

M3I in a Pedestrian Navigation & Exploration System

Rainer Wasinger, Christoph Stahl, and Antonio Krüger

DFKI GmbH,
66123 Saarbrücken, Germany
{rainer.wasinger, christoph.stahl, antonio.krueger}@dfki.de

Abstract. In this paper, we describe a near-complete Pocket PC implementation of a Mobile Multi-Modal Interaction (M3I) platform for pedestrian navigation. The platform is designed to easily support indoor and outdoor navigation tasks, and uses the combination of several modalities for presentation output and user input. Whereas 2D/3D-graphics and synthesized speech are used to present useful information on routes and places, fused input from embedded speech and gesture recognition engines allow for situated user interaction.

1 Introduction

Mobile navigation services are becoming one of the more successful applications of 3G mobile communication. However, the use of a navigation and exploration system on the go, poses special requirements on the design of appropriate presentation and interaction techniques. This paper presents our M3I-platform, which incorporates different presentation and interaction techniques to provide flexible and robust levels of user interaction in an indoor and outdoor environment. Our M3I-platform is unique in that it combines mobile 2D/3D graphics with synthesized speech generation, and the fusion of both speech and gesture input through the use of multiple recognizers. All of this functionality is embedded on a predominantly off-the-shelf mobile device.

Different approaches to mobile navigation and exploration services have been investigated in the past. Whereas first versions ran on laptops [1,2] with limited interaction possibilities, more recent approaches use PDAs [3] that are equipped with touchscreens and therefore allow for at least simple stylus gestures. Some of the newer approaches combine 2D and 3D representations of the environment [4]. A strong focus on distributed multi-modal interaction is demonstrated in [5]. Whereas all these systems focus on single aspects of mobile multi-modal user interactions, the M3I-platform presented here aims at combining all of these features on one mobile device to support users in navigating and exploring the real world.

2 Design Requirements

The navigation system allows a user to download predefined routes onto a PDA, and then select a route for *indoor* and *outdoor* pedestrian navigation and exploration. The *navigation* mode directs a user from start to destination through combined speech and graphics output, as shown in Fig.1. The *exploration* mode relaxes the direction information presented through speech and graphics, and instead allows the user to freely roam or explore a place. The navigation mode is best suited when a user is either under a high *cognitive load* (e.g. business people), or simply uninterested with their immediate surroundings, while the exploration mode is best suited to people with more time (e.g. travelers). For natural and flexible use, such a system must also be *multi-modal*. Route descriptions may be presented via 2D/3D graphic visualizations and an audio headset, and user input can take the form of combined speech and stylus requests, incorporating objects on the PDA's display, or objects in the real world around them (see Fig. 2). Possible user inputs are for example “what is that [gesture]?”, or “describe this [gesture] church”. The language and objects that the user refers to on the map, such as parks, churches, museums, and other buildings, must also be known to the system. This requires the incorporation of *dynamic speech grammars* and *graphics* that can adapt to a changing environment. Other features of the system include the ability to zoom and to rotate the map (reduces the user's cognitive load in associating the map with the environment [6]), a birdseye and egocentric view (egocentric shown in Fig.1 and 2), and a feature to record memos to geographical locations on the map.

3 The M3I Navigation Platform

The pedestrian navigation system comprises a navigation server and a Pocket PC. The Pocket PC component developed in C/C++, incorporates the IBM Embedded Via-Voice formant-based speech synthesizer and dynamic rule grammar based recognizer. The 2D- and 3D-graphics are generated via the embedded Cortona VRML¹-browser. A magnetic compass provides the user's (i.e. the PDA's) current facing direction used in determining gestures. GPS provides further sensor information such as velocity and direction, and is also required to locate the user when outside. Infrared beacons are used to locate a user when inside. PDA communication with the server is via a standard HTTP connection, for example through a bluetooth capable GPRS/UMTS mobile phone, WLAN, or a USB desktop connection.

