



A FAREWELL TO KEYWORDS

By Gary Stix

A picture may be worth a kilo of words, but typing into Google Image the single word “rosebud” returns about 60,000 pictures.

The power of an individual keyword is both good and bad. It can find a virtual stack of Web pages. But it is unable to differentiate between the flower in bloom and legendary film director Orson Welles’s scowl. Ideally, an Internet user should be able to use the likeness of a rose to tell a search engine to find others like it.

The idea of using images to search images is not new. About a decade ago software emerged that could match one photograph to another or take a graphic representation—say, a large red dot on a green background—and track down pictures of a rose in a database [see “Finding Pictures on the Web,” by Gary Stix; *SCIENTIFIC AMERICAN*, March 1997]. This type of search—which collectively came to be known as content-based image retrieval—has progressed only slowly beyond the graduate-project stage.

Major search engines have yet to implement this form of image retrieval to mill through their indexes of images, the largest of which contain links to billions of photographs and graphics. Still, research by both industry and academia has achieved some intriguing advances of late that sidestep the need for keywords—and address the challenge of analyzing the content of images in large databases.

Dial and Shoot

THE RECENT PROLIFERATION of Web-enabled camera phones and PDAs—and the still persistent difficulty of using thumb and forefingers on undersize keypads to input keywords—opens opportunities for those who can find a way to pull results off the Web by sending a query in the form of an image captured by a phone’s camera.

Microsoft Research has identified a list of uses for a cell phone camera as a Web input device. A prospective buyer interested in information about a new stove could

The reigning obsession with search technology has elicited new ways of using images to track down information on the Web

photograph the appliance in a department store and relay the image as a file to a server that

can return to the user’s cell phone a Web page from *Consumer Reports*. A picture of the *Mona Lisa* could yield an art history page. A shot of a nearby landmark building might produce a map showing the user’s present location. “It brings the Web closer to the real world,” says Larry Zitnick, a member of Microsoft Research’s Redmond, Wash., laboratory. “It’s kind of like having the Web look at what you’re looking at.” Another group at Microsoft’s Asian research laboratory in Beijing works on a similar project, and eventually the two efforts could merge.

Among the challenges faced by these investigators is finding ways to create search algorithms powerful enough to comb through the image content of the entire Web. Zitnick and his colleagues have demonstrated a system that can take a “query” photograph captured from a cell phone camera and then send it to a server that matches it with already captured, or “training,” images, each of which provides a link to the relevant Web site. Zitnick wants to craft a database populated with the billions of images captured by a search engine, such as MSN Search. Currently the system, which still lacks a name, can perform retrievals from among tens of thousands of images in two to four seconds, an interval that needs to be reduced to a fraction of a second.

SNAPSHOTS taken by a cell phone camera may be used to search the Web.

AARON GOODMAN

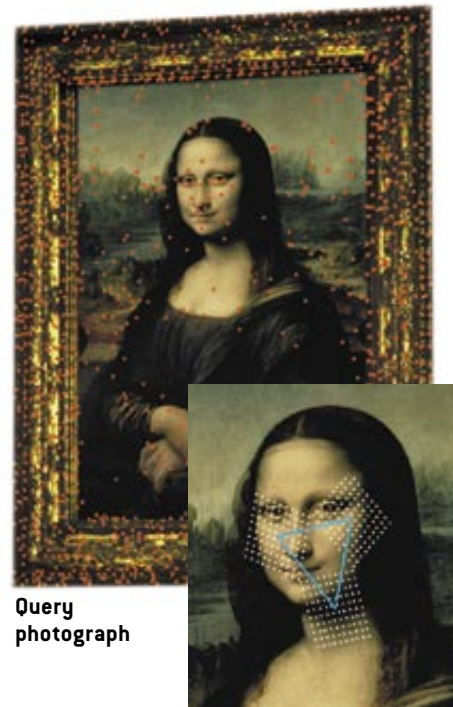
SEARCHING THE WEB BY PHOTO AND PHONE

In a project at Microsoft Research, a photograph snapped by a cell phone camera is used to look on the Web for matching images or information related to the picture.

1 A search begins when someone snaps a picture with a cell phone camera—a shot of the *Mona Lisa*, for instance—and sends the image to an image server via the Web. The server contains training images—copies of photographs gathered from all over the Web, which have been indexed and saved for matching with query images.



2 To speed the search for a match, the image server tries to find “features,” dark areas surrounded by light areas, or vice versa (*red dots*). Each feature forms the center of a square patch of pixels (often a 10-by-10 grid), and some of these features are grouped into sets of three that are a specified distance from one another (as indicated by *blue triangle at right*).



Query photograph



To prepare the system, a computer analyzes a training photograph from a Web page for distinctive features of the image that consist of dark areas surrounded by light areas, or vice versa. Some of the features are bunched into groups of three, based on a calculation of how far one is from the other. Each feature represents the center point of a 10-by-10 square patch of pixels. The grouping of three patches is called a triplet, some 5,000 of which are typically identified in a training image. The location of each triplet is stored as an entry in a huge table that is designed to minimize the amount of computation required to search any individual entry. A query image is also separated into triplets; these triplets are matched with those in the table, and then Web pages containing a matched image are sent to the user's cell phone [see box on these two pages]. The investigators chose triplets as the basis of comparison instead of single patches because a triplet encompasses a larger area of the overall image, a property that should reduce the risk of generating an erroneous match between a query and a training image.

