Finding Pictures of Objects in Large Collections of Images

David A. Forsyth[1], Jitendra Malik[1], Margaret M. Fleck[2], Hayit Greenspan[1,3], Thomas Leung[1], Serge Belongie[1], Chad Carson[1], Chris Bregler[1]

[1] Computer Science Division, University of California at Berkeley, Berkeley CA 94720
[2] Dept. of Computer Science, University of Iowa, Iowa City, IA 52240
[3] Dept. of Electrical Engineering, Caltech, Pasadena CA 91125

Abstract. 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 abstractly defined objects. 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 our approach to object recognition, which is 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; hierarchical grouping and learning strategies in the classification process; the ability to deal with rather general objects in uncontrolled configurations and contexts. We illustrate these properties with four case-studies: one demonstrating the use of color and texture descriptors; one showing how trees can be described by fusing texture and geometric properties; one learning scenery concepts using grouped features; and one showing how this view of recognition yields a program that can tell, quite accurately, whether a picture contains naked people or not.

1 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 indices 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 of the order of half-a-million images; a subset of this collection can be searched at `http://elib.cs.berkeley.edu`. 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 - computer programs that could automatically assess image content are a more practical alternative (Sclaroff, 1995).

Another reason that manual indexing is difficult is that it can be hard to predict future 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.

Classical object recognition techniques from computer vision cannot help with this problem. Recent techniques can identify specific objects drawn from a small (of the order of 100) collection, but no present technique is effective at telling, say, people from cows, a problem usually known as *classification*. This paper presents case studies illustrating an approach to determining 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.

1.1 Materials and Objects - "Stuff" vs "Things"

Many notions of image content have been used to organize collections of images (e.g. 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[4] and a similar distinction for actions, are 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 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.

[4] In computer vision, Ted Adelson has emphasized the role of filtering techniques in early vision for measuring stuff properties.

1.2 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 correspondences to backproject object features into an image. Appropriate feature correspondences can be obtained by various forms of search (for example, Huttenlocher and Ullman, 1986; Grimson and Lozano-Pérez, 1987; Lowe, 1987). A variant of this approach, due to Ullman and Basri (1991), uses correspondences to determine a new view of the object, which is defined by a series of images, and overlay that new view on the image to evaluate the comparison. 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 and Medioni, 1992; Taubin and Cooper, 1992; Liu *et al*, 1993; Kriegman and Ponce, 1994).

Each case described so far models object geometry exactly. An alternative approach, usually known as *appearance matching*, models objects by collections of images of the object in various positions and orientations and under various lighting conditions. These images are compressed, and feature vectors are obtained from the compressed images. Matches are obtained by computing a feature vector from a compressed version of the original image and then applying a minimum distance classifier (e.g. Sirovich and Kirby, 1987; Turk and Pentland, 1991; Murase and 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 (in the case of Murase and 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 vs. identification. Notable exceptions appear in Nevatia and Binford, 1977; Brooks, 1981; Connell, 1987; Zerroug and Nevatia, 1994, all of 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.

1.3 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 and Minka, 1995) as essentially textual matters (Price *et al.*, 1992), or as the proper domain of object recognition.

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 coocurrence 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, homogeneous frontally presented regions, but incorporates more sophisticated representations of texture and a degree of automatic segmentation. A version of Photobook (Pentland *et al.*, 1993; p. 10) incorporates a simple notion of object queries, using plane object matching by an energy minimization strategy. 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/`), Candid (home page at `http://www.c3.lanl.gov/ kelly/CANDID/main.shtml` and Kelly *et al.*, 1995) and Chabot (Ogle and Stonebraker, 1995). None of these systems code spatial organization in a way that supports object queries.

Variations on photobook (Picard and 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 a textured region. When a user is available to tune queries, supervised learning algorithms can clearly improve performance given appropriate object and image representations. In some applications of our algorithms, however, users are unlikely to want to tune queries.

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. To achieve an object-oriented query system there is a need to go to higher levels of the representation hierarchy and to encode spatial relationships using higher-level grouping features. Finally, there is a query mode that looks for images that are near iconic matches of a given image (for example, 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 - for example, it could not match a front and a side view of a horse.

2 A Grouping Based Framework for Object Recognition

Our approach to object recognition is to construct a sequence of successively abstract descriptors, at an increasingly high level, through a hierarchy of grouping and learning 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 together what might otherwise be regarded as separate entities. As a result, the grouper's activities become increasingly specialized as the object's identity emerges; constraints at higher levels 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 it presents a coherent view of combining bottom-up with top-down information flow that is richer than brute search;
– and that if grouping fails at some point, it is still possible to make statements about an object's identity.

Slogans characterizing this approach are: *grouping proceeds from the local to the global*; and *grouping proceeds 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 and Binford, 1977; Brooks, 1981; Connell, 1987; Zerroug and Nevatia, 1994), and the work of Zisserman *et al.*, 1995. Where we differ is in:

1. 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).
2. attributing much less importance to the recovery of generalized cylinders as the unifying theme for the recognition process.
3. attempting to combine learning with the hierarchical grouping processes.

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 four 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; however, most objects 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 confuse, for example, the English and the French flags. In what follows (Sec. 3), we show three important cases: in the first, features extracted from the orientation-histogram of the image are used for the extraction of coherent texture regions. This allows distinctions between uniform background and textured objects, and leads to higher-level information which can guide the recognition task. 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. Finally, measurements of the size and number of small blobs of color yield information about stuff regions - such as fields of flowers - that cannot be obtained from color histograms alone.
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 (eg Zisserman *et al.*, 1995); knowing the class of shape from which a region came allows other inferences. As we show in Sec. 4, 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. **Learning as a methodology for developing the relationship between object classes and color, texture and shape descriptors** Given the color, texture and shape descriptors for a set of labeled objects, one can use machine learning techniques to train a classifier. In section 5, we show results obtained using a decision tree classifier that was trained to distinguish among a number of visual concepts that are common in our image database. A novel aspect of this work is the use of grouping as part of the process of constructing the descriptors, instead of using simple pixel-level feature vectors. Interestingly, the output of a classifier can itself be used to guide higher level grouping. While this work is preliminary, it does suggest a way

to make less tedious the processes of acquiring object models and developing class-based grouping strategies.

