Library Declaration Form

University of Otago Library

Author's full name and year of birth: Nabeel Younus Khan,
(for cataloguing purposes) 17 December 1982

Title of thesis: Indoor scene positioning for blind people using vision

Degree: Doctor of Philosophy

Department: Department of Computer Science

Permanent Address: 85 Harrow Street, North Dunedin, NZ

I agree that this thesis may be consulted for research and study purposes and that reasonable quotation may be made from it, provided that proper acknowledgement of its use is made.

I consent to this thesis being copied in part or in whole for

i) a library

ii) an individual

at the discretion of the University of Otago.

Signature:

Date:

Indoor scene positioning for blind people using vision

Nabeel Younus Khan

a thesis submitted for the degree of

Doctor of Philosophy

at the University of Otago, Dunedin,

New Zealand.

30 June 2013

Put the abstract in this file

Acknowledgements

Put the acknowledgements in this file

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Navigation systems assist the process of localizing, planning and following a specific route where localization refers to identify one's position in space. The location information is firstly used by the navigation system to identify the current position on the map and secondly to guide the movement from one point to the other. A precise localization is therefore very important and is required for an effective performance of a navigation system. The navigation systems provide assistance either to the visually impaired people (especially blind) or robots. Blind people or robots both have no perception about the environment surroundings and require step by step instructions to navigate in unfamiliar environments via voice messages or control signals.

The localization is not a problem for humans without any visual impairment. They extract and use different pieces of information such as landmarks, sign boards, color cues, object etc from the current scene to recall and recognize the current place. But on the other hand, blind people cannot use any visual clues for the localization. The blind person has only access to the sound information of the surroundings but that information is either not enough or totally missing. Therefore, blind people rely totally on near by people for the desired information. Blind people need the positioning information quite often for a rough guidance during navigation in in unfamiliar environments which highlights the importance of a robust localization.

The basic localization can be achieved without any special instruments e.g. sailors have been using astronomical objects for localization for a thousand years during sea travel. Many specialized tools were developed for more accurate localization during the early centuries most notably the astrolabe, compass, sextant, marine chronometer etc. The major advancements in this area are observed in the 20th century and different electronic tools have been invented. In the 1920s, the localization based on radio signals from shore based transmitters was proposed which helped in further improving the sea based navigation . The first practical radar system was developed in 1935 which was capable of determining the presence and range of an object, its position in space, its size and shape etc. It was used for marine, controlling air traffic, detecting weather patterns and tracking spacecraft. In the late 1960s, the United States Department of Defense began developing a satellite based localization system for military purposes which eventually led to the Global Positioning System (GPS) in 1978 (Pace (1995)). This space based radio navigation system provided accurate positioning to within about 30 feet as well as velocity and time worldwide in any weather conditions. The modern GPS systems can provide accuracy up to 3 meters.

GPS is the most popular system to find the location and the position of the objects for navigation systems. It receives signals from multiple satellites and determines the physical location through a triangulation process. GPS work extremely well in outdoor but GPS satellites are often occluded by buildings or trees in urban environments. Therefore many systems use GPS information along with other sensors such as vision(Agrawal and Konolige (2006); Leung *et al.* (2008)), geographical information system (Bonnifait *et al.* (2008)), inertial measurement units (Georgy *et al.* (2011)) etc for better performance during urban navigation.

GPS is an outdoor positioning system and does not work in indoor because GPS signals cannot penetrate in the buildings. However high sensitivity GPS can provide indoor positioning but signals are heavily attenuated and reflected by the building materials making it highly unreliable. It's is also hard to extract reliable high level details such as floor level information etc from GPS signals. The indoor localization without GPS is still an open and universal research problem. The cameras have proven to be good sensors for indoor positioning in terms of cost and reliability (Filliat (2007); Kang *et al.* (2009)). The localization systems based on vision attempt to match the captured image of the current scene with a labeled database of trained images to identify the current location. There are two main problems with most of the existing indoor vision based solutions: (1) few indoor places are used for experiments, and (2) focus is kept on non-office buildings. The scene localization becomes quite challenging in office indoor buildings where different places look much similar as shown in Figure 1.1.

In order to address the mentioned limitations, we have proposed an "Indoor Positioning System (iPoS)" which is currently based on a client server model and targets the blind users. The smartphone application allows its user to take a picture of the

(a) Central hall of 3rd floor. (b) Central hall of second floor.

Figure 1.1: Images of two locations with in the office building with huge similarities.

indoor scene, sends it to the server to find the best match and then generates a voice message on the smartphone to indicate the current location. The algorithms which enable iPoS to operate robustly include firstly, a proposed visual Bag of Words (BoW) scheme for relevant image retrieval. Secondly a verification method to verify the location match followed by further validation if required. We have tested our system in large indoor environments and got good results. Our results show that single camera is a feasible sensor for indoor positioning in real time in challenging indoor environments.

1.1 Motivation

There are about 11,500 people in New Zealand who are either blind or have a low vision (RNZFB (2012)). By 2020, the population of blind people in New Zealand is estimated to rise up to 18,300. While population of blind people have risen to 39 million all over the world lately (WHO (2012)). Therefore, a cost effective solution is required for the blind people in coming years.

Indoor positioning systems based on non-vision sensors such as infrared light (Want *et al.* (1992a)), ultrasonic (Ko *et al.* (2008)), Wi-Fi (**?**) and Active-RFID (Ni *et al.* (2004)) have been proposed. The indoor localisation system based on any sensor should at least meet some requirements: (1) equipment cost should be less, (2) positioning accuracy should be high, (3) handling of multiuser should be there, and (4) easy to use for the blind people. The performance comparison of different sensors keeping in mind the mentioned requirements in summarized in Table 1.1 (Kawaji *et al.* (2010)). The image processing or vision solutions based on a mobile camera seem to offer a cost effective solution with reasonable accuracy. Therefore, the main motivation our

Table 1.1: Performance comparison of existing technologies

Technology	Positioning accuracy	Equipment cost
Infrared light	5-10m	High
Ultrasonic	1-10cm	High
RFID	5cm-5m	High
Wi-Fi	2-100m	Low-High
Audible sound	5-10m	Middle
Mobile camera	1-5m	Low

research work is to:

> *come up with a robust localization system based solely on vision which is easy to use and can provide rough guidance to blind people during navigation in office and non-office buildings.*

The other motivators for our work are:-

- Most vision based localization works focus on outdoor environments.
- The literature shows not much work is done in large scale indoor environments using vision. Most works limit experiments to few indoor places.
- Localisation is often performed in context of non office buildings. The positioning becomes a challenge in office buildings where places are pretty similar. Therefore, a localisation system performance needs to be analysed in office environment.
- Video stream is used as an input for the real time localization and mapping in many works (Gemeiner *et al.* (2008)). This is not ideal for blind people because video stream will drain the phone battery and will require often charging of the phone. This will only add more to the worries of blind people.
- The localization module is often embedded in the navigation systems and is not available as a separate independent unit.

Blind people carry smart phones for making calls, reading emails etc and are familiar with most of the functionality. The majority of the smartphones has a reasonable camera and a smartphone application based on vision will be really handy for the blind people. Blind people can use the application when they feel they are lost while navigating in indoor buildings.

1.2 Challenges and contributions

Most of the successful scene recognition work is done in outdoor environments. Comparatively, less work is done for office indoor buildings. It's quite hard to find a framework in which smartphone uses a single image of the current scene and performs a robust localization. Amongst, the challenges we faced were:-

- Identifying the reliable features which are capable of good image matching with different transformations in any environment.
- People propose and use different techniques to verify the images retrieved from visual BoW but on different data sets. So there is no standard way for images verification.
- The absence of office building data sets for the experimental purposes.
- Designing a user friendly interface for our smartphone application to ptovide maximum accessibility to the blind uers.
- Understanding the camera model coordinate systems and transformations between 2D features (query image) and 3D models (points) for effective pose estimation.

We followed a bottom up approach. We first identified the best features to be used for image matching in our work. We then started with the image matching in outdoor environment followed by the experiments in our indoor data set. After a careful analysis, we proposed a robust image matching system which can localize the current scene. Finally, we developed the smartphone application to send the query pictures to our system (running on a server) and evaluated the localisation performance on real indoor data sets. Our main contributions are:-

- We propose a shorter version of scale invariant feature transform (SIFT) features (Khan and McCane (2012)) for the image matching and compare it with other well known feature descriptors.
- We evaluate the performance of our shorter SIFT features thoroughly against different image transformations (Khan *et al.* (2011b)).
- We evaluate different ranking functions namely normalized term frequency (ntf), normalized term frequency inverse document frequency ($ntfidf$) and Okapi $BM25$ for the visual BoW for the analysis.

- We compare soft and hard assignments schemes in visual BoW for the indoor environment (Khan *et al.* (2011a)).

- We propose a visual BoW based on a voting module and a verification method. Voting module is not only efficient but also effective.

- We compare different verification methods with our proposed homography method for the analysis (Khan *et al.* (2011a)). Our proposed homography verification method gives comparable performance to the fundamental matrix based verification and is also efficient.

- We propose a track based approach to effectively reduce the features from large scale data sets up to a 50% Khan *et al.* (2012a).

- We develop a working android application (Khan *et al.* (2012b)). The application is refined over time in functionality and thoroughly tested on various indoor data sets. The current application offers the simplest interface which is suitable for blind users (Khan and McCane (2012)).

- We compare the localisaiton performance based on pose estimation and 2D image matching with different mobile devices for analysis. We also propose a hybrid algorithm for an effective indoor image matching with almost zero wrong matches.

