
Computer Vision Tools for Finding Images and Video Sequences

D. A. FORSYTH

ABSTRACT

VERY LARGE COLLECTIONS OF IMAGES are now common. Indexing and searching such collections using indexing languages is difficult. Computer vision offers a variety of techniques for searching for pictures in large collections. Appearance methods compare images based on the overall content of the image using such criteria as similarity of color histograms, texture histograms, spatial layout, and filtered representations. Finding methods concentrate on matching subparts of images, defined in a variety of ways, in the hope of finding particular objects. These ideas are illustrated with a variety of examples from the current literature.

INTRODUCTION

Some of the image collections described in Enser (1995) contain tens of millions of items. Accessing a collection of pictures is difficult because it is hard to describe a picture accurately. Indexing a large collection by hand involves a substantial volume of work. Furthermore, there is the prospect of having to re-index sections of the collection—e.g., if a news event makes a previously unknown person famous, it would be beneficial to know if the collection contained pictures of that person. Finally, it is very difficult to know what a picture is about—e.g., “the engineer requests an image of a misaligned mounting bracket . . . which only exists as an image described by a cataloguer as astronaut training . . . ” (Seloff, 1990, p. 686). These observations indicate that it would be ideal to have automated tools that can describe pictures and find them based on a description. IBM’s seminal QBIC (Query by Image Content) system demonstrated

D. A. Forsyth, Computer Science Division, University of California at Berkeley, Berkeley, CA 94720

LIBRARY TRENDS, Vol. 48, No. 2, Fall 1999, pp. 326-355
© 1999 The Board of Trustees, University of Illinois

that such tools could be built (a comprehensive description of QBIC appears in the system). This review will indicate the different approaches that computer vision scientists have taken in building tools—rather than listing available tools—and indicate the kind of tasks these tools will facilitate. Some ideas from computer vision will be introduced, but readers who would like a deeper discussion of the topic should consult texts by Forsyth and Ponce (in press) and Trucco and Verri (1998).

WHAT DO USERS WANT?

The most comprehensive study of the behavior of users of image collections is Enser's work on the Hulton-Deutsch collection (Armitage & Enser, 1997; Enser, 1993, 1995). This is a collection of prints, negatives, slides, and the like used mainly by media professionals. Enser (1993) studied the request forms on which client requests are logged; he classified requests into four semantic categories, depending on whether a unique instance of an object class is required or not and whether that instance is refined. Significant points are that the specialized indexing language used gives only a "blunt pointer to regions of the Hulton collections" (p. 35) and the broad and abstract semantics used to describe images. For example, users requested images of hangovers, physicists, and the smoking of kippers. All these concepts are well beyond the reach of current image analysis techniques. As a result, there are few cases where one can obtain a tool that directly addresses a need. For the foreseeable future, the main constraint on the design of tools for finding images will be our limited understanding of vision.

The Hulton-Deutsch collection is used largely by book, magazine, and newspaper publishers; Enser suggests that the most reliable measure of the success of their indexing system is that the organization is profitable. This author is unaware of any specific areas of computer vision technology that are profitable, though Virage appears to be thriving (the company is described at: <http://www.virage.com>. A description of their technology appears in Hampapur et al., 1997). The main source of value in any large collection is being able to find items. Potential application areas include:

- *military intelligence*: vast quantities of satellite imagery of the globe exist, and typical queries involve finding militarily interesting changes—e.g., concentrations of force occurring at particular places (e.g., Mundy, 1995, 1997; Mundy & Vrobel, 1994).
- *planning and government*: satellite imagery can be used to measure development, changes in vegetation, regrowth after fires, and so on (see, for example, Smith, 1996).
- *stock photo and stock footage*: commercial libraries—which often have extremely large and very diverse collections—sell the right to use particular images (e.g., Armitage & Enser, 1997; Enser, 1993, 1995).

- *access to museums*: museums are increasingly releasing collections, typically at restricted resolutions, to entice viewers into visiting the museum (e.g., Holt & Hartwick, 1994a, 1994b; Psarrou et al., 1997).
- *trademark enforcement*: as electronic commerce goes, so does the opportunity for automatic searches to find violations of trademark (e.g., Eakins et al., 1998; Jain & Vailaya, 1998).
- *indexing the Web*: indexing Web pages appears to be a profitable activity. Users may also wish to have tools that allow them to avoid offensive images or advertising. A number of tools have been built to support searches for images on the Web using techniques described later in this article (e.g., Cascia et al., 1998; Chang et al., 1997b; Smith & Chang, 1997).
- *medical information systems*: recovering medical images "similar" to a given query example might give more information on which to base a diagnosis or to conduct epidemiological studies (e.g., Congiu et al., 1995; Wong, 1998).

WHAT CAN TOOLS DO?

There are two threads discernible in current work on finding images: one can search for images based either on the appearance of the whole image or on object-level semantics.

The central technical problem in building appearance tools is defining a useful notion of image similarity; the section of this article entitled "Appearance" illustrates a variety of different strategies. Image collections are often highly correlated so that it is useful to combine appearance tools with browsing tools to help a user browse a collection in a productive way—i.e., by trying to place images that are "similar" near to one another and offering the user some form of dialogue that makes it possible to move through the collection in different "directions." Building useful browsing tools also requires an effective notion of image similarity. Constructing a good user interface for such systems is difficult. Desirable features include a clear and simple query specification process, and a clear presentation of the internal workings of the program so that failures can be resolved easily. Typically, users are expected to offer an example image or to fill in a form-based interface.

It is very difficult to cope with high-level semantic queries ("a picture of the Pope kissing a baby") using appearance or browsing tools. Finding tools use elements of the currently limited understanding of object recognition to help a user query for images based on this kind of semantics at a variety of levels. It is not known how to build finding tools that can handle high-level semantic queries, nor how to build a user interface for a general finding tool; nonetheless, current technology can produce quite useful tools (see the later section entitled Finding).

APPEARANCE

Images are often highly stylized, particularly when the intent of the artist is to emphasize a particular object or a mood. This means that the overall layout of an image can be a guide to what it depicts so that useful query mechanisms can be built by searching for images that “look similar” to a sample image, a sketched sample, or textual specification of appearance. The success of such methods rests on the sense in which images look similar. A good notion of similarity is also important for efficient browsing, because a user interface that can tell how different images are, can lay out a display of images to suggest the overall structure of the section of the collection being displayed. I will concentrate on discussing appearance matching rather than browsing because of the similarity of the technical issues.


The simplest form of similarity—two pictures are similar if all their pixels have similar values—is extremely difficult to use because it requires users to know exactly how a picture is laid out. For example, if one were trying, with a sketch, to recover the Van Gogh painting of sunflowers on a table, it would be necessary to remember on which side of the table the vase of flowers lay. It is possible to build efficient systems that use this iconic matching (e.g., Jacobs et al., 1995), but the approach does not seem to have been adopted, probably because users generally are unable to remember enough about the picture they want to find. It is important to convey to the user the sense in which images look similar, because otherwise mildly annoying errors can become extremely puzzling.

Histograms and Correlograms

A popular measurement of similarity compares counts of the number of pixels in particular color categories. For example, a sunset scene and a pastoral scene would be very different by this measure because the sunset scene contains many red, orange, and yellow pixels and the pastoral scene will have a preponderance of green (grass), blue (sky), and perhaps white (cloud) pixels (see Figure 1). Furthermore, sunset scenes will tend to be similar; all will have many red, orange, and yellow pixels and few others. This color histogram matching has been extremely popular; it dates back at least to Swain and Ballard (1991) and has been used in a number of systems (Flickner et al., 1996; Holt & Hartwick, 1994b; Ogle & Stonebraker, 1995). The usefulness of color histograms is slightly surprising, given how much image information the representation discards; Chapelle, Haffner, and Vapnik (1999) show that images from the Corel collection¹ can be classified by their category in the collection using color histogram information alone.

There is no record in a color histogram of where colored pixels are with respect to one another. Thus, for example, pictures of the French and British flags are extremely similar according to a color histogram

measure—each has red, blue, and white pixels in about the same number; it is the spatial layout of the pixels that differs. One problem that can result is that pictures taken from slightly different viewing positions look substantially different by a color histogram measure (see Figure 2). This effect can be alleviated by considering the probability that a pixel of some color lies within a particular pixel of another color (which can be measured by counting the number of pixels at various distances). Requiring this measure to be similar provides another measure of image similarity. Computational details are important because the representation is large (Huang et al., 1997; Huang & Zabih, 1998). For small movements of the camera, these probabilities will be largely unchanged so that similarity between these color correlograms yields a measure of similarity between images.



All Images
Berkeley Digital Library Project

Number of matches to your query: 22
Here are matches 1 through 20.

Use the Next and Previous buttons to see more.

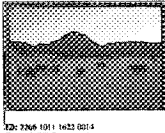
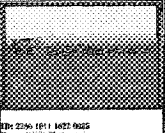


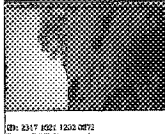

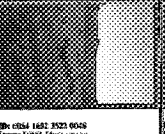



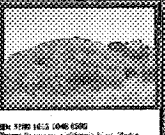

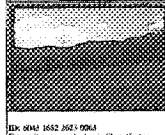







			
ID: 2768 1011 1423 0014 From: 1973, E. J. Smith	ID: 2768 1011 1423 0022 From: 1973, E. J. Smith	ID: 2768 1011 1423 0040 From: 1973, E. J. Smith	ID: 2768 1011 1423 0015 From: 1973, E. J. Smith
			
ID: 2317 0521 1202 0072 From: 1973, E. J. Smith	ID: 2317 0521 1202 0080 From: 1973, E. J. Smith	ID: 0254 1482 2523 0040 From: 1967, E. J. Smith	ID: 0439 1461 2523 0039 From: 1967, E. J. Smith
			
ID: 5185 1412 1040 0249 From: 1967, E. J. Smith	ID: 5186 1412 1040 0295 From: 1967, E. J. Smith	ID: 5780 1613 1040 0280 From: 1967, E. J. Smith	ID: 5210 1412 1040 0230 From: 1967, E. J. Smith
			
ID: 0041 1452 2051 0004 From: 1973, E. J. Smith	ID: 0040 1451 1751 0011 From: 1973, E. J. Smith	ID: 0189 1451 0932 0001 From: 1967, E. J. Smith	ID: 0480 1451 0932 0024 From: 1967, E. J. Smith
			
ID: 5181 1411 1041 0234 From: 1967, E. J. Smith	ID: 1909 1078 0041 0001 From: 1967, E. J. Smith	ID: 5207 1412 1040 0280 From: 1967, E. J. Smith	ID: 5207 1412 1040 0280 From: 1967, E. J. Smith

Figure 1. Results from a query to the Calphotos collection that sought pastoral scenes, composed by searching for images that contain many green and light blue pixels. As the results suggest, such color histogram queries can be quite effective.

