Communicated by Shimon Edelman

SEEMORE: Combining Color, Shape, and Texture Histogramming in a Neurally Inspired Approach to Visual Object Recognition

Bartlett W. Mel
Department of Biomedical Engineering, University of Southern California, Los Angeles, California 90089 USA

Severe architectural and timing constraints within the primate visual system support the conjecture that the early phase of object recognition in the brain is based on a feedforward feature-extraction hierarchy. To assess the plausibility of this conjecture in an engineering context, a difficult three-dimensional object recognition domain was developed to challenge a pure feedforward, receptive-field-based recognition model called SEEMORE. SEEMORE is based on 102 viewpoint-invariant nonlinear filters that as a group are sensitive to contour, texture, and color cues. The visual domain consists of 100 real objects of many different types, including rigid (shovel), nonrigid (telephone cord), and statistical (maple leaf cluster) objects and photographs of complex scenes. Objects were individually presented in color video images under normal room lighting conditions. Based on 12 to 36 training views, SEEMORE was required to recognize unnormalized test views of objects that could vary in position, orientation in the image plane and in depth, and scale (factor of 2); for nonrigid objects, recognition was also tested under gross shape deformations. Correct classification performance on a test set consisting of 600 novel object views was 97 percent (chance was 1 percent) and was comparable for the subset of 15 nonrigid objects. Performance was also measured under a variety of image degradation conditions, including partial occlusion, limited clutter, color shift, and additive noise. Generalization behavior and classification errors illustrate the emergence of several striking natural shape categories that are not explicitly encoded in the dimensions of the feature space. It is concluded that in the light of the vast hardware resources available in the ventral stream of the primate visual system relative to those exercised here, the appealingly simple feature-space conjecture remains worthy of serious consideration as a neurobiological model.

1 Introduction

The human visual system can recognize unprimed views of common objects at sustained rates in excess of 10 per second (Potter, 1976; Subramaniam,

Biederman, Kalocsai, & Madigan, 1995). In neurophysiological terms, the implication of this astonishing fact is that an object's identity can be computed by a 100-ms-wide time slice of neural activity propagating through the recognition "pipe" from the retina to the higher reaches of the visual system. At typical firing rates of visual neurons, we may surmise that under these special conditions, each neuron in the visual pathway devotes fewer than a dozen action potentials to the processing of any single object's image. While these figures bear primarily on the throughput of the recognition pipe (objects per second), a recent electroencephalogram study reveals that the length of the pipe from retina to recognition is probably no longer than 155 ms (DiGirolamo & Kanwisher, 1995). These results are also roughly consistent with known response latencies of neurons in the object recognition areas of inferotemporal cortex, which range around 100 ms (Oram & Perrett, 1992; Gochin, Colombo, Dorfman, Gerstein, & Gross, 1994). Beyond such feats of sheer speed, the human visual system easily recognizes a wide variety of stimuli, including rigid, articulated, and entirely nonrigid two-dimensional (2D) and three-dimensional (3D) objects, faces, statistical or fractal objects, surface textures, and views of complex natural scenes, while displaying remarkable insensitivity to changes in viewpoint, partial occlusion, and clutter.

How can a visual system work so well and so fast? The circumstances of biological vision suggest a solution of low time complexity (few steps) but high space complexity (many parallel processes)—synaptic connections among neurons in the cortical object recognition pathway probably number between 10 and 100 trillion.[1] One appealingly simple notion is that the visual system is organized as a feedforward feature-extracting hierarchy that builds progressively more complex and viewpoint-invariant features useful for identifying objects, where invariance over a group of transformations is achieved by summing over viewpoint-specific elemental representations (Pitts & McCullough, 1947). This class of algorithm has been brought to bear with great success, for example, in the realm of optical character recognition (Fukushima, Miyake, & Ito, 1983; LeCun et al., 1990). According to this view, and in correspondence with the axioms of statistical pattern recognition, visual object recognition in the brain is the process of mapping retinal pixels into a feature space that is better suited (than pixels) to the viewpoint-invariant classification and identification problems faced by visual animals. Within this feature space, represented by the activity of neurons at the top of the hierarchy, the similarity of one object view to another is given by a simple (e.g., Euclidean) distance calculation; recognition of an input is achieved by finding the identity or class of the most similar training view previously

[1] This assumes 10^{14} to 10^{15} synapses in the brain, 90 percent of these due to the cerebral cortex, one-third of these due to the visual system, and one-third of these accounting for the "what" pathway.

stored in memory. Feature dimensions are chosen such that large changes in object pose produce relatively small excursions in feature space, while small changes in object "quality" (shape, texture, color) produce relatively large excursions in feature space. Where 3D objects or scenes are to be recognized over large regions of the viewing sphere, a small set of reference views is stored during learning, leading to a "view-based" approach (Edelman & Bulthoff, 1992; Logothetis, Pals, Bulthoff, & Poggio, 1994; Murase & Nayar, 1995).

In evaluating the plausibility of this feedforward feature-extraction scenario as a model for biological vision, it is useful to consider several desirable properties of a feature-space representation as dictated by the "computational ecology" of natural vision. In particular, we might expect the brain to construct visual features that are:

1. *Large in number*, relating to the fact that sparse, high-dimensional feature representations provide fundamental advantages for fast, inexpensive recognition, especially large signal-to-noise ratios that allow objects to be recognized prior to explicit segmentation (Califano & Mohan, 1994; Kanerva, 1988).

2. *Useful*, that is, high-level features should be relatively sensitive to object quality—and hence identity—but relatively insensitive to an object's pose or configuration.

3. *Dominated by spatially localized measures*, relating to the need to cope with nonrigid object transformations, which often preserve local but not global structure; the need to cope with object textures, defined in large part by local relative-orientation structure; and the need to cope with occlusion and clutter, which are least disruptive to an object's internal code when derived from features with localized support.

4. *Driven by multiple visual cues*, relating to the need to maximize object discrimination power by using all available visual cues, the need to represent objects of many different types richly, and the need to buffer the visual representation of objects or scenes against a variety of forms of image degradation, to which different visual cues are by nature differentially sensitive.

In this context, recent neurophysiological results (Kobatake & Tanaka, 1994; Logothetis et al., 1994) that have extended classical results from other groups (Gross, Rocha-Mironda, & Bender, 1972; Perrett, Rolls, & Caan, 1982; Schwartz, Desimone, Albright, & Gross, 1983; Desimone, Albright, Gross, & Bruce, 1984) are intriguing, as they demonstrate a substantial population of neurons in the "object recognition areas" of the primate visual system (Mishkin, 1982) that respond best to specific complex minipatterns (e.g., localized conjunctions of contour, texture, and color elements). In many cases, these neurons exhibit considerable insensitivity to changes in viewpoint-

related parameters, such as stimulus position and scale (Ito, Tamura, Fujita, & Tanaka, 1995), while remaining selective for their preferred minipatterns.

Such empirical data are thus on their face suggestive of a feature-extraction hierarchy designed to cope with known ecological pressures. However, serious theoretical concerns persist regarding the limitations of bottom-up feature-space approaches, cropping up under a variety of guises:

1. Feature-space methods are essentially template-matching methods and so require an intractable number of templates to cope with inputs that have been rotated, scaled, partly occluded, nonrigidly transformed, or presented under varying lighting conditions.
2. Feature-space approaches lack a top-down component essential for resolving featural ambiguities present in real images.
3. Feature-space methods do not scale well to high dimensions (i.e., large numbers of features), either because it is not practical to learn or otherwise assemble a sufficiently large number of sufficiently useful features, too many dimensions of noise necessarily overwhelm too few dimensions of signal, or high-dimensional methods are computationally intractable or require too much data, even for a brain.

Most of these concerns have been addressed in recent years, as feature-space approaches and their variants have been applied with success in a variety of object recognition domains (Fukushima et al., 1983; Swain and Ballard, 1991; Viola, 1993; Lades et al., 1993; Califano & Mohan, 1994; Murase & Nayar, 1995; Rao & Ballard, 1995; Amit, Geman, & Wilder, 1995; Schiele & Crowley, 1996). However, the conjecture that a feature-space approach based on feedforward receptive-field-style computations could account for the prodigious recognition capacities of the primate brain as yet lacks direct support in the modeling literature. Existing approaches have generally involved one or more assumptions that place them squarely outside the biological "paradigm" for general-purpose visual recognition, such as use of small object corpus (typically no more than a few dozen objects), limited range of object types (e.g., faces or rigid volumetric objects), feature computations not amenable to receptive-field-style computations (e.g., use of sophisticated geometric invariants), or strong viewpoint assumptions, or corresponding explicit image prenormalization operations (typical in optical character recognition and face recognition).