- We develop five indoor data sets which can be used as a standard to test the classification performance in indoor.

1.3 Application

Our system is intended to assist the blind people whilst they navigate in indoor buildings. However, it is also applicable to other users:-

1. **Prospective Students :** It's hard to figure out the different places in the campus in the beginning. Our application can offer a natural way of learning information about the campus by taking pictures and getting pop up text messages. The new students can utilize this application to get familiar with the campus environment.

2. **Large malls:** It's easy to get lost in large shopping malls. This application could help shoppers to find where they are and where they want to go.

3. **Tourists:** Each city has a history reflected by the historic landmarks (e.g. buildings, statues, historic trees etc). Our application can serve a guide for tourists, giving information about the captured landmark images from the smartphone therefore promoting the tourism.

1.4 Limit of scope

In simple scenario, the system takes the picture of the current scene and identifies the current location. In real time, there are different factors which can affect the performance of such system. It's is not possible to address all challenges in a single PhD project. The limitations of our work are:-

1. The initial target of our work was a complete navigation system for blind people but we ended up with localization due to time constraint. We also did not get time to integrate our localisation module with any existing navigation system for further analysis.

2. We have not tested our system with blind users so far. However, the application may need to be voice powered (to launch/ close the application) to make it more accessible.

3. We use a non- incremental approach and the mapped images of the building are not updated whilst application performs localisaiton. The system does well with smaller changes with in the indoor places as shown in Figure 1.2. The maps will need updation if major changes happen in indoor places. This motivates the use of incremental approach to update the trained images during navigation (Angeli *et al.* (2008)).

4. Our system works well with a couple of people in the scene as shown in Figure 1.3. The system capability to handle more people or crowd is not tested and is not the focus of our work.

5. Our vision based localization system currently runs on the server. The smartphone application is just an interface for sending pictures and getting location information. We use this configuration for simplicity but whole system can be deployed on the smartphone.

(a) Trained Image.

(b) Query Image.

Figure 1.2: Images of the same locations taken after some time.

Figure 1.3: Query image taken at night.

6. The mapped images of the building intended for navigation must be be labeled. We have developed "Map builder" module which needs an excel file filled from the user to generate the labels. The automatic labeling can also be done by using scene semantics or by image classification (Ranganathan (2012); **?**) and is not the focus of our work.

1.5 Thesis Layout

This thesis describes the development of iPoS. To achieve this, the used algorithms and corresponding evaluations on data sets are discussed in the chapters one by one. This thesis consists of nine chapters and details are as follows:-

1. **Chapter 2** gives an overview about navigation system requirements in context of blind users and presents some navigation tools which are in use of blind people.

2. **Chapter 3** presents research works which perform vision based indoor navigaiton and can guide the blind people. The localisation importance in visual navigation

is further discussed followed by overview of some techniques used for indoor and outdoor scene recognition.

3. **Chapter 4** states all data sets and metrics used in this thesis for performance evaluation.

4. **Chapter 5** proposes a shorter versions of SIFT feature descriptor and compares it with different feature descriptors for general image matching on different data sets.

5. **Chapter 6** proposes a visual Bag of Words (BoW) approach based on the shorter features proposed in Chapter 5. The proposed system is compared with a normal system with different configurations for performance analysis.

6. **Chapter 7** is a comparison of different verification methods which can be used with our proposed BoW presented in Chapter 6.

7. **Chapter 8** presents ways to reduce the features proposed in Chapter 4. The target was to use reduced features with proposed BoW for more robust image matching.

8. **Chapter 9** is evaluation of image localisation using pose estimation and simple 2D image matching with different mobile devices for the analysis.

9. **Chapter 10** presents the framework used by us for the smartphone application. The system performance in real time data sets are reported in this chapter.

10. **Chapter 11** contains the final remarks and suggestions for possible future research work.

1.6 Publications

1. (Khan *et al.* (2011a)) Proceedings of Image and Vision Computing New Zealand. This paper presents a proposed visual BoW based on voting scheme and homography method for a robust image matching.

2. (Khan *et al.* (2011b)) Proceedings of International Conference on Digital Image Computing: Techniques and Applications (Australia). This paper discusses the proposed shorter version of SIFT features and compare it with SURF.

3. (Khan *et al.* (2012b)) Proceedings of International Conference of the NZ Chapter of the ACM's Special Interest Group on Human-Computer Interaction. We discuss the first prototype of our smart phone application and reported the preliminary results.

4. (Khan and McCane (2012)) Proceedings of 14th International ACM SIGACCESS conference on Computers and Accessibility (USA). We present the final prototype of our system and test it on larger images.

5. (Khan *et al.* (2012a)) Proceedings of Image and Vision Computing New Zealand. We propose the method to reduce the features from trained images by more than 50% and compare our reduced features against well known descriptors.

1.6.1 APPSTAR Competition

APPSTAR was a competition run by Otago Innovation Ltd, Otago University in 2012. The competition asked the University of Otagos academic and research staff and students all over the New Zealand to put forward an idea for a mobile or web Application. The competition received an incredible 108 entries and 5 ideas were selected by a panel of Judges. Our application in collaboration with William Levack 'Living it up/iPos" was one of the top 5 finalists in the competition. The basic theme of 'Living it up/iPos" was to help people with memory loss to know exactly where they are in their house and then guide them to perform daily life time activities. For example if someone is in the kitchen then he/she can be prompted with a list of things to do such as make a breakfast, cooking tips etc.

Chapter 2

Navigation tools

This chapter discusses the requirements of a navigation system for a blind person in context of indoor and outdoor navigation. The popular navigation tools in use of blind people are reviewed followed by some research projects which assist the navigation process.

2.1 Navigation system requirements

Blind people need different functionalities from a navigation system. They make choices when it comes to travel and use different navigation tools depending upon their preferences.The navigation tools help the blind people to move safely in an environment. Different navigation tools are designed which address blind people requirements during the navigation. It is important to have a good understanding about these requirements before designing any navigation tool. Therefore, we began our research by conducting interviews with blind people. The interviews took place in Disability Information and Support Center, Otago University[1]. The main purpose of the interview was the requirement analysis i.e. figure out the functionalities, a blind person expects from a navigation system. The requirements of a blind person from a navigation system are summarized as follows (Khan and James (2010)):

- **Pedestrian Crossing:** The navigation system should guide the blind person to remain in the center and to move in the right direction while on the pedestrian crossing. The crossing distance, intersection street name, traffic rules etc information should also be communicated to increase the surroundings perception.

[1]`http://www.otago.ac.nz/disabilities/`

- **Curbs Detection:** The possible curbs or holes in the path should be detected and blind user should be alerted so that precautionary steps can be taken to ensure the safety.
- **Sign boards:** The system should recognize the places or landmarks from sign boards. The location details along with the distance information to that landmark should be communicated to the user.
- **Temporary Hazards:** The temporary hazard signs such as wet floor signs, warning signs on a construction site etc must be detected and corresponding information should be communicated to the blind user.
- **Key destinations:** The blind user should be guided to important city destinations such as hospitals, supermarkets, bus stops, airport etc.
- **Speech accent:** Navigation systems often use voice to communicate with its user but . accent of the speech is sometimes not easy to understand. Therefore, a Braille display should also be available along with the system for further convenience.

  Braille displays are un-electro mechanical devices for displaying braille characters usually by means of raising dots through holes in a flat surface (Wikipedia (2012)). Blind users use it to read text output.
- **Indoor Localisation:** Blind people want to know about their current indoor location in unfamiliar buildings and feel lost in the absence of such information. Therefore, a system with such functionality is required by the blind people.
- **Obstacle avoidance:** The system should detect and alert user about the possible obstacles in the path to avoid the collision.
- **Toilets:** The system should be able to classify the male and female toilets in buildings.
- **Scene description:** Some sort of room description such as power socket location, exit doors information (opening inside/outside), wireless access points etc should be conveyed to the blind user. The location of power sockets is especially very important because blind people need to charge the carrying electronic devices.

- **Stairs:** Blind people prefer lifts and the system should guide the blind users towards the lift. However, use of stairs should also be incorporated to deal with the emergency situations.

- **Guidance in the corridors:** The system should indicate the steps information after which the turns are coming in the corridors of indoor buildings.

Above, we have discussed the desired requirements from a navigation system for blind people. We realized from the interview that *indoor localisation is one of the important requirement for blind people as they often visit unfamiliar buildings and have to rely on the near by people for the guidance.*

It is quite hard to address all requirements in a single research work. However, these problems are addressed separately in different research works. Few vision based research works addressing the mentioned requirements are mentioned below:-

- (Coughlan and Shen (2007)) developed a navigation tool based on vision algorithm to detect the curbs and holes for blind wheelchair users. The disparity map is retrieved from stereo images followed by the generation of an edge map. On edge map, each pixel is checked for an appropriate depth leading to detection of possible curbs or holes and user is alerted.

- (Deville *et al.* (2010)) presented a visual saliency based assistance system to point out areas of interest in a scene that present either particular interest or potential threat . The areas of interest refer to those objects that attract the visual attention of a human being.

  The system makes uses of a stereoscopic camera, a laptop and standard headphones. The user first defines an objective such as find an object, find a door etc. Depending upon the objective, system computes the specific feature maps (colors, orientations, edges etc) and corresponding Conspicuity maps which lead to focuses of attention (FoA). A conspicuity map contains information about regions of an image that differ from their neighborhood. When FoA is seen over number of frames, the user is informed with the global position of object or obstacle through a voice message.

