In: *Proc. International Conference on Mobile Learning*, MLEARN 2004
Lake Bracciano, Rome, 5-6 July 2004, *in print*.

Mobile Vision for Ambient Learning in Urban Environments

Gerald Fritz, Christin Seifert, Patrick Luley, Lucas Paletta, and Alexander Almer
JOANNEUM RESEARCH, Institute of Digital Image Processing
Wastiangasse 6, A-8010 Graz, Austria
Email: lucas.paletta@joanneum.at

Abstract

We describe a mobile vision system that is capable of automated object identification using images captured from a PDA or a camera phone. We present a solution for the enabling 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 GPS, W-LAN and a camera attached to a PDA or a 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, the architecture, or other related cultural context of historic or artistic relevance 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.

Keywords: Mobile vision, object recognition, location based services, learning in urban environments.

1. Motivation

Mobile learning 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 work 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 (Figure 1(a)). 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 (PDA, or camera phone) pointing towards the object of interest (e.g., a building or a statue) will capture images on demand to automatically find objects in the tourist user's view. The images are then transmitted to a server that automatically extracts the object information, associates it to m-learning content, and sends the resulting data back to the mobile user (Figure 1(b)). '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.

State-of-the art pattern recognition in mobile computer vision has advanced from indoors processing (see Aoki et al. (1999)) and time consuming outdoors recognition (see Coors et al. (2000)) to nomadic sign identification (Yang et al. (2001)) and attentive processing for robust recognition performance (Fritz et al. (2004)). Layered mobile location based services (see Hightower et al. (2002)) support in addition more accurate estimates about the immediate environment of the user, e.g., for the purpose of mobile tourist information services (see Almer & Luley (2003)). The goal is to annotate detected objects with characteristic data and to provide an interface to access even more detailed multimedia information about the selected object. Entries of a corresponding database on multimedia information of various kind (images, video, sound, texts) are interlinked to enable cross-references and access to different information spaces (Figure 1(c)-(f)).

