The Last 10 Metres: Using Visual Analysis and Verbal Communication in Guiding Visually Impaired Smartphone Users to Entrances

Anja Belz
Computing, Engineering and Maths
University of Brighton
Lewes Road, Brighton BN2 4GJ, UK
a.s.belz@brighton.ac.uk

Anil Bharath
Department of Bioengineering
Imperial College London
Prince Consort Road, London SW7 2BP, UK
a.bharath@imperial.ac.uk

1 Introduction

Blindness and partial sight are increasing, due to changing demographics and greater incidence of diseases such as diabetes, at vast financial and human cost (WHO, 2013). Organisations for the visually impaired stress the importance of independent living, of which safe and independent travel is an integral part. While existing smartphone facilities such as Apple's Siri are encouraging, the supporting localisation services are not sufficiently accurate or precise to enable navigation between e.g. a bus stop or taxi rank and the entrance to a public space such as a hospital, supermarket or train station.

In this paper, we report plans and progress to date of research addressing 'the problem of the Last 10 Metres.' We are developing methods for safely guiding users not just to the general vicinity of a target destination (as done by GPS-based services), but right up to the main entrance of the target destination, by a combination of semantically and visually enriched maps, visual analysis, and language generation.

2 Overview

The core task is to help users navigate approach paths to building entrances. Navigation guidance is delivered via a smartphone app with voice and haptic output. The app uses detailed, semantically tagged maps in which public buildings (museums, schools, hospitals, etc.) and the pavements, landmarks and other visual cues found in the approaches to their entrances (See Figure 2) are annotated. The maps differ from existing resources in that they have (i) more detailed information on pedestrian-relevant features, including obstructions and hazards, and (ii) computational descriptions of 'visual paths,' i.e. information about approach paths to entrances including image sequencess taken along the path (visual cues).

The navigation app provides guidance from the point where a GPS-based system drops the user: theoretically within 10m of a destination building, but in reality, anything up to a few hundred metres away from the actual building entrance. Our research is focused on developing a novel pedestrian guidance system that uses semantically and visually enriched maps, visual cues from user-generated live-feed video, and verbal and haptic communication to guide visually impaired pedestrians during the last few metres to the entrance of their destination, dropping them not just somewhere near, say, the British Museum, but more precisely and much more challengingly, right in front of the museum's main entrance.

3 Usage Scenario

The user employs their usual GPS-based app to get near a target destination, then our Last 10m app takes over: (1) User requests guidance to an entrance to their target building; (2) System retrieves relevant local map from server; (3) System converts guidance request to a specific target entrance T annotated on map; (4) Given location of T on map, system determines location U of user on map; (5) System computes approach path P from U to T; (6) System starts guiding user along P; at the same time system carries out continuous monitoring of user behaviour and surroundings, interacting with user as necessary: (a) System monitors that user stays on track; (b) System monitors path ahead to identify any obstacles; (c) System issues warnings and update information as necessary, and deals with user requests, e.g. information about an object detected by the user, location updates or output modality changes.

Proceedings of the 25th International Conference on Computational Linguistics, pages 118–120,
Dublin, Ireland, August 23-29 2014.

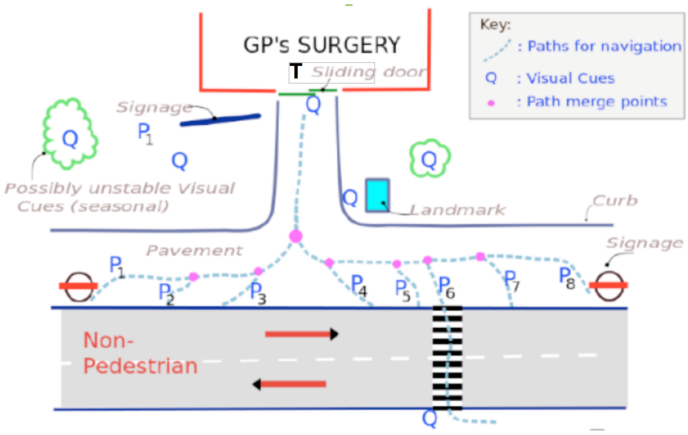

Figure 1: Illustration of the navigational context that we are addressing.

4 Key Challenges

4.1 Mapping Challenges

Semantically enriched local maps: Using OpenStreetMap,[1] which already includes many different kinds of relevant 'urban' tags such as 'tree', 'bus_stop', 'post_box', 'traffic_signals', etc., as a starting point, we are investigating ways of involving some of the 1.5 million volunteer mappers to create a new OSM layer of highly fine-grained local information and snapshots of visual cues.

Computing path from U to T: Adapting methods developed for similar purposes (Zeng et al., 2008), compute geometric paths from U to T; if necessary recompute these paths on the fly on the basis of obstacles that have been detected (see below).

4.2 Vision Challenges

Locating user on map based on visual cues: The task is to locate the user precisely on the map (within a given radius determined on the basis of GPS output) by identifying landmarks and visual cues in user-generated live feed and matching these to the tags and images in the semantically enriched local maps. In a pilot study (Rivera-Rubio et al., 2013), conducted within indoor, but highly ambiguous corridors, we have found that with relatively modest processes, paths can be distinguished with reasonable certainty using visual cues alone. In more extensive tests, verified with surveying equipment (Rivera-Rubio et al., 2014), we found that user location on a path can be inferred from hand-held and wearable cameras.