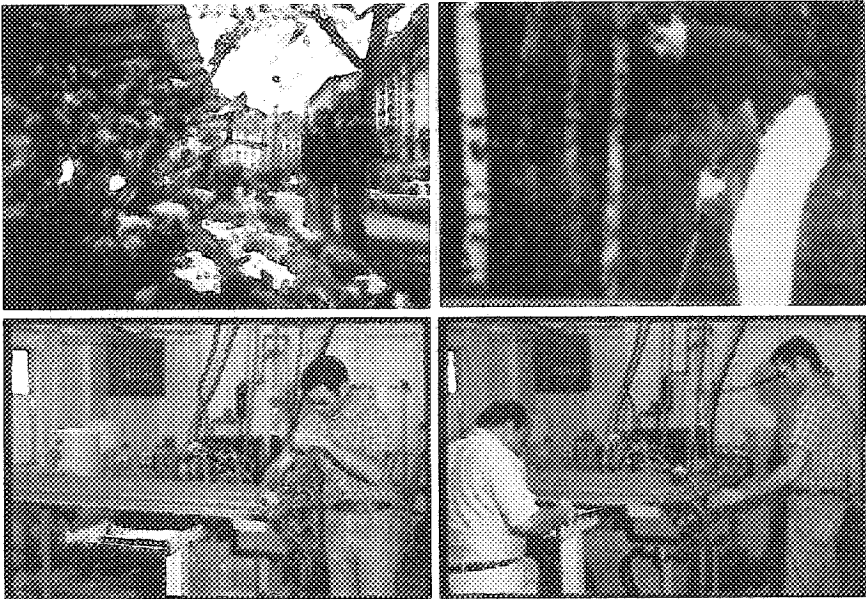


Figure 2. The top two figures have the same color histogram; the red patch on the golfer's shirt in the original print appears in the other image as brown looking flowers. These flowers were darker in color than they appear (their hue is changed somewhat by the fact that they are small and don't contrast strongly with their background), although not quite as dark in color as the shirt. However, in the scheme of color categories used, the colors are regarded as equivalent. The bottom two figures have similar content, but quite different color histograms; one person has a peach shirt and appears in only one figure and the one person wearing blue occupies more space in one than in the other. However, if one looks at a representation of the extent to which pixels of a given color lie near pixels of some other color, the pictures are quite similar—many blue pixels lie near either blue pixels, green ones, or white ones. This information is captured by the color correlogram. More information on color correlograms can be found in Huang and Zabih (1998). (Picture reproduced by kind permission of R. Zabih.)

TEXTURES

Color histograms contain no information about the layout of color pixels. An explicit record of layout is the next step. For example, a snowy mountain image will have bluer regions on top, whiter regions in the middle, then a bluer region at the bottom (the lake at the foot of the mountain), whereas a waterfall image will have a darker region on the left and right and a lighter vertical stripe in the center. These layout templates can be learned for a range of images and appear to provide a significant improvement over a color histogram (Lipson et al., 1997).

Looking at image texture is a natural next step, because texture is the difference between, for example, a field of flowers (many small orange blobs), a single flower (one big orange blob), or a dalmatian and a zebra. Most people know texture when they see it, though the concept is either difficult or impossible to define. Typically, textures are thought of as spatial arrangements of small patterns—e.g., a tartan is an arrangement of small squares and lines, and the texture of a grassy field is an arrangement of thin bars.


The usual strategy for finding these subpatterns is to apply a linear filter to the image where the kernel of the filter looks similar to the pattern element. From filter theory, we have that strong responses from these filters suggest the presence of the particular pattern; several different filters can be applied, and the statistics of the responses in different places then yield a decomposition of the picture into spotty regions, barred regions, and the like (Ma & Manjunath, 1997a; Malik & Perona, 1989, 1990).

A histogram of filter responses is a first possible description of texture. For example, one might query for images with few small yellow blobs. This mechanism is used quite successfully in the Calphotos collection at Berkeley.² As Figures 3, 4, and 5 illustrate, a combination of color and blob queries can be used to find quite complex images. The system is described in greater detail in Carson and Ogle (1996).

Texture histograms have some problems with camera motion; as the person in Figure 2 approaches the camera, the checks on his shirt get bigger in the image. The pattern of texture responses could change quite substantially as a result. This is another manifestation of a problem we saw earlier with color histograms—i.e., the size of a region spanned by an object changes as the camera moves closer to, or further from, the object.

A strategy for minimizing the impact of this effect is to define a family of allowable transformations on the image—e.g., scaling the image by a factor in some range. We now apply each of these transformations and measure the similarity between two images as the smallest difference that can be obtained using a transformation. For example, we could scale one image by each legal factor and look for the smallest difference between color and texture histograms. In Figure 7, the earth-mover's distance allows a wide variety of transformations. Furthermore, in Rubner, Tomasi, and Guibas (1998), it has been coupled with a process for laying out images that makes the distance between images in the display reflect the dissimilarity between images in the collection. This approach allows for rapid and intuitive browsing.

The spatial layout of textures is a powerful cue. For example, in aerial images, housing developments have a fairly characteristic texture, and the layout of this texture gives cues to the region sought. In the Netra system, textures are classified into stylized families (yielding a "texture thesaurus") which are used to segment very large aerial images; this



Query All Images

Berkeley Digital Library Project

This form issues content-based queries to a collection of over 50,000 images. The SQL query that was generated will be shown at the bottom of each page of pictures. For more information about the image analysis techniques that were used, see [Computer Vision Resources](#).

• [Home](#) • [Sample Queries](#)

Number of Photos to Display Per Page:

Show text ... Show blob info ...

Horizon? Yes No Text:

Things: For more info, see [Finding forms, text, and photos](#)


Collection: any air photos coral flowers habitats DWR

Color Percentages												Amount		
												Am	Partly	Mostly
OR	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
AND														
OR	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Colored Blobs												Size			Quantity					
												Any	S	M	L	Any	Few	Some	Many	
OR	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
AND																				
OR	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	

berkeley.digital.library/1/w/w3/dlhas/berkeley.asp


Figure 3. Specifying a query to the Calphotos collection using color and texture information. I have selected images that have a horizon and some red or yellow blobs with the intention of finding images of fields of flowers. The results appear in Figure 4.



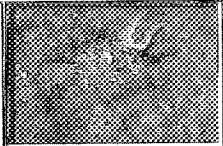
All Images

Berkeley Digital Library Project

Number of matches to your query: 2



ID: 2317 1021 1202 0041
From: DWR, Parangalita



ID: 5208 1611 1829 0072
From: Pomona, California, Eliza Prang

SQL: SELECT * FROM allimg WHERE horizon AND (tablename = 'flowers' OR tablename = 'habitats' OR tablename = 'photos') AND ((blob like '%yel_dot%' and (substring(blob from (position('yel_dot' in blob) + 9) for 2) > 20)) OR (blob like '%yel_dot%' and (substring(blob from (position('yel_dot' in blob) + 9) for 2) > 15))) OR ((blob like '%red_dot%' and (substring(blob from (position('red_dot' in blob) + 9) for 2) > 20)) OR (blob like '%red_dot%' and (substring(blob from (position('red_dot' in blob) + 9) for 2) > 15)))

Note: Care: images are for viewing only and may not be downloaded or saved.

Search All Images | U.C. Berkeley Digital Library | Comments? www.dlhas.berkeley.edu

Figure 4. Images obtained from the Calphotos collection using the query from Figure 3.



All Images
Berkeley Digital Library Project

Number of matches to your query: 229
Here are matches 1 through 20.

Next Use the Next and Previous buttons to see more.

