# AN INVESTIGATION ON SERVER-SIDE OBJECT-SCENE RECOGNITION PERFORMANCE USING COARSE LOCATION INFORMATION AND CAMERA PHONE-CAPTURED IMAGES

Christopher Mangion
*Vodafone Malta Limited,*
*and*
*St. Martin's Institute of I.T.*
*Email: christopher.mangion@vodafone.com*

Kenneth P. Camilleri
*Department of Systems and Control Engineering, University of Malta*
*and*
*St. Martin's Institute of I.T.*
*Email: kpcami@eng.um.edu.mt*

## ABSTRACT

*This paper presents a solution based on information already residing within a mobile network and aimed at the cultural tourist. It also demonstrates how scene (or landmark) recognition from an image can be achieved by combining local invariant image features, cell location information and classification based on Self-Organizing Map clustering. The proposed server-side approach makes the solution independent of the mobile platform and thus accessible to any camera-embedded mobile station having the Multimedia Messaging Service enabled.*

## 1. INTRODUCTION

Location-Based Services (LBS) designate any service that uses or provides location information. LBS refer to the capability of finding or estimating the geographical position of a mobile station (MS) within a wireless network, and subsequently provide services based on the location information.

Market estimates and forecasts from the ITU (International Telecommunication Union) suggest worldwide revenues from LBS would exceed $9.9 billion by 2010 [1]. The European Union is developing requirements for cellular operators for their *e112* emergency services [2]. This will push mobile network operators to build out the location determination infrastructure which can be exploited for other commercial purposes.

Different LBS technologies manifest different degrees of accuracy in MS position estimation. Their performance will also be influenced on several other factors, such as the geography of the area where LBS is considered, and the network characteristics. It was therefore important to review existing implementations and assess any that could be most suited to the Maltese geography in general.

With camera phones becoming ubiquitous imaging devices, new social practices of personal visual information will start to emerge. The applications for this sort of "augmented reality" are huge - from mapping and tourist information, to being able to give directions based on landmarks rather than road names and numbers. One could use one's phone to get information on almost any kind of consumer product - from CD covers, to movie posters and even wine in a local shop; by simply snapping a picture of the wine label and the phone could pull up reviews and sampling information.

Most available solutions require additional GPS (Global Positioning System) hardware for location determination. Apart from the additional hardware requirements these solutions require also the installation of specialized client software on the mobile platform.

The approach presented in this paper differs from the aforementioned in two significant ways: firstly, the proposed design does not require extra hardware to determine the location of a mobile station; and secondly, no client or third party software needs to be installed on the mobile platform.

## 2. CONTEXT AWARENESS

In cases where recognition cannot be accomplished quickly from just the target's physical attributes, contextual information can provide more relevant input for the recognition of that object than can its intrinsic properties.

Our brains can distinguish even subtle differences so that, for instance, physically similar objects are still labeled as different objects (for example, a cell phone and a calculator), and two visually different objects can still be given the same basic level name (for example, a cell phone and a telephone). These relationships are usually explored using *"priming"*, where improvement in performance after exposure to a stimulus can reveal the characteristics of underlying representations and their relations.

In this case, priming is provided by location context ("location-based") – in theory this should improve image retrieval precision when compared to a matching query with all images in the data set.

Existing solutions such as the *SnapToTell* system [3] use a combination of mobile image content and GPS coordinates to provide an "image-based mobile tour guide". It adopts a 3-tier (i.e. Client-Server-Database) approach whereby the *SnapToTell* client on the mobile phone creates an MMS message encompassing the acquired location coordinates as well as the image taken by the user. This data is subsequently sent to the *SnapToTell* server for further processing.

When a person seeks information about a monument or scene, a picture of that scene can be captured with the mobile phone. A GPS receiver acts as a "context source" from which the client application installed on the mobile phone acquires the GPS co-ordinates. These are in turn tagged to the image in order to create an MMS message which is then sent to the *SnapToTell* application server.

An empirical study [4] was carried out to demonstrate the effectiveness of context in such an application. The idea of using location context reduces the image search space and improves retrieval speed. However, such "mobile tourist information" service can only be offered to those customers having a high-end mobile phone handset with built-in GPS. Moreover, the aforementioned system requires the J2ME[1] platform in order to operate its client application.