Continuous route monitoring: (a) monitoring of path ahead to identify obstacles and other danger using computer vision techniques and map information, (b) monitoring actual path against target path, updating target path and adapting instructions to user as necessary. Monitoring is based on local maps, visual information obtained on the fly (Davison et al., 2007; Alcantarilla et al., 2010; Pradeep and Medioni, 2010) from smartphone camera live feeds, as well as information from inertial sensors, etc.

4.3 Communication Challenges

While 'smart canes' are promising technological improvements for visually impaired (VI) navigation, our research has shown that the VI community sharply divides into white cane users and guide dog owners, with the latter category in particular objecting to the use of a white cane. For this reason we are focusing on smartphone apps delivering verbal and haptic output (which is suitable for both types of users). We view the main communication challenges to be the following.

Interaction Management: Managing (a) the interaction between user and system, including allowing user interrupts and system alerts, and (b) any resulting changes to system behaviour. This includes allowing the user to input navigation and configuration options for the route before or during the journey.

Communicating navigation guidance: In the absence of interrupts from the continuous route monitoring processes described above, the system communicates route guidance along the target path to the user. We will carry out detailed requirements analyses to determine what kind of instructions and what

[1] http://www.openstreetmap.org

level of detail are most useful. While the assumption is that most instructions are best communicated via brief spoken outputs, a core question is what part of the guidance can be delivered by haptic output, e.g. different types/locations of vibration indicating different direction/speed of movement.

Communicating warnings: The properties required of warnings differ from navigation guidance, in that the nature of the danger and the required user reaction need to be conveyed as quickly and as efficiently as possible, with information ordered in terms of urgency. It is likely that a larger proportion of warnings (than of navigation instructions) are best conveyed by haptic and simple audio output.

Communicating uncertainty: If the system detects a hazard in the path ahead, identification of the type of hazard and appropriate user action will come with a confidence measure < 1. The degree of uncertainty in what the system has identified must be conveyed to the user. E.g. if a postbox is tagged in the map, and the continuous monitoring component has detected an object ahead that it has recognised with high confidence as a postbox, then it may be enough to simply steer the user around it. However, if the system detects an obstruction at head height which is not annotated in the map and which it classifies with similar confidence levels as several things, then this uncertainty has to be expressed in the verbal output, and the user may have to further investigate.

Communicating varying levels of detail: Similarly, when describing a hazard or verbalising route guidance, not all the detail about objects and routes available to the system needs to be conveyed to the user in every situation. For this purpose the system design incorporates a content selection component (CSC) which decides the appropriate level of detail given the context.

A suitable way to generate verbal output in line with the above communication requirements is probabilistic natural language generation (NLG) technology (Belz, 2008) which offers the possibility of automatically training the verbal output generator to adapt to different user requirements and usage contexts.

5 Current Work

We are currently in the early stages of developing the various components of the Last 10m system. We have carried out preliminary experiments in indoors path recognition identification (Rivera-Rubio et al., 2013; 2014), and conducted initial consultation sessions with VI people. The next step is to design Wizard-of-Oz experiments in order to obtain sizeable corpora of example instructions (produced by humans playing the role of the system) appropriate in a variety of contexts which is then used both for training NLG components and for other aspects of system design. At the same time we are improving the path computation algorithms (which provide important input to the CSC), using, for the time being, a small number of semantically and visually enriched local maps of entrances at our universities.

References

P. F. Alcantarilla, L. M. Bergasa, and F. Dellaert. 2010. Visual odometry priors for robust EKF-SLAM. In *Proceedings of the 2010 IEEE International Conference on Robotics and Automation*, pages 3501–3506.

A. Belz. 2008. Automatic generation of weather forecast texts using comprehensive probabilistic generation-space models. *Natural Language Engineering*, 14(4):431–455.

A. Davison, I. D. Reid, N. D. Molton, and O. Stasse. 2007. MonoSLAM: Real-time single camera SLAM. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1052–1067.

V. Pradeep and G. Medioni. 2010. Robot vision for the visually impaired. In *Proceedings of the 2010 Computer Vision and Pattern Recognition Workshop (CVPR)*, pages 15–22.

J. Rivera-Rubio, S. Idrees, I. Alexiou, L. Hadjilucas, and A.A. Bharath. 2013. Mobile visual assistive apps: Benchmarks of vision algorithm performance. In *New Trends in Image Analysis and Processing (ICIAP 2013)*, volume 8158 of *Lecture Notes in Computer Science*, pages 30–40.

J. Rivera-Rubio, I. Alexiou, A.A. Bharath, R. Secoli, Dickens, and E. Lupu. 2014. Associating locations from wearable cameras. In *Proceedings of the 25th British Machine Vision Conference*. To Appear.

WHO. 2013. Visual impairment and blindness. Fact Sheet No. 282, World Health Organization.

Q. Zeng, C. L. Teo, B. Rebsamen, and E. Burdet. 2008. Collaborative path planning for a robotic wheelchair. *Disability and Rehabilitation Assistive Technology*, 3(6):315–324.