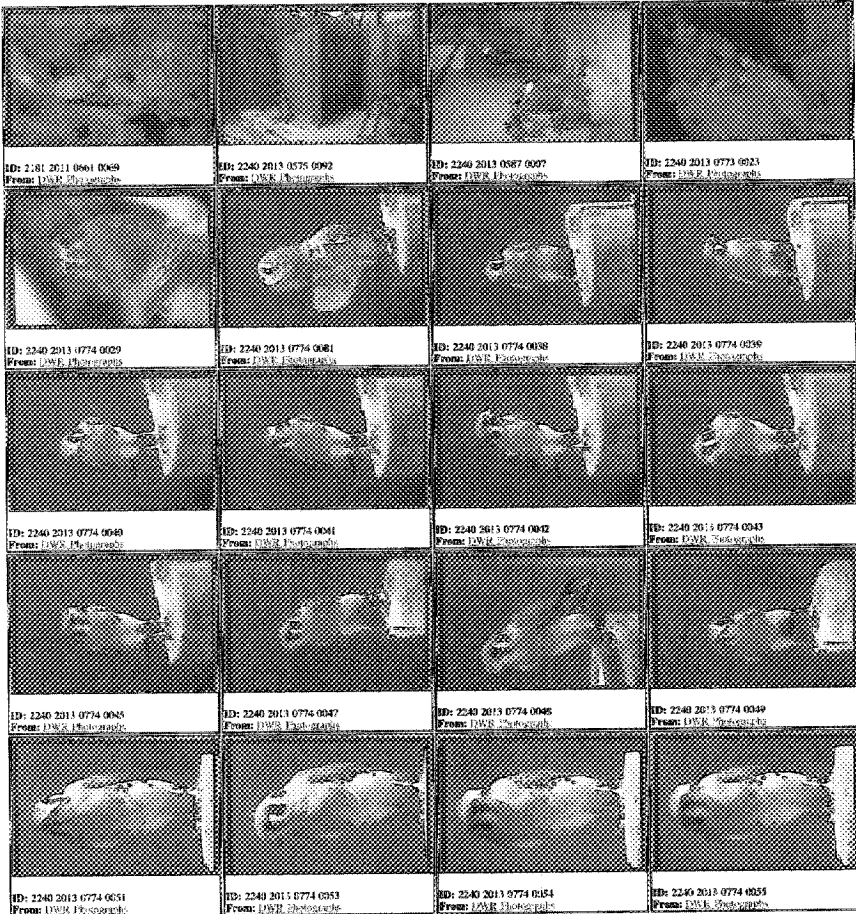


Figure 5. Response images obtained by querying the Calphotos collection for images with at least one large brown blob, one or more small black blobs, and some green; this query is intended to find animals or birds.

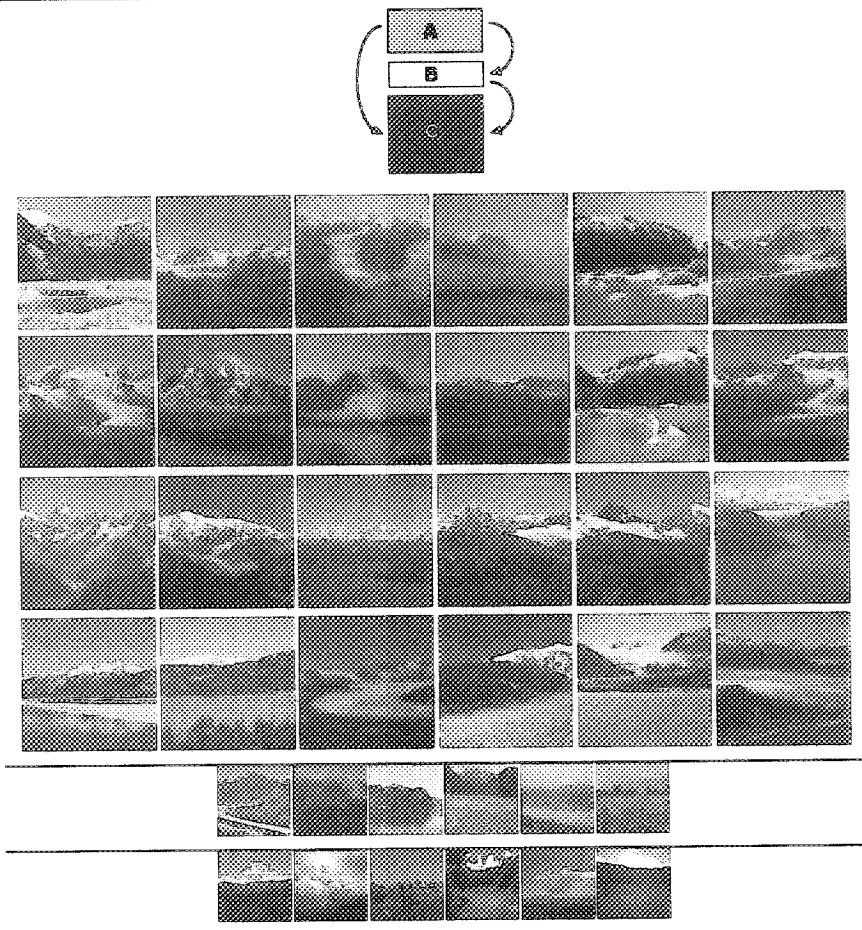


Figure 6. Spatial layout of colored regions is a natural guide to the content of many types of image. The figure on the top shows a layout of colored regions that suggests a scene showing snowy mountains; bottom center, views of mountains that were in the collection but not recovered, and bottom, images that meet the criterion but do not actually show a view of a snowy mountain. More detail appears in Lipson et al., 1997. (Figure reproduced by kind permission of W. E. L. Grimson and P. Lipson.)

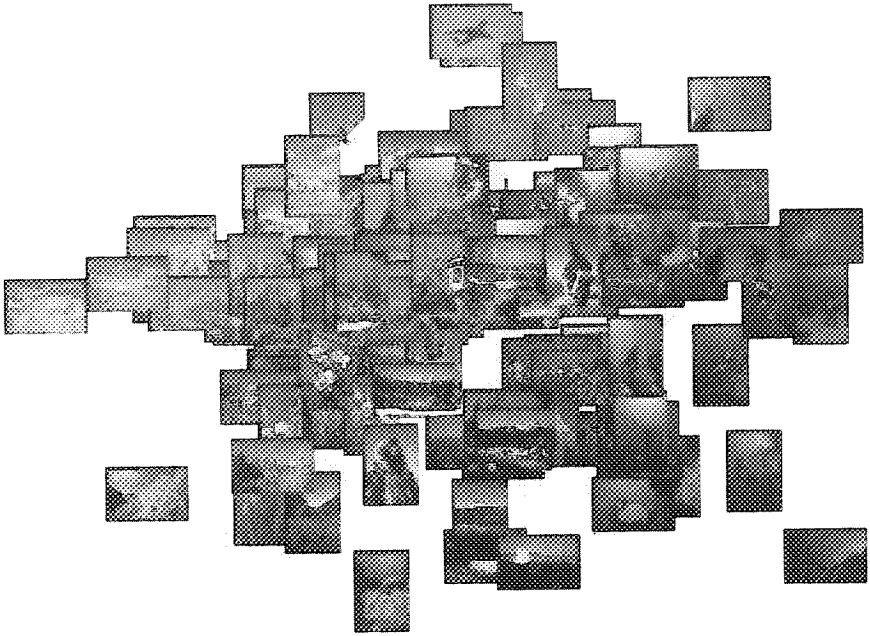


Figure 7. Images laid out according to their similarity using the earth mover's distance (EMD). The EMD can be computed quickly so that displays like this—where distances between images on the display reflect the EMDs between them as faithfully as possible—can be created online. Large numbers of pictures returned from a query into an image database can thus be viewed at a glance, and a mouse click in the neighborhood of pictures that look similar to what the user is looking for tells the retrieval system where to search next. With this technology, users browse and navigate in an image database just as they would browse through a department store. Because of the large number of images displayed, and their spatially intuitive layout, users quickly form a mental model of what is in the database, and rapidly learn where to find the pictures they need. More information on the EMD can be found in Rubner et al., 1998. (Figure reproduced by kind permission of C. Tomasi.)

approach exploits the fact that, while there is a very large family of possible textures, only some texture distinctions are significant. Users can utilize example regions to query a collection for similar views—e.g., obtaining aerial pictures of a particular region at a different time or date to keep track of such matters as the progress of development, traffic patterns, or vegetation growth (see Figure 8) (Ma & Manjunath, 1997a, 1998; Manjunath, 1997a; Manjunath & Ma, 1996a, 1996b).

Regions of texture responses form patterns, too. For example, if an image shows a pedestrian in a spotted shirt, then there will be many strong responses from spot detecting filters; the region of strong responses will look roughly like a large bar. A group of pedestrians in spotted shirts will look like a family of bars, which is itself a texture. These observations suggest applying texture-finding filters to the outputs of texture-finding filters—perhaps recurring several times—and using these responses as a measure of image similarity. This strategy involves a large number of features, making it impractical to ask users to fill in a form as in Figure 9. Instead, DeBonet and Viola (1998) use an approach in which users select positive and negative example images, and the system searches for images that are similar to the positive examples and dissimilar to the negative ones.

FINDING

The distinction between appearance tools and finding tools is somewhat artificial. Based on their differences in appearance, we can tell what objects are. The tools described in this section try to estimate object-level semantics more or less directly. Such systems must first segment the image—i.e., decide which pixels lie on the object of interest. Template matching systems then look for characteristic patterns associated with particular objects. Finally, correspondence reasoning can be used to identify objects using spatial relationships between parts.

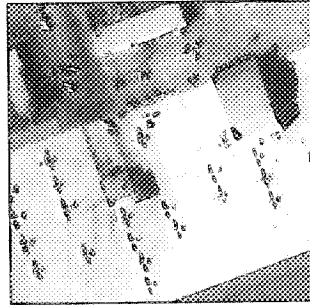
Structure in a collection is helpful in finding semantics because it can be used to guide the choice of particular search mechanisms. Photobook is a system that provides three main search categories: by shape (searches for isolated objects—e.g., tools or fishes) using contour shape measured as elastic deformations of a contour; by appearance (the program can find faces using a small number of principal components); and texture (the program uses a texture representation to find textured swatches of material) (Pentland et al., 1996).