4. **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 described in Sec. 6.

3 Case Study 1: Color and Texture Properties of Regions

Color and texture are two important low-level features in the initial representation of the input image; as such they form the initial phase of the grouping framework. Texture is a well-researched property of image regions, and many texture descriptors have been proposed, including multi-orientation filter banks (e.g. Malik and Perona, 1990; Greenspan *et al.*,1994), the second-moment matrix (Förstner, 1993; Gårding and Lindeberg, 1995), and orientation histograms (Freeman and Roth, 1995). We will not elaborate here on some of the more classical approaches to texture segmentation and classification- both of which are challenging and well-studied tasks. Rather, we want to introduce several new perspectives related to texture descriptors and texture grouping which were motivated from the content-based retrieval task; and which we believe present new problems in the field.

The first task that we present is that of identifying regions of uniform intensity vs. regions that are textured. This categorization enables the extraction of foreground vs. background regions in the image, guiding the search for objects in the scene. In addition, distinguishing among texture patterns which are singly-oriented, multiply-oriented, or which are stochastic in nature, can allow for further categorization of the scene and for the extraction of higher-level features to aid the recognition process (e.g. single-oriented flow is a strong characteristic of water waves, grass is stochastic etc). Finally, boundaries between a textured region and the background, or between differing texture segments, are an additional important feature which can facilitate the extraction of contour descriptors.

A view unifying region finding with contour extraction can be facilitated by extracting informative features from the orientation histogram of the gradient image. One such feature is a $180°$ normalized cross-correlation measure of the orientation histogram. An *edge*, which separates two uniform-intensity regions, is characterized by a single dominant orientation in the gradient image. Its cross-correlation figure will correspondingly be close to zero. A *bar*, on the other hand, can be thought of as the basic texture unit, and is characterized by its gradient

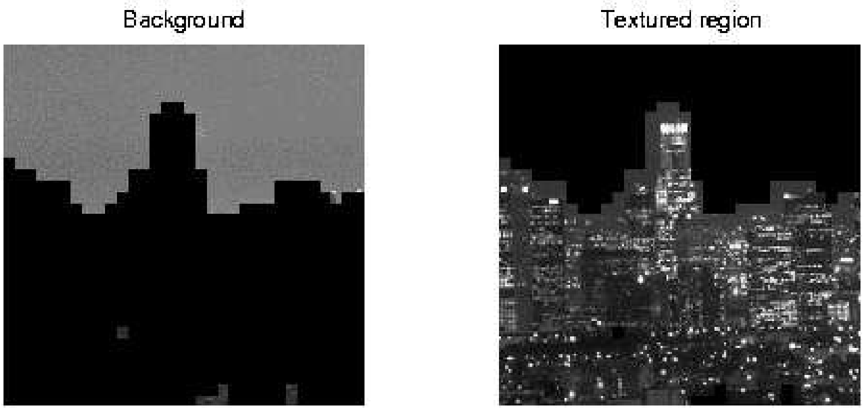

Fig. 1. *An example of non-textured vs. textured region categorization. The categorization is based on orientation histogram analysis of overlapping local windows. A non-textured window is characterized by a low DC component of the histogram. A window is labeled as textured based on a strong response to a* $180°$*-shift cross-correlation of the orientation histogram. The distinction into the two categories in this case allows for an important subdivision of the input image into the non-textured sky region and the textured city.*

map having two dominant orientations which are $180°$ phase shift apart. The normalized correlation figure is correspondingly close to one. Fitting the cyclic orientation histograms enables the extraction of additional informative features, such as relative peak energy as well as the corresponding peak angular localization. Finally, a frequency-domain analysis of the histogram harmonics can provide further region characterization.

In the following, features are extracted, characterizing the orientation histograms of two image resolutions ($8 * 8$ and $4 * 4$ overlapping windows). The combination of these features provide us with feature-vectors from which the desired categorization is enabled. Figs. 1 and 2 display preliminary results of the textured-region analysis.

A second problem of interest is the detection of periodic repetition of a basic tile, as a means for region grouping (Leung and Malik, 1996). Such regions can be described by a representation which characterizes 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 $I(\boldsymbol{x}_i)$ to $I(\boldsymbol{x}_j)$. This affine transform implicitly defines a correspondence between points on the image patches at $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$.
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 Förstner, 1993; Gårding and Lindeberg, 1995), 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 along 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 and Rosenholtz, 1994).

The final step is to group the elements together by a region growing technique. For each of the 8 windows neighboring an element, the patch which matches the element best, and the affine transform between them is computed. Two patches are grouped together by comparing the error between an element and its neighboring patch with the variation in the element. Of course, as the growth procedure propagates outward, the size and shape of the basic element in the image will change because of the slanting of the surface. An example of repetitive tile grouping is presented in Fig. 3. A more elaborate description of this work can be found in (Leung and Malik, 1996).

Of-course, texture need not be studied purely as a gray-scale phenomenon. Many interesting textures, such as fields of flowers, consist of a representative spatial distribution of *colored* elements. Color is yet another important cue in extracting information from images. 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. This example indicates the importance of fusing color with textural properties. The size and spatial distribution of blobs of color is a natural first step in such a fusion. It is also a natural stuff description - and hence, query - which is particularly useful for outdoor scenes in the case of hues ranging from red to yellow. We achieve this query by the following method:

- forming hue, saturation, and value (HSV) channels;
- coarsely requantizing these channels for various colors to form color maps, where an orange color map would reflect those pixels which fall within a certain range around orange in HSV space;
- forming a Gaussian pyramid (after Burt & Adelson, 1983) for each color map;

– filtering each level of the pyramid with a center-surround "dot" filter and several oriented "line" filters (all zero-mean);
– thresholding the filter outputs and counting the number of distinct responses to a particular filter.

