Towards a Computational Model for Object Recognition in IT Cortex

David G. Lowe
Computer Science Department
University of British Columbia
lowe@cs.ubc.ca

Abstract

There is considerable evidence that object recognition in primates is based on the detection of local image features of intermediate complexity that are largely invariant to imaging transformations. A computer vision system has been developed that performs object recognition using features with similar properties. Invariance to image translation, scale and rotation is achieved by first selecting stable key points in scale space and performing feature detection only at these locations. The features measure local image gradients in a manner modeled on the response of complex cells in primary visual cortex, and thereby obtain partial invariance to illumination, affine change, and other local distortions. The features are used as input to a nearest-neighbor indexing method and Hough transform that identify candidate object matches. Final verification of each match is achieved by finding a best-fit solution for the unknown model parameters and integrating the features consistent with these parameter values. This verification procedure provides a model for the serial process of attention in human vision that integrates features belonging to a single object. Experimental results show that this approach can achieve rapid and robust object recognition in cluttered partially-occluded images.

1. Introduction

During the past decade, there have been major advances in our understanding of how object recognition is performed in the primate visual system. There is now a broad body of evidence [21] showing that object recognition makes use of neurons in inferior temporal cortex (IT) that respond to features of intermediate complexity. These features are typically invariant to a wide range of changes in location, scale, and illumination, while being very sensitive to particular combinations of local shape, color, and texture properties.

This paper describes a computer vision system for performing object recognition that also makes use of local image features of intermediate complexity that are invariant to many imaging parameters. The approach is called a scale invariant feature transform (SIFT), as it transforms an image into a representation that is unaffected by image scaling and other similarity transforms. Features are located at peaks of a function computed in scale space. The features

First IEEE International Workshop on Biologically Motivated Computer Vision, Seoul, Korea (May 2000).

describe local image regions around these peaks using a representation that emphasizes gradient orientations while allowing small shifts in gradient position, in a manner modeled on the responses of complex cells in primary visual cortex. These features are bound to object interpretations through a process of indexing followed by a best-fit solution for object parameters. This achieves feature integration in a manner similar to the process of serial visual attention that has been shown to play an important role in object recognition in human vision [23]. The result is a system that is able to recognize 3D objects from any viewpoint under varying illumination in cluttered natural images with about 1 second of computation time.

There has been some previous research [14, 24] on building computer systems for object recognition that use similar intermediate features to those in IT cortex. One problem has been that these earlier systems use correlation to measure the presence of intermediate features in an image, which has a prohibitive computational cost due to the need to compare each feature at every location, scale, and orientation to the image. This paper describes a staged filtering approach in which stable key points in scale space are first identified, and then feature detection is performed only at these canonical points and with respect to a canonical scale and orientation. This reduces the cost by about a factor of 10,000, making it easily suitable for practical applications. Furthermore, the feature descriptors are modified to make them less sensitive to affine transformation, illumination change and local distortions as compared to image correlation. A final stage of solving for explicit model pose parameters allows for robust verification of object interpretations.

2. Related research

An understanding of the intermediate-level features in visual cortex was first obtained in a novel approach developed by Tanaka and his associates [7, 20]. They recorded from individual neurons in anesthetized monkeys while using a library of objects and a computer graphics editing system to characterize their responses. First, a collection of complex real-world objects were tested to obtain an initial response. An image of the best initial object was then subject to numerous attempts at simplification and modification to obtain an optimal response. Although some neurons in anterior IT cortex responded to very simple line or bar features, in most cases the optimal response was obtained by features of intermediate complexity, such as a dark five-sided star shape, a circle with a thin protruding element at a particular orientation, or a green horizontal textured region within a triangular boundary. Some neurons responded only to more complex shapes, such as moderately detailed face or hand images.