Annotation and Segmentation

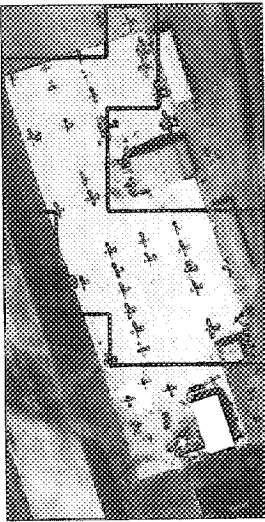
A natural step in determining image semantics is to classify the type of material that image patches represent—e.g., “sky,” “buildings,” and so on, as opposed to “blue” or “grey.” Generally, this kind of classification



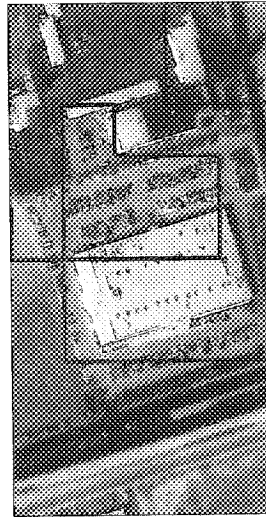
(a)



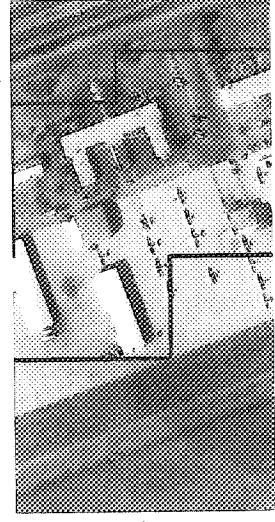
(b)



(c)



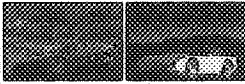
(d)



(e)

Figure 8. A texture-based search in an aerial image: (a) shows the down-sampled version of the aerial photograph from which the query is derived; (b) shows full-resolution detail of the region used for the query. The region contains aircraft, cars, and buildings. (c)-(e) show the ordered three best results of the query. Once again, the results come from three different aerial photographs. This time, the second and third results are from the same year (1972) as the query photograph but the first match is from a different year (1966). More details appear in Ma and Manjunath, 1997b. (Figure reproduced by kind permission of B. S. Manjunath.)

POSITIVE EXAMPLES



Eliminate Eliminate

NEGATIVE EXAMPLES

TOP 24 RETRIEVED IMAGES

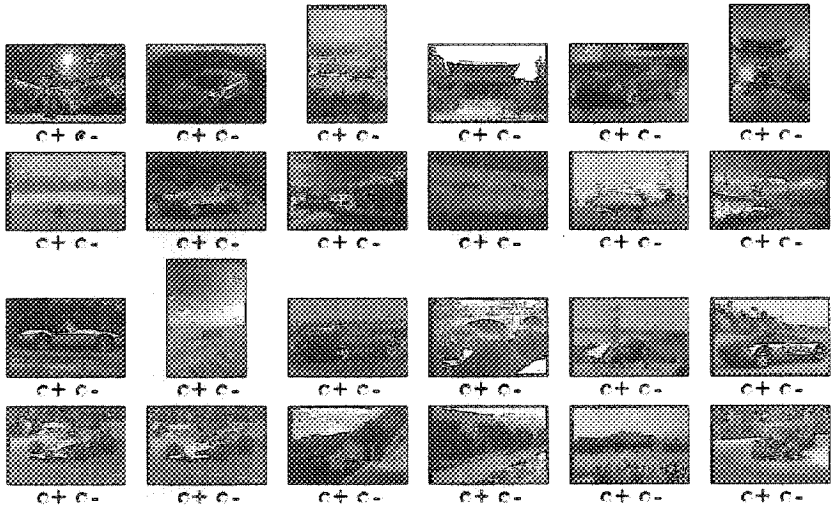
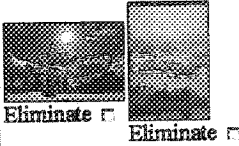


Figure 9. Querying using the "texture of textures" approach. The user has identified two pictures of cars as positive examples; these would respond strongly to large horizontal bar filters, among others. This query results in a number of returned images containing several images of cars. Figure 10 shows the effect of refining this query by exhibiting negative examples. (Figure reproduced by kind permission of P. Viola.)

POSITIVE EXAMPLES



NEGATIVE EXAMPLES



TOP 24 RETRIEVED IMAGES

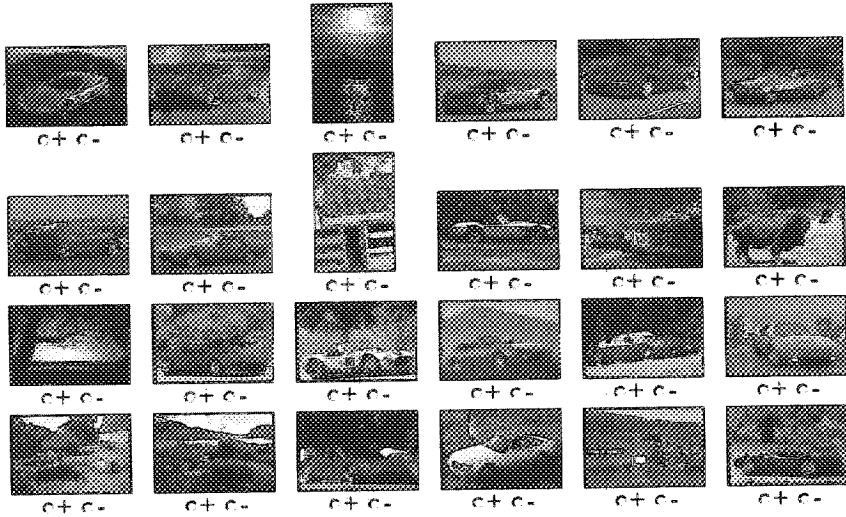


Figure 10. Querying using the "texture of textures" approach. The query has been refined by providing some negative examples, yielding a response set that contains more car images. More detail on this approach appears in De Bonet & Viola, 1998. (Figure reproduced by kind permission of P. Viola.)

would need to be done with a mixture of user input (to establish appropriate categories and provide examples from those categories) and automatic annotation (for speed and efficiency). A combination of color and texture features is often, but not always, distinctive of a region; a particular difficulty is knowing which features to use and which to ignore in classifying an image patch. For example, telling sky from grass involves looking at color; telling concrete from sky may require ignoring color and emphasizing texture. Foureyes (see Figure 12) uses techniques from machine learning to infer appropriate features from user annotation practices using across-image groupings (which image patches tend to be grouped together and which apart) and in-image groupings (which patches are classified as "sky" in this image) (Minka & Picard, 1997; Picard & Minka, 1995; Minka, 1996). As a result, a user annotating an image can benefit from past experience, as illustrated in Figure 12.

Humans decompose images into parts corresponding to the objects we are interested in, and classification is one way to achieve this segmentation. Segmentation is crucial because it means that irrelevant information can be discarded in comparing images. For example, if we are searching for an image of a tiger, it should not matter whether the background is snow or grass; the tiger is the issue. However, if the whole image is used to generate measures of similarity, a tiger on grass will look very different from a tiger on snow. These observations suggest segmenting an image into regions of pixels that belong together in an appropriate sense, and then allowing the user to search on the properties of particular regions. The most natural sense in which pixels belong together is that they originate from a single object; currently, it is almost never possible to use this criterion because of an inability to know when this is the case. However, objects usually result in image regions of coherent color and texture so that pixels that belong to the same region have a good prospect of belonging to the same object.

VisualSEEK automatically breaks images into regions of coherent color and allows users to query on the spatial layout and extent of colored regions. Thus, a query for a sunset image might specify an orange background with a yellow "blob" lying on that background (Smith & Chang, 1996).

Blobworld is a system that represents images by a collection of regions of coherent color and texture (Belongie et al., 1998; Carson, Thomas, Belongie, Hellerstein, & Malik, 1999; Carson et al., 1997). The representation is displayed to the user with region color and texture displayed inside elliptical blobs that represent the shape of the image regions. The shape of these regions is represented crudely since details of the region boundaries are not cogent. A user can query the system by specifying which blobs in an example image are important and what spatial relations should hold (see Figure 11).

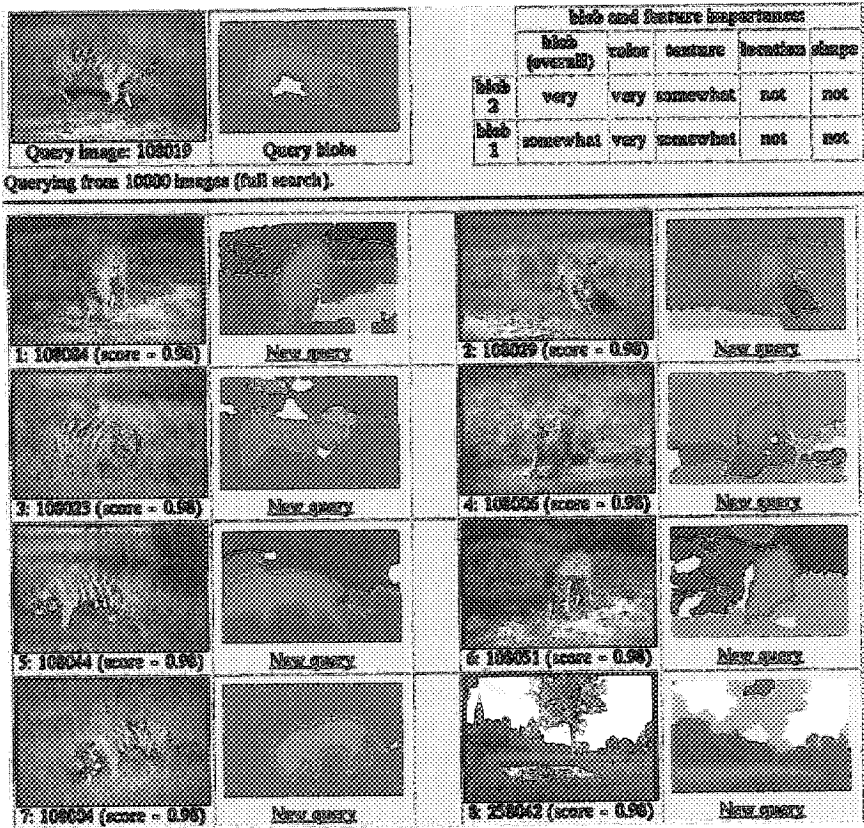


Figure 11. Blobworld query for tiger images. Users of image databases generally want to find images containing particular objects, not images with particular global statistics. The Blobworld representation facilitates such queries by representing each image as a collection of regions (or "blobs") which correspond to objects or parts of objects. The image is segmented into regions automatically, and each region's color, texture, and shape characteristics are encoded. The user constructs a query by selecting regions of interest. The Blobworld version of each retrieved image is shown, with matching regions highlighted; displaying the system's internal representation in this way makes the query results more understandable and aids the user in creating and refining queries. Experiments show that queries for distinctive objects, such as tigers and cheetahs, have much higher precision using the Blobworld system than using a similar system based only on global color and texture descriptions. Blobworld is described in greater detail in Belongie et al., 1998; Carson et al., 1999. (Figure reproduced by kind permission of C. Carson.)