Responses at a coarse scale indicate large blobs of the appropriate color; responses at finer scales indicate smaller blobs. The number of blobs at each scale and orientation for each color is returned. As Figs. 4 and 5 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/cypress`.

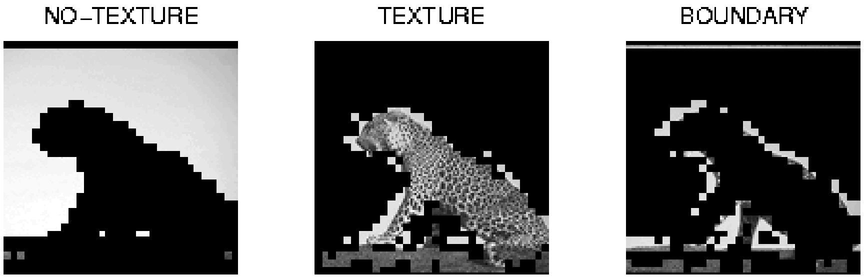

Fig. 2. *Detection of non-textured regions, textured regions and boundary elements on a Cheeta image. Local windows, of size $4*4$ are categorized into the 3 classes. Boundary elements correspond to both intensity-edges as well as textured-edges. Windows are labeled as boundary if they have a low histogram cross-correlation figure in both resolutions of analysis. We note the importance of detecting the no-texture region as a step which enables to focus further attention to the more-interesting, textured regions of the image - in this example, focusing the attention on the animal figure. The textured region extracts the entire cheeta, as well as the grass. Further, more detailed investigation on the extracted textured regions (using blob-finding or repetitive-pattern region-growing, for example, both schemes which will be described below), will enable a refined distinction between the cheeta and its surrounding. We note that the extraction of boundary information unified with textured-region information, helps eliminate the confusion between the true animal boundary and the many edges which exist within the cheeta's textured body (classic edge-finding schemes will detect all the circular edges as well).*

Fig. 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 - 2 tiles together still repeat to form a pattern.*

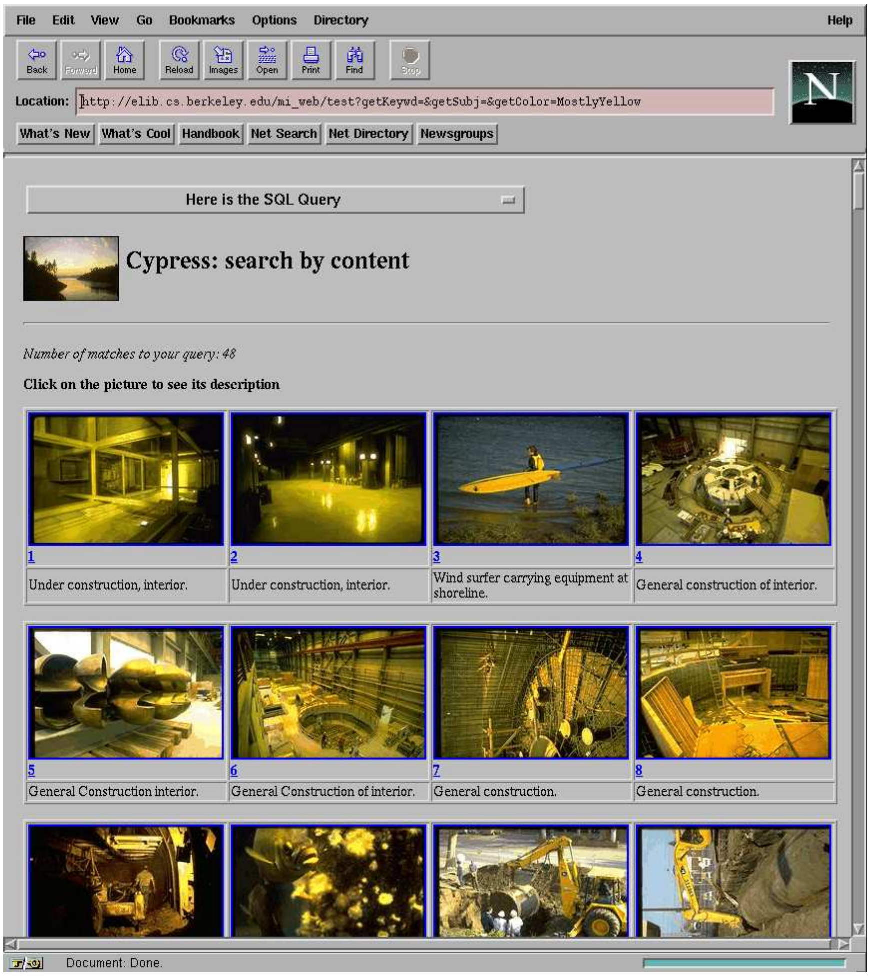

Fig. 4. *Querying the Cypress database for images that contain a large proportion of yellow pixels produces a collection of responses that is eclectic in content; there is little connection between the response to this query and particular objects. While these queries can be useful, particularly when combined with text information, they are not really concept or "thing" queries.*

Fig. 5. *Querying the Cypress database for images that contain a large number of small yellow blobs 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 blobs is a powerful and useful early cue to content. Note that the ordering of the images in response to the query (as presented in this figure) is arbitrary. No relative ranking is performed.*

4 Case Study 2: Fusing Texture and Geometry to Represent Trees

Generic grouping as studied in the previous subsection can only go so far; approaches can be made more powerful by considering classes of objects. We study trees as an interesting class. 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.
- 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.

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

An axis and an outline are important to 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. Figure 7 illustrates the axis finding procedure, and figure 8 shows how the outline follows from the axis.

