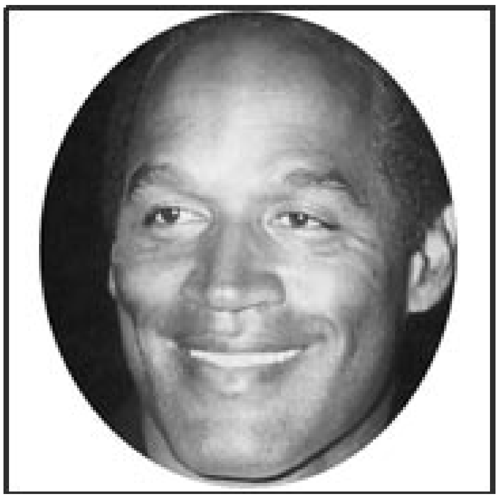

Figure 12. An example of two original grey level images of a famous person (O. J. Simpson) illustrating differences in expression and pose used in the face verification task of experiment 2. The images were collected such that the expression and/or the orientation of the two face images of a person were different.

great value in employment of a fiducial point representation is that it allows selective attention to be exercised over a holistic representation of the face.

(*c*) *The use of topological relations*

A second modification of the filter model would be the incorporation of the *distances* between the jets. This could be done either with the original lattice or with the fiducial points. Figures 7 and 8 show both arrangements with the nodes of the lattice (upper) connected to its nearest nodes and the fiducial points (lower) connected to their nearest fiducial points to form a set of triangles. A change in the image of a face produced by changes in orientation and expression, as in figures 7 and 8, results in distortion of the lattice or the triangles. A potentially important representational problem is whether the distances among the jets (or the distortions of these distances) should be incorporated into the representation or whether the jet similarities are sufficient to account for the accuracy of the model's performance in modelling human face recognition. Many issues remain concerning the possible inclusion of an explicit measure of distance (e.g. the sum of the squares of the differences in corresponding distances) as a component of similarity in the matching phase. The fiducial point model has a strong potential for serving as a research platform for addressing these and a number of the other issues in face recognition, such as norm-based coding.

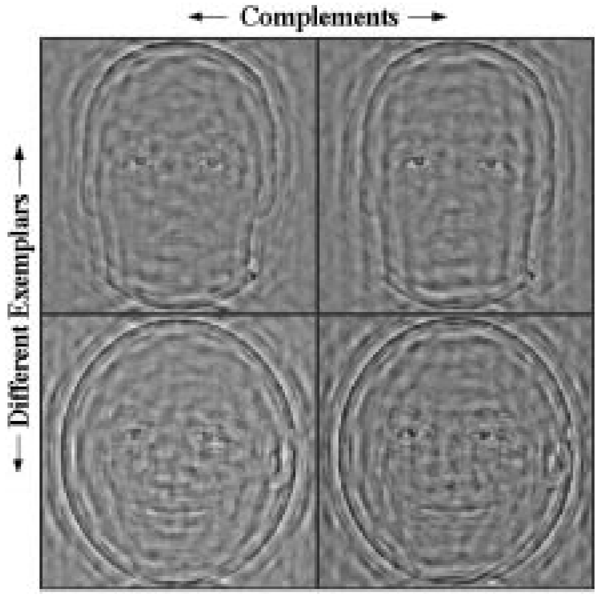

Figure 13. Filtered complementary images for the famous face verification task of experiment 2. Shown are the four images (two exemplars × two complements) created for the images of O. J. Simpson shown in figure 12. In the priming paradigm one of the four images was displayed on the first block of trials and either the identical image, its complementary pair, or a different exemplar image was displayed on the second block of trials.

(*d*) *Norm-based coding?*

In the current versions of the model, the match of a probe face to a face stored in the gallery is only a function of the similarity between the two. An alternative basis for matching could be to include not only the similarity of the two faces but their distances from the norms of a population of faces. There are several effects that would suggest some role of such norm-based coding in face recognition. Caricatures can be created by enhancing deviations (e.g. by 50%) of points on a particular face from the population values (see Rhodes & Tremewan (1994) for a recent review). Moreover, for famous faces the recognition accuracy of such caricatures does not suffer in comparison to—and can sometimes be found to exceed—the recognition accuracy of the original face (Rhodes & Tremewan 1994). Carey (1992) and Rhodes & Tremewan (1994) tested whether the caricature gains its advantage in recognition (or resists a loss) because of the increased 'distinctiveness' of the distortions in face space. They showed that 'lateral' caricatures, in which the distortions were made in a direction orthogonal to the direction of the deviation of a point, were recognized less well than 50% characters, which were recognized as well as the original,

