Visual Object Recognition in Mobile Imagery for Situated Tourist Information Systems

L. Paletta, G. Fritz; C. Seifert, P. Luley and A. Almer

JOANNEUM RESEARCH Forschungsgesellschaft mbH
Institute of Digital Image Processing, Mobile Vision Initiative
Wastiangasse 6, A-8010, Graz, Austria
lucas.paletta@joanneum.at

Abstract - We describe a system with a multi-sensor object awareness and positioning solution for augmented tourist information systems in urban areas. The system offers technology of outdoors vision based object recognition that will extend state-of-the-art location and context aware services towards object based awareness in urban environments. In the proposed application scenario, tourist pedestrians are equipped with a GPRS or UMTS capable camera-phone. They are interested whether their field of view contains tourist sights that would point to more detailed information. Multimedia type data about related history might be explored by a mobile user who is intending to learn within the urban environment. Ambient learning is in this way achieved by pointing the device towards the urban sight, capturing an image, and consequently getting information about the object on site and within the focus of attention, i.e., the user's current field of view. The described mobile system offers multiple opportunities for application in both mobile business and commerce, and is currently developed towards an industrial prototype.

Keywords: Mobile vision, outdoor object detection, tourist information systems, augmented reality.

I. INTRODUCTION

Location based services require the knowledge about the actual location of the user, the current user context and geo-referenced information about areas and points of interest. Different technologies can be used to fulfill these requirements. In general, one can distinguish to provide the digital service as either an offline or online solution. Location awareness can be provided based on GPS, using wireless network technologies such as GSM and WLAN or using self-location possibilities e.g. based on street names and house numbers. Furthermore, location awareness can be realized based on the knowledge of the location of the geo-referenced objects of interest, which allows determination of the user position.

The presentation of geo-referenced information on mobile devices requires a data transfer to the mobile device. This can be implemented as an offline service, by storing the required data on the phones' memory card, or as an online service, by transferring the data to the phone on-demand using wireless network infrastructure (e.g. GSM, GPRS, UMTS and WLAN. Both solutions have advantages and disadvantages and can be used for the development of mobile applications [1]. Location awareness for a mobile service in urban areas can, by the use of GPS only, not be assured everywhere and anytime, because of the known weaknesses of GPS signal availability in urban areas. Therefore, mobile systems operating in urban environments must take advantage of contexts arising from the spatial and situated information at a current location of the pedestrian user. Today, location based services are in principle able to provide access to rich sources of information and knowledge to the nomadic user. However, the kind of the location awareness that they do provide is not intuitive, requires reference to maps and addresses, i.e., the information is not directly mediated via the object of interest.

In contrast, the proposed system takes a decisive step towards getting in line with the user's current intention to relate information to its current sensorial experience, e.g., the object in its line of sight. In this way, the system can respond to the user's focus of attention, e.g., for the purpose of tourist information systems. A camera attached to the mobile system (camera phone, or PDA) pointing towards the object of interest (e.g., a building or a statue) will capture images on demand and would be capable of automatically finding objects in the tourist user's view. The images are then transmitted to a server that automatically extracts the object information, associates it to geo-referenced content, and sends the resulting data back to the mobile user. 'Mobile vision' is here referred to mobile visual data that are processed in an automated way to provide additional information to the nomadic client in real-time.

A location based service in urban areas has to offer area-wide location awareness to allow a spatial oriented access to information. A mobile application has to focus on the thematic requirements and also on the target groups. In the following chapters a mobile application system (see Fig. 1) will be described which is based on a common smart-phone and image based object recognition as a tool support the location awareness in combination with a GPS module.

II. USER SCENARIO

This chapter briefly describes a user scenario, in case of a city-tourist type pedestrian, focusing on the service of image based object recognition. The common way of doing city sightseeing is using a printed city map with integrated sightseeing-tours leading the tourist along a pre-defined path from one sight to the next. Brief descriptions

Fig. 1. Scenario: taking a picture of the object to be identified by the system.

of the sights can be found at the backside of the map. By the use of the image based object recognition service the tourist gets the freedom to explore the city without any pre-defined sightseeing tours.

The tourist moves completely free through an unknown area and if he is interested in any object (e.g. a historical building or a statue) he just has to take a picture of it with his camera-phone, with or without GPS device connected, and pressing the "Identify" button. As result he gets a detailed description of the object containing multi-media tourist information. As a second achievement he also gets the position of the identified object which can be used for navigation.