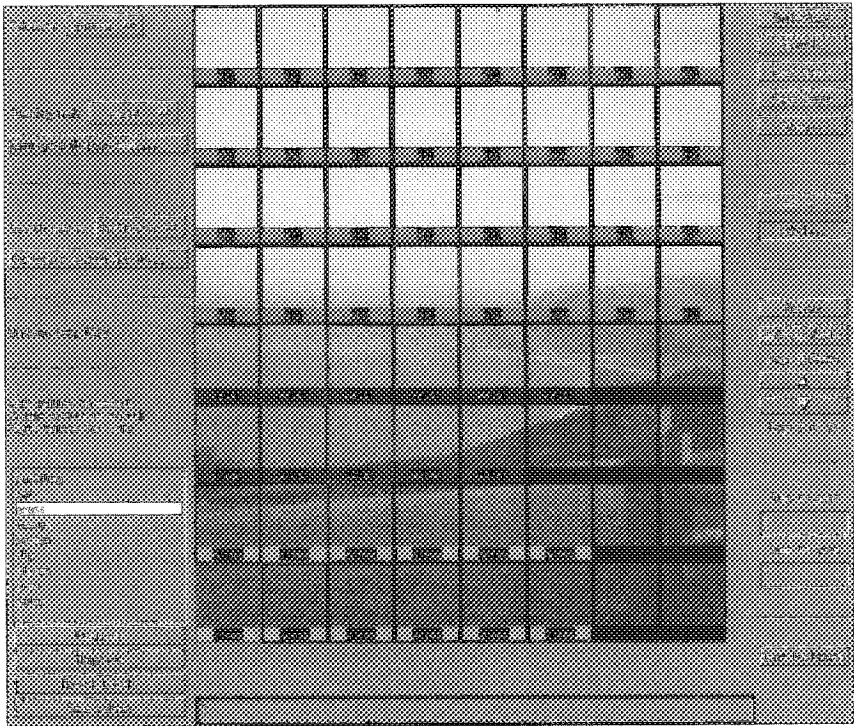


Figure 12. An image annotated with Foureyes. Red patches of image have been explicitly labeled "sky," "grass," or "water." Other labels are inferred by Foureyes from previous annotations and from the information given in these examples. More information appears in Minka and Picard, 1997. (Figure reproduced by kind permission of R. Picard.)

Template Matching

Some objects have such a distinctive appearance that they have a wide range of viewing directions and conditions. Template matching is an object recognition strategy that finds objects by matching image patches with example templates. A natural application of template matching is to construct whole-image templates that correspond to particular semantic categories (see Figure 13) (Chang et al., 1998b). These templates can be constructed offline and used to simplify querying by allowing a user to apply an existing template rather than compose a query.

Face finding is a particularly good case for template matching. Frontal views of faces are extremely similar, particularly when the face is viewed at low resolution—the main features are a dark bar at the mouth; dark blobs where the eyes are; and lighter patches at the forehead, nose, and near the mouth. This means that faces can be found, independent of the identity of the person, by looking for this pattern. Typical face finding systems extract small image windows of a fixed size, prune these windows to be oval, correct for lighting across the window, and then use a learned classifier to tell whether a face is present in the window (see Figure 14) (Rowley et al., 1996a, 1996b, 1998a; Poggio & Sung, 1995). This process works for both large and small faces because windows are extracted from images at a variety of resolutions (windows from low resolution images yield large faces and those from high resolution images yield small faces). Because the pattern changes when the face is tilted to the side, this tilt must be estimated and corrected for; this is done using a mechanism learned from data (Rowley et al., 1998b). Knowing where the faces are is extremely useful because many natural queries refer to the people present in an image or a video.

Shape and Correspondence

If object appearance can vary, template matching becomes more difficult as one is forced to adopt many more templates. There is a good template matching system for finding pedestrians, which appears to work because images of pedestrians tend to be seen at low resolution and with their arms at their sides (Oren et al., 1997). However, building a template matching system to find people is intractable because clothing and configuration can vary widely. The general strategy for dealing with this difficulty is to look for smaller templates—perhaps corresponding to “parts”—and then look for legal configurations.

One version of this technique involves finding “interest points”—points where combinations of measurements of intensity and its derivatives take on unusual values—e.g., at corners. The spatial arrangement of these points is quite distinctive in many cases. For example, as Figure 15 illustrates, the arrangement of interest points in an aerial view of Marseille is unaffected by the presence of cars; this means that one can recover and

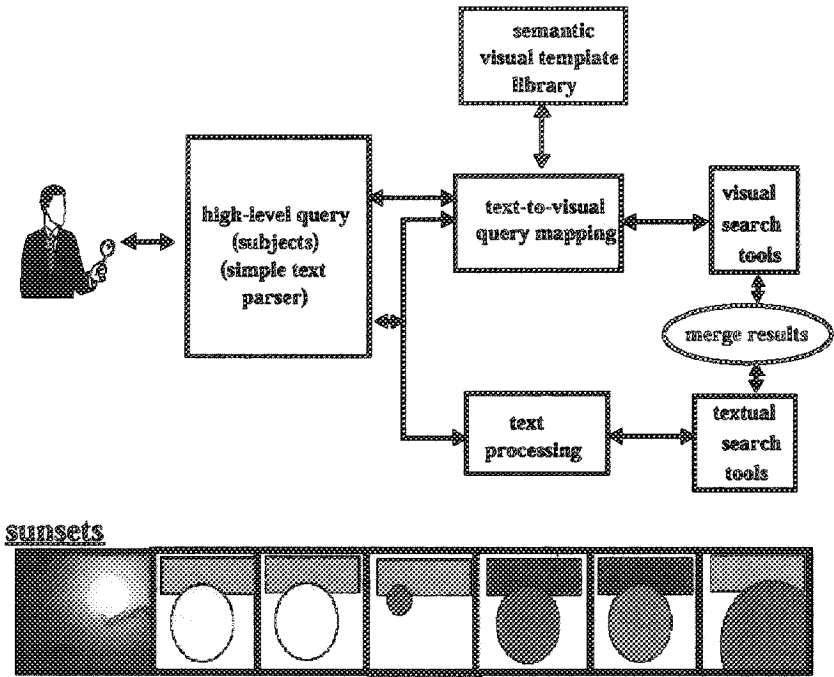


Figure 13. To remove the burden of drawing detailed low-level sketches from users, the Semantic Visual Template system helps users to develop personalized search templates. The library of semantic templates can then be used to assist users in high-level multimedia query. The top figure shows the overall structure of a system using semantic templates, and the lower figure shows a template for sunset images. More details appear in Chang et al., 1998b. (Figure reproduced by kind permission of S-F. Chang.)



Figure 14. Faces found by a learned template matching approach; the eye holes in the green mask are to indicate the orientation of the face. More details appear in Rowley et al., 1996b, 1998a, 1998b. (Figure reproduced by kind permission of T. Kanade.)

register aerial images of the same region taken at different times of day using this technique. Furthermore, once interest points have been matched, an image-image transformation is known, which can be used to register the images. Registration yields further evidence to support the match and can be used to compare, for example, traffic by matching the two images at specific points.

This form of correspondence reasoning extends to matching image components with object parts at a more abstract level. People and many animals can be thought of as assemblies of cylinders (corresponding to body segments). A natural finding representation uses grouping stages to assemble image components that could correspond to appropriate body segments or other components.

This representation has been used for two cases. The first example identifies pictures containing people wearing little or no clothing. This is an interesting example: first, it is much easier than finding clothed people because skin displays very little variation in color and texture in images, whereas the appearance of clothing varies widely; second, many people are interested in avoiding or finding images based on whether they contain unclad people. This program has been tested on an unusually large and unusually diverse set of images; on a test collection of 565 images known to contain lightly clad people and 4,289 control images with widely varying content, one tuning of the program marked 241 test images and 182 control images (more detailed information appears in Forsyth et al., 1996; Forsyth & Fleck, 1996). The second example used a representation whose combinatorial structure—the order in which tests were applied—was built by hand, but the tests were learned from data. This program identified pictures containing horses and is described in greater detail in Forsyth and Fleck (1997). Tests used 100 images containing horses and 1,086 control images with widely varying content—for a typical configuration, the program marks eleven images of horses and four control images.

VIDEO

While video represents a richer source of information than still images, the issues remain largely the same. Videos are typically segmented into shots—short sequences that contain similar content—and techniques of the form described applied within shots. Because a large change between frames is a strong cue that a shot boundary has been reached, segmenting video into shots is usually done using measures of similarity like those described in the earlier section on Appearance (e.g., Boreczky & Rowe, 1996).

