CDI-Type II: Cyber Enhancement of Spatial Cognition for the Visually Impaired

PI: Kostas Daniilidis, Department of Computer and Information Science, University of Pennsylvania,
coPI: Nicholas Giudice, Department of Psychology, University of California, Santa Barbara
coPI: Roberto Manduchi, Department of Computer Engineering, University of California, Santa Cruz
coPI: Stergios Roumeliotis, Department of Computer Science, University of Minnesota

1 Introduction

Most sighted people have never considered what information they use to navigate their environment: how to avoid obstacles, recognize landmarks, stay oriented, or build up a mental representation of the space. These are not tasks consciously learned, they are simply the product of an effective visually-guided perceptual-motor coupling. By contrast, non-visual navigation, as for blind people, relies on sensing from touch, audition or olfaction. This compensatory information is generally proximal, transient, and conveyed via a low-bandwidth sensory channel. Thus, much of the information about distal objects and global structure afforded by vision is unreliably specified, if not completely inaccessible, from non-visual sources of environmental access. Although behavioral studies generally find similar route-learning abilities between blind and sighted people, blind navigators often exhibit significantly greater difficulties on tasks requiring spatial updating (keeping track of self-to-object relations during movement), understanding the spatial relations between locations which are not directly connected by a route, and building up an accurate cognitive map (mental representation) of the space, [16, 19, 12]. Since we contend that most challenges of blind navigation stem from insufficient access to navigation-critical information, the obvious solution, as described in this proposal, lies in discovering a means of conveying the requisite cues to mitigate the problem. Several lines of evidence support the hypothesis that providing *compensatory spatial information aids blind spatial learning and navigation* [12]; evidence which is congruent with the research findings [6] and phenomenological experience of Giudice, one of the co-PIs who is himself blind.
This proposal deals with the most challenging components of blind navigation, i.e. learning and navigating a space beyond simple route travel, and involves the most vexing environments (complex indoor layouts). Let us start with a common scenario which illustrates some of the immediate challenges faced by a blind wayfinder. Jessica and her guide dog Zeek get out of her cab at Chicago's O'Hare International airport. She enters the terminal and pauses to figure out where her airline's check-in counter is located. This is a common scenario for all air travelers; however, the environmental information guiding Jessica's actions is likely very different from her sighted peers. She needs to not only find the check-in counter but also her place in the cue of people waiting in line. She then needs to negotiate the security checkpoint, determine the correct concourse, find her particular gate, and then determine an open seat in the waiting area. To enable return travel, she would also like to know about the landmarks passed along the way, such as cafes, bathrooms, etc.

Figure 1: Blind user of GPS and a Braille PDA.

In addition to the limitations of non-visual modalities, several factors contribute to the increased difficulty of indoor navigation for blind pedestrians compared to the same tasks performed in outdoor settings (cf. Fig. 1). Outdoor navigation is generally done in spaces with a clear and consistent spatial and semantic structure. For instance, knowledge of street naming conventions and address numbering sequences provide critical information about a navigator's position and direction of travel in the city. The analogs of these environmental regularities are generally not available (or poorly specified) within buildings. That is, indoor environments are rarely arranged with corridors organized according to a meaningful naming/numbering system. Sighted people benefit from information from you-are-here maps, signs and color coding to help solve the indoor navigation challenge but such cues are rarely accessible to a blind wayfinder.

2 Objectives

This proposal adopts a multi-faceted approach for solving the indoor navigation challenge. It leverages expertise from *behavioral studies* with *algorithm design* to guide the *discovery of information requirements and optimal delivery methods for an indoor navigation system*.
To better understand our goal, let us consider the following metaphor: Imagine that a blind person has a