As with most image-recognition applications, performance is far from perfect, with detection rates averaging about 80 percent. Detection of flat objects that have highly detailed surfaces—a description that fits many consumer products—has a better rate, however. And if performance constraints are loosened to allow for most people's expectation of a Web search—retrieval of not one but several links—the detection rate can be enhanced still further. Although an image search may return more than one Web site, its ability to match specific photographic features between the query images and the training images—a door, for instance—means that the number of images returned to the user would be fewer and better

targeted than the number found in the usual text search.

To expand the range of searchable objects, Zitnick wants to improve the system's ability to find mirrorlike surfaces or things with complex three-dimensional shapes, such as plants. The system, nonetheless, will never encompass the full range of visually detectable objects. “It's going to be useful for certain things, and for others, it's going to fail miserably,” he says.

Looking for Cheesecake

THE CHALLENGE of perusing the vast expanse of the Web for images remains a preoccupation for Google as well. The search-engine powerhouse does not let on the specifics of future plans, but its researchers have started to present papers on their doings at technical conferences. Full image-to-image matching or recognizing an individual object, such as a chair, takes a backseat, in the company's view, to the more pragmatic issue of how to provide simple generalizations about the content of billions of images. Is, for instance, that pinkish color in a particular photograph that of naked flesh or a snapshot of an Art Deco structure in Miami's South Beach? From the Web's earliest days, image-searching efforts have been plagued by the risk of having unwanted pornography turn up in the results.

“We want to make sure that images are classified as containing adult content by using not only keywords and URLs but also image analysis,” says Google researcher Shumeet Baluja. The Mountain View, Calif., company has developed—and actually implemented—a system that can differentiate with middling accuracy pictures that are naughty or nice, according to a paper from one conference. Eschewing shape-classification methods that can take from seconds to minutes

3 Group of three patches—called triplets—from the camera's query photograph is compared with triplets from training images to locate a match. An average of 5,000 triplets are collected for each query photograph and compared with all triplets from training images in the database. Comparing triplets, instead of the individual patches used by other search methods, enhances the likelihood of finding a match.



4 A mathematical algorithm ensures that all patches are depicted at the same scale and in the same orientation. When the triplet comparisons identify a match with a training image, the correspondence will be verified as correct if a set of pixels at the center points of the picture match (*green squares*).



5 Once a match is verified, a Web page on which the training image appears is sent to the user's cell phone.



to process, researchers have reported that they can detect half the adult pics among a test set of 1.5 billion thumbnail images during an eight-hour interval using 2,500 computers—a rate of about 20 images a second. A Web surfer does not have to wait eight hours: with such a tool, the user who wants to screen out porn simply instructs the search engine to omit the links that have already been tagged for questionable content.

The system works by combining modules for detecting 27 features, among them skin color, connected pixels (signifying a visual continuum of color that could represent flesh, for instance), skin texture and the presence of a face. Skin comprises a lot of colors, and many everyday objects assume the colors of flesh. One component of the detector looks for an object—perhaps a building—that often has the appearance of skin but can be distinguished by specific features, such as long, straight edges. The pictures turned up and flagged by the system can be filtered out. These tagged images serve as one component of Google's "image-safe search," a user-selected option in Google Images that also analyzes URLs and other text content to make decisions on what is inappropriate.

To be useful for Web-wide image perusal, any component algorithm in a larger search module would, above all, have to be fast and efficient. Two of the Google researchers on the adult-filtering project—Baluja and Henry Rowley—have dramatically reduced the amount of information required to determine the sex or orientation of a face. The resulting acceleration in processing time is important because users look for people perhaps more than any other type of image. Google would like a better way to determine whether Britney Spears or Tony Blair is really in a picture. It has crafted a variety of

image filters—one of which tries to identify a person's gender. Others examine a person's clothes or age.

The gender and facial-pose filters crafted by Baluja and Rowley are created by measuring the intensity (lightness and darkness) of pairs of pixels within a 20-by-20 patch taken from an image that includes the face of a man or a woman in one case and distinct facial poses in another. A separate algorithm first makes an educated guess as to the location of a face in an image before the pose-determination filter does its job. The filter has been trained so that it need examine only 150 pairs of pixels in a single 20-by-20 patch in the facial area before predicting with 99 percent accuracy whether a face is in one of five poses (frontal, right-half profile, and so on). Although they are suited to a variety of uses, both the gender and face-pose classifiers are being incorporated into Google's image-safe search feature.

The Google team tends to downplay technical virtuosity for its own sake. "We take a pragmatic approach to research," comments Peter Norvig, the company's engineering search quality director. "We prefer to finesse a problem if we can. We don't have to solve the object-recognition problem 100 percent." Given the density of information in any photograph, the key to success for an image-searching tool is to make each pixel count.

MORE TO EXPLORE

- Boosting Sex Identification Performance.** Shumeet Baluja and Henry Rowley. *Innovative Applications of Artificial Intelligence*, 2005.
 - Large Scale Image-Based Adult-Content Filtering.** Henry A. Rowley, Yushi Jing and Shumeet Baluja. *International Conference on Computer Vision Theory and Applications*, 2006.
- Larry Zitnick's home page: <http://research.microsoft.com/~larryz/>