DAVID A. FORSYTH, JITENDRA MALIK, THOMAS K. LEUNG, CHRIS BREGLER, CHAD CARSON, HAYIT GREENSPAN, & MARGARET M. FLECK

Finding Pictures of Objects in Large Collections of Images

Retrieving images from very large collections using image content as a key is becoming an important problem. Users prefer to ask for pictures using notions of content that are strongly oriented to the presence of objects, which are quite abstractly defined. Computer programs that implement these queries automatically are desirable but are hard to build because conventional object recognition techniques from computer vision cannot recognize very general objects in very general contexts.

This paper describes an approach to object recognition structured around a sequence of increasingly specialized grouping activities that assemble coherent regions of image that can be shown to satisfy increasingly stringent constraints. The constraints that are satisfied provide a form of object classification in quite general contexts.

This view of recognition is distinguished by far richer involvement of early visual primitives, including color and texture; the ability to deal with rather general objects in uncontrolled configurations and contexts; and a satisfactory notion of classification. These properties are illustrated with three case studies: one demonstrates the use of descriptions that fuse color and spatial properties; one shows how trees can be described by fusing texture and geometric properties; and one shows how this view of recognition yields a program that can tell, quite accurately, whether a picture contains naked people or not.

INTRODUCTION

Very large collections of images are becoming common, and users have a clear preference for accessing images in these databases based on the objects that are present in them. Creating indexes for these collections by hand is unlikely to be successful because these databases can be gigantic. Furthermore, it can be very difficult to impose order on these collections. For example, the California Department of Water Resources' collection contains approximately half-a-million images. Another example is the collection of images available on the Internet, which is notoriously large and disorderly. This lack of structure makes it hard to rely on textual

annotations in indexing. More practical alternatives are computer programs that could automatically assess image content (Sclaroff, 1995).

Another reason that manual indexing is difficult is that it can be hard to predict later content queries—for example, local political figures may reach national importance long after an image has been indexed. In a very large collection, the subsequent reindexing process becomes onerous.

Classic object recognition techniques from computer vision cannot help with this problem. Recent techniques can identify specific objects drawn from a small (on the order of 100 items) collection, but no present technique is effective at distinguishing, for example, people from cows, a problem usually known as classification. This discussion presents case studies illustrating an approach to determine image content that is capable of object classification. The approach is based on constructing rich image descriptions that fuse color, texture, and shape information to determine the identity of objects in the image.

MATERIALS AND OBJECTS—"STUFF" VERSUS "THINGS"

Many notions of image content have been used to organize collections of images (e.g., see Layne, 1994). Relevant here are notions centered on objects; the distinction between materials—"stuff"—and objects—"things"—is particularly important. A material (e.g., skin) is defined by a homogeneous or repetitive pattern of fine-scale properties but has no specific or distinctive spatial extent or shape. An object (e.g., a ring) has a specific size and shape. This distinction (in computer vision, Ted Adelson has emphasized the role of filtering techniques in early vision for measuring stuff properties) and a similar distinction for actions is well-known in linguistics and philosophy (dating back at least to Whorf [1941]) where they are used to predict differences in the behavior of nouns and verbs (e.g., Taylor, 1977; Tenney, 1987; Fleck, 1996).

To a first approximation, 3D materials appear as distinctive colors and textures in 2D images, whereas objects appear as regions with distinctive shapes. Therefore, one might attempt (following, for example, Adelson) to identify materials using low-level image properties and identify objects by analyzing the shape of, and the relationships between, 2D regions. Indeed, materials with particularly distinctive color or texture (e.g., sky) can be successfully recognized with little or no shape analysis, and objects with particularly distinctive shapes (e.g., telephones) can be recognized using only shape information.

In general, however, too much information is lost in the projection onto the 2D image for strategies that ignore useful information to be successful. The typical material, and so the typical color and texture, of an object is often helpful in separating the object from other image regions

and in recognizing it. Equally, the shapes into which it is typically formed can be useful cues in recognizing a material. For example, a number of other materials have the same color and texture as human skin at typical image resolutions. Distinguishing these materials from skin requires using the fact that human skin typically occurs in human form.

OBJECT RECOGNITION

Current object recognition systems represent models either as a collection of geometric measurements—typically a CAD or CAD-like model—or as a collection of images of an object. This information is then compared with image information to obtain a match. Comparisons can be scored by using a feature correspondence either to back-project object features into an image or to determine a new view of the object and overlay that on the image. Appropriate feature relationships can be obtained by various forms of search (e.g., Huttenlocher & Ullman, 1986; Grimson & Lozano-Perez, 1987; Lowe, 1987). Alternatively, one can define equivalence classes of features, each large enough to have distinctive properties (invariants) preserved under the imaging transformation. These invariants can then be used as an index for a model library (examples of various combinations of geometry, imaging transformations, and indexing strategies include Lamdan et al., 1988; Weiss, 1988; Forsyth et al., 1991; Rothwell et al., 1992; Stein & Medioni, 1992; Taubin & Cooper, 1992; Liu et al., 1993; Kriegman & Ponce, 1994).

Each case described so far models object geometry exactly. Systems that recognize an object by matching a view to a collection of images of an object proceed in one of two ways. In the first approach, correspondence between image features and features on the model object is either given a priori or is established by search. An estimate of the appearance in the image of that object is then constructed from the correspondences. The hypothesis that the object is present is then verified using the estimate of appearance (as in Ullman & Basri, 1991). An alternative approach computes a feature vector from a compressed version of the image and uses a minimum distance classifier to match this feature vector to feature vectors computed from images of objects in a range of positions under various lighting conditions (as in Murase & Nayar, 1995).

All of the approaches described rely heavily on specific detailed geometry, known (or easily determined) correspondences, and either the existence of a single object on a uniform known background (as in the case of Murase & Nayar, 1995) or the prospect of relatively clear segmentation. None is competent to perform abstract classification; this emphasis appears to be related to the underlying notion of model rather than to the relative difficulty of the classification versus identification. Notable exceptions appear in Nevatia and Binford (1977), Brooks (1981), Connell

(1987), and Zerroug and Nevatia (1994), which attempt to code relationships between various forms of volumetric primitive, where the description is in terms of the nature of the primitives involved and of their geometric relationship.

CONTENT-BASED RETRIEVAL FROM IMAGE DATABASES

Algorithms for retrieving information from image databases have concentrated on material-oriented queries and have implemented these queries primarily using low-level image properties such as color and texture. Object-oriented queries search for images that contain particular objects; such queries can be seen either as constructs on material queries (Picard & Minka, 1995) as essentially textual matters (Price et al., 1992) or as the proper domain of object recognition. A third query mode looks for images that are near iconic matches of a given image (e.g., Jacobs et al., 1995). This matching strategy cannot find images based on the objects present because it is sensitive to such details as the position of the objects in the image, the composition of the background, and the configuration of the objects—e.g., it could not match a front and a side view of a horse.