## 3. METHODS AND METHODOLOGY

GSM location-pertinent parameters are Location Area Code (LAC), Serving/Neighboring Cell-ID, Timing Advance (TA), and the measured signal strengths of the MS serving cell and possible neighbors. All these parameters are known at both the MS and the network end when the MS is in "dedicated mode" i.e. when the MS is busy in a call. However, this study assumes "idle mode" MS conditions (i.e. when the MS is free). In such a state only the LAC is known at the network's end – no details are relayed back to the network in idle mode conditions, even during a cell reselection process. The MS, on the other hand, continuously makes signal strength measurements and also knows the Cell-ID of the serving cell – the cell on which the phone is camped.

An "idle mode" MS that needs to be located in space, would need to transmit network parameters to the Location Server for its position estimation. One way to achieve this is to page the MS or initiate a dummy call to invoke the transmission of this information the moment the connection is established. LAC, Cell-ID, and TA can be used for rough position estimation.

Once a user takes a snapshot of a scene or object, an MMS with that picture is sent to a specific SMTP[2] address. Upon receipt, the MMSC[3] will take care of sending the message to the e-mail server using an MMS-to-Email gateway. Subsequently, the core application hosted on a backend application server – residing at the mobile network operator's end – acts like a daemon: whenever a new request comes in, the server-side application services it.

### 3.1 THE REFERENCE CORRELATING DATABASE

For the purpose of this study, the correlating database is the collation of empirical test measurements and cell information provided by the planning tool currently in use by Vodafone Malta, for cellular network planning and radio coverage prediction.

The data set used in all the experiments consists of 64 digital images representing 9 different sites or historical landmarks. In order to capture a wide variety of different conditions to which each site may be exposed to, pictures were taken as follows:

- three different viewing angles/positions (front, left, and right);
- for each view, three shots were taken under different lighting conditions (morning, noon, and afternoon);

[1] Java 2 Mobile Edition; [2] Simple Mail Transfer Protocol; [3] Multimedia Messaging Service Centre.

Twenty seven images were taken per site - therefore cumulatively two hundred and forty three images made up the preliminary image data set. All images – including the "test" images – were normalized to a standard resolution of 420 by 320 pixels.

To maximize performance without hindering recognition precision, a pre-processing step was introduced whereby the repertoire of 243 images was effectively represented by a "condensed" data set composed of 64 "salient" images.

### 3.2 THE SIFT OPERATOR

Numerous algorithms are employed in the field of computer vision to extract interest points (or "features") from an image – these can be subsequently used in object-scene recognition applications. These algorithms try to search for features that are relatively invariant to changes in orientation and lighting conditions, thus making it possible to find the same features in other images with different backgrounds or viewpoints.

Detection of local image regions is only the first part of the feature extraction process; the second part is the computation of descriptors to characterize the appearance of these regions. A good descriptor should be distinctive, in order to provide strong consistency constraints for image matching, yet robust to illumination changes and affine transformations (i.e. scale, rotation invariant etc.).

Some very promising results have been achieved by the *Scale-Invariant Feature Transform* (SIFT) [5, 6] developed by Lowe [7, 8], which is able to find stable image features that can be used for object recognition.

The SIFT algorithm analyses an image across scale-space by creating an image pyramid with successive Gaussian blur filters, and then calculating the difference-of-Gaussian (DoG) between two levels of the image scale-space pyramid. To generate the pyramid, the input image is repeatedly blurred and the difference between consecutive levels is then computed. The blurred image is down-sampled by a factor of two in each direction, and the process repeated generating an image with a smaller size hence the pyramid of images. Figure 1 illustrates this process in greater detail.

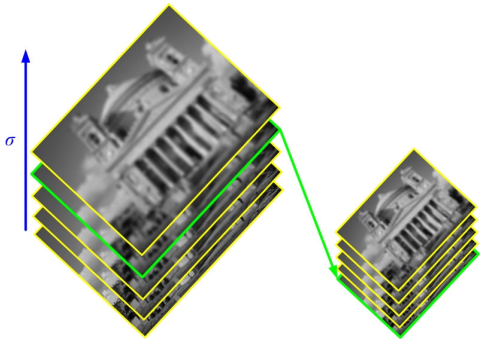


Figure 1. *A Gaussian scale-space image pyramid with s = 2 intervals.*

Keypoints are identified as local maxima or minima (extrema) of the DoG intervals (or images) across scales (or levels). Each pixel in the DoG images is compared to its 8 neighbors at the same scale, plus the 9 corresponding neighbors at neighboring scales (i.e. 9 in the interval above and 9 in the interval below). If and only if the pixel is greater than (local maximum) or less than (local minimum) all of its 28 neighbors, then it can be considered as a potential keypoint.

