8-2013

Indoor Scene Knowledge Acquisition Using Natural Language Descriptions

Saranya Kesavan

INDOOR SCENE KNOWLEDGE ACQUISITION USING

NATURAL LANGUAGE DESCRIPTIONS

By

Saranya Kesavan

B.E. (Geo Informatics), College of Engineering, Anna University, India

A THESIS

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

(in Spatial Information Science and Engineering)

The Graduate School

The University of Maine

August 2013

Advisory Committee:

Nicholas A. Giudice, Assistant Professor of School of Computing and Information Science, Advisor

Kate Beard-Tisdale, Professor of School of Computing and Information Science

Reinhard Moratz, Associate Professor of School of Computing and Information Science

THESIS ACCEPTANCE STATEMENT

On behalf of the Graduate Committee for Saranya Kesavan, I affirm that this manuscript is the final and accepted thesis. Signatures of all committee members are on file with the Graduate School at the University of Maine, 42 Stodder Hall, Orono, Maine.

Dr. Nicholas A. Giudice, Date

Assistant Professor of School of Computing and Information Science

LIBRARY RIGHTS STATEMENT

In presenting this thesis in partial fulfillment of the requirements for an advanced degree at The University of Maine, I agree that the Library shall make it freely available for inspection. I further agree that permission for “fair use” copying of this thesis for scholarly purposes may be granted by the Librarian. It is understood that any copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Signature:

Date:

INDOOR SCENE KNOWLEDGE ACQUISITION USING

NATURAL LANGUAGE DESCRIPTIONS

By Saranya Kesavan

Thesis Advisor: Dr. Nicholas A. Giudice

An Abstract of the Thesis Presented
in Partial Fulfillment of the Requirements for the
Degree of Master of Science
(in Spatial Information Science and Engineering)
August, 2013

The existing research addressing non-visual indoor navigation is limited to route guidance between locations (i.e., the corridor network). This focus ignores many critical regions contained within indoor spaces (e.g., rooms, lobbies, etc.), locations which are often as challenging to learn and navigate without vision as are the routes connecting them. To address this challenge, this thesis investigates the use of natural language (NL) descriptions as a non-visual medium for providing access to indoor scenes, including room structure, furniture placement, and location of salient landmarks. The work is part of a larger project to develop a system, called the Describer for Indoor Scenes (DISc) that uses automatically generated NL descriptions to represent indoor scenes based on photos taken in real-time from mobile devices. In order to develop cognitively comprehensible NL descriptions of indoor scenes, it is critical to first understand how humans describe and interpret the scene in order to support spatial behavior. To this end, six behavioral experiments were conducted to characterize scene descriptions

generated by human observers and to optimize these descriptions based on cognitive constraints and the structure of linguistic information to be included to best support non-visual learning, representation, and navigation.

The visual information that can be captured about a scene from photographs is potentially limited, both in quality and quantity, compared to the information apprehended from real time scene perception. Importantly for the DISc system, results from experiments 1, 2, and 3 converge to demonstrate that photographic observations are functionally equivalent to real time observations of indoor scenes in supporting spatial behavior and show that photographs can be used as information source in DISc. The data collected in these experiments showed that humans adopted different scene description strategies. To understand how the description strategy (i.e., order of objects) affected scene learning and reconstruction, a 4^{th} behavioral experiment was conducted. Results from this experiment suggest that following a cyclic path while describing an indoor scene (called a "Round-About strategy") was the most efficient approach for acquiring and representing spatial knowledge.

The results from the first four experiments elucidated that people used two different angular units (clock face and degree measurements) to describe directional information. However, it was not clear from the extant literature how angular units affect spatial apprehension of the listener or which measure yields the most accurate performance. As directional information is critical for specifying the location of objects in a scene, this question was addressed in a fifth experiment, with results demonstrating that the most

accurate performance manifested when angular directions were given as clock face units rather than degree measurements (i.e., 1:00 versus 30 degrees). Results also demonstrated that participants were equally accurate at producing angular values of 15 degrees or half hour increments (e.g., 1:30), which is meaningful as this is a 100% increase in precision from the standard clock face units employed in previous studies.

The sixth and final behavioral experiment was conducted to investigate whether the optimized NL scene descriptions support non-visual navigation of indoor scenes and if performance differs when using static or updated descriptions, meaning that they either were given from a fixed user perspective in the scene (as was done in the earlier experiments) or that the perspective changed based on the user's position and orientation. Results showed a clear advantage for updated NL descriptions on navigation accuracy, indicating that to be maximally effective, DISc should implement descriptions based on the user's real-time position and orientation as they move.

Taken together, the results of six human experiments extend earlier research with route navigation by showing that optimizing NL indoor scene descriptions based on perceptual and cognitive factors led to efficient spatial learning, representation, and navigation. These empirical results provide the much needed proof of concept for the efficacy of future development of DISc as a fully automated NL scene description system.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my mentor and advisor Dr. Nicholas A. Giudice for his kindness, support and patience throughout this work. Without his timely draft reviews, suggestions, and advice, it would have been impossible to perform this research and write this thesis. I have enjoyed my discussions with him and they have been instrumental in helping me grow as a researcher. I would also like to thank the advising committee members: Dr. Kate Beard- Tisdale and Dr. Reinhard Moratz for their support throughout my graduate career and in this work.

I also take this opportunity to thank all the professors in the Department of Spatial Information Science and Engineering for their support in the success of my work. Special thanks to my colleagues and friends at VEMI Lab, department, and the University: Avi Rude, Balaji Venkatesan, Bill Whalen, Christopher Bennett, Hengshan Li, J.C. Whittier, Jonathon Cole, Joshua Leger, Kate Cuddy, Liping Yang, Matt Dube, Meghan White, Monoj Raja, Nate Laspina, Rick Corey, Riju Shreshta, RJ Perry, Shravani Tadepalli, Sriram Bhuvnagiri, Sugandha Shankar, Tim McGrath, Upasana Pandey.

I would also like to acknowledge the support from the National Science Foundation (NSF grant CDI-0835689) and the National Institutes of Health (NIH grant EY017228-02A2) awarded to Nicholas A. Giudice without which my graduate studies would not have been possible.

Finally, I would like to thank my parents and my family's close friend Vimala Kannan for their unwavering support in all of my endeavors. More than everyone, I would like to thank Hari prasath Palani for everything he gave me in my life to make it meaningful.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 Introduction

Navigation involves the process of controlling and monitoring the movement of a physical entity from one place to another (Bowditch, 1802). Every human being performs this act of navigation, where the process is controlled by the brain with the help of myriad underlying cognitive processes. Various sensory inputs like visual apprehension of an environment, sound reflection by environmental entities, extraction of spatial information from smell, vestibular information, haptic information, and kinesthetic information are all involved in cognitive processes associated with navigation (D. Montello & Freundschuh, 1995). Our brain stores this collection of information as both short term and long term memory representations of space, each aiding in performing the process of navigation.

Although spatial knowledge is acquired as a result of information obtained from multiple sources of sensory perception, information obtained from vision is generally accepted as the primary source for acquiring spatial knowledge. According to (Thinus-Blanc & Gaunet, 1997): vision allows for simultaneous perception of multiple details about the external world. Visual information is an integral feature of all physical objects present in an environment. Unlike smell or sound, visual information is accurate and precise when compared to other modes of perceiving space (see section 2.1 for more discussion). Thus, the accessibility of visual cues plays an important role in the process of acquiring spatial information and hence is a critical source of input affecting the process of navigation (Golledge, Klatzky, & Loomis, 1996). However, blind and low-vision

people do not have access to information from visual cues about their surroundings and hence they must rely on alternative sensory information such as auditory, olfactory, haptic, and vestibular information. All these modalities are less intuitive, more proximal, less accurate and more attention demanding when compared with visual sensing of surrounding space. Hence navigation is often a challenging task for blind people as they must rely on spatial information obtained through non-visual sensory perception (Thinus-Blanc & Gaunet, 1997). In order to address this challenge, we should empower blind people by providing spatial information about their environment in an intuitive and information-rich manner, which will enable them to more easily plan and perform navigation tasks even in unknown spaces.

The research work presented in this thesis describes an intuitive way to provide spatial information to blind or low-vision people using natural language descriptions about their environment. Owing to recent technological advancements and promising results in the field of natural language processing and generation, we argue that natural language is one of the most efficient modes of information access for incorporation in non-visual interfaces and intelligent devices to be used by blind people. A critical problem with current natural language description generation is that it primarily concentrates on the semantic and syntactic aspects of linguistic descriptions in order to mimic human speech patterns. However, there is little formal research on understanding the ways in which humans verbally summarize spatial information about physical environments.

