A Picture is Worth a Thousand Keywords: Image-Based Object Search on a Mobile Platform

Tom Yeh[1], Kristen Grauman[1], Konrad Tollmar[2], Trevor Darrell[1]
MIT, CSAIL, USA[1] Lund University, Sweden[2]

ABSTRACT

Finding information based on an object's visual appearance is useful when specific keywords for the object are not known. We have developed a mobile image-based search system that takes images of objects as queries and finds relevant web pages by matching them to similar images on the web. Image-based search works well when matching full scenes, such as images of buildings or landmarks, and for matching objects when the boundary of the object in the image is available. We demonstrate the effectiveness of a simple interactive paradigm for obtaining a segmented object boundary, and show how a shape-based image matching algorithm can use the object outline to find similar images on the web.

Author Keywords

Content-based Image Retrieval, Object Recognition, Mobile Interface, Interactive Segmentation

ACM Classification Keywords

H.5.1.f Image/video retrieval, H.5.2.h Input devices and strategies, H.5.2.s Vision I/O

INTRODUCTION

Providing information to a user at the right time and the right place can be critical in a variety of situations. Today, it is generally possible to access the entire Internet from a mobile terminal, but to find a particularly relevant web page can be tedious. Considerable effort has been devoted to bandwidth and screen resolution issues, but comparatively little has been done to alleviate the difficulty of initiating and refining an information query within the constraints of a hand-held form factor. When querying about an object that has unique visual features, what can better describe these properties than a visual description? An image of the object can be used as a query that faithfully represents the visual properties of the item. By enabling searches on object appearance, we propose to offer a new, more convenient and direct means for finding information about objects encountered in everyday life.

With conventional mobile terminal interfaces it can be daunting to form complex textual queries and to browse through matches to find the right information. However, if a hand-held terminal has a camera, as do increasing numbers of cell phones and PDAs, a search can proceed using the appearance of the object directly. Such content-based queries have been the subject of much research, and systems for desktop searching of media libraries have been deployed (e.g., QBIC [6], Webseek [7], etc.). For the majority of these applications, however, image search has been less successful than traditional keyword search methods. Rather than matching based on appearance, images are typically pre-labeled with keywords or matching is performed based on image captions or filenames (e.g., Google Image Search).

When initiating a query from a camera-equipped mobile device, however, content-based queries can be very effective. In contrast to searching a multimedia database—where users know what they want categorically but not graphically—a user performing a mobile object search does not typically have the name or category label to use as a keyword search, but instead has the physical object in hand or nearby. Since a mobile terminal has natural limits to its ability to take complex text queries as input, we conjecture that the mobile platform has the potential to increase the popularity of image-based search engines—the mobile domain is one of the few in which images are often easier to obtain or enter than keywords.

The utility of mobile image search has been demonstrated for finding information about prominent landmarks by matching the whole image [8,9]. When the object of interest does not cover the majority of the image, the most common image distance metrics become less effective. Shape is also often a key feature for object retrieval, and obtaining an accurate estimate of the object contour is therefore important for object search applications.

To automatically discern the object shape is nontrivial when the image contains other objects or background structures. This is known as the problem of *image segmentation* and is widely considered to be an open problem in the computer vision literature. However, it is possible to drastically improve the segmentation performance of simple, fast algorithms by leveraging human perception with a suitable interactive environment.

CHI 2005, April 2-7, 2004. Portland, Oregon, USA.
ACM 1-59593-002-7/05/0004

We have developed an image-based object search system suitable for mobile platforms that uses an interactive segmentation technique as a front-end. Figure 1 shows an overview and example result from our system: the image of the object is taken by the user on the phone (1) and is interactively segmented (2). The segmented object image is used to query the database to find the most relevant object (3), which in turn retrieves the relevant web pages for the user's perusal (4).

For object search we exploit an interactive human-aided segmentation paradigm where human input is obtained online when the image of the object is being taken with the camera, as opposed to working on a static image offline with a mouse and editing tool such as Photoshop. A "two-shot" interface on a mobile device allows a user to specify the object of interest simply by taking two images, one with the object and one without the object. By comparing these two images, a segmented object image is extracted using simple computer vision techniques for foreground/background estimation [10].