The motion of individual pixels in a video is often called *optic flow* and is measured by attempting to find pixels in the next frame that correspond to a pixel in the previous frame (correspondence being measured by similarity in color, intensity, and texture). In principle, there is an

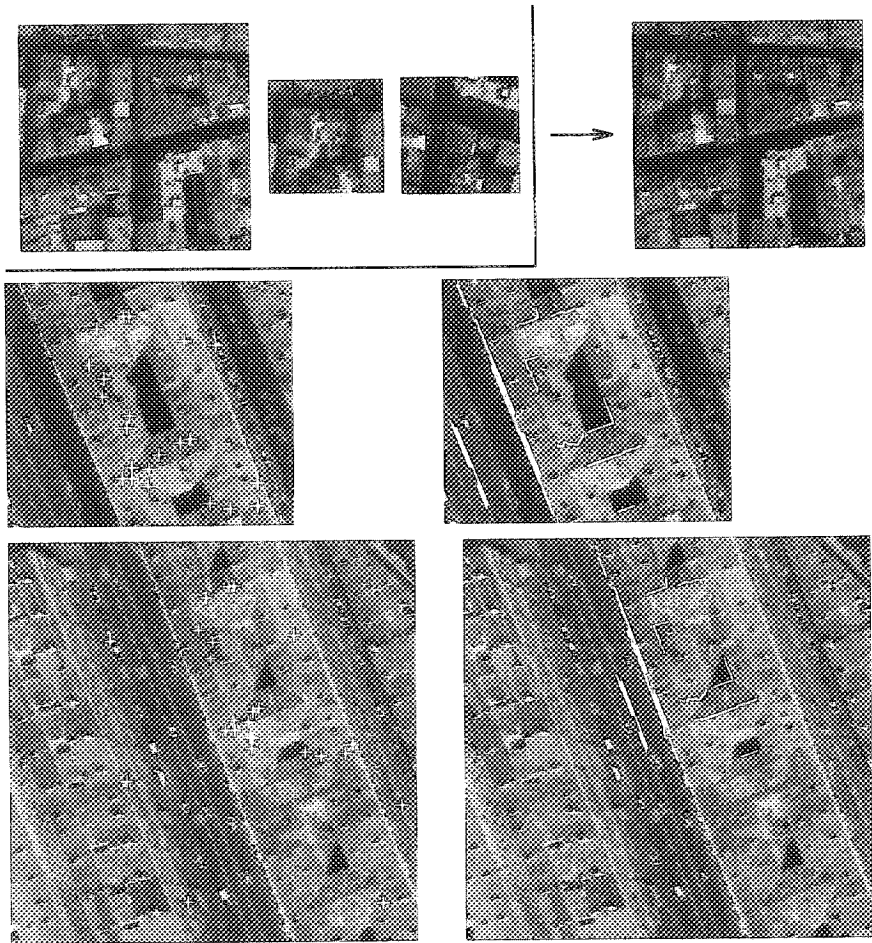


Figure 15. Images can be queried by detecting "interest points" on the image and then matching on configurations of these points based on their geometric distribution and the grey level pattern surrounding each point. The matching process is very efficient (it uses "indexing") and is tolerant of missing points in the configuration. In the example shown here, the image on the top right can be correctly retrieved from a collection of paintings, aerial images, and images of 3D objects using any of the images on the top left. Interest points used during the matching process, indicated by white crosses, for a query image (small inset on left) and the best match (bottom left). Additional evidence is obtained from the image-image transformation to confirm that the match is correct; on the right, edges which match under this transformation in the query (inset) and the result (bottom right). Notice that the two images have been taken from different viewpoints so that the building's shape differs between images. Also the scenes are not identical because cars have moved. Further details are given in Schmid and Mohr, 1997; Schmid et al., in press. (Figure reproduced by kind permission of A. Zisserman.)

optic flow vector at each pixel forming a *motion field*. In practice, it is extremely hard to measure optic flow reliably with featureless pixels because they could correspond to almost anything. For example, consider the optic flow of an egg rotating on its axis; there is very little information about what the pixels inside the boundary of the egg are doing because each looks like the other.

Motion fields can be extremely complex, however. If there are no moving objects in the frame, it is possible to classify motion fields corresponding to the camera shot used. For example, a pan shot will lead to strong lateral motion, and a zoom leads to a radial motion field. This classification is usually obtained by comparing the measured motion field with a parametric family (e.g., Sawhney & Ayer, 1996; Smith & Kanade, 1997).

Complex motion sequences are difficult to query without segmentation because much of the motion may be irrelevant to the query—e.g., in a soccer match, the motion of many players may not be significant. In VideoQ, motion sequences are segmented into moving blobs and then queried on the color and motion of a particular blob (see Figure 17) (Chang et al., 1997a; Chang et al., 1998).

The Informedia project at CMU has studied the preparation of detailed skims of video sequences. In this case, a segment of video is broken into shots, shots are annotated with the camera motion in shot, with the presence of faces, with the presence of text in shot, with keywords from the transcript, and with audio level (see Figure 18). This information yields a compact representation—the “skim”—which gives the main content of the video sequence (details in Smith & Kanade, 1997; Wactlar et al., 1996; Smith & Christel, 1995; Smith & Hauptmann, 1995).

CONCLUSION

For applications where the colors, textures, and layout of the image are all strongly correlated with the kind of content desired, a number of usable tools exist to find images based on content.

There has been a substantial amount of work on user interfaces and browsing, although this work is usually done to get a system up and running. Because color, texture, and layout are at best a rough guide to image content, puzzling search results are usually guaranteed. There is not yet a clear theory of how to build interfaces that minimize the impact of this effect. The most widely adopted strategy is to allow quite fluid browsing.

When queries occur at a more semantic level, we encounter deep and poorly understood problems in object recognition. Object recognition seems to require segmenting images into coherent pieces and reasoning about the relationships between those pieces. This rather vague view of recognition can be exploited to produce segmented representations that

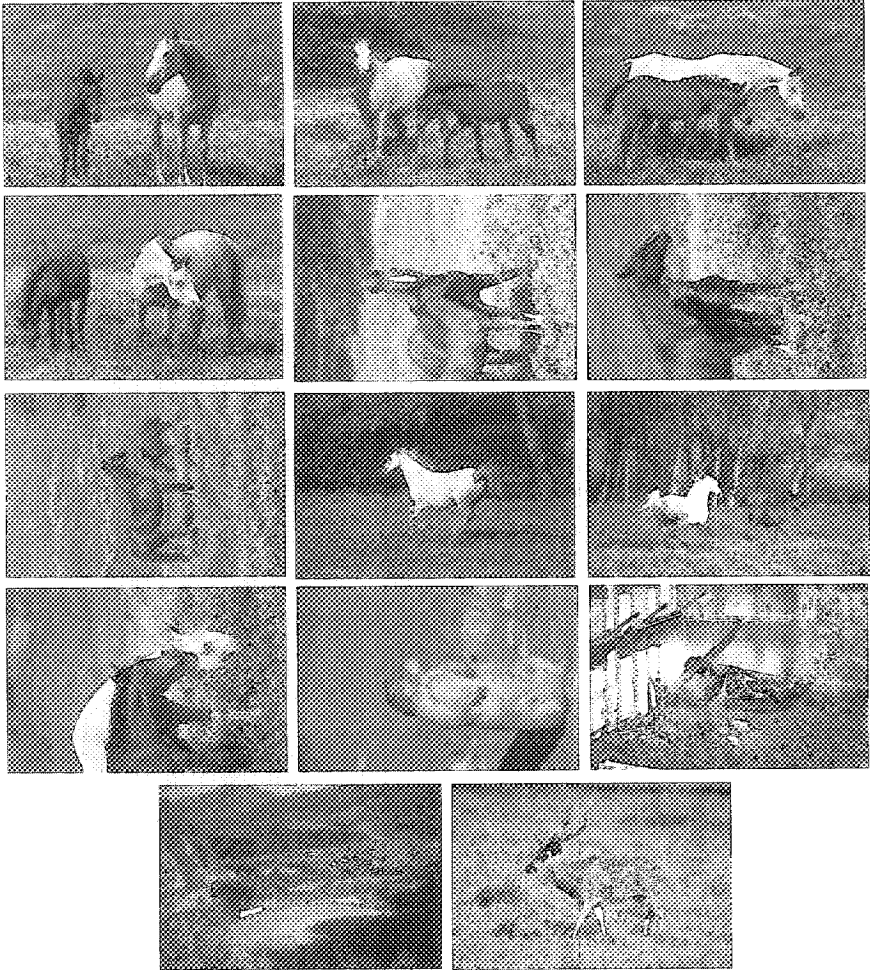


Figure 16. Images of horses recovered using a body plan representation from a test collection consisting of 100 images containing horses and 1,086 control images with widely varying content. Note that the method is relatively insensitive to aspect, but can be fooled by brown horse-shaped regions. More details appear in Forsyth and Fleck, 1997.

VideoQ - An Object-Oriented Content Based Video Search System

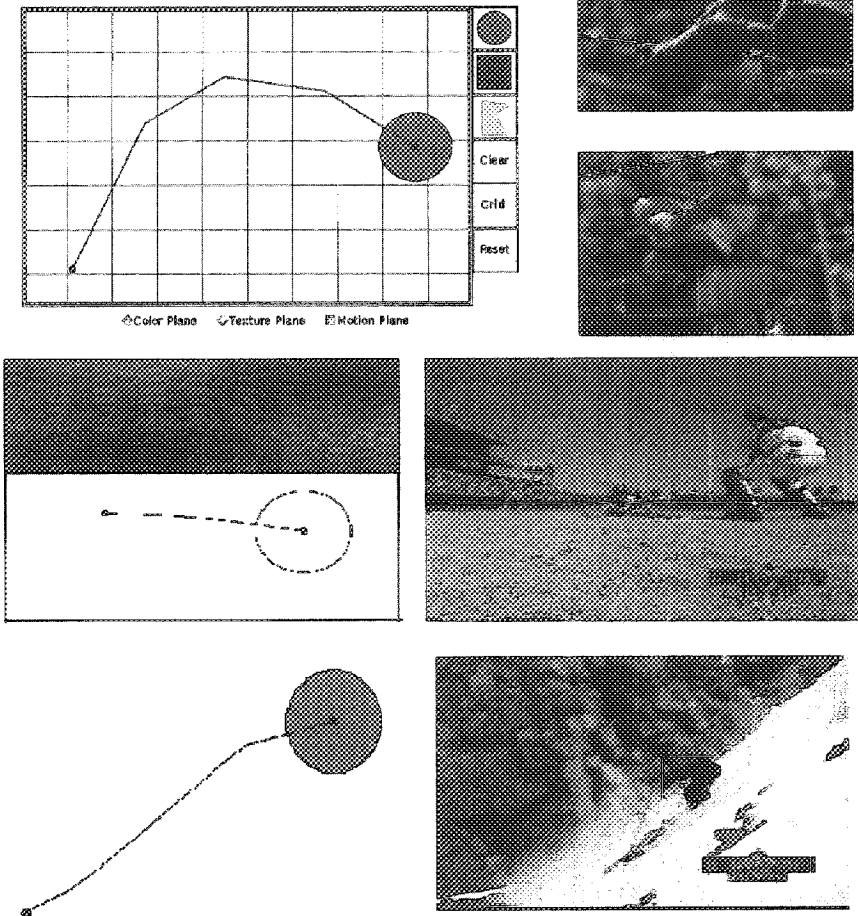


Figure 17. Video can be represented by moving blobs; sequences can then be queried by specifying blob properties and motion properties desired. The top left shows a query for a blob moving along a parabolic arc, sketched in the user interface for the VideoQ system. Top right shows frames from two sequences returned. As the center (baseball) and bottom (skiing) figures show, the mechanism extends to a range of types of motion. More details appear in Chang et al., 1997a. (Figure reproduced by kind permission of S-F. Chang.)

