IDEIXIS --- IMAGE-BASED DEIXIS FOR RECOGNIZING LOCATIONS

by

Pei-Hsiu Yeh

B.S., Computer Science (2000)

Simon Fraser University

Submitted to the Department of Electrical Engineering and Computer Science
in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Computer Science

at the

Massachusetts Institute of Technology

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
JUL 2 6 2004
LIBRARIES

February 2004
[June 2004]

© 2004 Massachusetts Institute of Technology
All rights reserved

Signature of Author .....
Department of Electrical Engineering and Computer Science
Feb 28, 2004

Certified by .....
Trevor Darrell
Professor of Computer Science
Thesis Supervisor

Accepted by .....
Arthur Smith
Chairman, Departmental Committee on Graduate Students

BARKER
GOOGLE EXHIBIT 1009

IDEIXIS --- IMAGE-BASED DEIXIS FOR RECOGNIZING LOCATIONS

by

Pei-Hsiu Yeh

Submitted to the Department of Electrical Engineering and Computer Science
on Feb 28, 2004 in partial fulfillment of the
Requirements for the Degree of Master of Science in
Computer Science

ABSTRACT

In this thesis, we describe an approach to recognizing location from camera-equipped mobile devices using image-based web search. This is an image-based deixis capable of pointing at a distant location away from the user's current location. We demonstrate our approach on an application allowing users to browse web pages matching the image of a nearby location. Common image search metrics can match images captured with a camera-equipped mobile device to images found on the World Wide Web. The users can recognize the location if those pages contain information about this location (e.g. name, facts, stories … etc). Since the amount of information displayable on the device is limited, automatic keyword extraction methods can be applied to help efficiently identify relevant pieces of location information.

Searching the entire web can be computationally overwhelming, so we devise a hybrid image-and-keyword searching technique. First, image-search is performed over images and links to their source web pages in a database that indexes only a small fraction of the web. Then, relevant keywords on these web pages are automatically identified and submitted to an existing text-based search engine (e.g. Google) that indexes a much larger portion of the web. Finally, the resulting image set is filtered to retain images close to the original query in terms of visual similarity. It is thus possible to efficiently search hundreds of millions of images that are not only textually related but also visually relevant.

Thesis Supervisor: Trevor Darrell
Title: Professor of Computer Science

Acknowledgements

I thank my advisor Trevor Darrell for his insights and guidance, and for many hours of tireless discussions on the topics relevant to this thesis.

I would like to thank all of my friends for their encouragement and assistance.

I thank my parents for all their sacrifices and encouragements, and for their belief in me.

Last but absolutely not least, I thank God for His daily spiritual supply and support, which allow me to face any difficulty during the completion of this thesis.

Table of Contents

CHAPTER 1 INTRODUCTION **7**

CHAPTER 2 RELATED WORK **11**

2.1 CONTENT-BASED IMAGE RETRIEVAL 11
2.2 LOCATION RECOGNITION 12
2.3 CAMERA-EQUIPPED MOBILE DEVICES 13
2.4 LOCATION-BASED INFORMATION RETRIEVAL 14

CHAPTER 3 IMAGE-BASED LOCATION RECOGNITION **15**

3.1 IMAGE MATCHING METRICS 16
3.1.1 Color Histogram *16*
3.1.2 Energy Spectrum *17*
3.1.3 Scene-context Feature *18*
3.2 FINDING MATCHING IMAGES 19
3.3 EXPERIMENTS AND RESULTS 19

CHAPTER 4 KEYWORDS BOOTSTRAPPING **23**

4.1 KEYWORD EXTRACTION 24
4.2 KEYWORD BOOTSTRAPPING IMAGE SEARCH 27
4.3 CONTENT-BASED FILTERING 27

CHAPTER 5 CONCLUSIONS AND FUTURE WORK **29**

REFERENCES **31**

List of Figures

Figure 1-1: Mobile Image-based Deixis – the client side application is running on the phone the girl is holding (left), while complicated CBIR algorithms are running a remote server. 8
Figure 1-2: Image-and-keyword hybrid search. 9
Figure 3-1: A walkthrough of the first prototype. User snaps an image to send as a query (left). The result is displayed as a thumbnail mosaic (center). Selecting a thumbnail image brings up a source webpage for browsing (right). 20
Figure 3-2: Average performance of three distance metrics presented in percentage of candidate images that are relevant or similar over the entire set of 150 test images..... 22
Figure 3-3: Average retrieval success for each distance metric shown in percentages of 150 test attempts that returned at least some relevant images on the first page of candidate images. 22
Figure 4-1: A user points the camera at a building and captures an image. 23
Figure 4-2: From the photo gallery the user can pick a query image (left) and send it to server (right). In this case, the trasmission method is through e-mail sent to a designated address. 24
Figure 4-3: Some examples of found webpages: (1) MIT Gallery of Hacks, (2) MIT Club of Cape Cod's official website, (3) Someone's picture gallery, and (4) Someone's guide on Boston. 25
Figure 4-4: Unigram and bigram frequency in the result of a sample query. 26
Figure 4-5: Search result. In this case, the result is contained in an email sent back to the user. It contains a set of extracted keywords as well as a list of relevant URLs. 26
Figure 4-6: A user can select a subset of keywords to search on Google (left). Google can return a list of relevant URLs for the search keywords "MIT DOME" (right). 27
Figure 4-7: Content-based filtering the search result. 28

Chapter 1

Introduction

Location-based information services offer many promising applications for mobile computing. For example, in the scenario of location-based content delivery service, a mobile user can be informed a list of restaurants nearby for a night of fine dining or the direction of the closest gas-station before some strenuous car-pushing is needed. This kind of service is useful only if the location of the mobile user can be automatically determined by the device.

Currently, many location recognition systems depend on radio or other signals (e.g., GPS, RFID tags). Technically speaking, the performance of GPS is often limited in indoor environments where the GPS satellite signal is greatly attenuated by the surrounding structures. RFID tags are suitable for indoor environments but tag deployment and infrastructure setup can be very expensive. As far as usability is concerned, these approaches are able to recover the current location but insufficient for recognizing locations distant from a device's actual position (e.g., "Tell me about that building over there."). There are no common or convenient means to make a pointing (deictic) gesture at a distant location with existing mobile computing interface or location-based computing infrastructure. We propose a new paradigm for such mobile image-based deixis, inspired by the growing popularity of camera-phones/PDAs and leveraging the fact that taking a picture is a natural and intuitive gesture for recording a location for all users who have ever taken a holiday snapshot.

The key idea is that with a camera phone, users can point at things by taking images, send images wirelessly to a remote server and retrieve useful information by matching the images to a multipurpose database such as the World Wide Web. Here we describe a scenario to illustrate an instance of the general idea:

> *"Mary is visiting campus for the first time ever. She is supposed to meet a friend at "Killian Court". She is uncertain if the building right in front is the "Killian Court". She*

takes an image of the building and sends it to the server. This image is then used to search the web for pages that also contain images of this building. The server returns the 5 most relevant web pages. By browsing these pages, she finds the name "Killian Court" and concludes that this is the right place."