Against this backdrop, SEEMORE was developed to assess more directly the plausibility of the feature-space account as a biological model for rapid, general-purpose object recognition. Key design goals prescribed a visual representation that (1) relied exclusively on geometrically simple receptive-field-style computations, (2) operated directly on input images without shift, scale, or other explicit object prenormalization steps, (3) could cope with a large number of real 3D objects of many different types, (4) could recognize objects over 6 degrees of freedom of viewpoint, gross nonrigid shape dis-

tortions, and partial occlusion, and (5) exhibited a significant capacity for generalization across natural object categories.

SEEMORE's visual representation is composed of 102 feature channels that emphasize spatially localized filter computations and are collectively sensitive to contour shape, color, and texture cues. SEEMORE's architecture is similar in spirit to the color histogramming approach of Swain and Ballard (1991) but includes spatially structured features that also provide for shape-based generalization. Experiments reveal good recognition performance in a viewpoint- and configuration-invariant 3D object recognition problem with 100 objects, including objects that are rigid, nonrigid, and "statistical" in nature and photographs of complex scenes. Perhaps most interesting, SEEMORE's patterns of generalization reveal the emergence of several striking object catagories that are not explicitly encoded in the 102 feature-space dimensions. (A short report describing this work appeared in Mel, 1996.)

2 Methods

2.1 SEEMORE'S Visual World. SEEMORE's database contains 100 common 3D objects and photogaphs of scenes, each represented by a set of presegmented color video images (see Figure 1). The training set consisted of 12 to 36 views of each object as follows. For rigid objects, 12 training views were chosen at roughly 60-degree intervals in depth around the viewing sphere (see Figure 2A), and each view was then scaled to yield a total of three images at 67 percent, 100 percent, and 150 percent. Image plane orientation was allowed to vary arbitrarily. For nonrigid objects, 12 training views were chosen in random poses (see Figure 2B).

During a recognition trial, SEEMORE was required to identify novel test images of the database objects. For rigid objects, test images were drawn from the viewpoint interstices of the training set, excluding highly foreshortened views (e.g., bottom of can). Each test view could therefore be presumed to be correctly recognizable but never closer than roughly 30 degrees in orientation in depth or 22 percent in scale to the nearest training view of the object, while position and orientation in the image plane could vary arbitrarily. For nonrigid objects, test images consisted of entirely novel random poses. Each test view depicted the object presented alone on a smooth background.

In some recognition trials, test views were systematically degraded. The five degradation conditions were (1) *scrambled*, in which the object view was cut up and the pieces rearranged and reflected (see Figure 3A), (2) *occluded*, in which 50 percent of the object view was blacked out (see Figure 3B), (3) *cluttered*, in which a second object—a randomly colored lowercase character—was superimposed onto the margin of the image to add limited pattern clutter while minimizing occlusion (see Figure 3C), (4) *colorized*, in which two of the three RGB channels were scaled either up or down by 30 per-

Figure 1: The database contains 100 objects of many different types, including rigid (soup can), nonrigid (necktie), statistical (bunch of grapes), and photographs of complex indoor and outdoor scenes.

cent while maintaining constant intensity (see Figure 3D), and (5) *noisy*, in which the images were subjected to uniform additive pixel noise with mean 30 percent of the maximum pixel value (see Figure 3E).

2.2 Feature Channels. SEEMORE's internal representation of a view of an object is encoded by a set of feature channels. The ith channel is based on an elemental nonlinear filter $f_i(x, y, \theta_1, \theta_2, \ldots)$, parameterized by position in the visual field and zero or more internal degrees of freedom (see Figure 4). Each channel is by design relatively sensitive to changes in the image that are strongly related to object identity, such as the object's shape, color, or

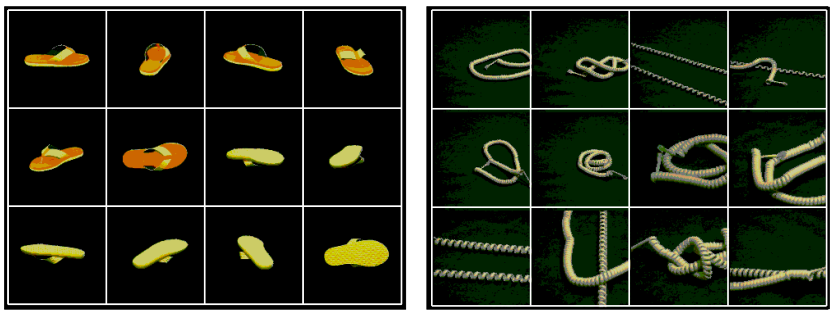

Figure 2: (A) Training views of rigid objects were sampled uniformly about the 3D viewing sphere. (B) Training views of nonrigid objects were drawn at random from the object's configuration space.

texture, while remaining relatively insensitive to changes in the image that are unrelated to object identity, such as are caused by changes in the object's pose. In practice, this invariance is achieved in a straightforward way for each channel by subsampling and summing the output of the elemental channel filter over the entire visual field and one or more of its internal degrees of freedom, giving a channel output $F_i = \sum_{x,y,\theta_1,\ldots}, f_i()$. For example, a particular shape-sensitive channel might "look" for the image-plane projections of right-angle corners, over the entire visual field, 360 degrees of rotation in the image plane, 30 degrees of rotation in depth, one octave in scale, and tolerating partial occlusion or slight misorientation of the elemental contours that define the right angle. In general, then, F_i may be viewed as a "cell" with a large receptive field whose firing rate is an estimate of the number of occurrences of distal feature i in the visual work space over a large range of viewing parameters.

SEEMORE's architecture consists of 102 feature channels, whose outputs form an input vector to a nearest-neighbor classifer (see Figure 5). Following the design of the individual channels, the channel vector $\mathbf{F} = \{F_1, \ldots, F_{102}\}$ is (1) insensitive to changes in image plane position and orientation of the object, (2) modestly sensitive to changes in object scale, orientation in depth, or nonrigid deformation, but (3) highly sensitive to object quality as pertains to object identity. Within this representation, total memory storage for all views of an object ranged from 1224 to 3672 integers.

As shown in Figure 6, SEEMORE's channels fall into in five groups: (1) 23 color channels, each of which responds to a small blob of color parameterized by "best" hue and saturation, (2) 11 coarse-scale intensity corner channels parameterized by open angle, (3) 12 "blob" features, parameter-

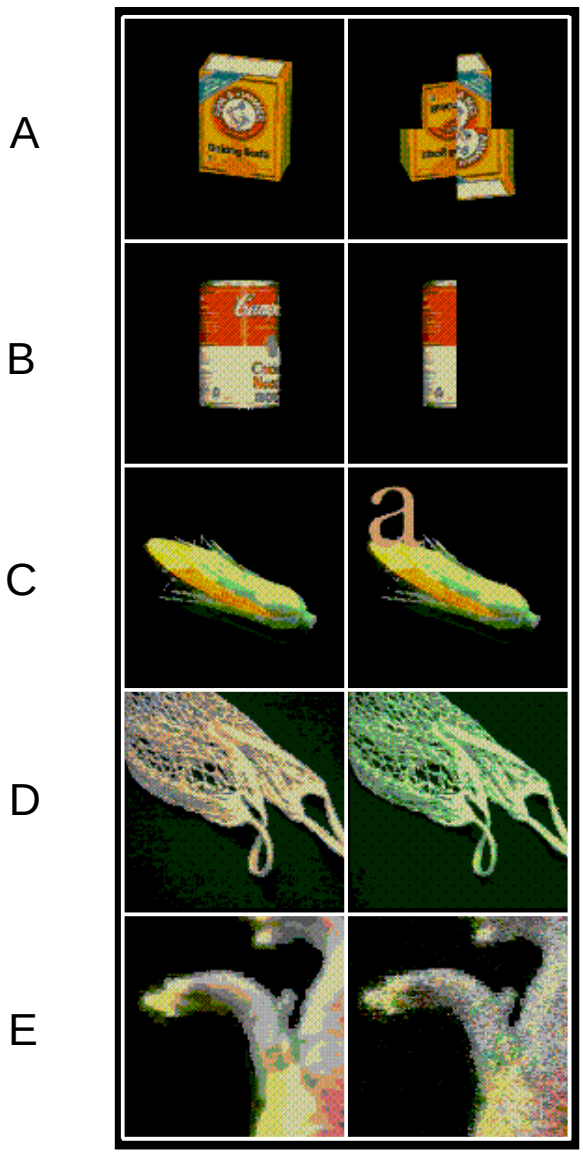

Figure 3: Sensitivity of recognition performance to various forms of image degradation was studied. Examples of the five degradation conditions are shown: (A) scrambled, (B) occluded, (C) cluttered, (D) colorized, and (E) noisy.

