This paper presents an approach that generates a set of search queries from a user’s query image. The approach provides the ability to query one’s environment in real-time, effectively allowing a user to ask "What is that?". This is achieved by capturing the user’s context through a geotagged photograph, and using it to filter a community image collection, for example Flickr, to retrieve a set of descriptive tags; these tags are processed and used as query terms. Having discussed the role of context in the application and the service’s architecture, an initial study of the benefits of this approach will be presented.

Categories and Subject Descriptors
H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—information filtering, query formulation; H.3.5 [Information Storage and Retrieval]: On-line Information Services—web-based services

General Terms
Human Factors

Keywords
Query-by-image, mobile search, context-in-search, location based search

1. INTRODUCTION

When we are in an unfamiliar place, we often query our environment - through a friend or guide - by pointing and asking "What is that?". When we are alone, we look up a guidebook for the area. These are usually organized by location; on examining the section for the searcher’s current location, a set of photographs and maps is examined until the subject in question is found; this process uses the visitor’s current context as a means of querying an information source. By combining inputs from smart-phones and shared community images (and their associated tags), the results of the latter technique can be achieved with the ease of use of the former. The phone (in particular its camera), equipped with a wide range of sensors, provides a context for the user; this can be used to filter an image collection and retrieve a set of tags that describe the subject of interest. These are then used to query some information source.

This type of application is called augmented reality (AR); these are programs that embellish a view of the physical world with digital information. Many of these are based on a browsing metaphor; information is serendipitously discovered by the user. Layar1, for example, provides a framework to place icons representing geo-tagged digital artifacts; each icon represents a piece of information related to the searcher’s environment, for example, youtube videos, or tweets. Recently, research has begun to explore how search might work with AR; applications such as Google Goggles2 and the work of Gammeter et al[1] have enabled users to query their environment - in both of those cases through images. While the goals of these contributions are similar in intent to that described herein, they have focussed on creating extensive searchable indices. This paper focusses on how context can be used to rationalize the search space on the fly. A relatively small subset of candidate images and their associated tags are retrieved and processed in real-time from the image-base, substantially reducing the need to implement expensive hardware and software infrastructures.

Having examined the role of context, the remainder of this paper will describe the service’s architecture before presenting the results of an evaluation of its performance and accuracy.

2. THE ROLE OF CONTEXT

The accuracy of the system described herein is achieved by using both the user’s context and that of those contributing knowledge to the community image-base. The seeker’s context encapsulates the circumstance of an information seeking task, capturing the relationship and impact of the task being undertaken, the system being used and the user’s existing knowledge that can influence his use of the search system[2]; it is this final component of context that is weakest when trying to formulate a query about an unfamiliar scene. Research mentioned in Section 4 suggests that a stranger to an area will typically describe a building, or some other point of interest, in terms of its physical characteristics and any location information.

1http://www.layar.com/
2http://www.google.com/mobile/goggles/
3. IMPLEMENTATION

Producing a list of terms that describe an object in a query photograph was accomplished by taking fast image matching algorithms and combining them with the power of social networks and smart-phones. The results of the service provide an information seeker with an entry point to conduct information seeking about a photographed object. Some restrictions apply - the object has to have a fixed position, for example, a building or statue and the object must be of interest enough to have been photographed and tagged a-priori. In addition, the camera that takes the query image must be capable of geo-tagging the image; many smart-phones are capable of this. These devices are also capable of uploading the image to the service, and provide a means of viewing results. This set of abilities allows an seeker to use their smart-phone to resolve serendipitous curiosities and explore their environment in-situ.

The process that translates a photograph into a set of terms is shown in Figure 1. In summary, the service takes an uploaded query image, extracts its geographical metadata and uses this to retrieve a set of images from an image-base. Because photographs of several objects of interest may be returned by the image-base, quick image matching is conducted to find a set of matching images. Tags associated with these images are then retrieved and ranked. These textual tags can then be used to query an information source. The authors will now examine each of these steps in greater detail.

Several metadata standards exist; the iPhone (the target device for the prototype client for this service) uses Exchangeable Image File Format (EXIF). Many attributes are recorded in this metadata when a photograph is taken on the iPhone, including the longitude and latitude of the device taking the photograph, the date when the photograph was taken and other camera attributes such as aperture and focal length. This admits many dimensions across which exploration can occur. Many smart-phones use a combination of signal triangulation and GPS (Global Positioning Service) to determine what position is recorded. While it is generally accurate, there are circumstances under which it is not. In addition, it is the location of the device is recorded when a photograph is taken; if the photographed object is large, this position might be some distance from the object. To account for these potential errors, images are retrieved from an area 100 meters square centered on the recorded location. This bounding box is submitted to several image-bases including Flickr and Panoramio and a set of candidate images are retrieved.