An image-based approach can recognize locations distinct from a device's actual position. (e.g., "Tell me about that building over there!", while taking a snapshot of the building.) We have found that many landmarks and other prominent locations already have images on relevant web pages, so that matching the mobile image to web documents can provide a reasonable location cue. We take images of an unknown location from a mobile camera, send them to a server for processing, and return a set of matching web pages.

The entire system consists of two major components: a client-side application running on the mobile device, responsible for acquiring query images and displaying search results, and a server-side search engine, equipped with a content-based image retrieval (CBIR) module to match images from the mobile device to pages in a generic database (Figure 1-1).

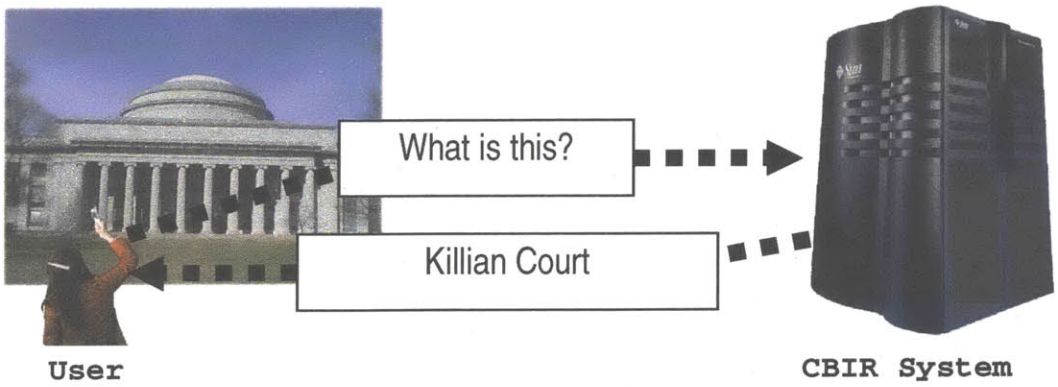


Figure 1-1: Mobile Image-based Deixis – the client side application is running on the phone the girl is holding (left), while complicated CBIR algorithms are running a remote server.

However, content-based search over the entire web is a computationally prohibitive proposition. Searching an image database on a scale similar to any commercial keyword-based image search engine, such as Google which boasts a collection of 425 million images, remains a research challenge. We propose a hybrid approach that can provide content-based image retrieval by combining both

image-based and text-based search. We leverage the fact that there will be redundant landmark images on the web and that search in a subset of the web will likely find a matching image.

Our approach is to first search a bootstrap set of images obtained by web-crawling a restricted domain. The domain is selected based on the expected application, for example tourism-related sites for a particular geographic location. To recover relevant pages across the full web, we exploit a keyword-based search followed by a content-based filtering step. Keywords are extracted from web pages with matching images in the bootstrap set. Instead of running CBIR over hundreds of millions of images, we only need to operate on a seed set of images and on the images returned from keyword-based search.

This idea is illustrated in Figure 1-2, breaking down into four steps. In step 1, the user takes an image with a camera phone, which is used as a query to find similar images from a small image database using CBIR techniques. In step 2, keywords are automatically extracted. In step 3, extracted keywords are sent to Google to find textually related images. In step 4, CBIR techniques are applied once again to filter out visually irrelevant images.

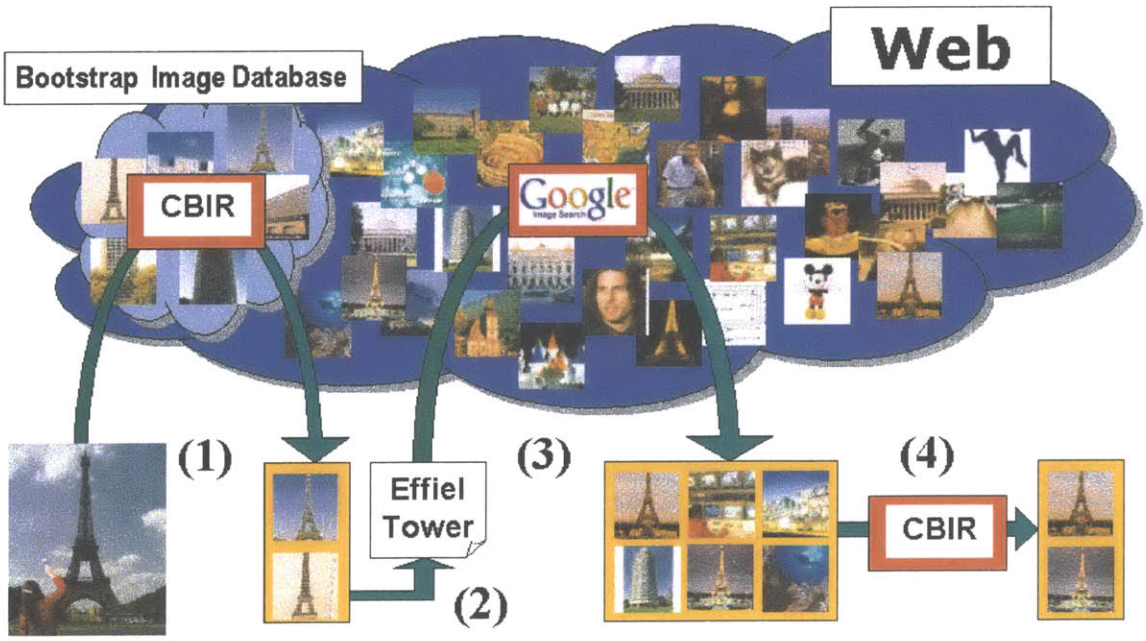


Figure 1-2: Image-and-keyword hybrid search.

To date, image-based web searching has seen only limited application despite considerable research effort in the area of content-based image retrieval (CBIR). Visual queries are hard to formulate on the desktop; unlike keywords, a visual query may require the user to sketch [17] or select parameter values for each low-level feature [21]. For recognizing real-world entities, searching with an actual image seems ideal. The advent of a new generation of camera-equipped phones and PDAs provides a unique platform for image-based web search.

In this thesis, we describe Image-based Deixis (IDeixis), an image-based approach to specifying queries for finding location-based information. We first review related work in the literature. We then describe in chapter 3 our approach to conducting image-based web search for finding matching location images for the purpose of location recognition. In chapter 4, we look into idea of using automatically extracted keywords to implement a hybrid text-and-image search method. Then, the thesis is concluded by offering a list of possible future works to extend the idea of image-based deixis for recognizing locations.

Chapter 2

Related Work

2.1 Content-based Image Retrieval

As imaging devices such as cameras and scanners are made cheaper and more powerful and the process of the digitalization of images is made easier than ever, the storage and archiving of a huge quantity of images has become a common practice in many areas such as journalism, meteorology and medicine. To be able to efficiently retrieve the desired images from any sizable image database is essential for productivity. An automatic approach is preferred since manually going through a huge collection of images is too time consuming. Therefore, an image search engine powered by fast computers and smart algorithms can be a very useful tool.