sighted human companion who would sense everything and verbally convey information without manually guiding the person. What information should the human guide convey so that a blind person builds the appropriate spatial representation? We realize that this is a non-trivial task as even in this ideal case, access to all sensory cues does not automatically mean perfect navigation. The questions of how to best understand and help wayfinding for blind persons and how to build an appropriate navigation system are irrevocably intertwined. We have three objectives in this proposal:
1. To determine the optimal information that would help a blind person find their way without confusing or overloading him/her. This will require to deal with the abundance of visual, range, and inertial information that a cyber-helper acquires from a scene.
2. To design algorithms for localization and semantic information extraction from the environment which will be implemented on an experimental portable device. Results from the first objective will be used as feedback to improve our algorithms and our output. In particular, we will use input from cameras, range scanners, and inertial sensors to perform path integration and recover a map of the environment as well as parse and semantically annotate the surroundings. Mapping and semantic annotation are daunting problems which are far from solved in robotics/vision, when dealing with complex indoor environments.
3. To create a community-based system of information exchange where blind people will contribute with metric maps as well as images. Landmark images can be associated with simple "tagging" phrases spoken by the blind user upon successful recognition of landmarks via sound or touch. We will design algorithms on how to merge online with pre-stored maps, as well as, on how to incrementally build a database of visual landmarks with their associations.
This is not a proposal about building a new artifact for an indoor navigation system for the blind. It is rather an experimental setup robust enough to conduct a series of behavioral studies in large scale indoor environments.
Our team combines expertise from computer vision, robotics, and psychophysics. It is rarely the case that a blind researcher like Nicholas Giudice (UC Santa Barbara, psychophysics and spatial cognition) is directly involved in studying assistive technologies and conducting behavioral studies. Kostas Daniilidis (UPenn, computer vision) has a long experience leading NSF funded projects on omnidirectional vision, tele-immersion, digital archaeology, and navigation. Roberto Manduchi (UC Santa Cruz, computer vision) has already experience in object and text recognition for the blind, in research funded by NSF and NIH. Stergios Roumeliotis (UMinnesota) is a roboticist with long expertise in estimation theory and localization, who has already designed a white-cane sensor and received funding from NSF and NASA.

3 Technical Approach

3.1 Discoveries in Spatial Cognition

Related Work Blind wayfinders have benefited enormously from the development of speech enabled GPS-based navigation systems allowing for precise world-wide positioning utilizing huge commercial databases of streets, addresses, and points of interest [9]. While these technologies have solved many of the challenges of outdoor navigation for blind people, the lack of similar technology for use in buildings means there is a glaring hole in supporting safe and efficient orientation, wayfinding and cognitive map development for indoor usage. The limited technology that is available either provides information about discrete landmarks only or requires significant modifications to the building infrastructure [9].

Preliminary Results Giudice and Legge discuss four factors that should be considered in the design and implementation of technology for blind navigation [9]. (1) Sensory Translation Rules: what is the optimal mapping between input and output modalities? (2) Selection of Information: Since the complexity of the display is directly proportional to the amount of information presented, what is the best output and resolution of spatial information to be displayed? (3) Device Operation: How does the environment and sensor characteristics influence performance? (4) Form and Function: what is the aesthetic impact of a device on

the user? Fortunately, as is described in [9], points 1 and 4 have solid precedent on which to base our work. Thus, our proposal is primarily concerned with points 2 and 3.

Challenges and Proposed Research There is little known about exactly what information should be conveyed, how it should be provided, or how different sensor configurations and environmental parameters will affect performance. Research on such issues, as is proposed here, is imperative to solve the indoor navigation challenge and must occur before a general purpose, commercial system can be developed. Although there is currently no existing device to aid in indoor navigation, we believe that the proposed research incorporating off-the-shelf technology, requiring no infrastructure investment, and utilizing rigorous behavioral studies, will lead to the most robust, usable, and effective system in the future.

We plan to conduct experiments that will enable us to better understand the spatiocognitive abilities of visually impaired and help us improve wayfinding systems to increase the safety and efficiency of indoor navigation. As was illustrated by the previous airport example, there are many challenges to indoor navigation without vision. The solution to this vexing problem involves understanding theories of spatial learning and cognitive map development by blind persons, as well as identifying how this basic knowledge can be applied to spatial displays to support navigation. Both issues have been addressed by Giudice and his colleagues in a series of experiments. Their work showed that verbal information, which is generally used for sequential tasks only, e.g. route following and direction giving, can be employed to support free exploration, wayfinding, and cognitive mapping behavior in complex buildings [7]. The critical factor for a verbal interface to work as a substitute for visual access is to describe the relevant spatial primitives of the environment in a dynamically-updated manner, i.e., the messages change with respect to the navigator's position and orientation [6, 8]. This work has also shown that dynamic virtual verbal displays (VVDs) can be used to learn and navigate computer-based virtual environments (VEs) and that learning with the VVD transfers to physical navigation of the corresponding real layout [4]. Furthermore, verbal learning in VEs is improved [10] and requires less cognitive load and working memory capacity [5] when spatialized audio (where objects are heard as coming from a specific 3-D location in space) are used vs. traditional spatial language descriptions (e.g., the door is at 10 o'clock in 20 feet). This proposal builds on earlier experimental work by addressing several unresolved questions about what information should be provided to a blind person during real-time navigation and how this information should be displayed to support independent spatial learning and wayfinding behavior. Specifically, we propose to provide answers to the following questions:

1) What are the minimum information requirements to describe an indoor environment so a wayfinder knows where they are and what is around them? Previous work has only looked at the basic geometric structure of a layout and has not addressed what environmental information will best support blind learning and navigation. 2) Does spatialized auditory information improve performance over the spatial language descriptions which are traditionally used in commercially available GPS systems? Our work suggests an advantage in virtual environments and this proposal extends the study to real environments. 3) What is the best output resolution from the display? That is, how much environmental information should be conveyed? 4) What is the best frequency of information delivery, i.e., how often should information be updated?

3.2 Localization and Mapping

Related Work and Current Limitations: Significant research has concentrated on mobile robot navigation [17]. However, there are only a few attempts to apply this knowledge to assist visually impaired people during their everyday navigation tasks. Most work on wayfinding aids has focused on outdoor navigation and has led to the development of speech-enabled GPS-based navigation systems. In contrast, the limited technology that is available for indoor navigation requires significant modification to the building infrastructure, whose high cost has prevented its wide use. Additionally, these systems are large and uncomfortable to wear (e.g., vests) or require to carry around robots meaning these devices have limited acceptability as indoor travel aids [9]. Despite the availability of a few indoor navigation systems for visually impaired humans, current research efforts suffer from relying on expensive insfrastructure (RFID, signs) and use

no information from the surroundings of the person to improve his/her positioning accuracy. In contrast to these approaches, we propose to incorporate both exteroceptive (e.g., micro-camera) and proprioceptive (e.g., nano-inertial measurement unit; nIMU) sensors so that we can simultaneously estimate the pose of the user and generate useful information about their surrounding environment.

Preliminary Results: In our recent work [11], we have designed and implemented algorithms that can estimate the position of a blind person using measurements from (i) a shoe-mounted pedometer, (ii) a 3-axial gyroscope, and (iii) a small-size laser scanner. These last two sensors are mounted under the handle of a white cane. We have chosen a white cane as the main sensor platform because it is light weight, portable, and unobtrusive to the user. Furthermore, it is a trusted tool, already in use by the target demographic, that allows the user to physically touch the environment.
They key idea behind our current map-based human localization technique is the following. We consider the case of an individual navigating inside a building whose layout is represented by a *sparse metric map* containing the location of salient features (wall corners). Since the main source of positioning uncertainty is the heading-estimate error, we integrate the rotational velocity measurements from the gyroscopes to predict the person's motion direction and provide periodic corrections from a "structural" compass (i.e., the laser scanner that detects large planar surfaces such as surrounding walls and the floor). The resulting heading estimates are used to integrate the distance measurements from the pedometer and compute the position of the person. Finally, by matching corner features detected by the laser scanner to the ones appearing on the map, periodic position corrections are provided that bound the positioning error to less than 16 cm.

Challenges and Proposed Research While maps are available for certain public buildings, for many others no map exists or available maps are out of date. Furthermore, most CAD drawings do not contain the location of large obstacles or landmarks such as desks, closets, and vending machines, which a visually impaired person must be aware of. Therefore a navigation aid must be able to adaptively create, maintain, and update a map of an area of interest. This is known as the Simultaneous Localization and Mapping (SLAM) problem. Accurately mapping the surroundings along the path traversed can significantly increase the accuracy of the position estimates when each of these areas is re-visited (e.g., when closing a loop around a corridor, or, tracing back a path along the same corridor).
The main challenges associated with this task are (i) reliably detecting and identifying unmapped landmarks, and (ii) reducing the quadratic, in the number of mapped features, computational complexity of Kalman filter-based SLAM. Although SLAM approaches have been successfully applied within small areas, their processing requirements makes their extension to large-size buildings quite challenging. Approximate algorithms proposed in the past to address this issue have serious limitations since they cannot quantify their information loss and can lead to divergence (loss of position track) due to estimator inconsistencies [14].
Our approach to address this problem is to design a hierarchical scheme that classifies and processes visual features based on their information content. Intuitively described, when humans navigate they memorize only a small amount of visual information (images) for creating a basic map of the area they move in. The rest of the features are used by our visual system for estimating how fast and in which direction we are moving. Our objective is to automate this process for supporting real-time precise human navigation. Specifically, the key idea behind our approach is to use the vast majority of the visual features detected – termed Opportunistic Features (OFs) – for directly estimating the person's motion through a linear complexity (in the number of features detected in a short sequence of images instead of the whole map) algorithm [13]. The remaining visual features – termed Persistent Features (PFs), i.e., the few that can be reliably detected once revisiting an area – are used to construct the skeleton of the map.