The best-known image database system is QBIC (Niblack et al., 1993) which allows an operator to specify various properties of a desired image. The system then displays a selection of potential matches to those criteria, sorted by a score of the appropriateness of the match. The operator can adjust the scoring function. Region segmentation is largely manual, but the most recent versions of QBIC (Ashley et al., 1995) contain simple automated segmentation facilities. The representations constructed are a hierarchy of oriented rectangles of fixed internal color and a set of tiles on a fixed grid, which are described by internal color and texture properties. However, neither representation allows reasoning about the shape of individual regions, about the relative positioning of regions of given colors, or about the cogency of geometric co-ocurrence information, and so there is little reason to believe that either representation can support object queries.

Photobook (Pentland et al., 1993) largely shares QBIC's model of an image as a collage of flat homogenous frontally presented regions but incorporates more sophisticated representations of texture and a degree of automatic segmentation. A version of Photobook incorporates a simple notion of object queries using plane object matching by an energy minimization strategy (Pentland et al., 1993, p. 10). However, the approach does not adequately address the range of variation in object shape and appears to require images that depict single objects on a uniform background. Further examples of systems that identify materials using low-level image properties include Virage (home page at <http://

www.virage.com> and elsewhere in this volume), Candid (home page at <http://www.c3.lanl.gov/~kelly/CANDID/main.shtml> and Kelly et al., 1995), and Chabot (Ogle & Stonebraker, 1995). None of these systems code spatial organization in a way that supports object queries.

Variations on Photobook (Picard & Minka, 1995; Minka, 1995) use a form of supervised learning known in the information retrieval community as "relevance feedback" to adjust segmentation and classification parameters for various forms of textured region. When a user is available to tune queries, supervised learning algorithms can clearly improve performance given appropriate object and image representations. In the most useful applications of our algorithms, however, users are unlikely to want to tune queries. Those who object to pictures of naked people are unlikely to want to spend time looking at such pictures to help tune a learning algorithm, though one might speculate that seekers could sell tuning services to avoiders.

More significantly, the representations used in these supervised learning algorithms do not code spatial relationships. Thus, these algorithms are unlikely to be able to construct a broad range of effective object queries. While relevance feedback can be effective at adjusting a metric by which image relevance is scored, it is hard to believe that user-supervised learning would be the technique of choice for establishing such intricate constructs as the variations in appearance associated with different views of a body plan.

A GROUPING-BASED FRAMEWORK FOR OBJECT RECOGNITION

Our approach to object recognition is to construct a sequence of sucessively abstracted descriptors, at an increasingly high level, through a hierarchy of grouping processes. At the lowest level, grouping is based on spatiotemporal coherence of local image descriptors—color, texture, disparity, motion—with contours and junctions extracted simultaneously to organize these groupings. There is an implicit assumption in this process that coherence of these image descriptors is correlated with the associated scene entities being part of the same surface in the scene. At the next stage, the assumptions that need to be invoked are more global (in terms of size of image region) as well as more class specific. For example, a group that is skin-colored, has an extended bilateral image symmetry, and has near parallel sides should imply a search for another such group nearby because it is likely to be a limb.

This approach leads to a notion of classification where object class is increasingly constrained as the recognition process proceeds. Classes need not be defined as purely geometric categories. For instance, in a scene expected to contain faces, prior knowledge of the spatial

configuration of eyes, mouth, etc. can be used to group what might otherwise be regarded as separate entities. As a result, the grouper's activities become increasingly specialized as the object's identity emerges; these constraints are evoked by the completion of earlier stages in grouping. The particular attractions of this view are:

- that the primary activity is classification rather than identification;
- that if grouping fails at some point, it is still possible to make statements about an object's identity;
- that it presents a coherent view of top-down information flow that is richer than a crude search; and
- that the model base now consists of information that is oriented primarily to vision (i.e., hints about grouping activities) rather than to CAD or graphics.

Slogans characterizing this approach are: grouping proceeds from the local to the global and grouping proceeeds from invoking generic assumptions to more specific ones. The most similar ideas in computer vision are those of a body of collaborators usually seen as centered around Binford and Nevatia (see, for example Nevatia & Binford, 1977; Brooks, 1981; Connell, 1987; Zerroug & Nevatia, 1994), and the work of Zisserman et al. (1995). Where we differ is in:

1. attributing much less importance to the recovery of generalized cylinders as the unifying theme for the recognition process; and
2. offering a richer view of early vision, which must offer more than contours extracted by an edge detector (an approach that patently fails when one considers objects like sweaters, brick walls, or trees).

A central notion in grouping is that of coherence, which is hard to define well but captures the idea that regions should (in some sense) "look" similar internally. Examples of coherent regions include regions of fixed color, tartan regions, and regions that are the projection of a vase. We see three major issues:

1. *Segmenting images into coherent regions based on integrated region and contour descriptors:* An important stage in identifying objects is deciding which image regions come from particular objects. This is simple when objects are made of stuff of a single fixed color. Most objects, however, are covered with textured stuff, where the spatial relationships between colored patches are an important part of any description of the stuff. The content-based retrieval literature cited above contains a wide variety of examples of the usefulness of quite simple descriptions in describing images and objects. Color histograms are a particularly popular example; however, color histograms lack spatial

cues and so must misidentify, for example, the English and the French flags. In what follows (see the Case Study 1 section), we show two important cases: in the first, measurements of the size and number of small areas of color yield information about stuff regions—such as fields of flowers—that cannot be obtained from color histograms; in the second, the observation that a region of stuff is due to the periodic repetition of a simple tile yields information about the original tile and the repetition process. Such periodic textures are common in real pictures, and the spatial structure of the texture is important in describing them.

2. *Fusing color, texture, and shape information to describe primitives*: Once regions that are composed of internally coherent stuff have been identified, 2D and 3D shape properties of the regions need to be incorporated into the region description. In many cases, objects either belong to constrained classes of 3D shapes—for example, many trees can be modeled as surfaces of revolution—or consist of assemblies of such classes—for example, people and many animals can be modeled as assemblies of cylinders. It is often possible to tell from region properties alone whether the region is likely to have come from a constrained class of shapes (e.g., see Zisserman et al., 1995); knowing the class of shape from which a region came allows other inferences. As we show in one of the following sections (see section on Case Study 2), knowing that a tree can be modeled as a surface of revolution simplifies marking the boundary of the tree and makes it possible to compute an axis and a description of the tree.

3. *Classifying objects based on primitive descriptions and relationships between primitives*: Once regions have been described as primitives, the relationships between primitives become important. For example, finding people or animals in images is essentially a process of finding regions corresponding to segments and then assembling those segments into limbs and girdles. This process involves exploring incidence relationships and is constrained by the kinematics of humans and animals. We have demonstrated the power of this constraint-based representation by building a system that can tell quite reliably whether an image contains naked people or not, which is briefly outlined in the later section describing Case Study 3.