- *See Color* is a navigation tool designed to provide the environment information to its blind users by transforming the colored pixels into musical instrument sounds (Deville *et al.* (2009). See ColOr encodes the colored pixels from frontal images by spatialised musical instrument sounds in order to represent the color and location of visual entities in their environment. The basic idea is to represent a pixel as a directional sound source with depth estimated by stereo-vision. Finally, each emitted sound is assigned to a musical instrument, depending on the color of the pixel. The experiments demonstrated that blind users can perform simple tasks like following a colored line, finding a colored object etc with this approach.

- (Ezaki *et al.* (2004)) presented a system capable of detecting the text from natural scenes to assist the visually impaired people. The system uses PDA, CCD-camera (placed on user shoulder) and the voice synthesizer. The image is captured and system automatically searches for text areas with smaller characters whilst person walks. In case of text area detection, the camera zooms to obtain a more detailed image and characters are recognized. The characters are then read out by the synthesizer to the visually impaired person.

  The morphology and binarization operations are used to detect the text regions from the captured image in the start. After zooming in, methods either based on edge or color are used to extract the larger characters in form of connected components. Finally, certain rules such as size, spacing between the characters etc are used to filter out the wrong text areas and come up with correct ones. The extracted characters then can be read out to visually impaired people.

2.2 Assistive Technology

It include assistive, adaptive and rehabilitative devices for people with disabilities. The assistive technology is provided by the navigation tools and is used by the individuals with disabilities to perform functions that might otherwise be difficult or impossible. The assistive technology has made possible for blind people to get the education and then pursue a career because of the use of computers and other devices. The assistive technology for blind people include (1) programs that run on the computers to speak the text on the screen, and (2) stand alone products designed to serve as a navigation or mobility tool including applications designed for smartphones, personal digital assistants (PDAs). There are two types of navigation tools :

- **Primary tools:** These tools aim to provide a safe navigation and are needed by blind people while navigating in an environment.

- **Secondary tools:** These tools only increase the surroundings perception and must be used with some primary tool.

285 million people in the world are visually impaired and about 14% of them are blind. The majority of the world's visually impaired people (90%) lives in developing countries. The last few years have therefore seen the developments of different navigation tools in developing countries to help the blind people. At any given time, blind person can travel using a human guide, which involves holding onto someone's arm or they use some navigation tools. This section lists some of the popular navigation tools used by the blind people in daily life.

2.2.1 White Cane

It serves as a primary tool and is commonly used by blind people. White canes are less expensive and offer a cheaper solution. Blind people have been using a stick, cane or shepherd's staff as an assistant tool for independent travel since centuries. It was not until last century, the importance of the white cane as a symbol was realized. The white cane was used as a symbol for the first time by James Briggs in 1921 in order to alert passing motorists that he is a blind traveler. After World War II, the number of returning blinded veterans were quite high which further shed light on the significance of white canes. Different types of white canes have been developed so far to address different requirements (Appendix A). With time, the awareness of white cane has increased and now it serves a dual role of both a travel tool and symbol identifying the user as a blind traveler in our society (Strong (Strong); Wikiepdia (2012b)). White cane is really useful for blind people to center themselves while walking in the corridors, detecting the obstacles on the way etc as shown in Figure 2.1. However the tactile information within the reach of the cane is only available. The situations that require route planning in unfamiliar environments will be difficult rather impossible with only white cane.

2.2.2 Guide Dogs

The guide dog is a useful primary navigation tool for blind people. The guide dogs are trained to guide their users around hazards, negotiate traffic, locate common destinations such as the supermarket, post shop, travel on buses etc (Wikiepdia (2012a)). However, guide dogs are not capable of guiding the people in unknown environments

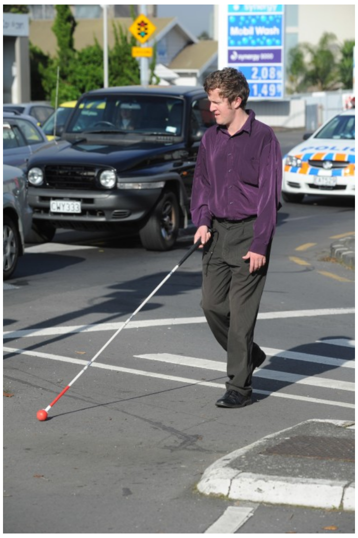

Figure 2.1: Blind person on the pedestrian crossing carrying a white cane. (Source :`http://www.rnzfb.org.nz`)

especially in buildings. The blind user therefore does the directing, based upon skills acquired through previous mobility training.

The first guide dog training school was established in Germany during World War I, to enhance the mobility of returning blinded veterans. Later years have seen the establishments of such schools all over the world such as Australia, New Zealand, England etc. There are about 240 working guide dogs in New Zealand (Webmaster (2012b)).

The main problem with the guide dog is the associated cost which goes beyond US$ 42,000 (Webmaster (2012a)). This cost includes training the dog and providing instructions to the guide dog user. Blind people organizations use donations to prepare guide dogs and provide to blind people for free which limits the availability of guide dogs. Due to this reason, white canes are alternatives for reasons of price and in case of some people with allergies. Today, most blind people still use canes at least sometimes, some prefer guide dogs only and some use both as shown in Figure 2.2.

2.2.3 Miniguide

Miniguide is a secondary travel aid to detect obstacles on the way. It is light weight, hand held and pocket size device as shown in Figure 2.3. The device uses ultrasonic echo-location to detect the obstacles or objects whilst blind person navigates. The aid vibrates to indicate the distance to objects i.e. a faster vibration means that the object is nearby. The device can be adjusted to detect obstacles within different

Figure 2.2: Blind person along with a guide dog and holding a white cane.
(Source :`http://theexistenceofournaturalenvironment.blogspot.co.nz/`)

distance ranges e.g. 1 meter, 2 meter etc. However, only large objects can be detected at 4 meters or beyond, for example, fences, walls.

Figure 2.3: Person holding a miniguide.
(Source :`http://www.gdp-research.com.au`)

Miniguide has assisted blind people in many ways such as avoiding obstacles, detecting overhanging obstacles (tree branches), locating door ways etc. With all these merits, it cannot detect drop offs and the cost of miniguide is up to US$ 400 which make sit hard for blind people to afford it (SenderoGroup (2012)).

2.2.4 Soundpost

Soundpost is a primary navigation tool. It is basically an orientation device and uses infrared technology to guide the blind person to move accurately from one landmark to the other (Povidi (2008)). It allows a blind person to cross up to 30 meters of an open space. Soundpost uses multiple base stations for generating signals and blind person carries a hand held transmitter to receive the signals as shown in Figure 2.4. The transmitter starts vibrating once it comes within range of a base station and keeps on vibrating until you move towards the landmark. The transmitter can also generate a voice message indicating the landmark information.

A small independent trial is undertaken at the University of Canterbury, New

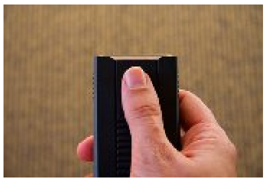

Figure 2.4: Soundpost device.
(Source :http://www.povidi.com/SoundPost.html)

Zealand to assess its utility for a blind citizen. It has been preferred by both guide dogs and white cane users as it further assists the navigation in any environment. However it suffers from following problems which limit its applicability:-

- The infrared signal is often broken which pose problem during the navigation.
- It's hard to estimate the movement direction towards the landmark once the vibration starts.
- The voice messages are limited.
- It costs about US$500 which includes the transmitter and two base stations.

2.2.5 Loadstone GPS

Loadstone GPS is a free open source software for satellite navigation for blind users (LoadstoneGPS (2010)). The software currently runs on different Nokia devices and requires a GPS receiver. This project was initiated by Monty Lilburn and Shawn Kirkpatrick back in 2004 who are themselves blind. The whole program is under the General Public License (GPL) and is financed entirely by the private developers and by the donations. It was made public in 2006 and has been a success since then.

The maps consist of way points imported from the common map data bases. In some cities, rural regions etc no exact map data is available in common map databases. In such scenarios, the software provides users an option to create and store their own way points for navigation and share it with others. There is also a website which allows users to share there custom points with each other and points can then be easily added to the map.

This application is useful to search for specific locations in a given area. However, it lacks some of the basic features such as automatic download of maps and a route planner. Still tool is very accessible and useful when travelling in an area.

2.2.6 Research Projects

In the above sections, we discussed some commercial products which provide mobility aid to the blind people. A lot of research has been carried out in the last few years to propose the reliable navigation tools. Some of the well known research projects which use computer vision and other technologies to aid the blind user mobility are discussed in the following sections.

Drishti

Drishti is a navigation system capable of guiding and helping the blind people to travel safely both in indoor and outdoor environments (Ran *et al.* (2004)). It uses a precise position measurement system, a wireless connection, a wearable computer, and a vocal communication interface. For outdoor, it uses differential GPS as its location system to keep the users as close as possible to the central line of sidewalks of campus and downtown areas. The user can switch the system mode from an outdoor to an indoor environment with a simple vocal command. An ultrasound positioning system is used to provide precise indoor location measurements with an accuracy of 22 cm. The user gets alerts in the form of vocal prompts to avoid the possible obstacles. It performs dynamic routing and rerouting to provide its blind users with an optimal route for navigation. A step-by-step walking guidance is provided for navigation in environments. The proposed system is good but it is not easy to use as blind person have to carry a number of devices.