III.SYSTEM OVERVIEW

This chapter will briefly describe how a common smart-phone with built-in digital camera can be used for image based object recognition. A GPS device, built-in or connected to the phone via cable or Bluetooth can help to accelerate the recognition process considerably. The overall concept consists of three main phases.

In the first phase a software client is activated by a user on his personal smart phone. The software can be directly downloaded from a website with the internet enabled smart-phone and then installed on the device. The software-client offers functionality to take a picture of an object the user wants to identify. Next, if available, the smart-phone reads the actual position from the GPS device. If the GPS cannot obtain a position for any reason, the cell information of the phone-network provider can be used to approximate the user location instead. The picture of the object and the position of the snapshot-location are put together into a SOAP Message and send to the image-recognition web-service, running on a dedicated server, over an common wireless internet connection like GPRS or UMTS.

In the second phase the web-service reads the request from the client (smart-phone) and extracts the picture and the GPS position. The picture is then analysed by an image recognition algorithm to obtain representative features of the picture like edges and significant surfaces or colour transitions. Next, these features are compared with an object-database. The database contains pictures of different objects, with the pre-processed features and the position of the snapshot-point. The object is then identified by matching the features of the user-picture with the features of in the database. This matching process is, in the case of a big database with many objects, very time consuming and that is where the GPS position can help. To accelerate this process the objects in the database can be filtered with the users GPS position. Only those database objects come into account for the matching process, which have corresponding pictures with a snapshot-point near the users-position. (see Fig. 3.) Once the object is identified the object-database is queried for some information containing text and pictures. This information is integrated into a SOAP-message and is then send to the smart-phone client as response.

In the third phase the web-service response is presented to the user on the smart-phone. Additional to the object quick information containing text and pictures there is an URL which can be used by the user to obtain detailed information about the object. The URL can be viewed in a common smart-phone internet browser like "Opera" or "Pocket Internet Explorer". The URL contains, as a parameter, the unique identification number of the desired object and links to a dynamic website, which is generated at runtime on the server. The layout of the website is optimized for the requesting hardware platform – because different smart-phones have different display resolutions.
The following Figure shows the whole technical concept and its three phases.

The aim of this client-server architecture is to bring the image based object recognition service to any person using a common camera-phone and to gain scalability in reference of the number of objects in the database and complexity of the image-recognition algorithms.

IV. IMAGE BASED OBJECT RECOGNITION

Our proposed image based service for object awareness requires both, robust and fast visual object recognition of typically low-quality outdoor images. We therefore applied a methodology that is highly suited for mobile vision applications, i.e., the Informative Features approach [2] on the state-of-the-art local SIFT descriptor [3], which is designed to be rotation and scale invariant and also invariant to illumination changes to some extent. Our resulting i-SIFT approach [4] tackles the standard SIFT bottleneck, i.e., extensive nearest neighbour indexing, by (i) significantly reducing the descriptor dimensionality, (ii) decreasing the size of object representation by one order of magnitude, and (iii) performing matching exclusively on attended descriptors, rejecting the majority of irrelevant descriptors [2]. Experiments on images of the TSG-20 database (download at http://dib.joanneum.at/cape/TSG-20), showed high re-

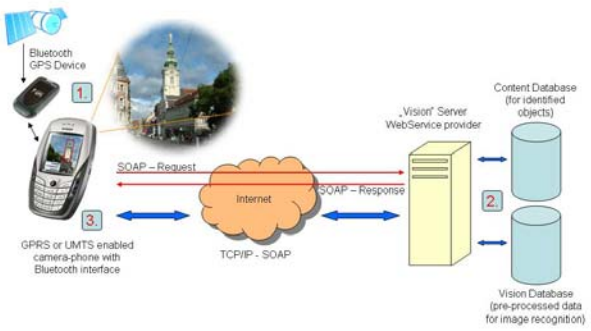

Fig. 2. Overview of the mobile application systems concept.