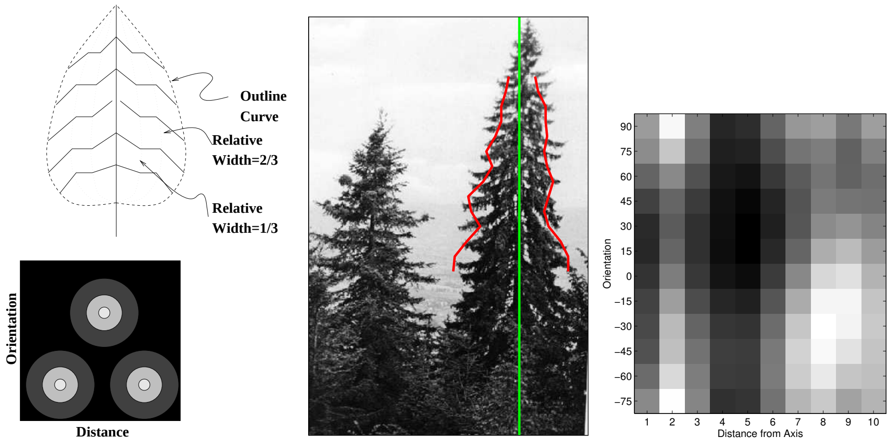

Fig. 6. *The orientation representation is obtained by computing the strength of response at various orientations with respect to the axis, at a range of perpendicular distances to the axis. These distances are measured relative to the width of the outline at that point, and so are viewpoint stable. 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 (bizarre!) 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* 30^0 *and some distance out indicates a preferred direction for the tree branches.*

5 Case Study 3: Learning Scenery Concepts Using Grouped Features

The previous case study demonstrated the power of a hand-crafted grouping strategy for an important class of objects– trees. However a legitimate concern might be that to generalize this approach would be quite cumbersome– does one hand-craft groupers for trees, buildings, roads, chairs? Our view is that given the

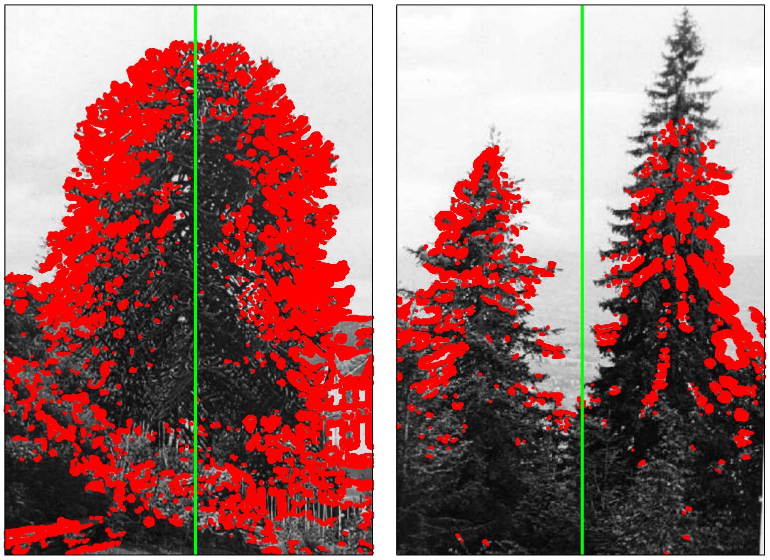

Fig. 7. *The viewing assumptions mean that trees have vertical axes and a reflectional symmetry about the axis. This symmetry can be employed to determine the axis by voting on its horizontal translation using locally symmetric pairs of orientation responses and a Hough transform. Left: The symmetry 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.*

appropriate set of visual primitives, a suite of grouping strategies and classified examples, it should be possible to use machine learning techniques to aid this process.

As a demonstration of this idea, we have implemented a simple system for automatic image annotation using images from the DWR image database. The system is capable of detecting concepts such as sky, water and man-made structure in color images. Our approach begins with an early-visual processing step to obtain color and texture information and then proceeds with a number of parallel grouping strategies. The grouping strategies seek to combine the results of the first processing step in a manner which lends itself to the task of classification. For example, a region of an image which is (1) coherent with respect to its light blue color and lack of texture, (2) is located in the upper half of the image and (3) is elongated horizontally suggests the presence of sky. The classification of concepts based on grouped features is accomplished by means of a decision tree, which was learned using C4.5 (Quinlan, 1993).

Figure 9 illustrates the first step of our approach, wherein the image is decomposed into a number of binary color/texture 'separations.' The separation images are then fed to three parallel grouping strategies. Each grouping strategy attempts to specify regions in a binary image which are coherent with respect to

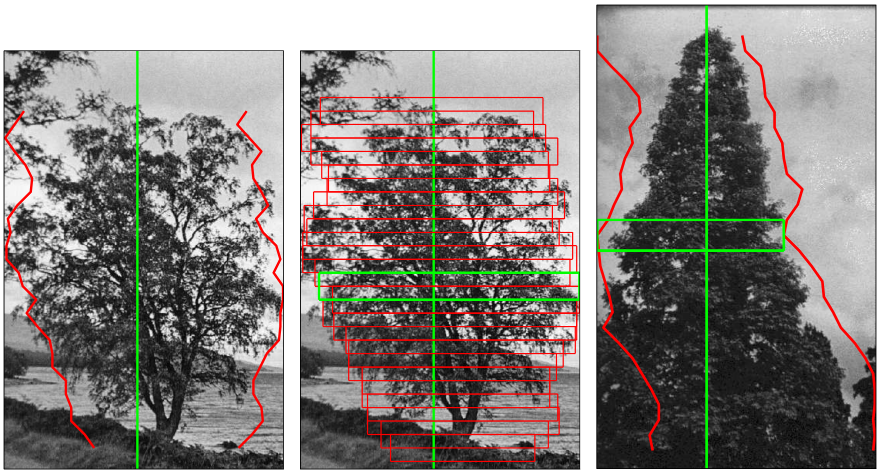

Fig. 8. *Once the axis is known, the outline can be constructed by taking a canonical horizontal cross-section and scaling other cross-sections to find the width that yields a cross-section that is most similar. Left: An outline and axis superimposed on a typical image. Center: The cross-sections that 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 blobs as well; such a representation would also generate an improved segmentation.*

one of the following rules: (1) solid contiguity, (2) similarity in local orientation and (3) similarity in diffuseness. The results of the three grouping strategies applied to the yellow, green, light blue and 'rough' separation images are shown in Figure 10.