CASE STUDY 1: COLOR AND TEXTURE PROPERTIES OF REGIONS

In the foreseeable future, it will be hard to provide users with a complete set of object concepts with which to query collections of images. To cover this omission, users can be provided with a query language that

manipulates combinations of the early visual properties that describe stuff regions. If these cues are properly chosen and can be automatically extracted, quite successful query mechanisms result. Their usefulness most probably follows because they represent a sensible choice of cues from the perspective of object recognition.

Color histograms have proven a useful stuff query but are poor at, for example, distinguishing between fields of flowers and a single large flower, because they lack information as to how the color is distributed spatially. The size and spatial distribution of areas of color is a natural stuff description—and hence, query—which is particularly useful for outdoor scenes in the case of hues ranging from red to yellow. Individual areas are hard to segment and measure, but a useful approximation can be obtained by:

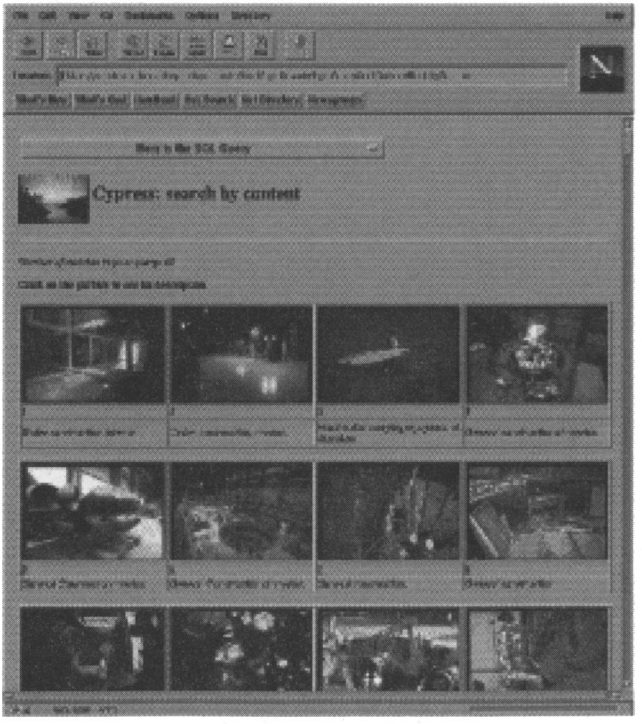

Figure 1. Querying the Cypress database for images that contain a large proportion of yellow pixels produces a collection of responses that is eclectic in content; there is little connection between the response to this query and particular objects. While these queries can be useful, particularly when combined with text information, they are not really concept or "thing" queries.

- forming R-G and B-Y opponent channels;
- coarsely requantizing these channels for various hues to form hue maps, where an orange hue map would reflect which pixels fall within a range of hues around orange;
- forming a Gaussian pyramid (after Burt & Adelson, 1983) for each hue map;
- thresholding the difference between pyramids at neighboring scales and summing to reflect the distribution of edge energy.

Finally, if an image has high energy at a coarse scale in, for example, the orange hue map, this is taken to mean it contains a large orange area;

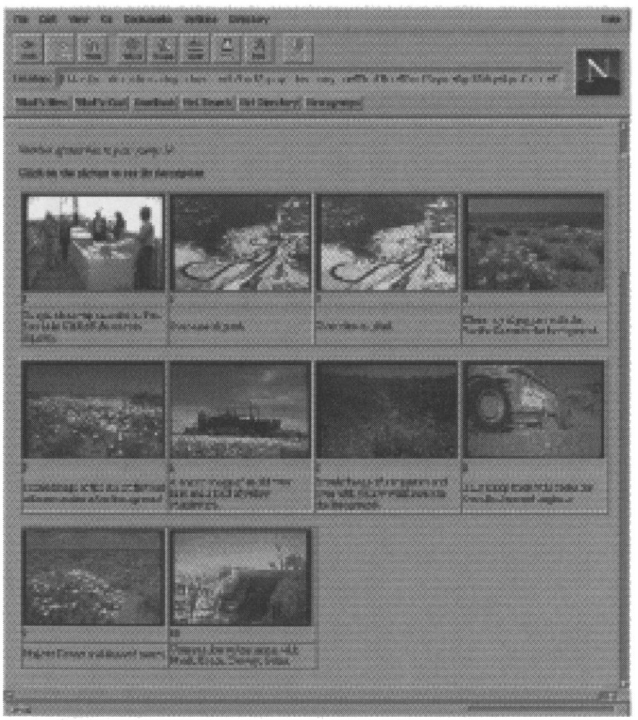

Figure 2. Querying the Cypress database for images that contain a large number of small yellow areas and a horizon yields scenic views of fields of flowers. The horizon is obtained by searching in from each boundary of the image for a blue region, extending to the boundary, that does not curve very sharply. In this case, the combination of spatial and color queries yields a query that encapsulates content surprisingly well. While the correlation between object type and query is fortuitous and relevant only in the context of the particular database, it is clear that the combination of spatial and chromatic information in the query yields a more powerful content query than color alone. In particular, the language of areas is a powerful and useful early cue to content.

comparison with the color histogram makes it possible to distinguish between few and many small areas. While this approximation is coarse, it provides extremely useful information about content. As figures 1 and 2 show, queries composed of a combination of this information with textual cues, or with an estimate of a horizon, correlate extremely strongly with content in the present Cypress database (this query engine is available on the World Wide Web at <http://elib.cs.berkeley.edu>).

A second important spatial ordering of color is the periodic repetition of a basic tile (see figure 3). Such regions can be a representation which describes the individual basic element and then represents the spatial relationships between these elements. Spatial relationships are represented by a graph where nodes correspond to individual elements and arcs join spatially neighboring elements. With each arc r_{ij} is associated an affine map A_{ij} that best transforms the image patch $I(x_i)$ to $I(x_j)$. This affine transform implicitly defines a correspondence between points on the image patches at x_i and x_j.