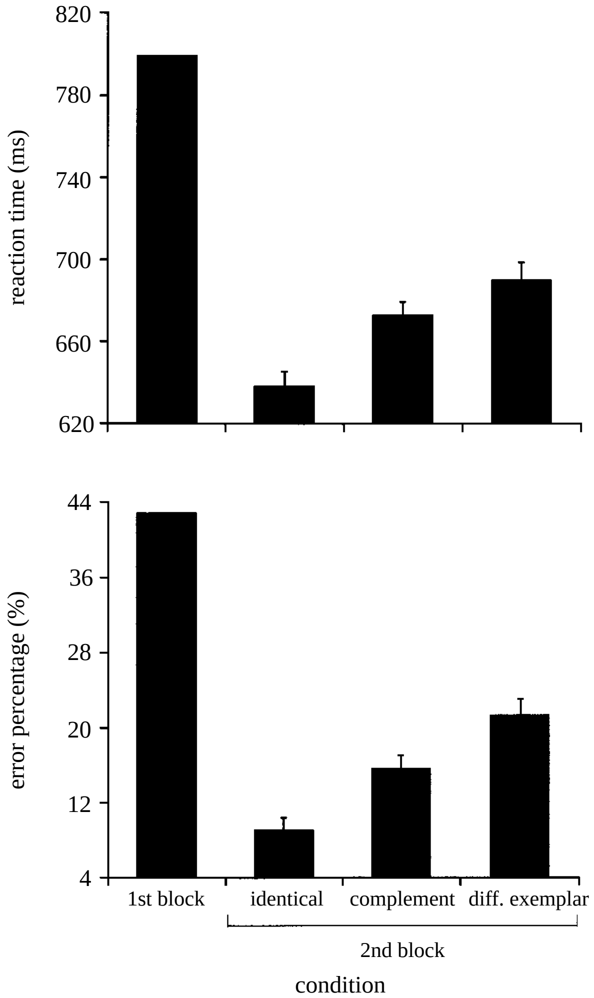

Figure 14. Mean correct naming RTs and mean error rates for the face verification task of experiment 2. The second block data are for those trials where the object was correctly named on the first block.

and even less well than *anticaricatures*, faces where the distortion was reduced by 50% towards the norm. Thus, it is not merely *any* distortion that produces an advantage, but only those that enhance the deviations from the norm.

The fiducial point model of Wiskott *et al*. (1997) would seem to be particularly well designed to incorporate norm-based coding. Whether the perception of caricatures differs from that of non-caricatured faces can be assessed with such a representation. A caricature matched against its original image will have a lower similarity value with the standard matching routines in the Wiskott *et al*. system. But it would be a simple matter to include deviations of both the jet locations and the kernel activation values from a normed face. One can also ask whether the advantage of the caricature is one of deviations from the norm or deviations from near neighbours? In general these two measures will covary. An explicit model also offers the possibility of more detailed tests of how caricatures function. When performed over a set of faces, would it be possible to predict which faces would enjoy a caricature advantage and which would not? Should greater weight in matching be given to kernels in proportion to their departure from their normed activation value? This last question raises a possible issue with respect to caricatures. People typically realize that they are looking at a caricature and not the original face. Is it possible that caricature perception alters the way in which faces are coded or matched? Specifically, do models that predict the distinctiveness of uncaricatured faces also serve to predict the distinctiveness of caricatured faces?

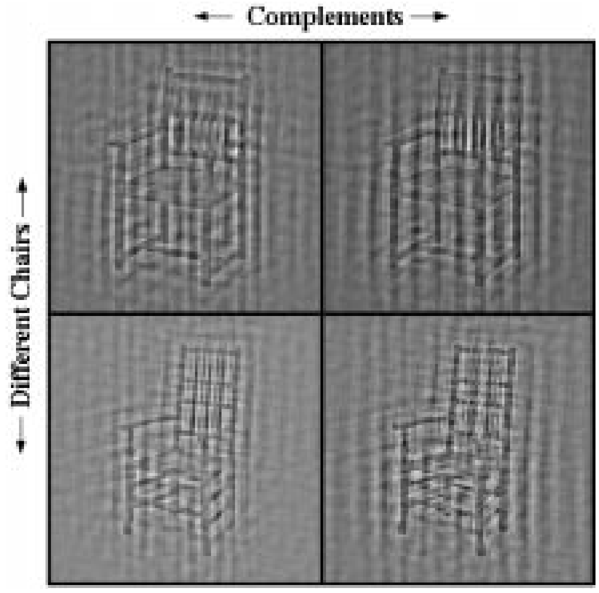

Figure 15. Same images for the chair matching task of experiment 3. Shown are the four images (two exemplars × two complements) created for two chair images from the stimuli set.

