Portable Navigations System with Adaptive Multimodal Interface for the Blind

Jacobus Lock, Grzegorz Cielniak, Nicola Bellotto
Lincoln Centre for Autonomous Systems (L-CAS)
School of Computer Science, University of Lincoln
Lincoln LN6 7TS, United Kingdom

Abstract

Recent advances in mobile technology have the potential to radically change the quality of tools available for people with sensory impairments, in particular the blind. Nowadays almost every smart-phone and tablet is equipped with high-resolutions cameras, which are typically used for photos and videos, communication purposes, games and virtual reality applications. Very little has been proposed to exploit these sensors for user localisation and navigation instead. To this end, the “Active Vision with Human-in-the-Loop for the Visually Impaired” (ActiVis) project aims to develop a novel electronic travel aid to tackle the “last 10 yards problem” and enable the autonomous navigation of blind users in unknown environments, ultimately enhancing or replacing existing solutions, such as guide dogs and white canes. This paper describes some of the key project’s challenges, in particular with respect to the design of the user interface that translate visual information from the camera to guiding instructions for the blind person, taking into account limitations due to the visual impairment and proposing a multimodal interface that embeds human-machine co-adaptation.

1 Introduction

Blindness and severe visual impairments are affecting a significant part of the world population. Such impairments diminish the quality of life of those affected by hindering their ability to move and perform many tasks independently. Besides its social impact, the problem has also a strong economic one. There are approximately 2 million VI residents in the UK only, and approximately 28 million in Europe, with an estimated annual cost to government of £2 billion and £27 billion respectively (Access Economics 2009; World Health Organisation 2010; Deloitte Access Economics). These numbers are expected to rapidly rise with an increase in obesity and diabetes incidences, as well as an ageing population. There is therefore significant market value and potential interest in devising a system that will allow the VI to independently navigate the world.

The UK’s Royal National Institute for the Blind (RNIB), which is also one of the leading world organisations in this area, has identified a number of challenges for the modern VI person. These include the latter’s ability to safely and independently use public transport services and navigate in unfamiliar environments (rni). Recent technological advances can facilitate the creation of new solutions to address these challenges. GPS technology, for example, has become almost ubiquitous in the Western world. However, its accuracy is quite low – typically a few meters – and is notoriously inaccurate in urban environments, where satellite signals are blocked by high rise buildings, usually unavailable indoor. GPS-based devices are therefore unsuitable to identify and direct a user to high resolution targets such as a building’s entrance or lift. To make a navigation system capable of successfully directing users to their desired location involves the creation of a system that can both localise the users within the world, i.e. by visually processing and understanding its surroundings, and communicate such information to them in a simple but effective way.

Figure 1: Navigation system components and application.

One promising avenue of research is making a hand-held mobile vision system to aid the navigation of the VI, extending concepts originally proposed as in (Bellotto 2013) and (Gallina, Bellotto, and Di Luca 2015). The main goal of the project “Active-Vision with Human-in-the-Loop for the VI” (ActiVis) can be thought as solving a classic control problem where the process is the navigation system and the actuator is the VI user. The idea here is to make a navigation system that can perceive its surroundings and convey safe navigation directions to the VI person in a way that minimises the control error (difference between desired and actual position of the user) as quickly as possible. To our knowledge, very little work has been done in this area, in particular with respect to control systems for humans (Jagacinski and Flach 2003).

To overcome the barrier of user acceptability, ActiVis adopts a standard hand-held mobile devices, such as the one

Copyright © 2017, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

illustrated in Fig. 1 (i.e. Google Project Tango). A key objective of the project is the development of an opportune user interface for this device, enabling the communication between the navigation system and the human. This should obviously take into account the visual impairment, but also the cognitive load which the user is exposed to when high-frequency sensor information are converted into navigation instructions to guide him/her. To this end, an important contribution of ActiVis will be the creation of a multimodal interface that exploit audio and vibration cues to transmit useful information to the VI person in the most effective way.