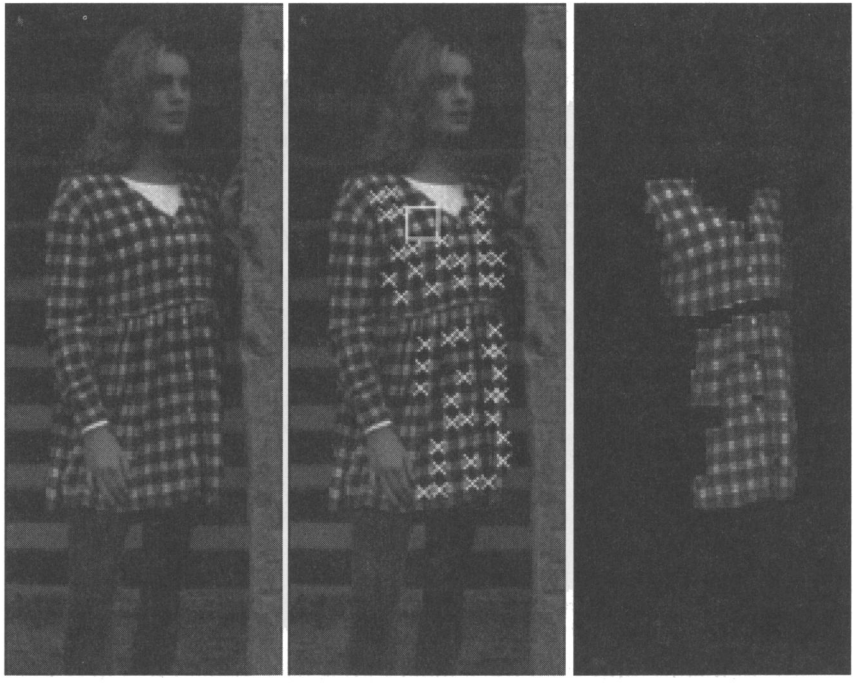

Figure 3. A textile image. The original image is shown on the left, and the center image shows the initial patches found. The crosses are the locations of units grouped together. The image on the right shows the segmented region is displayed. Notice that the rectangle includes two units in the actual pattern. This is due to the inherent ambiguity in defining a repeating unit— two tiles together still repeat to form a pattern.

Regions of periodic texture can be detected and described by:

- detecting "interesting" elements in the image;
- matching elements with their neighbors and estimating the affine transform between them;
- growing the element to form a more distinctive unit; and
- grouping the elements.

The approach is analogous to tracking in video sequences; an element is "tracked" to spatially neighboring locations in one image rather than from frame to frame. Interesting elements are detected by breaking an image into overlapping windows and computing the second moment matrix (as in Forstner, 1993; Garding & Lindeberg, 1994), which indicates whether there is much spatial variation in a window and whether that variation is intrinsically one- or two-dimensional. By summing along the dominant direction, "flow" regions—such as vertical stripes on a shirt—can be distinguished from edges. Once regions have been classified, they can be matched to regions of the same type.

An affine transform is estimated to bring potential matches into registration, and the matches are scored by an estimate of the relative difference in intensity of the registered patches. The output of this procedure is a list of elements which form units for repeating structures in the image. Associated with each element is the neighboring patches which match well with the element together with the affine transform relating them. These affine transforms contain shape cues as well as grouping cues (Malik & Rosenholtz, 1994).

The final step is to group the elements by a region-growing technique. For each of the eight windows neighboring an element, the patch which matches the element best and the affine transform between them is computed. Two patches are grouped by comparing the error between an element and its neighboring patch with the variation in the element. Of course, as the growth procedure propogates outward, the size and shape of the basic element in the image will change because of the slanting of the surface.

CASE STUDY 2: FUSING TEXTURE AND GEOMETRY TO REPRESENT TREES

Recognizing individual trees makes no sense; instead it is necessary to define a representation with the following properties:

- it should not change significantly over the likely views of the tree;
- it should make visual similarities and visual differences between trees apparent. In particular, it should be possible to classify trees into intuitively meaningful types using this representation; and

- it should be possible to determine that a tree is present in an image, segment it, and recover the representation without knowing what tree is present.

Trees can then be classified according to whether the representations are similar or not (see figure 4).

Branch length and orientation appear to be significant components of such a representation. Since trees are typically viewed frontally, with their trunks aligned with the image edges and at a sufficient distance for a scaled affine viewing model to be satisfactory, it is tempting to model a tree as a plane texture. There are two reasons not to do so: considering a tree as a surface of revolution provides grouping cues, and there is a reasonable chance of estimating parameters of the distribution of branches in 3D. Instead, we model a tree as a volume with a rotational symmetry with branches and leaves embedded in it. Because of the viewing conditions, the image of a tree corresponding to this model will have a bilateral

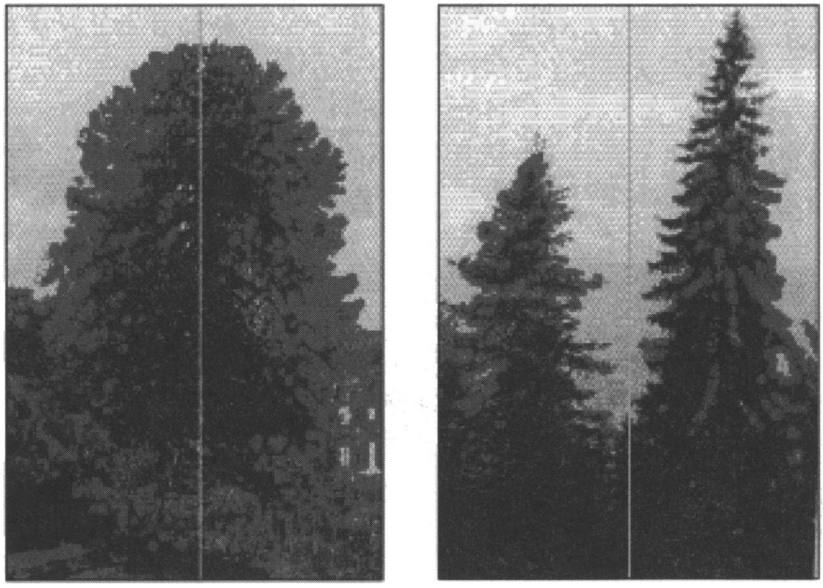

Figure 4. The viewing assumptions mean that trees have vertical axes and a reflectional symmetry about the axis. This symmetry can be employed to determine the axis by voting on its horizontal translation using locally symmetric pairs of orientation responses. **Left**: The symmetry axis superimposed on a typical image, showing also the regions that vote for the symmetry axis depicted. **Right**: In this image, there are several false axes generated by symmetric arrangements of trees; these could be pruned by noticing that the orientation response close to the axis is small.

symmetry about a vertical axis, a special case of the planar harmonic homology of Mukherjee et al. (1995). This axis provides part of a coordinate system in which the representation can be computed. The other is provided by the outline of the tree (see figure 5), which establishes scale and translation along the axis and scale perpendicular to the axis. A representation computed in this coordinate system will be viewpoint stable for the viewpoints described.