5. EMPIRICAL TESTS OF SPATIAL FILTER REPRESENTATIONS

The purpose of the foregoing set of experiments was to assess whether the identification or matching of faces and objects would be dependent on the original spatial filter values. We would expect such a dependence for faces but not for objects.

There is considerable evidence that the priming of objects cannot be dependent on a representation that retained the similarity space of the activation values of spatial filters (Fiser *et al*. 1997). For example, if contours are deleted from a line drawing of an object so that the geons cannot be recovered from the image, recognition becomes impossible (Biederman 1987). The same amount of contour deletion, when it permits recovery of the geons, allows ready recognition. Fiser *et al.* showed that the Lades *et al*. (1993) model recognized the two kinds of stimuli equally well. Similarly, the Lades *et al*. (1993) model failed to capture the differences in matching objects that did or did not differ in a NAP in the Cooper & Biederman (1993) experiment.

Biederman & Cooper (1991) showed that members of a complementary pair of object images in which every other line and vertex was deleted from each part (so that each image had 50% of the original contours) primed each other as well as they primed themselves. The measure of priming was the reduction in the naming RTs and error rates from the first to the

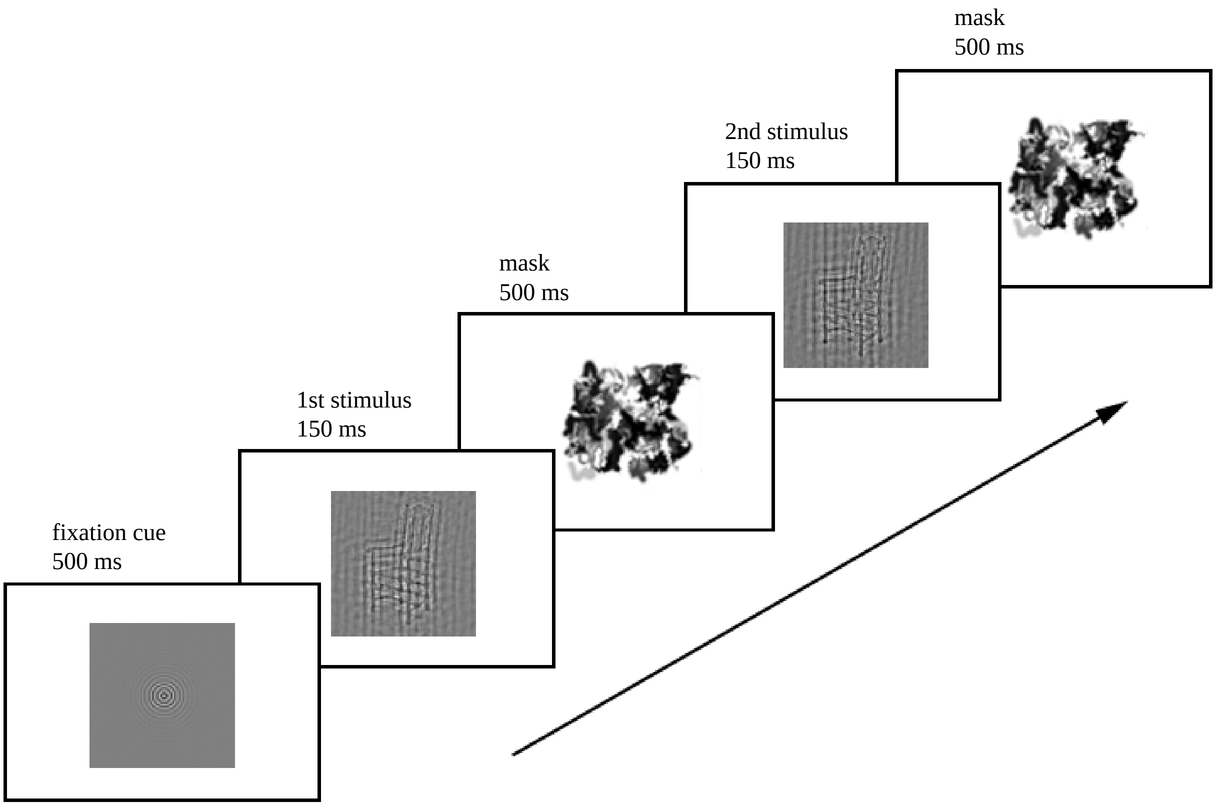

Figure 16. Sequence of images presented in the chair matching task of experiment 3. The correct response to this sequence is 'same' because both pictures are of the same chair, although different members of a complementary pair.

second brief exposure of an object picture. The priming was visual, and not just verbal or conceptual, because there was much less priming to an object that had the same name but a different shape (e.g. two different shaped chairs). In this case, humans treated the members of a complementary pair as equivalent, although the two members would have different spatial filter activation patterns (Fiser *et al*. 1997).

To test whether faces retain and objects do not retain the original spatial filter activation pattern, the first two experiments employed a similar design comparing the magnitude of priming of identical to complementary images. Rather than deletion of lines as in Biederman & Cooper's (1991) experiment, complementary pairs of grey-level images of objects and faces of celebrities were created by having every other Fourier component (eight scales × eight orientations) in one member and the remaining 32 components in the other, as illustrated in figure 9 (see also Appendix 1). In experiment 