Hence, in order to generate natural language descriptions that facilitate environmental learning and navigation, it is important to first understand the ways in which humans would naturally describe (that is to verbally narrate) the space around them. natural language generated without understanding of this narration is only a formal arrangement of words into sentences, which follows the syntactic and semantic rules of a language following a specific architecture. By contrast, this thesis approaches natural language descriptions from a cognitive standpoint while describing a space. That is, it analyzes the elements of a natural language description of an environment (see section 3.4 for a detailed discussion about the elements of natural language descriptions that are considered in this thesis) and their accuracy in imparting spatial knowledge to the listener.

Currently there are navigation systems available for assisting non-sighted individuals in navigating if their destination is a known street address or a building. There are also prototypical navigation assistance systems available for indoor navigation, e.g. if the destination is a room in the building (see (Ohkugo et al., 2005) for a prototypical wayfinding system). But once a non-sighted person enters a room, he is devoid of any assisting technologies to learn about that space, except for self-exploration, which includes various shortcomings, such as the time and risk involved (especially if unfamiliar indoor scenes are not safe for blind people to explore by themselves). Moreover, a human activity pattern survey indicates that people spends 87% of their time indoors. Hence, this thesis concentrates on describing spatial information only in

the context of indoor scenes (such as an office, kitchen, hotel lobby etc.), and not on generic indoor corridor layouts, or outdoor environments.

This introduction chapter, on the whole, discusses the reasons behind our interest to empower blind and low vision people to acquire indoor scene knowledge by means of natural language Descriptions.

1.1. Sample Scenario

Before understanding the importance of indoor scene knowledge acquisition, it is important to first understand how the knowledge about environmental entities affects the global navigation process, especially for blind people. In order to clearly illustrate the previous statement, a sample scenario is discussed below.

This scenario details an example by comparing and contrasting the ways in which a blind person and a sighted person might plan and follow a route between their starting and destination locations. This comparison is done in order to elaborate the differences between sighted and blind spatial learning and navigation processes and the information involved, so that we can gain insights about the associated challenges. It is important to note that although blind and low vision people lack access to visual information, they use well established techniques and methods to learn about their environment using other perceptual senses, but with increased cognitive effort (Shingledecker, C, 1983).

On reading this sample scenario, you will hopefully appreciate how navigating with vision is a perceptual process, while navigating without vision is a cognitively laborious

task. You can also likely comprehend from the discussion how technology plays a particularly important role in providing access to spatial information to a blind person. As of writing this thesis, there are no integrated way-finding or assistive travel aids available for a blind person to learn or navigate an indoor scene (Giudice & Legge, 2008).

1.1.1. Sample Task

The reader is invited to imagine a sample task being performed by two individuals: Jack (a non-sighted person) and Mick (a sighted person). Both Jack and Mick are 35-year-old software engineers, living in the same apartment complex. They both used an online appointment booking system to schedule an appointment with their tax consulting firm and both received a confirmation email stating the time and address for their visit. Jack used a screen reading software (see http: //www.freedomscientific.com/products /fs/JAWS-product-page.asp for details), which converts visual information presented as text in the email to auditory information, while Mick used visual information to understand the text in his confirmation email.

The tax consulting firm that they have to visit is in their neighboring city, which they are not familiar with and hence they wanted to acquaint themselves with the locality in which they have to navigate. So Mick opened Google maps for that area which included a wide range of visual information such as the location of important landmarks, street names, road networks, symbolic representations of subway stations and bus stations, street view of the roads, bird's-eye view of the city as a whole and so on. On the other

hand, Jack printed a tactile map of the city, which gives information about important landmarks and the road network connecting different parts of the city. The amount of information presented by blind navigation interfaces is an important factor to be considered (Giudice & Legge, 2008) as it affects the complexity of non-visual maps. The tactile map used by Jack only conveyed limited information about the city in which he had to travel. Also (Golledge, 1991) suggested that tactile maps with uncluttered displays are best when compared to those tactile maps which try to mimic all the information in visual maps. Hence Jack, using a limited information tactile map received only a rough idea of the location of landmarks and road networks, emphasizing the route he should take through the city the next day. On the other hand, Mick has a better idea about the overall city as he learned about the global area as well as the particular travel route.

With the help of more detailed online trip planning systems, they both knew that they had to take a bus from their apartment complex to reach the bus stop which is two blocks away from the building in which their tax consulting firm was located. From that bus stop they had to then walk to their destination building based on the directions given by their GPS navigation system. Jack used an accessible speech-based system that came along with his GPS (http://www.senderogroup.com /products /GPS/allgps.htm).

They both started their journey the next morning, separately. Mick started from his apartment and reached the bus stop near his apartment complex, without paying any attention to the cognitive processes that helped him to reach the bus stop. Jack knows

the location of the nearest bus stop from his apartment. So he started from his apartment and mentally updated his position and orientation by following auditory cues and simultaneously referring to prominent tactile landmarks that are present on his way, like the gate of his apartment complex, mailbox, and parking meter on the side of the road. (See (Lawson & Wiener, 2010; Long & Giudice, 2010; Loomis et al., 1993) for more discussion on strategies followed by non-sighted people while navigating from one place to another).

Both Jack and Mick performed position based navigation or piloting, which involves the use of external information about the environment in order to update their position and orientation. The difference is that Mick used visual information, which is cognitively trivial, yet a rich information source for performing the task of reaching the bus stop, while Jack simultaneously used auditory and tactile information input, which is less accurate and more cognitively demanding (Rieser, Guth, & Hill, 1986). So far, neither of them had any issues with navigation since they both have a well-defined cognitive map of the environment in which they navigated (i.e. familiar environment). As mentioned in section 1.1, we can see that Jack continuously accesses spatial information about his environment even in familiar surroundings, mainly to ensure safety in his path of travel and to update his location on-route to the destination. Fortunately the SONAR cane in his hand served its purpose by enabling him to access information about objects in his proximity.

Upon exiting the bus, Mick had no issues in orienting himself with the right direction in which their destination building was located. Based on environmental cues like the traffic pattern, position of the sun, auditory information from permanent landmarks, Jack was also able to orient and position himself in the right direction, (see (Lawson & Wiener, 2010; Long & Giudice, 2010; Wiener, Welsh, & Blasch, 2010) to learn more about orientation and mobility training). Jack and Mick used their GPS system and Google maps in their smartphone to get the route that they have to follow to reach the building where their tax consulting firm was located.

When Jack entered the building, he used a prototype grid based RFID positioning and routing system as proposed by (Chumkamon, Tuvaphanthaphiphat, & Keeratiwintakorn, 2008) in order to reach his tax consulting firm which was located on the second floor.

(The indoor positioning system mentioned here is totally a fictional system conceived from the idea proposed by (Chumkamon et al., 2008). Currently there are no commercially successful indoor navigation assistance systems available for assisting non-sighted individuals. A large body of research is currently being carried out for this purpose, most of them requiring a non-sighted traveler to wear portable computing devices and/or sensors (as discussed in (Cheok & Yue, 2011; Golding & Leash, 1999; Ran, Helal, & Moore, 2004)) . Considering the availability of Google Indoor maps and the research going on in the field of smart phone based indoor positioning and navigation assistance systems, the time is not too long to a full-fledged indoor navigation system that supports non-sighted people (see research projects by (Dekel & Schiller, 2010;

Lukianto, Christian, & Sternberg, 2010; Pei, Chen, Chen, Leppäkoski, & Perttula, 2009; Serra, Carboni, & Marotto, 2010) for indoor navigation assistance systems using smartphones. Refer to Chapter 2 of this thesis for more discussion about current indoor navigation systems.)

Jack had previously been in a situation like this, where he was lost in an unfamiliar building. The problem he has with indoor spaces is the lack of clear orientation cues compared to outdoor environments. For example, the corridors in the indoor spaces are not named as are streets in outdoor spaces and also there are no reference systems available for defining the space semantically (Giudice, Walton, & Worboys, 2010). Apart from the above mentioned problems, the vertical axis of the building adds complexity to the indoor navigation strategy used because of the differences in floor layout of multi-storied buildings. (See (Giudice & Li, 2012; Giudice et al., 2010) for more discussion on the reasons behind indoor navigation being relatively complicated when compared with outdoor navigation (both for sighted and non-sighted people)). Mick on the other hand reached the tax consulting firm by looking at the room signs, You-Are-Here maps and the like which Jack did not have access to.

Finally, both Jack and Mick reached their tax consulting firm. Mick entered the lobby of the tax consulting firm and went directly to talk to the representative at the information desk. On the other hand, Jack who is standing at the doorway had no idea about the shape of the room, the location of the information desk or the location of the waiting area. So far in his trip, Jack had access to environmental cues using navigation aids

(most of them working fine and commercially available, while the other prototypical navigation aids showed to work better in ideal environmental conditions). Now he does not have access to any technology that could help him to describe the indoor scene setting in which he has to move or to help him to navigate from one point to the other. This is primarily due to the lack of research on indoor scene descriptions, a gap this thesis aims to address.

