FINDING OBJECTS IN IMAGE DATABASES BY GROUPING

J. Malik, D.A. Forsyth, M.M. Fleck†, H. Greenspan‡,
T. Leung, C. Carson, S. Belongie & C. Bregler
Computer Science Division, University of California at Berkeley, Berkeley CA 94720

ABSTRACT

Retrieving images from very large collections, using image content as a key, is becoming an important problem. Finding objects in image databases is a big challenge in the field. This paper describes our approach to object recognition, which is distinguished by: a rich 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 three case-studies: one demonstrating the use of color and texture descriptors; one learning scenery concepts using grouped features; and one demonstrating a possible application domain in detecting naked people in a scene.

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 (DWR) 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. 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 the general classification task. In this short paper we will not attempt to cover all the related literature in the field (e.g. [1,2]). For a complete reference and comparison among the current systems in handling image databases, please refer to [3].

This paper presents case studies illustrating an approach to determining image content that is capable of object classification. Our approach 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. 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. Slogans characterizing this approach are: *grouping proceeds from the local to the global*; and *grouping proceeds from invoking generic assumptions to more specific ones.*

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; however, most objects are covered with textured stuff, where the spatial relationships between colored patches are important. The content-based retrieval literature 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 (section 2), 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. 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. 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. 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 3, 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.

3. Classifying objects based on primitive descriptions and relationships between primitives: Once regions have been described as primitives, the relationships between primitives become important. 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 explor-

†Department of Computer Science, University of Iowa, Iowa City, IA 52240
‡also with the Dept. of Electrical Engineering, CALTECH, Pasadena CA 91125

0-7803-3258-X/96/$5.00 © 1996 IEEE

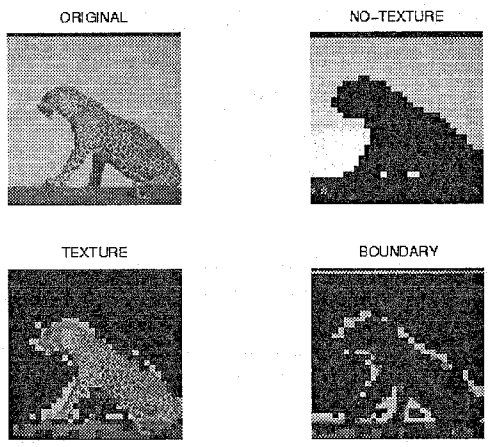

Figure 1: *Local windows, of size 4 × 4 pixels,are classified as non-textured regions, textured regions and boundary elements in this image of a leopard by analysing the orientation histogram of the gradient image in the window. A small DC component in the orientation histogram identifies no-texture regions, permitting us to focus further attention on the more interesting, textured regions of the image. Boundary elements correspond to both brightness edges as well as texture edges;. they are to be distinguished from the interiors of textured regions. This distinction is made by computing the 180° normalized cross-correlation measure of the orientation histogram. Windows are labeled as boundary if they have a small histogram cross-correlation figure at both high and low resolutions; textured otherwise. Note that the textured region extracts the entire leopard, as well as the grass. Extracting boundary information unified with textured-region information helps eliminate the confusion between the true animal boundary and the many edges caused by the animal's spots.*

ing incidence relationships, and is constrained by the human and animal kinematics. 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 section 4.

2. CASE 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. 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,

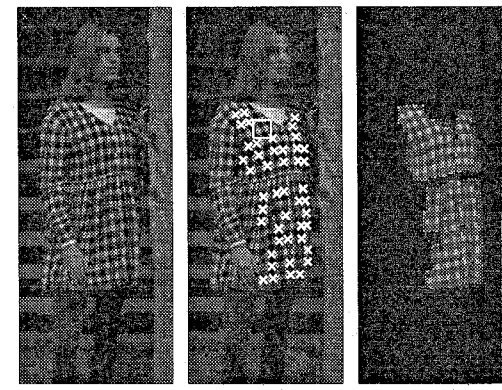

Figure 2: *A textile image. The original image is shown on the left, and the center image shows the initial patches found. The crosses are the locations of units grouped together. The image on the right shows the segmented region is displayed.*

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 and shape understanding. Figure 1 displays 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. 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 $I(x_i)$ to $I(x_j)$. This affine transform implicitly defines a correspondence between points on the image patches at x_i and x_j. An example of repetitive tile grouping is presented in figure 2. A more elaborate description of this work can be found in [4].