Most commercially successful image search engines are text-based. Corbis features a private database of millions of high-quality photographs or artworks that are manually tagged with keywords and organized into categories [6]. Google has indexed more than 425 millions web pages and inferred their content in the form of keywords by analyzing the text on the page adjacent to the image, the image caption, and other text features [11]. In both cases, the image search engine searches for images based on text keywords. Since the visual content of the image is ignored, images that are visually unrelated can be returned in the search result. However, this approach has the advantage of text search—semantically intuitive, fast, and comprehensive.

The alternative is to search images by image content. This is generally referred to as content-based image retrieval (CBIR) and has been an ongoing research topic for many years. Although the algorithms for analyzing image content in general run significantly slower than those for analyzing text, the search result can exhibit stronger visual relevance; we can have greater confidence as to whether the images retrieved fit the intended query. One of the first such systems was IBM's Query-

By-Image-Content (QBIC) system [21]. It supported search by example images, user-drawn pictures, or selected color and texture patterns, and was applied mainly to custom, special-purpose image databases. In contrast, the Webseek system [28] searched generically on the World Wide Web for images. This system incorporated both keyword-based and content-based techniques; the keyword search returned a set of images among which users could further search by color histogram similarity with relevance feedback. The Diogenes system used a similar bi-modal approach for searching images of faces on the web [3]. A face detector operated as a screening process to filter out nonface images from the set of initial images returned by a keyword-based query. Then, a face recognition module processed the remaining images to identify particular faces. These systems have not been applied to the task of recognizing locations from mobile imagery.

Describing what images to search can be tricky. QBIC lets users choose the relative weights of each feature in determining image similarity. [17] allows users to sketch the outline of the image. [28] features an interface to paint a color template for template-based image matching. However, these systems failed to provide an intuitive interface. The correspondence between low-level parameter values and high-level visual concept is not so obvious to ordinary users. Drawing can require too much effort and artistic skill. One solution is the search-by-example approach. For example, ImageRover presents the user with a set of starting images as examples and uses a "relevance feedback" framework to refine the search iteratively based on what the user says about the relevance of each image [27].

Due to the complexity typically involved in CBIR algorithms, there is a tradeoff between interactivity and comprehensiveness. Many existing CBIR systems use simple nearest-neighbor techniques to find matching images, which cannot scale to a large and comprehensive image database and still perform fast enough to be interactive. Hierarchical indexing with Self-Organizing Maps [16] and related techniques is possible, but the speed improvement is very limited. So far, practical CBIR systems operate on a scale much smaller than their text-based counterparts. We were not aware of any CBIR system that had gone beyond the million images mark in scale [16],[17],[28],[27],[3],[21].

2.2 Location Recognition

The notion of recognizing location from mobile imagery has a long history in the robotics community, where navigation based on pre-established visual landmarks is a well-known technique. The task of simultaneously localizing robot position and mapping the environment (SLAM) has

received considerable attention [19]. Similar tasks have been addressed in the wearable computing community, with the goal of determining the environment a user is walking through while carrying a body-mounted camera [34].

Our approach to image-based location-recognition is closely related to efforts in the wearable-computing and robotics literature on navigation and scene recognition. The wearable-museum guiding system built by Starner and colleagues uses a head-mounted camera to record and analyze the visitor's visual environment [30]. Computer vision techniques based on oriented edge histograms were applied to recognize objects in the field of view. Based on the objects seen, the system estimates the location in the museum and displays relevant information. The focus of this system was on recalling prior knowledge of locations—which item is exhibited where—rather than finding new information about locations. Torralba et al. generalized this type of technique and combined it with a probabilistic model of temporal dynamics, so that information from a video sequence from a wearable device could be used for mapping and navigation tasks [34]. In these robotics and wearable computing systems, recognition was only possible in places where the system had physically been before. In our system location-relevant information can be found from a single image of a novel place.

2.3 Camera-equipped Mobile Devices

Camera-equipped mobile devices are becoming commonplace and have been used for a variety of exploratory applications. The German AP-PDA project built an augmented reality system on camera-equipped iPAQ to assist electricians in appliance repair [10]. 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. At HP, mobile OCR applications are being developed which use a pen-size camera to capture images of text for archiving storage or machine translation [24]. FaceIT ARGUS is a commercial system that provides remote face recognition to law enforcement agency on mobile devices [7]. Mobile image matching and retrieval has been used by insurance and trading firms for remote item appraisal and verification with a central database [4]. These systems are successful 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.

2.4 Location-based Information Retrieval

There are already location information services offering on-line map (e.g. www.mapquest.com), traffic reports, and marketing opportunities on mobile devices. An often-discussed commercial application of location-based information is proximity-based coupon delivery. In a typical scenario, a merchant is notified when a potential customer visits nearby a retail outlet, upon which the customer can be delivered a coupon or offered a special promotional deal. The comMotion project [18] extended this idea to allow user-side subscription-based and location-specific content delivery. The GUIDE system was designed to provide city visitors location specific information customized to their personal needs [5]. The identity of the location is provided by the underlying cell-based communication infrastructure, which is also responsible for broadcasting relevant tourist information.

In [5], Cheverst concluded that the ability to filter information based on location made such electronic tour guide a very useful tool. However, care needed to be given when constraining the scope of available information to a specific location. For example, some users reported frustration when they were interested in something visible in the distance but were unable to find anything about it in the tour guide because the available information was restricted to their current physical location.

In [14], Kaasinen examined the usage of location-based information through interviews. What he found was that even if new technologies, such as GPS, can somewhat reliably provide users with their current location, the real need for location-based information often goes beyond merely an answer to "where am I". For instance, a GPS-enabled mobile Yellow Page might tell users a list of nearby restaurants but users realized they would want to know more, such as menus and reviews. This study highlighted the need for more comprehensive services in terms of geographic coverage, variety (number of services offered) and depth (amount of information available).

Chapter 3

Image-based Location Recognition

Web authors tend to include semantically related text and images on web pages. Thus, there exists a high correlation between semantic relevancy and spatial proximity that can be exploited on the web; pieces of knowledge close together in cyberspace tend to be also mutually relevant in meaning [2]. To find information about a well-known landmark, we can thus find web pages with images which match the image of the current location, and analyze the surrounding text. Using this approach, the location-recognition problem can be cast as a CBIR problem — if methods can be found to match mobile images to the web despite time, pose, and weather variation, it can serve as a useful tool for mobile web search (and in the particular application we consider here, location-based computing.)

Given an image taken with a camera phone, similar images can be found on the web. Relevant keywords can be found in the surrounding text and used directly as a location context cue, or used for further interactive browsing to find relevant information resources. A very attractive property of this method is that it requires no special-purpose communications infrastructure or prior location database, and that from a user-interface perspective users can specify nearby locations with a very natural interface action–taking a picture of the intended place.