In the following section we review existing applications of mobile image matching and related work in image segmentation. We then demonstrate the benefits of interactive segmentation with a user study that compares the two-shot approach to a baseline direct contour-drawing approach. We then describe our system for image matching via a mobile device and show examples of searching databases from the web.

RELATED WORK

Camera-equipped mobile devices are becoming commonplace and have been used for a variety of exploratory applications. The AP-PDA project built an augmented-reality system on a camera-equipped iPAQ to assist electricians in appliance repair [3]. Images of the problematic appliance are taken, sent to a remote server for geometry-based model recognition, and finally augmented with useful information pertinent to that particular model. Mobile image matching and retrieval has been used by insurance and trading firms for remote item appraisal and verification with a central database [2]. These systems are cases of information retrieval made possible by camera-equipped mobile devices, but they require specific models (e.g. for appliances) and are unable to perform generic matching of new images.

Interfaces for image segmentation are, of course, common in many desktop environments, and methods for automatic image partitioning along object boundaries are a core topic of research in computer vision. Photoshop and GIMP are tools known to many computer users, as is the general concept of tracing an object, painting an image mask, or performing a region fill based on a specified threshold or example value.

A variety of "smart" segmentation tools for interactive use have been developed, including a method for dynamic contour optimization often called *snakes* or *active contours* [5]. Systems for automatic segmentation into regions based on color and texture cues have been developed for CBIR using state-of-the-art methods in graph partitioning or spectral clustering (e.g., [1]). Ideally, these vision methods could be used to find the precise segmentation of the scene into objects, and a single pointer gesture could specify which object image to use as a search query. In practice, object segmentation is still an imperfect technique, and an interactive method (or top-down model) is typically needed to refine segmentation. Presently, implementation of these computer vision methods is impractical in real time on available mobile terminal platforms. As the speed of mobile platforms increases, we believe it will be advantageous to include these types of methods in an interactive paradigm.

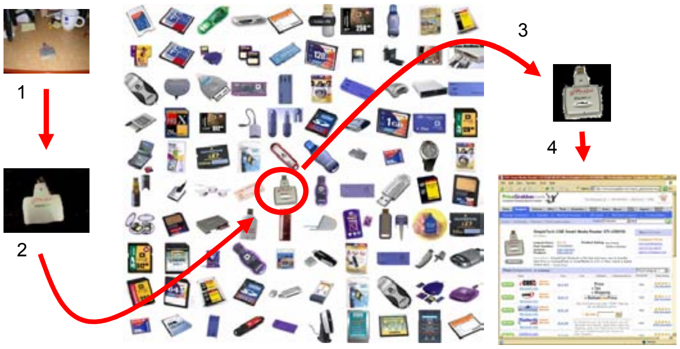

Figure 1: Mobile image-based object search. The result of a real search session for a flash memory reader is shown.

MOBILE IMAGE-BASED OBJECT SEARCH

Our system for image-based search on a mobile terminal (e.g., phone) allows users to specify an object of interest with images, submit the segmented object as a query to a shape-based matching engine, and then receive on the mobile terminal the images from a stored database that are most similar to the query object, together with a link to a relevant web page for each similar object. In the following sections we describe each component of the system and provide a user study to evaluate the two-shot interactive segmentation technique.

To see what types of image-based searches users would be interested in performing, we recruited subjects and instructed them to take pictures of objects with a camera phone over the course of several days and write down the information about these objects for which they would like to search on the web. Some of the most commonly raised questions were: "what is this", "where can I buy this", and "is there a review for this product." It is worth noting that a good number of questions were related to shopping. This suggests that besides general web search, a special mode targeted to online shopping could be very useful to mobile users. For example, an image database that only consists of images from an online merchant's website can be built, and by finding similar images on the website the user can be directed to the webpage describing the product (see Figure 1).

SPECIFYING IMAGES OF OBJECTS AS QUERIES

When building a mobile object search system, the first technical challenge is how to allow the mobile user to specify the image of the object of interest in a precise and intuitive way. One solution is to provide an interactive user-aided segmentation tool. Our method requires the user to take only two images—one with the object and one without the object, but against the same background—and it then automatically determines what area in the image this object must occupy (i.e., the area of the scene that changes between the two images). The interface guides the user to take both images from the same precise viewpoint by showing a semi-transparent overlay of a grayscale copy of the first image while the user is positioning the camera for the second image [10].