Assuming that the axis and outline have been marked, the orientation representation is obtained by forming the response of filters tuned to a range of orientations. These response strengths are summed along the axis at each orientation and for a range of steps in distance perpendicular to the axis relative to width. The representation resulting from this process (which is illustrated in figure 6) consists of a map of summed strength of response relative to orientation and distance from the axis. As the figure shows, this representation makes a range of important differences between trees explicit. Trees that have a strongly preferred branch orientation (such as pine trees) show a strong narrow peak in the representation at the appropriate orientation; trees, such as monkey puzzle trees, which have a relatively broad range of orientations of branches, show broader peaks in the representation. Furthermore, the

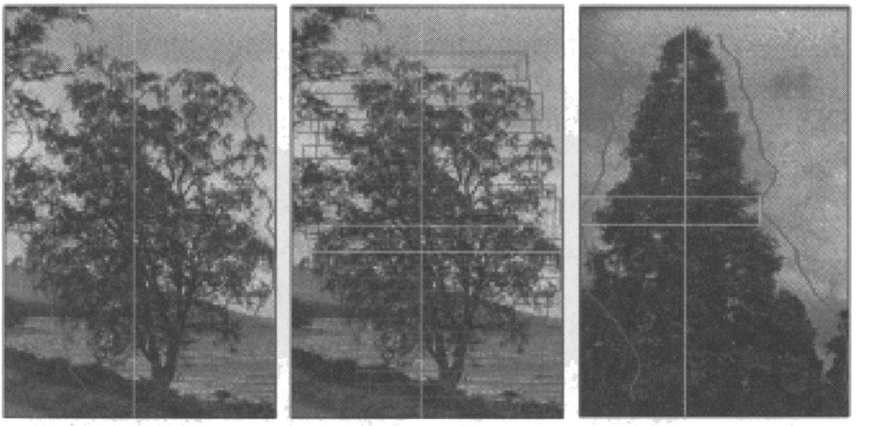

Figure 5. The outline can be constructed by taking a canonical horizontal cross-section and scaling other cross-sections to find the width that yields a cross-section that is most similar. **Left:** An outline and axis superimposed on a typical image. **Center:** The cross-sections that make up the outline superimposed on an image of the tree. **Right:** The strategy fails for trees that are poorly represented by orientations alone, as in this case, as the comparisons between horizontal slices are inaccurate. Representing this tree accurately requires using filters that respond to areas as well; such a representation would also generate an improved segmentation.

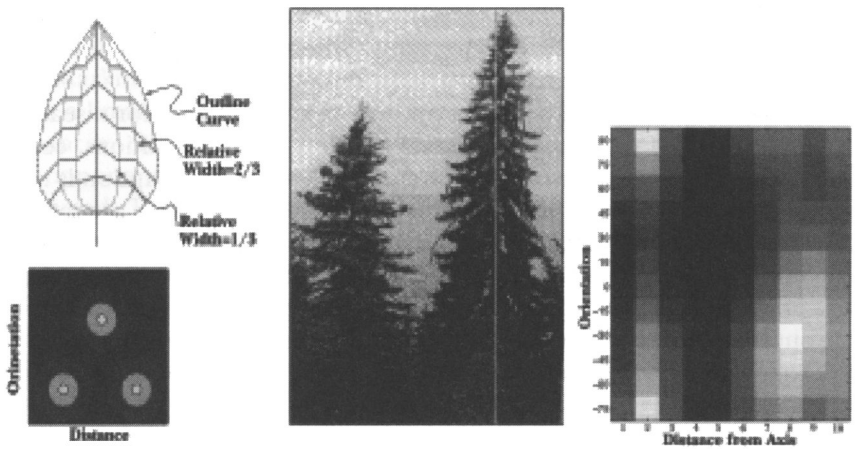

Figure 6. The orientation representation is obtained by computing the strength of response at various orientations with respect to the axis, at a range of perpendicular distances to the axis. These distances are measured relative to the width of the outline at that point and so are viewpoint stable. Responses at a particular orientation and a particular distance are summed along the height of the outline. The figure on the left illustrates the process; the representation has three clear peaks corresponding to the three branch orientations taken by the illustrative tree. The image on the extreme right shows the representation extracted for the tree in the center image. In our display of the orientation representation, brighter pixels correspond to stronger responses; the horizontal direction is distance perpendicular to the tree axis relative to the width of the tree at the relevant point, with points on the tree axis at the extreme left; the vertical direction is orientation (which wraps around). In the given case, there is a sharp peak in response close to the axis and oriented vertically, which indicates that the trunk of the tree is largely visible. A second peak oriented at about 30° and some distance out indicates a preferred direction for the tree branches.

representation distinguishes effectively between trees that are relatively translucent—such as the monkey puzzle—and those that are relatively opaque.

Finding the Axis and the Outline of a Tree

As the discussion of the representation shows, an axis and outline are important in forming the representation. Both can be found by exploiting the viewing assumptions, known constraints on the geometry of volumetric primitives, and the assumed textural coherence of the branches. The axis, which by the assumptions on viewing, is vertical and is found using a Hough transform, where pairs of strong orientation responses that display a reflectional symmetry (i.e., angles to the vertical are symmetric, and the vertical coordinates are similar) generate votes for a vertical axis. Local maxima of the accumulator array, which obtain

more than a specified minimum number of votes, are accepted as axes. Symmetric arrangements of trees generate false axes, which can be pruned by observing that regions near the axes, while symmetric, do not have strong orientation responses.

Once found, the axis serves as an organizing principle for a search for the outline of the tree. In particular, the viewing conditions, combined with an assumption of spatial homogeneity (this assumption is not always true, but is useful), imply that horizontal "slices" of tree are scaled versions of the same statistical process. In turn, this means that the outline of the tree can be generated from a single good cross-section region by a process of a search up the axis. The good section is found by searching out from the axis, at a variety of heights, to find a cross-section where a sharp change in orientation properties signals that the boundary of the tree is found.

CASE STUDY 3: FUSING COLOR, TEXTURE, AND GEOMETRY TO FIND PEOPLE AND ANIMALS

A variety of systems have been developed specifically for recognizing people or human faces. There are several domain-specific constraints in recognizing humans and animals: humans and (many) animals are made out of parts whose shape is relatively simple; there are few ways to assemble these parts; the kinematics of the assembly ensures that many configurations of these parts are impossible, and, when one can measure motion, the dynamics of these parts are limited, too. Most previous work on finding people emphasizes motion, but face-finding from static images is an established problem. The main features on a human face appear in much the same form in most images, enabling techniques based on principal component analysis or neural networks proposed by, for example, Pentland et al. (1994), Sung and Poggio (1994), Rowley et al. (1996), and Burel and Carel (1994). Face-finding based on affine covariant geometric constraints is presented by Leung et al. (1995).

However, segmentation remains a problem; clothed people are hard to segment because clothing is often marked with complex colored patterns, and most animals are textured in a way that is intended to confound segmentation. Attempting to classify images based on whether they contain naked people or not provides a useful special case that emphasizes the structural representation over segmentation, because naked people display a very limited range of colors and are untextured. Our system for telling whether an image contains naked people:

- first locates images containing large areas of skin-colored region;
- then, within these areas, finds elongated regions and groups them into possible human limbs and connected groups of limbs.

Images containing sufficiently large skin-colored groups of possible limbs are reported as potentially containing naked people. No pose estimation, back-projection, or verification is performed.