In this situation Jack could have explored the indoor scene by himself to learn about it. But manually exploring an indoor scene has many disadvantages. At this point, try closing your eyes and imagine yourself in the position of Jack, who is standing at the doorway of an unfamiliar indoor scene and think about the downsides of manually and randomly exploring the lobby of the tax consulting firm. First of all, self-exploration is not at all efficient owing to the risk and time of searching the space with no cues to provide guidance or orientation, effort that increases proportionally with the complexity of the indoor scene involved. Moreover, manually exploring an indoor scene might lead to embarrassment and to feel undignified. Also, it could be invasive if you have to interact with others while exploring the scene.

Considering the downsides of self-exploration, Jack wished to have his smartphone give a route to the information desk (if present) from his current position as it did in outdoor environments. He also prefers his smartphone to describe the spatial location of objects present in the room, in order to understand the global layout of the indoor scene. This global layout knowledge could help Jack to form a mental map of the room, which

would enable him to perform accurate position-based navigation or piloting from one point to the other in an efficient and rapid manner.

1.2. Research Goal

Jack felt that to be useful, a complete navigation system for him would include a system that would assist him to navigate from an origin to the destination, where the destination could be anywhere from a postal address, a building, a room in the building, or a particular location or object inside a room. From the sample scenario discussed above, we know there are navigation systems available for assisting non-sighted individuals if their destination is a street address or a building (for example Google street maps with the text to speech feature available on a smartphone) and there are prototypical navigation assistance systems available if the destination is a room in the building. But once a non-sighted person enters a room, he is devoid of any assistive technologies except for self-exploration, which includes a number of downsides as described in the previous section.

Also, as mentioned in the sample scenario, the use of different types of interfaces (like the RFID enabled SONAR cane for travelling in indoor environments, and the Smartphone based navigation system for travelling in outdoor environments) involves learning to use different types of technologies and user interfaces. For example in Jack's case, he has to learn how to interpret the route directions he gets from his RFID based route directions and he must also be able to understand the GPS based routing directions. Hence it would be useful to have a single user interface for providing

navigation assistance across both indoor and outdoor spaces. Designing a single user interface for both these spaces is an intricate task, mainly because both indoor and outdoor spaces have different navigational challenges.

Fortunately, the advancement of smartphones has lessened the burden of designing an intricate user interface that could be used to solve both indoor and outdoor navigational needs. Smartphones, embedded with multiple sensors and accessibility features are considered one of the dominant devices of choice in the blind / low-vision demography. There are a number of successful smartphone applications available in the market to assist non-sighted people in navigating through outdoor spaces (see (www.ariadnegps.eu) for example) and a large amount of research in indoor navigation has included the use of smartphone devices. This emphasis clearly shows that smartphones will play an important role in navigation assistance systems for non-sighted people in the future. This trend also suggests that it would be best to use the sensors of a smartphone system to develop an indoor scene description system.

The research work described in this thesis aims to fill the gap that exists in current navigation assistance systems for non-sighted individuals using smartphone devices. That is, the current research work proposes a scene description system called DISc (Describer for Indoor Scenes) for assisting non-sighted individuals to navigate in any unfamiliar indoor scene.

1.3. Describer for Indoor Scenes (DISc) – An Overview

In order to better understand about DISc and its usage, consider Jack's situation in the sample scenario when he is standing at the doorway of the tax consulting firm. If he wants to navigate to the information desk, he should walk towards the information desk by avoiding all obstacles in his path without any idea about its spatial location with respect to his position in the doorway. One way to navigate easily in the room is to become familiarized with the objects and the spatial relationship between those objects present in the room prior to navigating there. Hence it can be clearly seen that it is important to have a personal guidance system to navigate in indoor scenes like the guidance systems available to navigate in outdoor environments and indoor corridor networks. Also it is important to remember that indoor scene navigational needs are different from outdoor and indoor corridor navigational needs. For example, in outdoor and indoor corridor navigation people are used to route descriptions, while in indoor scene navigation it is important to have survey descriptions owing to the proximity of objects in a physically bounded space (see section 3.2 for more discussion which contrasts survey and route descriptions). Hence the personal guidance system to be used in indoor scenes should be able to convey spatially relevant information in the simplest and most efficient medium to support the task of navigation.

Thus, DISc will help Jack to get familiarize with the lobby of the tax consulting firm in which he has to navigate. For example, assume that Jack's smartphone has an application called DISc. When he enters the tax consulting firm, where he has to find the

information desk, he starts the DISc application and takes strategic pictures, roughly capturing the entire scene in which he has to navigate, as shown in Figure 1.1.

In the background, the image processing component of DISc analyzes the images taken by Jack to identify the objects present in that indoor scene and extracts the spatial information of those objects. Then the natural language Generation component of DISc converts the spatial information extracted from the photographs to a more meaningful and easily interpretable natural language Description of the tax consulting firm's lobby. With this description, Jack could easily interpret the route that he has to follow from his current position at the doorway to the information desk without any trial and error method.

Apart from determining the route to the place of his interest, he could mentally develop the spatial relations of the space in general—so he also can develop a cognitive map of the lobby which will assist him in making subsequent actions, like finding the seating area, the elevator, or the door out, or in performing these tasks on a subsequent visit as he now has the opportunity to develop a mental representation of the global spatial relations based on the verbal descriptions received.

It is important to note here that with the development in current imaging technologies it is promising to use other imaging sensors like Kinect or even multispectral images to be used to acquire spatial information of indoor scenes (for example see (Khoshelham & Elberink, (2012)). But this thesis concentrates only on the spatial information that is being collected and not on the source sensor that is used for collecting that spatial

information. Hence the technical details of using other information sources, such as Kinect and other multispectral images, are not discussed as part of this thesis.

Figure 1.1 Image describing the scope of our scene description system

1.4. Perception of Scene Layout Grant

This thesis is a part of the research work funded by a grant from the National Science Foundation (Beard, Giudice, Latecki, Moratz, & Daniilidis, 2010). There are three different components involved in the design of DISc. These include object detection and localization component, a spatial representation and reasoning component and a spatial description production component.

The object Detection and Localization component involves image analysis to recognize and detect objects in an indoor scene followed by the graphical representation of objects and their spatial relation to construct an occupancy grid with imprecise

boundaries (see (Moratz, 2006; O'Shaughnessy, 2012) for discussion about object Detection and Localization). The spatial representation and reasoning component deals mainly with the analysis of how spatial representation supports spatial reasoning. For example, this component deals with how to calculate new perspectives from the available object perspectives (which are similar to human mental rotation) and so on.