The aforementioned benefit does not come without any difficulty. CBIR systems with decent retrieval performance often involve complex image matching algorithms. They might run into problem with scalability when a very large number of images need to be considered. Matching against the millions of images on the web in real-time is currently computationally infeasible. Therefore, we can device a hybrid keyword-and-image query system where we effectively implement CBIR over the entire 425 million images without having to apply a content-based metric on every single image. Such a hybrid design benefits from both the power of keyword-based search (speed and comprehensiveness) and image-based search (visual relevancy).

We leverage an existing keyword-based image search engine, Google, which has indexed more than 425 millions images. We extract keywords from web pages found in a content-based search in a bootstrap database, and use these keywords on Google to search its larger database of images for images we want. Recall that one shortcoming of keyword-based search is the existence of visually unrelated images in the result set. To overcome this, we apply a filtering step, where we run CBIR on this small set of images to identify visually related images. In this way we retrieve images that are not only visually relevant but also textually related. In this chapter, we discuss in details several components involved in this approach to carrying out image-based web search.

3.1 Image Matching Metrics

Having the right feature set and image representation is very crucial for building a successful CBIR system. The performance of general object matching in CBIR systems is far from perfect; image segmentation and viewpoint variation are significant problems. Fortunately, finding images of landmarks requires analysis over the entire image, making general image segmentation unnecessary. (A simpler, robust filtering step can remove small regions with foreground objects: this is easier than segmenting a small or medium sized object from a large image). Also, users ask about a location most likely because they are physically there and there are a much smaller number of physically common viewpoints of prominent landmarks than in the entire viewsphere of a common object. Consequently, the simplicity of recognizing location images naturally makes it an attractive test-bed to test the idea of an image-and-text hybrid search approach.

We have experimented with three image-matching metrics on the task of matching mobile location images to images on the World Wide Web. The simple color histogram is first used as the preliminary study. Then, we adopt the metrics described in [23] and [34], originally developed for image-based location recognition applications in robotics and wearable computing, to fit our specific purpose of retrieving similar landmark images from the web.

3.1.1 Color Histogram

Techniques based on global color histogram have advantages of simplicity and insensitivity to scale and orientation. The application of color histogram to related areas such as image indexing and retrieval has been shown to have good results [31]. To match images, we compute the color histogram over the entire image database. Then, given a query image, we simply compute its color histogram and find images in the database whose color histograms are most similar to that of the query image. Similarity can be measured as the Euclidean distance between two color histograms

expressed as vectors. This is perhaps the most naïve method. Being a global metric, it is insensitive to orientation and scale. But it is sensitive to illumination variation.

Not too surprisingly, color histogram is not discriminative enough to achieve a reasonable performance. The image domain in our application is images that contain landmarks, which typically consist of man-made structures against certain natural background. Intuitively, the structural information should be important for telling different landmarks apart. Pure global color histogram completely ignores such local structural information.

3.1.2 Energy Spectrum

Energy spectrum is a more advanced metric based on the work in [23]. It is the squared magnitude of the Fourier transform of an image. It contains unlocalized information about the image structure. This type of representation was demonstrated by [34] to be invariant to object arrangement and object identities. The energy spectrum of a scene image stays fairly constant despite the presence of minor changes in local configuration. For instance, different placements of people in front of a building should not affect its image representation too dramatically. Several previous studies can be seen to exploit energy spectrum in simple classification tasks [12].

In more detail, to compute the energy spectrum *ES* of an image *i*, we first compute the discrete Fourier transform (DFT) of an image:

$$I(f_x, f_y) = \sum_{x,y=0}^{N-1} i(x,y)h(x,y)e^{-j2\pi(f_x x + f_y y)} = A(f_x, f_y)e^{j\Phi(f_x, f_y)}$$

The $i(x,y)$ term in the intensity distribution of an image along two axis x and y. The spatial frequency is represented by f_x and f_y. $h(x,y)$ is a circular Hanning window to reduce boundary effects. The resulting sum $I(f_x, f_y)$ has a complex form, which can be decomposed into two real terms: $A(f_x, f_y)$ and $\Phi(f_x, f_y)$. The former is the amplitude spectrum of the image and the later is the phase function of the Fourier transform.

The information contained in the phase function $\Phi(f_x, f_y)$ is relative to the local properties of the image such as the form and the position of image components [20]. The amplitude spectrum $A(f_x, f_y)$ on the other hand represents the spatial frequencies spread everywhere in the image. In effect, it embeds unlocalized information about the image structure such as orientation, smoothness,

length and width of the contours that compose the scene picture. Finally, we square $A(f_x, f_y)$ to obtain an absolute value to represent the energy spectrum for image *i*.

$$ES(i) = A(f_x, f_y)^2 = | I(f_x, f_y) |^2$$

3.1.3 Scene-context Feature

The second metric is based on wavelet decompositions. Local texture features are represented as wavelets computed by filtering each image with steerable pyramids [9] with 6 orientations and 2 scales to its intensity (grayscale) image. Since this gives us only the local representation of the image, we take the mean values of the magnitude of the local features averaged over large windows to capture the global image properties.

For a given image *i*, its local representation contains sub-bands λ_i, is computed by convolving the image with the filter *G* oriented at each angle θ_i.

$$s(i) = G(\theta_i) * I$$

We use the second-derivative of Gaussian for *G* and X-Y separable basis filters to synthesize *G* at an arbitrary angle. The basis filters are:

$$G_a = 0.9213(2x^2 - 1)e^{-(x^2+y^2)},\ G_b = 1.843xye^{-(x^2+y^2)},\ G_a = 0.9213(2y^2 - 1)e^{-(x^2+y^2)}$$

To rotate *G* along to an angle θ_i, we compute the following:

$$G(\theta_i) = (\cos^2(\theta_i)G_a - 2\cos(\theta_i)\sin(\theta_i)G_b + \sin^2(\theta_i)G_c$$

This is computed over two different scales: 75 by 75 and 150 by 150. Therefore, a total number of sub-bands is 12.

Global characteristics of the image is captured by computing the mean value of the magnitude of the local features averaged over large spatial regions:

$$m(x) = \sum_{x'} | \lambda(i) | \cdot w(x' - x)$$

where $w(x)$ is the averaging window. The final representation combines both global and local features and is downsampled to have a spatial resolution of 4 by 4 pixels. Thus, the dimensionality of $m(x)$ is $4 \times 4 \times 12 = 96$.

3.2 Finding Matching Images

Given a query mobile image of some landmark, similar images can be retrieved by finding the k nearest neighbors in terms of the Mahalanobis distance between two feature vectors computed using the selected metric. However, the high dimensionality (d) of the feature involved in the metric can be problematic. To reduce the dimensionality, we computed principal components (PCs) over a large number of landmark images on the web. Then, each feature vector can be projected onto the first n principal components. Typically, $n << d$. The final feature vector will be the n coefficients of the principal components.

