Mobile vision for ambient learning in urban environments

Gerald Fritz
Christin Seifert
Patrick Luley
Lucas Paletta
Alexander Almer

lucas.paletta@joanneum.at
Institute of Digital Image Processing
Joanneum Research
Wastiangasse 6
A–8010 Graz
Austria

Abstract

We describe a mobile vision system that is capable of automated object identification using images captured from a personal digital assistant (PDA) or a camera phone. We present a solution for a technology that will enable outdoors vision-based object recognition, which will extend state-of-the-art location and context aware services towards the achievement of object-based awareness in urban environments. In the proposed application scenario, tourist pedestrians are equipped with a Global Positioning System (GPS), a wireless local area network (WLAN) and a camera attached to a PDA or a camera phone. They are asked to look and see whether their field of vision contains tourist sights that would prompt them to seek more detailed information. A mobile user who is intending to learn within the urban environment might want to explore multimedia-type data about the history, architecture or other related cultural context of historic or artistic relevance. Accordingly, ambient learning is achieved by pointing the device towards the sight, capturing an image and consequently getting on-site information about the object within the focus of attention – that is, the user's current field of vision.

Keywords
mobile vision, object recognition, location-based services, learning in urban environments

1 Motivation for the project

Mobile learning systems operating in urban environments must take advantage of contexts arising from the spatial and situated information at the pedestrian user's current location. Location-based services today are, in principle, able to provide access to rich sources of information and knowledge to the nomadic user. However, the kind of location awareness that they do provide is not intuitive, requiring reference to maps and addresses – in other words, the information is not directly mediated via the object of interest.

In contrast, the work presented here takes a decisive step towards getting in line with the user's current intention to relate information to his or her current sensorial experience – the object in his/her line of sight (**Figure 1a**). In this way, the system can respond to the user's focus of attention; for example, for the purpose of tourist information systems. A camera attached to the mobile system (PDA or camera phone) pointing towards the object of interest (eg a building or a statue) will capture images on demand, 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 m-learning content and sends the resulting data back to the mobile user (**Figure 1b**). 'Mobile vision' here refers to mobile visual data that is 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 (Aoki, Schiele and Pentland 1999) and time-consuming outdoors recognition (Coors, Huch and Kretschmer 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 (Hightower, Brumitt and Borriello 2002) support, in addition, more accurate estimates about the immediate environment of the user; for example, to provide mobile tourist information services (Almer and 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.