These intermediate-complexity neurons were often highly sensitive to small variations in shape, such as the degree of rounding of corners or relative lengths of elements. On the other hand, the neurons exhibited a wide range of invariance to other parameters, such as retinal location, size, and contrast. Detailed studies of size and position invariance [6, 16] have shown that about 55% of the neurons have a size invariance range of greater than 2 octaves and 20% have a range of more than 4 octaves. Most neurons have a receptive field covering a large portion of the image (an average 25 degrees of visual angle and usually including all of the fovea). These properties would seem ideally suited to determining object identity without needing to replicate the neuron for each combination of values of the imaging parameters.

Neurons that were close together in cortex often responded to small variations of the same feature. Based on the average size of these related feature columns as a proportion of the total

size of the brain region, Tanaka [20] estimated that there was room for about 1300 such feature columns. However, if the column sizes vary, then the large ones would be preferentially sampled and the total number of feature types could be far greater.

The feature responses have been shown to depend on previous visual learning from exposure to specific objects containing the features. Logothetis, Pauls & Poggio [10] examined the responses of neurons in monkeys that had been trained to classify views of wire-frame and spheroidal shapes. They discovered many neurons that responded only to particular views of these shapes, while exhibiting the usual invariance to large ranges of scale and location. Booth & Rolls [3] found similar results for 10 plastic objects that had been placed in the monkey's cage for a period of weeks or months without any training. In addition to the usual view-dependent neurons, they found a small population of neurons that responded to any view of a particular object (these responded to a conjunction of shape views, rather than to a simple feature such as color that was shared between views). In a dramatic illustration of learning, Tovee, Rolls & Ramachandran [22] showed that a face sensitive neuron could learn to recognize degraded face images (that were previously unrecognizable) by exposure to 5 seconds of training images that showed the transition between normal and degraded images.

It is also known that object recognition in human vision uses a serial process of attention to bind features to object interpretations, determine pose, and segment an object from a cluttered background [23]. A wide range of psychophysical experiments [25] have shown that preattentive object descriptions consist of only a collection of isolated features, and serial attention is necessary to represent shape relationships and integrate features into a common object description. In this paper we will describe the use of a Hough transform to generate object hypotheses, followed by best-fit parameter solving and selection of consistent features to perform binding and model verification.

Within the computer vision field, there has been recent work on using dense collections of local image features for object recognition. One approach has been to use a corner detector (more accurately, a detector of peaks in local image variation) to identify repeatable image locations, around which local image properties can be measured. Schmid & Mohr [18] used the Harris corner detector to identify interest points, and then created a local image descriptor at each interest point from an orientation-invariant vector of derivative-of-Gaussian image measurements. These image descriptors were used for robust object recognition by looking for multiple matching descriptors that satisfied object-based orientation and location constraints. This work was impressive both for the speed of recognition in a large database and the ability to handle cluttered images. However, the corner detector examines an image at only a single scale. As the change in scale becomes significant, the detector responds to different image points. Also, since the detector does not provide an indication of the object scale, it is necessary to create image descriptors and attempt matching at a large number of scales.

Other approaches in computer vision to appearance-based recognition include eigenspace matching [15], color histograms [19], and receptive field histograms [17]. These approaches have all been demonstrated successfully on isolated objects or pre-segmented images, but due to their more global features it has been difficult to extend them to cluttered and partially occluded images.

3. Key localization

Rather than searching all possible image locations for particular features, we obtain far better efficiency by first selecting key locations and scales of interest and describing the local image region around each location. The key locations are selected in a manner that is invariant with respect to image translation, scaling, and rotation, and is minimally affected by noise and small distortions.

Lindeberg [8] has shown that under some rather general assumptions on scale invariance, the Gaussian kernel and its derivatives are the only possible smoothing kernels for scale space analysis. To achieve rotation invariance and a high level of efficiency, we have chosen to select key locations at maxima and minima of a difference of Gaussian function applied in scale space. This can be computed very efficiently by building an image pyramid with resampling between each level. Furthermore, it locates key points at regions and scales of high variation, making these locations particularly useful for characterizing the image. Crowley & Parker [4] and Lindeberg [9] have previously used the difference-of-Gaussian in scale space for other purposes.

