To What Extent Can Matching Algorithms Based on Direct Outputs of Low Level Generic Descriptors Account for Human Object Recognition?

József Fiser, Irving Biederman University of Southern California

Eric E. Cooper, Iowa State University

Abstract: A number of recent successful models of face recognition posit only two layers, an input layer consisting of a lattice of spatial filters and a single subsequent stage by which those descriptor values are mapped directly onto an object representation layer by standard matching methods such as stochastic optimization. Is this approach sufficient for modeling human object recognition? We tested whether a highly efficient version of such a two-layer model would manifest effects similar to those shown by humans when given the task of recognizing images of objects that had been employed in a series of psychophysical experiments. System accuracy was quite high overall, but was *qualitatively* different from that evidenced by humans in object recognition tasks. The discrepancy between the system's performance and human performance is likely to be revealed by all models that map filter values directly onto object units. These results suggest that human object recognition (as opposed to face recognition) may be difficult to approximate by models that do not posit hidden units for explicit representation of intermediate entities such as edges, viewpoint invariant classifiers, axes, shocks and/or object parts.

INTRODUCTION

Recently several successful face recognition systems have been developed that map the output of a retinotopic array of Gabor filters *directly* onto a similar array of "representation" units, without activating any intermediate representations (King & Xu, 1994; Lades, Vorbruggen, Buhmann, Lange, von der Malsburg, Würtz, & Konen 1993; Weiss & Edelman, 1995; Phillips, 1994). At first glance, it may seem that because face recognition (i.e., determining *whose* face it is) is a relatively difficult task compared to object recognition, a successful face recognizer would be a good candidate for an object recognizer[1].

We tested whether one such system, the Face Recognition System (FRS) proposed by Lades, et al. (1993), would yield the qualitative pattern of effects well documented in several human object recognition experiments. Although the tests were run against only this single model, the implications are most likely characteristic of the class of models that attempt to achieve object recognition directly from the output of a lattice of low level descriptors, (e.g., Edelman & Bülthoff, 1992; Poggio & Edelman, 1990; Turk & Pentland, 1991). In the discussion we consider the generality of the results, and suggest some of the additional characteristics that

---

[1]Thorough this paper “object recognition” refers to categorization of common objects with two or more parts, such as a chair, a lamp, or a teapot. Our conclusions might not apply to identification of blobby, single part objects that do not have characteristic part shapes, such as beanbags, crumpled clothes, or single part objects that require surface specification, such as a tennis ball. Indeed, Biederman, Hilton and Hummel (1991) reported substantial gain from the presence of texture contours in the recognition speed of blobby and single part objects, an effect not evidenced with multipart objects (Biederman & Ju, 1988).

might be required to extend two-stage networks so that they might better model the psychophysical data presented here. It is undecided--indeed, it may be an implementational option--whether these additions can be incorporated as modifications of the input layer, the positing of additional layers, or both.

In noting that the FRS does not activate any intermediate representations, we mean that it does not posit units that, broadly speaking, would represent the results of perceptual organization. Thus, edges representing orientation and depth discontinuities, viewpoint-invariant properties, parts or various manifestations of parts such as "shocks" (cf. Kimia, Tannenbaum and Zucker, 1995) are not made explicitly by the FRS and other two-stage models. All that is activated are units performing a Gabor filtering of the luminance array, at various scales and orientations. The model assumes a *deformable template*, in that it performs a similarity calculation between the values of the units in an (original) stored image and these same units in a probe image, after deformation to find an optimal match.

Although there is widespread belief that intermediate representations are useful and important in object recognition, these beliefs have, for the most part, developed out of computer vision implementations, and only recently has evidence for their role been obtained from biological observations (see Biederman, 1995 for a summary). It has never been established what psychophysical effects, if any, could not be accomplished through a direct matching of deformable templates based on the initial filter array and what effects (or large proportion of their variance) might require the existence of intermediate representations. Although there is some belief that faces and objects are represented differently, it has never been established just how these representations might differ. To the extent that the FRS might have some merit as a face recognizer, but not as an object recognizer, we have some computational basis for understanding the differences in the representation of faces and objects.

Why the Lades et al. Face Recognition System?

Our selection of the FRS was motivated by six considerations. The system was: 1) successful at recognizing faces, 2) successful at modeling certain aspects of human face recognition data, 3) a good representation of some of the early cortical neural coding of shape, 4) a complete and maximally efficient basis set for the representation of any image, 5) designed to use a widely accepted method for matching, and 6) available.

To elaborate:

1. The FRS works. In fact, it performs remarkably well, recognizing at over 95% accuracy, translated, depth-rotated and moderately distorted faces from a gallery of hundreds of faces (Lades et al., 1993).

2. The similarity values derived from the model show extremely high correlations with psychophysical data from face recognition experiments (Kalocsai, Biederman, & Cooper, 1994). Specifically, correlations of -.90 to -.95 have been obtained between the similarity values, derived from the FRS, of two images of a person and the costs in reaction times (RTs) and error rates in determining that these two images (presented sequentially) are, in fact, the same person, when the orientation in depth or of different and expression of the two faces have been varied.

3. A substantial body of evidence supports the idea that part of the earliest cortical representation of shape information in mammals specifies values of a multiscale-multiorientation representation embodied in simple cells in V1 (Daugman, 1984; DeValois & DeValois, 1988; Jones & Palmer, 1987a; Jones & Palmer, 1987b; Kulikowski, Marcelja, & Bishop, 1982). Formal analyses of such representations have led to numerous vision system implementations consisting of a lattice of Gabor (or Gabor-like) spatial filters at a variety of scales and

orientations (Daugman, 1988; Zhong & Mallat, 1990). The array of Gabor filters of the model's input layer thus mimics the arrangement of V1 hypercolumn simple cells.

4. The columns of Gabor filters essentially perform a wavelet decomposition of the image (Daugman, 1988) and thus offer a basis set for the representation of the image. Given enough resolution (filters at different scales and orientations), the representation is *complete* in that the image--any image--could be restored from the filter values. The representation is also efficient in that the Gabor filters can subserve an optimal sampling with minimal uncertainty in the dimensions of scale and position (Daugman, 1985). Moreover, an appropriate family of self-similar Gabor-functions (also called a Morlet-wavelet) is optimally suited to process natural scenes (Field, 1987; Field, 1994). Is such a system adequate for modeling human entry-level object recognition? We note here that a low correlation between the model and human performance would not *prove* that intermediate representations are directly employed in recognition but would raise the question as to why such a complete representation was insufficient.

5. The FRS uses stochastic optimization with a normalized distance metric to obtain a measure of similarity between the input and the stored templates. While there exist different methods to compute similarities (such as multidimensional scaling), and different measures to define similarities between two templates, the method used by the FRS (or some slight variant of it) is reasonably standard.

6. We had easy access to the system and good cooperation from the system's managers.

Overview and Rationale of the Research

We compared the performance of the FRS against psychophysical data from four experiments by Biederman and his colleagues on the speed and accuracy of naming contour-deleted object pictures (Biederman, 1987; Biederman & Cooper, 1991a,b). We selected these experiments for three reasons:

1) The experiments had exceptionally clear and replicable results whose qualitative nature could be readily assessed.

2) We had all the images and the data for the individual images.

3) The experiments were of theoretical importance in that their results had been used to argue for a role of intermediate representations in human object recognition. Specifically, the experiments were originally designed to test a principle of geon recovery (Biederman, 1987) which holds that recognition of degraded, occluded, or rotated objects will be possible if two or three geons, in their defined relations, can be recovered from the image. The present investigation, however, was not designed to test this particular theory of intermediate representations.

The images from the different conditions used in the experiments were collected in the system's gallery by frame grabbing, and the various comparisons or identifications asked of the human subjects were requested of the system. The experimental stimuli were line drawings. There is ample evidence that a line drawing whose contours reflect the orientation and depth discontinuities of the parts of an object can provide an input that is sufficient for real-time, basic-level object recognition by humans and yields the similarity between multipart objects evident with full color images (Biederman & Ju, 1988). Moreover, line drawings and gray level images are virtually identical in their manifestation of the invariances characteristic of human object recognition (Bartram, 1974; Fiser & Biederman, 1995). However, the FRS achieved its success with faces by a computation based on the extended frequency range in gray-level images of

faces. The continuous variation in gray scale characterizes smooth changes in surface depth and curvature (Horn, 1975; Koenderink & van Doorn, 1980; Marr, 1982). Could such a system be expected to recognize line drawings? Prior to our tests, the system developers judged that the answer would be no. But if the system is to be taken as a general model of image recognition, it should be able to handle line drawings as well as gray level images. To anticipate some of the results, the system *did* recognize line drawings at very high accuracy. This allowed us to assess whether the model would reflect the systematic variation in human performance produced by various image manipulations.

We present first a description of the face recognition system, then an assessment of the system's functioning with faces under conditions that resembled those used for object images. We then present four "experiments" with the system on images employed in psychophysical experiments on object recognition. A general discussion follows.

THE FACE RECOGNITION SYSTEM

Here we will provide a brief overview of the face recognition system that is sufficient to understand the present investigation. A more elaborate description can be found in Lades, et al. (1993).

The system consists of two layers, the image layer and the object layer (Figure 1). In both layers the information is represented by 2D graphs, that is, by nodes and connections (edges) between nodes. The nodes represent complex information about the underlying image: A node at a given location in the image represents a convolution at that location with 2D Gabor filters (kernels) in eight different orientations distributed at equal intervals between 0 and π, and on five different spatial frequency scales at equal intervals on a log scale between

$$\frac{\pi}{8} \le \vec{k} \le \frac{\pi}{2} ,$$

where $\vec{k}'$ is the center frequency of the Gabor kernels (and also controls the Gaussian window). The family of filters defined above are often referred to as "Morlet wavelets" (Mallat, 1989). Since the system uses even (cosine) and odd (sine) types of the Gabor-function, the family of kernels can be described as a family of complex functions in the following form:

$$\Gamma_{\vec{k}}(\vec{x}) = C_{k,\sigma} * \exp\left(-\frac{\vec{k}'^2 \vec{x}^2}{2\sigma^2}\right) * \exp\left(j\vec{k}\vec{x}\right). \qquad (1)$$

Here $C_{k,\sigma}$ is constant, $\vec{k}'$ controls not only the frequency of the sine/cosine part and the position/size of the window, but also the orientation of the Gabor-kernel. The constant parameter σ assures that the ratio of the wavelength and the window size is such that in all cases the shape of the Gabor kernels are similar, and resemble the simple cell receptive field profiles found in V1 (DeValois & DeValois, 1988; Jones & Palmer, 1987b).

