Searching the Web with Mobile Images for Location Recognition

Tom Yeh, Konrad Tollmar, Trevor Darrell
MIT CSAIL
Cambridge, MA 02319

Abstract

In this paper, we describe an approach to recognizing location from mobile devices using image-based web search. We demonstrate the usefulness of common image search metrics applied on images captured with a camera-equipped mobile device to find matching images on the World Wide Web or other general-purpose databases. 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. It is thus possible to efficiently search hundreds of millions of images that are not only textually related but also visually relevant. We demonstrate our approach on an application allowing users to browse web pages matching the image of a nearby location.

1. Introduction

Content-based image retrieval (CBIR) has been an ongoing research topic for the past decade. One of the difficulties reported by users is that queries are hard to specify with the conventional input modalities; unlike keywords, a visual query may require the user to sketch or select parameter values for each low-level feature [10]. 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.

We apply this paradigm to location recognition. Location recognition with radio or other signals (e.g., GPS, RFID tags, etc.) is possible, but 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 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.

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 with 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 a 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 the keyword-based search (see Figure 1).

In this paper, we focus on one particular class of images—landmarks for recognizing location. The goal is to recognize the location by finding web pages with an image matching the image of the location. In the next section we will review relevant related work. Following that, we present our approach to keyword and image-based web search and detail the component modules of our method. We present experiments on a tourist location recognition task to evaluate the components of our approach. We conclude with summary of our contribution.

2. Previous Work

Most commercially successful image search engines are text-based. Corbis features a private database of millions of high-quality photographs and artwork that are manually tagged with keywords and organized into categories. Google has indexed more than 425 million web pages and

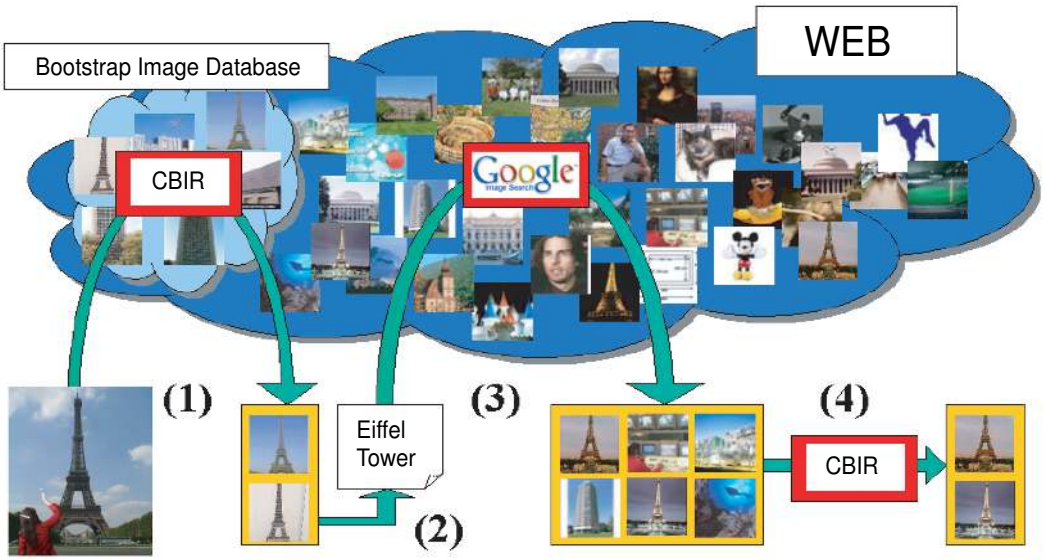


Figure 1: Image-and-keyword hybrid search. In step 1, the user takes an image with a camera phone, which is used as a query to find similar images from a small image database using CBIR techniques. In step 2, keywords are automatically extracted. In step 3, extracted keywords are sent to Google to find textually related images. In step 4, content-based image matching techniques are applied to identify visually relevant images.

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. 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.

Searching for images from images is often called content-based image retrieval and has been an ongoing research topic for many years [10]. 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 [3]. 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 [11] 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 [2]. 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.

Describing what images to search for can be tricky. QBIC lets users choose the relative weights of each feature in determining image similarity. Lew et. al proposed a system that allows users to sketch the outline of the image [6]. VisualSeek features an interface to paint a color template for template-based image matching [11]. 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 [9].

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 [5] 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 [6, 11, 9, 2, 3].

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 [12]. 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 [13]. 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.