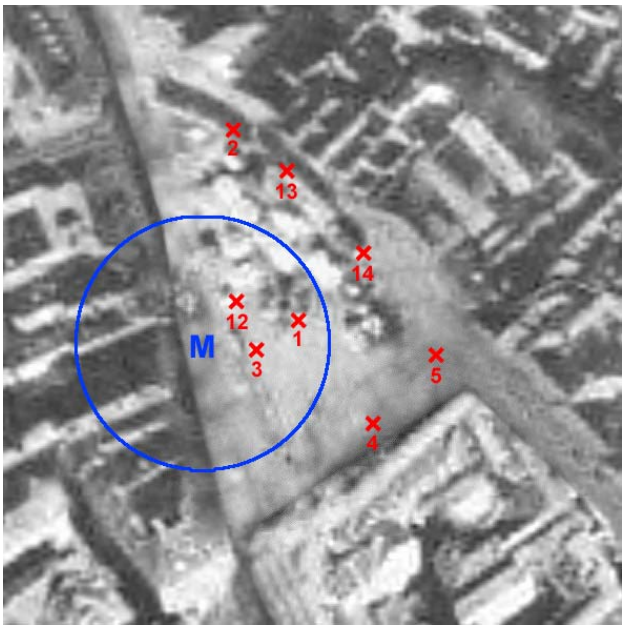

Fig. 3. Geo-Context, e.g., from GPS based position estimates ('M' with blue uncertainty radius), can set priors by geographically indexing into a number of object hypotheses ('X's are coordinates of user positions while capturing images about objects of interest).

cognition accuracy, even on low resolution images (320x240 pixel).

A. The Informative Visual Features Approach

According to the Informative Features approach, the information content of a descriptor with respect to a specific task, i.e. object recognition, is determined from the posterior distribution. In contrast to costly *global* optimization, we expect that it is sufficiently accurate to estimate *local* information content, by computing it from the posterior distribution within a sample test point's local neighbourhood in feature space [5]. Using only the informative features in further processing steps leads to a significant speed-up compared to the normal SIFT based object recognition.

The object recognition task is applied to sample local descriptors f_i in feature space F, $f_i \in \Re^{|F|}$, where o_i denotes an object hypothesis from a given object set Ω. We need to estimate the entropy $H(O \mid f_i)$ of the posteriors

$$P(o_k \mid f_i), k = 1..\Omega \ , \qquad (1)$$

Ω is the number of instantiations of the object class variable O. Shannon conditional entropy denotes

$$H(O \mid f_i) \equiv -\sum_k P(o_k \mid f_i) \log P(o_k \mid f_i) \qquad (2)$$

We approximate the posteriors at f_i using only samples g_i inside a Parzen window of a local neighbourhood ε,

$$\| f_i - f_j \| \leq \varepsilon, j = 1..J \qquad (3)$$

We weight the contribution s of specific samples $f_{j,k}$ - labelled by object o_k - that should increase the posterior estimate $P(o_k \mid f_i)$ by a Gaussian kernel function value $N(\mu, \sigma)$ in order to favour samples with smaller distance to observation f_i, with

$$\mu = f_i \text{ and } \sigma = \varepsilon / 2 \ . \qquad (4)$$

The estimate about the conditional entropy $\hat{H}(O \mid f_i)$ provides then a measure of ambiguity in terms of characterizing the information content with respect to object identification within a single local observation f_i.

We receive sparse instead of extensive object representations, in case we store only selected feature information that is relevant for classification purposes, i.e., discriminative f_i with

$$\hat{H}(O \mid f_i) \leq \Theta \ . \qquad (5)$$

A specific choice on the threshold Θ consequently determines both, storage requirements and recognition accuracy. For efficient memory indexing of nearest neighbour candidates we use the adaptive K-d tree method.

B. Object Recognition Using the Informative SIFT Descriptors

The SIFT based descriptor showed good performance, measured in terms of repeatability, with respect to matching distinctiveness, invariance to blur, image rotation, and illumination changes [6], [7]. As described in [3] the descriptor is calculated in four stages:

1. Determine scale-space extrema as keypoint canditates.
2. Keypoint localization with subpixel accuracy and rejection of unstable keypoints.
3. Orientation assignment, calculated from local keypoint context.
4. Calculation of the orientation histogram in a local keypoint environment relative to keypoint orientation.

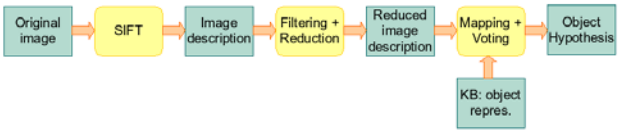

Fig. 3. Object Recognition using i-SIFT descriptors. Standard SIFT descriptors are first extracted within the test image. Then the entropy of the descriptors is determined and decision making is performed only on informative descriptors. Majority voting is then used to integrate local votes into a global classification.