As the 2D Gaussian function is separable, its convolution with the input image can be efficiently computed by applying two passes of the 1D Gaussian function in the horizontal and vertical directions:

$$g(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2}$$

For key localization, all smoothing operations are done using $\sigma = \sqrt{2}$, which can be approximated with sufficient accuracy using a 1D kernel with 7 sample points.

A pyramid of Gaussian images is computed at all scales differing by factors of 1.5, using smoothing followed by resampling. Images at adjacent scales are subtracted prior to resampling to obtain the difference of Gaussian function. See [13] for details on how this can computed very efficiently. Maxima and minima of this scale-space function are determined by comparing each pixel in the pyramid to its immediate neighbours in location and scale.

4. Key orientation and stability

To characterize the image at each key location, the smoothed image at each level of the pyramid is processed to extract image gradients and orientations. At each pixel, A_{ij}, the image gradient magnitude, M_{ij}, and orientation, R_{ij}, are computed using pixel differences:

$$M_{ij} = \sqrt{(A_{ij} - A_{i+1,j})^2 + (A_{ij} - A_{i,j+1})^2}$$

$$R_{ij} = \mathrm{atan2}\left(A_{ij} - A_{i+1,j}, A_{i,j+1} - A_{ij}\right)$$

The pixel differences are efficient to compute and provide sufficient accuracy due to the substantial level of previous smoothing.

Robustness to illumination change is enhanced by thresholding the gradient magnitudes at a value of 0.1 times the maximum possible gradient value. This reduces the effect of a change in illumination direction for a surface with 3D relief, as an illumination change may result in large changes to gradient magnitude but will probably have less influence on gradient orientation.

Each key location is assigned a canonical orientation so that the image descriptors are invariant to rotation. In order to make this as stable as possible against lighting or contrast changes,

the orientation is determined by the peak in a histogram of local image gradient orientations. The orientation histogram is created using a Gaussian-weighted window with σ of 3 times that of the current smoothing scale. These weights are multiplied by the thresholded gradient values and accumulated in the histogram at locations corresponding to the orientation, R_{ij}.

The stability of the resulting keys can be tested by subjecting natural images to affine projection, contrast and brightness changes, and addition of noise. The location of each key detected in the first image can be predicted in the transformed image from knowledge of the transform parameters. Testing on a collection of natural images [13] shows that about 70% of the key locations will be stable to these changes and will be detected at the correct corresponding location, orientation, and scale.

5. Local image description

Given a stable location, scale, and orientation for each key, it is now possible to describe the local image region in a manner invariant to these transformations. In addition, it is desirable to make this representation robust against small shifts in local geometry, such as arise from affine or 3D projection. One approach to this is suggested by the response properties of complex neurons in the visual cortex, in which a feature position is allowed to vary over a small region while orientation and spatial frequency specificity are maintained. Edelman, Intrator & Poggio [5] have performed experiments that simulated the responses of complex neurons to different 3D views of computer graphic models, and found that the complex cell outputs provided much better discrimination than simple correlation-based matching. This can be seen, for example, if an affine projection stretches an image in one direction relative to another, which changes the relative locations of gradient features while having a smaller effect on their orientations and spatial frequencies.

This robustness to local geometric distortion can be obtained by representing the local image region with multiple images representing each of a number of orientations (referred to as orientation planes). Each orientation plane contains only the gradients corresponding to that orientation, with linear interpolation used for intermediate orientations. Each orientation plane is blurred and resampled at fewer locations to allow for larger shifts in positions of the gradients.

This approach can be efficiently implemented by using the same precomputed gradients and orientations for each level of the pyramid that were used for orientation selection. For each keypoint, we use the pixel sampling from the pyramid level at which the key was detected (to maintain scale invariance). The pixels that fall in a circle of radius 8 pixels around the key location are inserted into the orientation planes. The orientations are measured relative to that of the key by subtracting the key's orientation. For our experiments we used 8 orientation planes, each sampled over a 4×4 grid of locations, with a sample spacing 4 times that of the pixel spacing for that pyramid level. The blurring is achieved by allocating the gradient of each pixel among its 8 closest neighbors in the sample grid, using linear interpolation in orientation and the two spatial dimensions. This implementation is much more efficient than performing explicit blurring and resampling, yet gives almost equivalent results.