Convolving the image with 40 even and 40 odd kernels (8 orientations * 5 scales) centered at the same position in the image gives two 40-dimension vectors (sine and cosine families). Computing the Euclidean norm of the corresponding elements in the two vectors we obtain 40 coefficients that describe how much luminance waves of different orientation and scales are present at the given point in the image. Such a 40 dimensional "power spectrum vector " (called "jet") at each node will be the information associated with a given node. The memory

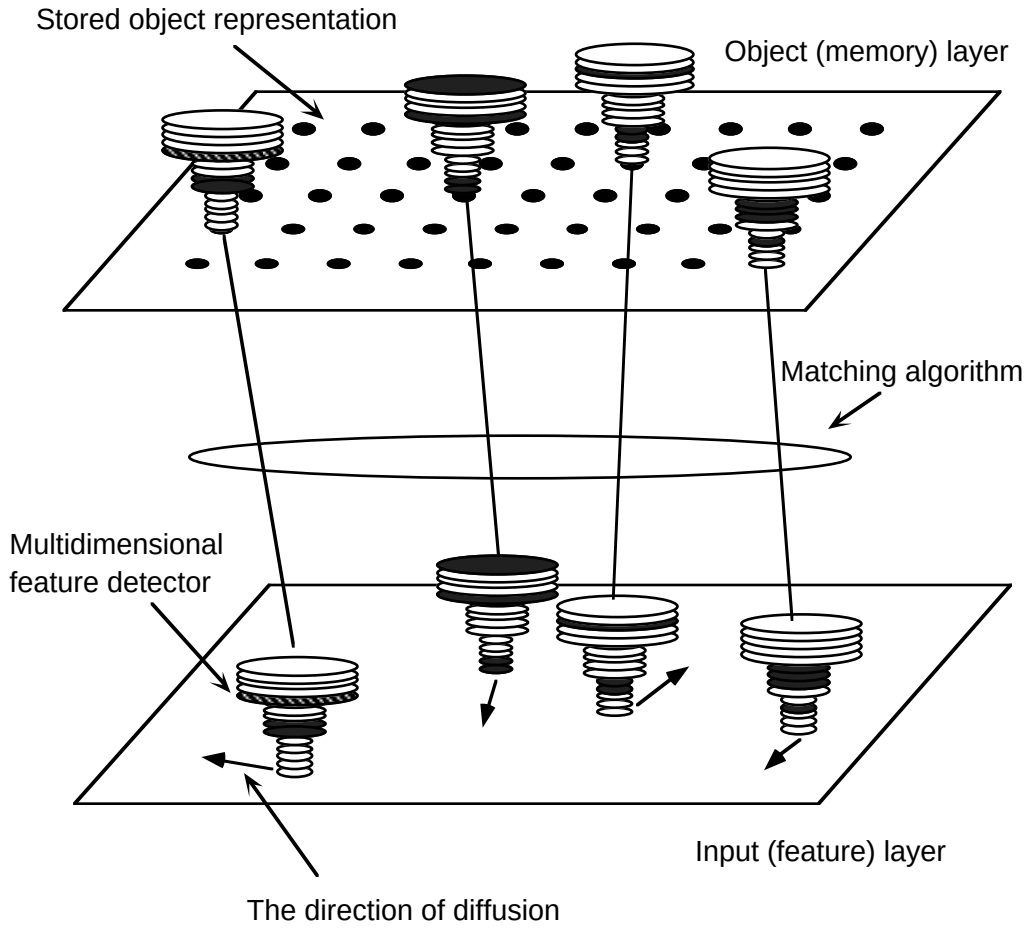

Figure 1. Schematic representation of the FRS. In the upper layer is the stored representation of an object consisting of 40 jets positioned on an 8*5 grid (only a few of the jets are shown in the picture). The lower layer is the input layer where jets positioned on each spatial location are computed from the input image. Starting from the original position of the node a given jet in the upper layer is compared to a jet in the lower layer. Based on the similarity between the two jets and the deformation of the original grid the position in the lower layer is slightly changed (diffusion), and next the jet in the new position is compared to the jet in the upper layer. The position of all the 40 jet is optimized with a Monte-Carlo method. After the best positions are found the computed similarity value is used against similarity values with other stored templates to select the best match between the input and the gallery of stored objects.

representation of a face is a 5x9 rectangular grid positioned on the face where at each grid point the "jets" are calculated and stored. The smallest kernels in the jet use sine waves of approximately 4 pixels/cycle, the largest kernels use sine waves with 16 pixels/cycle. All kernels have three prominent positive peaks under the Gaussian envelope, thus the extent of the smallest kernels is approximately 13 pixels in one direction. The extent of the largest kernels is approximately 52 pixels.

The comparison of a stored image representation with a new incoming image involves two measurements: determining the similarity between stored jets and jets computed in the new image and measuring the necessary distortion of the grid in the new image in order to find similar jets. Similarity ($S_{vertices}$) between an input jet (J_{inp}) and an output jet (J_{out}) is measured by the dot product of the two 40 dimensional jets

$$S_{vertices}\left(J_{inp}, J_{out}\right) = -\frac{J_{inp} * J_{out}}{\left\|J_{inp}\right\| * \left\|J_{out}\right\|} \qquad (2)$$

The distortion of an edge segment between two nodes (S_{edges}) is measured by the quadratic difference in lengths between corresponding edges in the input and the stored image

$$S_{edges}\left(D_{inp}^{i,j}, D_{out}^{i,j}\right) = \left(D_{inp}^{i,j} - D_{out}^{i,j}\right)^2 \qquad (3)$$

Here $D^{i,j}$ gives the distance between nodes i and j.

The graph matching occurs in the following manner. First the 40 dimensional node representation centered on each pixel of the input image is obtained. Next, the same grid that was used for the stored pattern is moved around *rigidly* on the input image searching for the best initial position for the grid. Rigidity means that the distance between two nodes of the graph does not change. The search is performed by random walk gradient descent method with arbitrary but less than a maximum step size. At each step a cost function is evaluated which is a combined measure of jet similarities and the grid distortion.

The cost function is:

$$C_{total} = \lambda * \sum_{i,j} S_{edges}\left(D_{inp}^{i,j}, D_{out}^{i,j}\right) + \sum_{i} S_{vertices}\left(J^{i}{}_{inp}, J^{i}{}_{out}\right) \qquad (4)$$

where λ is a constant determining the relative importance of the two type of costs, and i,j takes each possible integer values between 1 and the number of horizontal and vertical vertices, respectively.

When the grid is rigid, the first term in the cost function is zero. If the cost in the new position is lower than in the old one, the grid is repositioned. Otherwise it remains in the old position. After the optimal initial position is explored in this way, the second phase of optimization begins where the individual nodes can "diffuse" independently constrained by topographical neighborhood. The nodes take a randomly selected new position if by this step the cost function is reduced by more than a predefined threshold. In this phase the grid gets distorted, therefore the first term in the cost function also contributes to the total cost. The process stops when no improvement happens during a given number of trials. This optimization procedure corresponds to simulated annealing at zero temperature (Kirkpatrick, Gelatt, & Vecchi, 1983). The energy landscape of a local jet is smooth enough to allow the gradient descent method to find its minimum.

The matching process described above is performed sequentially for each stored face in the gallery ranking them according to their cost. The best match (lowest cost) will be designated as

the system's decision as to which face in the gallery corresponds to the input image. The range of possible cost function values runs from 0 to -45, where -45 is the score for a perfect match. Values generally range between -35 (an extremely poor match) to -45 (the best possible match). Figure 2 shows two typical results of successfully matched (rank = 1) faces.

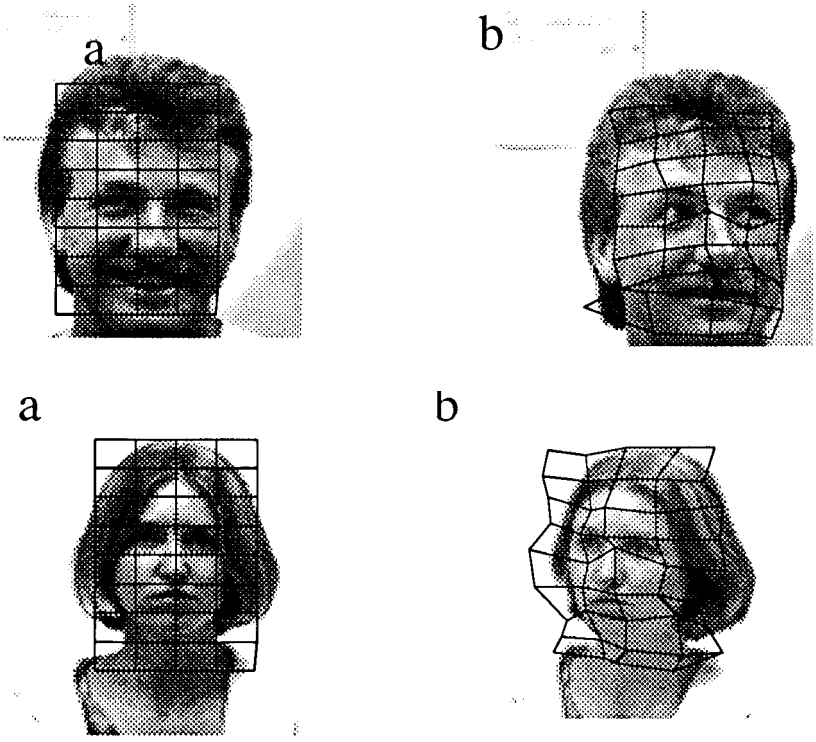

Figure 2. Two examples of successfully matched faces (rank = 1). The intersections of a grid represent the positions of the jets in the image. The grids on the matched faces are distorted because the individual jets stored from the original images diffused to different position in the test images to find their best match.

Experimental method and results with face stimuli

In order to have a basis to relate the results of object recognition to face recognition, we performed a set of comparisons of faces to a gallery of images for 56 people, equal to the maximum number of objects in the object experiments. This was a challenging problem in that typically there were sizable differences in the two images of a person, as described below and illustrated in Figures 3 and 4. Yet the FRS performed very well. The mean ranking of the cost function value for one image of a person to another image of that same person among the 55 distractors for all 56 faces was 1.45**.** Chance would have been a rank of 27.5. The correct face was recognized -- ranked first -- in all but 11 of the 56 comparisons. When a correct response was made, it tended to be made "confidently," in that there was a large difference in the value of the cost function between the first (correct) rank and the next highest rank. In the 11 faces that were not ranked first, it was never the case that the false alarm received high confidence. On these error trials, the correct face always had a cost that was only slightly below that of the face that was ranked first and the ranking for the correct face was almost always among the top five.

Figure 3. Six examples of pairs of faces successfully identified (lowest cost function assigned to the correct face) by the Face Recognition System. The lower picture of each pair taken of the individuals was ranked as the most similar to the upper image among all the other 57 faces in the gallery.

Figure 4. Six examples of pairs of faces where the Face Recognition System did not assign the lowest cost value to the correct individual in the gallery of 57 faces. Typically, the correct face when it did not have the first rank, was ranked among the best two or three candidates with the lowest cost values.

The 19.6% error rate (i.e., instances where the correct face was not ranked first) in the present test was considerably higher that the 5% error rate reported for the system in a previous investigation (Buhmann, Lange, & von der Malsburg, 1989). As noted in the preceding

paragraph, we posed a difficult test in that the members of a pair typically differed in expression, viewpoint, lighting, and gross features. To illustrate the difficulty of the tests, Figure 3 presents 6 pairs of successfully identified faces, and Figure 4 shows 6 pairs of the 11 failures, which represent those typical circumstances when the system failed. Figure 3 shows that the system is robust with respect to size, orientation and lighting changes, as well as to changes in the background structure up to a certain degree. Some of the differences in image pairs in Figure 4 are such that humans can overcome the problem easily (as with the more extreme lighting condition in the case of the lower middle person). In some other cases however, such as with the images of the upper middle person, even humans might have some difficulty in establishing correspondence between the two images of the same person.

What accounted for variations in grid distortion?

Figure 5 shows two comparison samples of human faces with typical grid distortions. In the upper comparison a second view of a person is matched against his/her stored view; in the lower

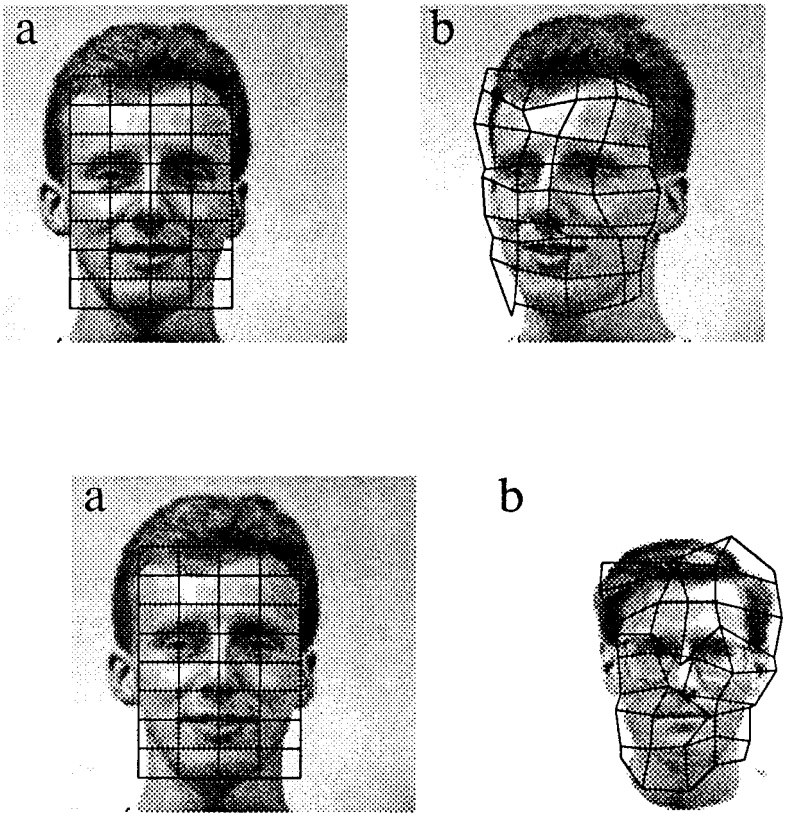

Figure 5. Examples of grid distortion with the face of the same person as the second image (upper row) *versus* a different person's face (lower row). The correct match was therefore ranked first.

row, a different person is compared to the same stored view. In the first case the grid distortion followed the rotation of the face, with the neighboring elements of the same initial orientation remaining mostly parallel. In the second case, which illustrates a high level of

distortion for matching faces, although the grid lost its regular appearance, the topographic structure (i.e., adjacency relations and local parallelism of edges) is still fairly well preserved. (A case where much of the topographic structure is lost is illustrated in Figure 10, lower panel middle image.) After optimal positioning of the rigid grid but before nodes started to diffuse independently the costs were -41.78 and -39.92 for the upper and lower comparisons, respectively. After finishing the matching process completely (i.e., after the individual diffusion of nodes was completed) the cost values changed to -42.91 and -42.15 for the same person (upper) and for the different one (lower), respectively. Recall, that a cost value of -45 indicated the best possible match between two images, and the closer the cost to this value, the better the match between an image and a stored view. For easier interpretation, in the rest of this paper an error measure based on the percent of the difference from the best possible match (viz. -45) will be used instead of the actual cost values. Maximum matching (with the cost value of -45) will be represented by 0% error. In our experiments there were no matches with cost values higher than - 25, thus 100% error was set to -25. According to this encoding a match with a cost value of -35 will be represented as 50%.