Marking skin involves stuff processing; skin regions lack texture and have a limited range of hues and saturations. To render processing invariant to changes in overall light level, images are transformed into a log-opponent representation, and smoothed texture and color planes are extracted. To compute texture amplitude, the intensity image is smoothed with a median filter; the result is subtracted from the original image, and the absolute values of these differences are run through a second median filter. The texture amplitude and the smoothed R-G and B-Y values are used to mark as probably skin all pixels whose texture amplitude is no larger than a threshold, and whose hue and saturation lie in a fixed region. The skin regions are cleaned up and enlarged slightly to accommodate possible desaturated regions adjacent to the marked regions. If the marked regions cover at least 30 percent of the image area, the image will be referred for geometric processing.

The input to the geometric grouping algorithm is a set of images in which the skin filter has marked areas identified as human skin. Sheffield's implementation of Canny's (1986) edge detector, with relatively high smoothing and contrast thresholds, is applied to these skin areas to obtain a set of connected edge curves. Pairs of edge points with a near-parallel local symmetry (as in Brady & Asada, 1984) and no other edges

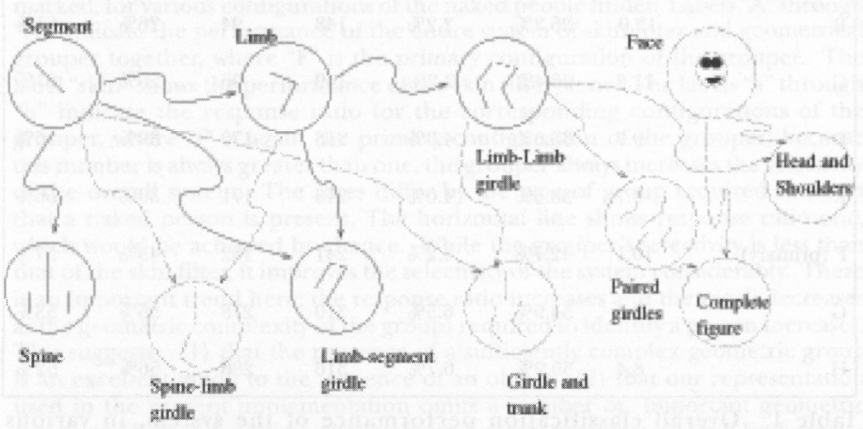

Figure 7. The grouping rules (arrows) specify how to assemble simple groups (e.g., body segments) into complex groups (e.g., limb-segment girdles). These rules incorporate constraints on the relative positions of 2D features, induced by geometric and kinematic constraints on 3D body parts. Dashed lines indicate grouping rules that are not yet implemented. Notice that this representation of human structure emphasizes grouping and assembly but can be comprehensive.

between them are found by a straightforward algorithm. Sets of points forming regions with roughly straight axes (termed "ribbons" by Brooks, 1981) are found using a Hough transformation.

Grouping proceeds by first identifying potential segment outlines, where a segment outline is a ribbon with a straight axis and relatively small variation in average width (see figure 7). Pairs of ribbons whose ends lie close together, and whose cross-sections are similar in length, are grouped to make limbs. The grouper then proceeds to assemble limbs and segments into putative girdles. It has grouping procedures for two classes of girdle—one formed by two limbs and one formed by one limb and a segment. The latter case is important when one limb segment is hidden by occlusion or by cropping. The constraints associated with these girdles are derived from the case of the hip girdle and use the same form of interval-based reasoning as used for assembling limbs. It is not possible to reliably determine which of two segments forming a limb is the thigh: the only cue is a small difference in average width, and this is unreliable when either segment may be cropped or foreshortened.

system configuration	response ratio	test response	control response	test images marked	control images marked	recall	precision
skin filter	7.0	79.3%	11.3%	448	485	79%	48%
A	10.7	6.7%	0.6%	38	27	7%	58%
B	12.0	26.2%	2.2%	148	94	26%	61%
C	11.8	26.4%	2.2%	149	96	26%	61%
D	9.7	38.6%	4.0%	218	170	39%	56%
E	9.7	38.6%	4.0%	218	171	39%	56%
F (primary)	10.1	42.7%	4.2%	241	182	43%	57%
G	8.5	54.9%	6.5%	310	278	55%	53%
H	8.4	55.9%	6.7%	316	286	56%	52%

Table 1. Overall classification performance of the system, in various configurations, to 4,289 control images and 565 test images. Configuration F is the primary configuration of the grouper, fixed before the experiment was run, which reports a naked person present if either a girdle, a limb-segment girdle, or a spine group is present, but not if a limb group is present. Other configurations represent various permutations of these reporting conditions—e.g., configuration A reports a person present only if girdles are present. There are fewer than fifteen cases because some cases give exactly the same response.

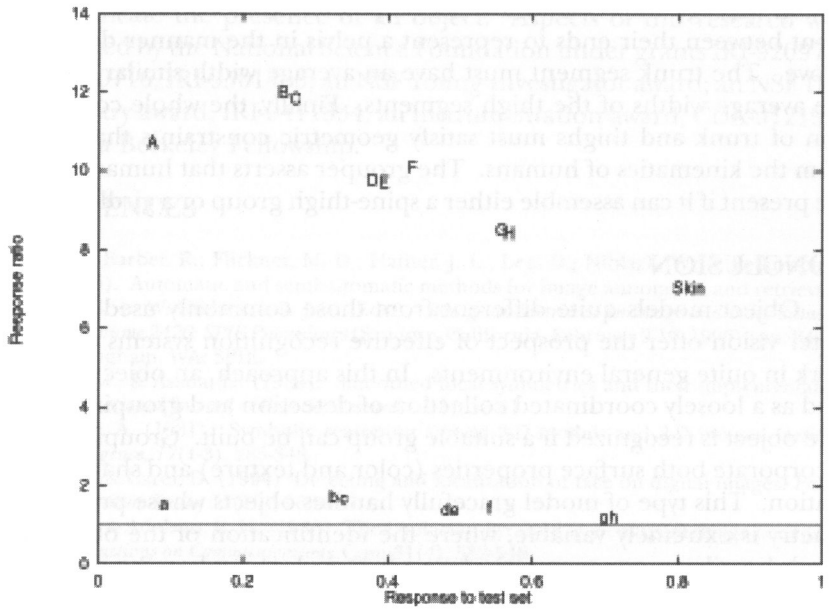

Figure 8. The response ratio (percent incoming test images marked/percent incoming control images marked), plotted against the percentage of test images marked, for various configurations of the naked people finder. Labels "A" through "H" indicate the performance of the entire system of skin filter and geometrical grouper together, where "F" is the primary configuration of the grouper. The label "skin" shows the performance of the skin filter alone. The labels "a" through "h" indicate the response ratio for the corresponding configurations of the grouper, where "f" is again the primary configuration of the grouper; because this number is always greater than one, the grouper always increases the selectivity of the overall system. The cases differ by the type of group required to assert that a naked person is present. The horizontal line shows response ratio one, which would be achieved by chance. While the grouper's selectivity is less than that of the skin filter, it improves the selectivity of the system considerably. There is an important trend here; the response ratio increases and the recall decreases as the geometric complexity of the groups required to identify a person increases. This suggests: (1) that the presence of a sufficiently complex geometric group is an excellent guide to the presence of an object, (2) that our representation used in the present implementation omits a number of important geometric structures. **Key:** A: limb-limb girdles; B: limb-segment girdles; C: limb-limb girdles or limb-segment girdles; D: spines; E: limb-limb girdles or spines; F: (two cases) limb-segment girdles or spines and limb-limb girdles, limb-segment girdles or spines; G, H each represent four cases, where a human is declared present if a limb group or some other group is found.