Each blob is represented by a feature vector containing its area, coordinates in the image, eccentricity, principle orientation, mean saturation and mean intensity, as well as the color/texture separation and grouping strategy which gave rise to it. These feature vectors are the input to a decision tree classifier. The decision tree attempts to assign a label to each blob according to these characteristics. The class confusion matrix obtained in our experiments is shown in Figure 12. The performance of the system as summarized in the confusion matrix was obtained using 10-fold cross validation. For the most part, the confusion matrix contains large diagonal entries, indicating a tendency to classify concepts correctly. Notice that when errors do occur, the incorrectly chosen class tends to share some salient characteristics with the correct class (e.g. tree and vegetation).

While the current system's use of shape (i.e. area and principle axes) is somewhat primitive, there is a natural way to proceed within the same general framework. The additional use of symmetry features and repeating patterns, for example, promises to extend the capabilities of the system beyond simple blob-

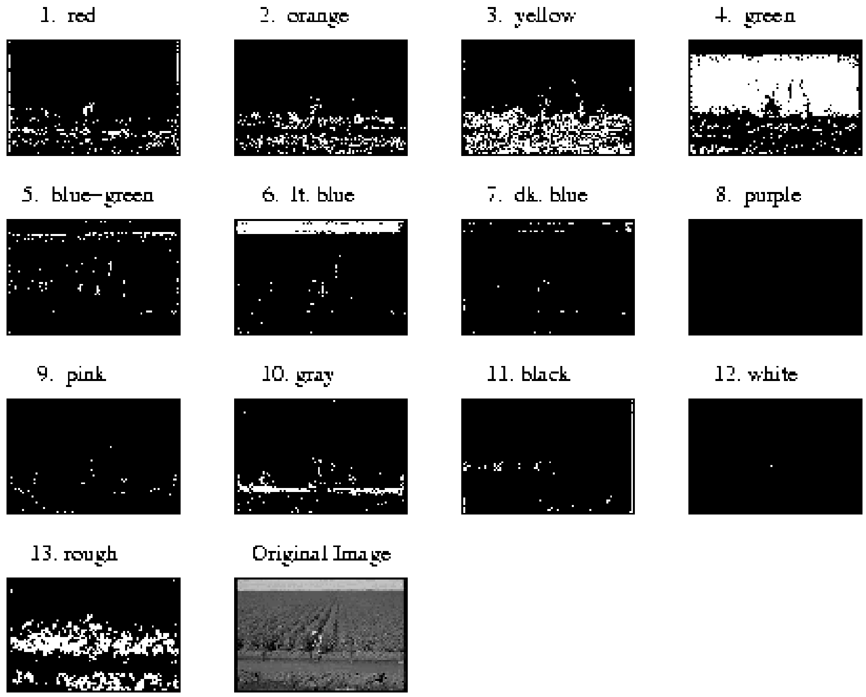

Fig. 9. *Illustrating the color/texture separations for a test image. Separations 1-9 were formed strictly by gating the hue; e.g., a hue in the interval* $[.07, .1)$ *is labeled as 'orange.' Separations 10-12 each made use of saturation and/or intensity. Lastly, separation 13, 'rough,' made use of the eigenvalues of the windowed-image second moment matrix computed for the intensity component of the original image.*

like scenery concept detection. An interesting question to address is, given the component features needed to detect a tree in an image, can the performance of the hand-crafted tree recognizer of case study 2 be matched by a special instance of a general grouping-based concept learner?

In future work, we intend to answer this question by augmenting the toolbox of early vision descriptors and by adding more grouping strategies. We also intend to investigate additional learning strategies such as Bayes' nets and decision graphs for improved handling of spatial relationships between simpler concepts.

6 Case Study 4: 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

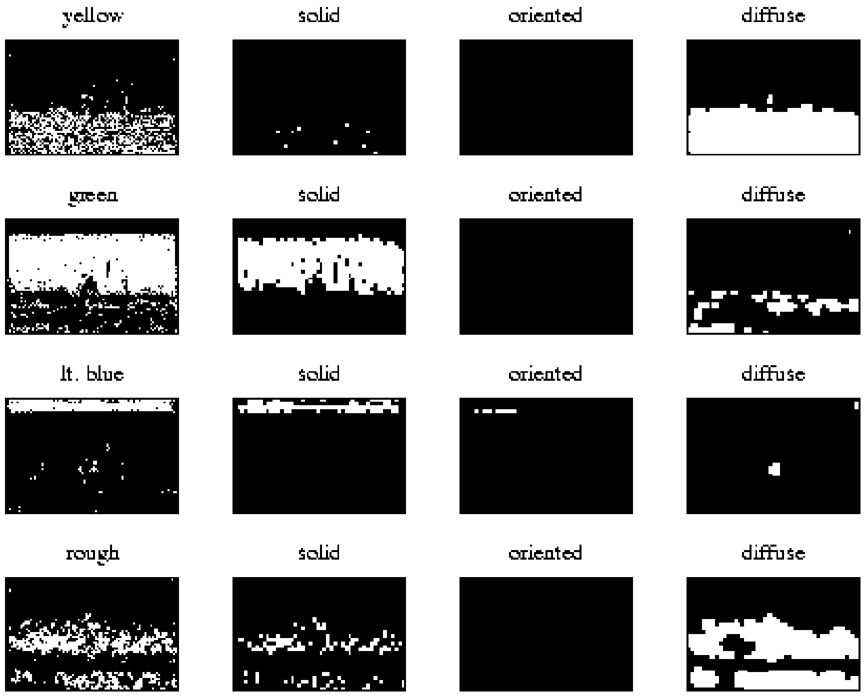

Fig. 10. *The results of the three grouping strategies for four of the color/texture bin images from the preceding figure. For example, the top row indicates that nearly all of the pixels in the 'yellow' separation were accounted for as a 'diffuse' region. Since there was no strongly oriented structure in the original image underlying the 'yellow' separation, no pixels were labeled as 'oriented.'*