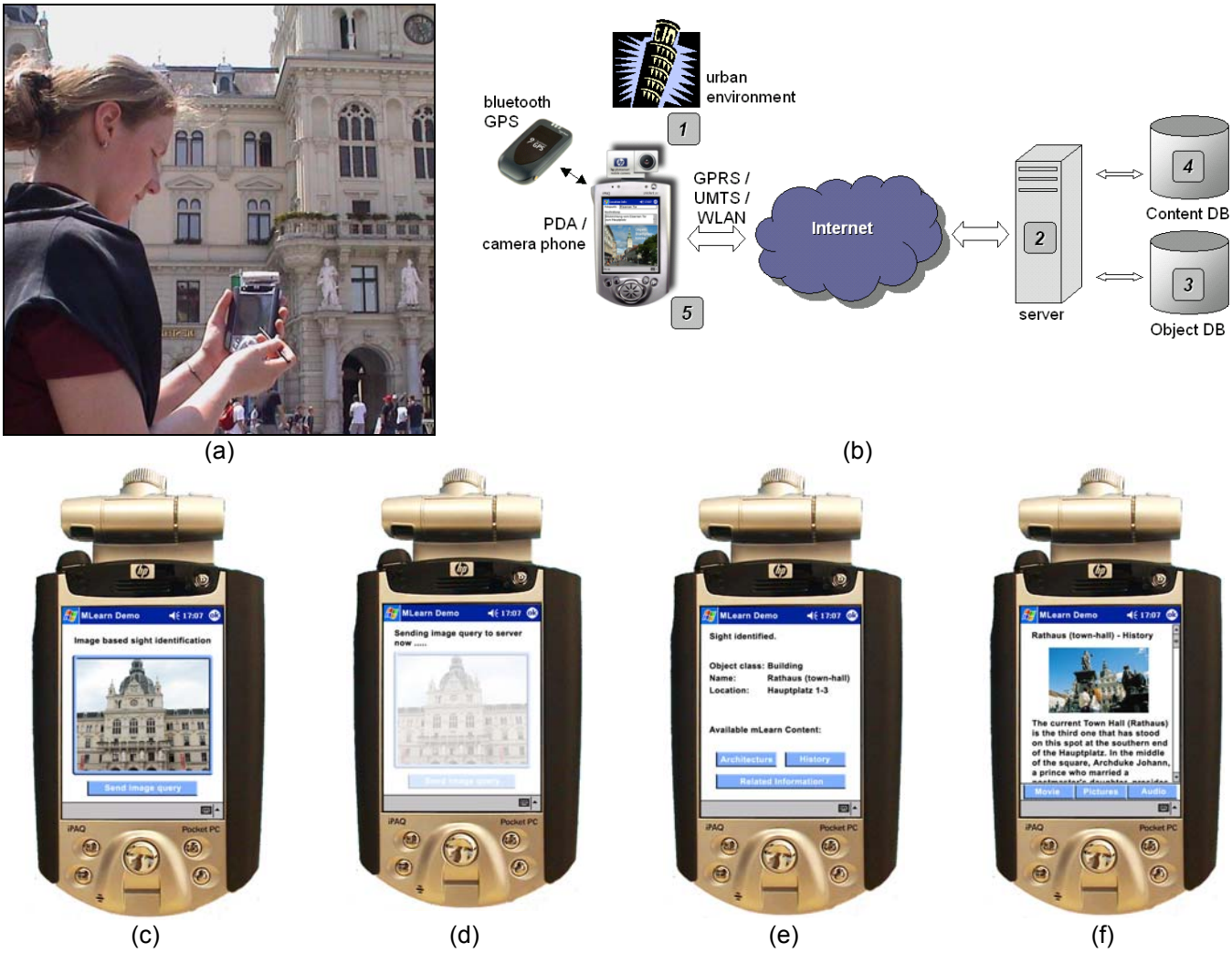

Figure 1. *(a) The nomadic tourist explores the city in interaction with its vision enhanced ambient learning system. (b) The image capture about the sight is transmitted via internet to the server that performs image processing and automated object recognition. (c)-(f) Screen shots from the mobile learning system: (c) the user initiates sight identification via image query, (d) the image is transmitted to the server, (e) the resulting information is displayed at the client site, (f) contextual information is provided together with a menu that links to additional information sources (movies, pictures, audio, etc.)*

2. Mobile Learning using Visual Object Detection

The user equipment of the ambient learning system (Figure 1(b)) consists of a camera based PDA or a camera phone, and a GPS receiver connected via blue-tooth. The client device should provide an integrated or connected camera with a resolution of at least 640x480 pixels as supported by the industrial standard today. The mobile device is linked via wireless connection to a server. We assume that at least a cell phone network can be used to establish an internet connection; alternatively, different data transfer technologies like GPRS, UMTS or WLAN could be used as well, depending on the choice of local network providers. Object identification is performed at the server site and hidden to the user who just receives the result (Figure 1(e)) in about 2-3 seconds. The server receives a raw GPS based position estimate and concludes a selection of relevant sights who potentially might appear in the tourist's field of view. The user initiates an image capture to start the visual object recognition (Figure 2). The system first extracts a grey level pixel pattern (Figure 2(a)) and identifies