3.3 Semantic annotation of scenes

Related Work and Current Limitations: Objects are visually recognized on the basis of their shape and their color appearance. Current recognition approaches are trained on samples of the appearance and the 2D

outline of objects in annotated databases. Appearance is captured by local features which might be locally affine invariant but are still sensitive to viewpoint and illumination changes [15]. Object outline detection relies on successful contour extraction and segmentation and captures only the shape of the occluding contours. Even when successfully detected they are classified using 2D templates capturing a subset of the possible object poses [3].
In assistive systems, semantics have been conveyed to the user by reading signs or building an appropriate infrastructure with uniquely identifiable markers. Written information and signs represent a typical approach to providing localized directional information to sighted persons. Signage is particularly useful for quickly figuring out which direction to move to in order to reach a certain destination (e.g., a gate in an airport), as well as for characterizing a particular location (e.g., the name of a store). Automatic text reading and sign interpretation are therefore important features of the proposed technology. Existing technology for text/sign detection and reading [1, 20] has given promising results but does not suffice for navigation in an indoor enviroment with clutter and complex illumination.

Preliminary Results In ongoing work we have focussed on image matching and how to detect that the same place has been revisited. We introduced the concept of co-saliency [18] as a similarity measure between images that rewards not only intra-image feature similarity but also intra-image grouping. We have applied it successfuly in place recognition in outdoors scenes. In addition, we have been working on completely automatic systems for registration of 3D metric maps with a lot of clutter, a problem that appears when the system tries to find existing maps among those available from community based contributions. With respect to text recognition, we have developed and tested a cellphone-based sign recognition and bar code reader, specifically designed to support wayfinding for a blind person [2].

Challenges and Proposed Research The very first challenge is to find signage and text, a problem associated with two fundamental issues: (i) unfortunate camera setting (worn by the user or mounted on a white cane) exhibiting motion blur, incorrect exposure, and low resolution, (ii) selection of relevant text. Images taken in a public space may contain text from a whole variety of sources: signs posted in the environment, but also advertisements, titles from a book stand, price listings from a coffee shop, etc. Semantic characterization of a space thus requires sorting through all detected text strings and signs in order to determine the most stable (non-transient) and meaningful ones, such as directional signs or room labels. This can be obtained in two ways: by ranking the content based on a-priori knowledge of the expected labels; and by comparing data across video taken at different times, in order to single out transient, and therefore less reliable, semantic content.
Looking only at text and signage would be the same as navigating in a dark room with only signs lighted. We want our system to be able to parse the scene into components which it has seen before or lie in model databases. Indoor environments like airports are crowded with moving humans and contain a lot of static visual clutter irrelevant to the purpose of wayfinding but possibly useful in recognizing a specific place. While we will exhaust all the possibilities of pure appearance-based approaches [15] we will use pre-stored geometric models because many components in indoor environments are man-made. The challenges here are to select what descriptors to use and how to match models and scene segments when only few parts are visible and possibly corrupted by clutter. To this end, we propose to make use of multiple views since the camera is moving and build descriptors that can efficiently compare models to reconstructions from multiple views. Every object is associated with a model as well as neighboring text labels.
When entering a new place we will register it with the previous instances to best exploit previously recognized objects by expecting them after registration at known positions. Given an online image, we will compare not only the appearance through histogram comparison of "bags of features" but also their 3D metric maps based on the modes of the histograms of normals. Even if large parts of the scene, such as standing people, obstruct the scene, our approach selects inliers by comparing histograms at multiple offset positions.