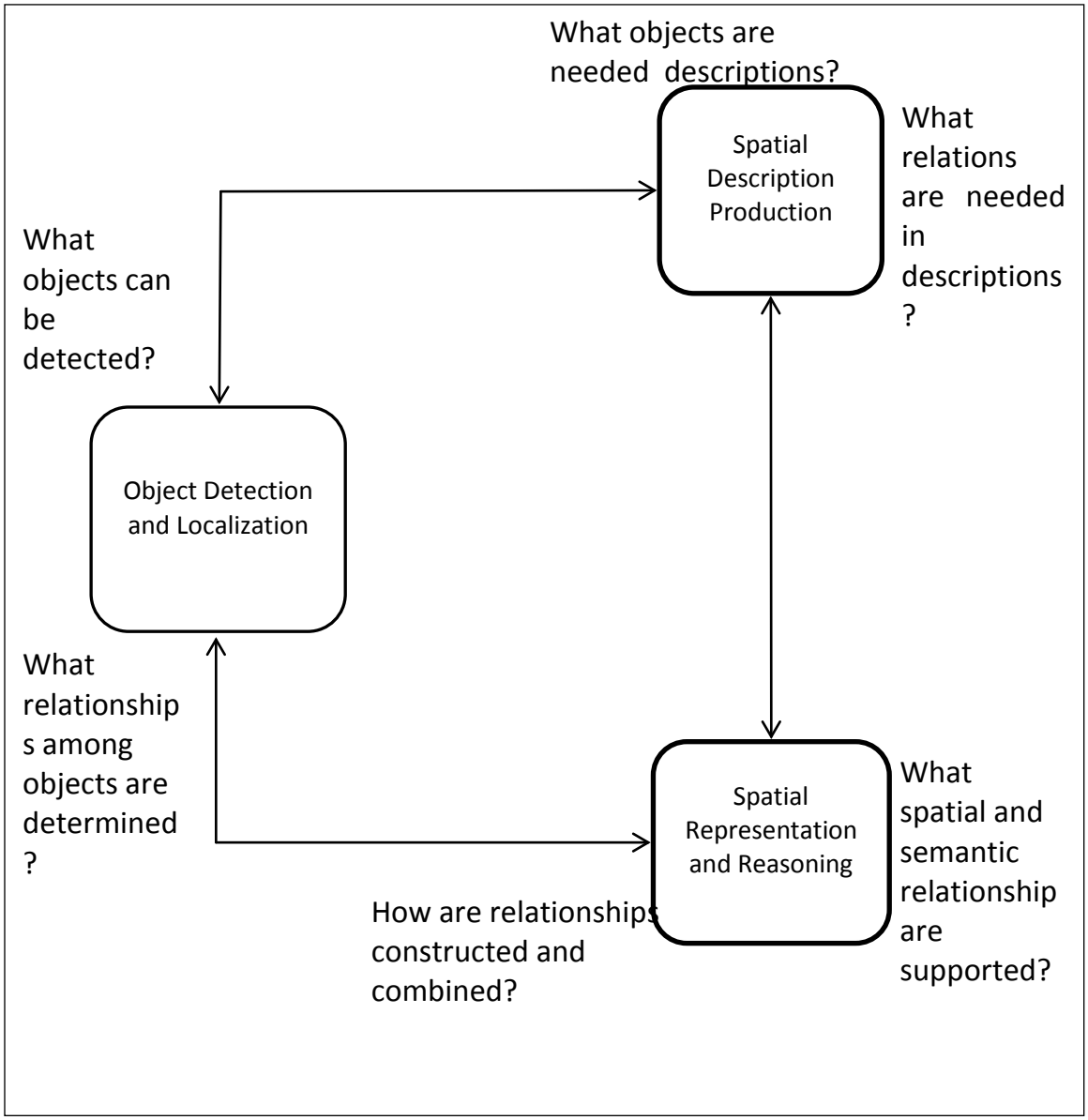

Figure 1.2 Components involved in fully functional screen description system

The spatial description production component analyzes the ways in which the metric details obtained from the object detection and localization component will be converted into a natural language Description of the indoor scene based on the understanding from the spatial representation and reasoning component. Figure 1.2 shows the relationship between these three components (courtesy of: NSF grant CDI-1028895).

1.5. Thesis Concentration

This thesis concentrates only on the Spatial Description Production Component of the Scene Description system discussed above. Hence the main goal of the thesis is to analyze the ways in which the abstract numerical spatial information obtained from the Object Detection and Localization component of the Scene Description system can be converted to a meaningful natural language description of Indoor scenes. The output should save accurate learning of an indoor scene, representation of the spatial relations between objects present in that scene, and support spatial activities in the scene using this non-visual information. Figure 1.3 highlights the areas that are covered in this thesis and the overall architecture of the Scene Description system.

See (Liu, Yang, & Jan, 2012; Ma, Latecki, & Sciences, 2011; Wang, Bai, You, Liu, & Latecki, 2012; Xinggang Wang, Bai, Liu, & Jan, 2011; Yang, Adluru, & Latecki, 2011), which were the publications produced in other components as a result of this research grant.

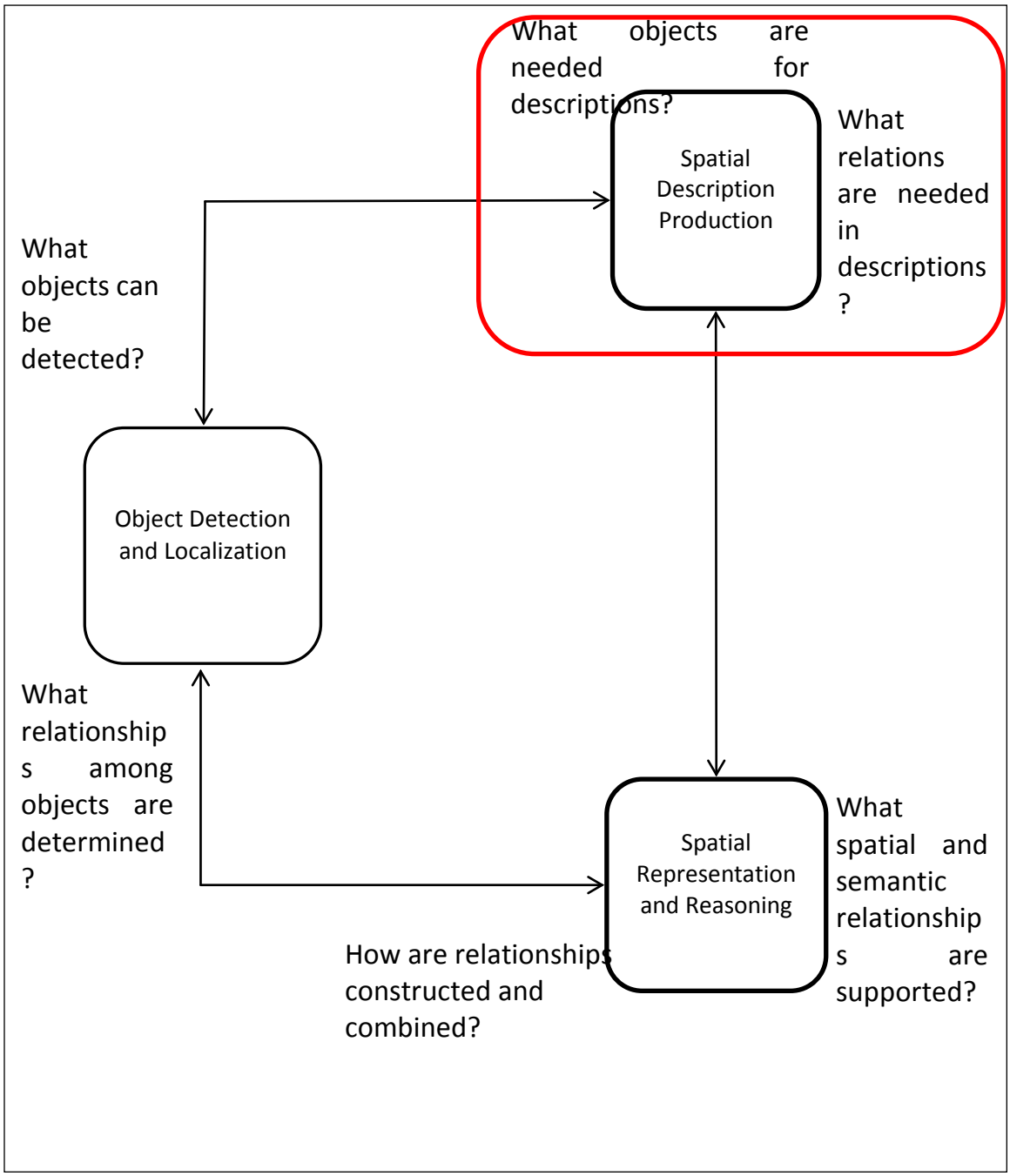

Table 1.3 Component focused in this thesis

1.6. Structure of the Thesis

Chapter 2 of this thesis discusses previous research related to the work presented here. Chapter 3 explains current challenges related to natural language Generation of Indoor scenes. Chapter 4,5,6, and 7 discusses behavioral experiments conducted to find solutions for the challenges mentioned in Chapter 3. Chapter 8 describes the findings from the behavioral experiments presented in chapters 4,5,6 and 7, and discusses how those findings are relevant for development of a scene description system. Chapter 8 also provides conclusions and discusses some future work.

2 Related Work

This chapter presents literature in areas of research related to the problems and solutions dealing with the development of a non-visual navigation system called the Describer for Indoor Scenes (DISc). The work described in this thesis derives knowledge from various domains, such as, non-visual navigation, spatial cognition, experimental psychology, natural language processing, and robotic wayfinding. As non-visual indoor navigation systems and natural language generation systems are the broad fields associated with the research work presented in this thesis, this chapter concentrates on reviewing the literature related to those topics in order to present an idea of what has been previously done with respect to indoor scene descriptions.

2.1. Need for Navigation aids for Non-Visual Navigation

As illustrated in the sample task performed by Jack (a non-sighted person) and Mick (a sighted person) in Chapter 1, it is evident that there are many differences between navigating with visual cues and without visual cues. Ungar, (2000) reviewed a number of studies focusing on the differences in spatial processing capabilities between congenitally blind, early blind, and sighted people. In general, tasks like route finding are similar between groups, but there is more difficulty in more complex tasks like wayfinding and cognitive map development. Giudice & Legge, (2008) and Long & Giudice, (2010) suggest that this problem stems from lack of access to the environment, rather to any underlying cognitive deficit. If this is true, then providing spatial information, whether about an outdoor environment or a scene, should improve performance as it affords access to this critical spatial information. That is one of the

core rationale for this research, as these natural language descriptions are the easiest, most intuitive, and most parsimonious to produce and implement. While, other information is possible to get from existing assistive technology, e.g. outdoor GPS-systems and other technology that can provide access to the network nature of spatial structure. None of this technology describes scenes (which are the nodes in the spatial network), the output necessary for effective navigation and cognitive map development.

Is there any advantage of having visual cues in acquiring the spatial representation, which makes the task easier? Millar, (1994) and Thinus-Blanc & Gaunet, (1997) suggest the following advantages of having visual cues in the process of acquiring spatial representations:

- People who have access to visual cues have a coincidence of body-centered and external reference frames during locomotion, which makes the task of spatial learning easy.