Figure 1 The nomadic tourist using a vision-enhanced ambient learning system

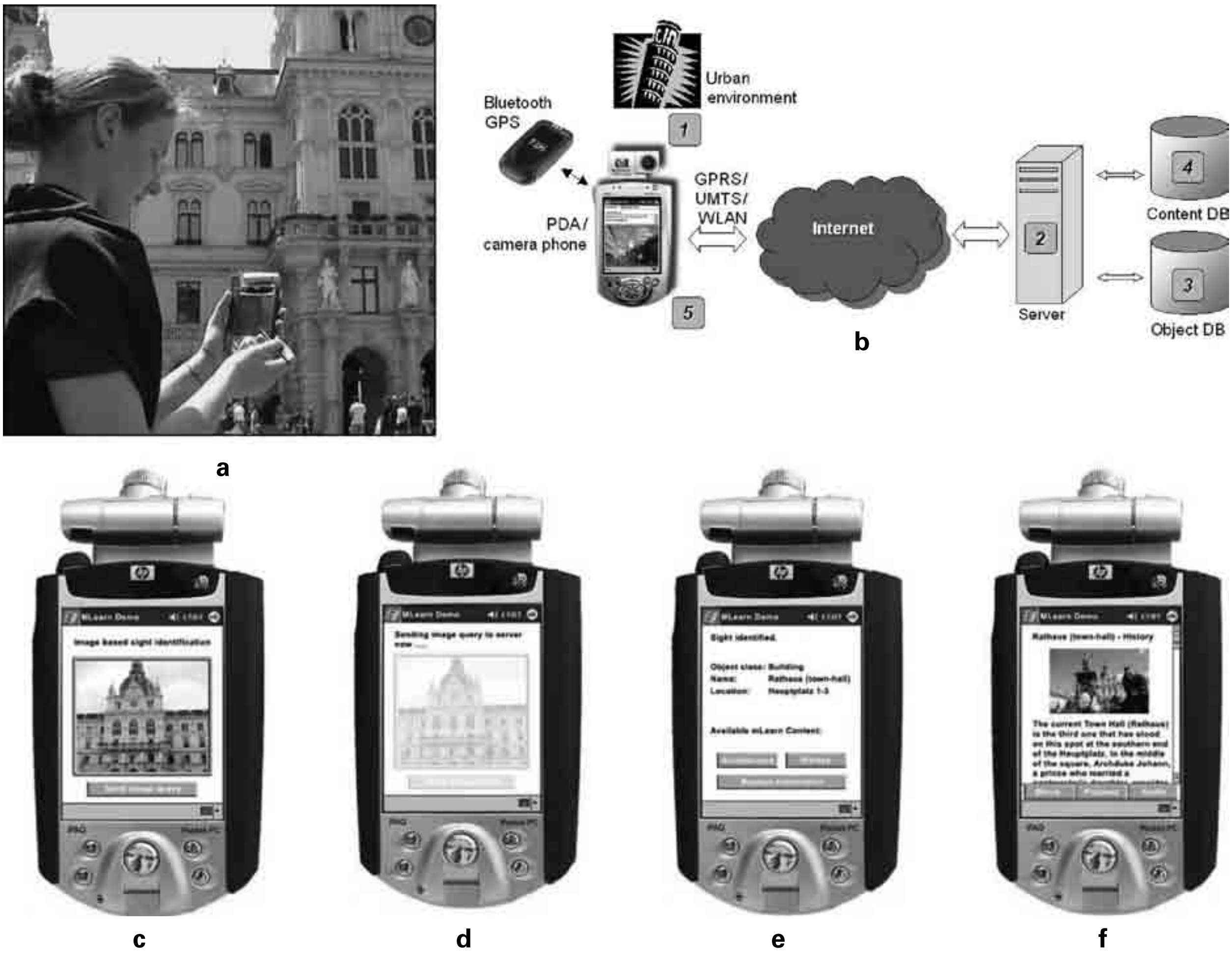


Image based sight identification

Send image query

Sending image query to server now

Sight identified.

Object class: Building
Name: Rathaus (town-hall)
Location: Hauptplatz 1-3

Available mLearn Content:

Rathaus (town-hall) - History

The current Town Hall (Rathaus) is the third one that has stood on this spot at the southern end of the Hauptplatz, in the middle of the square. Archduke Johann, a prince who married a

1a The nomadic tourist explores the city in interaction with his/her vision-enhanced ambient learning system.

1b The image capture relating to the tourist sight is transmitted via the internet to the server that performs image processing and automated object recognition.

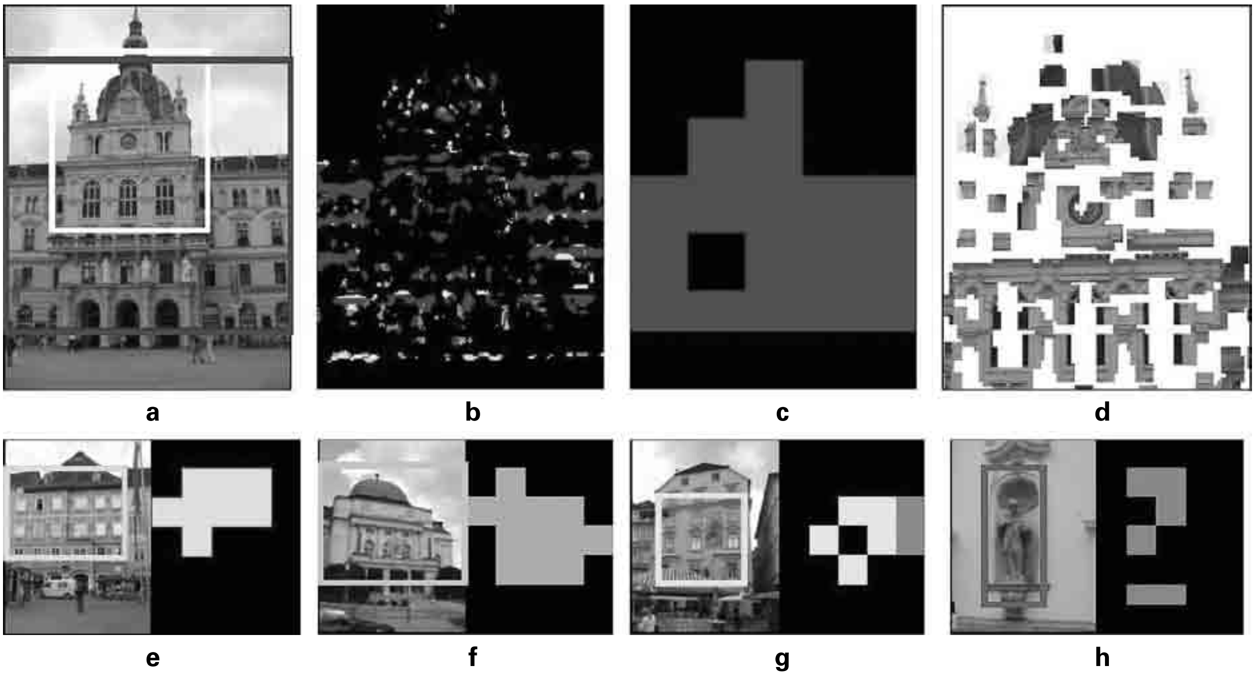
1c–f Screenshots from the mobile learning system: (**c**) the user initiates sight identification via an image query; (**d**) the image is transmitted to the server; (**e**) the resulting information is displayed at the client site; (**f**) contextual information is provided together with a menu that links to additional information sources (movies, pictures, audio, etc).