The demonstration in the preceding paragraph that the matching costs of the image of a *different* person can reveal such large improvements when individual diffusion is allowed illustrates a fundamental trade-off implemented in the FRS. If topological constraints were absent, the individual nodes could easily find positions in the image with jets having information content that would be similar to their own. However, such a system would produce many false matches. The topological constraints provide a relational structure to the stored information so that the search space is dramatically reduced. Although the face recognition model is not a structural description, in that the spatial relations are *implicit* in the positions of the jets rather than being *explicit* (such as "ABOVE" in the Hummel & Biederman (1992) network), the topological constraints revealed by the grid distortion prevent the representation from being just a feature list, insensitive to the spatial relations of luminance variations[2].

GENERAL METHODOLOGY IN TESTING THE SYSTEM AGAINST LINE DRAWINGS OF OBJECTS

In all the experiments, the same Cricket Draw line drawings of common objects were used that were used in the experiments with humans. The line drawings were printed out by a laser printer and recorded with a camera. As with the experiment with faces, a video signal was obtained by an Ikegami CCD camera (Model: ICD-200), and this signal was digitized to 640*512 pixels with 8 bits of resolution by a transputer-based frame grabber. A 512*512 pixel section of this data was low-pass filtered and converted to a 128*128 pixel image. There were slight variations in size, position, and lighting conditions within and across every set of images.

A control experiment (detailed in Experiment II) was run to assess the effect of these variations. No noticeable differences were found due to these effects. Another test was run in which hand positioning of the rigid grid was compared to the system's automatic positioning. Again, there were no differences in the results.

To avoid the possibility of easy classification by global orientation, all of the drawings with elongated shapes were positioned at the same orientation to make their appearance similar, as

---

[2]Insofar as the receptive fields of the individual filters are extended, the medium and low frequency kernel coefficients also carry implicit information about spatial relations in the image.

shown in Figure 6. The grid size (but not the number of nodes) was increased because our drawings typically occupied a larger area of the scene than the faces.

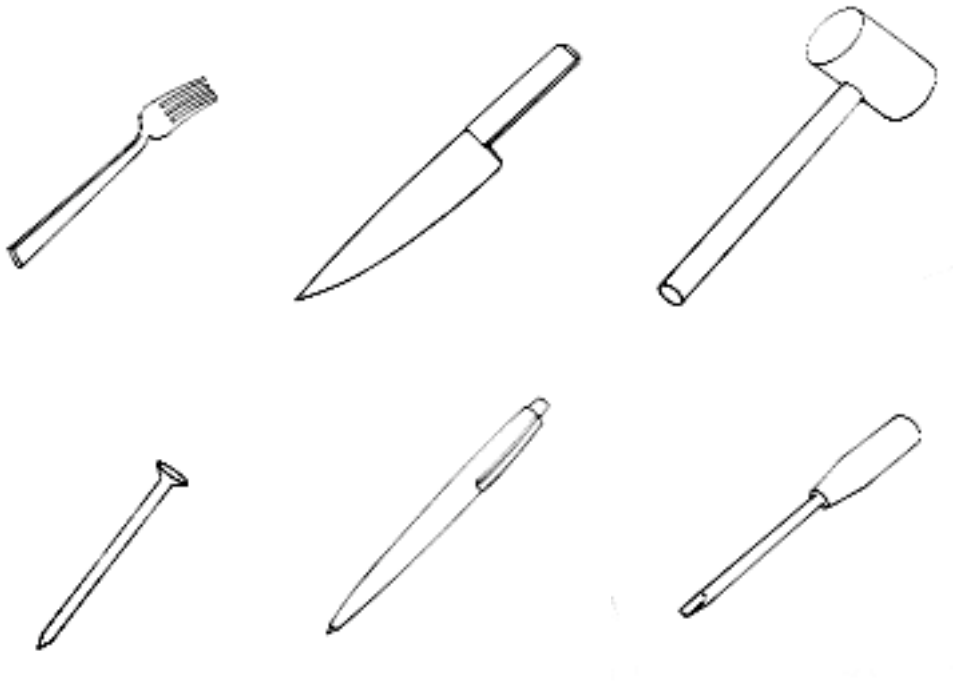

Figure 6. The elongated objects in the experiment. All were given the same orientation as shown here.

EXPERIMENT I: COMPARING RECOVERABLE VERSUS NONRECOVERABLE CONTOUR-DELETED IMAGES AGAINST THEIR INTACT ORIGINALS

In this experiment (described in Biederman 1987) contour was deleted from 48 images in such a way that the geons were either *recoverable* or *nonrecoverable*.[3] An equal amount of contour was deleted from each type of image, as illustrated in Figure 7. The results of this experiment were clear: Subjects had extraordinary difficulty in naming the nonrecoverable stimuli, even when given an opportunity to view the intact pictures prior to the experiment: The

---

[3]"Recoverable" contour-deleted images are those where the contour is removed, typically at midsegment, such that the nonaccidental (i.e., viewpoint invariant) distinguishing features of the volumes, and the information needed for their segmentation can still be determined so that the geons canbe activated (or recovered) from the image. With nonrecoverable contour-deleted images, the deletion alters the nonaccidental characteristics distinguishing the geons and their segmentation from each other are altered and hence the geons cannot be activated from the image. This is accomplished by altering, deleting, or suggesting inappropriate vertices, deleting concavities in such a manner that the remaining contour from different geons are joined by good continuation, and deleting opposite sides of a geon along with the associated vertices so that the boundaries of the geon cannot be determined. (See Biederman, 1987, or Blickle 1989, for a more extensive discussion.)

median error rate was 100% even after five seconds of viewing. Given sufficient exposure duration, the naming of the recoverable images was perfect.

METHOD

The 48 intact images of objects from the original experiment comprised the gallery. The face recognition system was tested by computing the quality of the match between the filtered representation of each of the contour-deleted images (both recoverable and nonrecoverable), and the filtered representations of the original intact images. It will be recalled that the "quality" of a match is a measure that combines the similarity of the activation of the jets (viz., vectors containing information at different orientations and scales at a given node) and the degree of distortion of the mesh of jets required for a jet to find a similar jet. Given a set of images, the quality of a match could be expressed as the rank order of the computed match against the true image. The question posed here was whether the mean rank order of matching the recoverable images against the intact images would be greater than that for the nonrecoverable and intact images.

Figure 7. Examples of test images with recoverable and nonrecoverable deletions. The left column shows examples of intact images of common objects. The recoverable (middle column) and nonrecoverable (right column) versions contain the same amount of contour. (Modified from Biederman, 1987).

RESULTS

Figure 8 shows that the mean rankings for the recoverable and nonrecoverable images are almost identical (slightly under 8) and both were well above chance, which would have been 24. Compared to humans, the system does much too well on the nonrecoverable images. One might conjecture that if humans had perfect memory (as does the FRS) they could recognize all the objects. There are three points here. First, the system, with its perfect memory, was *not* perfect

in recognizing the objects, so there was plenty of room for the system to reveal an advantage for the recoverable versions--which it did not do. Second, in one condition in the original experiment, noted previously, subjects *were* shown the original intact objects prior to the experiment. Only a modest improvement in performance resulted from that manipulation, with only a slight reduction in the enormous advantage in the recognition of the recoverable over the nonrecoverable images. Third, even if the original, intact versions were studied extensively so that subjects could accurately identify the nonrecoverable images, given that there was a sizable set of objects, they would undoubtedly still be markedly slower in making such identifications.

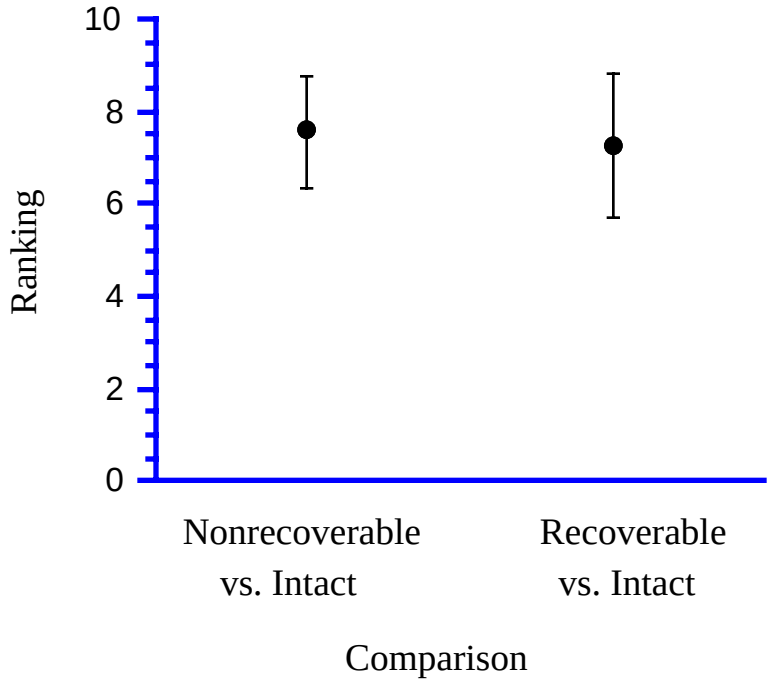

Figure 8. Average ranking of nonrecoverable (left) and recoverable (right) contour-deleted object images when matched against a gallery of 48 intact images by the Face Recognition System (FRS). Chance ranking would have been 24. (Bars indicate the standard error of the mean.) The average ranks are identical indicating that the FRS recognized the nonrecoverable images as well as the recoverable images.

EXPERIMENT II: RECOGNIZING MEMBERS OF COMPLEMENTARY IMAGE PAIRS

The speed and accuracy of naming a briefly presented picture of an object is facilitated by its prior presentation (e.g., Bartram, 1974; Biederman & Cooper, 1991a; Biederman & Cooper, 1991b). A sizable portion of this facilitation is visual and not just verbal or conceptual in that much less priming is observed when a different shaped exemplar from the same category is presented on the primed trial, as, for example, when a stuffed chair is shown on the second presentation when the first presentation was of a kitchen chair.