- With vision, it is possible to look around (unlike the haptic sense, where the focus of interest is limited to proximal information at the point of touch). This access to distal information helps sighted people to connect spatially separated landmarks

- Visual cues are ubiquitous in nature with respect to all physical objects (as every physical object has visual components like shape, size and color, unlike auditory or olfactory cues)

- With vision, it is possible to see most things simultaneously, and quick eye movements allow you to capture distal information over a large field of view at an extended distance compared to its sister senses. With non-visual senses, information encoding is most often sequential, with information from touch being very proximal and all non-visual sensing being less spatially accurate than vision.

- Visual cues help in estimating direction and distance with the greatest accuracy, while sound and olfactory cues are often inaccurate and limited in specifying this information.

Also, results from various behavioral studies show that visual cues help in understanding and mentally representing the spatial structure of the environment more accurately than non-visual sensing. For example, (Casey, 1978) conducted a study where he asked blind and sighted students to create a tactile map of their school. The results showed that the tactile maps created by blind students were less organized and integrated, when compared to those tactile maps created by sighted students. (Golledge et al., 1996) later analyzed the tactile maps generated by non-sighted students and found that the blind students could not distinguish between curved paths and linear paths, which resulted in segmented tactile maps.

Apart from the problems developed as a reason of not understanding spatial nuances (for example, linearizing a curved path by non-sighted individuals), non-sighted individuals perform more cognitive computations and use more effort in spatial problem

solving as compared with sighted individuals. This problem of increased use of cognitive processes versus perception and the effortful nature of blind navigation compared to sighted people was discussed in (Rieser et al., 1986; Rieser, Guth, & Hill, 1982). Also Passini and Proulx conducted a study where they took non-sighted and sighted individuals on a guided tour of a two storied building (Passini & Proulx, 1988). After two such guided tours, both sighted and non-sighted individuals were asked to perform a navigation task. After their navigation task they were asked to create a model of the route they took to reach the destination while thinking aloud about their decision making process. The analysis of results showed that non-sighted individuals used significantly more decision factors demanding higher cognitive load during their way-finding process with a large number of tangible landmarks, when compared with the sighted individuals.

Hence non-visual navigation aids are being designed to assist blind or low-vision people to increase the accuracy of their spatial apprehension about their environment and to reduce their cognitive effort while performing spatial tasks. The next section of this chapter discusses more about the non-visual navigational aids that are commercially available.

2.2. Non-Visual Navigation Aids

Non-visual navigation aids can broadly be classified based on learning mode (real time or offline learning), based on sensory cues used (haptic or auditory), and type of environments traveled (outdoor or indoor). This section reviews the logic and

technology behind current non-visual navigation assistance systems and discusses how the indoor scene description system proposed in this thesis (DISc) fills the gap created by the current navigation aids available to blind and low vision people.

2.2.1. White Canes and Guide Dogs for Indoor Scenes

The white cane and guide dog are the traditional mobility aids used by non-sighted individuals. The main purpose of the white cane is to detect and avoid obstacles in the path of travel. Guide dogs also perform these tasks but also help their non-sighted partner to walk in a straight line from one point to another and warn of the presence of any elevation changes (like curbs and stairs), overhead obstacles, street junctions, unsafe walking conditions (see (Guide dogs for the blind.inc, 2012) for more information about the capabilities of a guide dog with respect to navigation).

Later, with the advancement in technology traditional canes were improvised by fitting additional sensors like SONAR, Gyroscopes, and ultrasonic sensors. Refer (Borenstein, 1997; Debnath, Hailani, Jamaludin, Syed, & Kader, 2001; Faria, Lopes, Martins, & Barroso, 2010; Grow, 1999) for more information about the advanced guide canes used by blind and non-sighted people.

Although independent mobility can be achieved by most non-sighted people, with the use of advanced white canes and guide dogs, there are some major disadvantages associated with them: they are restricted to avoiding only proximal obstacles; they will not help a blind person to learn an environment from the obstacles encountered, because they cannot give any useful information about the obstacle apart from merely

indicating its presence; canes and guide dogs will not determine a route for a non-sighted user to move from one point to the other. More importantly, both white canes and guide dogs can only solve mobility problems, not orientation problems, which deal with navigation decisions, wayfinding and cognitive map development (Giudice & Legge, 2008; Long & Giudice, 2010).

For the same reasons, when manually exploring an indoor scene, canes and guide dogs could merely be used to find obstacles. Also, they cannot ease the process of indoor scene exploration because of their inability to provide more information about encountered object types and the overall spatial representation of the room.

2.2.2. Can tactile maps solve Indoor scene knowledge acquisition needs?

While the use of canes and guide dogs are not sufficient for exploring an indoor scene because they cannot provide information about its global elements (for example, inter-object relations and architectural components) and they cannot provide overall spatial relationships between objects that are present in the room. On the other hand, maps are able to provide an overall view of a space in a desired scale and they can help to learn the overall spatial representation. Hence, non-visual maps could be a solution for the problem, which canes and guide dogs could not answer for a non-sighted person.

There are multiple ways to present maps to a non-sighted person. Loomis & Lederman, (1994) define haptic as one of those different ways to perceive non-visual spatial elements like maps and graphics, by using cutaneous perception and kinesthetic information for tactual learning.

Rowell & Ungar, (2003a, 2003b, 2003c) conducted an international survey about the usage of tactile maps and tactile map elements in geographical space. The survey showed that most of the hard copy tactile maps were designed and created to assist non-sighted individuals by providing access to visual map elements like road networks and landmarks that represent geographical space. From the survey results it could be seen that the majority of tactile maps have only been used for outdoor usage, although their benefits are just as relevant for indoor usage, even for the indoor scenes studied in this thesis. But there are three main reasons for not using tactile maps for learning indoor scenes in real time.

The first reason is that indoor scenes in general constitute a greater number of objects that are not fixed in position (like furniture). Hence even if a non-sighted person learns an indoor scene with these hard copy tactile maps, there is a high probability that the actual spatial arrangement of objects in an environment might not match the spatial representation as depicted in its hard copy tactile map rendering– since it takes only a little time and effort to physically modify the actual spatial location of objects by adding, removing or displacing furniture. Also based on the results of a survey conducted by (Jonathan Rowell & Ungar, (2005), it is evident that most of the tactile map users preferred to learn environments with tactile maps when they are at home, where there are limited distractions and few other things demanding their attention. Unlike for learning indoor scenes, this offline learning mode with tactile maps would benefit a non-sighted individual to learn about outdoor environments where there are more prominent and fixed landmarks or even in indoor spaces to learn building corridors,

entrances, exits and other architectural components, which are more prominent and fixed.

Cost is the second main reason for not using tactile maps to learn indoor scenes, as it would require more time and expensive equipment (for example see (www.viewplus.com) which would cost thousands of dollars, to design and produce hard copy tactile maps. For all the efforts spent on producing tactile maps, they are useful for only a limited period of time owing to the lesser prominent spatial elements present in an indoor scene. Hence this mandates the design and development of new tactile maps to be produced more frequently for an indoor scene.

The third reason is that it will be hard to provide information (like object type and dimensions) about every object present in an indoor scene, because of the limited variations that could be shown in tactile map symbols and labels. See (Rowell & Ungar, 2003c) for more information about the type of symbols and labels used currently while designing a tactile map and (Rowell & Ungar, 2005) for more information about current constraints and user preferences for tactile map labels and symbols).

All of the above mentioned reasons suggest that hard copy tactile maps do provide useful access to global structure, which is different and useful compared to the mobility information afforded by white canes and guide dogs. However, there are still problems mainly due to the wide variety of movable objects found in an indoor scene and to the expense of creating these static maps. Thus, tactile maps are not the ultimate solution for providing access to indoor scene knowledge to blind people, even if the information

they provide can be useful. What is needed is something that can provide this global information but in a dynamic manner, which is easy to update, inexpensive, and efficient.

2.2.3. Electronic Travel Aids for blind Indoor Scene Navigation

There are a variety of assistive technologies available on the market through which blind and low vision people can access indoor and/or outdoor maps. In most situations, electronic technology for rendering indoor and/or outdoor maps for blind people use haptic and auditory signals for presenting map elements. Force Feedback Devices and Touch screen devices are some off-the-shelf devices through which a non-sighted individual can access dynamic haptic and audio maps.

Specialty devices (hardware designed for specific purpose) are not discussed as many such specialty devices involve the purchase of expensive hardware and this thesis is interested in using off-the shelf, multi-purpose and less expensive interfaces for indoor scene knowledge acquisition.