Spine-thigh groups are formed from two segments serving as upper thighs and a third which serves as a trunk. The thigh segments must have similar average widths, and it must be possible to construct a line segment between their ends to represent a pelvis in the manner described above. The trunk segment must have an average width similar to twice the average widths of the thigh segments. Finally, the whole configuration of trunk and thighs must satisfy geometric constraints that follow from the kinematics of humans. The grouper asserts that human figures are present if it can assemble either a spine-thigh group or a girdle group.

CONCLUSION

Object models quite different from those commonly used in computer vision offer the prospect of effective recognition systems that can work in quite general environments. In this approach, an object is modeled as a loosely coordinated collection of detection and grouping rules. The object is recognized if a suitable group can be built. Grouping rules incorporate both surface properties (color and texture) and shape information. This type of model gracefully handles objects whose precise geometry is extremely variable, where the identification of the object depends heavily on nongeometrical cues (e.g., color), and on the interrelationships between parts.

In this view of an object model and of the recognition process, model information is available to aid segmentation at about the right stages in the segmentation process in about the right form. As a result, these models present an effective answer to the usual critique of bottom up vision—i.e., that segmentation is too hard in that framework. In this view of the recognition process, the emphasis is on proceeding from general statements ("skin color") to particular statements ("a girdle"). As each decision is made, more specialized (and thereby more effective) grouping activities are enabled. Such a model is likely to be ineffective at particular distinctions ("John" versus "Fred") but effective at the kind of broad classification required by this application—an activity that has been, to date, very largely ignored by the object recognition community.

Much work is required to fully elaborate and test this model of modeling and recognition, but there is reason to believe that it will extend to cover a wide range of objects, including at least animals assembled according to the same basic body plan as humans. Our view of models gracefully handles objects whose precise geometry is extremely variable, where the identification of the object depends heavily on nongeometrical cues (e.g., color), and on the interrelationships between parts. While the present model is handcrafted and is by no means complete, there is good reason to believe that an algorithm could construct a model of this form, automatically or semi-automatically, from a 3D object model or from a range of example images.

ACKNOWLEDGMENTS

We thank Joe Mundy for suggesting that the response of a grouper may indicate the presence of an object. Aspects of this research were supported by the National Science Foundation under grants IRI-9209728, IRI-9420716, IRI-9501493; an NSF Young Investigator award; an NSF Digital Library award, IRI-9411334; an instrumentation award, CDA-9121985; and by a Berkeley Fellowship.

REFERENCES

Ashley, J; Barber, R.; Flickner, M. D.; Hafner, J. L.; Lee, D.; Niblack, W.; & Petkovich, D. (1995). Automatic and semiautomatic methods for image annotation and retrieval in QBIC. In W. Niblack & R. Jain (Eds.), *Storage and retrieval for image and video databases III: Volume 2420: SPIE Proceedings* (San Jose, California, February 9-10, 1995) (pp. 24-35). Bellingham, WA: SPIE.

Brady, J. M., & Asada, H. (1984). Smoothed local symmetries and their implementation. *International Journal of Robotics Research, 3*(3), 36-61.

Brooks, R. A. (1981). Symbolic reasoning among 3-D models and 2-D images. *Artificial Intelligence, 17*(1-3), 285-348.

Burel, G., & Carel, D. (1994). Detecting and localization of face on digital images. *Pattern Recognition Letter, 15*(10), 963-967.

Burt, P. J., & Adelson, E. H. (1983). The Laplacian Pyramid as a compact image code. *IEEE Transactions on Communications.* Com-31(4), 532-540.

Canny, J. F. (1986). A computational approach to edge detection. *IEEE Transactions on Pattern Analysis Machine Intelligence, 8*(6), 679-698.

Connell, J. H., & Brady, J. M. (1987). Generating and generalizing models of visual objects. *Artificial Intelligence, 31*(2), 159-183.

Fleck, M. (1996). The topology of boundaries. *Artificial Intelligence, 80*(1), 1-27.

Förstner, W. (1993). Image matching. In R. Haralick & L. Shapiro (Eds.), *Computer and robot vision* (vol. 2). Reading, MA: Addison-Wesley.

Forsyth, D. A.; Mundy, J. L.; Zisserman, A. P.; Heller, A.; Coehlo, C.; & Rothwell, C. A. (1991). Invariant descriptors for 3D recognition and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 13*(10), 971-991.

Gårding, J., & Lindeberg, T. (1994). Direct estimation of local surface shape in a fixating binocular vision system. In *Third European conference on computer vision—ECCV '94: Vol. 1: Proceedings* (Stockholm, Sweden, May 2-6, 1994) (pp. 365-376). Stockholm, Sweden: ECCV.

Grimson, W. E. L., & Lozano-Pérez, T. (1987). Localising overlapping parts by searching the interpretation tree. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 9*(4), 469-482.

Huttenlocher, D. P., & Ullman, S. (1986). Object recognition using alignment. In *First international conference on computer vision: Vol. 1: Proceedings* (London, June 8-11, 1987) (pp. 102-111). Washington, DC: IEEE Computer Society Press.

Jacobs, C. E.; Finkelstein, A.; & Salesin, D. H. (1995). Fast multiresolution image querying. In R. Cook (Ed.), *SIGGRAPH-95 conference proceedings* (August 6-11, 1995) (pp. 277-285). New York: ACM SIGGRAPH.

Kelly, P. M.; Cannon, M.; Hush, D. R. (1995). Query by image example: The comparison algorithm for navigating digital image databases (CANDID) approach. In W. Niblack & R. Jain (Eds.), *Storage and retrieval for image and video databases III: Vol. 2420: SPIE Proceedings* (San Jose, California, February 9-10, 1995) (pp. 238-249). Bellingham, WA: SPIE.

Kriegman, D., & Ponce, J. (1994). Five distinctive representations for recognition of curved surfaces using outlines and markings. In M. Hebert, J. Ponce, T. Boult, & A. Gross (Eds.), *Object representation in computer vision: International NSF-ARPA workshop proceedings* (New York, December 5-7, 1994) (Lecture Notes in Computer Science: 994, pp. 89-100). Berlin: Springer-Verlag.

Lamdan, Y.; Schwartz, J. T.; & Wolfson, H. J. (1988). Object recognition by affine invariant matching. In *IEEE Computer Society Conference on Computer Vision and Pattern recognition (CVPR '88) Proceedings* (Ann Arbor, Michigan, University of Michigan) (pp. 335-344). Washington, DC: IEEE Computer Society Press.