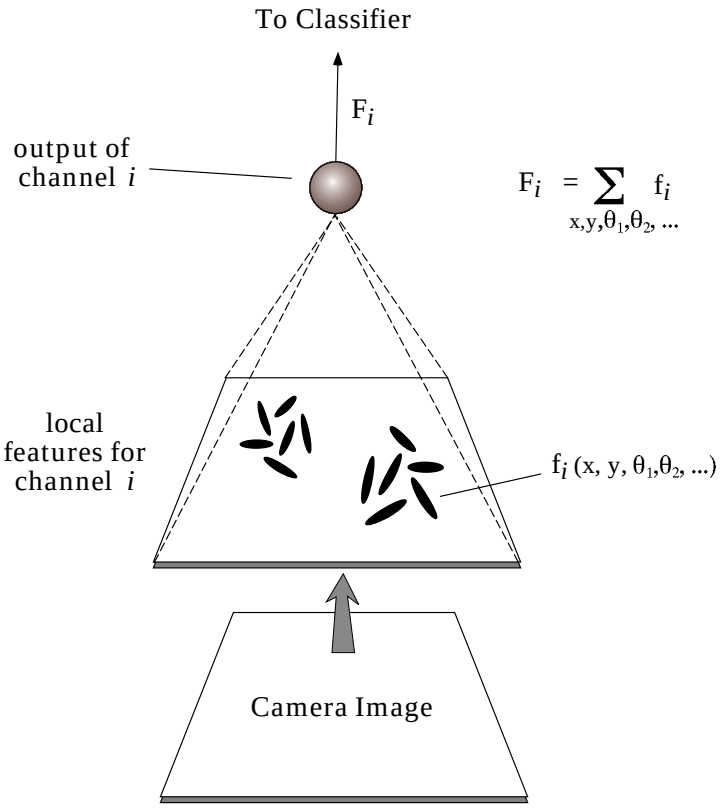

Figure 4: The ith feature channel is based on an elemental nonlinear filter $f_i(x, y, \theta_1, \ldots)$, which is subsampled across the image, at a range of image-plane orientations, and over any of the filters' additional internal degrees of freedom. The output of the ith channel F_i is the sum over all elemental filter outputs within that channel.

ized by the shape (round and elongated) and size (small, medium, and large) of bright and dark intensity blobs, (4) 24 contour-shape features, including straight angles, curve segments of varying radius, and parallel and oblique line combinations, and (5) 16 shape- and texture-related features based on the outputs of Gabor functions at five scales and eight orientations. The implementations of the channel groups were crude, in the interest of achieving a working, multiple-cue system with minimal development time. Images were grabbed using an off-the-shelf Sony S-Video Camcorder and SunVideo digitizing board that provided 11 color bits per pixel (YUV); images were

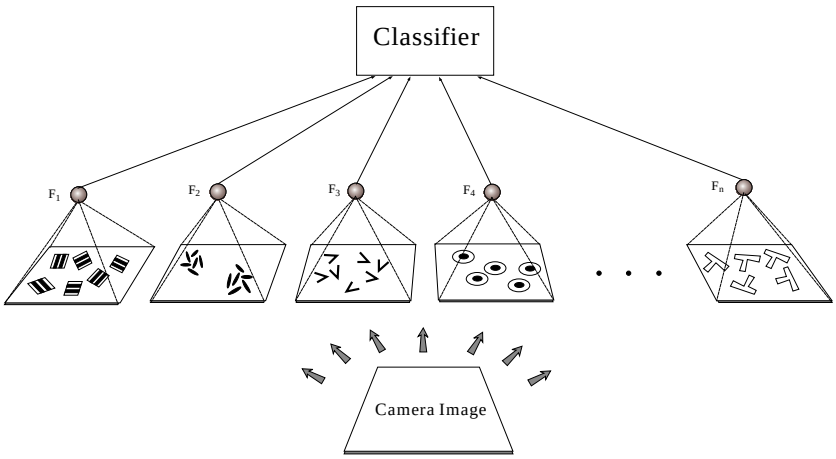

Figure 5: SEEMORE's architecture consists of a set of channels that provides input to a nearest-neighbor classifier. Since each channel is by itself insensitive to translation and rotation in the image plane and is slowly varying as an object rotates in depth, the output vector **F** for the complete set of channels is similarly invariant to these transformations. However, **F** remains highly sensitive to changes in object quality, and hence object identity.

converted to RGB format for all subsequent processing. Implementation details of the 102 feature channels are described below.

2.2.1 Twenty-three Color Channels. The image was first low-pass-filtered by two octaves using a five-element separable mask (1/16, 1/4, 3/8, 1/4, 1/16), and subsampled to a size of 120×120 color pixels. In the following, for $R, G, B \in [0, 1]$, hue, saturation, and intensity (*HSI*) are defined as $H = \arctan(\sqrt{3}(G-B)/((R-G)+(R-B)))$, $S = 1.0 - \min(R, G, B)/I$, $I = (R+G+B)/3$, and sigmoids g and h with threshold θ and gain s are defined as $g(x, \theta, s) = 1/(1+\exp((\theta - x) * s))$, and $h = 1 - g$. Of the 23 color channels, 22 (11 hues at 2 levels of saturation) were derived as follows. Around the hue circle, 1D "receptive field" centers were placed at 11 evenly spaced hues, including 0 degree. Response profiles hue_i, $i \in \{1, \ldots, 11\}$ were triangular and symmetrical, returning a peak score of 1.0 in response to the center hue, with a linear decay to 0 over a 45-degree radius. The 11 high- and low-saturation unit pairs shared the same hue centers. The high-saturation unit outputs y_i^{hisat} were computed as a product of the hue score and two sigmoidal factors that (1) fell to 0 as the saturation crossed into the range of the low-saturation units and (2) fell to 0 at low-intensity values where chromaticity