2.2.3.1. Force Feedback Devices

Force Feedback devices are one of those widely used Haptic-based electronic technologies, which are used more by haptic researchers in general, but with a clear and obvious application to blind users for map learning. PHANToM Omni, a force feedback device, with its ability to provide six degrees of freedom in exploring maps is commonly used in research related to map exploration (see(Geomagic.Inc, 2012) for product specifications). By coupling PHANToM Omni with a computer interface, a blind user can

explore maps and landmarks by means of haptic and auditory feedback. For example (Kostopoulos et al., 2007), described a framework to generate haptic – audio representation of street maps from image analysis of visual street maps, in order to be used by blind people.

While most of the research was meant to learn outdoor maps, Yu & Habel, (2012) used a force feedback device (PHANToM Omni) coupled with speech and/or sonification cues to render the floor plan of small scale apartments. There are three main drawbacks of using force feedback devices coupled with auditory cues to provide indoor scene knowledge for blind users: Like hard copy tactile maps, this requires offline learning of indoor scenes which is not suitable for the need for learning indoor scenes in real-time (due to reasons mentioned in section 2.2.2 when discussing the disadvantages of hard copy tactile maps for learning indoor scenes). Second, indoor scene maps should be converted to a specific format which could be used by force feedback devices and there are no current technologies available to code the elements of indoor scenes and their spatial representation to a map that could be read using force feedback devices. The third main reason is that these haptic devices are not portable, which is one of the main reasons for them to be used in an offline mode. These represent some current restrictions limiting these devices for the rendering of indoor scenes for use in non-visual exploration.

2.2.3.2. Touch Screen Devices

Touch Screen Devices are currently dominating the field of mobile navigation assistance devices designed for blind and low vision people. A survey on the preferred methods for accessing spatial information suggest that blind users prefer using mobile phone based technology to assist their everyday way finding tasks and to learn spatial representation of environmental entities (Jonathan Rowell & Ungar, 2005). Smartphone and touch screen tablets are the preferred touch screen devices by the blind community for the following reasons: portability, multi-purpose nature, availability of different sensors (in-built Camera, GPS, accelerometer, Gyroscope, WiFi, and Proximity Sensor), different sensory outputs (haptic and audio). A number of research projects are being carried out to explore ways and means for presenting outdoor and indoor spatial knowledge with these devices. Haptic, Audio and Spatial language are three common mediums used to communicate spatial knowledge with touch screen devices.

Haptic and Audio cues are mostly used together as sensory inputs when touch screen devices are used to learn non-visual maps. For example, see research work by (Raja, 2012; Su, Rosenzweig, Goel, Lara, & N.Truong, 2010; Yatani, Banovic, & Truong, 2012) in which haptic and audio cues are used to provide spatial information about an environment. The above cited research work concentrates only on offline learning modes. Because of the reasons discussed in section 2.2.2, the offline-learning mode is less useful for learning indoor scenes.

2.3. Role of Language in Spatial Representation

Currently, there are a number of research projects testing the efficiency of using language as a communication medium to provide route and survey knowledge both for sighted and non-sighted people. For example, a behavioral study conducted by Loomis, Marston, Golledge, & Klatzky, (2005) evaluated both speech based guidance system and haptic information based guidance system in terms of their ability to support blind navigation. The results of their experiment suggested that a virtual speech based guidance system resulted in shortest travel times when compared with its haptic counterparts apart from receiving a higher subject preference rating. Also, it has been demonstrated that the spatial representations based on vision and language are functionally equivalent (Avraamides, Loomis, Klatzky, & Golledge, 2004), encouraging the use of language in a non-visual navigation assistance system. The reason behind this functional equivalence of vision and language is demonstrated based on the behavioral and neural findings, stating that spatial image is generally formed in supramodal areas of the brain which can integrate data from both language and vision (Struiksma, Noordzij, & Postma, 2009)

On the other hand, a survey conducted by Golledge, Marston, Loomis, & Klatzky, (2004), asked 30 blind people about their preferences for input and output modes while using non visual travel aids. Results from the survey suggested that blind participants preferred speech based inputs and outputs to be used in their non-visual navigation aids over all other perceptual interfaces using haptic and audio cues while navigating. Taken

together, these findings suggest that language is an efficient and reliable medium to communicate spatial information to blind users.

The cognitive load imparted during non-visual navigation when instructions are given as spatial language is known to be greater when compared with the cognitive load imparted during non-visual navigation when instructions are given as spatialized sound (Klatzky, Marston, Giudice, Golledge, & Loomis, 2006) . This could be because the language used for giving spatial information about environmental entities in all the research cited above (and in fact, all the other research works cited in this chapter), was developed just by following the syntactic and semantic rules to convey metric details like distance and angular information and the language of this type is called spatial language. They are less intuitive, mainly because they are not natural sounding and hence they are harder to comprehend unlike other perceptual senses. Even if the language is perfect, they would require more cognitive effort to interpret when compared with the perceptual senses.

To the best of my knowledge, there are no scientific investigations available to analyze the accuracy of spatial apprehension based on the cognitive standpoint of linguistic elements like description order, vocabulary of the language etc., (See chapter 3 of this thesis for more information about the discussion of linguistic elements). For example, there is no scientific discussion to analyze the qualitative aspect of a spatial mental map, with respect to the linguistic elements involved in generating it. Had there been such a discussion, it would be easier to develop a formal way to generate more intuitive and

naturally sounding language, generically called natural language. Hence the research work mentioned in this thesis analyzes the accuracy of spatial apprehension of indoor scenes from a cognitive standpoint of linguistic elements. This will pave way for developing a formal structure for generating natural language descriptions, especially in the context of indoor scenes.

Chapter 3 discuss language and the problems involved in generating natural language description systems for indoor scenes in more detail.

2.4. Chapter Summary

This chapter discussed literature and background of traditional non-visual navigation aids. Section 2.2 followed with a discussion of literature about modern electronic travel aids available for blind and low-vision people. In Section 2.3, I reviewed literature discussing the use of language as a communication medium in providing spatial knowledge to blind and low-vision people. From the literature reviewed in this chapter, we could see that natural language is an efficient communication medium to describe spatial relations in order to support non-visual navigation in unfamiliar indoor scenes.

3 Natural Language and Spatial Cognition

Natural Language is the most common and innate medium used by human beings to communicate with each other. It is intuitive, flexible, and versatile, but at the same time, it can be ambiguous and hard to interpret. The lexical richness of a natural language is enormous, enabling human beings to express their thoughts easily and effectively. Hence natural language is a widely used and preferred communication medium not only between humans, but also between humans and machines (Roush, 2010).

Chapter 2 we discussed the pros and cons of commercially available and prototypical non-visual travel aids, which could be used by blind people. It discussed the importance of using an intuitive communication medium to be used in those travel aids. This chapter presents reasons for considering natural language as the communication medium in the proposed scene description system (DISc).

3.1. Why Natural Language is used in DISc?

Both sighted and non-sighted people are accustomed to giving verbal route directions between spatially separated locations and to use those verbal route directions in their way finding behaviors, to virtually walk their listeners through an imaginary space with the help of spatial descriptions and to create a mental map of a space based on those descriptions. All the above-mentioned space related tasks suggest that it is not an uncommon notion to use natural language for sharing and acquiring spatial knowledge. Hence this thesis considers natural language as a natural and intuitive information

source for acquiring spatial knowledge about indoor scenes using DISc – A non-visual navigation assistance system.

The interesting part here is that the natural language descriptions used for acquiring indoor scene knowledge is to be generated by machines in the proposed scene description system (smartphones, in our case) and not by humans. Hence it is not only important to design machines which could generate naturally sounding indoor scene descriptions, but also to design machines which could generate indoor scene descriptions that are easily and accurately comprehensible by the intended end users. In order to teach our intelligent machines about the ways to generate accurate and intuitive spatial descriptions, it is important to first understand and analyze the ways in which people generally describe indoor scenes. That is, studying this issue will provide guidance for the fully automated scene description system that is eventually envisaged but it is not yet possible from current knowledge.

It is important to note here that DISc – a non-visual system is proposed to use natural language as a communication medium. One of the main goals of this thesis is to address the issue of how to best describe a space using natural language, as there is little research to provide guidance in this area. The most likely end-users are blind people, but the system is developed with an intention to support multiple user groups, including robots and people working in dark and zero visibility conditions. Hence in order to support all of the potential user groups, it is important to understand the ways in which humans generally describe space using natural language.

As discussed in the previous section, the versatile and flexible nature of natural language allows multiple ways to present spatial information. Spatial apprehension and cognitive mediation will not be the same across these multiple ways of presenting spatial information. The research work presented in this thesis is an attempt to identify the best way to present indoor scene descriptions through a non-visual system (DISc) to enable easy and accurate spatial apprehension of indoor scenes.

The following sections, review literature about current natural language generation systems, and their limited interest or awareness of human cognition. It also reviews the limited literature intended to strengthen the link between natural language and Spatial Cognition. Finally, this section describes the questions which have motivated this research and discusses the behavioral experiments conducted to find answers to these motivating questions.