To assess whether the priming was mediated by the specific features in the image (viz., the lines and vertices), Biederman and Cooper (1991b) prepared complementary pairs of images in which half the contour of each of 48 object pictures was deleted by removing every other edge

and vertex from each *geon* (simple viewpoint invariant volumetric primitives presumed to form the representation for human basic level object recognition, Biederman, 1987), as illustrated in Figure 9.[4] If the two members of a complementary pair were superimposed they would produce

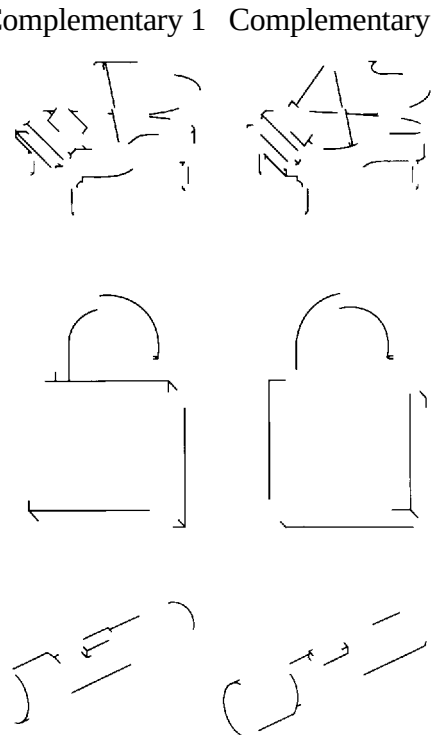

Figure 9. Three examples of pairs of complementary images, in which only half of the vertices and edges of each part are presented in each member. The two complements have no pixels in common so if they were superimposed they would produce an intact image with no overlap in contour. (Modified from Biederman and Cooper, 1991b.)

the original, intact image with no overlap in contour. Thus, the same geons would be activated from each member of a complementary pair, albeit by completely different *local* image features such that no pixel was common to the two images.

The 48 pictures were composed of 24 pairs with the same name but a different shape. On a first block, the subjects named 24 images, one with each of the 24 names, each presented for 500 msec followed by a mask. On a second (primed) block of trials (approximately 7 min later), subjects saw (for 200 msec each) pictures that were either: (a) identical to those viewed on the first block, (b) the complements which had the missing contours, or (c) same name-different shaped exemplars of the object class (e.g., a grand piano when an upright piano had been shown on the first block). The speed and accuracy of naming identical and complementary images on the second block was equivalent, indicating that none of the priming could be attributed to the features actually present in the image. Performance with both types of image enjoyed a sizable advantage over that with the different exemplars, establishing that the priming was visual, rather than verbal or conceptual. Biederman and Cooper (1991b) interpreted these results as indicating that none of the visual priming was mediated by the vertices and edges in the image. Other control experiments established that all the priming could be attributed to the activation of the

[4]Very long edges were split in two and each half assigned to a different member of a complementary pair.

parts (and the relations among the parts), none to top-down activation of the basic-level class (concept or name priming) or to the global shape to the image.

The equivalence of complementary images is not just an effect that is only manifested in the statistical analyses of the experiment: *The images look identical.* Scrutiny is required to distinguish them.

The equivalence of complementary images in priming provides a clear challenge to direct matching models: Would the system be more likely to match an image to its complement than to images of other objects?

METHOD

One of the complementary images (Complement A) from each pair was arbitrarily designated as the memory image, representing the priming effect of the image shown in the first block in the human experiment. For each A image, the 5*9 grid Morlet-representation was included in a gallery (Gallery A), and all the comparisons were made to this gallery.

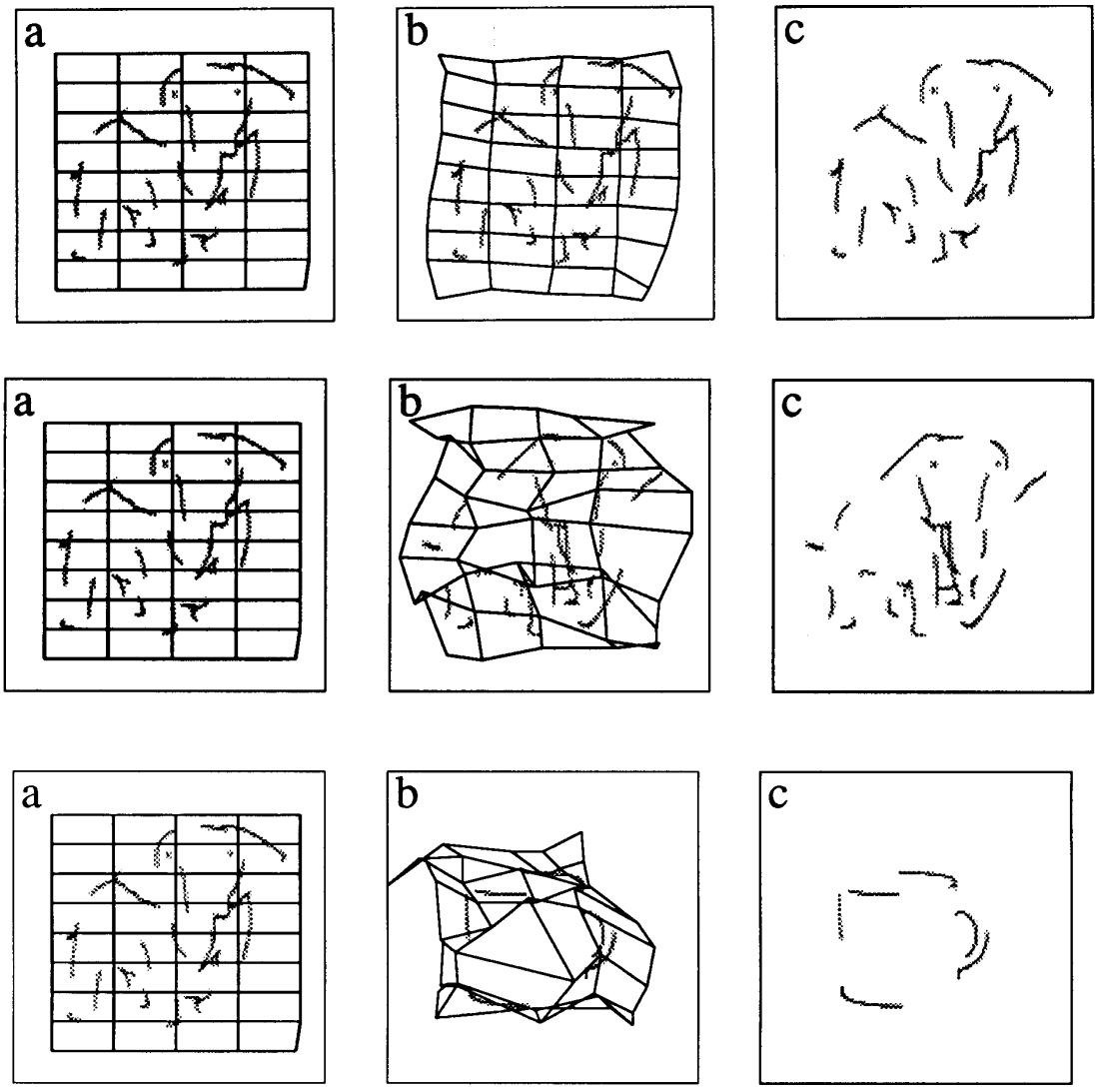

Figure 10. Illustration of the grid distortions in the complementary image matching experiment. The leftmost (a) picture of the elephant was the original one stored in the gallery so it shows no grid distortion.

Top row. The right figure (c) shows a second frame grab of the picture in A. The middle (b) image shows the slight degree of grid distortion (cost = 9%) due to differences in positioning, lighting, etc. Middle row. Matching a complement (c) against the other member of a complementary pair produces modest grid distortion (cost = 25%). Bottom row. Matching a complement of a different object against the original produces high grid distortion (cost = 40%).

As a control test to see how well the FRS performed in the face of the minor variations in lighting, distance, size, and 2D orientation, that would occur from image capture to image capture, we used an enlarged set of 56 complementary image pairs. We compared a set of new samples (frame grab) of one member of each complementary pair against the same set of images from an earlier frame grab.

The middle panel (b) in the top row of figure 10 shows the (slight) degree of grid distortion when a given A image, the elephant, was compared against itself in two different frame grabs shown in panels a (which displays the original grid position) and c.[5] The results for this comparison are shown as the "Original" point in Figure 11: The performance of the system for this comparison was perfect: All the 56 images were correctly matched against themselves. Moreover the matching was with high confidence in that there was a much larger difference in cost values between the first and second best matches (mean difference μ_{1-2}=2.417) than between the second and third best matches (mean difference μ_{2-3}=0.578). The difference between the two mean differences ($\mu_{1-2} - \mu_{2-3}$) was highly significant $t(110)=12.72$, $p<0.001$).

The middle row of figure 10 illustrates the grid distortion for the complementary images of the elephant. It is clear that there is considerably more grid distortion than in the top row. The costs for the comparisons illustrated in the top and middle rows were 9% and 25%, respectively. In both cases the system selected these correct images from the corresponding complementary image gallery. The bottom row of figure 10 illustrates the grid distortion for an image of a cup when it was matched against the elephant that differed substantially in shape from the stored (a) view of the elephant. The cost for this particular match was 60%, close to what would be obtained against random noise (for detailed analysis see the Appendix).

The point labeled "Complements" in figure 11 shows the results for matching the 56 members of complementary pairs (Complement B) against the gallery consisting of the other (A) members of each pair. Compared to the “Original” comparison (A matched to Gallery A), there was a significant drop in recognition accuracy with a mean rank of 4.5 (only 26 correct selections out of 56). We concluded that there was thus a significant difference between the system's representations of the two members (A and B) of a complementary pair.

There was also a drop in the confidence of the correct matches. That is, the mean difference in cost between the correct first and the second choices became much smaller (though still significant), (μ_{1-2}=0.853, μ_{2-3}=0.258, t (50) =3.49, p<0.01). With the incorrect selections this difference vanished (μ_1=0.281, μ_2=0.23, t(58) =0.873, ns.).

---

[5]In all the following line drawings with grids, when there are three panels, the left image (a) shows the stored image and the position of the grid on it. The nodes of the grid denote the position of jets whose information is stored in the memory. The middle image (b) shows the compared image with the distorted grid after the matching process is completed. The right image shows the same line drawing as in (b), but without the grid so that it can be seen more clearly.

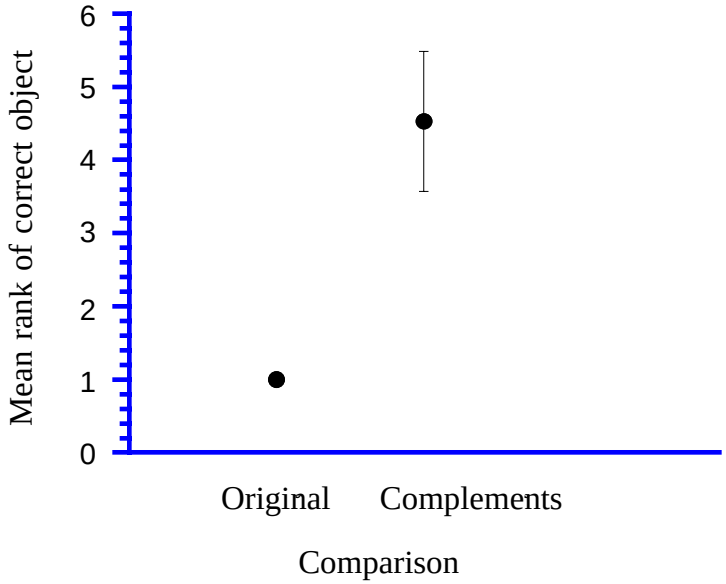

Figure 11. Left. Mean rank of the correct image by the Face Recognition System (FRS) in matching one contour-deleted complementary image (taken from a Gallery A consisting of one member, arbitrarily chosen, of each of 56 complementary pairs of images) against a second frame grab of the same image (left). Right. Mean rank of the correct image by the FRS in matching each of the contour-deleted complementary images in Gallery A against Gallery B, composed of the complementary pair members that were not arbitrarily assigned to Gallery A. Chance would have been 28. (Bars indicate the standard error of the mean.)

Although the system did not treat Original and Complements equivalently, complements were recognized well above chance, with a mean ranking of 4.5. Chance would have been 28. The implications of this result will be considered in the discussion.