A SIFT descriptor consists of 128 floating point numbers. For performance reasons in nearest neighbour calculation we applied PCA on the SIFT descriptors. Thus, the informative approach is applied as described above, on vectors in the 40 dimensional sub-eigenspace. We applied then the Informative Visual Features approach to the SIFT descriptors (resulting in the i-SIFT approach), selecting only those SIFT responses from the image that provided sufficient information content with respect to the task of object recognition.

In Figure 3 the whole object recognition process is schematised. First, the SIFT descriptors are calculated for a given input image. After applying PCA, the lower dimensional features are fed into an decision tree [8] which has been trained to estimate the local entropy of a feature. Third, the so determined informative features are used for nearest neighbour matching against the image database, whereby the object hypotheses is build only on features with a distance to their nearest neighbour below a given threshold.

C. Experimental Results

The TSG-20 database includes images from 20 objects, i.e. facades of buildings from the city of Graz, Austria. Most of these images contain a tourist sight, together with background information from surrounding buildings, trees, pedestrians, etc. The images contain severe changes in 3D viewpoint, partial occlusion, scale changes by varying distances from exposure, and various illumination changes due to different weather situations and changes in daytime. The images were first subsampled to size 320x240. For each object, we then selected 2 images taken by a viewpoint change of $\approx 30\%$ of a similar distance to the object for training to determine the i-SIFT based object representation. 2 additional frontal views with different distances were taken for testing purpose, given 40 images in total. Figure 4 shows one view of each of the 20 objects contained in the database.

In the experiments on the average only 178 out of 711 (31%) descriptors per object were retained for object representation. Also the descriptor dimensionality was reduced from 128 to 40 dimensions. The threshold on the entropy criterion for attentive matching was defined by Θ = 1.8, which leads to only 40 % of nearest neighbor processing's. The recognition accuracy according to MAP

Fig. 4. The TSG-20 database, consisting of images from 20 tourist relevant buildings in the city of Graz.

(Maximum A Posteriory) classifications was 100%, the average entropy in the posterior distribution was $\hat{H}(O \mid f_i) \approx 1.4$. Figure 5 shows two sample building recognitions, whereas figure 6 shows results for background images (buildings not in training database).

V. CONCLUSIONS AND OUTLOOK

Location based services will be a default service for many themes in the future. Location awareness, concerning the presented theme, is a key issue for such mobile applications. Shortcomings in the provision of position information ("everywhere and anytime") related to the mobile user may be a crucial requirement for the acceptance of mobile services in urban areas.

The integration of contexts arising from the spatial information at a current location will support the area-wide location awareness in combination with GPS functionality or the cell information of the phone-network provider. This allows an individual and also spontaneous information service. Such a mobile service will offer new ways of visualising spatial information and a customized data access for a mobile user in an urban environment.