Layne, S. S. (1994). Some issues in the indexing of images. *Journal of the American Society of Information Science, 45*(8), 583-588.

Leung, T. K.; Burl, M. C.; & Perona, P. (1995). Finding faces in cluttered scenes using random labelled graph matching. In *Proceedings of the Fifth international conference on computer vision* (Cambridge, Massachusetts, June 2-23, 1995) (pp. 637-644). Los Alamitos, CA: IEEE Computer Society Press.

Liu, J.; Mundy, J. L.; Forsyth, D. A.; Zisserman, A. P.; & Rothwell, C. A. (1993). Efficient recognition of rotationally symmetric surfaces and straight homogenous generalized cylinders. *IEEE Computer Society conference on computer vision and pattern recognition (CVPR '93) Proceedings* (New York, June 15-18, 1993). Los Alamitos, CA: IEEE Computer Society Press.

Lowe, D. G. (1987). The viewpoint consistency constraint. *International Journal of Computer Vision, 1*(1), 57-72.

Malik, J., & Rosenholtz, R. (1994). Recovering surface curvature and orientation from texture distortions: A least squares algorithm and sensitivity analysis. In J. -O. Eckland (Ed.), *Third European conference on computer vision (ECCV '94) Proceedings* (Stockholm, Sweden, May 2-6, 1994) (Lecture Notes in Computer Science: 800, pp. 353-364). Berlin: Springer-Verlag.

Minka, T. (1995). *An image database browser that learns from user interaction.* Unpublished article by Massachusetts Institute of Technology media lab report # TR 365.

Mukherjee, D. P.; Zisserman, A.; & Brady, J. M. (1995). Shape from symmetry—detecting and exploiting symmetry in affine images. *Philosophical Transactions of the Royal Society of London* [Series A: Physical Sciences and Engineering], 351(1695), 77-106.

Murase, H., & Nayar, S. K. (1995). Visual learning and recognition of 3D objects from appearance. *International Journal of Computer Vision, 14*(1), 5-24.

Nevatia, R., & Binford, T. O. (1977). Description and recognition of curved objects. *Artificial Intelligence, 8,* 77-98.

Niblack, W.; Barber, R.; Equitz, W.; Flickner, M.; Glasman, E.; Petkovic, D.; & Yanker, P. (1993). The QBIC Project: Querying images by content using colour, texture and shape. In W. Niblack (Ed.), *Storage and retrieval for image and video databases: Vol. 1908: SPIE Proceedings* (San Jose, California, February 2-3, 1993) (pp. 173-187). Bellingham, WA: SPIE.

Ogle, V. E., & Stonebraker, M. (1995). Chabot: Retrieval from a relational database of images. *IEEE Computer, 28*(9), 40-48.

Pentland, A.; Picard, R. W.; & Sclaroff, S. (1993). P*hotobook: Content-based manipulation of image databases.* Unpublished MIT Media Lab Perceptual Computing TR No. 255.

Pentland, A.; Moghaddam, B.; & Starner, T. (1994). View-based and modular eigenspaces for face recognition. In *Proceedings of the IEEE Computer Society Conference on computer vision and pattern recognition* (CVPR '94, Seattle, Washington, June 21-23, 1994) (pp. 84-91). Los Alamitos, CA: IEEE Computer Society Press.

Picard, R. W., & Minka, T. (1995). Vision texture for annotation. *Journal of Multimedia Systems, 3*(1), 3-14.

Polana, R., & Nelon, R. (1993). Detecting activities. In *Proceedings of the IEEE Computer Society Conference on computer vision and pattern recognition* (CVPR '93, New York, June 15-18, 1993) (pp. 2-13). Los Alamitos, CA: IEEE Computer Society Press.

Price, R.; Chua, T.-S.; & Al-Hawamdeh, S. (1992). Applying relevance feedback to a photo-archival system. *Journal of Information Science, 18*(3), 203-215.

Rothwell, C. A.; Zisserman, A.; Mundy, J. L.; & Forsyth, D. A. (1992). Efficient model library access by projectively invariant indexing functions. In A. Rosenfeld (Chair), *Proceedings of the IEEE Computer Society conference on computer vision and pattern recognition* (CVPR '92, Champaign, Illinois, June 15-18, 1992) (pp. 109-114). Los Alamitos, CA: IEEE Computer Society Press.

Rowley, H.; Baluja, S.; & Kanade, T. (In press). Human face detection in visual scenes. *Neural Information Processing Systems, 8.*

Sclaroff, S. (1995). *World wide web image search engines* [White paper for NSF Workshop on Visual Information Management, June 1995]. Unpublished manuscript by Boston University Computer Science Department (Report #TR95-016).

Stein, F., & Medioni, G. (1992). Structural indexing: Efficient 3D object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 14*(2), 125-145.

Sung, K. K., & Poggio, T. (1994). *Example-based learning from view-based human face detection.* Unpublished MIT Artificial Intelligence Lab Memo No. 1521.

Taubin, G., & Cooper, D. B. (1992). Object recognition based on moment (or alegebraic) invariants. In J. L. Mundy & A. P. Zisserman (Eds.), *Geometric invariance in computer vision.* Cambridge, MA: MIT Press.

Taylor, B. (1977). Tense and continuity. *Linguistics and Philosophy, 1,* 199-220.

Tenny, C. L. (1987). *Grammaticalizing aspect and affectedness.* Unpublished doctoral dissertation, Department of Linguistics and Philosophy, Massachusetts Institute of Technology.

Ullman, S., & Basri, R. (1991). Recognition by linear combination of models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 13*(10), 992-1006.

Weiss, I. (1988). Projective invariants of shapes. *Proceedings of the DARPA Image Understanding Workshop* (pp. 1125-1134). San Francisco, CA: Morgan Kaufmann Publishers.

Whorf, B. L. (1941). The relation of habitual thought and behavior to language. In L. Spier, A. Hallowell, & S. Newman (Eds.), *Language, culture, and personality, essays in memory of Edward Sapir.* Unpublished paper of the Sapir Memorial Publication Fund, Menasha, WI.

Zerroug, M., & Nevatia, R. (1994). From an intensity image to 3D segmented descriptions. In *Proceedings of the 12th international conference on pattern recognition* (Jerusalem, October 9-13, 1994) (pp. 108-113). Los Alamitos, CA: IEEE Computer Society.

Zisserman, A.; Mundy, J. L.; Forsyth, D. A.; Liu, J. S.; Pillow, N.; Rothwell, C. A.; & Utcke, S. (1995). Class-based grouping in perspective images. In *Proceedings of the fifth international conference on computer vision.* (Cambridge, Massachusetts, MIT, June 20-23, 1995) (pp. 183-188). Los Alamitos, CA: IEEE Computer Society Press.