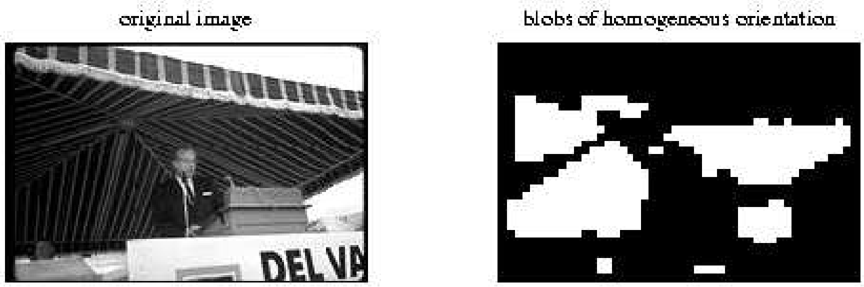

Fig. 11. *The homogeneously oriented regions in the striped canopy in the image on the left show up as distinct blobs in the result of the second grouping strategy, shown on the right. (The color/texture separation was equal to one everywhere, thus defining the entire image as a potential area of interest.)*

(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)	(l)	<-classified as
5				1	2							(a) cloud
	9	1		1			2					(b) dirt
		12						1				(c) flower
										8		(d) lawn
	1			27			2		1	1		(e) manmade
1					28				1	1		(f) sky
						6		1		1		(g) snow
							13					(h) tan-ground
				1				8		2	1	(i) tarmac
1	1								8	11		(j) tree
				1						36		(k) vegetation
									1	1	10	(l) water

Fig. 12. *The class confusion matrix. Off-diagonal entries correspond to misclassifications. Notice that all 8 of the lawn examples were misclassified as vegetation. Similarly, 8 out of 21 trees were misclassified as vegetation. This suggests that our features were not rich enough to discriminate these two classes. Notice also that 2 out of 8 cloud examples were misclassified as sky, probably because of noise in the training data.*

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; 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 (Fleck *et al.*, 1996) 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.

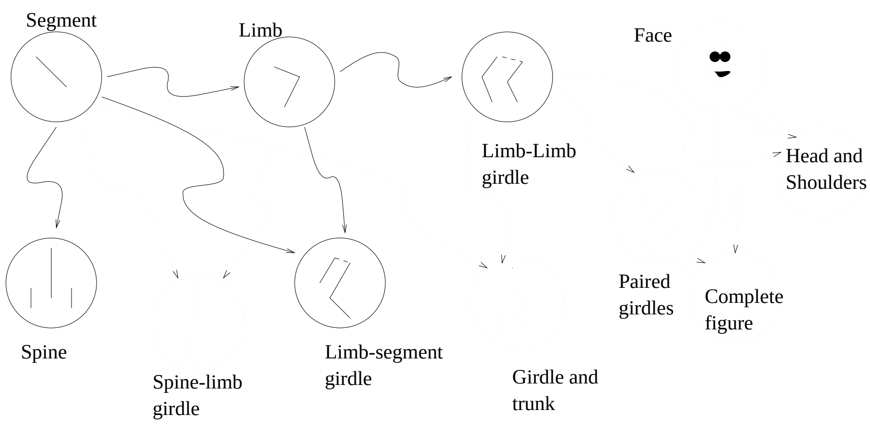

Fig. 13. *The grouping rules (arrows) specify how to assemble simple groups (e.g. body segments) into complex groups (e.g. limb-segment girdles). These rules incorporate constraints on the relative positions of 2D features, induced by geometric and kinematic constraints on 3D body parts. Dashed lines indicate grouping rules that are not yet implemented. Notice that this representation of human structure emphasizes grouping and assembly, but can be comprehensive.*

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% 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 and Asada, 1984) and no other edges between them are found by a straightforward algorithm. Sets of points forming regions with roughly straight axes ("ribbons"; 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. Pairs of ribbons whose ends lie close together, and whose cross-sections are similar in length, are grouped together 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 use the same form of interval-based reasoning as used for assembling limbs. Finally, the grouper can form spine-thigh groups from two segments serving as upper thighs, and a third, which serves as a trunk.

In its primary configuration, the system uses the presence of either form of girdle group or of a spine-thigh group to assert that a naked human figure is present in the image. This yields a system that is surprisingly accurate for so abstract a query; as figure 14 shows, using different groups as a diagnostic for the presence of a person indicates a significant trend. The selectivity of the system increases, and the recall decreases, as the geometric complexity of the groups required to identify a person increases, suggesting that our representation used in the present implementation omits a number of important geometric structures and that the presence of a sufficiently complex geometric group is an excellent guide to the presence of an object.

7 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. The primary focus is on *classification* instead of *identification*. The central process is that of hierarchical grouping. Initially, the grouping is based on quite local (short range in the image) measurements of color and texture coherence; as it proceeds more global and more specific models, e.g. surfaces of revolution, are invoked. In this approach, the object database is modeled as a loosely coordinated collection of detection and grouping rules. An 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 non-geometrical cues (e.g. color and texture) and on the interrelationships between parts. Learning can be incorporated into the framework as a convenient way of associating object class labels with color, texture and shape descriptors.

We demonstrated the paradigm with four case studies that are prototype implementations of modules of such a grouping based recognition system.