Finally, we note that many interesting textures, such as fields of flowers, consist of a representative spatial distribution of *colored* elements. The size and spatial distribution of blobs of color is a natural first step in fusing color and texture. 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 a Gaussian pyramid for each color map obtained by coarsely requantizing hue, saturation, and value (HSV) channels. For example, an orange color map would reflect those pixels which fall within a certain range around orange in HSV space;
- thresholding the output of a system of a center-surround

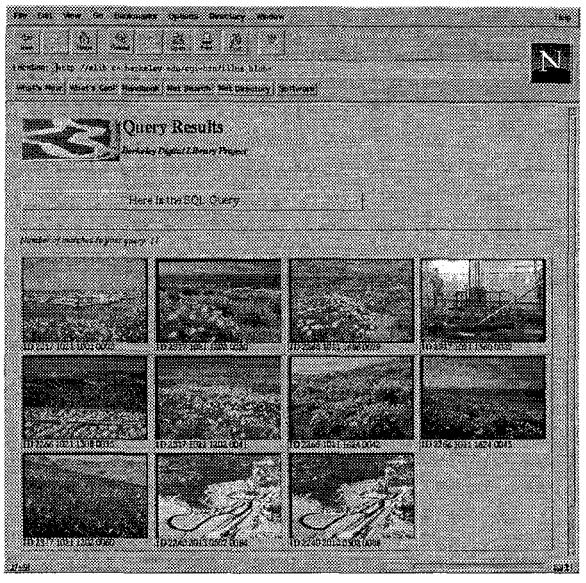

Figure 3: *Querying the Cypress database for images that contain a large number of small yellow blobs and a horizon yields scenic views of fields of flowers (available on the Web at* `http://elib.cs.berkeley.edu/cypress`*). The horizon is obtained by searching in from each boundary of the image for a 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. This example demonstrates that 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.*

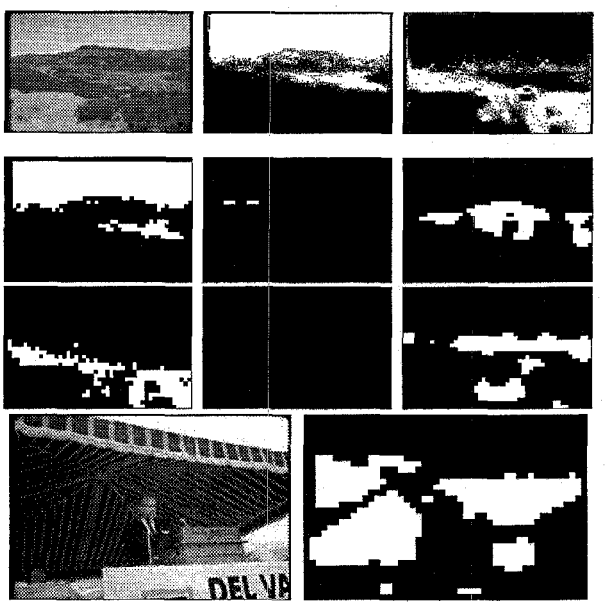

Figure 4: *Illustrating the color/texture channels and grouping strategies for a test image. Top row: original image, lt. blue separation, green separation. Rows 2 and 3: results of the three grouping strategies (solid, oriented, diffuse) for the lt. blue and green channels, respectively. Bottom row: 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 (oriented) grouping strategy, shown on the right. Combinations of these properties that are diagnostic of various types of stuff are then learned; typical classification results appear in table 1.*

"dot" filters and oriented "line" filters (all zero-mean), 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. Figure 3 shows the results of one such query.

3. CASE 2: LEARNING TO GROUP SCENERY

Given a set of visual primitives, we can learn higher level concepts by appropriate grouping strategies. As a demonstration of this idea, we have implemented a simple system for automatic image annotation. 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, a commercial system for learning decision trees from labeled examples).