Crosswatch

(Ivanchenko *et al.* (2008)) presented Crosswatch, a real time mobile application to detect the zebra crossings in outdoor environments. At the proposed time, this was the first portable system capable of providing the real time orientation information at urban traffic intersections. The system first extracts the straight line segments as features from the mobile image. A factor graph model is then used to group the features into figure and ground, representing crosswalk and background respectively. The detection of enough features having sufficient length indicates the detection of a cross walk and a brief audio tone is sounded for frames in which a crosswalk is detected. Nokia N95 mobile phone is used in experiments which can process approximately three frames per second. The application is tested with blind subjects to test the usability and the subjects answered correctly whether a zebra crosswalk is present or absent at all used intersections.

(Ivanchenko *et al.* (2010)) further extended Crosswatch to incorporate the detection of walk lights on the pedestrian crossings. Crosswatch is now capable to alert the user once the green walk light is illuminated at the intersections. Corsswatch is robust, easy to use and performs well in urban intersections in the United States.

Electro Neural Vision System (ENVS)

(Meers and Ward (2005)) presented ENVS, a system capable to avoid obstacles, perceive landmarks and assist in navigation via visual sensors, GPS and electro-tactile stimulation. ENVS first extracts the depth information from the environment via disparity map obtained from the head mounted stereo cameras as shown in Figure 2.5. This range data is then delivered to the fingers via electro-neural stimulation to indicate the range of objects being viewed by the cameras to the user. To perceive the location of obstacles and the 3D structure of the environment, user imagines that the hands are held in the direction viewed by the cameras, with fingers extended and the amount of stimulation felt by each finger indicates the range of objects in the pointed direction.

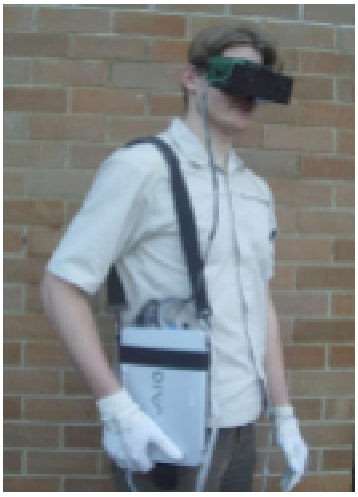

Figure 2.5: The Electro- Neural Vision System.
Source: (Meers and Ward (2005))

In order to perceive landmarks, ENVS uses a GPS, a digital compass and a database of landmarks. The relative location of significant landmarks is determined (using GPS and stored GPS coordinates) and delivered to the fingers via encoded pulses when the landmarks are in the field of view. The system does quite well in conveying the 3D structure of the environment to the blind people. The experimental results indicate that ENVS enables the user to achieve localization, obstacle avoidance and navigation without using the eyes. However, it is not easy to use because blind user needs to carry a laptop and wear gloves for ENVS.

2.3 Summary

The purpose of this chapter was to provide a overview of navigation system requirements and highlight the importance of indoor localisation for blind people. Some popular navigation tools along with the research projects which aid the navigation process are discussed. In the next chapter, we discuss the navigation systems which mainly use computer vision techniques such as features, topological maps, scene localisation to provide navigation in indoor environments. Research works addressing the scene localisation problem in indoor and outdoor environments are presented towards the end in the next chapter.

Chapter 3

Image based localisation

The previous chapter mentioned some navigation tools which assist the navigation. This chapter moves to a next level and reviews works which offer visual navigation for robotics or visually impaired people. The localisation based on different technologies is reviewed to understand the importance of vision based localisation. Finally, we present some good research works which use robust computer vision techniques to perform scene matching in indoor and outdoor environments.

3.1 Visual Navigation

The navigation based on vision (images or videos) is referred as "walk through problem". The vision sensor seems to be the best choice for navigation if we consider its nice features like low price, low power, non contact and high potential information contents. However, its really hard to extract useful and reliable information from the image feeds to base mobility decisions. The visual navigation focuses on guidance with the goal of automatically reproducing the tasks performed by humans without any visual impairment. The important tasks are to detect the current location, plan the route and then guide the blind person safely while avoiding the possible obstacles on the way. Francisco *et al.* provides a very good survey on visual navigation for mobile robots and mention key research contributions which are made from nineties till 2008 (Bonin-Font *et al.* (2008)). The visual navigation systems are divided into two main categories:

3.1.1 Map based systems

Map based systems need some representation of the environment built before the start of navigation process. There are two systems mainly (1) systems which use an existing map of the environment usually referred as *map-using*, and (2) systems that build the map of the environment themselves usually referred as *mup-building*. The navigation starts only after the environment representation or a map is available.

For a true autonomous visual navigation, the system needs to do automatic exploration, localisation and mapping of the environment via techniques known as Simultaneous localization and mapping (SLAM). SLAM basically involves the simultaneous estimation of a map and the robot pose concurrently. Recently, cameras have been used as the sole sensor to yield SLAM based on vision i.e. visual SLAM (Chen *et al.* (2007); Davison *et al.* (2007); Silveira *et al.* (2008)). A standard scheme to visual SLAM consists of first extracting a sufficiently large set of features and robustly matching them between successive frames. These corresponding features are used for estimating the camera pose and scene structure. In this section, we discuss some well known SLAM systems which make use of a map with a particular focus on indoor buildings.

Davison presented MonoSLAM which is based on monocular vision and is one of the most successful schemes for a single camera (Davison (2003)). Landmark positions and camera pose are estimated and refined directly from observations in the image using an Extended Kalman Filter (EKF). The camera can be positioned accurately for extended periods provided that sufficient landmarks are tracked. However, the positioning is limited to small scale environments with 100 landmarks in total. Clement *et al.* (Clemente *et al.* (2007)) extended MonoSLAM to deal with larger environments by using the submaps framework proposed by (Estrada *et al.* (2005)). MonoSlam is used on small scale and new map is started once the number of landmarks grows beyond some limit. The local submaps are combined into an accurate global map by opimising transformations between submaps. The system is able to close 250m loop with an accurate global map.

Sim *et al.* presented a system which maps a large and complex environment in real time using a pair of stereo cameras (Sim and Little (2006)). The main contribution of the work is a fully automatic mapping system which operates on-line and consistently produces accurate maps of large scale environments. The system uses a hybrid approach consisting of 3D landmark extraction based on SIFT features (Lowe. (2004)) for image localisation and occupancy grid to obtain a map for a safe navigation. The occupancy grid represents the environment as an evenly spaced field of binary random variables each representing the presence of an obstacle at that location in the environment. The occupancy grid offers a reliable spatial representation of the world for a safe navigation. The proposed system is deployed on a robot and it does well in a laboratory environment with of two rooms during the testing.

Topological maps are graph based representation of the environment and are parsimonious representation because they are simple and easy to scale. Winters *et al.* (Winters and Santos-Victor (1999) proposed a system for robot which uses omnidirectional camera images to generate a topological map of the indoor structured environment during a training phase. The links in the graph are sequence of images belonging to a place (e.g. corridor) and nodes represent specific places associated with actions such as turning, crossing a door etc. The image set is represented by a low dimensional eigenspace and robot determines its position by projecting the current image into eigenspace during movement. After position estimation, the robot uses ground plane images to extract corridor guidelines to control its trajectory. The system does well in corridors of Institute for Systems and Robotics, Portugal (ISR) [1].

Kidono *at al.* developed a system which needs training from a human to build a map (Kidono *et al.* (2002)). As user guides a mobile robot to a destination by remote control, the robot constructs a 3D model on line incrementally frame by frame from stereo images. For autonomous movement, the robot utilizes the map and past experience on observation to safely navigate from the starting point to the goal point. The robot is navigated through an indoor room with different obstacles and did well.

Thrun *et al.* developed a museum guiding robot MINERVA[2] that uses laser scans, camera images and odometry readings to to build map of the environment for the navigation (Thrun *et al.* (1999)). The robot is first navigated via joystick in the museum to generate occupancy and ceiling texture maps. The occupancy grid ensure safe navigation and texture maps are used for localisation in case of a huge crowd because frontal images do not offer sufficient information for feature based tracking. The robot remained operational for two weeks at Smithsonians National Museum of American History and successfully interacted with thousands of people. Shen *et al.* proposed ATLAS, a more advanced museum guiding robot (**?**). The laser data and images are used to match the map data of the environment for localisation. A visual appearance based algorithm is than used for normal topological navigation. In visual appearance algorithms, the current image is compared with stored templates or models

[1] `http://welcome.isr.ist.utl.pt/home/`

[2] `http://www.cs.cmu.edu/~minerva/`

for the recognition. ATLAS is tested in the foyer of London County Hall for 03 months. The robot can (1) detect the nearby visitors and interact with them via voice and a touching screen, (2) uses a human face detection to approach to new visitors and (3) can detect the battery level and charge itself automatically. This makes ATLAS possible to run continuously in a museum environment.

Some navigation systems support on line construction of a local occupancy grid which is the portion of the environment surrounding the robot. This local information is used for a subsequent map construction frame by frame for on line safe navigation. Gartshore *et al.* presented a framework for map building based on online images captured by a single camera (Gartshore *et al.* (2002)). The main contribution of the work is combination of occupancy grid and visual data from a single camera. The system uses feature detector algorithm to identify the object boundaries based on features in occupancy grids from the current frame. While positioning module computes robot position using odometry data and detected features. The color or gradient from features from past frames are used to further verify the object presence at a specific location. The experiments were conducted in indoor environment and robot was able to navigate safely while avoiding the obstacles.