3.1 GIS Server

The only part of this system that is not embedded on the Pocket PC is a custom made GIS server, based on the open source GRASS² project. It generates the graphic and

¹ Virtual Reality Markup Language

² More information on the grass project can be found at <http://grass.baylor.edu>

text-based route descriptions for a particular trip, and contains street names and very limited landmark information. This is not currently being processed on the PDA due to performance issues. Outdoor navigation is based on commercially available material from NavTech, but data on indoor floor plans and detailed landmark information (e.g. opening hours, cost and description) have to be modeled by hand. All information that is necessary for the presentation and interaction is collected into a set of XML files and passed to the embedded components of the M3I-platform.

3.2 Speech

The speech synthesizer receives navigational data such as street names, distance, turning angle and landmark information from the XML files, from which it then generates appropriate route descriptions. Long, middle and short phrases are created for each segment and presented to the user in combination with graphics.

The recognizer currently implements the use of static rule grammars that cover command-and-control functionality (e.g. zoom in), trip queries (e.g. where is my start?), and simple multi-modal interaction (e.g. what is *that* [gesture]). It is now intended to extend this functionality by incorporating dynamic rule grammars that can more closely model objects on the map such as landmarks and streets. These objects are generally difficult to manage because they change when a new map is loaded. Aside from the base functionality provided by our static pre-compiled grammars, three sets of dynamic grammars (see Fig. 1) provide extended interaction for the user.

The first dynamic grammar will allow for interaction with landmark types, while the second will allow for interaction with specific street and landmark names. These grammars will be created each time a map is loaded, but due to the large amount of specific street and landmark information, the latter will only be activated upon request. A third grammar analogous to information found in tourist pamphlets (e.g. description, opening hours, cost) will allow for detailed interaction with the individual landmarks. This interaction is landmark dependent as the different landmarks including parks, churches and museums all display significantly different characteristics. The above strategy, in combination with our relatively small map sizes, and acoustic models designed for hand-held devices will allow for robust rates of user-recognition.

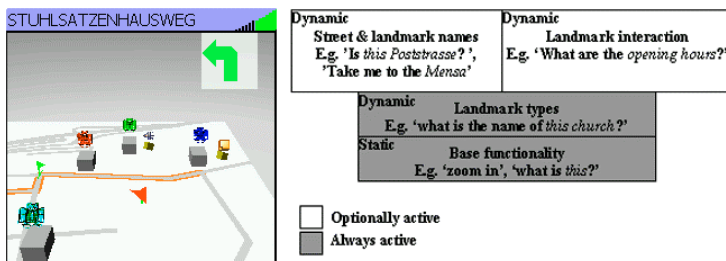


Fig.1. Static and dynamic speech grammars, and a PDA screen-shot of a route containing landmark interaction objects

3.3 Gesture and Sensor Fusion

The fusion of speech with gesture increases the flexibility and naturalness of a user's interaction. It also supports a more robust level of interaction in that non-conflicting but overlapped verbal and gesture segments (e.g. describe this [gesture] church!) can reduce the recognition search space. This also gives rise to the need to unify results, as described in [7]. We use a blackboard architecture to capture the results of gesture, speech and sensor input, and a media fusion module to combine these inputs based on timestamps. Gesture input can be either *intra*, in that the user points to an object on the display through the use of a stylus, or it can be *extra* in that the PDA is used as a pointing stick to point at an object in the real world. Intra gestures are currently limited to 'point-like' gestures that can be used to query landmarks, but we plan to extend this to include 'line-like' gestures that will enable the querying of street names. Intra and extra gestures are both illustrated in Fig. 2.

Extra gestures are currently detected with the help of a magnetic compass/GPS CF-Card from Pointstar that is inserted into the PDA. Based on the movement and the position of the PDA, the system is able to answer requests on the landmarks that a user is currently looking at. These sensors can also provide information on the user's velocity (e.g. stopped, walking, running), and whether the device is being looked at by the user. This allows for better adaptation of the presentation, for example by placing higher importance on either the graphical or speech modality. Sensor fusion is also used to improve localization by combining information from infrared beacons, GPS and the magnetic compass. A further improvement is achieved through map-matching, i.e. snapping the user to the closest point on a route segment of their current path.