3. Image-based Web Search

Web authors tend to include semantically related text and images on web pages. 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 [1]. To find information about a well-known landmark, we can search for web pages with images matching 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, they can serve as a useful tool for mobile web search (and in the particular application we consider here, location-based computing.)

With a camera phone, a user can take an picture of a prominent landmark around the current location. This picture can be used as a query to find images similar to this landmark on the web using content-based image matching techniques. 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.

For a pure CBIR system to match against the millions of images on the web in real-time is currently computationally infeasible. However, using a hybrid keyword-and-image query system, 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 the power of both 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 million 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.

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. 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 two image-matching metrics on the task of matching mobile location images to images on the World Wide Web. The first metric is based on the *energy spectrum*, the squared magnitude of the windowed Fourier transform of an image. Given an image i, the discrete Fourier transform can computed as

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

where $h(x, y)$ is a circular hanning window to reduce the boundary effect. The amplitude component of the above term represents the spatial frequency spread everywhere in the image. It embeds unlocalized information about the image structure such as orientation, smoothness, length, and width of the contours that compose the entire scene in the

image. This type of representation was demonstrated by [13] to be invariant to object arrangement and object identity.

The second metric is based on wavelet decompositions. We compute the local texture features of an image I by convolving it with steerable filters [4] over 2 different scales on the intensity (grayscale) image to obtain a representation as follows:

$$\lambda = G_{\theta_i}(S_j(I))$$

where G is the steerable filter for 6 different θs, evenly spaced between 0 and 2π, and S is the scaling operator. This can produce $6 \times 2 = 12$ subbands. 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 global image properties

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

where w is the averaging window. The final representation is downsampled to have a spatial resolution of 4 by 4 pixels resulting in a dimensionality of $4 \times 4 \times 12 = 96$.

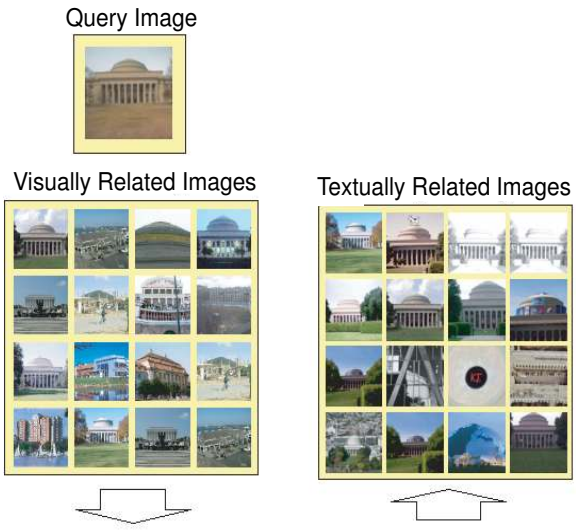
Given a mobile image of some landmark, similar images can be retrieved by finding the k nearest neighbors in the database using either of the two metrics (e.g. see Figure 2, where $k = 16$). However, the high dimensionality (d) of the feature involved in the metric can be problematic. To reduce the dimensionality, we computed principal components over a large number of landmark images on the web. Then, each feature vector can be projected onto the first 32 (a number picked empirically) principal components which form the final feature vector for the images.

3.2. Extracting Useful Textual Information

After similar landmark images are found, we need to analyze the text of the web documents containing these images with the intention to uncover any useful information about this particular landmark. We devise a simple frequency-based approach to identifying named-entities (e.g. name, location, telephone number) by computing the *term frequency inverse document frequency* metric μ for each bigram and trigram found in the document:

$$\mu(w) = \frac{df(w)}{tf(w)}$$

where $tf(w)$ is the global frequency of w over a large corpus and $df(w)$ is the local frequency of w within the document [8]. The underlying intuition is that people tend to find the sudden occurrence (high df) of uncommon things (low tf) interesting. A good estimate of the global word frequency can be the number of links Google indexes for this word. The top n word combinations with the highest f can be reported as the potential named-entities to the users.



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

Figure 2: A typical search scenario: query image to visually similar images to named-entities to textually relevant Google images.

Figure 2 illustrates a typical search scenario. A user initiates a search request by taking an image of some prominent landmark (e.g. MIT Dome) at the current location. Sixteen visually similar images to the query image are found in the web image database. Five of them are correct (1,3,4,9, and 14). The content of the source HTML documents of these 5 images are analyzed and potential named-entities are found and ranked. These named-entities can be submitted as keywords to a text-based search engine (e.g. Google) to find more images. However, it can not be guaranteed that these images are visually similar to the questioned landmark.