Another objective of the project is the implementation of a control system to guide the VI user that adapts to his/her skills and varying performance, taking into account the "co-adaptation" factor, i.e. the reciprocal adaptation of humans to the inherent limitations of any user interface. Considering the user as the plant adds an interesting aspect to the control problem: since controllers are typically designed for well-characterised electro-mechanical systems, while humans are known for the variability in their behaviour, it would be impossible to design a "one-size-fits-all" controller. A significant contribution of ActiVis to the active vision and human control fields will be a generic template controller that can adapt its internal parameters over time to its human operator's habits to enhance the complete system's (human operator included) navigation performance.

The remainder of the paper is as follows: Sec. 2 reviews some of the most relevant work, focusing in a particular on navigation aids, interfaces and adaptive systems for the VI; Sec. 3 presents some key aspects of the multimodal interface in ActiVis; Sec. 4 illustrates the co-adaptation concept applied to the proposed navigation system; finally Sec. 5 describes the current progress of the project and future steps towards its objectives.

2 Related Work

Navigation Aids

Producing a navigation and obstacle avoidance system for the VI is the main goal of this project. GPS technology, while undeniably useful, is not applicable here since it does not solve the so-called "last 10 yards (or meters) problem", where the VI user needs to be directed to a specific location 10m from him/her, e.g. the changing rooms in an unfamiliar department store. Different approaches have been investigated in literature; from so-called "smart canes" to RFID guidance and computer vision-based systems.

Early approaches to the problem of safe navigation for the VI involved augmenting the classic white cane with sensors such as ultrasound or laser sensors, to improve obstacle avoidance and warn the VI user of oncoming objects. The GuideCane system created in (Borenstein and Ulrich 1997), for example, is a wheeled apparatus equipped with ultrasound sensors and motor encoders that the VI user pushes along the ground. The system is made in such a way that when an obstacle is detected, the wheels steer away to avoid the object and the VI person avoids the obstacle by adjusting his/her walking direction accordingly. The authors report successful tests with safe walking speeds of up to 1.5m/s. However, the system is clunky and clearly visible, which are two big obstacles to user acceptability and significant market penetration among the VI.

Using Radio Frequency Identification (RFID) technology for localisation and navigation has also been proposed. The authors in (Willis and Helal 2005) describe the system they made where a set of RFID tags encoded with spatial coordinates and location information are strategically placed around the environment. The VI users are then equipped with a cane or similar apparatus augmented with an RFID tag reader. Then, as the VI approaches an encoded tag, the RFID reader extracts the information which is then conveyed to the user. This allows the VI person to localise himself and learn about the descriptive features of the immediate environment. However, even though RFID tags and readers are relatively cheap and compact, installing a number of them across a location could be time consuming and costly since infrastructure might need to be adapted. Also, since the environment is typically dynamic and can change over time (for example a shop can close down), maintaining and updating such a system is also a concern.

More recently, with the rise in interest in computer vision, RGB-D sensors have become popular for navigation and object recognition tasks thanks to their relatively low cost, accuracy and their ability to output a scene's depth information. Various systems have been tested with varying degrees of success. One approach has allowed a computer to scan a scene for door-like shapes (e.g. bathroom doors, fire escapes, elevators, etc.) (Tian et al. 2013) and extract contextual information from the door in the form of text or signage, which can then be conveyed to the VI user. Other approaches involve using RGB-D information to build a 3D map of the environment and update it as the user traverses through it. In (Lee and Medioni 2015) the authors developed such a system, which is also capable of active collision avoidance, generating and continuously updating the shortest route to the VI user's desired destination. The person is directed along the route by a haptic feedback vest, which again can pose problems in terms of usability and user acceptability.

The project described in (Chessa et al. 2016) created a system framework that carries out different computer vision tasks to assist a VI user. These tasks include recognising faces it has seen before, so that it can tell the VI user if some known person is approaching, recognise and count cash amounts, and extract text from a scene to contextualise the environment. This is a promising framework, although only tested on videos recorded by a mobile phone. However it is a "passive" approach, in the sense that it does not actively guide the VI in performing a particular action, and where only a rudimentary user interface has been implemented.