3.2. Current Natural Language Generation Systems

Natural Language Generation is the process of generating natural language output from non-linguistic data sources. It is used in a wide variety of applications serving different user groups. This section discusses natural language Generation systems that are currently used by people in their day-to-day life. Although there is no working or prototypical natural language Generation system for converting scene related spatial data to a meaningful spatial description, there are a number of commercially successful natural language Generation systems that are used routinely. Reviewing these systems helps us to understand the basic design needs of DISc.

Following is an interesting natural language generation system proposed by Goldberg & Driedger, (1994), which is currently used in a Canadian weather service center. In this paper, the authors describe a weather forecast system called Forecast Generator (FOG), which takes non-linguistic data such as wind speed, atmospheric pressure, temperature and other meteorological parameters as inputs. The role of FOG is to convert the metric input data to a comprehensible weather forecast, using natural language as the output medium. It can be seen from their research that an important requirement for the development of the FOG system is to collect an extensive corpus of weather forecast generated by humans based on the analysis of non-linguistic meteorological data. The corpus is collected in order to understand the ways in which incomprehensible non-linguistic data could be expressed as meaningful natural language utterances.

There are several other natural language generation systems which use a corpus of textual inputs generated by humans as a prototype for generating natural language utterances (see works by (Buchanan et al., 1995; Iordanskaja, Kim, Kittredge, Lavoie, & Polgu, 1992)). These research projects suggest that in order to generate an indoor scene description system, as is the purpose of this research, it is important to first collect and analyze how people generally describe indoor scenes, so that the machine made natural language descriptions will be natural sounding and intuitive to the end user and that they will support safe and accurate spatial behaviors, e.g. spatial learning, navigation, and cognitive map development.

Reiter and Dale authored a book titled “Building natural language Generation Systems” primarily as an attempt to set ground rules for building a natural language Generation system. The importance of collecting extensive corpora from domain experts is stated clearly in the book (Reiter & Dale, 2000). They suggest that analyzing a corpus of natural language utterances obtained from domain experts is the best approach for acquiring knowledge about “soon-to-be-generated” natural language utterances. See (Ratnaparkhi, 2000; Williams & Reiter, 2003) for research work which supports these corpus based natural language generation methods.

3.3. Natural Language Understanding – With Space Sense

Although only a few researchers are dealing with the exact issue of scene descriptions that motivate this thesis work, there are a group of researchers who share common research interests dealing with understanding natural language utterances which were created with an objective to communicate spatial information to the listeners (for example, see (Johannsen, Swadzba, Aiegler, Wachsmuth, & Ruiter, 2013)) .

A behavioral experiment conducted by Taylor and Tversky shows that humans generally describe an environment as route descriptions, survey descriptions or sometimes mixing both route and survey descriptions (Taylor & Tversky, 1996). The primary difference between a route description and a survey description of an environment is that in a route description, the location of landmarks is defined with respect to the moving user, while in a survey description, the location of landmarks are defined from a single fixed perspective (Tversky, 1993).

Building on these findings, the indoor scene description system (DISc) should use either of these three perspectives (route, survey or mixed). In this section, I will review literature that are related to both route and survey descriptions to understand more about their usage.

3.3.1. Route Descriptions

There are a number of research projects trying to bridge the gap between Geographic Information Systems and Linguistics. One such research initiative has been conducted by Dale and his group with an intention to generate route descriptions by applying general natural language generation principles based on the input datasets obtained from commercially available Geographic Information Systems (Dale, Geldof, & Prost, 2005).

For people with little knowledge about their destination, route descriptions need not just be 'turn-by-turn' directions given to the listener. But the location of the destination could be explained to the listener by describing where the destination is located, instead of explaining how to reach the destination. This notion is proposed by (Tomko & Winter, 2009). For example in order to describe the location of something using a route description, one might explain, "A is in the city center, to the right of B". But DISc could not use this kind of route descriptions as the primary goal of DISc is to provide spatial knowledge of unfamiliar indoor scenes to its end user.

Although there are a handful of researchers with the intent of converting non-linguistic geographic data to comprehensible route descriptions, only a small subset of them are working on route descriptions with the context described below. A route description

constitutes a number of distinct spatial and linguistic components that could affect the spatial cognition of a way-finder. One such linguistic component is the way in which the description is divided into segments.

A cognitive analysis of the route directions suggested that the quality of a route direction is correlated with increased number of segments and increased detail about the turns of the route in the description (Lovelace, Hegarty, & Montello, 1999). Landmarks are another type of spatial component that influences the cognition of route apprehension. It has been shown that considering the structural salience of landmarks and formalizing the route descriptions by integrating the landmarks based on their shape and saliency will complement the cognitive apprehension of the route description and result in improved behavior or the representation of the route in memory (Klippel & Winter, 2005). Specifically considering the importance of landmarks (Hansen, Richter, & Klippel, 2006) proposed a data structure to understand the conceptual semantics of a landmark and thus enable facilitation of automatic generation of route directions. Although all the research projects mentioned in this section analyzed the relationship between linguistic elements and their associated spatial entities, they did not analyze how those linguistic elements affected the ways in which the listeners (who are the end users of the spatial description) actually perceived the spatial information conveyed through the description. By contrast, the research work presented in this thesis is an attempt to meet this need by understanding the effect of linguistic elements in perceiving spatial entities, from the standpoint of the listeners.

There is another group of researchers who were also interested to develop a complete framework for automating the process of generating route descriptions by considering the cognitive properties of the spatial and linguistic components. For example, (Denis, 1997) developed a framework to analyze route descriptions in order to understand how people verbally narrate a route. His analysis mainly focused on the use of landmarks and actions prescribed by the participant to follow the route. Another such framework was developed by Klippel and his group to generate route directions by chunking spatial elements of route directions, to enrich route directions with information about landmarks and to enhance descriptions about possible ambiguous spatial situations (Klippel, Hansen, Davies, & Winter, 2005). The frameworks mentioned in this section analyzed how processes underlying spatial cognition can be externalized through route descriptions. But as mentioned before, none of these frameworks were motivated by analyzing whether the spatial cognition acquired by the end users of route descriptions was affected by the linguistic component of the route description itself.

Also it is important to note here that the research projects mentioned in this section were dealing with outdoor environments and indoor corridor networks and not on indoor scenes. The main purpose of DISc, as discussed in section 1.3, is to provide a global overview of an indoor scene to the end user. But the main disadvantage of route descriptions as seen from the literature discussed in this chapter has clearly demonstrated that route descriptions were not associated with understanding the global structure of an environment. Since DISc is interested to provide global structure

of an environment to the end user, it cannot use route descriptions, and the associated formalizations, to describe an indoor scene.

3.3.2. Survey Descriptions

A Survey description of a scene, as used in the scope of this thesis, is conceptualized as the natural language description of the spatial representation of objects from a fixed perspective in an indoor scene. Thus, unlike a route description, survey descriptions can provide a global overview of an indoor scene to the end user, and this is why we adopted the use of survey descriptions in DISc. Unlike route descriptions, there is only a minimal amount of formal research being carried out in the area of Survey verbal descriptions, (see section 3.3.1 for more information about research projects related to route descriptions).

An important research study investigating survey descriptions was reported in (Ehrich & Koster, 1983), where the authors conducted a behavioral study in which they asked the subjects to describe the spatial location of objects in an indoor scene. They were interested to see how the spatial description of a room is organized and also how the linguistic components are distributed in natural verbal discourse. Their research has commonalities with this research focus. For example, they have shared interests in indoor scene descriptions, and analyzing those descriptions for understanding the pattern in which the spatial knowledge is arranged in the discourse. Ehrich and Koster analyzed the indoor scene descriptions collected in their experiment to understand the order of objects or path used for which the subjects described an indoor scene. As a

result they found that their participants either adopted a strategy that followed a round-about path or a parallel line path in an attempt to linearize the 3 dimensional room to a 1 dimensional natural language medium (see section 5.2.1 and 5.2.2 for more discussion about these strategies). They also analyzed the sentence patterns found in the discourse, based on the grammatical perspective, and concluded that the sentence and word patterns adopted depend on the type and arrangement of object clusters in the room.

Another relevant study was conducted by Linde and Labov (1975). In their study they asked their subjects to describe the floor plan of their apartment. There were no restrictions in the selection of description strategy (Linde & Labov, 1975). One of the description strategies extracted from the descriptions collected from their experiment is the Tour strategy, which is similar to a route description (discussed in section 3.3.1), where the describer took the listener on a virtual tour of their apartment. A map strategy is another description strategy that was used by participants while describing the floor plan of their apartment. In DISc, we are using the map strategy as the route strategy is considered to be inefficient for use for indoor scene descriptions due to various reasons (refer to section 3.3.1 for details).

More recently, a behavioral study conducted by Coventry and Tenbrink were interested to see what type of content and what type of objects were preferred for describing the spatial location of objects (called as reference objects) and what types of perspectives were used most commonly in natural language descriptions of indoor scenes. They also

were interested to see if the content, relatums and perspective selection were dependent on each other (Tenbrink & Coventry, 2011). The results suggested that the describers preferred an observer-based perspective and they also found that the choice of perspective was affected by the direction in which the objects were facing. It was also found that the scale, configuration and task at hand influenced survey descriptions.