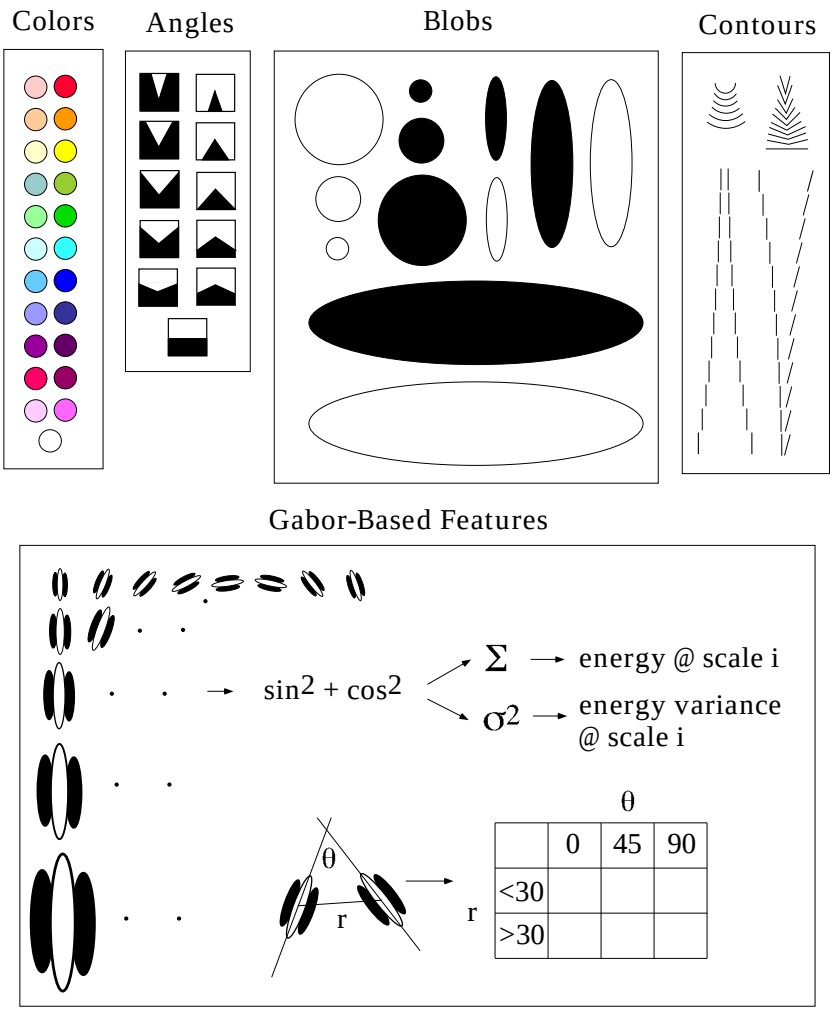

Figure 6: SEEMORE's 102 channels fall into five groups: (1) 23 circular hue/saturation channels, (2) 11 coarse-scale intensity corner channels, (3) 12 circular and oriented-intensity blobs, (4) 24 contour-shape features, including curves, junctions, and parallel and oblique contour pairs, and (5) 16 oriented-energy and relative-orientation features based on the outputs of Gabor functions at several scales and orientations.

information is unreliable, giving $y_i^{hisat} = \text{hue}_i \cdot g(S, 0.4, 10) \cdot g(I, 0.2, 20)$. The low-saturation unit outputs y_i^{lowsat} were computed as a product of the hue score and three sigmoidal factors—the additional sigmoid was needed to bound the saturation valued both from above and below—giving $y_i^{lowsat} = \text{hue}_i \cdot h(S, 0.4, 10) \cdot g(S, 0.15, 20) \cdot g(I, 0.2, 20)$. A separate white unit selected for pixels of high intensity and low saturation, with an output given by $y^{white} = g(I, 0.6, 20) \cdot h(S, 0.15, 20)$. Black pixels were ignored. Each of the 23 color functions was evaluated at every pixel in the image; any response value exceeding 0.1 generated a count within a histogram containing 23 bins, one for each "color."

2.2.2 Eleven Intensity Corner Channels. The image was first low-pass-filtered and subsampled by two octaves to a size of 120 × 120 intensity pixels. A crude-oriented intensity edge detector was run in 12 increments of 30 degrees around the circle centered at each pixel; the length of the edge was approximately 20 pixels when mapped back into the full-resolution image (480 × 480). The criterion for detection of an oriented edge was twofold: (1) An intensity gradient of a consistent sign should exist across the edge at multiple sites along the length of the edge, and (2) the mean deviation of intensity (from its average) should be small along the length of the edge on both sides.

Both criteria were computed based on pairs of pixels symmetrically straddling the edge along its length, as shown in Figure 7. The numerical score for an edge defined in terms of left and right pixels from three pixel pairs indexed by i was given by $y = \prod_i g(|I_i^R - I_i^L|, 0.08, 30) - \lambda \sum_i (|I_i^R - \overline{I_i^R}| + |I_i^L - \overline{I_i^L}|)$, where $\lambda = 0.8$ controlled the relative importance of the gradient versus smoothness constraints. The result was passed through a hard binary threshold giving 1 for $y > 0$, 0 otherwise. The 12 oriented binary edge maps were used to find intensity corners, which were sought at every pixel. Corners were parameterized only by open angle (from 30 to 330 degrees in 30-degree increments), yielding 11 corner types and 11 corresponding histogram bins. A corner was scored whenever two edge segments of the proper relative orientation and offset were found in the image at any absolute orientation, and the corresponding histogram bin was incremented.

2.2.3 Twelve Intensity Blob Channels. The image was first low-pass-filtered and subsampled by two, three, and four octaves to yield image sizes of from 120 × 120 to 30 × 30 intensity pixels. The operations described below were carried out at each of the three scales. An intensity blob consisted of a set of intensity gradients of the proper sign in an elliptical configuration about a central pixel. They were thus elliptical versions of the straight edge segments of Figure 7, based on several sigmoidally modulated pixel pair intensity differences. However, in lieu of three gradient terms straddling a straight edge combined multiplicatively, a *sum* of 12 gradient terms

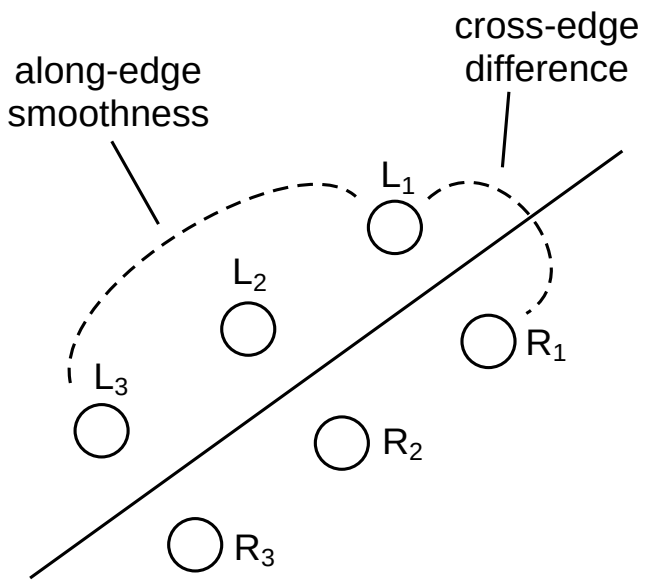

Figure 7: Straight-oriented edge segments were detected in the intensity image based on (1) sign consistency of cross-edge intensity differences, measured at several sample pairs of pixels, and (2) along-edge intensity "smoothness," measured as mean deviation in intensity on either side of putative edge.

across an elliptical contour was computed. A smoothness term was again given by the mean deviation of intensity along the inside and outside of the circular contour; this term was subtracted from the across-contour term with $\lambda = 1.0$. If a threshold of 4.0 was exceeded, then a blob was scored, and the corresponding histogram bin was incremented. Twelve total bins corresponded to light or dark, round or elongated intensity blobs, at three scales.

2.2.4 Forty Generalized Contour Channels. The image was low-pass-filtered as before but maintained at full resolution (480×480). An oriented edge detector was run that differed in three ways from that underlying the intensity corner channels. First, edges consisted of five pairs of across-edge pixels spanning a total length of approximately 18 original image pixels at full resolution, and were computed at 7.5 degree intervals around the half-circle. Second, an edge was scored iff all five gradient terms exceeded a hard threshold (in lieu of a product of sigmoids), and no along-contour smoothness penalty term was used. However, the gradient threshold differed from

edge to edge, given by an estimate of the along-edge gradient (average difference of end pixel values on both sides of edge). Third, an edge was scored if it existed according to the above criteria in any of the RGB channels, but neither the RGB channel nor the polarity of the edge was stored. As such, the output of the edge-detecting stage was a set of 24 binary contour maps representing orientations from 0 to 172.5 degrees, in 7.5-degree intervals at a resolution of 240 $\times$ 240 pixels. Every edge in the binary edge maps was "generalized" (copied) into a 2×2 block of neighboring pixels, in order effectively to relax the position tolerances in the subsequent compound-feature detection step. Forty compound contour features were then tested for at every pixel (240 $\times$ 240) and orientation (48 steps of 7.5 degrees). A compound feature consisted of six oriented contours with proper offsets and relative orientations; when at least five were present, the compound feature was scored, and its histogram bin was incremented. Various optimizations were used to speed computation, including heuristics for early rejection of both pixels and compound features. As shown in Figure 6, features included seven long, simple contours (straight plus 6 degrees of curvature), nine corner features with angles from 30 to 150 degrees, in 15-degree steps (analogous to intensity corners but with longer limbs, higher angular resolution, and tolerating up to one missing link), and parallel and oblique line pairs (12 each at a range of separations from 16 to 82 pixels in increments of 6 pixels).