Botterril *et al.* presented a BoWSLAM scheme which enables real-time navigation of robots in dynamic environments using a single camera alone (Botteril *et al.* (2010)). The robot can position itself in real-time while exploring a previously unknown environment. The system works by representing every frame as a Bag of Words (BoW). BoW representation is then used to select multiple nearby frames to compute relative positions and latest frame is positioned relative to each of these frames. A subset of these multiple position hypotheses with minimum gross errors is selected to accurately position the robot in a global map. The proposed system allows robot navigation in challenging dynamic and self-similar environments.

3.1.2 Mapless systems

These systems refer to those techniques which do not need any knowledge of the environment but they navigate as they perceive the environment. These techniques often grab video frames to produce enough information about the unknown environment for a safe navigation.

Optical flow is one of the technique commonly used for mapless systems. Taludker *et al.* presented a novel optical flow based solution for robots to detect the presence of dynamic objects in the camera field of view during navigation (Talukder and Matthies

(2004)). The algorithm assumes that moving objects cause discontinuity in optical flow orientation and changes its magnitude with respect to the background pixels orientation and magnitude. The system is initially tested with single camera and later on with stereo cameras to extract the depth information.

Green *et al.* presented a Closed Quarter Aerial Robot (CQAR) whose navigation is controlled by an insect inspired optical flow based system (Green *et al.* (2003)) . The robot can fly into buildings, take off and land based on optical flow computed from the images of the environment. The minimum flying speed of CQAR is 2 m/s and it needs to turn 5 meters before to avoid an obstacle.

Zhou *et al.* used appearance based strategy for mapless navigation (Zhou *et al.* (2003)). Such strategies has two phases (1) first in a pre-training phase, the images or prominent features from the environment are observed and stored as model templates. These models are labeled with a certain location information, and (2) in the navigation stage, the robot matches the current image with the stored templates to recognize the environment and self-localize. Zhou *et al.* utilized color, gradient, edge density and texture histograms to describe the appearance of pre-recorded indoor images. The navigation is performed by matching the multidimensional histogram of current image with the stored templates. The use of histograms save computation resources and are simple to use. The main problems with appearance based strategies are to identify the ways to represent the environment and perform on-line matching.

The image qualitative characteristics and their interpretation are often used by visual techniques to navigate and avoid obstacles. The navigation systems based on qualitative information avoid as much as possible the obstacles in the environment. Fasola *et al.* proposed to use image color segmentation techniques to detect the possible objects from frames followed by classification of opponent robots via gray scale umage processing (Fasola and Veloso (2006)). The integral images which were introduced for real time face detection have been used for a robust robot classification (Viola and Jones (2001)). The proposed system is developed with the annual RoboCup Competition in mind where teams of Sony AIBO 4-Legged robots compete in the game of soccer. From a set of 327 test images, the system was able to achieve a 97% classification accuracy yielding only one false positive.

The techniques for tracking moving elements in a video sequence have become robust enough and useful for navigation. Such techniques divide a tracking task into two sub-tasks (1) motion detection, which refers to the identification of most likely region in next frame to find the feature, and (2) feature matching, by which the feature tracked is identified within the determined region (Trucco and Plakas (2006)). Such techniques do not handle the obstacle avoidance. SIFT features (Lowe. (2004)) are the popular features for the navigation so far. During the navigation, the SIFT features are observed from different viewpoints, angles, distances and with different illumination changes. The detected features serve as appropriate landmarks to be trace over time for navigation, global localisation and a robust vision based SLAM performance (Se *et al.* (2001, 2005)).

Saeedi used stereo vision in a novel navigation strategy applicable to unstructured indoor/outdoor environments (Saeedi *et al.* (2006)). The main emphasis of this work is to estimate the robot motion independently from any prior scene or landmark knowledge. The system uses a new, faster and more robust corner detector and detected features are 3D positioned. The 3D positions of scene features and the robot are refined by a Kalman filtering over time. The results indicate a good tracking and localization performance covering outdoor environment of 5 cm by the system.

3.2 Localisation

In the previous section, we have reviewed some key contributions in area of visual navigation for robotics. However, same techniques can be utilized for blind people by replacing the control signals with appropriate voice messages.

However, the focus of our work is the scene localisation in indoor environment only and not the navigation. Technologies other than vision have been used for indoor positioning, such as infrared, ultrasonic etc. It is important to briefly review these techniques to understand the corresponding limitations to identify the benefits of vision based localisation. In this section, we briefly discuss indoor positioning based on different techniques for the analysis (Gu *et al.* (2009); Song *et al.* (2011)).

3.2.1 Infrared

Infrared (IR) positioning technology needs infra-red emitters and measures the position of the object according to the receiving time of infrared signals. Want *et al.* developed Active Badge system, one of the first successful IR based indoor positioning system at AT&T Cambridge in 1990s (Harter *et al.* (2002); System (2008); Want *et al.* (1992b)). The person needs to carry an active badge and system provides room level indoor localisation with active badges. An active badge transmits a globally unique IR signal every 15 seconds. In each located place, one or more sensors are fixed and detect the IR signal sent by an active badge. The position of an active badge can be specified by the information from these sensors and forwards the location information of the tracked active badges to a central server. The price of active badges and networked sensors are cheap but the cables connecting the sensors raise the cost of the Active Badge system.

OPTOTRAK system is designed for positioning in congested shops and workspaces Digital (2008). The OPTOTRAK uses a system of three cameras as a linear array to track 3- D positions of numerous markers on an object. The markers mounted on different parts of a tracked object emits IR light which is detected by the camera of the system to estimate the location of them via triangulation technique. The system offers a high accuracy of 0.1 mm to 0.5 mm with 95% success probability. A disadvantage of OPTOTAK system is the line-of-sight requirement between the objects and the tracking system. However this can be partially solved by using a large number of IR markers at the expense of cost.

IR based systems provide a precise position estimation However, IR signals have some limitations for sensing location, for example, interference from florescent light and sunlight. This problem can be solved by using optical and electronic filters but it raises the cost of the positioning system. Although, the IR emitters are cheap but whole system using camera array, transmitters, IR device and wire connectivity for each isolated place increases the cost of system. connected via wires is expensive when used in large areas. IR device taken by a person is covered by his/her clothes, the system fails to work since the IR wave can not penetrate opaque materials.

3.2.2 Ultrasonic

Using ultrasound signal is another way to measure the position. Ultrasound signals are used by bats to navigate in the night which inspire people to design a similar navigating system. Ultrasonic positioning technology uses mainly reflective distance method to determine the location of the object.

Active Bat system is a well known ultrasonic indoor positioning system researched by AT&T Labs (Ward *et al.* (2004)). It is a low-power, wireless indoor location system accurate up to 3 cm. It relies on multiple ultrasonic receivers embedded in the ceiling and measures time-of-flight to them. It works by finding the distance to minimum of three reference nodes and then using multilateration technique to find the exact position. The performance of the system is influenced by the reflection and obstacles between tags and receivers, which degrades the system accuracy. The scalability of the system is affected due to deployment of a large number of sensors on the ceiling.

The receivers also need to be accurately placed, which results in complex and costly installation.

Cricket system is another location system which offers efficient performance and low cost (Priyantha (2005); Priyantha *et al.* (2000)). The cricket system uses time of arrival measuring method and triangulation location technique to locate a target. The system includes ultrasound emitters as infrastructure attached on the walls or ceilings at known positions, and a receiver mounted on each object to be located. The cost of the whole system is low and it can provide a position estimation accuracy of 10 cm. The receivers in the cricket system consume more power and its power supply needs to be efficiently designed to avoid frequent charging. The Sonitor ultrasound IPS is an indoor tracking and positioning solution provided by Sonitor Technologies Inc (Website (2008)). The Sonitor system can locate and track people and devices in real-time and offer very good room level accuracy.

Usually the ultrasound signals used to locate objects need to be combined with RF signals, which perform synchronization and coordination in the system. These ultrasound positioning systems increase the system coverage area. However, Ultrasonic positioning system requires large-scale layout with many of receiver hardware leading to a overall higher cost. The positioning result is very sensitive with the environment and suffers from reflected signals, metal objects etc.

3.3 RFID

Radio frequency (RF) technologies use frequency signals which can travel through walls and human bodies easier therefore offering a larger coverage area and less requirement of hardware. The RFID positioning systems are commonly used in complex indoor environments and offer cheap identification.

LANDMARCE positioning system is a RFID-based indoor positioning system developed by Michigan State University and the Hong Kong University of Science and Technology (Ni *et al.* (2004)). The system uses fixed position reference tags and and RFID readers to determine the nearest reference tag from the target tag followed by locaiton estimation. Jin *at al.* further improved LANDMARC with respect to system's energy consumption and costs (Jin *et al.* (2006)).

WhereNet positioning system offers indoor and outdoor real-time positioning (Wbesite (2008)). RFID technology is used to identify various located tags mounted on the target located objects, such as a device or a person. The system uses differential time of arrival algorithm to calculate the precise locations of these tags. The tags are powered by batteries which can last up to 7 years depending on the transmission rate of the tags. However, the WhereNet offers an error range around 2 m to 3 m, which is not very accurate in indoor situations. The system is complex and the installation of these devices is time consuming.

The advantage associated with RFID positioning system is light and small tags that can be taken by people to be tracked. The RFID system can uniquely identify equipment and persons tracked in the system. However, the absolute positioning techniques need numerous infrastructure components installed and maintained in the working area of an RFID positioning system which increases the overall cost.