This set of candidate images may contain photographs whose subject is something other than that of the query photograph’s. To ensure that only the query image’s subject is considered, the candidates are filtered using an implementation of the Scale Invariant Feature Transform (SIFT) image matching algorithm[3]. The algorithm uses distinctive invariant features in an image to match against different views of a scene. It is an efficient approach, robust to distortion, scale and noise. Once this step has executed, the set of candidate images is reduced to a set of matching images.

Once a set of matching images has been found, tags associated with these images are retrieved from the image-base. While tags include descriptions of the location, they often reflect the personal context of the user including their names, details of the camera used and the occasion on which the photograph was taken; the tags describing the photograph should reflect generic information and filter out those that reflect personal context. After analysis of the tagging habits of users, it was determined that tags that describe personal contexts are common across numerous photographs, though they are unique to the photographer. As a result, tags were grouped into meta-documents on a per-user basis, and the TF/IDF statistical information retrieval technique[4] was

Figure 1: Service Architecture.
applied to these. This resulted in terms that were common across the collection of documents (and thus users) emerging; these are ranked by popularity. While some relevant terms’ significance was easily determined, there were others that were ranked lower, but were considered very descriptive. A second pass of relevance measurement is completed to capture these terms. For each of these terms, a Yahoo! Boss\textsuperscript{5} query is formed by AND-ing each term with the significant terms, and measuring the number of documents that are returned. While tags that describe personal context are rarely associated with the significant descriptive tags, minor descriptive terms appear in many documents with the significant terms. A list of the top ten descriptive terms is returned by the service. This list can be used to form queries to an information source; this can be a general source, Google or Wikipedia, or a custom information source.

The implementation of the service instance used during the evaluation was architected thus; a service mediator marshaled each request; it was responsible for accepting requests, making the initial request to the image-base (Flickr), sending requests for work to distributed agents and processing their responses into a query. The distributed agents, reserved lab machines, were responsible for processing each candidate image and its tags. Both the mediator and ten agents were Apple iMac 2.66 GHz Intel Core Duo machines running 2GB of RAM. The distributed network was enabled using XServe.

### 4. RESULTS

During development of the service, a simple iPhone client was developed and deployed to several devices. This application allowed a user to take a query image, upload it to the service server and displayed the set of query tags that were returned for the query. Two examples are shown, one instance from Ireland (Figure 2), and one instance that was submitted when visiting Croatia (Figure 3). When the returned terms are AND-ed and submitted to Google, the first page of results contains Wikipedia articles about each location. The service was formally evaluated based on its ability to efficiently produce an accurate query. The quality of the service’s queries were also evaluated by comparing their precision to those created by human ‘visitors’ to a location.

**Performance**

Queries executed with an average time of 9.1 seconds. This time is reduced as more distributed agents are attached to the service. The greatest limiter to performance was the XML handling that resulted from interactions with the image-bases. This code is currently not fully optimized so performance gains are expected.

**Quality**

During the development of this service, the authors sought to understand how users would formulate a query when tasked with identifying a building. Using the crowd-sourcing tool, Amazon mTurk\textsuperscript{6}, a series of ten images of local landmarks were presented along with their general location (Cork City, Ireland). Two hundred participants we asked to provide the initial query they would use to identify the building. When these queries were submitted to Google’s search engine, 40% of queries returned a site on the first page of results.
Development testing exposed some interesting social habits. It demonstrated a tendency towards a social curiosity. While testers did try to foil the service by submitting images of buildings that caused no social interest, they admitted that when a location caused a genuine curiosity, the service would respond with an accurate set of query terms. This is borne out by the work of Eric Fischer7. The images he created, one of which is shown in Figure 4 (reproduced under the terms of the creative common’s license8), shows the locations of geo-tagged photographs taken in the Greater London area; images taken by local people are shown in blue, those by visitors are shown in red and those of an indeterminate provenience are shown in yellow. Red clusters visibly demonstrate that there is a natural tendency for visitors’ photographs to cluster around popular tourist locations. This is also apparent for locals, though their range of locations is much wider. The authors’ own research has shown that tags generated by local photographers tend to demonstrate greater precision. The role that context – both the user’s and contributor’s – has in improving the precision and accuracy must also be investigated; initial evaluations suggest that tags generated by local contributors are superior to those of visitors. This observation might be used to weight the terms when determining those of greater descriptive significance.

Finally, the application currently does not examine the context of the location. It too has a story to tell, for instance a location’s history and its association with historical events; already others have begun to provide a historic context for photographs9. Community image-bases also provide sufficient metadata to implement this type of exploration.

6. REFERENCES


7http://www.flickr.com/photos/walkingsf/4671589629/in/set-7215762420158632/
8http://creativecommons.org/licenses/by-sa/2.0/
9http://www.historypin.com

Figure 4: A portion of one of Eric Fischer’s Geotaggers World Atlas Maps - showing London. Locals’ images are in blue, visitors in red.