Interfacing

One of the main challenges in to designing a navigation interface for the VI, besides easy ways to input desired destinations, is how to receive simple, unambiguous and accurate navigation instructions and environmental information conveyed in a way the VI user can understand and make sense of it. Feedback media such as vibration, spatialised sound, voice commands and image sonification have previously been considered in literature.

Alves and de Souza (2014) performed a survey on VI's preferred feedback media. They found that the respondents overwhelmingly preferred systems which do not make the VI person stand out from the crowd and do not interfere with their daily functions. Furthermore, the respondents generally selected haptic and audio (including speech) feedback as their preferred feedback media, as opposed to electrostimulation, for example. Interestingly, a haptic feedback-based navigation iPhone app (SpaceSense) received the most favourable reviews from the study's respondents. However, the study also highlighted that none of the systems tested satisfied all of the VI's requirements.

With these results in mind, Frauenberger and Noisterig (2003) created a so-called virtual audio reality (VAR) and implanted it with "earcons" (derived from the words "ear" and "icons"), which represent pre-programmed notable features in the environment. These earcons convey the features' position relative to the user with 3D spatialised sound. The authors report favourable results, but have not formally tested their system with VI users and they do not specify on how they plan to scale the system to be useful in unfamiliar environments where the position of features are unknown to the system. A similar system would be Pieter Meijer's "The Voice" (Ward and Meijer 2010), which translates the image of a scene into a set of sound waves of varying pitch and volume.

On the front of user input, Kane, Wobbrock, and Ladner (2011) found that if the VI user was limited to a typical touch screen interface, he/she showed a strong preference for gestures using screen corners, edges and multi-touch gestures, i.e. two or more fingers. They also found that gestures produced by the VI are larger, slower and display more variation in size than their sighted counterparts. Interestingly, when asked to invent their own gestures, the VI users came up with a technique not commonly seen in typical interfaces designed for sighted users: they would touch the screen with a second finger to activate a specific mode and touch areas of the screen corresponding to QWERTY keys. Furthermore, Arditi and Tian (2013) found that the VI typically prefer to give input to the system through a QWERTY keyboard interface, rather than special icons. They also strongly prefer querying the system via voice commands for output rather than receiving a constant stream of feedback.

None of these factors are typically considered when designing an app interface for a normally sighted user, but as Kane, Wobbrock, and Ladner (2011) and Arditi and Tian (2013) show, they become very significant when the designer is trying to cater for non or partially-sighted users. Recent work (Bellotto 2013), upon which the current research is partly based, has looked therefore at combinations of multimodal interfaces that can provide different sensory cues, audio- and vibration-based, to guide a VI person in pointing a mobile camera sensor towards landmarks and features of the environment useful for navigation (see Fig. 2).

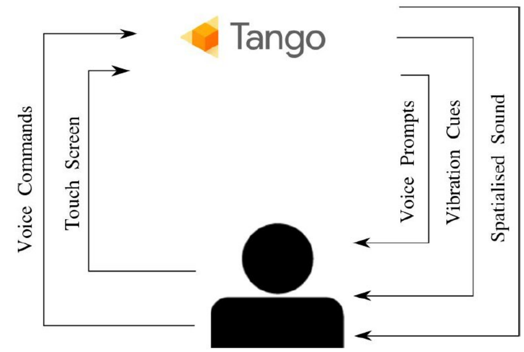

Figure 2: Multimodal interface.

Human-Machine Co-adaptation

Since task performance varies between different users, or even for the same user at different times, a mechanism for adapting to the user skills and actual performance must be implemented to interact effectively with a navigation system for the VI. However, most of the solutions proposed so far do not take into account these factors.

Several adaptive human-machine interfaces exists, which are typically classified as *human-centred* (or *user-centred*) and *goal-oriented*. The aim of the first is to create pleasant adaptive interfaces that maximize usability (Dixon 2012). In this case, the design focuses on user skills and expectations, developing interactive systems that prioritise their use, where human factors/ergonomics are applied together with usability knowledge and techniques (Jokela et al. 2003).