In order to sample the image at a larger scale, the same process is repeated for a second level of the pyramid one octave higher. However, this time a 2×2 rather than a 4×4 sample region is used. This means that approximately the same image region will be examined at both scales, so that any nearby occlusions will not affect one scale more than the other. Therefore, the total

Figure 1: Top row shows model images for 3D objects with outlines found by background segmentation. Images below show the recognition results for these objects with superimposed model outlines and image key locations used for matching.

number of samples in the SIFT key vector, from both scales, is $8 \times 4 \times 4 + 8 \times 2 \times 2$ or 160 elements, giving enough measurements for high specificity.

6. Indexing and matching

For indexing, we need to store the SIFT keys for sample images and then identify matching keys from new images. The problem of identifying the most similar keys for high dimensional vectors is known to have high complexity if an exact solution is required. However, a modification of the k-d tree algorithm called the best-bin-first search method (Beis & Lowe [2]) can identify the nearest neighbors with high probability using only a limited amount of computation.

An efficient way to cluster reliable model hypotheses is to use the Hough transform [1] to search for keys that agree upon a particular model pose. Each model key in the database contains a record of the key's parameters relative to the model coordinate system. Therefore, we can create an entry in a hash table predicting the model location, orientation, and scale from the match hypothesis. We use a bin size of 30 degrees for orientation, a factor of 2 for scale, and 0.25 times the maximum model dimension for location. These rather broad bin sizes allow for

clustering even in the presence of substantial geometric distortion, such as due to a change in 3D viewpoint. To avoid the problem of boundary effects in hashing, a hypothesis that is close to a bin boundary is hashed into bins on both sides of the boundary for all dimensions (analogous to a feature neuron activating multiple object-selective neurons with overlapping receptive fields).

The hash table is searched to identify all clusters of at least 3 entries in a bin, and the bins are sorted into decreasing order of size. Each such cluster is then subject to a verification procedure in which a least-squares solution is performed for the affine projection parameters relating the model to the image [12, 13].

Outliers can now be removed by checking for agreement between each image feature and the model, given the parameter solution. Each match must agree within 15 degrees orientation, $\sqrt{2}$ change in scale, and 0.2 times maximum model size in terms of location. If fewer than 3 points remain after discarding outliers, then the match is rejected. If any outliers are discarded, the least-squares solution is re-solved with the remaining points.

Knowledge of the model pose allows us to perform top-down matching, in which all image keys within the model region are checked for whether they are consistent with the model pose. Features are often added at this stage that were not in the original hash bin due to errors in predicting the model parameters from a single feature or due to previous ambiguity in the object match for a feature.

7. Experiments

The top row of Figure 1 shows three model images of typical objects that were used to test recognition. The models were photographed on a black background, and object outlines were extracted by segmenting out the background region (the outlines are used only for display purposes and play no role in recognition). Examples of recognition in occluded, cluttered images are shown below the model images. The SIFT key locations used for recognition are shown superimposed on the test images. The object outlines are also projected onto the image using the best-fit affine parameter solution. Since only 3 keys are needed for robust recognition, it can be seen that the solutions are often highly redundant and can survive substantial occlusion.

Although the model images and affine parameters do not account for rotation in depth of 3D objects, they are still sufficient to perform robust recognition of 3D objects over about a 20 degree range of rotation in depth away from each model view. The images in these examples are of size 384×512 pixels. The computation times for recognition of all objects in each image are about 1.5 seconds on a Sun Sparc 10 processor, with about 0.9 seconds required to build the scale-space pyramid and identify the SIFT keys, and about 0.6 seconds to perform indexing and least-squares verification. This does not include time to pre-process each model image, which would be about 1 second per image, but would only need to be done once for initial entry into a model database.