Entries within a corresponding database of multimedia information of various kinds (images, video, sound, texts) are interlinked to enable cross-referencing and access to different information spaces (**Figure 1c–f**).

2 Mobile learning using visual object detection

The user equipment of the ambient learning system (**Figure 1b**) consists of a camera-based PDA or camera phone and a GPS receiver connected via Bluetooth. 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 the very least, a mobile phone network can be used to establish an internet connection;

Figure 2 Operation of the mobile vision system



2a The mobile vision system operates on grey-level pixel patterns; for example, with reference to the Town Hall in Graz.

2b It first identifies distinctive 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'.

2c describes a tiled image partition, being colour coded from a more abstract voting process.

2e–h Colour coding of correct object classifications for several sights and a statue (**h**) found among the sights of interest to tourists in Graz.

alternatively, different data transfer technologies like General Packet Radio Service (GPRS), Universal Mobile Telecommunications System (UMTS) or wireless local area network (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 1e**) in about 2–3 seconds. The server receives a raw GPS-based position estimate and collects a selection of relevant sights which potentially might appear in the tourist's field of vision. The user initiates an image capture to start the visual object recognition (**Figure 2**). The system first extracts a grey-level pixel pattern (**Figure 2a**) and identifies image regions that are highly discriminating with respect to object identification (**Figure 2b**). Larger regions of local object votes (**Figure 2c**) are integrated in a second processing stage, resulting in a single object vote (eg the blue rectangle in **Figure 2a**).

The demonstration system contains 10 local objects of interest from the city centre of Graz in Austria (buildings, statues, etc; see sample results in **Figure 2e–h**), but is currently being extended to display up to 50 objects of interest to tourists. The PDA will then display associated m-learning information from the content database corresponding to the object-search result (**Figure 1b and f**). The multimedia information will describe the cultural context relating to the object in the field of vision, enabling the user to learn within the urban environment. The presented experimental results seem to be promising and are being extended to evaluate larger object sets.

References

Almer A and Luley P (2003). Location-based tourism information on mobile systems. In *Proceedings of the European Navigation Conference, Graz, Austria, 22–25 April*.

Aoki H, Schiele B and Pentland A (1999). Real-time personal positioning system for wearable computers. In *Proceedings of the IEEE International Symposium on Wearable Computing, San Francisco, California*. Los Alamitos, CA: IEEE Computer Society, 37.

Coors V, Huch T and Kretschmer U (2000). Matching buildings: pose estimation in an urban environment. In *Proceedings of the IEEE and ACM International Symposium on Augmented Reality, Munich, 5–6 October*, 89–92.

Fritz G, Seifert C, Paletta L and Bischof H (2004). Rapid object recognition from discriminative regions of interest. In *Proceedings of the National Conference on Artificial Intelligence, San Jose, California*.

Hightower J, Brumitt B and Borriello G (2002). The location stack: a layered model for location in ubiquitous computing. In *Proceedings of the IEEE Workshop on Mobile Computing Systems and Applications, Callicoon, New York*. Los Alamitos, CA: IEEE Computer Society, 22–28.

Yang J, Gao J, Zhang Y and Chen X (2001). An automatic sign recognition and translation system. In *Proceedings of the Workshop on Perceptive User Interfaces, Lake Buena Vista, Florida, 15–16 November*.