1, subjects named pictures of common objects on two blocks of trials. On the second block, for each object viewed on the first block, subjects would see either the identical filtered image that was shown on the first block, its spatial complement, or a different shaped exemplar with the same name, as illustrated in figure 10. The results of this experiment are shown in figure 11. Visual priming was evidenced on the second block of trials because the same shaped object was named more quickly and accurately than an image with the same name but a different shape. However, naming RTs and error rates for identical and complementary images were virtually equivalent, indicating that there was no contribution of the original Fourier components compared to their complements to the magnitude of visual priming.

Experiment 2 employed the same general priming design with faces except that subjects verified rather than named the images of famous people. Before each trial the subject was given the name of a famous person. If the image was that person the subjects were to respond 'same'. In half of the trials the picture did not correspond to the target and the subjects were to respond 'different'. In these cases the picture was a face of the same general age, sex, and race as the target. The verification task was used, rather than a naming task, because the naming of faces is slow and error prone. As in experiment 1, two pictures with the same name but a different shape (differences in pose, expression, orientation, etc.), as illustrated in figure 12, were used to assess that the priming would be visual and not just verbal or conceptual. As in experiment 1, for the 'same' trials on the second block, for each face viewed on the first block, subjects would see either the identical image, its complement or the different image of the same person as illustrated in figure 13. In contrast to the result for object naming, in this experiment complementary images were verified significantly more slowly and less accurately than those in the identical condition, as shown in figure 14. The difference between the complementary and the different exemplar faces was not significant, indicating that the visual system represented complementary face images almost as differently from the original as it did the different exemplar images. This result indicates that the representation of a face, unlike that of an object, is specific to the original filter values.

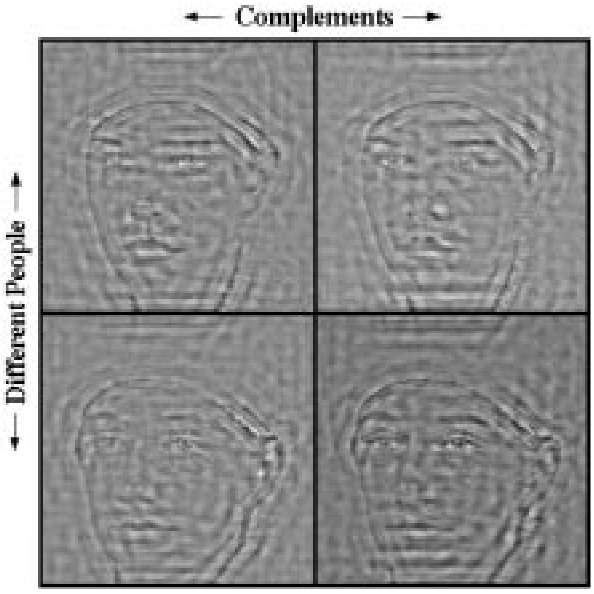

Figure 17. Example images for the unfamiliar face matching task of experiment 4. Shown are the four images (two exemplars × two complements) created for two face images from the stimuli set.