The design of goal-oriented adaptive interfaces, instead, focus on maximising the system's potential, assuming that the user is very skilled in performing the task. In this case, only well trained users can benefit from the adaptability of the interface. An example of goal-oriented interface is the one implemented in "The Voice" (Ward and Meijer 2010), where long training periods of the VI person are required to use the navigation system, due to the amount of visual information translated by the latter to complex auditory signals for the user. More recently, a new "progressive co-adaptation" paradigm has been proposed by Gallina, Bellotto, and Di Luca (2015) that combines human-centred and goal-oriented advantages, from which ActiVis could benefit to overcome the limitations of previous VI interfaces.

3 Multimodal Interface

A key aspect that differentiates ActiVis from most of the active vision systems in the literature (Rivlin and Rotstein 2000) is the human-in-the-loop part, that is, the fact that the process to be controlled is not an electro-mechanical actuator panning and tilting the camera, but the hand, arm and body of the user holding it. Consequently, the control signal u (see Fig. 3) that instructs the movement of the human-actuator is not a particular current or voltage, but a sensory input (audio-tactile interface) for the VI user.

In ActiVis, the multimodal navigation system is based on Google's Project Tango device (see Fig. 1), developed

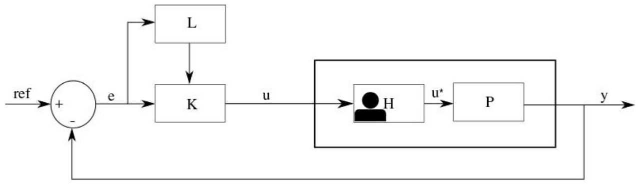

Figure 3: Progressive co-adaptation with human-in-the-loop.

with the Android and Tango SDKs. The latter gives access to three core capabilities: motion tracking, depth perception and area learning. These technologies allow the Tango to accurately localise itself and perceive its surroundings.

The navigation system guides the person toward a particular target using voice commands and spatialised sound, while vibration is used for obstacle avoidance. The spatial sound has 3 parameters that communicate the target's position: the pitch (elevation angle to the target), gain (distance to the target) and 3D panning (whether the target is to the left or right of the user). The solution is partially based on previous work in multimodal interfaces for the blind (Bellotto 2013), extending it here to take into account the augmented perception capabilities of the Tango device.

Thanks to the Tango's localisation ability, getting from point A to B is relatively simple; the real challenge lies in directing the VI user along a route to B as quickly and safely as possible. Building a navigation app with an interface optimised for the visually impaired is obviously crucial here.

The multimodal interface enable also the implementation of a co-adaptive module, responsible to learn the behaviour of the user over time and adapt the navigation system's parameters to improve performance, i.e. get the user to his desired location quicker by conveying its navigation instructions more effectively. Examples include learning the frequency of the user's voice feedback requests, the power of the vibrations required and the user's most common location requests. The following section discusses more in detail the approach that will be used in ActiVis.

4 Progressive Co-adaptation

Most of the existing co-adaptive techniques assume that there is a well-defined link between the optimal interface Ψ and the user' skill Φ to accomplish a task, so in theory it is possible to associate a particular interface, or interface parameter, to pre-determined (static) skill levels. However, in many cases, the interface should also adapt dynamically to possible changes of the user skills over time in order to maximise the overall performance. Therefore, co-adaptive approaches have been proposed that take into account time t as key parameter of the interface, paying attention at the actual adaptation rate of the interface $d\Psi/dt$ to guarantee the stability of the system (Merel et al. 2013).

Recent work extended this further by considering, besides time and current user skills, their changing rate $d\Phi/dt$ (Gallina, Bellotto, and Di Luca 2015). The idea is that, by observing the rate at which the skills vary over time, it is possible to detect when they reach a plateau – i.e. the user is not able to improve the task performance any more. The interface could be regulated then by the variations $d\Phi/dt$ to implement a "progressive co-adaptation" where the user-navigation system achieves gradual but steady improvements in task performance that goes beyond local maxima.