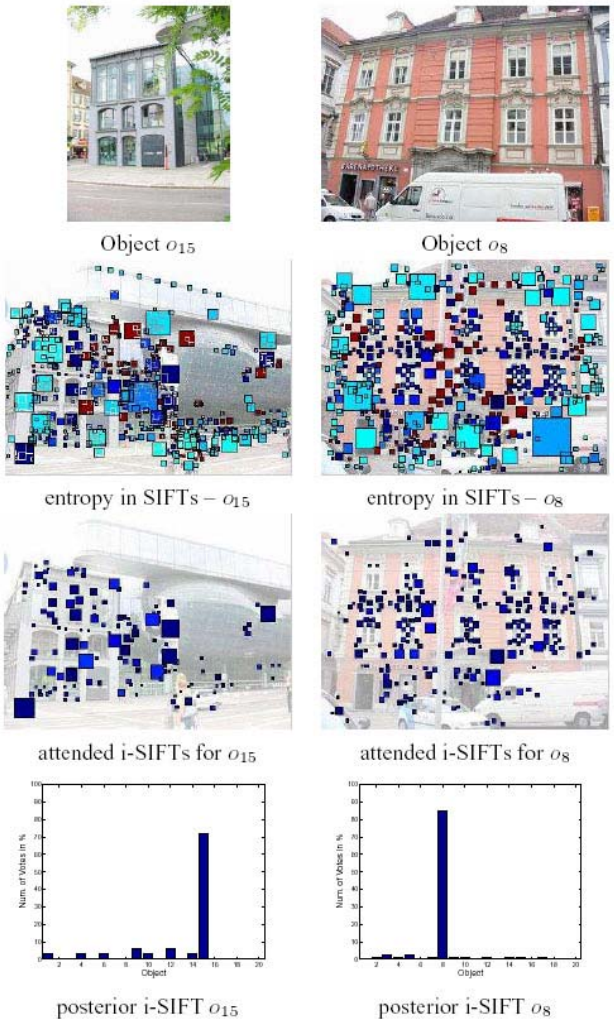

Fig. 5 Sample building recognition for objects o_{15} and o_{18}, (top down) training images, entropy coded (blue: low, red: high) SIFT descriptors (without selection), entropy coded informatitive SIFT descriptors, and corresponding posterior distribution for i-SIFT based descriptor recognition.

We provide an overview in this paper about a mobile system, which includes GPS functionality and vision enhanced context awareness to offer user friendly digital services for pedestrians in urban areas. The client-server architecture of the system allows bringing the image based object recognition service to any person using a common camera-phone. This mobile system enables its ubiquitous use in many business and commerce relevant situations.

Further work will focus on the integration of different sensors for the realization of smart mobile vision services in urban environments. These developments will be realized in frame of both the EU-project *MOBVIS* (Vision Technologies and Intelligent Maps for Mobile Attentive Interfaces in Urban Scenarios) and the national, ASA funded project *Mobile City Explorer*.

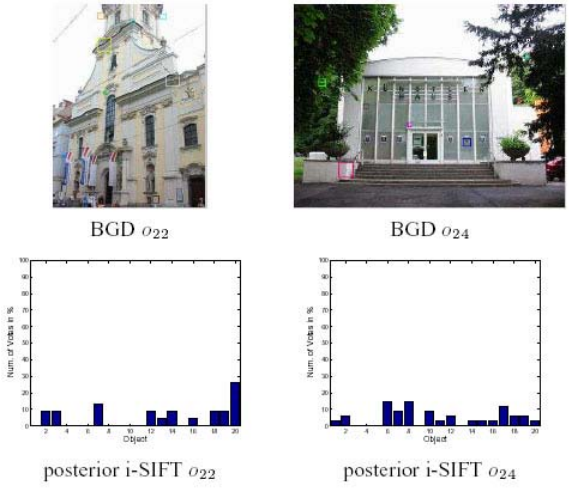

Fig. 6 Detection of background from high entropy in the posteriors.

VI. ACKNOWLEDGMENTS

This work is funded by the European Commission's project MACS under grant number FP6-004381, and by the Austrian Joint Research Project Cognitive Vision under sub-projects S9103-N04 and S9104-N04.

VII. REFERENCES

[1] Baldzer J., Boll S., Krösche J., Rump N., Scheibner H. and Thieme S.: Location-Based Geodata Broadcast. 1st Workshop on Positioning, Navigation and Communication 2004 – WPNC'04; March, 26th 2004. Shaker Verlag Aachen (2004); pp 67-73.

[2] Fritz, G., Seifert, C., Paletta, L., and Bischof, H., Rapid Object Recognition from Discriminative Regions of Interest, Proc. 19th National Conference on Artificial Intelligence, AAAI 2004, San Jose, CA, July 25-29, 2004, pp. 444-449.

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

[4] Paletta, L., Fritz, G., Seifert, C. Informative SIFT Descriptors for Object Detection, submitted to CVPR 2005, San Diego, CA

[5] Fritz, G., Paletta, L., and Bischof, H., Object Recognition using Local Information Content, Proc. International Conference on Pattern Recognition, ICPR 2004, Cambridge, UK, August 22-26 2004, Vol. II, pp. 15-18.

[6] Mikolajczyk K, and Schmid, C. A performance evaluation of local descriptors. Proc. Computer Vision and Pattern Recognition, CVPR, Madison, WI, 2003.

[7] Mikolajczyk K, and Schmid, C. A performance evaluation of local descriptors. submitted to PAMI, http://lear.inrialpes.fr/pubs/2004/MS04a, 2004.

[8] Quinlan, J. R. C4.5 Programs for Machine Learning. Morgan Kaufmann, San Mateo, CA, 1993.