After assessing the detected keypoints for stability, descriptors are created at each stable location. The descriptors represent the local image data around a keypoint without compromising between "geometric invariance" and "discriminative power" requirements. The SIFT descriptor divides a square patch into a 4×4 grid and computes a 128 bin histogram of gradient orientations in each sub-region. Eight gradient orientations are used, resulting in a 4×4×8 = 128-dimensional feature vector. This vector is normalized to enhance invariance to changes in illumination. The histogram method provides stability against deformations of the image pattern, while the region sub-division offsets the potential loss of spatial information.

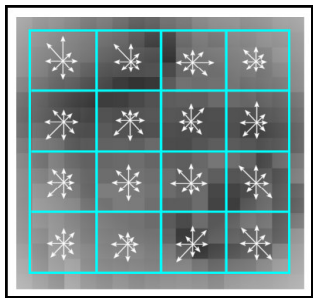

Figure 2. *The SIFT feature descriptor covers a 16×16 pixel sample area, and each histogram cell covers a 4×4 sub-region within the 16×16 sample area.*

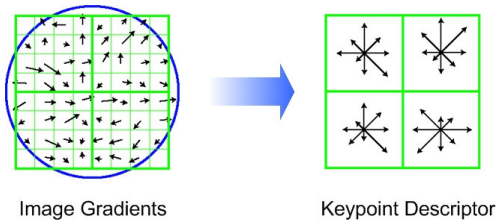


Figure 3. *A 2×2 descriptor array computed from an 8×8 set of samples.*

### 3.3 SIFT FEATURE MATCHING

The best candidate match for each keypoint is found by identifying its nearest neighbor in the database of keypoints from training images. The nearest neighbor is defined as the keypoint with minimum Euclidean distance for the invariant descriptor vector. However, many features from an image will not have any correct match in the training data set since they may contain background clutter or were not detected in the training images. For robustness, a more effective measure is obtained by comparing the distance of the closest neighbor to that of the second-closest neighbor. According to Lowe [8], if there are multiple training images of the same object, then we define the second-closest neighbor as being the closest neighbor that is known to come from a different object than the first, such as by only using images known to contain different objects.

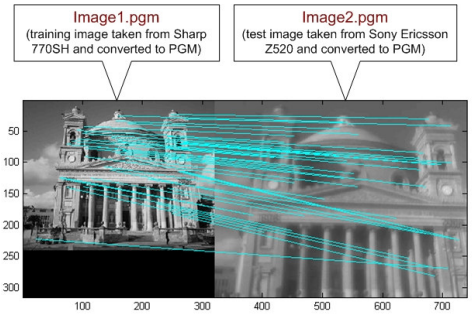


Figure 4. *The results from the matching program as visualized from MATLAB*

### 3.4 SELF-ORGANIZING MAP

Kohonen's Self-Organizing Maps are a special class of artificial neural networks that are based on competitive learning [9]. The SOM was inspired by an interesting feature of the human brain: as discovered by neuroscientists, some areas of brain tissue can be ordered according to an input signal. Therefore the brain is organized in such a way that different sensory inputs are represented by *topologically ordered computational maps* [9]. Basically, the SOM is a computer program simulating this biological ordering process. Applied to electronic datasets, the algorithm is capable of producing a map that shows similar input data items appearing close to each other.

To get activated, the network's output neurons compete among themselves, with the result that only one output neuron at a time is active. An output neuron that wins the competition is called a *winner-takes-all neuron*. As an unsupervised neural network algorithm, a SOM projects high-dimensional data onto a two-dimensional map. The projection preserves the topology of the data such that similar data items are mapped to nearby locations on the map. Therefore a SOM identifies patterns in data, clusters them into a predefined number of classes, and orders the classes in a two-dimensional output space such that nearby neighbors in the input (data) space are also close in the output space.

The Kohonen map model includes an input layer and an output layer (Figure 5). The input layer is just a flow-through layer for the input vectors, whereas the output layer consists of a two-dimensional network of neurons (or *nodes*) arranged on a grid.