3.3. Content-filtered Keyword Search

The name-entities extracted from the web pages can provide the users with not only the basic notion about this location but also the possibility to acquire further information from a powerful text-based search engine (e.g. Google).

One possibility is to search for more web pages that might lack any visual content but nevertheless indirectly share similarity with the query image through text. We can describe such similarity inference notationally:

$$I_q \underset{image}{\longleftrightarrow} I_v, P_v \underset{text}{\longleftrightarrow} P_t \Longrightarrow I_q \longleftrightarrow P_t$$

where I_q denotes the query image, I_v denotes the found similar image, P_v denotes the web page containing I_v, P_v denotes a web page containing similar keywords as P_t, and $\leftrightarrow$ denotes a similarity relation based on either text or image. Here, P_t would not have been found otherwise if either of the pure image-based search or pure text-based search were employed.

Another possibility is to use the keywords to search for images I_t. The similarity inference can be denoted as:

$$I_q \underset{image}{\longleftrightarrow} I_v, P_v \underset{text}{\longleftrightarrow} I_t \Longrightarrow I_q \longleftrightarrow I_t.$$

While I_q and I_t are textually related, to guarantee visual similarity, a content-based filtering step needs to be applied to keep keep only those images visually close to the query image. As shown in Figure 4, a search on the query image I_q (a 15 floor building) yields the keywords "MIT Green Building". These keywords are submitted to Google to retrieve 16 such I_ts that are textually relevant but not quite visually similar. Sorting each I_t according to its distance to I_q, we are able to boost the ranks of the visually similar images to the top.

Moreover, there might exist images visually distant but conceptually close to the query image; they can help us learn more about this location. We can use a bottom-up, opportunistic clustering technique that iteratively merges data points to uncover visually coherent groups of images. Initially, each image is considered as the center in its own cluster. At each iteration, two closest clusters, in terms of the same distance metric used in other parts of the system, are merged. The process will continue until the distance between every cluster pair has exceeded a certain threshold. If a cluster is reasonably large, it means the images in this group represent some potentially significant common concept. In the example shown in Figure 4, four images are removed from the result set because they can not be clustered with any other image.

4. Experiments

A series of experiments was conducted to evaluate the effectiveness of our algorithms for location recognition by matching mobile images to a web database. We chose the task of recognizing landmark images on a university campus. We ran a web crawler on $mit.edu$ to collect images and their source URL's to build an image database as the bootstrap set. Images too small to contain interesting location information, less than 100 by 100, were discarded. The resulting database contained roughly 33,000 images. Since dealing with images of various dimensions and sizes can be tricky, we resized each image to 100 by 100 for uniformity.

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 nonlandmark. The nonlandmark 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 [7] that was demonstrated to be successful in classifying indoor-outdoor images by examining color and texture characteristics.

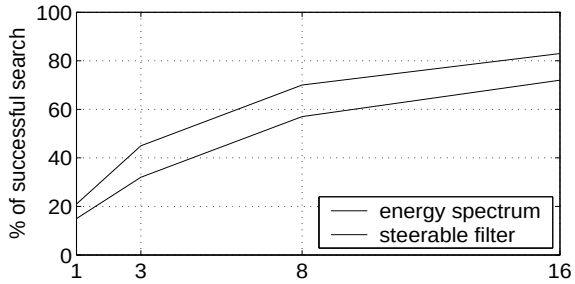


Figure 3: The performance of the matching metrics with k nearest neighbors.

Since we were concerned with location recognition in a mobile setting, the test queries consisted of landmark images acquired with a camera-equipped mobile device. We selected 5 prominent landmarks and asked volunteers to take 20 images of each of these landmarks using a Nokia 3650 camera phone. A total of 100 query images were obtained for 100 test runs.

The matching metrics discussed in section 3.1 were used in our experiments. For each query image, the k nearest images were retrieved from the database to form the search result. The same volunteers were asked to evaluate the search result. A success was defined as the occurrence of at least one correct landmark image among the first k nearest images. The findings are summarized in Figure 3, which plots the percentage of test cases that were successful over different k.

At $k = 16$, the likelihood of finding correct images in the result set is reasonably high. This indicates that even with these standard feature sets it is possible to find similar location images for recognition tasks.