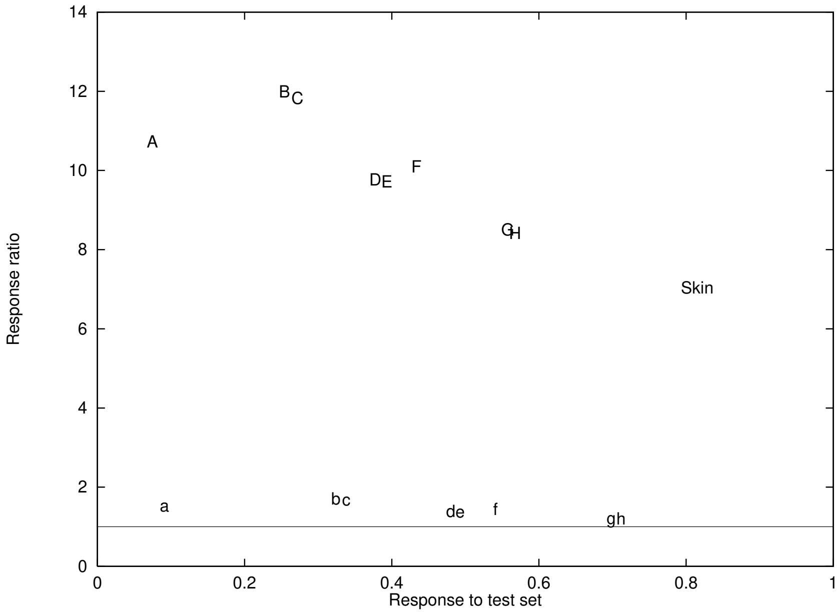

Fig. 14. *The response ratio, (percent incoming test images marked/percent incoming control images marked), plotted against the percentage of test images marked, for various configurations of the naked people finder. Labels "A" through "H" indicate the performance of the entire system of skin filter 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 guided 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.*

Acknowledgments

We would like to thank R. Blasi and K. Murphy who collaborated with S. Belongie in the work on learning decision trees for visual concept classification. 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

1. Ashley, J., Barber, R., Flickner, M.D., Hafner, J.L., Lee, D., Niblack, W. and Petkovich, D. (1995) "Automatic and semiautomatic methods for image annotation and retrieval in QBIC," *SPIE Proc. Storage and Retrieval for Image and Video Databases III*, 24-35.
2. Belongie, S., Blasi, R., and Murphy, K. (1996) "Grouping of Color and Texture Features for Automated Image Annotation," Technical Report for CS280, University of California Berkeley.
3. Brady, J.M. and Asada, H. (1984) "Smoothed Local Symmetries and Their Implementation," *Int. J. Robotics Res.* 3/3, 36–61.
4. Brooks, R.A. (1981) "Symbolic Reasoning among 3-D Models and 2-D Images," *Artificial Intelligence* 17, pp. 285–348.
5. Burel, G., and Carel, D. (1994) "Detecting and localization of face on digital images" *Pattern Recognition Letters 15* pp 963-967.
6. Burt, P.J., and Adelson, E.H., (1983) "The Laplacian Pyramid as a Compact Image Code," *IEEE Trans. on Communications*, vol. com-31, no. 4.
7. Canny, J.F. (1986) "A Computational Approach to Edge Detection," *IEEE Patt. Anal. Mach. Int.* 8/6, pp. 679–698.
8. Connell, J.H., and Brady, J.M. (1987) "Generating and Generalizing Models of Visual Objects," *Artificial Intelligence*, **31**, 2, 159–183.
9. Fleck, Margaret M. (1996) "The Topology of Boundaries," in press, *Artificial Intelligence.*
10. Fleck, M.M., Forsyth, D.A., and Bregler, C. (1996) "Finding Naked People," *Fourth European Conference on Computer Vision*, Cambridge, UK, Vol 2, pp. 593-602.
11. Förstner, W. (1993) Chapter 16, in Haralick, R. and Shapiro, L. *Computer and Robot Vision*, Addison-Wesley.
12. Forsyth, D.A., Mundy, J.L., Zisserman, A.P., Heller, A., Coehlo, C., and Rothwell, C.A. (1991) "Invariant Descriptors for 3D Recognition and Pose," *IEEE Trans. Patt. Anal. and Mach. Intelligence*, **13**, 10.
13. Freeman, W., and Roth, M. (1995) Orientation histograms for hand gesture recognition. *International Workshop on Automatic Face- and Gesture-Recognition.*
14. Garding, J., and Lindeberg, T. (1996) Direct computation of shape cues using scale-adapted spatial derivative operators. *Int. J. of Computer Vision*, 17, February 1996.
15. Greenspan, H., Goodman R., Chellappa, R., and Anderson, S. (1994) "Learning Texture Discrimination Rules in a Multiresolution System," in the special issue on "Learning in Computer Vision" of the *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, Vol. 16, No. 9, 894-901.