These first examples used only single training images and tested recognition for views within about 20 degrees of the training view. We have recently developed an approach to integrating images from many different viewpoints. Large changes in pose are handled by noticing when the least-squares solution has a high residual (above 0.1 of model image size) and creating a new model view. Points that match across views are linked so that indexing can consider both possible model views for images near the boundaries. For small changes in view (with a low least-squares residual) the key points from all images are combined into a common model view.

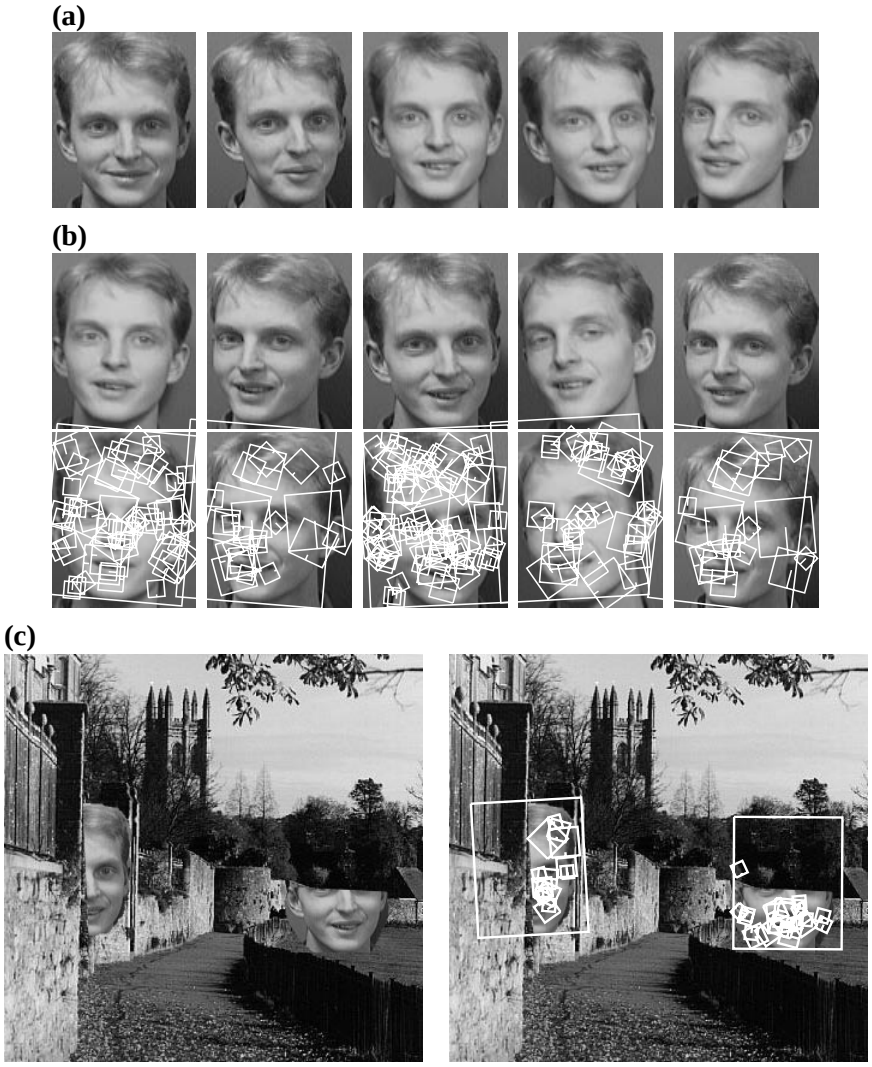

Figure 2: (a) Five training images of a face were used to build a model. (b) The model was matched to five test images, as shown with the overlay of matched keys below each image. (c) Parts of a test image could be recognized following insertion into a cluttered scene. (Face images are courtesy of AT&T Laboratories Cambridge).