2.2.5 Sixteen Gabor-Derived "Texture" Channels. The largest central square in the original image was reduced to 128 $\times$ 128 pixels. Forty Gabor filters were applied to the image at eight orientations and five scales using a Fast Fourier Transform (FFT) (scales varied from X pixels per cycle to Y pixels per cycle in powers of $\sqrt{2}$). Energy images were computed by squaring and summing corresponding pixels in the sine and cosine components of the Gabor output. The energy pixel values were then summed across all eight orientation images at each scale, giving the total oriented energy at each scale as the values of the first five texture channels. The next five texture channels consisted of the variances of the oriented energy at each scale, gotten by computing the mean squared deviation of the total energy at each orientation from the mean total energy for all orientations, for that scale. High variances signified images with energy distributed nonuniformly across the orientation channels, such as an image with one or two dominant orientations. The Gabor energy images were then passed through a fixed binary thresholding operation chosen for each scale to emphasize localization of perceptually relevant oriented image structure. The remaining six texture channels measured probability of relative orientation energy in the image, parameterized by orientation difference (0 degrees, 45 degrees, 90 degrees) and image distance ($d < 30$ pixels, $d \geq 30$ pixels). All relevant suprathreshold pixel pairs were considered and used to generate counts in the six histogram bins.

2.3 The Learning Rule. While nearest-neighbor classification techniques are remarkably powerful given their simplicity (Friedman, 1994), the problem of scaling feature dimensions in order to minimize classification error in high dimension remains an experimental art. The need for such optimization is particularly acute in cases, including the present case, where feature dimensions are grossly inhomogeneous in terms of their variances, entropy, and individual classification power and where strong, poorly characterized correlations exist among the features dimensions. The notion that features should be simply normalized by their variances (i.e., "sphering" the data) does not take account of the individual utility of the features for the classification problem at hand or of the correlations between features.

One approach to this problem has been to project a high-dimensional feature space onto the low-dimensional subspace in which the exemplars of each class are as tightly clustered as possible while the class means are as widely dispersed as possible (e.g., Fisher's linear discriminant; Duda & Hart, 1973). Where dimensionality reduction is not needed for computational or other reasons, classification in all available dimensions in principle allows for improved classification performance. Thus, related methods involve finding a "clustering transformation" of the input space that maximizes between-class distances while minimizing within-class distances (Fukunaga, 1990).

In the approach here, an objective function that predicts classification performance using all dimensions is maximized using gradient ascent. The formulation is based on the standard assumption that, on average, two views of the same object are more similar to each other than two views of different objects. Thus, measurements are taken centered at every view j in the database, comparing the average distance from j to views of the same object with the average distance from j to views of different objects, where the comparison is normalized by a local measure of the dispersion of interclass distances. In this way, the mean distance to views of the same object can be treated as a z-score with respect to the distribution of distances to different objects (see Figure 8). The normalization is local to each view since the dispersion of object views can vary substantially in different regions of the feature space. Classification performance is expected to increase as these z-scores increase, averaged over the entire database of views. We begin by writing the weighted city-block distance between two views represented as N-dimensional vectors j, k as

$$D_{jk} = D(j, k, \mathbf{w}) = \sum_{i=1}^{N} w_i D_{jk}^i,$$

where i is the feature index, w_i is a weight, and $D_{jk}^i = |j_i - k_i|$ is the distance along single feature dimension i. The mean distance between view j and all views k of the same object, or of different objects, is written, respectively, as

$$\overline{D}_{\equiv j} = \langle D_{jk} \rangle, \quad \forall k \equiv j$$

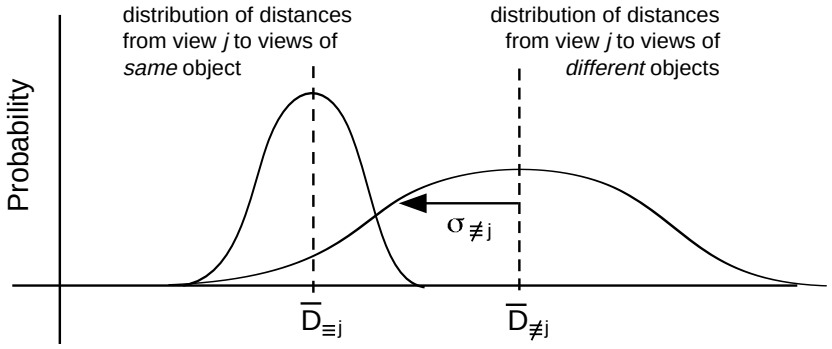

Figure 8: Graphical representation of $\overline{D}_{\equiv j}$, $\overline{D}_{\not\equiv j}$, and $\sigma^i_{\not\equiv j}$. Measured distances from j to members of same object class (shown as solid lines) give rise to distribution 1 whose mean is $\overline{D}_{\equiv j}$. Similarly, measured distances from j to members of all other classes give rise to distribution 2 whose (larger) mean is $\overline{D}_{\not\equiv j}$. The predicted "goodness" of feature i in the neighborhood of view j, G_j, is then given by the z-score $(\overline{D}_{\not\equiv j} - \overline{D}_{\equiv j})/\sigma_{\not\equiv j}$.

or

$$\overline{D}_{\not\equiv j} = \langle D_{jk} \rangle, \quad \forall k \not\equiv j,$$

where the identity sign is used to indicate all views k of the same ($\equiv j$) or different ($\not\equiv j$) object as j. Analogous mean distance quantities are defined for the single feature dimension i, that is,

$$\overline{D^i}_{\equiv j} = \langle D^i_{jk} \rangle, \quad \forall k \equiv j$$

and

$$\overline{D^i}_{\not\equiv j} = \langle D^i_{jk} \rangle, \quad \forall k \not\equiv j \; .$$

The dispersion about view j of distances to views k of different objects is given by the standard deviation, for both the one-dimensional and N-dimensional cases,

$$\sigma^i_{\not\equiv j} = \sqrt{\langle (D^i_{jk} - \overline{D^i}_{\not\equiv j})^2 \rangle}, \quad \forall k \not\equiv j$$

or

$$\sigma_{\not\equiv j} = \sqrt{\langle (D_{jk} - \overline{D}_{\not\equiv j})^2 \rangle}, \quad \forall k \not\equiv j \; .$$

Both one-dimensional and N-dimensional feature "goodness" measures may then be defined in the neighborhood of view j as

$$G_j^i = \frac{\overline{D^i}_{\not\equiv j} - \overline{D^i}_{\equiv j}}{\sigma^i_{\not\equiv j}}$$

and

$$G_j = \frac{\overline{D}_{\not\equiv j} - \overline{D}_{\equiv j}}{\sigma_{\not\equiv j}}.$$

The global N-dimensional feature goodness measure is then given by

$$G = \langle G_j \rangle.$$

G is maximized when, averaged over all views j in the database, the average distance to views of the same object is many standard deviations smaller than the average distance to views of different objects. G was optimized using gradient ascent, using the following weight-update rule:

$$\Delta w_i \propto \frac{\delta G}{\delta w_i} = \left\langle \frac{\sigma^i_{\not\equiv j}}{\sigma_{\not\equiv j}} [G_j^i - G_j \cdot \mathrm{Cov}_{\not\equiv j}(D^i, D)] \right\rangle, \quad \forall j \tag{2.1}$$

where

$$\mathrm{Cov}_{\not\equiv j}(D^i, D) = \frac{\langle (D^i_{jk} - \overline{D^i}_{\not\equiv j})(D_{jk} - \overline{D}_{\not\equiv j}) \rangle, \quad \forall k \not\equiv j}{\sigma^i_{\not\equiv j} \sigma_{\not\equiv j}}$$

is the covariance of the one-dimensional distance (for feature i) and the weighted N-dimensional city-block distance from view j to views of different objects. Loosely speaking, the square bracketed term of equation 2.1 indicates that a feature's weight tends to increase until its individual goodness accounts for a specific fraction of the global feature goodness, where the fraction is given by the covariance term.