One possible explanation for the above results is what we have been positing: face representations preserve the activation pattern of early filter values, whereas object representations do not. Alternatively, it could be that it is the necessity for distinguishing among highly similar entities, such as faces, that produces a dependence on the original early filter outputs. Two additional experiments were conducted to assess whether the dependence on the precise filter values were a consequence of the greater similarity of the face stimuli (or the verification task itself) as opposed to being a phenomenon specific to the representation of faces. In these experiments, subjects viewed a sequence of two highly similar chairs (experiment 3, figures 15 and 16; or two highly similar faces (experiment 4), figure 17). Subjects performed a same–different matching task in which they judged, 'same' or 'different', whether the two chairs or persons were the same, ignoring whether the image was identical or complementary. The mean similarities of the complementary pairs of faces and chairs were approximately equivalent, as was the mean similarity of target and distractor faces and chairs as assessed by the Lades *et al.* (1993) model (see also Appendix 2). In both experiments 3 and 4, in half the same trials the second presented image was identical to the first, and in the other half it was the complementary image.

Performance on identical and complementary chair images on 'same' trials was virtually identical, as shown in figure 18, indicating that there was no effect of changing the specific spatial components of the chair images. However, for faces, the complementary images were significantly more difficult to match than identical ones (figure 19), indicating a strong contribution of the specific spatial components in the image.

In summary, this set of experiments showed equivalent priming and matching performance for identical and complementary images of objects. However, faces revealed a striking dependence on the original filter values. There was virtually no visual priming across members of a complementary pair of faces, and face complements were far more difficult to match than identical images. These results indicate that faces are represented as a more direct mapping of the outputs of early filter values. One likely reason why the objects were unaffected by varying the filter values is that object representations employ non-accidental characterization of parts or geons based on edges at depth or orientation discontinuities. Different spatial filter patterns can activate the same units coding edges, non-accidental characteristics, part structures, and relations, as discussed by Hummel & Biederman (1992).

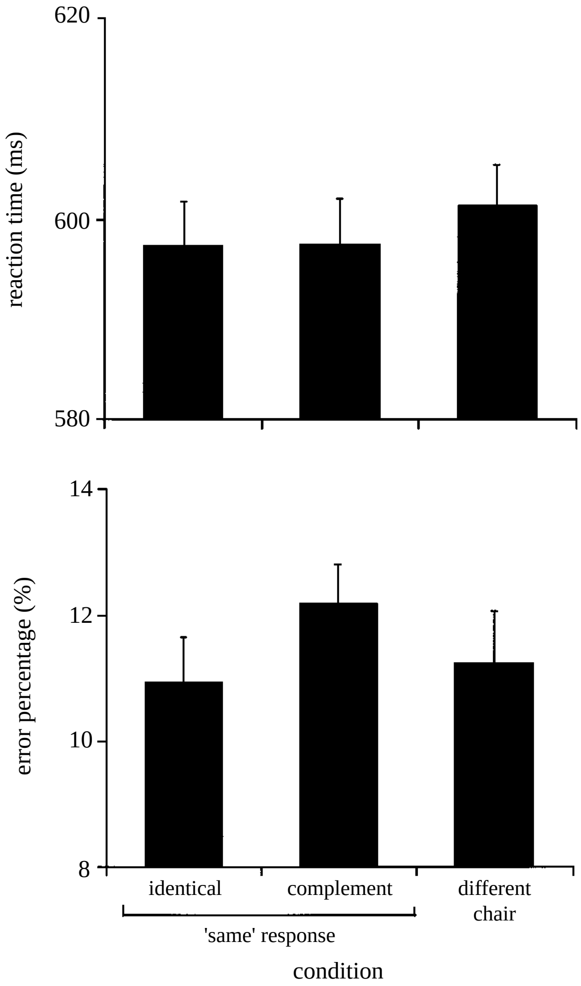

Figure 18. Mean correct RTs and mean error rates for the chair matching task of experiment 3.