3.3.1 WLAN

WLAN technology is very popular and is widely used in business districts, universities, airports etc. WLAN-based positioning systems reuse the existing WLAN infrastructures in indoor environments, which lower the cost of indoor positioning. In case of unavailability of WLAN infrastructure, the associated cost increases. This positioning technology uses the WLAN client such as laptop, PDA, phone etc to get the received signal strength (RSS) or signal to noise ratio (SNR) from wireless network interface card.

RADAR positioning system was proposed by a Microsoft research group as an indoor position tracking system (Bahl and Padmanabhan (2000)). It uses signal strength and SNR with the triangulation location technique. The RADAR system can provide 2-D absolute position information and thereby enable location-based applications for users. The major advantages of RADAR system are that the use of existing indoor WLAN infrastructures and requirement of few base stations to perform location sensing. However, the limitation is that the located object needs to be equipped with WLAN technology which is difficult for some lightweight and energy-limited devices. There is also no consideration of privacy issues where a person using a device with WLAN interface may be tracked, even he/she does not want any one know his/her location. In addition, the RADAR system suffers from the limitations of RSS positioning methodology.

The COMPASS system uses WLAN infrastructures and digital compasses to provide low cost and high accurate positioning services to locate a user carrying a WLAN-enabled device (King *et al.* (2006)). The COMPASS system uses fingerprinting location technique and a probabilistic positioning algorithm to determine the location of a user.

During position estimation, the users orientation is measured by a digital compass to reduce the human body blocking influence. For the tracking of a mobile user, the orientation impact is highly addressed by the designers of both RADAR and COMPASS system. As human body contain more than 50% water, which absorbs the 2.4 GHz radio signal, the clocking effect of human body influences the measurement accuracy. The COMPASS system achieves an accuracy of about 1.65 m compared to RADAR system which has an error distance of 2.26 m in the same indoor place. However, the COMPASS system only considers tracking a single user which reduces its scalability.

WLAN systems provide very cheaper indoor position estimation due to re-usability of existing infrastructures in indoor environments. WLAN technology is widely used and integrated in various wireless devices such as PDAs, laptops, mobile phones, etc. However, because of complex indoor environments consisting of various influenced sources, the performance of the positioning systems are not very accurate with an accuracy of several meters. And using the stored information and fingerprinting technique in the location estimations is complex and costly if the number of users of the positioning system is increasing significantly. The fingerprinting requires a costly and time consuming signal strength calibration at the kick off. Some recent solutions do not require calibration and an accuracy of 2-7 m is achievable which is good (Chintalapudi *et al.* (2010)).

3.4 Audible Sound

Audible sound is a possible technology for indoor positioning as almost every mobile device has that ability. Beep was designed as a cheap positioning solution based on audible sound technology (Lopes *et al.* (2006)). Triangulation location technique is used in Beep with a standard 3-D multilateration algorithm based on time of arrival measured by the sensors in Beep system. The 3-D location information determined by Beep can be used by various practical applications. The system gives an accuracy of 0.4 m with 90% . In addition, the effect of sound noise and obstacles reduce the positioning accuracy by 6-10%. One of the benefits brought by the Beep system is that the privacy of the users is considered by avoiding them being tracked automatically. The users can stop their devices from sending audible sound; if they do not want the system knows their location.

Audible sound is an available service in various mobile devices used in our daily lives. Thus the users can use their personal devices in an audible sound positioning system to get their positions. Because of properties of audible sound, using it for indoor positioning has some limitations. The audible sound can be interfered by the sound noises in the dynamic changing and public indoor situations. Audible sound does not have high penetration ability, so the scope of an infrastructure component is within a single room. The associated cost will increase with more rooms. Transmitting audible sound is a kind of noise to indoor environments, where people would not like to hear audible sound made by the positioning services.

3.5 Vision

The main merit of vision basd system is a low price camera can covera large area. However, these systems have some drawbacks. Firstly, the privacy of people is not provided by the vision-based positioning. Secondly, the system is not reliable in a dynamic changing environment. However, the problem is often solved by using an incremental approahc in which data base is updated with new images. The visionbased positioning is influenced by light such as turning on and off a light. Our system targets to provide guidance to blind person who often visits the building during day time and can take multiple pictures to localisae the place.

3.6 Image based Localisation

Our work is based on scene localisation using computer vision which deals with the images. The basic idea is to match the current image with stored images and get the corresponding location information. Image based indoor location recognition is an active research problem and has been widely studied by computer vision researchers. In this section, we discuss different good techniques which offer a robust scene recognition in different environments.

The visual Bag of Words (BoW) approach offers robust image matching and is used in many research works for location recognition. In Sivic and Zisserman (Sivic and Zisserman (2003)), the idea of visual BoW was introduced for the first time for object retrieval. Given a query object, the system is able to search and localize all the occurrences of that object in a video. The inverse document frequency (*idf*) weighting scheme was used for image retrieval.

Nister and Stewenius (Nister and Stewenius (2006)) used a scheme based on visual BoW scalable to large image databases is presented. Their scheme is based on hierarchical clustering and uses term frequency-inverse document frequency ($tfidf$) weighting, which is shown to be very effective for retrieval in large scale data sets. The system is able to match an image robustly in about one second.

Tom *et al.* presented a real-time approach for the detection of identical scenes (Botterill *et al.* (2008)). The work is based on visual BoW which uses the SURF features and hue information for a robust image matching. The system matches a query image from a standard data set of 10200 images in about 0.036 seconds.

Filliat (Filliat (2007)) used a two stage voting scheme for indoor matching. SIFT, hue and texture features are used for visual BoW and a room is recognized only if a quality threshold is reached. The work is only tested for a small scale indoor environment but it has not been shown to work in office buildings which have similar color/texture schemes in many places.

In Kang *et al.* (Kang *et al.* (2009)), visual BoW is used to perform matching on a large scale indoor office environment. The top eight images most similar to the query image are retrieved via visual BoW. A potential localization is suggested if there are a cluster of pre-recorded images less than 3 meters from each other among the retrieved images. The proposed system performs well in an office like indoor building.

Robertson and Cipolla (Robertson and Cipolla (2004)) used homography and rectification for scene recognition in an outdoor environment. In their work, camera's are assumed calibrated (or at least approximately so), and database images are assumed rectified. Features are identified using the Harris corner detector and a RANSAC based algorithm for image registration is applied. The query image is matched against each database image and the closest match is returned as the location.

In Zhang *et al.* Zhang and Kosecka (2006), the image localisation is performed in the urban environment. The system computes the GPS location of the novel query image from a database of city street scenes tagged with GPS locations. The two closest views relative to the query image are first retrieved from the database based on SIFT Lowe. (2004) feature matching followed by the motion estimation. The location of the query image is finally estimated by the triangulation of GPS positions of the closest views.

Skrypnyk *et al.* Skrypnyk and Lowe (2004) presented one of the first systems to estimate the pose of the query image with respect to the 3D model. A sparse 3D model from objects of interest is reconstructed using multi view geometry and the associated SIFT descriptors are stored in a kd-tree. A robust algorithm is then used to estimate the pose of a query object with respect to the 3D models via 2D-3D correspondences.

Sattler *et al.* Sattler *et al.* (2011) presented a direct 2D-3D matching framework based on visual vocabulary for image matching. A novel approach is proposed to find 2D-3D correspondences efficiently followed by the pose estimation. The system performs well on city scale large data sets.

In Irschara *et al.* Irschara *et al.* (2009), a fast location recognition technique based on sfm is presented. A minimal set of "synthetic" images which represent the 3D models of places is derived first to compress the 3D data. Given a query image, the system then uses the vocabulary tree to retrieve the similar images from this compressed data based on the pose estimation.

Most research works focus on direct matching of 2D features (query) to 3D points (3D models) for localisation. However Li *et al.* Li *et al.* (2010), proposed an opposite approach based on 3D-2D matching. The work uses a priority scheme to effectively compute the 2D-3D correspondences followed by the pose estimation. The proposed scheme works quite fast on large scale data sets compared to the direct matching which is slow.

Arth *et al.* Arth *et al.* (2009) used an efficient method to localise the image on the mobile phone based on 3D models. The 3D models of the places are reconstructed and part of it is loaded in the phone memory. The system extracts the features from the camera query image, loads the relevant part of the feature data base in the phone memory and performs the pose estimation to match the image. The experiments are conducted in indoor rooms.

Mulloni *et al* Mulloni *et al.* (2009) proposed an indoor positioning system where the smart phone application sends the query photo to a server. The server uses SURF features **?**, distance estimation and calibration to estimate the indoor position. The system shows good performance on a long corridor.

Hile *et al* Hile and Borriello (2007) presented a smart phone application which sends the query image along with WIFI fingerprints to a server. The server computes feature correspondences with the floor plan to estimate the pose and returns the position information to the user. The system does well with hallway images.

Kawaji *et al.* Kawaji *et al.* (2010) proposed an image based indoor localization system using omnidirectional panoramic images. The system uses SIFT features Lowe. (2004) to match the query image with a database of mapped images by using nearest neighbor search based on local sensitive hashing. The system is shown to work in a Railway museum.

Ruf *et al.* Ruf *et al.* (2010) integrated SIFT and SURF for image matching based on nearest neighbor search leading to precise position estimation inside a museum.

However the proposed system produced acceptable accuracy at the expense of high computational cost.

More recently, object detection and probabilistic semantics have been used in a small scale indoor environment to identify the place type Espinace *et al.* (2010). Their work should be applicable to indoor environments of any type/scale but will perform slower as objects segmentation, objects classification and then use of semantics is usually slow.