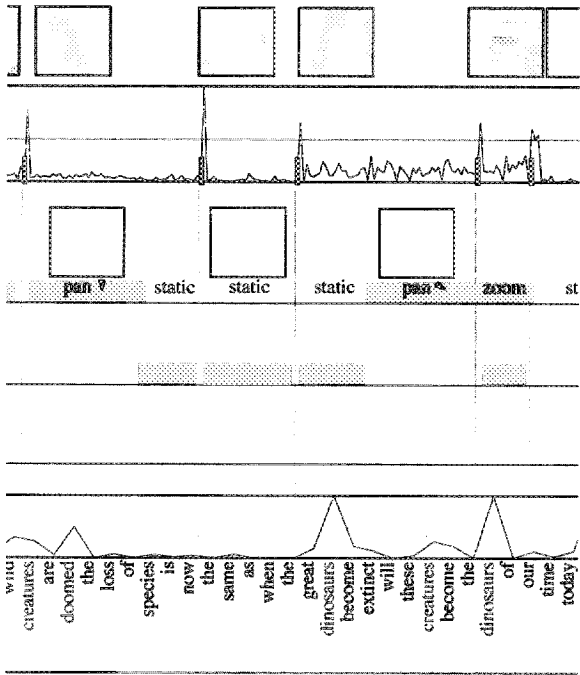


Figure 18. Characterizing a video sequence to create a skim. The video is segmented into scenes. Camera motions are detected along with significant objects (faces and text). Bars indicate frames with positive results. Word relevance is evaluated in the transcript. More information appears in Smith and Kanade, 1997. (Figure reproduced by kind permission of T. Kanade.)

allow searches for objects independent of their backgrounds. Furthermore, some special cases of object recognition can be handled explicitly. It is not known how to build a system that could search for a wide variety of objects; building a user interface for such a system would present substantial problems, too. Some images have text associated with them, either because content providers have explicitly described the image or because images can be associated with captions or document text. In this case, it is natural to use these terms to obtain semantic descriptions. There is currently no clear set of principles to use to combine available text descriptions with images. What lines of research are more promising? The answer depends on what application one has in mind. Some applications—e.g., reviewing images or video in libraries to choose a good background picture for an advertisement—really need images to be arranged by their appearance so that all the soothing blue pictures are in about the same place. Other applications demand semantic descriptions that are very difficult to supply; there is no way to answer a question like “show trends in the pose of subject in portraits in the seventeenth century” without well developed processes for finding people in pictures. General semantic descriptions are a long way off, but one natural focus is the activities of people. I expect that, in the next few years, using entirely automatic methods, it may be possible to find pictures of, for example, a politician kissing a baby.

ACKNOWLEDGMENTS

Many people kindly allowed me to use their figures and suggested captions. Thanks to: Shih-Fu Chang of Columbia; Takeo Kanade, Henry Rowley, and Michael Smith of CMU; Jitendra Malik and Chad Carson of U.C. Berkeley; B.S. Manjunath of U.C. Santa Barbara; Roz Picard of the MIT Media Lab; Cordelia Schmid of INRIA; Carlo Tomasi and Leo Guibas of Stanford; Paul Viola of the MIT AI Lab; Ramin Zabih of Cornell; and Andrew Zisserman of Oxford University. Margaret Fleck read a draft under trying circumstances. Portions of the research described in this article were supported by a Digital Library grant (NSF-IRI-9411334).

NOTES

- ¹ A collection of 60,000 images quite commonly used in vision research available in three series from the Corel corporation, whose head office is at 1600 Carling Avenue, Ottawa, Ontario, K1Z 8R7, Canada.
- ² At <http://elib.cs.berkeley.edu/photos>, there are many thousands of images of California natural resources, flowers, and wildlife.

REFERENCES

- Armitage, L. H., & Enser, P. G. B. (1997). Analysis of user need in image archives. *Journal of Information Science*, 23(4), 287-299.
- Belongie, S.; Carson, C.; Greenspan, H.; & Malik, J. (1998). Color and texture-based image segmentation using EM and its application to content based image retrieval. In *Proceedings of the IEEE 6th International Conference on Computer Vision* (Bombay, India, January 4-7, 1998) (pp. 675-682). New Delhi, India: Narosa Publishing House.
- Boreczky, J. S., & Rowe, L. A. (1996). Comparison of video shot boundary detection

- techniques. *Journal of Electronic Imaging*, 5(2), 122-128.
- Carson, C., & Ogle, V. E. (1996). Storage and retrieval of feature data for a very large online image collection. In S. Y. W. Su (Ed.), *Proceedings of the 12th International Conference on data engineering*. Los Alamitos, CA: IEEE Computer Society.
- Carson, C.; Thomas, M.; Belongie, S.; Hellerstein, J. M.; & Malik, J. (1999). Blobworld: A system for region-based image indexing and retrieval. In D. P. Huijsmans & A. W. M. Smeulders (Eds.), *Visual information systems* (Proceedings of the 3rd International Conference, Visual '99, June 2-4, 1999) (pp. 509-516). New York: Springer.
- Chang, S.-F.; Chen, W.; Meng, H. J.; Sundaram, H.; & Zhong, D. (1997a). VideoQ: An automatic content-based video search system using visual cues. In *Proceedings of the Fifth ACM International Multimedia Conference* (November 9-13, 1997, Seattle, WA) (pp. 313-324). New York: Association for Computing Machinery Press.
- Chang, S.-F.; Smith, J. R.; Beigi, M.; & Benitez, A. (1997b). Visual information retrieval from large distributed online repositories. *Communications of the ACM*, 40(12), 63-71.
- Chang, S.-F.; Chen, W.; Meng, H. J.; Sundaram, H.; & Zhong, D. (1998a). A fully automated content based video search engine supporting spatiotemporal queries. *IEEE Transactions on Circuits & Systems for Video Technology*, 8(8), 602-615.
- Chang, S.-F.; Chen, W.; & Sundaram, H. (1998b). In *ICIP '98: Proceedings of the 1998 International Conference on Image Processing* (October 4-7, 1998, Chicago, IL). Los Alamitos, CA: IEEE Computer Society.
- Chapelle, O.; Haffner, P.; & Vapnik, V. (1999). Support vectors for histogram-based classification. *IEEE Transactions on Neural Networks*, 10(5), 1055-1064.
- Congiu, G.; Del Bimbo, A.; & Vicario, E. (1995). Iconic retrieval by contents from databases of cardiological sequences. In *Visual database systems 3: Visual information management* (Proceedings of the 3rd IFIP 2.6 Working Conference on Visual Database Systems, March 27-29, 1995, Lausanne, Switzerland) (pp. 158-174). London: Chapman & Hall.
- De Bonet, J. S., & Viola, P. (1998). Structure driven image database retrieval. In M. I. Jordan, M. J. Kearns, & S. A. Solla (Eds.), *Advances in neural information processing systems* (vol. 10, pp. 866-872). Cambridge, MA: MIT.
- Eakins, J.; Boardman, P.; & Graham, M. E. (1998). Similarity retrieval of trademark images. *IEEE Multimedia*, 5(2), 53-63.
- Enser, P. G. B. (1993). Query analysis in a visual information retrieval context. *Journal of Document and Text Management*, 1(1), 25-52.
- Enser, P. G. B. (1995). Pictorial information retrieval. *Journal of Documentation*, 51(2), 126-170.
- Flickner, M.; Sawhney, H.; Niblack, W.; Ashley, J.; Huang, Q.; Dom, B.; Gorkani, M.; Hafner, J.; Lee, D.; Petkovic, D.; & Steele, D. (1996). Query by image and video content: The QBIC system. *Computer*, 28(9), 23-32.
- Forsyth, D. A., & Fleck, M. M. (1996). Identifying nude pictures. In *Proceedings of the 3rd IEEE Workshop on Applications of Computer Vision, WACV '96* (December 2-6, 1996, Sarasota, FL) (pp. 103-108). Los Alamitos, CA: IEEE Computer Society.
- Forsyth, D. A., & Fleck, M. M. (1997). Body plans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (San Juan, PR) (pp. 678-683). Los Alamitos, CA: IEEE Computer Society.
- Forsyth, D. A.; Fleck, M. M.; & Breger, C. (1996). Finding naked people. In B. Buxton & R. Cipollo (Eds.), *Computer Vision, ECCV '96* (Proceedings of the 4th European Conference on Computer Vision, Cambridge, United Kingdom, April 14-18, 1996) (pp. 593-602). Berlin, Germany: Springer-Verlag.
- Forsyth, D. A., & Ponce, J. (in press). *Computer vision: A modern approach*. Upper Saddle River, NJ: Prentice-Hall.
- Hampapur, A.; Gupta, A.; Horowitz, B.; Shu, C.-F.; Fuller, C.; Bach, J.; Gorkani, M.; & Jain, R. (1997). Virage video engine. In *Storage and retrieval for image and video databases V* (Proceedings of SPIE, The International Society for Optical Engineering) (vol. 3022, pp. 188-198). Bellingham, WA: SPIE.
- Holt, B., & Hartwick, L. (1994a). Quick, who painted fish?: Searching a picture database with the QBIC project at UC Davis. *Information Services and Use*, 14(2), 79-90.
- Holt, B., & Hartwick, L. (1994b). Retrieving art images by image content: The UC Davis QBIC project. *ASLIB Proceedings*, 46(10), 243-248.