2.4 Assessing Recognition Performance. SEEMORE's recognition performance was assessed quantitatively as follows. A test set consisting of 600 novel views (100 objects $\times$ 6 views) was culled from the database, as previously described. In some cases, the raw images were preprocessed according to one of the five degradation conditions (scrambled, occluded, colorized, noisy, or cluttered). Then the intact or degraded images were presented to SEEMORE for identification.

In the course of this work, it was noted empirically that a compressive transform on the feature dimensions (histogram values) led to improved classification performance; prior to all learning and recognition operations, therefore, each feature value was replaced by its natural logarithm (0 values

were first replaced with a small positive constant to prevent the logarithm from blowing up).

For each test view j, the distance D_{jt} was computed for every training view t in the database, and the nearest neighbor was chosen as the best match.[2] The logarithmic transform of the feature dimensions thus tied D to the ratios of individual feature values in two images rather than their arithmetic differences.

3 Results

3.1 Intact Images. Recognition time on a Sparc-20 was 1–2 minutes per view; the bulk of the time was devoted to shape processing, with under 2 seconds required for matching.

Recognition results are summarized in Table 1, reported as the proportion of test views that were correctly classified. Performance using all 102 channels for the 600 novel object views in the intact test set was 96.7 percent; the chance rate of correct classification was 1 percent. Across recognition conditions, second-best matches usually accounted for approximately half the errors. Results were broken down in terms of the separate contributions to recognition performance of color-related versus shape-related feature channels. Performance using only the 23 color-related channels was 87.3 percent, and using only the 79 shape-related channels was 79.7 percent. Remarkably, very similar performance figures were obtained for the subset of 90 test views of the nonrigid objects, which included several scarves, a bike chain, necklace, belt, sock, necktie, maple leaf cluster, bunch of grapes, knit bag, and telephone cord. Thus, a novel random configuration of a telephone cord was as easily recognized as a novel view of a shovel.

3.2 Degraded Images. Results for the scrambled, occluded, colorized, noisy, and cluttered test sets are also shown in Table 1. Under the scrambling manipulation, overall recognition performance was only modestly affected, due primarily to the stability of the color channels under this manipulation, while the shape-related channels showed a significant drop in performance as expected (from 80 to 62 percent). When test views were half-occluded, recognition performance fell to 79 percent, due to the disruptive effect of occlusion on both color- and shape-related channels; the

[2] In informal experimentation with a number of variants of the nearest-neighbor classifier, including a variety of different distance metrics, typically only small differences in performance were seen, and the reasons for these changes in performance were obscure. This work emphasized the power of the visual representation rather than the benchmarking of standard classifiers and their variants; given that nearest-neighbor classifiers are fast and conceptually simple, and are known to perform well often in high-dimensional spaces in comparison to more sophisticated classifiers (Friedman, 1994), this single method was used throughout the trials reported in this article.

Table 1: Summary of Results.

	Intact	Nonrigid	Scrambled	Occluded	Cluttered	Colorized	Noisy
Shape only	79.7	76.7	62.2	38.2	57.3	43.5	35.8
Color only	87.3	94.4	86.5	72.2	61.2	6.8	47.2
Color and shape	96.7	97.8	93.7	79.0	79.0	19.8	58.3

relatively severe disruption of the shape representation was in part due to the more spatially heterogeneous distribution of distinctive (i.e., necessary) shape cues and in part to the introduction of spurious contours and other accidental features at the occluding boundary. The color shift manipulation cut recognition performance to under 20 percent, an indication of the complete lack of color constancy in SEEMORE's feature channels. Note that while shape-related channels did not explicitly indicate the color of origin of their respective features, they depended on both color and intensity and were thus affected by the colorization manipulation as well. The clutter condition, which simulated the presence of a distractor object, dropped correct recognition rate to just below 80 percent. The additive noise condition disrupted recognition performance along both shape and color axes, cutting overall performance to 58 percent, indicative of poor high-frequency noise tolerance in the underlying feature channels.

3.3 Generalization Behavior and Recognition Errors. Numerical indexes of recognition performance are useful but do not explicitly convey the similarity structure of the underlying feature space. A more qualitative but extremely informative representation of system performance lies in the sequence of images in order of increasing distance from a test view. Records of this kind are shown in Figure 9 for several recognition trials. In each, a test view is shown at the far left highlighted in red (gray), and the ordered set of nearest neighbors is shown to the right. When a test view's nearest neighbor (second image from left) was not the correct match, the trial was classified as a recognition error in Table 1.

In Figure 9A, generalization behavior can be seen when only color channels were used for recognition, illustrating the limits of a simple "color histogramming" system. Errors of this kind occurred for objects with very similar color content and were caused by changes in lighting conditions that tipped the balance in favor of an incorrect object match.

In Figure 9B, generalization behavior can be seen when only shape-related channels were used for recognition. Thus, as shown in row 1, a view of a book is judged most similar to a series of other books (or the bottom of a rectangular cardboard box)—each a view of a rectangular object

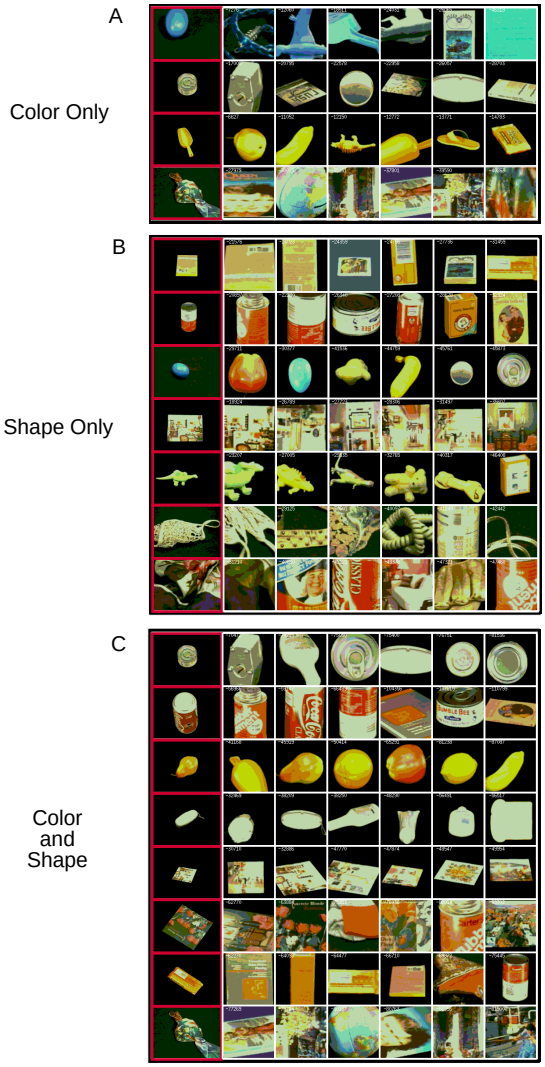

Figure 9: Generalization and errors. In each row, a novel test view is shown at the left (outlined in red). The sequence of best matching training views (one per object) is shown to the right, in order of decreasing similarity. (A) Examples of color generalization and errors (i.e., excluding shape-related features). (B) Examples of shape and texture generalization and errors, excluding color features. (C) Examples of generalization and errors using all 102 color- and shape-related channels.

with high-frequency surface markings. A similar sequence can be seen in subsequent rows for (2) a series of cans, each a right cylinder with detailed surface markings, (3) a series of smooth, not-quite-round objects, (4) a series of photographs of complex scenes, and (5) a series of dinosaurs (followed by a teddy bear). In certain cases, SEEMORE's shape-related similarity metric was more difficult to interpret visually or verbalize (last two rows), or was substantially different from that of a human observer.

When all 102 color and shape-related channels were used for recognition, the few remaining errors occurred among views of objects that shared common shape and color features (see Figure 9C). These errors plainly illustrate the limitations of SEEMORE's visual discrimination power, exemplified by chronic confusion among the Coke can, rubber cement can, and a Campbell's soup can. The confusion among these three similarly proportioned right cylinders with red and white surface markings accounted for 3 of the 20 errors for the intact test set. An additional 3 errors were explained by confusion between two similar photographs.

4 Discussion

4.1 Recognition Performance and Generalization Behavior. As can be seen in Table 1, color alone is a remarkably powerful cue for object recognition, even in the total absence of shape information. This result is consistent with the results of Swain and Ballard (1991), who successfully recognized objects using color histograms alone and first drew the tentative analogy between histogram bins and neural receptive fields. However, when the color signatures of objects mimic each other and lead to recognition errors, the errors are generally "inexcusable" to a human observer (see Figure 9A).