This allows a single model to incorporate features from training views that have different illumination or changes in model shape, such as differing expressions on a face.

Figure 2 shows this recognition approach being applied to face images. Five training images shown in Figure 2(a) were used to build a model of the face. These were easily matched to one another in spite of variations in pose and illumination, and the features combined into a common model. This model could then be matched to subsequent images, such as the five test images in Figure 2(b). Under each image are shown the keys that were used to match the model to the image, as well as an outer rectangle showing the least-squares solution for the model affine fit. The robustness of the approach is illustrated by taking portions of a test image and inserting them into a natural cluttered image, as shown in Figure 2(c). These face parts are

easily identified, as shown by the superimposed keys and solution for the model frame. In fact, the face portion could be considerably smaller and still produce at least 3 key matches. While these experiments are still at a preliminary stage, they show promise for face recognition from arbitrary pose in cluttered scenes.

8. Conclusions and discussion

We have described an efficient method for the detection of dense local features of intermediate complexity. These SIFT features are sensitive to local image shape properties while being partially invariant to many common imaging transformations. Object matching is performed by identifying clusters of features that agree on a model interpretation, solving for best-fit model parameters, and integrating features that agree with this interpretation. This approach is broadly similar in its use of features and their integration to what is known of object recognition in visual cortex.

One topic that we have not considered is the discrimination and categorization of similar instances within an object class (for example, identifying a particular person or expression from matching a face image). This could probably be performed using the same set of SIFT features by maintaining the correlations between each feature type and the categorizations of interest, but this approach remains to be developed and tested. In IT cortex, neurons are organized in columns that respond to close variations on each feature [20], and these feature columns likely play the role of assisting fine discriminations between similar shapes.

Human vision is sometimes able to identify clear, isolated objects so rapidly that it seems likely that recognition is performed in a purely bottom-up manner [23]. We might model this as consisting of just the Hough transform stage that generates object hypotheses as a conjunction of features without detailed verification of consistent model parameters. However, human object recognition in cluttered scenes appears to require a process of serial attention to one object at a time [25] (a familiar scene can itself be recognized as a single object). This appears to involve the determination of object pose and other parameters, as well as selection and integration of features consistent with these parameters. This process appears to be consistent with the verification component described above, including best-fit parameter solving, outlier detection, and integration of new consistent features that were not sufficiently distinctive to contribute to the initial object hypothesis.

One clear difference to IT cortex is that the SIFT features are invariant to image rotation, while neurons in IT cortex are usually not. However, the rotation invariance brings significant improvements in efficiency and is easily accounted for by the subsequent indexing and least-squares verification. Less is known about how orientation constraints are enforced in primate vision.

Another area for further development is to add new SIFT feature types to incorporate color, texture, and edge groupings, all of which play an important role in primate vision. Scale-invariant edge groupings that make local figure-ground discriminations would be particularly useful at object boundaries where background clutter can interfere with other features. The indexing and verification framework allows for all types of scale and rotation invariant features to be incorporated into a single model representation.

References

[1] Ballard, D.H., "Generalizing the Hough transform to detect arbitrary patterns," *Pattern Recognition,* **13,** 2 (1981), pp. 111-122.

[2] Beis, Jeff, and David G. Lowe, "Shape indexing using approximate nearest-neighbour search in high-dimensional spaces," *Conference on Computer Vision and Pattern Recognition,* Puerto Rico (1997), pp. 1000–1006.

[3] Booth, Michael C.A., and Edmund T. Rolls, "View-invariant representations of familiar objects by neurons in the inferior temporal cortex," *Cerebral Cortex,* **8** (1998), pp. 510–523.

[4] Crowley, James L., and Alice C. Parker, "A representation for shape based on peaks and ridges in the difference of low-pass transform," *IEEE Trans. on Pattern Analysis and Machine Intelligence,* **6,** 2 (1984), pp. 156–170.