- Huang, J., & Zabih, R. (1998). *Combining color and spatial information for content-based image retrieval*. Retrieved July 12, 1999 from the World Wide Web: <http://www.cs.cornell.edu/html/rdz/Papers/ECDL2/spatial.htm>.
- Jacobs, C. E.; Finkelstein, A.; & Salesin, D. H. (1995). Fast multiresolution image querying. In *Proceedings of SIGGRAPH '95* (August 6-11, 1995, Los Angeles, CA) (pp. 277-285). New York: Association of Computing Machinery Press.
- Jain, A. K., & Vailaya, A. (1998). Shape-based retrieval: A case study with trademark image databases. *Pattern Recognition*, 31(9), 1369-1390.
- La Cascia, M.; Sethi, S.; & Sclaroff, S. (1998). Combining textual and visual cues for content based image retrieval on the World Wide Web. In *IEEE workshop on content based access of image and video libraries* (pp. 24-28). Los Alamitos, CA: IEEE Computer Society.
- Lipson, P.; Grimson, W. E. L.; & Sinha, P. (1997). Configuration based scene classification and image indexing. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (June 17-19, 1997, San Juan, PR) (pp. 1007-1013). Los Alamitos, CA: IEEE Computer Society.
- Ma, W. Y., & Manjunath, B. S. (1997a). Edge flow: A framework for boundary detection and image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (June 17-19, 1997, San Juan, PR) (pp. 744-749). Los Alamitos, CA: IEEE Computer Society.
- Ma, W. Y., & Manjunath, B. S. (1997b). NeTra: A toolbox for navigating large image databases. In *Proceedings of the IEEE international conference on image processing* (October 26-29, 1997, Santa Barbara, CA) (pp. 568-571). Los Alamitos, CA: IEEE Computer Society.
- Ma, W. Y., & Manjunath, B. S. (1998). A texture thesaurus for browsing large aerial photographs. *Journal of the American Society for Information Science*, 49(7), 633-648.
- Malik, J., & Perona, P. (1989). A computational model of texture segmentation. In *Proceedings of the 22nd Asilomar Conference on Signals, Systems, and Computers* (October 31-November 2, 1988, Pacific Grove, CA, Naval Postgraduate School, San Jose State University) (pp. 490-494). San Jose, CA: Maple Press.
- Malik, J., & Perona, P. (1990). Preattentive texture discrimination with early visual mechanisms. *Journal of the Optical Society of America: A-Optics & Image Science*, 7(5), 923-932.
- Manjunath, B. S., & Ma, W. Y. (1996a). Browsing large satellite and aerial photographs. In *Proceedings of the 3rd IEEE international conference on image processing* (September 16-19, 1996, Lausanne, Switzerland) (pp. 765-768). New York: IEEE Computer Society.
- Minka, T. P., & Picard, R. W. (1997). Interactive learning with a "society of models." *Pattern Recognition*, 30(4), 565-581.
- Minka, T. P. (1996). *An image database browser that learns from user interaction* (MIT Media Laboratory Perceptual Computing Section Tech. Rep. No. 365). Cambridge, MA: MIT.
- Mundy, J. L., & Vrobel, P. (1994). The role of IU technology in radius phase II. In *Proceedings of the 23rd Image Understanding Workshop* (November 13-16, 1994, Monterey, CA) (pp. 251-264). San Francisco: Morgan Kaufmann.
- Mundy, J. L. (1995). The image understanding environment program. *IEEE Intelligent Systems and their Applications*, 10(6), 64-73.
- Mundy, J. L. (1997). IU for military and intelligence applications: How automatic will it get? In *Emerging applications of computer vision* (Proceedings of SPIE, the Society for Optical Engineering, vol. 2962) (pp. 162-170). Bellingham, WA: SPIE.
- Ogle, V. E., & Stonebraker, M. (1995). Chabot: Retrieval from a relational database of images. *Computer*, 28(9), 40-48.
- Oren, M.; Papageorgiou, C.; Sinha, P.; & Osuna, E. (1997). Pedestrian detections using wavelet templates. In *IEEE Computer Society conference on computer vision and pattern recognition* (June 17-19, 1997, San Juan, PR) (pp. 193-199). Los Alamitos, CA: IEEE Computer Society.
- Pentland, A.; Picard, R.; & Sclaroff, S. (1996). Photobook: Content-based manipulation of image databases. *International Journal of Computer Vision*, 18(3), 233-254.
- Picard, R. W., & Minka, T. (1995). Vision texture for annotation. *Journal of Multimedia Systems*, 3(1), 3-14.
- Poggio, T., & Sung, K.-K. (1995). Finding human faces with a gaussian mixture distribution-based model. In *AACV '95* (Proceedings of the 2nd Asian Conference on Com-

- puter Vision) (pp. 435-440). Singapore: Nanyang Technological University.
- Psarrou, A.; Konstantinou, V.; Morse, P.; & O'Reilly, P. (1997). Content based search in medieval manuscripts. In *TENCON '97* (Proceedings of the IEEE TENCON '97, IEEE Region 10 Annual Conference: Speech and image technologies for computing and telecommunications (December 2-4, 1997, Queensland University of Technology, Brisbane, Australia) (pp. 187-190). New York: IEEE Computer Society.
- Rowley, H. A.; Baluja, S.; & Kanade, T. (1996a). Human face detection in visual scenes. In D. S. Touretzky, M. C. Mozer, & M. E. Hasselmo (Eds.), *Advances in neural information processing 8* (Proceedings of the 1995 conference) (pp. 875-881). Cambridge, MA: MIT.
- Rowley, H. A.; Baluja, S.; & Kanade, T. (1996b). Neural network based face detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (June 18-20, 1996, San Francisco, CA) (pp. 203-208). Los Alamitos, CA: IEEE Computer Society.
- Rowley, H. A.; Baluja, S.; & Kanade, T. (1998a). Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1), 23-38.
- Rowley, H. A.; Baluja, S.; & Kanade, T. (1998b). Rotation invariant neural network-based face detection. In *Proceedings of the 1998 IEEE conference on computer vision and pattern recognition* (June 23-25, 1998, Santa Barbara, CA) (pp. 38-44). Los Alamitos, CA: IEEE Computer Society.
- Rubner, Y.; Tomasi, C.; & Guibas, L. J. (1998). A metric for distributions with applications to image databases. In *Proceedings of the 6th international conference on computer vision* (January 4-7, 1998, Bombay, India) (pp. 59-66). New Delhi, India: Narosa Publishing House.
- Sawhney, H., & Ayer, S. (1996). Compact representations of videos through dominant and multiple motion estimation. *IEEE Transactions on Analysis and Machine Intelligence*, 18(8), 814-830.
- Schmid, C., & Mohr, R. (1997). Local gray value invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5), 530-534.
- Schmid, C.; Zisserman, A.; & Mohr, R. (in press). Integrating geometric and photometric information for image retrieval. In *International workshop on shape, contour, and grouping in computer vision*.
- Seloff, G. A. (1990). Automated access to the NASA-JSC image archives. *Library Trends*, 38(4), 682-696.
- Smith, J. R., & Chang, S.-F. (1996). VisualSEEK: A fully automated content-based image query system. In *Proceeding of ACM multimedia '96* (November 18-22, 1996, Boston, MA) (pp. 87-98). New York: Association for Computing Machinery Press.
- Smith, J. R., & Chang, S.-F. (1997). Visually searching the Web for content. *IEEE Multimedia*, 4(3), 12-20.
- Smith, M. A., & Christel, M. G. (1995). Automating the creation of a digital video library. In *Proceedings of ACM multimedia '95* (November 5-9, 1995, San Francisco, CA) (pp. 357-358). New York: Association for Computing Machinery Press.
- Smith, M. A., & Hauptmann, A. (1995). Text, speech and vision for video segmentation: The Informedia project. In *AAAI Fall 1995 symposium on computational models for integrating language and vision*. Menlo Park, CA: AAAI Press.
- Smith, M., & Kanade, T. (1997). Video skimming for quick browsing based on audio and image characterization. In *1997 IEEE computer society conference on computer vision and pattern recognition* (June 17-19, 1997, San Juan, PR). Los Alamitos, CA: IEEE Computer Society.
- Smith, T. R. (1996). A digital library for geographically referenced materials. *Computer*, 29(5), 54-60.
- Swain, M. J., & Ballard, D. H. (1991). Color indexing. *International Journal of Computer Vision*, 7(1), 11-32.
- Trucco, E., & Verri, A. (1998). *Introductory techniques for 3-D computer vision*. Upper Saddle River, NJ: Prentice-Hall.
- Wactlar, H.; Kanade, T.; Smith, M.; & Stevens, S. (1996). Intelligent access to digital video: The Informedia project. *Computer*, 29(5), 46-52.
- Wong, S. T. C. (1998). CBIR in medicine: Still a long way to go. In *Proceedings of the workshop on content-based access of image and video libraries* (June 21, 1998, Santa Barbara, CA) (p. 114). Los Alamitos, CA: IEEE Computer Society.