The idea of visual BoW is used by different authors for robust object recognition in large scale data sets. Phibin *et al.* used spatial information to verify the consistency of retrieved images to re-rank the retrieved results (Philbin *et al.* (2007)). The initially returned result list is re-ranked by estimating afne homographies between the query image and each of the top-ranking results from the initial query. The score used in re-ranking is computed from the number of verified inliers for each result. Philbin *et al.* later extended his work by exploring techniques to map visual region to a weighted set of words to include features which get lost in the quantization stage (Philbin *et al.* (2008)). The proposed method of visual assignment is found to perform better than state of the art techniques on different large data sets.

Aly *et al.* analysed the performance of visual BoW with two leading methods i.e. inverted file and mini-hash on four diverse real world large data sets (Aly *et al.* (2011)). The effect of the different parameters of these methods on the recognition performance and the run time are analysed in the paper. The work basically targets object recognition.

Lazebnik *at al.* presented the idea of a spatial pyramid with visual BoW (Lazebnik *et al.* (2006)). The spatial pyramid is a simple and computationally efficient extension of an orderless bag-of-features image representation to get improved performance in scene categorization tasks. The proposed method works by repeatedly subdividing an image and computing histograms of image features over the resulting subregions. The system does well on standard datasets for image matching.

Ji *et al.* proposed a framework to perform landmark search via smart phone efficiently without the need to send query photo to the server (Ji *et al.* (2012)). When user starts the application, the GPS information is used to download the generalized visual vocabulary codebook on the phone. When user selects a query photo, local features are extracted from the phone followed by the generation of a compact Location Discriminative Vocabulary Code (LDVC) compact descriptor from the generalized vocabulary. The information normally in 10-50 bits is sent to the server for matching and results are returned to the phone. The proposed approach results in lower transmission rates and provide very good performance. The descriptor is sent to the server,

3.7 Conclusion

In this chapter, we have discussed some key works in area of visual navigation. This was followed by comparison of different localization techniques. Finally, we mention different research works which have used visual BoW to perform image matching in indoor and outdoor environments. We also mention the limitations of these works to highlight the need of an indoor localisation system based on vision.

Chapter 4

Conclusion

I will state the conclusion of my research work.

4.1 Future Work

I will mention the future work in this section.

References

Agrawal, M. and Konolige, K. (2006). Real-time Localization in Outdoor EnFilli-avironments using Stereo Vision and Inexpensive GPS. In *Proceedings of the 18th International Conference on Pattern Recognition - Volume 03*, ICPR06, 1063–1068.

Aly, M., Munich, M. E., and Perona, P. (2011). Bag of Words for Large Scale Object Recognition - Properties and Benchmark. In *International Conference on Computer Vision Theory and Applications*, 299–306.

Angeli, A., Filliat, D., Doncieux, S., and Meyer, J.-A. (2008). Fast and Incremental Method for Loop-Closure Detection Using Bags of Visual Words. Volume 24, 1027–1037.

Arth, C., Wagner, D., Klopschitz, M., Irschara, A., and Schmalstieg, D. (2009). Wide area localization on mobile phones. In *8th IEEE International Symposium on Mixed and Augmented Reality*, ISMAR '09, 73–82.

Bahl, P. and Padmanabhan, V. N. (2000). RADAR: An In-Building RF-Based User Location and Tracking System. In *INFOCOM*, 775–784.

Bonin-Font, F., Ortiz, A., and Oliver, G. (2008). Visual Navigation for Mobile Robots: A Survey. Volume 53, 263–296.

Bonnifait, P., Jabbour, M., and Cherfaoui, V. (2008). Autonomous Navigation in Urban Areas using GIS-Managed Information. Volume 6, 84–103.

Botteril, T., Mills, S., and Green, R. (2010). Bag-of-Words-driven Single Camera Simultaneous Localisation and Mapping.

Botterill, T., Mills, S., and Green, R. (2008). Speeded-up Bag-of-Words algorithm for robot localisation through scene recognition. In *Proc. Image and Vision Computing New Zealand*.

Chen, Z., Samarabandu, J., and Rodrigo, R. (2007). Recent advances in simultaneous localization and map-building using computer vision. Volume 21, 233–265.

Chintalapudi, K., Padmanabha Iyer, A., and Padmanabhan, V. N. (2010). Indoor localization without the pain. In *Proceedings of the sixteenth annual international conference on Mobile computing and networking*, MobiCom '10, 173–184.

Clemente, L., Davison, A., Reid, I., Neira, J., and Tardós, J. D. (2007). Mapping Large Loops with a Single Hand-Held Camera. In *Proceedings Robotics: Science and Systems Conference*.

Coughlan, J. and Shen, H. (2007). Terrain Analysis for Blind Wheelchair Users: Computer Vision Algorithms for Finding Curbs and other Negative Obstacles. In *Proceedings of the Conference and Workshop on Assistive Technologies for People with Vision and Hearing Impairments: Assistive Technology for All Ages*, Volume 415 of *CEUR Workshop Proceedings*.

Davison, A. J. (2003). Real-Time Simultaneous Localisation and Mapping with a Single Camera. In *Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2*, ICCV '03.

Davison, A. J., Reid, I. D., Molton, N. D., and Stasse, O. (2007). MonoSLAM: Real-Time Single Camera SLAM. Volume 29, 1052–1067.

Deville, B., Bologna, G., and Pun, T. (2010). Detecting objects and obstacles for visually impaired individuals using visual saliency. In *Proceedings of the 12th international ACM SIGACCESS conference on Computers and accessibility*, ASSETS '10, 253–254.

Deville, B., Bologna, G., Vinckenbosch, M., and Pun, T. (2009). See ColOr: Seeing Colours with an Orchestra. In *Human Machine Interaction*, 251–279.

Digital, N. (2008). Optotrak.

Espinace, P., Kollar, T., Soto, A., and Roy, N. (2010). Indoor Scene Recognition Through Object Detection. In *Proc. IEEE International Conference on Robotics and Automation*.

Estrada, C., Neira, J., and Tardos, J. D. (2005). Hierarchical SLAM: Real-Time Accurate Mapping of Large Environments. Volume 21, 588–596.

Ezaki, N., Bulacu, M., and Schomaker, L. (2004). Text Detection from Natural Scene Images: Towards a System for Visually Impaired Persons. In *17th International Conference on the Pattern Recognition*, ICPR ’04, 683–686. IEEE Computer Society.

Fasola, J. and Veloso, M. M. (2006). Real-time Object Detection using Segmented and Grayscale Images. In *International Conference on Robotics and Automation*, 4088–4093.

Filliat, D. (2007). A visual bag of words method for interactive qualitative localization and mapping. In *International Conference on Intelligent Robotics and Systems.*, 3921–3926.

Gartshore, R., Aguado, A., and Galambos, C. (2002). Incremental map building using an occupancy grid for an autonomous monocular robot. In *ICARCV*, 613–618. IEEE.

Gemeiner, P., Davison, A., and Vincze, M. (2008). Improving Localization Robustness in Monocular SLAM Using a High-Speed Camera. In *Proceedings of Robotics: Science and Systems IV*, 183–188.

Georgy, J., Karamat, T., Iqbal, U., and Noureldin, A. (2011). Enhanced MEMS-IMU/odometer/GPS integration using mixture particle filter. Volume 15, 239–252.

Green, W. E., Ch, P. Y., Sevcik, K., and Barrows, G. (2003). Autonomous Landing for Indoor Flying Robots Using Optic Flow. In *ASME International Mechanical Engineering Congress and Exposition*, 1347–1352.

Gu, Y., Lo, A., and Niemegeers, I. (2009). A Survey of Indoor Positioning Systems for Wireless Personal Networks. In *IEEE Communications Surveys & Tutorials In Communications Surveys & Tutorials,*, Volume 11, 13–32.

Harter, A., Hopper, A., Steggles, P., Ward, A., and Webster, P. (2002). The anatomy of a context-aware application. Volume 8, 187–197.

Hile, H. and Borriello, G. (2007). Information overlay for camera phones in indoor environments. In *3rd international conference on Location-and context-awareness*, 68–84.

Irschara, A., Zach, C., michael Frahm, J., and Bischof, H. (2009). From structure-from-motion point clouds to fast location recognition. In *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR*, 2599–2606.

Ivanchenko, V., Coughlan, J., and Shen, H. (2008). Detecting and locating crosswalks using a camera phone. Volume 0, 1–8.

Ivanchenko, V., Coughlan, J., and Shen, H. (2010). Real-time walk light detection with a mobile phone. In *Proceedings of the 12th international conference on Computers helping people with special needs*, 229–234.

Ji, R., Duan, L.-Y., Chen, J., Yao, H., Yuan, J., Rui, Y., and Gao, W. (2012). Location Discriminative Vocabulary Coding for Mobile Landmark Search. In *International Journal of Computer Vision*, Volume 96, 290–314.

Jin, G.-y., Lu, X.-y., and Park, M.-S. (2006). An Indoor Localization Mechanism Using Active RFID Tag. In *IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing -Vol 1 (SUTC'06) - Volume 01*, SUTC '06, 40–43.

Kang, H., Efros, A., Kanade, T., and Herbert, M. (2009). Image Matching in Large Scale Indoor Environment. In *Computer Vision and Pattern Recognition*, 33–40.

Kang, H., Efros, M. H., and Kanade, T. (2009). Image Matching in Large Scale Indoor Environment. In *Proc. Computer Vision and Pattern Recognition*, 33–40.