It should be noted that our statistics for measuring system performance, the average rank order of the correct image, imposes a weaker criterion for obtaining invariance than Biederman and Cooper's (1991a) psychophysical performance criterion of equivalent reaction times and error rates for complementary and identical images. Our criterion for the system merely requires that no other image be more similar to a member of a complementary pair than the other member of that pair. Complete invariance could have been obtained between members of a complementary pair even though the complementary image had lower similarity values than the identical image, as long as no other image was more similar to the identical image. This occurred in only 26 cases out of the 56 possible pairs. If similarity values were directly related to RTs and error rates, then the stronger system criterion would have required that the similarity values themselves be equal. Given that the system has perfect memory of the original image, there would have been virtually no chance for the system to demonstrate invariance by the strong criterion. The weaker criterion allows invariance to be demonstrated even though the similarity values for identical and complementary pairs are not equal.

Why was more distortion evidenced for the cup than the complement of the elephant? It is not simply that the cup had different pixels than the elephant because there was not a single pixel in common between Complements A and B for the elephant, whereas there was some overlap between the contours of the elephant and the cup. Yet the cost for matching the Complement elephant against the Original was markedly lower than the cost of the cup against that same image of the elephant (25% vs. 40%). Although there was no pixel overlap in a pair of

complementary images, the manner in which contours were deleted in making the complementary images insured that much local information would be shared with respect to oriented filter outputs at any scale. Specifically, the members of a complementary pair would each have one of the sides of each geon, as with the two curves comprising the right tusk or right hind leg of the elephant. Thus members of complementary pairs tended to share edges of approximately the same position, orientation, and length. Because there are filters within each jet with receptive fields sufficiently large to be activated by the opposite sides of a geon, these edges in complementary images could readily produce substantially similar "jet-descriptions." This was the reason why the system performed so much better than chance. But such a representation was inadequate in that in over half the comparisons some other objects was computed to be more similar to the complement than the other member of a complementary pair.

EXPERIMENT III. COMPARING CONTOUR-DELETED IMAGES (COMPLEMENTS) AGAINST INTACT ORIGINALS.

In this experiment, we compared all the stimuli from the previous experiment (i.e., the complementary sets A and B) to the original intact versions. One goal of this comparison was to provide a replication of the result of Experiment I that the FRS is able to recognize line drawings, in general, and contour deleted images, in particular. More important, we employed these data to test whether the FRS similarity values can predict the degree to which contour deletion of specific images results in an increase in naming RTs and error rates over that for their intact versions.

Method

A gallery of the 56 intact images, Gallery INT, was created and compared to both the A and B complementary images.

Results

Recognition performance overall was quite high, considering that each image had only half the contour of the original. The mean rank for both A and B complementary sets was approximately 2 for each set. The A-Gallery INT comparison had 40 correct (first place) selections; the B-Gallery INT comparison 44 out of 56 trials. This result shows that the FRS is quite adept at recognizing contour deleted line drawings

An explanation similar to the one proposed for how the system was able to achieve above chance-recognition in matching members of a complementary pair can also account for the success of the FRS in recognizing a contour-deleted complementary image against its intact parent: The partial image, of course, contained much of the identical contour that was in the original. What was missing would not cause a large collapse in the grid because, as noted previously, the deleted contour was similar in position, length, and orientation, to the retained contour. Figure 12 shows that when the similarity between nodes of the original image (intact elephant) and the two compared images (Complementary A and B) is at about the same level, then similar cost functions imply similar degrees of distortion in their grids. For Complementary image A, the cost for comparison with its intact parent was 20%; for Complementary image B it was 16%.

Does the system's similarity correlate with behavioral similarity across intact and contour deleted pictures?

Deleting contour from an image of an object, as was done in creating the complementary images, markedly increases the naming reaction times and error rates for that object by human subjects (Biederman, 1987). Can the increase in naming reaction times be predicted for the

individual complementary images by the cost functions from the comparison of the system's match of a complement against its intact version?

To investigate this question, differences in the naming RTs and error rates between the contour deleted (complementary) images from Biederman & Cooper (1991b) and the intact images from Biederman & Cooper (1993) were calculated. It is necessary to substract the RTs of the responses with intact objects, because of name length, frequency, quality of the depiction,

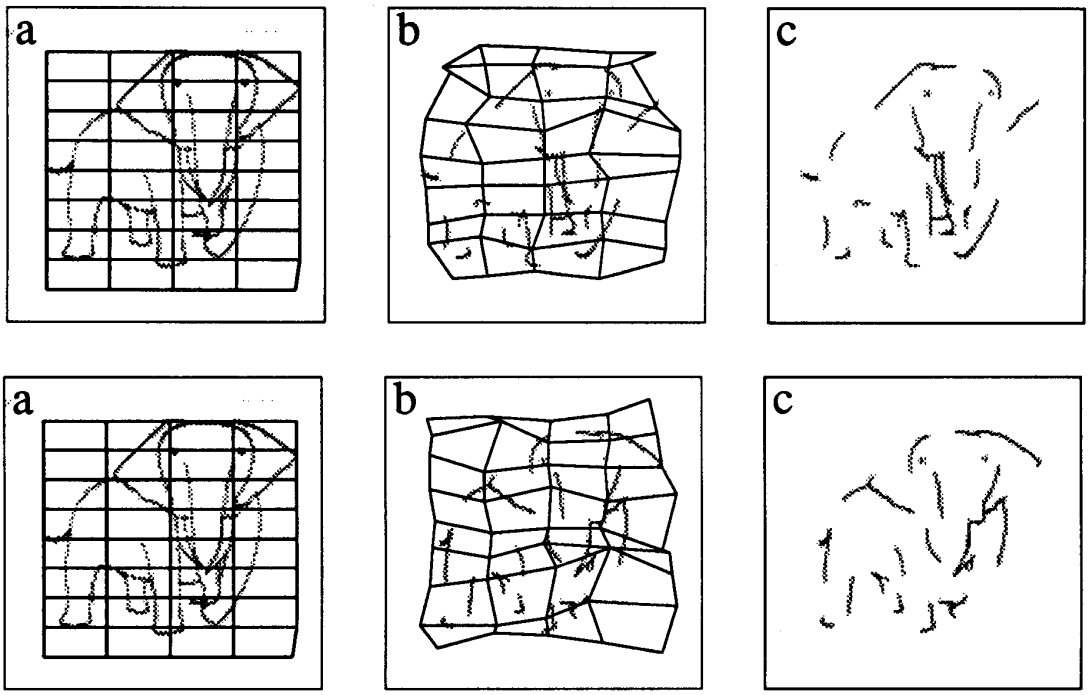

Figure 12. Illustration of the grid distortions in matching the complementary images against their intact version. The leftmost (a) picture of the elephant was the original one stored in the gallery, so it shows no grid distortion. The right figure (c) in the top row shows one image of the complementer pair of the picture in (a), the bottom row shows the other complementer image. The degree of distortion of the grid is similar in the two cases as shown in the two figures (b).

etc. Only the data from the first trials where a subject saw a particular briefly presented image were included. These raw scores were converted to z-scores (normalized) producing four sets of z-scores (RTs and error rates for each contour-deleted and intact image). For each stimulus type the z-scores for the RTs and error rates were added[6], yielding an overall normalized performance score for each image with larger values indicating a picture that was slower and/or produced more errors than a picture with a low value. For each object, the normalized performance score for the intact image was subtracted from each of the complements to yield a similarity score, with high values indicating low similarity.

---

[6]In some cases there was a reduction in "range of talent" such that, for example, there would be low variance in error rates across intact objects but there would be high variance in error rates for contour-deleted objects. The opposite pattern could occur with reaction times. Insofar as difference among objects in the low variance conditions might be unreliable, it could spuriously reduce correlations. The combined measure offered some correction of this problem. As pointed out by an anonymous reviewer, our measure assumes equal weights of RTs and error rates.

The effect of contour deletion of individual images on the FRS's performance was calculated by the following method: For each target image matched against a gallery, the difference in cost function values between the best match (apart from the correct match) in the gallery for the given image and the correct match was calculated. The logic here was that images incorrectly matched by a large margin would give large positive values, that should correlate with larger RTs and higher error rates. Images incorrectly matched by a smaller margin should be less difficult, whereas correctly matched images which give negative difference values should be the easiest to identify. The difference value was normalized (i.e. divided) by the variance of all cost values obtained by comparing the same target image to each image in the gallery. This normalized difference, called *difficulty*, was assigned to the target image as a measure corresponding to RTs and error rates in human performance. That is, to the extent that the model captured human picture naming difficulty, images which had higher difficulty values would be expected to have longer RTs and higher error rates. Difficulties were computed for all complementary and intact images. For each image, similarity of the complementary image to its intact representation in the gallery was calculated by taking the difference between difficulty values of the corresponding intact and complementary images of the object, and using the z-scores of these values.

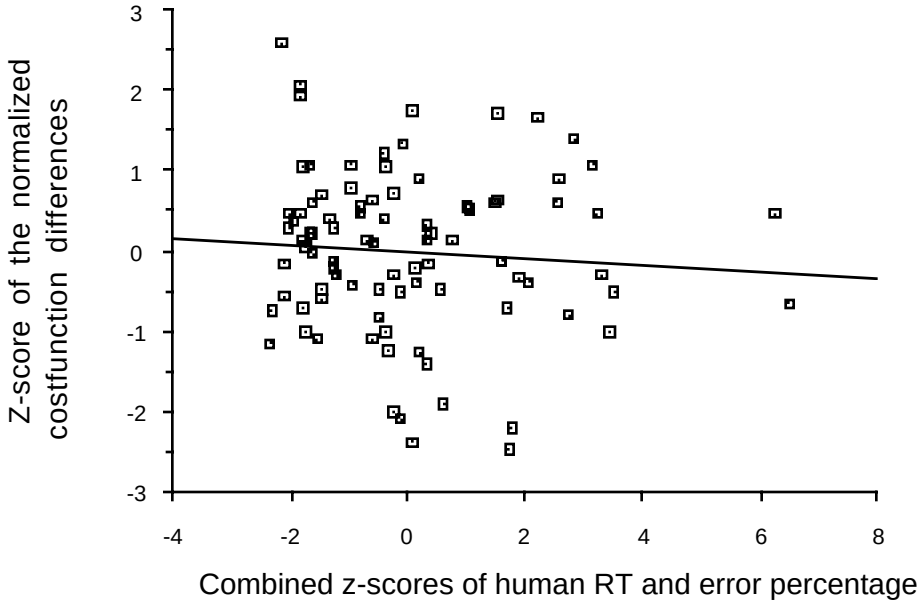

Figure 13. Scatterplot and linear regression line relating the increased difficulty of recognizing individual objects attributable to their contour deletion for humans and the Face Recognition System (FRS). For humans, increased difficulty was measured as the increase in RTs plus error rates (z-score values) when naming contour-deleted image images minus their intact originals. For the FRS, the difficulty was the similarity of the contour deleted image to the intact original. The contour deleted images, N = 56, were taken from the complementary-image experiment. Each point in the scatterplot represents the increased difficulty for humans and the FRS for a single object.

A strong positive correlation between human and the FRS similarity values across the images would suggest that the FRS cost function values reflect the psychological similarity between a contour deleted image and its intact original. However, the results shown in the scatter plot (Figure 13) indicate that there was no relation (Product moment correlation = 0.002) between system similarity and the similarity measures based on human performance. We note here that this test places the FRS at a disadvantage in that sources of unreliability in human data would serve to lower the correlation.

EXPERIMENT IV: REFLECTED IMAGES

Biederman and Cooper (1991a) investigated translation and reflection invariance with a priming paradigm. In that experiment, in each of two blocks of trials, each subject named 48 intact pictures, each shown for 150 msec, followed by a mask. The images could be inscribed in a circle with a radius of 2 degrees and they were centered 2.4 degrees either to the left or to the right of fixation. On the second (primed) block, the images could be presented either at the same position or the opposite position and at the same orientation or mirror reversed from their original orientation. Neither translation nor reflection reduced the magnitude of priming, as assessed by the speed and accuracy of naming. The test posed for the model was the effect of reflection: Would it produce the same equivalence of reflected images as revealed in the priming data?