6. CONCLUSION

A number of differences are apparent in the behavioural and neural phenomena associated with the recognition of faces and objects. Readily recognizable objects can typically be represented in terms of a geon structural description which specifies an arrangement of viewpoint invariant parts based on a non-accidental characterization of edges at orientation and depth

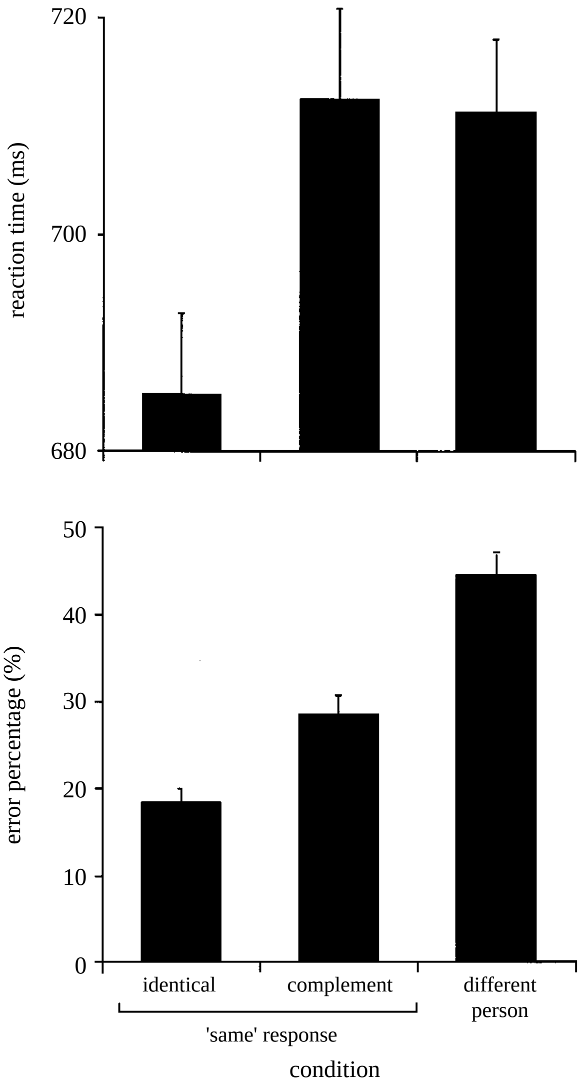

Figure 19. Mean correct RTs and mean error rates for the unfamiliar face matching task of experiment 4.

discontinuities. The parts and relations are determined in intermediate layers between the early array of spatially distributed filters and the object itself, and they confer a degree of independence between the initial wavelet components and the representation. The units in a structural description of an object allow ready verbalization. The non-accidental characterization of discontinuities endows the representation with considerable robustness over variations in viewpoint, lighting, and contrast variables. Finally, object experts discover mapping of small non-accidental features. Individuation of faces, by contrast, requires specification of the fine metric variation in a holistic representation of a facial surface. This can be achieved by storing the pattern of activation over a set of spatially distributed filters. Such a representation will provide evidence for many of the phenomena associated with faces, such as holistic effects, unverbalizability, and great susceptibility to metric variations of the face surface, as well as to image variables such as rotation in depth or the plane, contrast reversal, and direction of lighting. Face experts represent the whole face. A series of experiments demonstrated that the recognition or matching of objects is largely independent of the particular spatial filter components in the image, whereas the recognition or matching of a face is closely tied to these initial filter values.

The authors express their deep appreciation to Professor Christoph von der Malsburg for providing the general theoretical context by which biological face recognition might be modelled, his help in using the face recognition systems that he and his associates have developed, and for a number of stimulating and helpful discussions. We also thank Dr Wolfgang Konen for his assistance and support in our implementation of ZN-Face, and Pearl Fang and Nancy Wang for their assistance in creating the stimuli and running the subjects. This research was supported by ARO NVESD grant no. DAAH04-0065 and ONR N00014-95-1-1108.

APPENDIX 1.

Complementary image pairs were created by the following procedure: eight-bit greyscale images were Fourier-transformed and bandpassed filtered cutting off the highest (above 181 cycles/images) and lowest (below 12 cycles/image) spatial frequencies. The remaining part of the Fourier domain was divided into 64 areas (eight orientations × eight spatial frequencies). The orientation borders of the Fourier spectrum were set-up in succession of 22.5°. The spatial frequency range covered four octaves in step of 0.5 octaves. By this operation the two complementary images had no common information about the objects in the Fourier domain.

APPENDIX 2.