3.3 Experiments and Results

We built our prototype on a Nokia 3650 phone taking advantage of its built-in camera (640×480 resolution) and the support for Multimedia Messaging Service (MMS), using C++ on Symbian OS [32]. To initiate a query, the user points the camera at the target location and takes an image of that location, which is sent to a server via MMS.

We designed our system with an interactive browsing framework, to match users' expectations based on existing web search systems. For each query image, the search result contains the 16 most relevant candidate images for the location indicated by the query image. Selecting a candidate image brings up the associated web page. The user can browse this page to see if there is any useful information (Figure 3-1).

To evaluate whether CBIR can match location images from mobile devices to the pages on the web, we construct an image database consisting of 12,000 web images collected from the `mit.edu` domain by a web crawler. Test query images were obtained by asking student volunteers to take a total of 50 images from each of three selected locations: Great Dome, Green Building and Simmons Hall. Images were collected on different days and with somewhat different weather conditions (sunny/cloudy); users were not instructed to use any particular viewpoint when capturing the images.

Figure 3-1: A walkthrough of the first prototype. User snaps an image to send as a query (left). The result is displayed as a thumbnail mosaic (center). Selecting a thumbnail image brings up a source webpage for browsing (right).

To find similar landmark images, it would not be useful to search images that do not contain any landmark (e.g. faces, animals, or logos). Thus, an image classifier was used to classify these 33,000 images as landmark or non-landmark. The non-landmark images were then removed from the database to reduce the search-space to 2000+ images. The image classifier was trained using a method similar to [33] that was demonstrated successfully in classifying indoor-outdoor images by examining color and texture characteristics

The same volunteers were then asked to evaluate the performance of these algorithms. For each test query, the search result consisted of the 16 closest images in the database to the query image using the selected distance measure (color histogram or Fourier transform). Each candidate image was evaluated to be *relevant* if it was pertinent to the same location as the query image, or *similar* if it belonged to a different location but exhibited similar characteristics as the query image.

Unfortunately, the underlying MMS-based communication infrastructure had in practice an average turnaround time of 25sec, which required much patience of the user. We expect as MMS becomes more popular, with better technology implemented by wireless service carriers, the turnaround time can be drastically improved. We are also currently exploring image compression methods and better network protocols to improve interactive performance.

Figure 3-2 summarizes the performance we obtained from the tested image matching metrics; on average each image in the top 16 was relevant over 1/3 of the time, which is encouraging given the simplicity of our CBIR techniques. Even with such a relatively low individual relevance rate it is very likely that a user will find a relevant match interactively on the first result screen. Users can quickly discount the approximately 1/3 of images which are irrelevant and not visually similar. Figure 3-3 shows the percentage of time a relevant image appears on the first page. Since there were an approximately equal number of relevant and similar images, we would expect that after browsing at most two or three links from the initial screen a relevant page would be found.

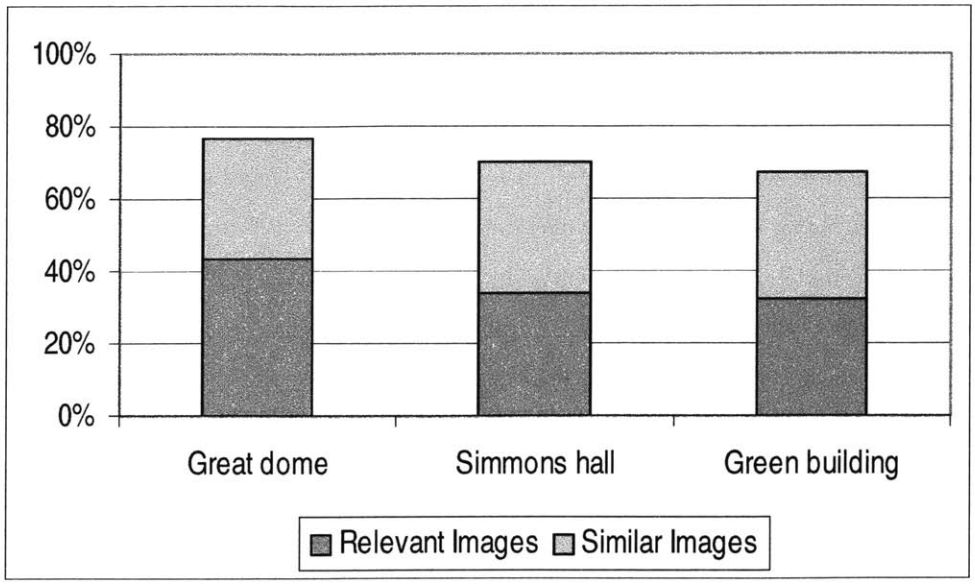


Figure 3-2: Average performance of three distance metrics presented in percentage of candidate images that are relevant or similar over the entire set of 150 test images.

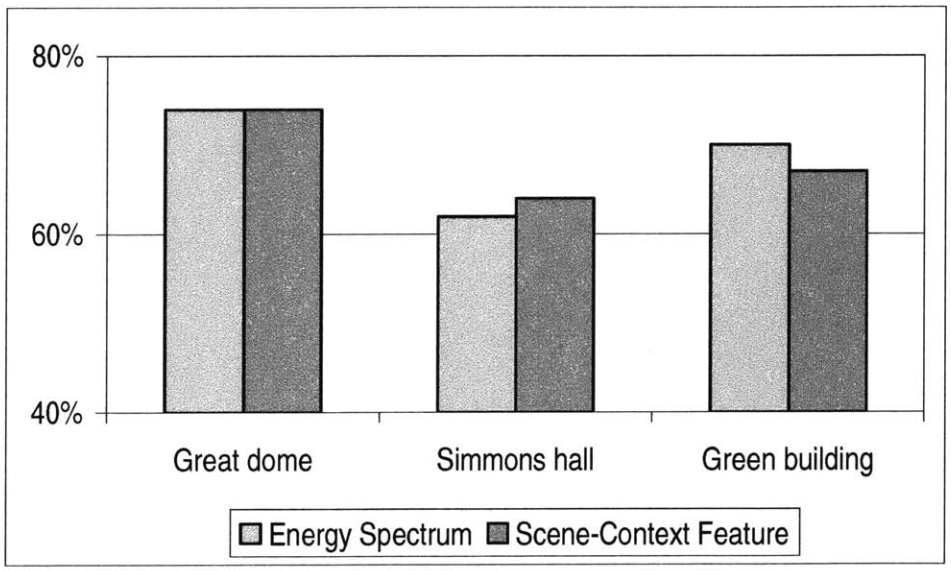


Figure 3-3: Average retrieval success for each distance metric shown in percentages of 150 test attempts that returned at least some relevant images on the first page of candidate images.