In ActiVis, the latter form of co-adaptation will be implemented in a parameterised user interface, including variable amplitudes and frequencies of voice instructions, sounds and vibrations, in which the measurable user skills are those related to his/her capability to move and point the mobile device as required by the navigation system. The adaptive module L in Fig. 3 will include therefore a control law that is a function of the user skills (measured, for example, as average error between actual and desired position of the device) and the rate at which they change.

5 Current and Future Work

The current implementation of the ActiVis system is an Android app based on a Tango device and an AfterShokz Sportz 3 bone-conducting headset (see Fig. 1). Through the latter, the user can listen to sounds and voice instructions generated by the Tango without affecting normal hearing. Along with a set of sensors commonly found in modern smart phones (vibration, 3-axis gyroscope and accelerometer, touch screen, GPS, etc.), the Tango comes equipped with a set of RGB-D camera's that allows it to perceive its environment in 3D.

Developing a navigation app on a mature, widely used and well-supported Android platform provides a multitude of input/output options. For example, speech recognition and screen gesture support are built into it. Furthermore, Android's native code support makes it easy to integrate C/C++ libraries such as OpenCV (for computer vision tasks), OpenAL and OpenSL (sound generating libraries with sound spatialisation support) into an app, presenting an opportunity to create a unique, interesting and rich input/output interface for a VI user. Adding a peripheral device (e.g. via USB) to produce more diverse vibrations can also be considered.

The app is mostly written in Java using the standard Android SDK with extra Tango libraries and support, and it uses the OpenAL native library to generate spatial sound signals. The system is capable of recording various parameters in real-time, including 3D position and rotation of the device. The Android SDK also allows the streaming, recording and visualisation of these parameters in real-time using a PC connected to the Tango's WiFi network.

Virtual Cane

The app's current implementation is inspired by the white cane widely used by the VI. By using the depth data from the Tango's RGB-D sensor, the system can tell the user when an obstacle is approaching by using the Tango's built-in vibration interface. The obstacle's distance can also communicated by adjusting the vibration intensity.

Furthermore, a navigation feature is built into the Virtual Cane app, which allows the user to program a destination. The person is guided towards the latter using the multimodal interface described in Sec. 3, which includes voice commands and spatialised sounds.

For testing purposes, the system takes advantage of the Tango's place recognition capabilities, which allow for more robust and reliable position tracking over longer distances. It provides also a platform for testing the multimodal feedback with VI users and record their performances in navigation tasks for further studies in human-machine co-adaptation.

Planned Experiments

To create the co-adaptation module, it is necessary to know how different feedback modes affect each individual user's navigation performance, e.g. whether more intense vibration results in more effective obstacle avoidance. A set of experiments have been planned to determine how the four feedback parameters of the multimodal interface affect user performance. These parameters are the pitch gain rate, the volume gain rate, vibration period and voice command frequency. The sound panning is controlled by a separate library, so it is not include as a variable in these tests.

The first set of experiments is to have the user traverse through an obstacle course on a 2D plane where the exit of the course will be the final target. Since the latter is at a fixed height, the pitch variable is not considered here, which simplifies the experiment. Only the vibration, voice commands and volume gain are used for feedback.

A separate set of tests will be conducted to determine the effect of the pitch, where the user will be directed to point the Tango to look at targets floating at different levels and distances to the user. Here the vibration parameter will be eliminated and only the voice commands, pitch and volume gain will be used as feedback parameters.

Next Steps

After the test data has been analysed to have a better understanding of how the different feedback parameters affect a user's navigation performance, the project will continue to create and integrate an autonomous co-adaptation module into the Tango. This will be responsible for monitoring a user's navigation performance and adapt the feedback parameters in real-time to improve such performance.

Acknowledgment

This research is partially funded by a Google Faculty Research Award (Winter 2015).