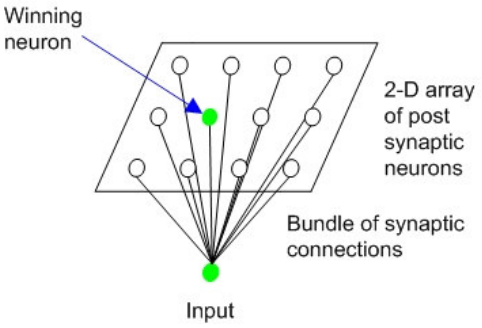


Figure 5. *Kohonen's Self-Organizing feature map model [9].*

### 3.5 CLUSTERING OF THE SELF-ORGANIZING MAP

To be able to effectively utilize the information provided by the SOM, methods that give good candidates for map unit (or map node) clusters are required. In this paper, clustering is carried out using a two-level approach. The first abstraction level is obtained by creating a set of prototype vectors using the SOM. The second abstraction level is clustering of the SOM. Clustering the SOM rather then clustering the sample data directly

has the advantage of considerably decreasing computational load, thus making it possible to cluster large data sets and to consider several pre-processing strategies in a limited time [10].

### 3.6 THE CORE JAVA SERVER-SIDE APPLICATION

Once the image file is received and the Cell-ID is resolved from the served party mobile number, the following process can then be initiated:

An image *f* and its instantaneous Cell-ID $\omega$ are queried in order to determine which site *S* corresponds to image *f*. Each site has a finite number of Cell-IDs (or a neighborhood of Cell-IDs) associated to it, which, in turn, carry the same amount of probability in terms of being a potential serving cell i.e. at any moment in time a site might have a particular Cell-ID (e.g. 10146), but in another instance a different serving Cell-ID (e.g. 10163) may be associated to it. Therefore the whole set of possible Cell-IDs related to this site must be considered. Thus, given a Cell-ID $\omega$, to determine the whole set of neighboring cells $\omega_j$ have to be determined such that the following holds:

$$( \omega \in \Omega_{Si} ) \cap ( \omega_j \in \Omega_{Si} ) \; \forall \; S_i$$

where $S_i$ is a specific site.

Two different algorithms, namely "*Match SUBSET*" and "*Match ALL*", were developed. Both methods employ the same matching algorithm but yield different output files with the intent of determining what kind of recognition performance and precision can be attained in each case.

## 4. EXPERIMENTAL RESULTS

Four different experimental approaches were setup. Two SOMs were created; one for each "Match SUBSET" and "Match ALL" experimental pair. The first SOM was trained using an input data file containing both the match counts and the relevant geo-coordinates i.e. latitude and longitude of a site, while the second SOM was trained using just the match counts as input data, that is, excluding any location-based information. Each experiment was repeated twice for the same test image using the two previously mentioned matching techniques. In each case, a pre-trained SOM was used and the time taken to generate the necessary SOM sample input data file was recorded for every run. A confusion matrix was generated to assess the accuracy of the sites' classification. Moreover, the classifier's performance was assessed by plotting each site's discrete classification in ROC (Receiver Operating Characteristic) space.

The best results were obtained for the "Feature-Location" SOM using the "Match SUBSET" technique giving an overall recognition rate of 74%. This result derives from the confusion matrix depicted in Figure 6. Site recognition was obtained in an average time of 14.301 seconds per test image

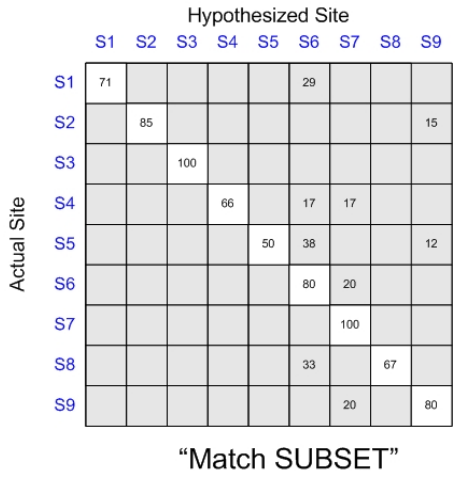


Figure 6. *Confusion matrix for the test image data set using the "Match SUBSET" technique.*