[5] Edelman, Shimon, Nathan Intrator, and Tomaso Poggio, "Complex cells and object recognition," Unpublished Manuscript, preprint at `http://kybele.psych.cornell.edu/~edelman/abstracts.html#ccells`

[6] Ito, Minami, Hiroshi Tamura, Ichiro Fujita, and Keiji Tanaka, "Size and position invariance of neuronal responses in monkey inferotemporal cortex," *Journal of Neurophysiology,* **73,** 1 (1995), pp. 218–226.

[7] Kobatake, Eucaly, and Keiji Tanaka, "Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex," *Journal of Neurophysiology,* **71,** 3 (1994), pp. 856–867.

[8] Lindeberg, Tony, "Scale-space theory: A basic tool for analysing structures at different scales", *Journal of Applied Statistics,* **21,** 2 (1994), pp. 224–270.

[9] Lindeberg, Tony, "Detecting salient blob-like image structures and their scales with a scale-space primal sketch: a method for focus-of-attention," *International Journal of Computer Vision,* **11,** 3 (1993), pp. 283–318.

[10] Logothetis, Nikos K., Jon Pauls, and Tomaso Poggio, "Shape representation in the inferior temporal cortex of monkeys," *Current Biology,* **5,** 5 (1995), pp. 552–563.

[11] Lowe, David G., "Three-dimensional object recognition from single two-dimensional images," *Artificial Intelligence,* **31,** 3 (1987), pp. 355–395.

[12] Lowe, David G., "Fitting parameterized three-dimensional models to images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **13,** 5 (1991), pp. 441–450.

[13] Lowe, David G., "Object recognition from local scale-invariant features," *International Conference on Computer Vision,* Corfu, Greece (September 1999), pp. 1150–1157.

[14] Mel, Bartlett W., "SEEMORE: Combining color, shape, and texture histogramming in a neurally-inspired approach to visual object recognition," *Neural Computation,* **9,** 4 (1997), pp. 777-804.

[15] Murase, Hiroshi, and Shree K. Nayar, "Visual learning and recognition of 3-D objects from appearance," *International Journal of Computer Vision,* **14,** 1 (1995), pp. 5–24.

[16] Perrett, David I., and Mike W. Oram, "Visual recognition based on temporal cortex cells: viewer-centered processing of pattern configuration," *Zeitschrift für Naturforschung C,* **53c** (1998), pp. 518–541.

[17] Schiele, Bernt, and James L. Crowley, "Recognition without correspondence using multidimensional receptive field histograms," *International Journal of Computer Vision,* **36,** 1 (2000), pp. 31–50.

[18] Schmid, C., and R. Mohr, "Local grayvalue invariants for image retrieval," *IEEE PAMI,* **19,** 5 (1997), pp. 530–534.

[19] Swain, M., and D. Ballard, "Color indexing," *International Journal of Computer Vision,* **7,** 1 (1991), pp. 11–32.

[20] Tanaka, Keiji, "Neuronal mechanisms of object recognition," *Science,* **262** (1993), pp. 685–688.

[21] Tanaka, Keiji, "Mechanisms of visual object recognition: monkey and human studies," *Current Opinion in Neurobiology,* **7** (1997), pp. 523–529.

[22] Tovee, Martin J., Edmund T. Rolls, and V.S. Ramachandran, "Rapid visual learning in neurones of the primate temporal visual cortex," *NeuroReport,* **7** (1996), pp. 2757–2760.

[23] Treisman, Anne M., and Nancy G. Kanwisher, "Perceiving visually presented objects: recognition, awareness, and modularity," *Current Opinion in Neurobiology,* **8** (1998), pp. 218–226.

[24] Viola, Paul, "Complex feature recognition: A Bayesian approach for learning to recognize objects," *MIT AI Memo 1591,* Massachusetts Institute of Technology (1996).

[25] Wolfe, Jeremy M., and Sara C. Bennett, "Preattentive object files: shapeless bundles of basic features," *Vision Research,* **37**, 1 (1997), pp. 25–43.