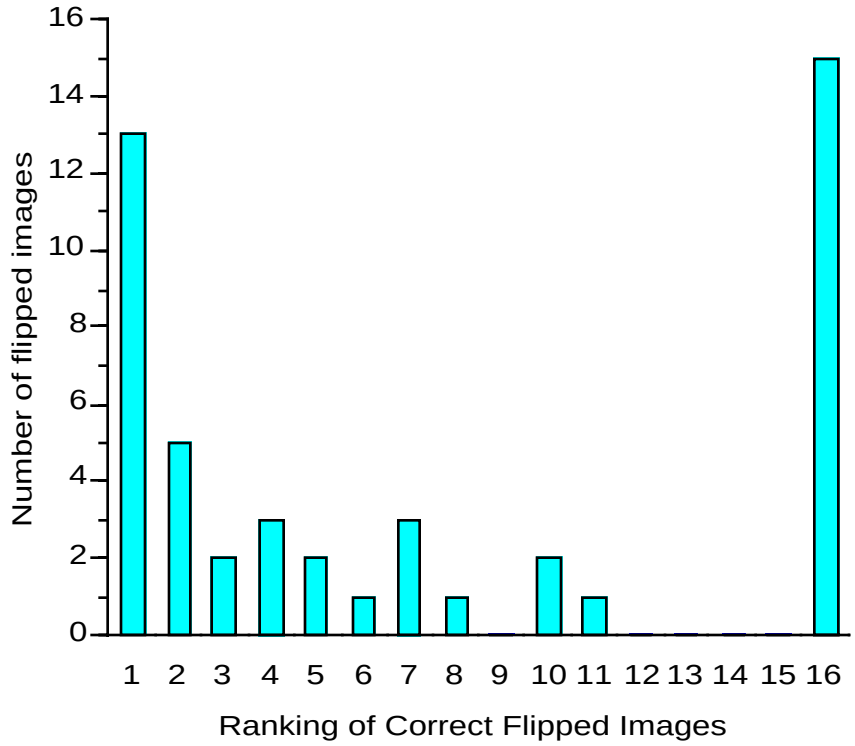

Figure 14. Distribution of ranks in matching 48 left-right mirror reversed (flipped) images against their original orientations. The flipped version of only 13 objects was correctly identified, and ranked first. The 15 cases ranked below the 16th position were grouped together.

Figure 14 shows the results. All comparisons which assigned the correct image below the 16th position in the rank were grouped at the 16th position accounting for the peak at that point. The majority of the images were recognized correctly (i.e. those with rank 1) were mirror symmetric images. The overall ranking (Mean = 7.16) was markedly lower than the results obtained by comparing Complementary A images to galleries which contain the object in their correct position (Mean = 1.45). Unlike humans, the FRS does not manifest reflection invariance.

It would be possible, of course, to endow the FRS with a capacity to recognize reflected images by simply testing a reflected version of the image and imposing zero cost on the reflection, or by storing each object in the original and in reflected positions. However, allowing easy matches for reflected versions of images would reduce the topological constraints of spatial relations in the cost function, and produce more false matches. Storing reflected templates would double the representation of any object. Reflection is only one of the possible

transformations of the image of an object. Strong transformations e.g. those for rotation in depth or partial occlusion would further increase the required representations for each object, which in turn would increase the possibilities for false matches.

DISCUSSION

Comparing the System's Performance to Human Recognition

There were marked differences in the performance of the Face Recognition System and the pattern of results from the human psychophysical experiments. Perhaps the most striking and important discrepancy was the failure of the FRS to differentiate between recoverable and nonrecoverable images with respect to their identifiability. Whereas median identification accuracy by humans for recoverable and nonrecoverable images is 100 percent and 0 percent, respectively (Biederman, 1987), the FRS recognized both at equivalent--and well above chance--levels of accuracy.

It was not that the FRS could not recognize line drawings. Given another frame grab of a line drawing of an object, with slight changes of position, lighting and orientation, the system's performance was perfect, as shown in Experiment II. If anything, the overall performance of the system was better than that of human subjects in that it recognized the nonrecoverable images at high accuracy levels, as it did with the complementary images. The extraordinarily large effect of nonrecoverable contour deletion is thus not captured by the system's direct and efficient mapping of filter activation values to object representations.

Whereas the nonrecoverability effect was a difference to which the system was insensitive, the equivalence in the priming of complementary images (Biederman & Cooper, 1991a) was an invariance (an equivalence) that the system did not manifest. Although complementary images were recognized against each other with reasonably high accuracy, they were not recognized as well as an image against itself, which would have been the comparable human psychophysical result. As noted in the discussion to that experiment, the criterion for the FRS to demonstrate invariance -- equal mean rank order -- was a weak one in that it required only that no other image be more similar to the identical image than the complement. As noted previously, complementary images look identical to humans. The FRS failed to achieve this subjective equivalence.

Humans readily recognize the equivalence of mirror reversed images of the same object (e.g., Biederman & Cooper, 1991a,b) but the worst case of the system's performance was obtained with the test for reflection invariance.

These three results demonstrate qualitative differences between humans and the FRS--not just modest quantitative differences. To summarize the results: The system completely failed to reveal the enormous effect of recoverable versus nonrecoverable contour deletion shown by humans, and the system failed to manifest the invariance shown by humans in their recognition of complementary or reflected images.

Could the FRS be Modified to Handle the Present Results?

Insofar as we believe that the first stage of the FRS provides a good model of initial V1 Gabor-type filtering, we are of the opinion that the general answer to this question is yes. But Gabor-type filtering is only one of the many operations performed by V1 and other relatively

early stages, e.g., V2 and V4, that maintain some spatial representation. We consider two types of additions to the FRS model:

a) An extention of the first layer so that it performs additional computations that are known to be characteristic of V1 but are currently omitted in the FRS. Such localized nonlinear computations would include size- and phase-independent complex cell-like tuning, end-stopping, pooling normalizations, etc. (Shapley, 1994, Heeger, 1994; Heitger, Rosenthaler, von der Heydt, Peterhans and Kubler, 1992) The question is whether such capabilities would be able to produce the difference in recognition accuracy between recoverable and nonrecoverable images and, at the same time, produce the equivalence of recognition for the complementary images. Although we cannot prove that such developments would endow the system with the needed capabilities, we see no clear path by which such capabilities could actually be achieved with such local operations.

b) An additional layer (or layers) to make explicit intermediate representations that might mediate human object recognition. Different theories have posited different entities for such representations, such as edges representing orientation and depth discontinuities, viewpoint invariant features, parts, and nonaccidental relations, as was done by Hummel and Biederman (1992), shocks (Kimia, Tannenbaum and Zucker, 1995), superquadrics (Pentland, 1986), cores (Burbeck & Pizer, 1995). The key aspect of such representations is that the same representation can be activated by different images that are well distinguished by V1-type operations. So, on the one hand, the system would have the strong capability of distinguishing between images that, on the other hand, were treated equivalently.

We are of the opinion that an adequate model of human shape recognition will require both kinds of additions: The intermediate representations will require for their activation the additional image computations but the modeling of the effects of nonrecoverability and the equivalence of complementary images will require the positing of intermediate representations, likely corresponding to the parts of the object.

In discussing these extentions we have avoided consideration of special case kludges that might handle some aspects of the data but may not be generally useful nor motivated by biological plausibility. For example, it would have been simple to achieve recognition of reflected images by merely reversing the order of the columns in the lattice of jets. Apart from obvious drawbacks of such kludges discussed previously, such a “bag of tricks” might handle some results, but could lower recognition rates as it doubles the opportunity of false matches.[7]

Prima facie, the apparent translation and scale invariance of face recognition would appear to pose a challenge to our contention that aspects of face recognition may be well represented by a direct mapping of spatial filter values to a person-pose unit in a representation layer.[8] There are a number of nonexclusive mechanisms which would allow these invariances (for objects as well), while preserving the *relative* orientation and scale activation values. We will mention just a few. One such mechanism would be that attentional selection of a position in the visual field only allows activation of units coding that region. The output of a pattern of these units then

---

[7] A reviewer noted that doubling the number of false matches might, in fact, mean that only a few additional errors would result, insofar as the system performs so well. However human object recognition performance is generally so accurate that any error is of concern. Moreover, the increased competition from close matches might increase the time for recognition, even if there were only a few of these close matches.

[8] We thank Steven Zucker, one of the reviewers, for forcing us to address this problem.

would then converge on the same face cell later in the visual pathway. Indeed, face cells (as well as nonface complex feature cells) in IT often reveal translation and scale invariance (e.g., Kobatake & Tanaka, 1994). There also can be dynamic retuning across scales, i.e., zooming in or zooming out, to produce maximal sensitivity to a given pattern, such as that of a face. The implication of the preceding discussion is that whereas the filter activation values can be generated as a pattern of excitation over a particular set of simple cells (in V1, V2, or V4), there is no need to reinstate the identical pattern of activation over the identical cells to recognize a face. Indeed, that never happens. Instead, different cells may produce the same or highly similar patterns that converge on the same downstream cells, as noted previously. With respect to scale changes, it is likely the case that retuning may either convey the *relative* spectrum of activation or a normalization mechanism might allow the same cell to be responsive to information at a variety of absolute scales. Dynamic returning appears to be a common phenomenon in neural systems.

The version of the FRS that we employed did achieve a mild degree of translation invariance in that there was a global match prior to the diffusing of the jets. Newer versions of the FRS have far more robustness in handling such transformations in that each jet recognizes a particular facial feature and diffuses to it.

What are the implications of these results for models of human object recognition?

A representation suitable for modeling entry-level object recognition can probably be encoded in either two-layer or multi-layer structures. In two-layer networks additional stages are not available to transform the input, which makes encoding of new representations inefficient. Multi-layer models can achieve a greater degree of independence between the initial representation and features at less cost (Barron, 1993). Further support for multi-layer implementations comes from the hierarchy of areas existing in the ventral visual system i.e., V1, V2, V4, IT (TEO + TE), which is believed to subserve visual object recognition in humans (Desimone & Ungerleider, 1989; Felleman & Van Essen, 1991) .

There are two classes of proposals for the hidden units of the intermediate layers. One would map the filter values directly onto a modest number of object representations, as described, for example, by Edelman (1995). A new object would then be represented as a weighted linear combination of the existing object models.

The alternative proposal derives from the supposition and evidence that intermediate layers make explicit various properties, such as orientation and depth discontinuities, viewpoint invariant properties, and parts-based organization (e.g., Biederman, 1987; Dickinson, Pentland, and Rosenfeld, 1993)[9].

---

[9]Note that this is not an exhaustive classification of existing object recognition models in the computer vision literature. We consider only the two extreme approaches of deriving intermediate representations for hierarchical systems. Approaches with full "view-based" representations and approaches with specific higher level, 'feature-based' descriptors depict the low and high ends of the spectrum in abstraction of intermediate descriptions. Different computer vision algorithms (such as ones using skeletons, codons, superquadratics, etc.) represent intermediate variations between 'full-view-based' and 'feature-based' methods. Dickinson et al. (1993) provide an extensive discussion of variations of intermediate representations.

Although the present investigation was not designed to test between these alternatives, the consistency of the results of the psychophysical experiments with geon theory suggests that the intermediate units posited by that account (Hummel & Biederman, 1992) might at least define the goals of the intermediate units. At its most general level, geon theory assumes that objects are represented as an arrangement of parts defined by edges at orientation and depth discontinuities. The edges are specified with respect to a restricted set of nonaccidental features, such as curved or straight segments, vertex type, parallel and collinear contours and axes. The spatial relations between geons, such as above, perpendicular to, larger than, are explicitly coded by units representing these relations. Reflection invariance is simply obtained by not specifying left-right relations (Hummel & Biederman, 1992).

Cooper and Biederman (1993) directly investigated whether the FRS would manifest the greater sensitivity to differences in viewpoint invariant properties compared to differences in metric properties (viz., aspect ratio). From a standard simple object composed of only one or two parts, such as a lamp with a vertical cylinder as a base and a truncated cone (shade), they created one version in which one of the parts differed in aspect ratio, e.g., the cylinder was more elongated than in the standard, and another version differing in a geon of that part, e.g., a brick of the same aspect ratio was substituted for the original cylindrical base. The difference in aspect ratio was selected so as to be slightly less similar, according to the FRS, compared to the geon differences.

In one experiment subjects had to judge whether a simultaneously presented pair of these objects--always with the same name--were physically identical or not. This task required neither memory nor classification, and performance was consistent with the FRS similarity scaling: "different" responses to the aspect ratio differences were faster than the responses to the geon differences.