Figure 4 (top row) illustrates the first step of our approach, wherein the image is decomposed into a number of binary color/texture 'channels.' The color channels are based on hue, saturation and intensity, while the texture channels are based on the eigenvalues of the second moment matrix. 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 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 light blue and green separation images are shown in the center rows of figure 4.

Each blob (or grouped region) is represented next 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 class confusion matrix obtained in our experiments is shown in Table 1. The performance of the system as summarized in the confusion matrix was obtained using 10-fold cross

labeled as:	a	b	c	d	e	f	g	h	i	j	k	l
a.cloud	5	0	0	0	1	2	0	0	0	0	0	0
b.dirt	0	9	1	0	1	0	0	2	0	0	0	0
c.flower	0	0	12	0	0	0	0	0	1	0	0	0
d.lawn	0	0	0	0	0	0	0	0	0	0	8	0
e.manmade	0	1	0	0	27	0	0	2	0	1	1	0
f.sky	1	0	0	0	0	28	0	0	0	1	1	0
g.snow	0	0	0	0	0	0	6	0	1	0	1	0
h.ground	0	0	0	0	0	0	0	13	0	0	0	0
i.tarmac	0	0	0	0	1	0	0	0	8	0	2	1
j.tree	1	1	0	0	0	0	0	0	0	8	11	0
k.veg.	0	0	0	0	1	0	0	0	0	0	36	0
l.water	0	0	0	0	0	0	0	0	0	1	1	10

Table 1: *The class confusion matrix of our learned stuff classification system, showing few off-diagonal entries (which correspond to misclassifications). 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. Some cloud examples were also misclassified as sky, probably because of noise in the training data.*

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 principal axes) is somewhat primitive, there is a natural way to proceed within the same general framework, using symmetry features and repeating patterns, for example, to extend the capabilities of the system. Future work, will augment the toolbox of early vision descriptors, add more grouping strategies, and investigate additional learning strategies (such as Bayes' nets and decision graphs) for improved handling of spatial relationships between concepts.

4. CASE 3: GROUPING TO FIND PEOPLE

There are several domain specific constraints in recognizing humans and animals, all of which suggest specialized grouping activities: humans and (many!) animals are made out of parts whose shape is relatively simple, implying specialized part groupers; there are few ways to assemble these parts into limbs and girdles, as the kinematics of the assembly ensures that many configurations of these parts are impossible, implying specialized limb and girdle groupers; and, when one can measure motion, the dynamics of these parts are limited, implying that motion information strongly constrains segmenting and identifying body parts.

However, clothed people are hard to segment, because clothing is often marked with complex colored patterns. Attempting to classify images based on whether they contain naked people or not provides a useful special case. We have built a system for telling whether an image contains naked people that: first locates images containing large areas of skin-like pixels (stuff processing); then, within these areas, finds elongated regions and groups them into possible human limbs and connected groups of limbs using a simplified kinematic model.

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. Detailed experimental data on very general images, demonstrate the approach is effective [5]. Fusing color, texture and shape in our paradigm makes a number of other application domains, including describing and detecting trees [3], practical. Other work in this paradigm involves detecting animals and describing human action by grouping motion-coherent regions (see `http://HTTP.CS.Berkeley.EDU/~bregler`).

5. 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 local measurements of color and texture coherence; as it proceeds more global and more specific models 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 three case studies that are prototype implementations of modules of such a grouping based recognition system.

Acknowledgments

We would like to thank R. Blasi and K. Murphy who collaborated in the work on learning decision trees for visual concept classification. This research was supported by NSF Digital Library award IRI-9411334

6. REFERENCES

[1] 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.*

[2] Pentland, A., Picard, R.W., and Sclaroff, S. (1993) "Photobook: content-based manipulation of image databases," MIT Media Lab Perceptual Computing T R No. 255.

[3] Forsyth, D. Malik, J., Fleck, M., Greenspan, H., Leung, T.,Belongie, S., Carson, C., and Bregler, C., "Finding Pictures of Objects in Large Collections of Images," UCBerkeley, CS Division Technical Report, CSD-96-905

[4] 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.

[5] 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.