A recent study (Biederman & Subramaniam 1997) provides strong documentation that the Lades *et al.* (1993) system can provide an a priori measure of shape similarity when the pairs of shapes only differ in metric properties. In a same–different sequential matching task, subjects judged whether two highly similar, blobby, asymmetric toroidal free-form shapes were identical or not. A family of 81 such shapes has been generated by Shepard & Cermack (1973). On different trials, the shapes varied in similarity, as assessed by the Malsburg system. For intermediate to highly similar shapes, RTs and error rates in judging that two shapes were different correlated by 0.95 with the model's similarity measure.

REFERENCES

Baylis, G. C., Rolls, E. T. & Leonard, C. M. 1987 Functional subdivisions of the temporal lobe neocortex. *J. Neurosci.* **7**, 330–342.

Behrmann, M., Winocur, G. & Moscovitch, M. 1992 Dissociation between mental imagery and object recognition in a brain-damaged patient. *Nature* **359**, 636–637.

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

Biederman, I. & Bar, M. 1995 One-shot viewpoint invariance with nonsense objects. Paper presented at the Annual Meeting of the Psychonomic Society, November 1995, Los Angeles.

Biederman, I. & Cooper, E. E. 1991 Priming contour-deleted images: evidence for intermediate representations in visual object recognition. *Cognit. Psychol.* **23**, 393–419.

Biederman, I. & Gerhardstein, P. C. 1993 Recognizing depth-rotated objects: evidence and conditions for three-dimensional viewpoint invariance. *J. Exptl Psychol.: Human Perception and Performance* **19**(6), 1162–1182.

Biederman, I. & Subramaniam, S. 1997 Predicting the shape similarity of objects without distinguishing viewpoint invariant properties (VIPs) or parts. *Investigative Ophthalmol. Visual Sci.* **38**, 998.

Biederman, I., Gerhardstein, P. C., Cooper, E. E. & Nelson, C. A. 1997 High level object recognition without an anterior inferior temporal cortex. *Neuropsychologia* **35**, 271–287.

Biederman, I., Subramaniam, S., Kalocsai, P. & Bar, M. 1997 Viewpoint-invariant information in subordinate-level object classification. In *Attention and performance XVII. Cognitive regulation of performance: interaction of theory and applications* (ed. D. Gopher & A. Koriat). Cambridge, MA: MIT Press. (In the press.)

Bruce, V. 1988 *Recognizing faces*. Hove and London, UK: Erlbaum.

Bruce, V. & Humphreys, G. W. 1994 Recognizing objects and faces. *Visual Cognition* **1**, 141–180.

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

Cesa, I. L. 1994 Of attractive librarians and lop-sided faces: development and testing of a training procedure for improving the accuracy of eyewitnesses. Ph.D. thesis, Department of Psychology, University of Southern California.

Cooper, E. E. & Biederman, I. 1993 Metric versus viewpoint-invariant shape differences in visual object recognition. *Investigative Ophthalmol. Visual Sci.* **34**, 1080.

Cooper, E. E. & Wojan, T. J. 1996 Differences in the coding of spatial relations in faces and objects. *Investigative Ophthalmol. Visual Sci.* **37**, 177.

Damasio, A. R., Damasio, H. & Van Hoesen, G. E. 1992 Prosopagnosia: anatomic basis and behavioural mechanisms. *Neuropsychologia* **2**, 237–246.

Davidoff, J. B. 1988 Prosopagnosia: a disorder of rapid spatial integration. In *Perspectives on cognitive neuropsychology* (ed. G. Denes, C. Semenza & P. Bisiachi), pp. 297–309. Hillsdale, NJ: Erlbaum.

Desimone, R. & Ungerleider, L. G. 1989 Neural mechanisms of visual processing in monkeys. In *Handbook of neuropsychology, vol. 2* (ed. F. Boller & J. Grafman), ch. 14, pp. 267–299. Amsterdam: Elsevier.

Edelman, S. 1995 Representation of similarity in 3D object discrimination. *Neural Computation* **7**, 407–422.

Enns, J. T. & Shore, D. I. 1997 Separate influences of orientation and lighting in the inverted-face effect. *Perception Psychophysics* **59**, 2331.

Farah, M. J. 1990 *Visual agnosia: disorders of object recognition and what they tell us about normal vision*. Cambridge, MA: MIT Press.

Fiser, J., Biederman, I. & Cooper, E. E. 1997 To what extent can matching algorithms based on direct outputs of spatial filters account for human shape recognition? *Spatial Vision* **10**, 237–271.

Garner, W. R. 1966 To perceive is to know. *American Psychologists* **31**, 11–19.