Chapter 4

Keywords Bootstrapping

In chapter 3, we saw how matching images for a location can be retrieved from an image database. From a set of candidate images, users can follow the backward links to the source webpages of these matching images to find useful information. However, it is only first part of the story. To recognize place, one has to acquire relevant knowledge of the location in more tangible, such as a set of keywords suggesting the identity of the location. This implies that the content of the web pages need to be examined. Besides helping users recognize the location, these keywords can be submitted to a text-based image search engine such as Google to take advantage of its vast image database. More images about the location can be returned; potentially, they can lead us to more information about the location. Here, we switch the platform to T-Mobile Sidekick device to develop our prototype. The first step is to capture and send an image of the location (Figure 4-1 and Figure 4-2).

Figure 4-1: A user points the camera at a building and captures an image.

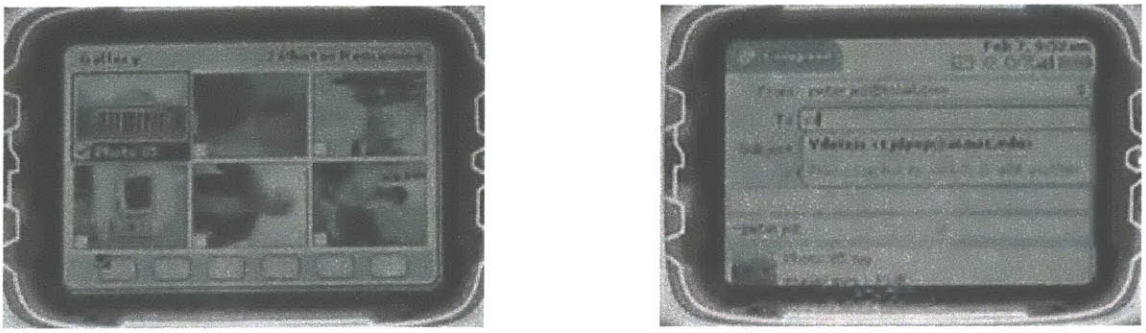


Figure 4-2: From the photo gallery the user can pick a query image (left) and send it to server (right). In this case, the trasmission method is through e-mail sent to a designated address.

4.1 Keyword Extraction

Relevant matches represent the actual location, but the content of the underlying web page can be variable (Figure 4-3). After a set of matching web pages are presented to the user in the query result, the user needs to verify whether any of the pages actually contains the information he or she is seeking. Recognizing the location in this context can as simple as identifying the name of the place or as involving as knowing interesting facts and stories about this place. One has to read the content of the page in order to determine the existence of such relevant information. A search is considered a success only if such information can be found. However, browsing web pages on a small screen can be cumbersome. The amount of information that can be displayed to the user is very limited by the relatively small screen size seen in most mobile devices on the market. With automatic text extraction techniques, a condensed version of the entire page can be derived either in the form of a short summarizing paragraph or a set of keywords. The amount of information is reduced to a more manageable manner; less cognitive effort is required of the user to identify relevant information.

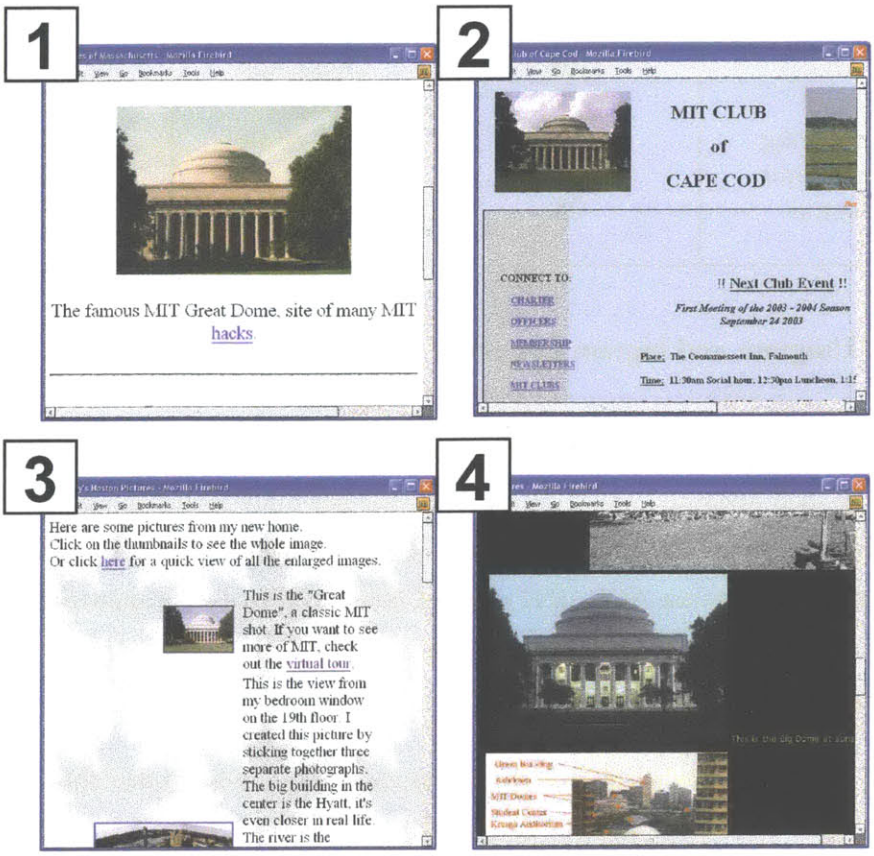


Figure 4-3: Some examples of found webpages: (1) MIT Gallery of Hacks, (2) MIT Club of Cape Cod's official website, (3) Someone's picture gallery, and (4) Someone's guide on Boston.

Simple word count is sometimes sufficient for document overview. In this simplest method, we extract the *k* most frequent terms in a set of documents. Since names of locations often take the form of bi-grams or tri-grams, we also look for *n*-grams made up of the *k* most frequent unigrams (Figure 4-4).

Unigram	Freq	Bigram	Freq
MIT	128	foundation	3
story	59	relations	2
engineering	33	MIT dome	2
kruckmeyer	29	da lucha	2
boston	28	view realvideo	2

Figure 4-4: Unigram and bigram frequency in the result of a sample query.

To be more sophisticated, a word can be considered a keyword if it is repeated several times in the document (high local frequency) and if it is normally infrequent (low global frequency). One simple estimate of the global frequency of a word *w* is the number of pages Google returns when a search on *w* is performed. We also use AQUAINT unigram dataset collected from approximately 1 million news articles, which contains 1.7 million unique tokens. This method is commonly known as *tfidf* (term frequency inverse document frequency)[25].

A number of simple heuristics are also used. First, we use a simple stop list that filters words and nouns that are obviously useless (e.g. HTML tags). Second, we pay attention to capitalized words since location names are usually written in capital letters. Third, some words might bear meanings strongly suggestive of locations, such as "building" and "town." We can device a simple grammar rule: whenever the word "building" is encountered, check if the word before has the first letter capitalized. This allows us to identify names such as "Green Building."