16. Greenspan, H., Belongie, S., Perona, P., and Goodman, R. (1994) "Rotation Invariant Texture Recognition Using a Steerable Pyramid," *12th International Conference on Pattern Recognition (ICPR)*, Jerusalem, Israel.
17. Grimson, W.E.L. and Lozano-Pérez, T. (1987) "Localising overlapping parts by searching the interpretation tree", *PAMI*, **9**, 469-482.
18. Huttenlocher, D.P. and Ullman, S. (1986) "Object recognition using alignment," *Proc. ICCV-1*, 102-111.
19. Jacobs, C.E., Finkelstein, A., and Salesin, D.H. (1995) "Fast Multiresolution Image Querying," *Proc SIGGRAPH-95*, 277-285.
20. Kelly, P.M., Cannon, M., Hush, D.R. (1995) "Query by image example: the comparison algorithm for navigating digital image databases (CANDID) approach," *SPIE Proc. Storage and Retrieval for Image and Video Databases III*, 238-249.
21. Kriegman, D. and Ponce, J. (1994) "Representations for recognising complex curved 3D objects," *Proc. International NSF-ARPA workshop on object representation in computer vision*, LNCS-994, 89-100.
22. Lamdan, Y., Schwartz, J.T. and Wolfson, H.J. (1988) "Object Recognition by Affine Invariant Matching," Proceedings CVPR, p.335-344.
23. Layne, S.S. (1994) "Some issues in the indexing of images," *J. Am. Soc. Information Science*, **45**, 8, 583-588.
24. Leung, T.K., Burl, M.C., Perona, P. (1995) "Finding faces in cluttered scenes using random labelled graph matching, " *International Conference on Computer Vision* pp 637-644.
25. Leung, T.K., and Malik, J., "Detecting, localizing and grouping repeated scene elements from an image," (1996) *Fourth European Conference on Computer Vision*, Cambridge, UK, Vol 1, pp. 546-555.
26. Liu, J., Mundy, J.L., Forsyth, D.A., Zisserman, A.P., and Rothwell, C.A. (1993) "Efficient Recognition of rotationally symmetric surfaces and straight homogenous generalized cylinders," *IEEE Conference on Computer Vision and Pattern Recognition '93*.
27. Lowe, David G. (1987) "The Viewpoint Consistency Constraint," *Intern. J. of Comp. Vis,* 1/1, pp. 57–72.
28. Malik, J., and Perona, P. (1990) "Preattentive texture discrimination with early vision mechanisms," *J. Opt. Soc. Am. A*, 7(5):923–932.
29. Malik, J., and Rosenholtz, R. (1994) "Recovering surface curvature and orientation from texture distortion: a least squares algorithm and sensitivity analysis," *Proc. of Third European Conf. on Computer Vision*, Stockholm, published as J.O. Eklundh (ed.) LNCS **800**, Springer Verlag, pp. 353-364.
30. Minka, T. (1995) "An image database browser that learns from user interaction," MIT media lab TR 365.
31. Mukherjee, D.P., Zisserman, A., and Brady, J.M. (1995) "Shape from symmetry - detecting and exploiting symmetry in affine images," *Proc. Roy. Soc.*, **351**, 77-106.
32. Murase, H. and Nayar, S.K. (1995) "Visual learning and recognition of 3D objects from appearance," *Int. J. Computer Vision*, **14**, 1, 5-24.
33. Nevatia, R. and Binford, T.O. (1977) "Description and recognition of curved objects," *Artificial Intelligence*, **8**, 77-98, 1977
34. Niblack, W., Barber, R, Equitz, W., Flickner, M., Glasman, E., Petkovic, D., and Yanker, P. (1993) "The QBIC project: querying images by content using colour, texture and shape," *IS and T/SPIE 1993 Intern. Symp. Electr. Imaging: Science and Technology, Conference 1908, Storage and Retrieval for Image and Video Databases.*

35. Ogle, Virginia E. and Michael Stonebraker (1995) "Chabot: Retrieval from a Relational Database of Images," *Computer* 28/9, pp. 40–48.
36. Pentland A., Moghaddam, B., Starner T., (1994) "View-based and modular eigenspaces for face recognition," *Computer Vision and Pattern Recognition*, pp 84-91.
37. Pentland, A., Picard, R.W., and Sclaroff, S. (1993) "Photobook: content-based manipulation of image databases," MIT Media Lab Perceptual Computing TR No. 255.
38. Picard, R.W. and Minka, T. (1995) "Vision texture for annotation," *J. Multimedia systems*, **3**, 3-14.
39. Polana, R., Nelon, R. (1993) "Detecting Activities" *Computer Vision and Pattern Recognition* pp 2-13.
40. Price, R., Chua, T.-S., Al-Hawamdeh, S. (1992) "Applying relevance feedback to a photo-archival system," *J. Information Sci.*, **18**, 203-215.
41. J. R. Quinlan, *C4.5 Programs for Machine Learning*, Morgan Kauffman, 1993.
42. Rothwell, C.A., Zisserman, A., Mundy, J.L., and Forsyth, D.A. (1992) "Efficient Model Library Access by Projectively Invariant Indexing Functions," *Computer Vision and Pattern Recognition 92*, 109-114.
43. Rowley, H., Baluja, S., Kanade, T. (1996) "Human Face Detection in Visual Scenes" *NIPS*, volume 8, 1996.
44. Sclaroff, S. (1995) "World wide web image search engines," Boston University Computer Science Dept TR95-016.
45. Sirovitch, L. and Kirby, M., "Low-dimensional procedure for the characterization of human faces," *J. Opt. Soc. America A*, **2**, 519-524, 1987.
46. Stein, F. and Medioni, G. (1992) "Structural indexing: efficient 3D object recognition," *PAMI-14*, 125-145.
47. Sung, K.K, Poggio, T., (1994) "Example-based Learning from View-based Human Face Detection" MIT A.I. Lab Memo No. 1521.
48. Taubin, G. and Cooper, D.B. (1992) "Object recognition based on moment (or algebraic) invariants," in J.L. Mundy and A.P. Zisserman (ed.s) *Geometric Invariance in Computer Vision*, MIT Press.
49. Taylor, B., (1977) "Tense and Continuity" Linguistics and Philosophy 1 199–220.
50. Tenny, C.L. (1987) "Grammaticalizing Aspect and Affectedness," Ph.D. thesis, Linguistics and Philosophy, Massachusetts Inst. of Techn.
51. Turk, M. and Pentland, A., "Eigenfaces for recognition," *J. Cognitive Neuroscience*, **3**, 1, 1991.
52. Ullman, S. and Basri, R. (1991) "Recognition by linear combination of models," *IEEE PAMI*, **13**, 10, 992-1007.
53. Weiss, I. (1988) "Projective Invariants of Shapes," Proceeding DARPA Image Understanding Workshop, p.1125-1134.
54. Whorf, B.L. (1941) "The Relation of Habitual Thought and Behavior to Language," in Leslie Spier, ed., *Language, culture, and personality, essays in memory of Edward Sapir*, Sapir Memorial Publication Fund, Menasha, WI.
55. Zerroug, M. and Nevatia, R. (1994) "From an intensity image to 3D segmented descriptions," *Proc 12'th ICPR*, 108-113.
56. Zisserman, A., Mundy, J.L., Forsyth, D.A., Liu, J.S., Pillow, N., Rothwell, C.A. and Utcke, S. (1995) "Class-based grouping in perspective images", *Intern. Conf. on Comp. Vis.*

This article was processed using the LaTeX macro package with LLNCS style