References

Access Economics. 2009. Future Sight Loss UK(1): The Economic impact of Partial Sight and Blindness in the UK Adult Population. Technical report, RNIB.

Alves, R. D. S., and de Souza, A. S. 2014. Multimodal and Alternative Perception for the Visually Impaired: A Survey Structured. *Igarss* 1 – 5.

Arditi, A., and Tian, Y. 2013. User Interface Preferences in the Design if a Camera-Based Navigation and Wayfinding Aid. *Journ. of Visual Impairments and Blindness* 118 – 129.

Bellotto, N. 2013. A Multimodal Smartphone Interface for Active Perception by Visually Impaired. In *IEEE SMC International Workshop on Human Machine Systems, Cyborgs and Enhancing Devices (HUMASCEND)*.

Borenstein, J., and Ulrich, I. 1997. The GuideCane - A Computerised Travel Aid for the Active Guidance of Blind Pedestrians. In *Proceedings of IEEE International Conference on Robotics and Automation*, 1283 – 1288.

Chessa, M.; Noceti, N.; Odone, F.; Solari, F.; and Sosa-García, J. 2016. An Integrated Artificial Vision Framework for Assisting Visually Impaired Users. *Computer Vision and Image Understanding* 209 – 228.

Deloitte Access Economics. The Cost of Blindness. Report, to be published in 2016.

Dixon, J. 2012. *Human Factors in Reliable Design*. John Wiley & Sons, Inc. 137–155.

Frauenberger, C., and Noisterig, M. 2003. 3D Audio Interface for the Blind. In *Proceedings of the International Conference on Auditory Display*, 280 – 283.

Gallina, P.; Bellotto, N.; and Di Luca, M. 2015. Progressive Co-Adaptation in Human-Machine Interaction. In *12th International Conference on Informatics in Control Proceedings (ICINCO 2015)*, 2362 – 368.

Jagacinski, J. R., and Flach, M. J. 2003. *Control Theory for Humans*. Laurence Erlbaum Associates, Inc.

Jokela, T.; Iivari, N.; Matero, J.; and Karukka, M. 2003. The standard of user-centered design and the standard definition of usability: Analyzing iso 13407 against iso 9241-11. In *Proceedings of the Latin American Conference on Human-computer Interaction*, CLIHC '03, 53–60.

Kane, S.; Wobbrock, J.; and Ladner, R. 2011. Usable Gestures for Blind People: Understanding Preference and Performance. In *Proc. of the Annual Conf. on Human Factors in Computing Systems (CHI)*, 413 – 422.

Lee, Y. H., and Medioni, G. 2015. RGB-D Camera Based Wearable Navigation System for the Visually Impaired. *Computer Vision and Image Understanding* 3 – 20.

Merel, J.; Fox, R.; Jebara, T.; and Paninski, L. 2013. A multi-agent control framework for co-adaptation in brain-computer interfaces. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, NIPS'13, 2841–2849. USA: Curran Associates Inc.

Rivlin, E., and Rotstein, H. 2000. Control of a camera for active vision: foveal vision, smooth tracking and saccade. *International Journal of Computer Vision* 81 – 96.

Uk vision strategy. `https://goo.gl/BTdLPp`. Accessed: 19-07-2016.

Tian, Y.; Yang, X.; Yi, C.; and Arditi, A. 2013. Toward a Computer Vision-Based Wayfinding Aid for Blind Persons to Access Unfamiliar Indoor Environments. *Machine Vision and Applications* 521 – 535.

Ward, J., and Meijer, P. 2010. Visual Experiences in the Blind Induced by an Auditory Sensory Substitution Device. *Conciousness and Cognition* 425 – 500.

Willis, S., and Helal, S. 2005. RFID Information Grid for Blind Navigation and Wayfinding. In *Proc. of the 9th IEEE International Symposium on Wearable Computers*, 34 – 37.

World Health Organisation. 2010. Global Data on Visual Impairment. Technical report, World Health Organisation.