To test the usability of the interactive segmentation tool, we conducted a study in which we compared it to a simple baseline approach. The baseline approach employs a direct contour-drawing method whereby a user can specify the outline of the object by manipulating a cursor with a built-in joystick or number keys on an ordinary camera phone. We let 15 subjects perform image segmentation on five different kinds of objects (stuffed chicken doll, a shoe, a watch, a computer mouse, and a flash memory reader) using both tools and recorded the time taken to accomplish the task. The subjects were explicitly informed of the real purpose of the image segmentation and were allowed unlimited time and trials until they were satisfied with the segmentation result as shown in the review screen.

Users took considerably more time with the baseline method than with the interactive tool to obtain comparable segmentation qualities. Figure 2 shows the average time spent segmenting each object using each method. It is clear that the interactive approach has a time advantage over the naïve contour-drawing method.

In most of the cases, it took only one attempt for the participant to obtain a satisfactory segmented image using direct contour-drawing, but the interactive method took anywhere from one to five attempts. Although contour-drawing is straightforward, its time-consuming nature per trial discouraged users from trying multiple times. In contrast, the interactive segmentation mode required the users to take only two images, which can be very quick to perform; the time penalty to reapply the segmentation steps is relatively low.

Another important observation is that with direct contour-drawing, the time spent is independent of the types of background. In contrast, with the two-shot approach, more time is spent when the object is situated in a more complex background. Some participants recognized the difficulty a complex background can pose and learned to switch to a simpler background and redo the segmentation.

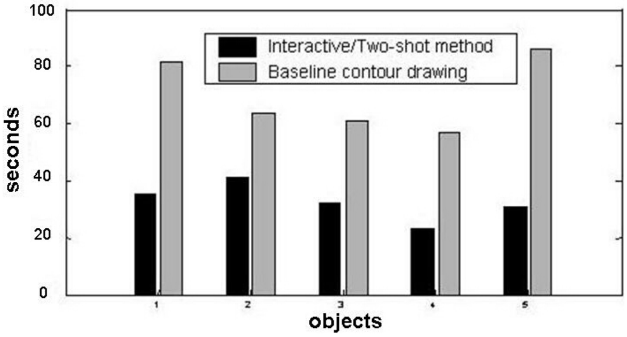

Figure 2 Average time subjects spent on segmentation for five different objects.

Each session concluded with an open-ended interview collecting user feedback. The most general response was that the two-shot system felt like playing a game. Most users felt they quickly got the hang of it and were surprised how the object image could be segmented out so easily. They could also quickly identify the main weakness of the system: it is not applicable to immobile objects such as a statue in a museum. Some said they would like to try it on their personal objects. None of them, given the choice between the two systems, preferred contour-drawing. All also agreed with the proposition that it would be occasionally useful in their daily experience to search with object appearance, if image matching was accurate enough to find images of the same object on the web.

FINDING IMAGES OF OBJECTS ON THE WEB

A segmented query image can be used to find similar objects, especially if an easily segmented database is available. We have experimented with several sources of segmented images available on the web. One useful source of segmented online images is the catalogues of major online merchants. These images are typically pre-segmented for display reasons. Moreover, they tend to be organized into meaningful categories and annotated with keywords and descriptive text containing useful information such as price, specification, reviews, and peer comments.

When the user has a general idea of the object category but wants to find information about a particular instance of the object, another useful image source is an image database indexed by text-based search engines (e.g., Google Image Search). For example, imagine a child wants to find stories about her teddy bear. The category keywords (e.g., “teddy bear”) can be used to retrieve a large number of images associated with the object category, and many will have simple backgrounds. We can search among these images based on visual characteristics to find an image with a similar appearance (e.g., a teddy bear that looks like the child’s).

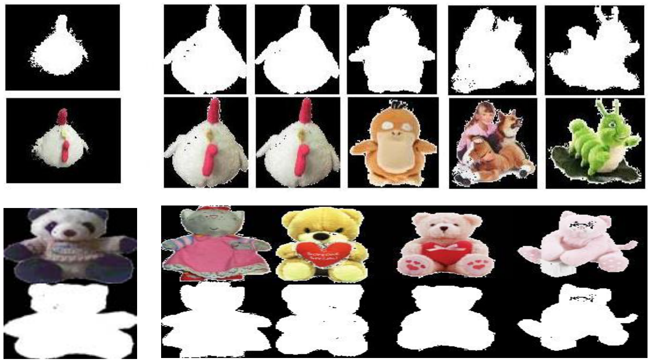

Figure 3: Example search results where the target object is present in (top, chicken) or absent from (bottom, bear) the database. Images were segmented with the interactive tool by our study participants.