However, in another experiment where subjects judged whether the same pair of images, presented sequentially had the same name (and basic-level classification) ignoring whether a part was different in aspect ratio or geon, the results were opposite to that based on the FRS similarity. Now on the same trials geon differences produced longer RTs and higher error rates than aspect ratio differences[10].

The general implication of these results is that processes involved in object recognition and memory, presumably occuring later in the cortical stream, are more sensitive to viewpoint invariant properties of images than the processes by which physical identityof simultaneous stimuli are judged. V1 simple cell similarity scaling seems to predict only the latter. Neural support for this conjecture derives from Kobatake and Tanaka's (1994) recent report of cells in IT that show maximal responses to complex patterns and little or no responding to their simple orientation components. The complex patterns that have been identified, generally, differ from each other in viewpoint invariant specification and show little sensitivity to viewpoint dependent manipulations, such as scale or position changes, and what would be transformations produced by rotation in depth (though this has not been directly tested). Cells earlier in the ventral pathway, e.g., V1, V2, and V4, do not show these preferences to complex patterns, and the responses to a complex pattern at these stages can be well predicted from its spatial components .

---

[10]When making "same-different" judgements of physical identity larger differences result in faster different responses, but when making "same-different" judgements of object classes larger "to-be-ignored" differences produce slower "same" responses. E. E. Cooper (unpublished) has shown that both the need to retain image information over an interval--even intervals as brief as a second--as well as the requirements of object classification increase the saliency of nonaccidental properties relative to metric properties.

An additional comparison to emphasize the necessity of viewpoint invariant descriptors of orientation and depth discontinuities would be to compare gray-scale images of objects to their line drawings which express the contours at these discontinuities. Humans would perform this task with ease, whereas the FRS would be expected to perform poorly.

Implications for the relations between face recognition and object recognition.

What do our results imply about the general issue of interest which was whether two-layer standard matching approaches, so successful in face recognition, could be a good candidate as a model of human object recognition?

Identification of faces (at similar orientations) and object recognition pose different challenges to a recognition system. Face identification requires making relatively fine distinctions among highly similar images, whereas entry-level object recognition requires distinguishing among highly dissimilar entities. Faces are highly similar in that they all have the same parts in approximately the same positions. Representing individual faces, after scaling and translation, by storing the output of a two dimensional lattice of simple filters is effective for two reasons. First, it facilitates easy matching. The image variations produced by a face at a given position in a two dimensional representation can be compared to the stored information in the same position of the other image: the left corner of the left eye at a particular orientation always tends to be at approximately the same location in the face. The activity from the left corner of the left eye of a particular probe face can therefore be compared with left corners of left eyes for all faces in memory. The total variance of the activity at this region will be relatively small with well-defined landmarks, e.g., the left eye. Second, the slight metric variations important for identification can be readily expressed by the outputs of simple filters as in the FRS or even with the values of the pixels themselves, as in Turk and Pentland's (1991) face recognition system. The high correlation between human psychophysical data in judging facial identity and the performance of the FRS found in Kalocsai, et al. (1994) suggests that the human visual system might exploit these advantages for face recognition. However, this strategy comes at a price: human face recognition performance is drastically disrupted by changes in the filter values produced by changes in orientation, direction of contrast, lighting, and rotation in depth (Bruce, Valentine, & Baddeley, 1987; Johnston, Hill, & Carman, 1992; Kalocsai et al., 1994).

For entry-level object recognition there would appear to be little advantage in using a two dimensional coordinate space, or coding precise metric features. Neither comparing the upper right regions of the two images, nor relating slight metric differences between the images helps particularly when a telephone has to be distinguished from a chair. Instead, there is some evidence that intermediate, viewpoint invariant representations of the parts of the objects are activated as part of a structural description (Biederman & Cooper, 1991b). Structural descriptions might be derived from classifiers that distinguish viewpoint invariant differences from relatively small image variations, such as that produced by a straight vs. curved edge, at depth discontinuities. As long as two or three parts in their specific invariant relations can be extracted from the image, entry-level classification will almost always be successful despite drastic variations in viewpoint, lighting, orientation, and the actual features presented in the image (Biederman & Cooper, 1991a; Biederman & Gerhardstein, 1993; Ellis & Allport, 1986; Srinivas, 1995) Such a representation, to a large extent, can neglect metric changes of the object, as was shown by Cooper and Biederman (1993) .

Kalocsai, Biederman, Fiser, and Fang (1996) reported a direct test of the conjecture that the representation of faces may be accomplished by a direct mapping of relative spatial filter values

whereas the intermediate representation of objects allow invariance of these filter values (although the filter values are required to activate the intermediate representations themselves).[11] A space composed of spatial frequency (8 levels) x orientation (8 levels) was defined. Kalocsai et al. (1996) prepared *Fourier complements* of chairs and faces by deleting every other combination of frequency and orientation from each image. That is, if the space is regarded as an 8 x 8 checkerboard, then one member of a complementary pair was composed of the red squares and the other member the black squares. In the critical experiment, subjects performed a sequential same-different matching task in which they judged whether a pair of images, each followed by a mask, was the same or different chair or the same or different person, ignoring whether they were identical or complementary images. The distractors were highly similar chairs or faces (same sex and age with hair and adornments eliminated). The similarity of the distractors and the similarity of the members of the complementary pairs were the same for both faces and objects according the FRS. Whereas RTs and error rates in matching a pair of mianges of a chair were unaffected by whether the pair was identical or complementary, there was a marked increase in RTs and error rates when matching complementary images of the same faces compared to identical images.

In conclusion, activation of a lattice of simple spatial filters may be sufficient for the representation mediating face recognition and judgments of physical identity of objects, but additional representations are required to model how humans recognize multipart objects.

Acknowledgment

We express our deepest thanks to Jean-Marc Fellous, for his invaluable support in managing the FRS and facilitating our work with the system. We thank Cristoph von der Malsburg for the opportunity to access the FRS and for many constructuve discussions. We also thank Trish Seumanutafa for her help in reading in the images. This work was supported by AFORS grant 90-0274, McDonnell-Pew Foundation Program in Cognitive Neuroscience grant T89-01245-029, and Night Vision and Electronics Sensors Directorate (NVESD) ARO grant DAAH04-94-G-0065. Parts of this work were presented at the Annual Meeting of the Association for Research in Vision and Ophthalmology (ARVO'94), and at the Third Annual Computation and Neural Systems Meeting (CNS'94).

[11]The filter values are termed "relative" to allow for size and position invariance of faces.

References

Barron, A. R. (1993). Universal approximation bounds for superposition of a sigmoidal function. *IEEE Transactions on Information Theory, 39*, 930-945.

Bartram, D. J. (1974). The role of visual and semantic codes in object naming. *Cognitive Psychology, 6*, 325-356.

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

Biederman, I. (1995). Visual object recognition. In D. N. Osherson & S. F. Kosslyn (Eds.), *An Invitation to Cognitive Science* (pp. 121-165). Cambridge, MA: MIT Press.

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

Biederman, I. & Cooper, E. E. (1991b). Priming contour-deleted images: Evidence for intermediate representations in visual object recognition. *Cognitive Psychology, 23*, 393-419.

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

Biederman, I., Hilton, H. J. and Hummel, J. E. (1991). Pattern goodness and pattern recognition. In J. R. Pomerantz & G. R. Lockhead (Eds.), *The Perception of Structure* Ch. 5. pp 73-95. Washington, D.C: APA.

Biederman, I. & Ju, G. (1988). Surface vs. Edge-Based Determinants of Visual Recognition. *Cognitive Psychology, 20*, 38-64.

Blickle, T. W. (1989) *Recognition of contour deleted images*. Unpublished doctoral dissertation, State University of New York at Buffalo.

Bruce, V., Valentine, T. and Baddeley, A. (1987). The basis of the 3/4 view advantage in face recognition. *Applied Cognitive Psychology, 1*, 109-120.

Buhmann, J., Lange, J. and von der Malsburg, C. (1989). Distortion invariant object recognition by matching hierarchically labeled graphs. In *IJCNN'89*, 1 (pp. 155-159). Washington.

Burbeck, C. A. & Pizer, S. M. (1995). Object representation by cores - identifying and representing primitive spatial regions. *Vision Research, 35*, 1917-1930.

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

Daugman, J. G. (1984). Spatial visual channels in the Fourier plane. *Vision Research, 24*(9), 891-910.

Daugman, J. G. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America., 2*(7), 1160-1179.

Daugman, J. G. (1988). Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression. *IEEE Transactions on Acoustics, Speech and Signal Processing, 36*, 1169-1179.

Desimone, R. & Ungerleider, L. G. (1989). Neural mechanisms of visual processing in monkeys. In F. Boller & J. Grafman (Eds.), *Handbook of Neurophysiology* . Elsevier Science Publishers B. V. (Biomedical Division). pp. 267-299

DeValois, R. L. & DeValois, K. K. (1988). *Spatial vision*. New York, N. Y. Oxford Press.

Dickinson, S. J., Pentland, A. P. and Rosenfeld, A. (1992). From volumes to views: An approach to 3-D object recognition. *Computer Vision, Graphics, and Image Processing: Image Understanding, 55*, 130-154.

Edelman, S. (1995). Representation of similarity in 3-dimensional object discrimination. *Neural Computation, 7*, 408-423.

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

Ellis, R. & Allport, D. A. (1986). Multiple levels of representation for visual objects: A behavioural study. In A. G. Cohn & J. R. Thomas (Eds.), *Artificial intelligence and its applications* New York: Wiley. pp. 245-257.

Felleman, D. J. & Van Essen, D. C. (1991). Distributed Hierarchical Processing in the Primate Cerebral Cortex. *Cerebral Cortex, 1*, 1-47.

Field, D. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America. A, Optics and Image Science, 4*, 2379-2394.

Field, D. (1994). What is the goal of sensory coding? *Neural Computation, 6*, 559-601.

Fiser, J. & Biederman, I. (1995). Size invariance in visual object priming of gray scale images. *Perception, 24*, 741-748

Heeger, D. J. (1994). The representation of visual stimuli in primary visual cortex. *Current Directions in Psychological Science, 3*, 159-163.

Heitger, F., Rosenthaler, L., von der Heydt, R., Peterhans, E. and Kubler, O. (1992). Simulation of neural contour mechanisms: from simple to end-stopped cells. *Vision Research, 32*, 963-981.

Horn, B. K. P. (1975). Obtaining shape from shading information. In P. H. Winston (Eds.), *The Psychology of Computer Vision*. New York: McGraw-Hill. pp. 115-155

Hummel, J. E. & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review, 99*, 480-517.

Johnston, A., Hill, H. and Carman, N. (1992). Recognising faces: effects of lighting direction, inversion, and brightness reversal. *Perception, 21*, 365-375.

Jones, J. P. & Palmer, L. A. (1987a). An evaluation of the two dimensional Gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology, 58*, 1233-1258.

Jones, J. P. & Palmer, L. A. (1987b). The two-dimensional spatial structure of simple receptive fields in cat striate cortex. *Journal of Neurophysiology, 58*, 1187-1211.

Kalocsai, P., Biederman, I. and Cooper, E. E. (1994). To what extend can the recognition of unfamiliar faces be accounted for by a representation of the direct output of simple cells. *Investigative Ophthalmology & Visual Science, 35*, 1626-1626.

Kalocsai, P., Biederman, I., Fiser, J. and Fang, P. (1996). Do complementary images (in the Fourier-domain) of faces and objects prime each other? *Investigative Ophthalmology & Visual Science, 37*, 841-841

Kimia, B. B., Tannenbaum, A. R., & Zucker, S. W. (1995). Shapes, shocks and deformations: 1. The components of 2-dimensional shape and the reaction-diffusion space. *International Journal of Computer Vision,* 15, pp. 189-224.

King, I. & Xu, L. (1994). Global and local PCA for face feature extraction. Presented at the Workshops *at 8th Annual Conference on Neural Information Processing Systems,* Vail, CO

Kirkpatrick, S., Gelatt, C. D. and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science, 220*, 671-680.

Kobatake, E. & Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *Journal of Neurophysiology, 71*(3), 856-867.

Koenderink, J. J. & van Doorn, A. J. (1980). Photometric invariants related to solid shape. *ACTA OPTICA , 27*, 981-996.