Gauthier, I. & Tarr, M. J. 1997 Becoming a 'greeble' expert: exploring mechanisms for face recognition. *Spatial Vision.* (In the press.)

Gibson, J. J. 1947 Motion picture testing and research. Rep. No. 7, Army Air Forces Aviation Psychology Program Research Reports. pp. 113–168.

Grüsser, O.-J. & Landis, T. 1991 *Visual agnosias and other disturbances of visual perception and cognition*. Boca Raton: CRC.

Hosie, J. A., Ellis, H. D. & Haig, N. D. 1988 The effect of feature displacement on the perception of well-known faces. *Perception* **17**, 461–474.

Hummel, J. E. & Biederman, I. 1992 Dynamic binding in a neural network for shape recognition. *Psychol. Rev.* **99**, 480–517.

Johnston, A., Hill, H. & Carman, N. 1992 Recognizing faces: effects of lighting direction, inversion, and brightness reversal. *Perception* **21**, 365–375.

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

Kalocsai, P., Biederman, I. & Cooper, E. E. 1994 To what extent can the recognition of unfamiliar faces be accounted for by a representation of the direct output of simple cells. *Investigative Ophthalmol. Visual Sci.* **35**, 1626.

Kanwisher, N., Chun, M. M. & McDermott, J. 1996 fMRI in individual subjects reveals loci in extrastriate cortex differentially sensitive to faces and objects. *Investigative Ophthalmol. Visual Sci.* **37**, 193.

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

Lades, M., Vortbrüggen, J. C., Buhmann, J., Lange, J., von der Malsburg, C., Würtz, R. P. & Konen, W. 1993 Distortion invariant object recognition in the dynamic link architecture. *IEEE Trans. Computers* **42**, 300–311.

Logothetis, N. K., Pauls, J., Bülthoff, H. H., & Poggio, T. 1994 View-dependent object recognition by monkeys. *Current Biol.* **4**, 401–414.

Phillips, P. J. & Rauss, P. 1997 The face recognition technology (FERET) program. *Proc of the Office of National Drug Control Policy, CTAC Int. Technol. Symp* (In the press.)

Rhodes, G. & Tremewan, T. 1994 Understanding face recognition: caricature effects, inversion, and the homogeneity problem. *Visual Cognition* **1**, 275–311.

Rolls, E. T. 1992 Neurophysiological mechanisms underlying face processing within and beyond the temporal cortical visual areas. *Phil. Trans. R. Soc. Lond.* B **335**, 11–21.

Rumiati, R. I., Humphreys, G. W., Riddoch, M. J. & Bateman, A. 1994 Visual object agnosia without prosopagnosia or alexia: evidence for hierarchical theories of visual object recognition. *Visual Cognition* **1**, 181–225.

Sergent, J., Ohta, S. & MacDonald, B. 1992 Functional neuroanatomy of face and object processing: a PET study. *Brain* **115**, 15–29.

Sergent, J., Ohta, S., MacDonald, B. & Zuck, E. 1994 Segregated processing of facial identity and emotion in the human brain: a PET study. *Visual Cognition* **1** 349–369.

Shepard, R. N. 1964 Attention and the metric structure of the stimulus space. *J. Math. Psychol.* **1**, 54–87.

Shepard, R. N. & Cermak, G. W. 1973 Perceptual-cognitive explorations of a toroidal set of free-from stimuli. *Cognit. Psychol.* **4**, 351–377.

Subramaniam, S. & Biederman, I. 1997 Does contrast reversal affect object identification. *Investigative Ophthalmol. Visual Sci.* **38**, 998.

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

Tootell, R. B. H., Dale, A. M., Sereno, M. I. & Malach, R. 1996 New images from human visual cortex. *Trends Neural Sci.* **19**, 481–489.

Tversky, B. & Hemenway, K. 1984 Objects, parts, and categories. *J. Exp. Psychol.: General* **113**, 169–193.

Wiskott, L., Fellous, J.-M., Krüger, N. & von der Malsburg, C. 1997 Face recognition by elastic bunch graph matching. *PAMI.* (In the press.)

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

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

Young, A. W., Hellawell, D. & Hay, D. C. 1987 Configural information in face perception. *Perception* **16**, 747–759.

Young, A. W., Hay, D. C., McWeeny, K. H., Flude, B. M. & Ellis, A. W. 1985 Matching familiar and unfamiliar faces on internal and external features. *Perception* **14**, 737–746.