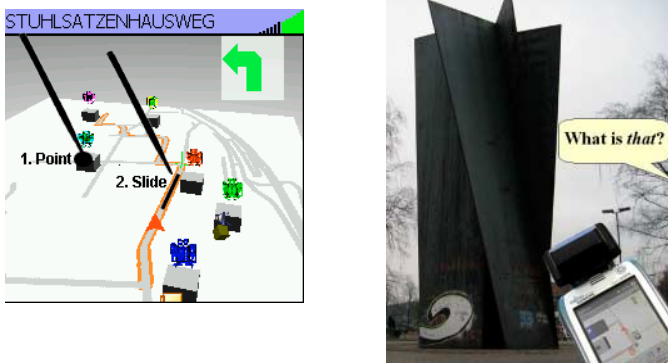


Fig. 2. Intra and extra gestures respectively

3.4 3D Graphics

The graphics architecture of the M3I-platform is based on the Cortona VRML 3D component, which consists of three layers. The bottom layer deals with the loading and rendering of the 3D-VRML scene. All visible objects are represented as nodes in a scene graph, and interaction through gesture is provided by adding touch-sensors to geometry nodes. The middle layer provides a generic C++ class that implements the functionality to model interactive 3D objects as a complex collection of basic VRML

nodes, and provides simple methods for creation and localization. On the upper layer, the application derives classes to represent landmarks, users and media-icons from the generic 3D object class. These classes store all symbolic information about the object such as name and description, and also define the individual 3D geometry in VRML code. The consistent object-oriented representation of all localized objects in the application, allows for the handling of both data and graphics by a single object reference. This simplifies the unification of object references in the media fusion module, based on timestamps.

4 Conclusions

In this paper we have described the design issues arising in the implementation of our M3I-platform for pedestrian navigation and exploration. The platform combines 2D/3D-graphics and synthesized route descriptions, with combined speech and gesture recognition. It is unique in embedding all of this functionality on a mobile device. The system is near completion and has been demonstrated in the German cities of Saarbrücken and Munich for both indoor and outdoor use. Future work will take the form of user studies, and we plan to make parts of the M3I platform publicly available on completion.

References

- [1] K. Cheverst, N. Davies, K. Mitchell, A. Friday, and C. Efstratiou. Developing a context-aware electronic tourist guide: Some issues and experiences. ,CHI 2000, Amsterdam, 2000.
- [2] J. Baus, A. Krüger, and W. Wahlster. A Resource-Adaptive Mobile Navigation System. In: IUI2002: International Conference on Intelligent User, ACM Press, 2002.
- [3] G. Popischil, M. Umlauf, E. Michlmayr: Designing Lol@, a mobile tourist navigation guide, In: Proceedings of HCI with mobile devices 02, Springer LCNS 2411, 2002.
- [4] C. Kray, K. Laakso, C. Elting, V. Coors: Presenting Route Instructions on mobile devices, In: Proceedings of IUI 03: Intelligent User Interfaces, ACM Press, 2003.
- [5] Cohen, P., Johnston, M., McGee, M., Oviatt, S., Pittman, J., Smith, I., Chen, L., Clow, J.: Quickset: Multimodal interaction for distributed applications, In: Proceedings of the Fifth ACM International Multimedia Conference, 1997, pp 31-40.
- [6] 2. Hunt, E., Walter, D.: Orientation and wayfinding: A Review (Tech. Rep. Nr.N00014-96-0380, Arlington: Office of Naval Research, 1999.
- [7] Alexandersson, J., Becker, T.: Overlay as the Basic Operation for Discourse Processing in a Multimodal Dialogue System, Proceedings of the IJCAI Workshop Knowledge and Reasoning in Practical Dialogue Systems, 2001.