Given a segmented image of the object of interest, that image is used to query the database of web images. Since we are focusing on segmented images, we are able to directly take advantage of the discriminative power of the object shape. We use the image matching technique presented in [4], since it is well-suited for comparing segmented shapes and allows efficient retrievals from large databases. Given an image search result with associated web pages, we could additionally extract salient keywords from those pages and use the resulting keywords to find more images to test against the search image, or to find relevant web pages directly [9].

EXAMPLES

To demonstrate our system we have tested it with two different databases of object images. We generated both databases automatically, using a keyword search on the web to initially detect candidate images, and then filtering these candidates to detect which examples were already well-segmented based on the uniformity of the background. Once the segmented database of web images is obtained in this manner, each database example's shape features are extracted and then embedded into a L_1 metric space using the method in [4]. The processing up to this point may be done offline. Then the embedded shape features from the image region segmented interactively on the camera phone are used to do a content-based search on the prepared database. The most similar images in the database are retrieved, and the search result is displayed on the phone screen.

The first database consists of 139 segmented stuffed animal images that were obtained by using Google Image Search with the keywords "plush toy". Our study participants used the interactive segmentation method on the phones to produce segmented images of a couple real toys as the queries. Figure 3 shows some example retrievals using input shapes produced by our study participants with the interactive segmentation technique. Because the real chicken toy that our users photographed was present in the toy database, our system retrieves these images as the top most relevant search results. In the event that the exact query object is not represented in the database, our system will return the objects whose shapes are most similar to the query, as in the bear doll example shown in Figure 3. This is a scenario where the user performs a general web search on an object by appearance; for example, a father wants to find a replacement for his child's broken toy.

The second database contains images that were obtained from web pages pulled up by a keyword query on PriceGrabber.com using the words "flash cards and accessories". After filtering the results to keep only segmented images, we were left with a database of 741 images. Figure 1 shows an example query result using our system with this database.

CONCLUSION

In this paper, we described a system for mobile web searches using object appearance, and demonstrated the advantage of using an interactive segmentation tool to let users specify objects of interest. An example application searching for an object such as a toy or computer accessory was shown, using segmented image databases collected automatically with a simple keyword search.

REFERENCES

1. Belongie S., Carson C., Greenspan H., and Malik J., Color- and Texture-Based Image Segmentation Using EM and its Application to Content-Based Image Retrieval. In *Proc. Int. Conf. Comp. Vision* 1998.
2. Built-in Camera Mobile Phones Increasingly Used for Business Purposes. *Nikkei Net Business*, Dec. 9 2002.
3. Gausemeier I., and Bruederlin B. Development of a Real Time Image Based Object Recognition Method for Mobile AR-devices. In *Proc. of AFRIGRAPH*, 2003.
4. Grauman K., and Darrell T. Fast Contour Matching Using Approximate Earth Mover's Distance. In *Proc. IEEE Conf. on Comp. Vision and Pattern Recognition*, June 2004.
5. Kass M., Witkin A.P., and Terzopoulos D., Snakes: Active Contour Models, *Int. Journal. Computer Vis.*(1), No. 4, January 1988.
6. Niblack W., Barber R., Equitz W., Flickner M., Glasman E., Petkovic D., Yanker P., Faloutsos C., and Taubin G.. The QBIC project, *SPIE Storage and Retrieval for Image and Video Databases*, 1993.
7. Smith J. and Chang S., Image and Video Search Engine for the World Wide Web. In *Proc. of SPIE*, 1997.
8. Tollmar K., Yeh T., and Darrell T., IDeixis - Image-Based Deixis for Finding Location-Based Information, In *Proc. Mobile HCI*, 2004.
9. Yeh T., Tollmar K., and Darrell T., Searching the Web with Mobile Images for Location Recognition. In *Proc. IEEE Conf. on Comp. Vision and Pattern Recognition*, June 2004.
10. Yeh T., Darrell T., DoubleShot: an Interactive User-Aided Segmentation Tool, In *Proc. IUI*, 2005.