5. Conclusions and Future Works

In this paper, we explored the potential of web search with mobile images using both images and text. We have shown it is possible to conduct fast and comprehensive CBIR over hundreds of millions of images using a text-based search engine on keywords generated from an initial search. We presented an approach to recognizing location from mobile devices using image-based web search, and showed that common image search metrics can match images captured with a camera-equipped mobile device to images found on the

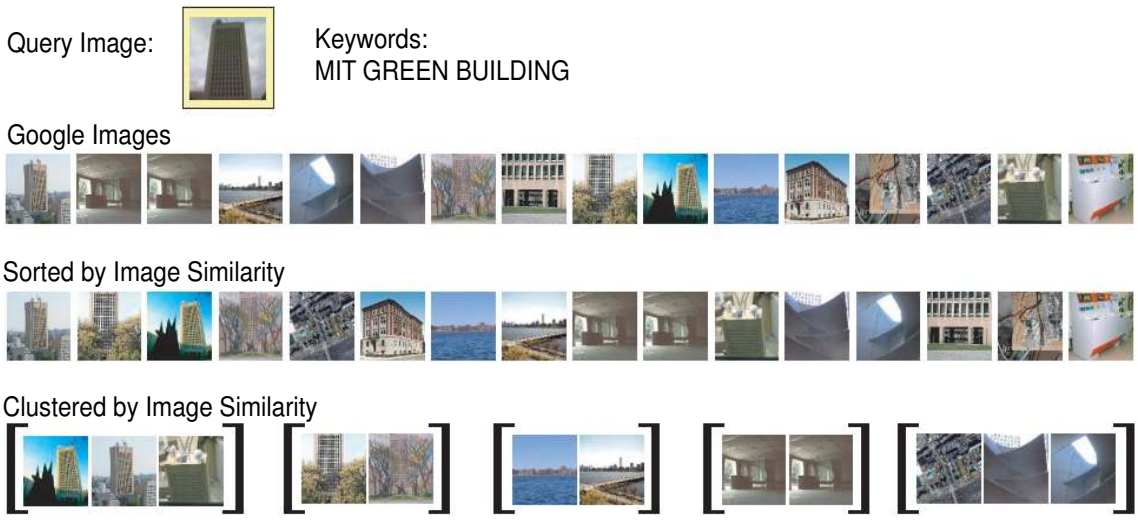


Figure 4: Content filtered keyword search for images. The keywords found for the query image are submitted to Google and a set of textually released images are found. These images can be further filtered according to their visual similarity to the query image either by sorting or clustering.

World Wide Web or other general-purpose databases.

A hybrid image-and-keyword searching technique was developed that first performed an image-based search over images and links to their source web pages in a bootstrap database that indexes only a small fraction of the web and a simple frequency-based procedure was employed to extract relevant keywords from these web pages. These keywords can be submitted to an existing text-based search engine (e.g. Google) that indexes a much larger portion of the web. The resulting image set is then filtered to retain images close to the original query. It is thus possible to efficiently search hundreds of millions of images, retrieving those that are not only textually related but also visually relevant. We demonstrated our approach on an application allowing users to search information on the web by matching images.

In the future, we would like to explore more sophisticated image matching techniques and adapt them to the landmark image domain. In addition, we plan to carry out an extensive quantitative evaluation of our approach and an overall usability user study of the system.

References

[1] A. Arasu, J. Cho, H. Garcia-Molina, A. Paepcke, and S. Raghavan. Searching the web. *ACM Transactions on Internet Technology*, 1(1):2–43, 7 2001.

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

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

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

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

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

[7] Szummer M. and R. W. Picard. Indoor-outdoor image classification. In *IEEE Intern. Workshop on Content-based Access of Image and Video Databases*, pages 42–51, 1998.

[8] G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.

[9] Stan Sclaroff, Leonid Taycher, and Marco LaCascia. ImageRover: A content-based image browser for the world wide web. Technical Report 1997-005, 24, 1997.

[10] Worring M. Santini S. Gupta A. Smeulder, A. and R. Jain. Cbir at the end of the early years, Dec 2000.

[11] John R. Smith and Shih-Fu Chang. Image and video search engine for the World Wide Web. In *Proc. of SPIE)*, pages 84–95, 1997.

[12] T. Starner, B. Schiele, and A. Pentland. Visual contextual awareness in wearable computing. In *Proc. of Visual Contextual Awareness in Wearable Computing*, pages 50–57, 1998.

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