A set of keywords can be discovered in this way and scores based on frequency and application-specific heuristics can be assigned. We then pick the top *n* keywords to be returned to the user for review (Figure 4-5).

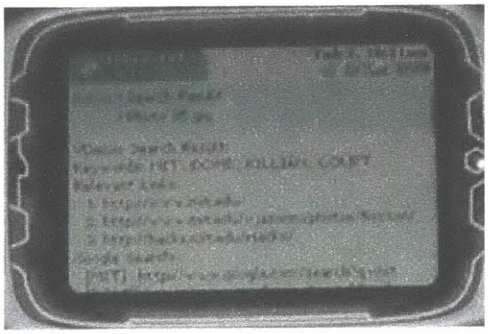
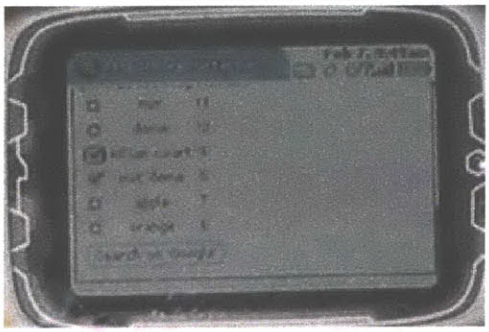


Figure 4-5: Search result. In this case, the result is contained in an email sent back to the user. It contains a set of extracted keywords as well as a list of relevant URLs.

4.2 Keyword Bootstrapping Image Search

Having uncovered a set of keywords, the users can use them to search Google either for more web pages or images. Searching for more web pages allows us to find other web pages that might share conceptual similarity with the query image but do not contain any similar image. These web pages would not have been found otherwise if only the pure image-based search were employed. This has the flavor of a hybrid approach combing both image-based and text-based search for images. We make use the observation that there will be redundant landmark images on the web and that search in a subset of the web will likely find a matching image.



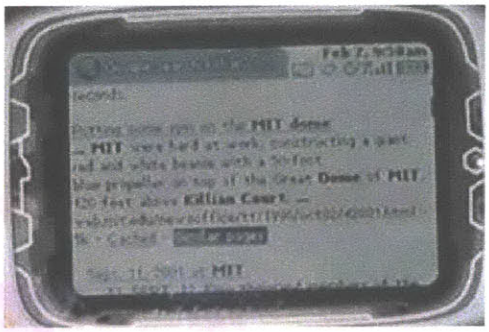


Figure 4-6: A user can select a subset of keywords to search on Google (left). Google can return a list of relevant URLs for the search keywords "MIT DOME" (right).

4.3 Content-based Filtering

Searching for more images using Google might return many visually unrelated images. Therefore, we apply a CBIR filter step to the result and keep only those images visually close to the query image under the same matching metric. Also, we can do a bottom-up, opportunistic clustering technique that iteratively merges data points to uncover visually coherent groups of images. Initially, each image is considered as a point in the feature space. First, we compute the distance between every pair of points. We then find the two closest pair of images and merge them into a cluster. Only when the distance is below a certain threshold are two points combined. These two points are removed. Their average are computed and added to the set. Another round of merging then proceeds. We repeat this process until no pair of points is close enough to be merged. If a group is reasonably large, it means the images in this group represent some potentially significant common concept

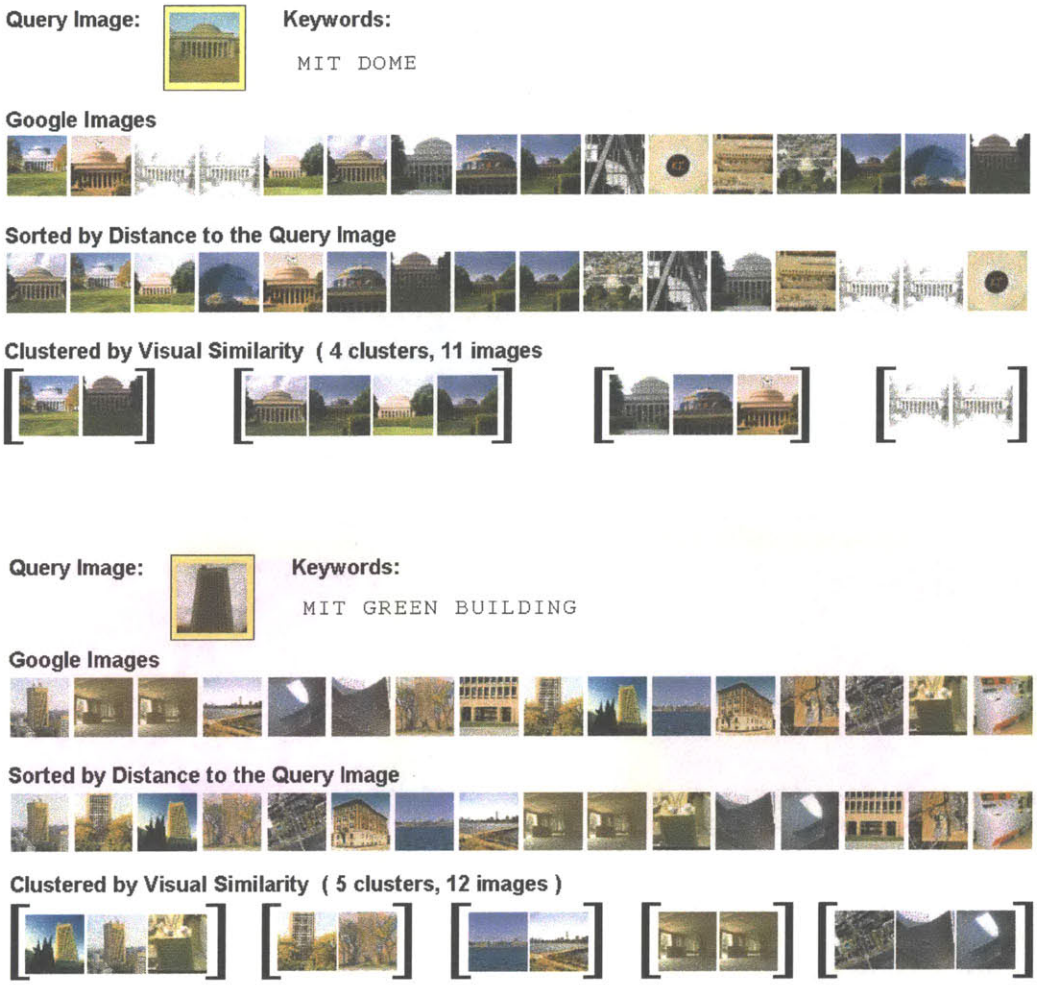


Figure 4-7: Content-based filtering the search result.

Two examples are shown in Figure 4-7. The keywords are selected by the user from the *k* best keywords suggested by the automatic keyword extraction module. They are submitted to Google to retrieve a set of images that are textually relevant but not quite visually similar. The distance metric between the query image and each Google image is computed. We sort the result by distance in increasing order. Alternatively, visually similar images in the Google set can be clustered. Some of the images are left out of any cluster because they are too distinct.