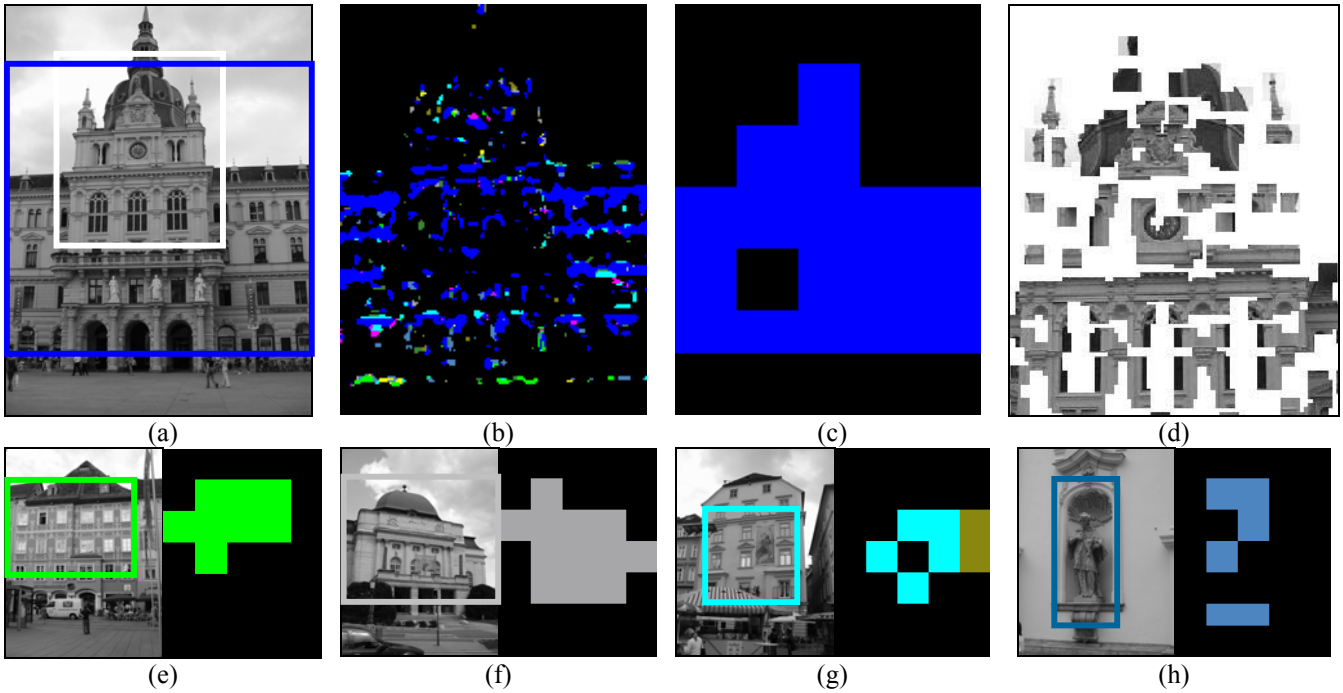

Figure 2. *(a) The mobile vision system operates on grey level pixel patterns, e.g., about the town hall of Graz. (b) It first identifies discriminative local patterns (Fritz et al. (2004)) in the complete image ((d) depicts the corresponding local patterns from the region in (a) circumscribed by the white rectangle). Blue pixels were classified as voting for the 'town hall object'. (c) describes a tiled image partition, being colour coded from a more abstract voting process. (e)-(h) Colour coding of correct object classifications for several sights and a statue (h) found at tourism relevant sites of Graz.*

image regions that are highly discriminative with respect to object identification (Figure 2(b)). Larger regions of local object votes (Figure 2(c)) are integrated in a second processing stage resulting in a single object vote (e.g., blue rectangle in Figure 2(a)). The demonstrator system contains 10 local objects of interest from the city centre of Graz, Austria (buildings, statues; see sample results in Figure 2(e)-(h)) but is currently extended to up to 50 tourist relevant objects. Corresponding to the object result, the PDA will display associated m-learn information from the content database (Figure 1(b),(f)). The multimedia information will describe the cultural context with respect to the object in the field of view and enables the user to learn within the urban environment. The presented experimental results seem to be promising and are extended to evaluate larger object sets.

3. References

Almer A, Luley P (2003). Location Based Tourism Information on Mobile Systems. *Proc. European Navigation Conference,* GNSS 2003, Graz, Austria.

Aoki H, Schiele B, Pentland A (1999). Real-time Personal Positioning System for Wearable Computers. Proc. *IEEE International Symposium on Wearable Computing,* San Francisco, CA.

Coors V, Huch T, Kretschmer U (2000). Matching buildings: Pose Estimation in an Urban Environment. *Proc. IEEE and ACM International Symposium on Augmented Reality*, Munich, Germany, pp. 89-92.

Fritz G, Seifert C, Paletta L, Bischof H (2004). Rapid Object Recognition from Discriminative Regions of Interest. *Proc.* National Conference on Artificial Intelligence, AAAI 2004, San Jose, CA, *in print*.

Hightower J, Brumitt B, Borriello G (2002). The Location Stack: A Layered Model for Location in Ubiquitous Computing. *Proc. IEEE Workshop on Mobile Computing Systems & Applications*, Callicoon, NY, pp. 22-28.

Yang J, Gao J, Zhang Y, Chen X (2001). An Automatic Sign Recognition and Translation System. *Proc. Workshop on Perceptive User Interfaces*, PUI 2001, Lake Buena Vista, FL.