Thus, despite the utility of color for identification of specific instances of objects, the preeminence of shape information in object vision is clear, both for the definition of object categories (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976) and in the process of rapid object naming (Biederman & Ju, 1988). In this context, several aspects of SEEMORE's shape representations deserve closer examination.

First, the ecology of natural object vision gives rise to an apparent contradiction: Generalization in shape space must in some cases permit an object whose global shape has been grossly perturbed to be matched to itself, such as the various tangled forms of a telephone cord, but quasi-rigid basic-level shape categories (e.g., chair, shoe, tree) must be preserved as well and distinguished from each other.

A partial resolution to this conundrum lies in the observation that locally computed shape statistics are in large part preserved under the global shape deformations that nonrigid common objects (e.g., scarf, bike chain) typically undergo. A feature-space representation with an emphasis on locally derived shape channels will therefore exhibit a significant degree of invariance to global nonrigid shape deformations. The principal sources

of change to the local image-plane shape statistics under nonrigid shape deformations include (1) actual bending, rending, or warping of the object surfaces; (2) visual foreshortening as object surfaces change orientation in depth; (3) the addition and subtraction of object features that move in and out of self-occlusion; and (4) the addition and subtraction of spurious local shape statistics across accidental self-occluding boundaries. (Each of the latter three categories of feature changes applies to rigid objects as well.) The representational challenge, then, is to accumulate a sufficiently rich set of local shape statistics, which individually "tolerate" some degree of bending, warping, and foreshortening (categories 1 and 2) so that as a group they overwhelm in importance the feature-space excursions due to "capricious" feature changes (categories 3 and 4). The situation for rigid objects is somewhat less problematic, in that shape statistics are stable over longer spatial scales, and the problems associated with self-occlusion are generally less severe.

The definition of shape similarity embodied in the present approach is that two objects are similar if they contain similar profiles (histograms) of their shape measures, which emphasize locality. One way of understanding the emergence of global shape categories, then, such as "book," "can," and "dinosaur," is to view each as a set of instances of a single canonical object whose local shape statistics remain quasi-stable as it is warped into various global forms. In many cases, particularly within rigid object categories, exemplars may share longer-range shape statistics as well.

It is useful to consider one further aspect of SEEMORE's shape representation, pertaining to an apparent mismatch between the simplicity of the shape-related feature channels and the complexity of the shape categories that can emerge from them. Specifically, the order of binding of spatial relations within SEEMORE's shape channels is relatively low, consisting of single simple open or closed curves, or conjunctions of two oriented contours or Gabor patches. The fact that shape categories, such as "photographs of rooms" or "smooth, lumpy objects," cluster together in a feature space of such low binding order would therefore at first seem surprising. This phenomenon relates closely to the notion of "wickelfeatures" (Wickelgren, 1969; see also Rumelhart & McClelland, 1986, Chap. 18), where features that bind spatial information only locally—before being globally pooled—are nonetheless able to represent global patterns (words) with little or no residual ambiguity. For example, features sensitive to individual letters—but not their positions—are not sufficient to distinguish words such as *ngtcnrooiie* and *recognition*. In contrast, features sensitive to letter pairs, analogous to SEEMORE's shape features, are typically sufficient to eliminate representational ambiguity (i.e., multiple inverse images), even in the absence of relative positional information among the detected pairs. Thus, not only does no other English word contain the same set of adjacent letter pairs as *recognition (co, ec, gn, io, it, ni, og, on, re, ti)*, but no other possible word contains the same letter pairs (conjecture without proof!). One mechanism

underlying this phenomenon is that the overlapping subcomponents of the various features provide a form of implicit glue or binding that probabilistically constrains the spatial relations between features, even though these are not given explicitly within the representation. Thus the *e* in *ec* is likely to be the same *c* as in *re*, producing a high a posteriori probability for the compound *rec*. It is important to recall, however, that in direct opposition to the pressure for precise representation of shape is the pressure to generalize across different shapes within the same shape category or between views of the same object in different configurations. One strategy for coping with this trade-off may be to maintain separate representations at a range of specificities, coupled with a source of expertise to choose between them.

4.2 Rationale for Choice of Feature Channels. Following two main guidelines, SEEMORE's feature channels were designed as abstractions of known feature types seen in the ventral stream of the primate visual system and to represent information in images that is known to be important for object perception. The level of abstraction was roughly as follows: color-sensitive cells, texture-sensitive cells, cells responding to contour conjunctions, cells distinguishing curved from straight, and so forth. Within these general design guidelines, the specific selectivities and spatial invariances that defined SEEMORE's feature channels were largely shaped by two additional forces: (1) the contents of the object database, which implicitly rendered certain image cues more salient than others for purposes of classification, and (2) the relatively severe, monolithic nature of the recognition task, which entailed nearly complete viewpoint uncertainty from trial to trial.

Both of these influences make it likely that the details of SEEMORE's feature channels differ from those of any real animal, whose image statistics inevitably differ from those contained in SEEMORE's database, and for whom very different viewpoint assumptions, requiring different degrees of invariance, hold in different behavioral contexts. As such, SEEMORE's "neurophysiology" (feature channel responses) and "psychophysics" (recognition performance and patterns of generalization) do not yet provide a detailed or comprehensive model for the neurophysiological or psychophysical record for any particular animal species. Indeed, the fitting of existing empirical data along these lines has not been a goal of this work so far. On the other hand, SEEMORE's representations fall easily within the family of receptive-field-based representations seen in biological visual systems and, under many of the same challenges faced by "real" visual systems, have exhibited levels of recognition performance and patterns of generalization that seem surprising given the simplicity of the underlying computations. Given that the scale of visual hardware in the ventral stream of the primate visual system devoted to object recognition likely outstrips SEEMORE's 102 feature channels by multiple orders of magnitude, it seems a reasonable possibility that the gap between the performance of a visual "simpleton" like SEEMORE

and his much more gifted cousins in the animal kingdom is largely a matter of scale.

In contrast to the engineering-driven approach adopted here, features could have in principle been learned. Unsupervised learning approaches have been used to discover statistical structure in natural images that is strikingly close to the structure of simple cell-receptive fields in the mammalian visual system (Olshausen & Field, 1996). In the context of recognition, however, the problem of feature learning is more difficult in that statistical structure present in an ensemble of images may or may not be useful. Utility of image features depends on the specific classification problem that the animal faces, which can change from moment to moment; consider the processing of faces for identity versus age versus gender versus emotional state, and so forth. The discovery of useful structure is thus a problem requiring task-dependent supervision and requiring potentially large quantities of labeled data. If the features needed for recognition are highly descriptive, as needed in difficult recognition domains, then the search spaces for either unsupervised or supervised approaches to feature learning are huge. High representational biases could perhaps make such approaches more tractable.

4.3 Limitations and Extensions. The presegmentation of objects in the test sets used here is a simplifying assumption that is clearly invalid in the real world. The advantage of the assumption from a methodological perspective is that the object similarity structure induced by the feature dimensions can be studied independent of the problem of segmenting or indexing objects embedded in complex scenes. Thus, a representation that does not permit good discrimination among a large number of objects over a range of 3D viewpoints and internal configurations and does not impose appropriate category structure among objects need not be considered further as a component of a system that deals with objects embedded within scenes.

The cluttered test condition (see Figure 3C) was designed to assay explicitly the sensitivity of SEEMORE's visual representations to additive pattern noise. While the figures in Table 1 show a seven-fold increase in error rate for cluttered test images relative to the intact condition (21 percent versus 3 percent), these figures must be interpreted with care. First, as for the case of intact images, approximately half (44 percent) of the "errors" in the cluttered case were second-best matches from a database of 100 objects and so do not represent total recognition failure. Moreover, the act of cluttering (or occluding or color shifting) a test image can introduce legitimate accidental similarities between the modified test view and a view of another (i.e., "incorrect") object. Since SEEMORE is entirely unsuspecting of these manipulations and is asked simply to compare incoming views to familiar views at face value, some of the committed errors may be excusable, to the

extent that the degraded object view does look most similar to a view of an incorrect object to the eye of a human observer.