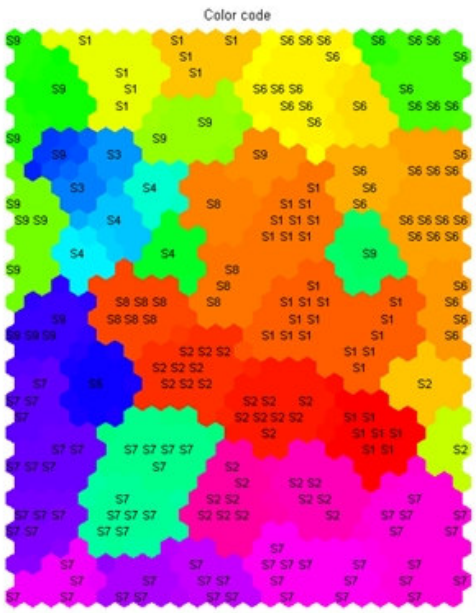


Figure 7. *SOM-based coloring for a given data/map highlights the various clusters.*

## 5. CONCLUSION AND FUTURE WORK

The obtained results are encouraging, and indicate that the proposed use of Cell IDs can improve the overall site recognition.

Since this study was investigating the feasibility of the concept, further tests would need to be carried out in order to justify whether this approach can be drawn into a fully-blown commercial solution. A broader test plan and further optimization need to be pursued in order to build a larger sample of test measurements and correlations before such a decision can be taken in the light of higher confidence in the performance of such a solution.

Furthermore, it was noticed that erroneous results are more frequently obtained for rural sites. Experimental results showed that a Mobile Station (MS) taking pictures of a landmark situated in a rural location can pick up a serving Cell-ID of a BTS installed in an urban locality. Such geo-location information from rural areas introduced noise into the input space causing incorrect hits by the classifier and consequently false results. To mitigate this adverse effect, a method which removes the geo-location information whenever a "rural" landmark is detected might help reduce such misclassifications.

Another possible improvement which could enhance recognition speed is the introduction of an automatic Region of Interest (ROI) detection algorithm. The *Statistical Region Merging Segmentation* (SRM) algorithm developed by Nielsen and Nock [11] in conjunction with Prewitt edge detection enhances the edges of the relevant objects defining the boundaries of the region of interest allowing the original image to be cropped as shown in Figure 8. This can reduce the number of descriptors generated by the SIFT algorithm, hence reducing the number of keypoints that need to be matched.

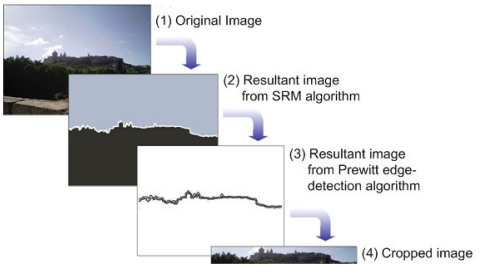


Figure 8. *Combining the SRM and Prewitt edge-detection algorithm to crop the ROI.*

Finally, from the point of view of deploying this solution commercially, additional real-life network issues, such as outages of cells, sites or parts of the network, have to be factored in.

## REFRENCES

[1] F. Liete, J. Pereira (2001) Location-based services and emergency communications in IMT- 2000. *ITU News 7, accessed October 21, 2007*, http://www.itu.int/itunews/issue/2001/07/mobility.html

[2] T. D'Roza, G. Bilchev (2003) An overview of location-based services. *BT Technology Journal*, 21(1): 20-27

[3] J.H. Lim, J.P. Chevallet, S.N. Merah. SnapToTell: Ubiquitous Information Access from Camera. *International Workshop on Mobile and Ubiquitous Information Access* (September 2004)

[4] Ramnath V., Joo-Hwee Lim, Chevallet J.P., Daqing Zhang. Harnessing location-context for content-based services in vehicular systems. *IEEE 61st Vehicular Technology Conference*, 2874 - 2878 Vol. 5, 2005

[5] Lowe, D. G. *Perceptual Organization and Visual Recognition* (Kluwer, Boston, 1985).

[6] Ullman, S. Aligning pictorial descriptions: an approach to object recognition. *Cognition* 32, 193–254, 1989.

[7] D.G. Lowe. Object recognition from local scale-invariant features. *International Conference on Computer Vision, Corfu, Greece* (September 1999), pp. 1150-1157

[8] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60, 2 (2004), pp. 91-110.

[9] Haykin S. 2005. Neural Networks, 465-473, Ed. 2, 2005.

[10] Varfis A., Versino C., Clustering of Socio-Economic Data with Kohonen Maps, *Neural Network World*, Vol. 2, no. 6, pp. 813-834, 1992.

[11] Nielsen F., Nock R. Statistical Region Merging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1452-1458 Vol. 26, Issue 11, 2004.