Kulikowski, J. J., Marcelja, S. and Bishop, P. O. (1982). Theory of spatial position and spatial frequency relations in the receptive fields of simple cells in the visual cortex. *Biological Cybernetics, 43*, 187-198.

Lades, M., Vorbruggen, J. C., Buhmann, J., Lange, J., von der Malsburg, C., Würtz, R. P. and Konen, W. (1993). Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers, 42*, 300-311.

Mallat, S. G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Journal of Pattern Analysis and Machine Intelligence, 11*, 674-693.

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

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

Phillips, P. J. (1994). Matching pursuit filters applied to face identification. In *SPIE Conference on Automatic Systems for the Identification and Inspection of Humans*, San Diego, CA:

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

Poggio, T. & Girosi, F. (1990). Regularization algorithms for learning that are equivalent to multilayer networks. *Science, 247*, 979-981.

Shapley, R. (1994). Linearity and nonlinearity in cortical receptive fields. In G. R. Bock & J. A. Goode (Eds.), *Higher-order Processing in the Visual System* John Wiley & Sons.

Srinivas, K. (1995). *Contrast and illumination effects on explicit and implicit measures of memory* Unpublished ms. Boston College.

Turk, M. & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience, 3*, 71-86.

Weiss, Y. & Edelman, S. (1995). Representation of similarity as a goal of early visual processing. *Network, 6*, 19-41.

Zhong, S. & Mallat, S. (1990). Compact image representation from multiscale edges. In The Proceedings of the 3rd *International Conference on Computer Vision*, Osaka, Japan

APPENDIX:

Detailed Analysis of System Performance: Effects of Global Shape and Local Contour

In this section the performance of the FRS was analyzed by investigating two issues:

1) Why did the grids manifest more distortion with line drawings of objects than with faces?
2) What are the relations between the amount of contour deleted, the magnitude of grid distortion, and the FRS's performance?

The reason why grids of line drawings distort more than grids of faces is that many of the high frequency detectors, when encoding a line drawing, will be positioned over blank space and consequently will have little activation at any of the orientations. Because the activity of such detectors will be similar over large regions of the image, they can be readily moved to different positions at low cost, thus distorting the grid. Although there will always be activity in the low frequency detectors with their larger receptive fields, these will tend to be activated by many different lines of (most likely) different orientations. This low frequency activity will thus tend not to be diagnostic for a particular object. In the case of faces, there will always be local surface information that will tend to render some part of the frequency domain diagnostic for that region, thus anchoring the node positions.

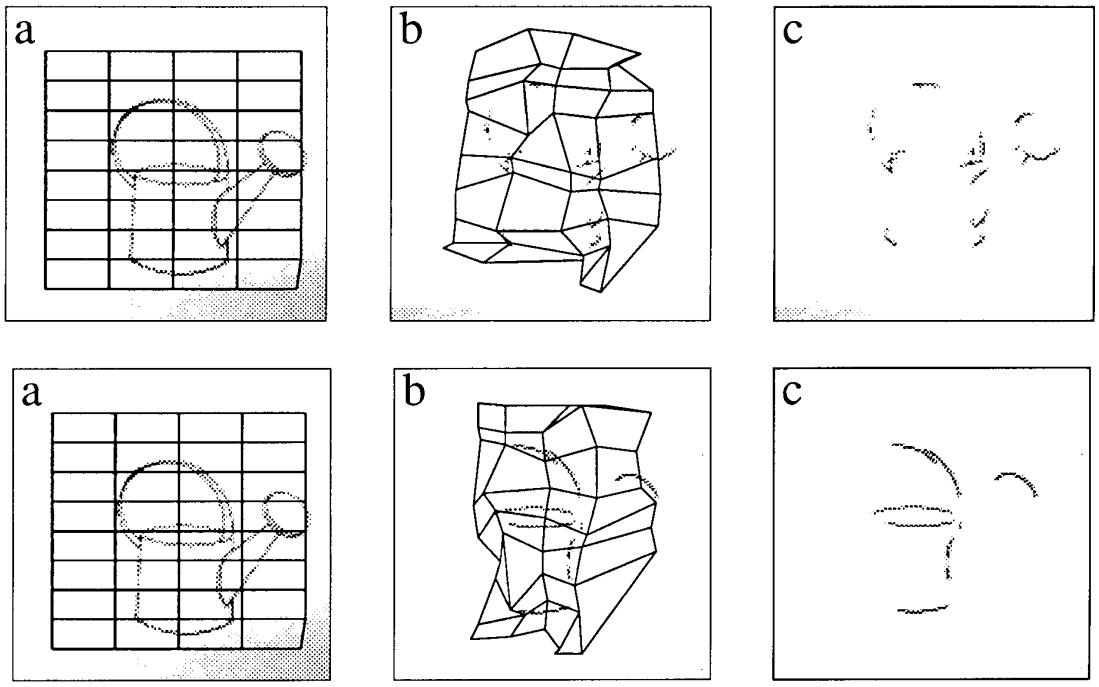

Figure 15. Comparison of the grids for the images with nonrecoverable (lower) and recoverable (upper) deletion to the intact version. The right side of the grids (in column b) are similarly distorted. The left side of the grid on the line drawing with nonrecoverable deletion (top figure in column b) shows greater distortion than the recoverable version because the nonrecoverable version has no contour from the leftmost edge of the object.

The regularity of a grid is not a good indicator of the confidence of the match, or the perceptual significance of a given segment for humans, it merely signals the global shape of the input object. The sensitivity of the system to global shape is illustrated by Figure 15, which shows the results of comparing a nonrecoverable and recoverable version of a watering can to the intact image. As is clear from the right sides of the two grids, it does not matter whether the deletion of the contours is recoverable or nonrecoverable: As long as information about the *extension* of the parts is present, the grids are fairly similar. However, on the lower left side of the can there is a sizable difference in the extent of the two partially deleted images. This difference is obviously detected by the system, and the collapse of the nodes is much more pronounced in the nonrecoverable case than in the recoverable case. Because, in general, the recoverable deletions tended to match the global shape of the object slightly better than the nonrecoverable versions, the performance of the system with such deletions was slightly better (Figure 8). In contrast, Biederman (1987) has demonstrated that even when much more contour is deleted from the recoverable versions than from the nonrecoverable versions (they differed more in global shape from the original than the nonrecoverable versions), humans can still identify the recoverable images much more readily than the nonrecoverable pictures.

The modification of the information in the jets that results from the deletion of contour in line drawings shows that whenever contour is deleted, the system's performance monotonically deteriorates with increasing amounts of contour deletion as occurred with the Nonrecoverable - Intact Gallery comparisons relative to highly accurate results of the Complement A - Gallery A comparison. The gradual deterioration of the system's performance with increasing proportions of contour deletion nicely reflects the increase in RTs and error rates evidenced by human subjects attempting to identify briefly presented images (Biederman, 1987; Blickle, 1989). However, with respect to the importance of real-time measures of human performance, we note that humans, given enough time, would be able to recognize all but the 100% contour-deleted images (Biederman, 1987).

To illustrate the points above Figure 16 shows comparisons of the intact version of the elephant with itself (0% contour deleted), and with 10%, 40%, 70%, and 100% (noise from the prevailing luminance distribution over the white paper) contour-deleted versions. For the line segments were removed evenly from the midsegment of each individual contour, the global shape of the images do not change. The 100% comparison used a noisy background to illustrate the matching of the system to random input with nonzero jet elements. As it is shown in Figure 16, the grids remain fairly regular, even in the last case, despite the obvious lack of any similarity between the noise background and the stored pattern.

Table 1 shows the comparison costs after the rigid positioning phase and after finishing the matching process. Both the continuous, monotone deterioration of performance and the gradual increase of improvement in the individual diffusion phase are clearly demonstrated as the percentage of contour deletion increases. The last row of Table 1 together with the regularity of the grid over noisy background in Figure 16 illustrates that it is mostly the discrimination capacity of the nodes that deteriorates with contour deletion.

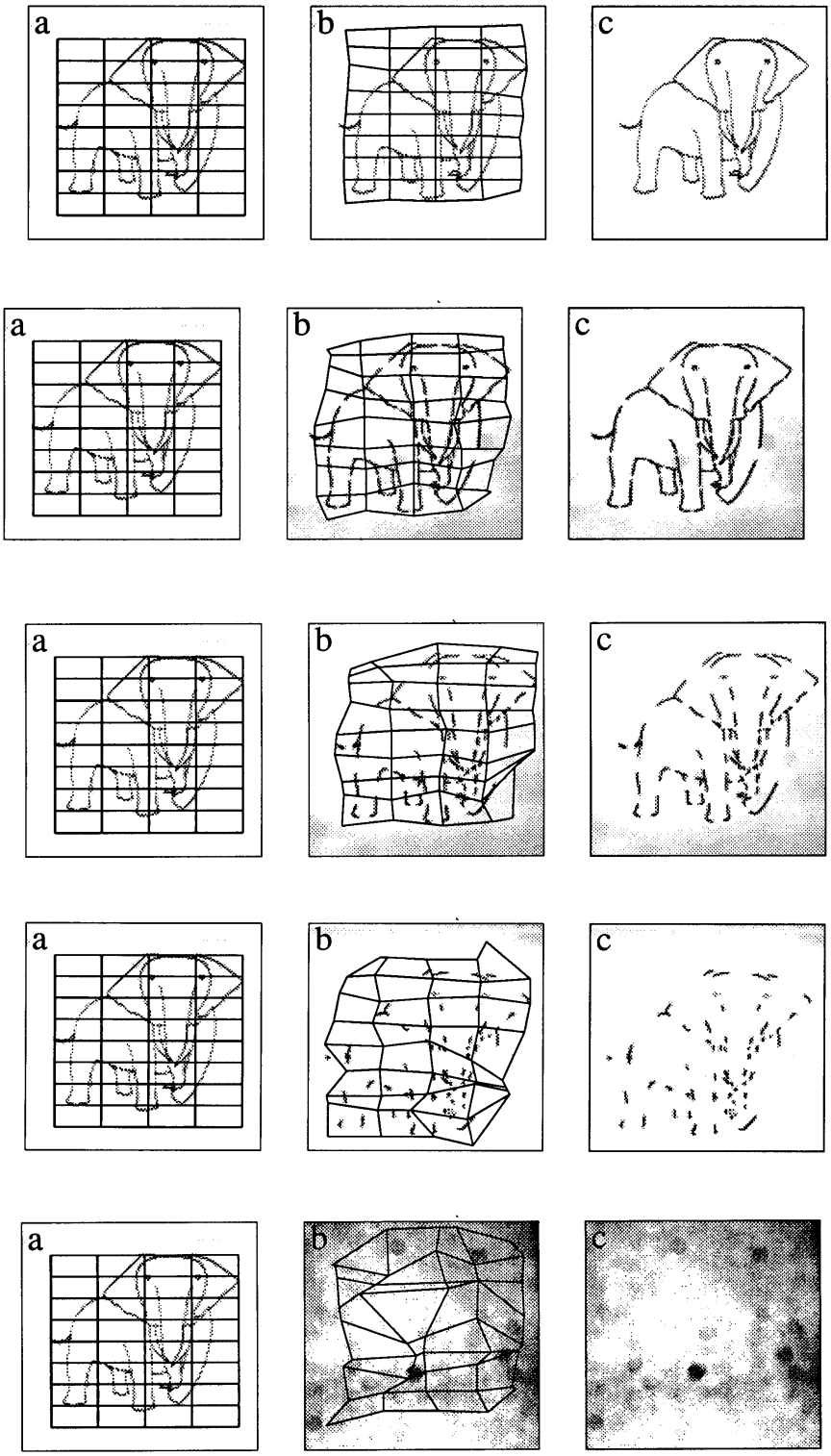

Figure 16. Results of comparing intact line drawings to 0%, 10%, 40%, 70% and 100% contour-deleted versions (top to bottom, respectively). The 100% contour-deleted image consists only of the variation from the luminance distribution on the page from the prevailing light.

Contour Percentage Deleted	Before Individual Diffusion	After Finishing The Matching Process
0	0%	0%
10	25%	9%
40	34%	16%
70	41%	19%
100	84%	62%

Table 1. Costs of matching line drawings of the elephant shown in Figure 16, with different amounts of their contour deleted at midsegment against the intact version.