When pattern clutter is made more severe than that illustrated in Figure 3C, however, recognition performance suffers dramatically. One of the main reasons for this is that SEEMORE's feature channels are currently far from sparse in their activity profiles across images; most feature channels are activated to some degree by most object views. Under these circumstances, the representations of multiple objects in the visual field additively collide within the same feature channels, making separate recognition difficult or impossible.

Two kinds of remedies for this problem have been suggested by biology and prior work. The first is crude image segmentation based on a movable fovea or other focal attention system (Koch & Ullman, 1985; Niebur & Koch, 1994). This type of strategy typically minimizes, but cannot eliminate, the additive visual noise induced by extraneous objects outside the focus of attention. The second, more potent remedy is the leap to sparse, very-high-dimensional space, whose mathematical advantages for vision and recognition have been discussed at length elsewhere (Kanerva, 1988; Califano & Mohan, 1994). For example, in a domain of 2D line drawings, Califano and Mohan (1994) have demonstrated position, orientation, and scale-invariant recognition in multiple-object scenes with no prior segmentation, using on the order of 10^6 highly descriptive contour-based features (based on conjunctions of triplets of contour segments). Intuitively, an object is recognized when a sufficient number of sufficiently distinctive features "vote" for the presence of a particular object, regardless of whether other objects also get votes. The leap into high-dimensional feature space is easily arranged combinatorially, by conjoining existing (or other) low-order features into compound features of higher order. The issue of feature locality remains crucial, however. If the desired combinatorial explosion is achieved by conjoining elemental shape tokens over relatively long distances in the image, the resulting shape-based similarity metric can permit extremely fine shape distinctions but may lead to poor generalization under nonrigid shape deformation and partial occlusion. This approach is most relevant in feature-poor images, in which the variety of local token configurations is limited. In feature-dense images, such as natural images, the combinatorial explosion can be achieved while maintaining locality within elemental features configurations, thus preserving the constellation of desirable representational properties that locality confers.

One means of effectively increasing the richness of localized feature configurations is to include elemental filters simultaneously sensitive to multiple visual cues, such as contour, color, shading, and texture cues. Interestingly, the preferred responses of pattern-selective cells in the primate IT cortex frequently involve explicit binding of contour shapes with particular colors and/or textures (Kobatake & Tanaka, 1994). In this context, we may interpret these data as evidence of an implicit neural strategy whose

goal is to achieve sparse high dimensionality in order to simplify object indexing in cluttered scenes, while retaining locality, in order to facilitate viewpoint-, configuration-, and occlusion-insensitive shape-based generalization. In continuing work, we are pursuing this course.

Acknowledgments

Thanks to József Fiser for useful discussions and for development of the Gabor-based channel set, to Dan Lipofsky and Scott Dewinter for helping in the construction of the image database, and to Christof Koch for providing support at Cal Tech where this work was initiated. This work was funded by the Office of Naval Research, and the McDonnell-Pew Foundation.

References

Amit, Y., Geman, D., & Wilder. K. (1995). *Recognizing shapes from simple queries about geometry* (Tech. Rep.). Chicago: University of Chicago, Department of Statistics.

Biederman, I., & Ju, G. (1988). Surface vs. edge-based determinants of visual recognition. *Cognitive Psychology, 20*, 38–64.

Califano, A., & Mohan, R. (1994). Multidimensional indexing for recognizing visual shapes. *IEEE Trans. on PAMI, 16*, 373–392.

Desimone, R., Albright, T. D., Gross, C., & Bruce, C. (1984). Stimulus selective properties of interior temporal neurons in the macaque. *J. Neurosci., 4*, 2051–2062.

DiGirolamo, G., & Kanwisher, N. (1995). *Accessing stored representations begins within 155 ms in object recognition*. Paper presented at the 36th Annual Meeting of the Psychonomics Society, Los Angeles.

Duda, R. O., & Hart, P. (1973). *Pattern classification and scene analysis*. New York: Wiley.

Edelman, S., & Bulthoff, H. (1992). Orientation dependence in the recognition of familiar and novel views of three-dimensional objects. *Vision Res., 32*, 2385–2400.

Friedman, J. H. (1994). *Flexible metric nearest neighbor classification* (Tech. Rep.). Stanford University, Department of Statistics and Stanford Linear Accelerator Center.

Fukunaga, K. (1990). *Introduction to statistical pattern recognition*. New York: Academic Press.

Fukushima, K., Miyake, S., & Ito, T. (1983). Neocognitron: A neural network model for a mechanism of visual pattern recognition. *IEEE Trans. Sys. Man and Cybernetics, SMC-13*, 826–834.

Gochin, P. M., Colombo, M., Dorfman, G. A., Gerstein, G., & Gross, C. (1994). Neural ensemble coding in inferior temporal cortex. *J. Neurophysiol., 71*, 2325–2337.

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

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

Kanerva, P. (1988). *Sparse distributed memory*. Cambridge, MA: MIT Press.

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

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiol., 4*, 219–227.

Lades, M., Vorbruggen, J., Buhmann, J., Lange, J., Malsburg, C., Wurtz, R., & Konen, W. (1993). Distortion invariant object recognition in the dynamic link architecture. *IEEE Trans. Computers, 42*, 300–311.

Le Cun, Y., Matan, O., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., Jackel, L., & Baird, H. (1990). Handwritten zip code recognition with multilayer networks. In *Proc. of the 10th Int. Conf. on Patt. Rec.* Los Alamitos, CA: IEEE Computer Science Press.

Logothetis, N., Pauls, J., Bulthoff, H., & Poggio, T. (1994). View-dependent object recognition by monkeys. *Current Biology, 4*, 401–414.

Mel, B. (1996). Seemore: A view-based approach to 3-D object recognition using multiple visual cues. In D. S. Touretzky, M. C. Mozer, & M. E. Hasselmo (Eds.), *Advances in Neural Information Processing Systems* (Vol. 8, pp. 865–871). Cambridge, MA: MIT Press.

Mishkin, M. (1982). A memory system in monkey. *Philos. Trans. R. Soc. Lond. B., 298*, 85–95.

Murase, H., & Nayar, S. (1995). Visual learning and recognition of 3-D objects from appearance. *International Journal of Computer Vision, 14*, 5–24.

Niebur, E., & Koch, C. (1994). A model for the neuronal implementation of selective visual attention based on temporal correlation among neurons. *Journal of Computational Neuroscience, 1*(1), 141–158.

Olshausen, B., & Field, D. (1996). Emergence of simple-cell receptive-field properties by learning a sparse code for natural images. *Nature, 381*, 607–609.

Oram, M., Perrett, D. (1992). Time course of neural responses discriminating different views of the face and head. *J. Neurophysiol., 68*(1), 70–84.

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

Pitts, W., & McCullough, W. (1947). How we know universals: the perception of auditory and visual forms. *Bull. Math. Biophys., 9*, 127–147.

Potter, M. (1976). Short-term conceptual memory for pictures. *J. Exp. Psychol.: Human Learning and Memory, 2*, 509–522.

Rao, R., & Ballard, D. (1995). An active vision architecture based on iconic representations. *Artificial Intelligence, 78*, 461–505.

Rosch, E., Mervis, C., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8*, 382–439.

Rumelhart, D., & McClelland, J. (1986). *Parallel distributed processing.* Cambridge, MA: MIT Press.

Schiele, B., & Crowley, J. (1996). Probabilistic object recognition using multidimensional receptive field histograms. In *Proc. of the 13th Int. Conf. on Patt. Rec.* (Vol. 2, pp. 50–54). Los Alamitos, CA: IEEE Computer Society Press.

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

Subramaniam, S., Biederman, I., Kalocsai, P., & Madigan, S. R. (1995). Accurate identification, but chance forced-choice recognition for rsvp pictures. Poster presented at the Annual Meeting of the Association for Research in Vision and Ophthalmology, Ft. Lauderdale, FL.

Swain, M. J., & Ballard, D. (1991). Color indexing. *Int. J. Computer Vision, 7*, 11–32.

Viola, P. (1993). Feature-based recognition of objects. In *Proc. of the AAAI Fall Symposium on Learning and Computer Vision*, 60–65.

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

Received April 29, 1996; accepted September 19, 1996.