Chapter 5

Conclusions and Future Work

We have proposed a new image-based paradigm for location-aware computing in which users select a desired place by taking a photograph of the location. Relevant web pages are found using automatic image-search on the web or other large database. In contrast to conventional approaches to location detection, our method can refer to distant locations and does not require any physical infrastructure beyond mobile Internet service.

While we have demonstrated the success of several CBIR techniques for our IDeixis system, it remains a topic of ongoing research to find the optimal algorithm to support searching the web with images acquired by camera-equipped mobile devices. We are exploring machine learning techniques which can adaptively find the best distance metric in a particular landmark domain. Additional study is needed to evaluate the performance of mobile CBIR with a range of different camera systems, and under a wider range of weather conditions (e.g., we did not consider snow in our studies to date.)

Currently the cameras on our prototypes have a fixed focal length and there is no simple way to specify a very distant location that occupies a small field of view in the viewfinder. We plan to augment our interface with a digital zoom function to overcome this, and/or use camera phones with adjustable zoom lenses when they become available.

As discussed earlier, even with image-based search of location-based information, additional context will be needed for some specific searches. A human-voice driven input module is the most attractive option since we envision such system will materialize in camera phones in the future, making possible for users to specify exactly what they want to know about this location (e.g. "Show me a directory of this building!").

Moreover, it is likely that mobile devices in the near future will incorporate both camera technology as well as GPS technology, and that geographic information of some form may become a common meta-data tag on certain web homepages. Given these, we could easily deploy a hybrid system, which would restrict image matching to pages that refer to a limited geographic region, dramatically reducing the search complexity of the image matching stage and presumably improving performance.

Further study is needed to evaluate the usability of the overall method in various contexts beyond tourist and travel information browsing, including how to best present browsing interfaces for particular search tasks.

References

[1] Arasu, A., Cho, J., Garcia-Molina, J., Paepcke, A., and Raghavan, S., Searching the web. *ACM Transactions on Internet Technology*, 1( ):2–43, 2001.

[2] Arasu, A., Searching the Web. *ACM Transactions on Internet Technology*, 1(1), Aug. 2001, Pages 2–43.

[3] Aslandogan, Y.A., and Yu, C., Multiple evidence combination in image retrieval: Diogenes searches for people on the web. In *Proc. of ACM SIGIR 2000*, 88-95.

[4] Built-in camera mobile phones increasingly used for business purposes. *Nikkei Net Business*, Dec. 9 2002.

[5] Cheverst, K., Davies, N., Mitchell, K., and Efstratiou, C., Developing a context-aware electronic tourist guide: some issues and experiences. *Proc. of CHI 2000*, 17-24.

[6] Corbis, Corbis: stock photography and digital pictures, http://www.corbis.com/

[7] FaceIT, www.visionics.com.

[8] Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., and Dom B. Query by image and video content: the QBIC system. *IEEE Computer*, 28(9):23–32, 1995.

[9] Freeman, W.T., and Adelson, E. H., The design and use of steerable filters. IEEE Trans. Pattern Analysis and Machine Intelligence, 13(9):891–906, 1991.

[10] Gausemeier I., and Bruederlin B., Development of a real time image based object recognition method for mobile AR-devices. In *Proc. of the 2nd AFRIGRAPH*, 2003, 133-139.

[11] Google, Google Technology, http://www.google.com/technology/.

[12] Gorkani, M.M. and Picard, R.W. Texture orientation for sorting photos "at a glance". In *Proc. ICPR*, 1:459-464, 1994.

[13] Jansen, B. J., Spink, A., Bateman, J., and Saracevic, T., Real life information retrieval: A study of user queries on the Web. *ACM SIGIR Forum 32*, 1 (1998), 5-17.

[14] Kaasinen, E., User needs for location-aware mobile services, *Personal and Ubiquitous Computing* (2003) 7: 70–79.

[15] Kosecka, J., Zhou, L., Barber, P. and Duric, Z., Qualitative Image-Based Localization in Indoors Environments, *CVPR 2003*, (To Appear).

[16] Laaksonen, J., Koskela, M., Laakso, S., and Oja, E., Pic-SOM: Content-based image retrieval with self-organizing maps. *Pattern Recognition Letters*, 21(13-14):1199–1207.

[17] Lew, M., Lempinen, K., and Huijsmans, N., Webcrawling using sketches. pages 77–84, 1997.

[18] Marmasse N. and Schmandt C., Location-aware information delivery with commotion. *HUC 2000 Proceedings*, 157-171.

[19] Montemerlo, M., Thrun, S., Koller, D. and Wegbreit, B. FastSLAM: A factored solution to the simultaneous localization and mapping problem, In *Proc. of the AAAI National Conference on Artificial Intelligence*, 2002.

[20] Morgan, M.J., Ross, J., and Hayes, A. The relative importance of local phase and local amplitude in patchwise image reconstruction. *Biological Cybernetics,* 65:113-119, 1991.

[21] 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, 173-187.

[22] Nokia, Next Generation Mobile Browsing, http://www.nokia.com/nokia/0,,32912,00.html

[23] Oliva, A. and Torralba, A. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):147–175.

[24] Pilu, M. and Pollard, S. A light-weight text image processing method for handheld embedded cameras. *Tech. Report*, HP Laboratories, March 2002.

[25] Salton, G., Buckley, C., Weighting Approaches in Automatic Text Retrieval, *Information Processing and Management*, Vol. 24, No. 5, pages 513-523, 1988.

[26] Schiele, B., Jebara, T., and Oliver, N. Sensory-augmented computing: wearing the museum's guide. *Micro. IEEE*, 21(3), 2001, 44-52.

[27] Sclaroff, S., Taycher, L., and LaCascia, M., ImageRover: A content-based image browser for the world wide web. Technical Report 1997-005, 24, 1997.

[28] Smith, J. and Chang, S., An Image and Video Search Engine for the World-Wide Web, *Symposium on Electronic Imaging*, IS&T/SPIE, Feb. 1997.

[29] Smith, J.R., and Chang, S.F., Searching for images and videos on the WWW, *Multimedia Systems*, 1995, 3-14.

[30] Starner, T., Schiele, B. and Pentland, A. Visual Contextual Awareness in Wearable Computing. In Proc. of ISWC, 1998, 50-57.

[31] Swain, M., & Ballard D., Indexing via color histograms. Intern. Journal of Computer Vision, 1990, 390-393.

[32] Symbian OS – The Mobile Operating System, http://www.symbian.com/.

[33] Szummer, M., and Picard, R. W., Indoor-outdoor image classification. In IEEE International Workshop on Content-based Access of Image and Video Databases, in conjunction with ICCV'98, pages 42–51, 1998.

[34] Torralba, A., Murphy, K.P., Freeman, W.T., and Rubin, M.A., Context-based vision system for place and object recognition. In *Proc. of ICCV*, pages 273–280, 2003.