Kawaji, H., Hatada, K., Yamasaki, T., and Aizawa, K. (2010). Image-based indoor positioning system: fast image matching using omnidirectional panoramic images. In *1st ACM international workshop on Multimodal pervasive video analysis*, 1–4.

Khan, N. and James (2010). Interview wih blind person: Navigation System Requirements. Technical report.

Khan, N. and McCane, B. (2012). Smartphone application for indoor scene localization. In *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility*, ASSETS '12, 201–202.

Khan, N., McCane, B., and Mills, S. (2012a). Feature set reduction for image matching in large scale environments. In *Proc. International Conference on Image and Vision Computing*, 68–72. ACM.

Khan, N., McCane, B., and Mills, S. (2012b). Vision based indoor scene localization via smart phone. In *Proceedings of the 13th International Conference of the NZ Chapter of the ACM's Special Interest Group on Human-Computer Interaction*, CHINZ '12, 88–88.

Khan, N., McCane, B., and Wyvill, G. (2011a). Homography based Visual Bag of Word Model for Scene Matching in Indoor Environments. In *Proc. International Conference on Image and Vision Computing*.

Khan, N., McCane, B., and Wyvill, G. (2011b). SIFT and SURF Performance Evaluation Against Various Image Deformations on Benchmark Dataset. In *Proc. Digital Image Computing and Techniques*, 501–506.

Kidono, K., Miura, J., and Shirai, Y. (2002). Autonomous Visual Navigation of a Mobile Robot Using a Human-Guided Experience. Volume 40, 2–3.

King, T., Kopf, S., Haenselmann, T., Lubberger, C., and Effelsberg, W. (2006). In *COMPASS: A Probabilistic Indoor Positioning System Based on 802.11 and Digital Compasses*.

Ko, S.-i., Choi, J.-s., and Kim, B.-h. (2008). Indoor Mobile Localization System and Stabilization of Localization Performance using Pre-filtering. Volume 6, 204–213.

Lazebnik, S., Schmid, C., and Jean, P. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision Pattern and Recongition*, 2169–2178.

Leung, K. Y. K., Clark, C. M., and Huissoon, J. P. (2008). Localization in urban environments by matching ground level video images with an aerial image. In *International Conference on Robotics and Automation*, ICRA, 551–556.

Li, Y., Snavely, N., and Huttenlocher, D. P. (2010). Location recognition using prioritized feature matching. In *Proceedings of the 11th European conference on Computer vision: Part II*, ECCV'10, Berlin, Heidelberg, 791–804. Springer-Verlag.

LoadstoneGPS (2010). LOADSTONE GPS - free GPS software for your mobile phone.

Lopes, C. V., Haghighat, A., Mandal, A., Givargis, T., and Baldi, P. (2006). Localization of Off-the-Shelf Mobile Devices Using Audible Sound: Architectures, Protocols and Performance Assessment. In *ACM SIGMOBILE Mobile Computing and Communication Review*, Volume 10, 38–50.

Lowe., D. (2004). Distinctive Image features from scale invariant keypoints. *International Journal of Computer Vision*, *60*, 91–110.

Meers, S. and Ward, K. (2005). A Substitute Vision System for Providing 3D Perception and GPS Navigation via Electro-Tactile Stimulation. In *Proc. International Conference on Sensing Technology*.

Mulloni, A., Wagner, D., Barakonyi, I., and Schmalstieg, D. (2009). Indoor Positioning and Navigation with Camera Phones. *IEEE Pervasive Computing*, *8*(2), 22–31.

Ni, L. M., Liu, Y., Lau, Y. C., and Patil, A. P. (2004). LANDMARC: indoor location sensing using active RFID. *Wireless Networks*, *10*(6), 701–710.

Nister, D. and Stewenius, H. (2006). Scalable recognition with a vocabulary tree. In *Proc. Computer Vision and Pattern Recognition*, Volume 2, 2161–2168.

Pace, S. (1995). *The global positioning system: Assessing national policies.* Santa Monica, CA: RAND.

Philbin, J., Isard, M., Sivic, J., and Zisserman, A. (2007). Object Retrieval with Large Vocabularies and Fast Spatial Matching. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Philbin, J., Isard, M., Sivic, J., and Zisserman, A. (2008). Lost in quantization: Improving particular object retrieval in large scale image databases. In *Computer Vision and Pattern Recognition*.

Povidi (2008). Soundpost A revolution in blindness orientation.

Priyantha, N. B. (2005). The Cricket Indoor Location System. In *PhD Thesis, MIT*.

Priyantha, N. B., Chakraborty, A., and Balakrishnan, H. (2000). The Cricket location-support system. In *Proceedings of the 6th annual international conference on Mobile computing and networking*, MobiCom '00, 32–43.

Ran, L., Helal, S., and Moore, S. (2004). Drishti: an integrated indoor/outdoor blind navigation system and service. 23–30.

Ranganathan, A. (2012). PLISS: Detecting and Labeling Places Using Online Change-Point Detection. In *Proceedings of Robotics: Science and Systems*, 1–8.

RNZFB (2012). Royal New Zealand Foundation of the Blind, New Zealand.

Robertson, D. and Cipolla, R. (2004). An image-based system for urban navigation. In *Proceedings British Machine Vision Conference*, 819–828.

Ruf, B., Kokiopoulou, E., and Detyniecki, M. (2010). Mobile museum guide based on fast SIFT recognition. In *International conference on Adaptive Multimedia Retrieval: identifying, Summarizing, and Recommending Image and Music*, 170–183.

Saeedi, P., Lawrence, P. D., and Lowe, D. (2006). Vision-based 3-D trajectory tracking for unknown environments. *Transactions of Robotics*, *22*(1), 119–136.

Sattler, T., Leibe, B., and Kobbelt, L. (2011). Fast image-based localization using direct 2D-to-3D matching. In *ICCV*, 667–674.

Se, S., Lowe, D., and Little, J. (2001). Vision-based Mobile Robot Localization And Mapping using Scale-Invariant Features. In *In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA*, 2051–2058.

Se, S., Lowe, D., and Little, J. (2005). Vision-based global localization and mapping for mobile robots. Volume 21, 364–375.

SenderoGroup (2012). Miniguide.

Silveira, G., Malis, E., and Rives, P. (2008). An Efficient Direct Approach to Visual SLAM. *IEEE Transactions on Robotics*, *24*(5), 969–979.

Sim, R. and Little, J. J. (2006). Autonomous vision-based exploration and mapping using hybrid maps and Rao-Blackwellised particle filters. In *IROS*, 2082–2089. IEEE.

Sivic, J. and Zisserman, A. (2003). Video google: A text retrieval approach to object matching in videos. In *Proc. International Conference on Computer Vision.*, 1470–1477.

Skrypnyk, I. and Lowe, D. G. (2004). Scene Modelling, Recognition and Tracking with Invariant Image Features. In *Proceedings of the 3rd IEEE/ACM International Symposium on Mixed and Augmented Reality*, ISMAR '04, 110–119.

Song, Z., Jiang, G., and Huang, C. (2011). A Survey on Indoor Positioning Technologies. In *Second International Conference on Theoretical and Mathematical Foundations of Computer Science*, Volume 164, 98–106.

Strong, P. The History of the White Cane.

System, A. B. (2008). Active Badge System, Website.

Talukder, A. and Matthies, L. (2004). Real-time detection of moving objects from moving vehicle using dense stereo and optical flow. In *Conference on Intelligent Robots and Systems*. IEEE.

Thrun, S., Bennewitz, M., Burgard, W., Cremers, A., Dellaert, F., Fox, D., Haehnel, D., Rosenberg, C., Roy, N., Schulte, J., and Schulz, D. (1999). MINERVA: A second generation mobile tour-guide robot. In *Proc. of the IEEE International Conference on Robotics and Automation (ICRA'99).*

Trucco, E. and Plakas, K. (2006). Video tracking: A concise survey. Volume 31, 520–529.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition*, 511–518.

Want, R., Hopper, A., Falcão, V., and Gibbons, J. (1992a). The active badge location system. *ACM Transactions Information Systems*, *10*(1), 91–102.

Want, R., Hopper, A., Falcão, V., and Gibbons, J. (1992b). The active badge location system. *ACM Trans. Inf. Syst.*, *10*(1), 91–102.

Ward, A., Jones, A., and Hopper, A. (2004). New Location Technique for the Active Office. In *IEEEPersonal Communications*, Volume 4, 98–106.

Wbesite, W. (2008). WhereNet.

Webmaster (2012a). MAKING A DIFFERENCE.

Webmaster (2012b). Royal Foundation of blind New Zealand.

Website, S. S. (2008). Sonitor Systems.

WHO (2012). Visual impairment and blindness. Technical report.

Wikiepdia (2012a). Guide Dog.

Wikiepdia (2012b). White Cane.

Wikipedia (2012). Refreshable braille display.

Winters, N. and Santos-Victor, J. (1999). Omni-directional Visual Navigation. In *In Proceedings of the 7th International Symposium on Intelligent Robotics Systems*, 109–118.

Zhang, W. and Kosecka, J. (2006). Image Based Localization in Urban Environments. In *Proceedings of the Third International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT'06)*, 3DPVT '06, 33–40.

Zhou, C., Wei, Y., and Tan, T. (2003). Mobile Robot Self-Localization Based on Global Visual Appearance Features. In *Proceedings of the 2003 IEEE Int. Conf. on Robotics and Automation, ICRA 2003, September 14-19, 2003*, 1271–1276.