A behavioral study conducted by (Xin Wang, Matsakis, Trick, Nonnecke, & Veltman, 2008), collected nearly 2000 spoken descriptions about scenes which were composed of abstract two dimensional objects. The collected descriptions were then analyzed to understand how human subjects describe the spatial relations that existed between those abstract objects and also they analyzed if there were any commonalities or variation in the selection of spatial relations for a specific object configuration. The results suggested that the spatial perception was the same across different subjects for a particular object configuration. Based on the natural language descriptions collected in the previous experiment, a framework was developed to extract spatial relations from natural language descriptions (Veltman, 2010). Again, all these frameworks approached the spatial descriptions from the describer's perspective and not from the user's perspective.

As can be seen from the various literature discussed in this chapter and also based on the literature search done in the background for this research (as discussed in chapter 2), it was observed that only a limited number of researchers are interested in understanding the ways in which a natural language survey description of an indoor

scene is organized. Regarding survey descriptions, there are many important questions that remain open. For example, “Is there a specific order that should be followed while describing an indoor scene in order to increase the accuracy of spatial apprehension?”, “Will the choice of metric units used in indoor scene descriptions affect distance and direction estimation of the listener?” and so on.

It is important to mention here that scene description is a generic term used widely among the research community. Although there are a group of researchers who are working on indoor scene descriptions, there are another group of researchers who are interested in natural language description of scenes with respect to events in a video sequence (as in a game scene, for example, see (Bergen, Lindsay, Matlock, & Narayanan, 2007; Herzog et al., 1989) in which the term “scene description” means the verbal description of a video scene). This thesis is interested only in static indoor scenes and not on a video sequence.

3.4. Indoor scene Description in Natural Language

Starting from Chapter 2 until section 3.2.2, I have reviewed various research literature discussing non-visual navigation, natural language, and spatial cognition, and their relationship to each other. In this section, I will present a brief summary of the reviewed literature. I will then delineate the gaps that exist in current non-visual indoor navigation systems. Finally I will describe the research goals of this thesis, which are aimed at filling the gaps that exist in current non-visual indoor navigation systems.

We have discussed that current navigation assistance systems for blind people are not functionally complete because of the absence of an indoor scene description and navigation assistance system. Hence, as an attempt to fill the gap, a non-visual indoor scene navigation assistance system called DISc (Describer for Indoor Scenes) is proposed in this thesis. We proposed to use natural language as an interface for DISc because it is innate and intuitive to humans. At the same time, it has several complexities because of its flexible and versatile nature. This means that verbal descriptions could be narrated in multiple ways to describe a spatial representation of the scene. But the spatial apprehension and cognitive mediation will not be the same across these multiple ways of presenting a spatial representation. Hence in order to overcome these complexities, linguistic elements in an indoor scene description should be understood completely from the standpoint of an end user in order to make DISc a more efficient scene description system. Hence this thesis studied the relationship between linguistic elements of an indoor scene description and its effects in the end user's spatial apprehension.

3.5. Motivation

Based on my literature review from Chapter 2, it can be seen that the majority of natural language related research (dealing with cognitive analysis) has analyzed linguistic elements from the standpoint of its describer, the ways in which the discourse is structured and the factors influencing the structure of discourse. The drawback with this type of approach to a natural language description is that the end users for who the descriptions were generated were not considered. That is, little emphasis was given for

evaluating listener's spatial apprehension when using the descriptions in order to understand the effects of linguistic elements and discourse structures on spatial cognition. The absence of this approach has resulted in natural language descriptions that are syntactically and semantically correct, but whose purpose will not be met if the end user's find it hard to apprehend the spatial information conveyed in the descriptions. From the existing literature, only a little information is available on whether the language components of spatial descriptions affect the spatial cognition of the end user. But for constructing an applied scene description system like DISc, it is important to explicitly consider the perspective of the end user. Hence this thesis is motivated to approach the generation of spatial description from the perspective of the end users of those descriptions. Thus, the research work presented in this thesis is aimed at gaining knowledge about the effects of linguistic components like the sentence structure and information content on the spatial apprehension acquired by end users of indoor scene descriptions and if these descriptions support accurate spatial behavior. This knowledge will help us to generate better indoor scene descriptions through DISc in order to support the most accurate spatial learning and behavior possible based on using purely verbal descriptions.

It is important to note here that the listener's knowledge needs about the space is as important as the describer's notion of space. So far the research projects mentioned in this thesis discussed space only from the perspective of a describer. But it is equally important to conduct behavioral experiments to understand how to make it easier for listeners to understand space when visual information is not available.

There are different aspects of an indoor survey description that could affect spatial cognition of blind users. To name a few, vocabulary knowledge of the describer and the listener, size and type of room, discourse strategy, reference frames used in discourse, requirements of mental rotation, spatial updating capability, granularity of spatial information, metric units used in discourse, all may or may not affect the listener's spatial apprehension based on the survey description of an indoor scene.

Although there are different aspects of an indoor scene description that could affect listener's spatial apprehension, only a few of them showed greater variability across the corpus of indoor scene descriptions collected as a part of this thesis (refer to chapter 4 for information about the scene description corpus). Hence the research work described in this thesis concentrates only on studying the effects of the following linguistic elements of a survey description of an indoor scene,

- Discourse strategy
- Metric units of angular information
- Effect of mental rotation and/or spatial updating process

Six behavioral experiments were conducted in order to understand how the above listed linguistic elements affect an end user's spatial cognition about the scene being described in the verbal description.

3.5.1. Direct Observation Versus. Photographic Observation of an indoor scene.

The photos that are taken using smartphones were used as the basis of spatial information about objects in the indoor scene (see figure 1.1 for an image depicting the

scope of DISc). In other words, visually impaired users will be using these photos as a substitute for their vision in order to obtain information about the spatial configuration of an indoor scene and the location of its constituent objects via natural language descriptions. Photos taken using smartphone cameras will have some limitations based on the color sensitivity of image sensors, stereoscopic property of the sensor and also because of the limited field of view of the lens being used. Hence it is important to compare the spatial information that is obtained from photographic observations of an indoor scene against the spatial information that is obtained from direct observations of the same scene. This comparison is done to evaluate whether the limitation of photographs leads to the exclusion of important environmental details in the ensuing spatial verbal descriptions. A behavioral experiment was conducted to evaluate whether there is a significant difference between the observation modes by comparing the accuracy of scene recreation based on previously generated scene descriptions from both direct and photographic observation modes (Refer to Chapter 4 for details).

3.5.2. Comparing Description Strategies

Flexibility is one of the most important features of a natural language. An important challenge owing to this flexible nature is that natural language descriptions about the spatial location of objects in an indoor scene could be structured in different ways following different strategies as discussed in (Ehrich & Koster, 1983). For example, a description could begin by describing the name and spatial locations of objects in one corner of the room and then follow a cyclic clockwise strategy of describing the other objects around the room. In another case, a description could combine objects based on

their functionality, e.g. describing the spatial location of all the tables that are present in a room, then describing the chairs, etc.

It is important to first identify different scene description strategies which people might use if they have to describe an indoor scene from a fixed perspective and then to determine which description strategy will lead to the most accurate spatial apprehension in the end user with limited cognitive effort required for the process.

The indoor scene descriptions, collected from the behavioral study to compare the photo and direct observation modes, were analyzed to understand different types of description strategies used by humans (refer to chapter 4 for more information about the study). This was followed by experiment 4 on 'evaluating description strategies', which was another behavioral study conducted to evaluate different description strategies used by humans. This strategy evaluation study was conducted in order to select the best indoor scene description strategy to be used in DISc. The evaluation of description strategies was done based on their ability to help the end user in order to gain the most accurate spatial information to support spatial learning, spatial representation in memory, and behavior in the space (i.e., navigation) (refer to Chapter 5 for more details).

3.5.3. Experiment on Evaluating Presentation units of Directional Cues

For a linguistic scene description to be on par with that derived from visual perception, it is extremely important to have an accurate method for specifying directional cues about the spatial locations of objects within the scene. As with scene description

strategies, there are different ways to verbally present these directional cues to the user. The behavioral study conducted by (Ishikawa & Kiyomoto, 2008) suggests that people best understand directional cues when they are presented using relative directions rather than using only absolute directions. However, there is no formal research investigating the best way to present directional cues with the highest precision within a relative reference frame.

Degree measurements and clock face directions are the most common ways to present angular information using a relative frame of reference. For example, “a desk is at your 1 o’ clock position” and “a desk is at 30 degrees on your right” both specify the same spatial location of the desk. But it is important to know which of these presentation methods leads to the most accurate perception of directional information. To address this question, a behavioral study was conducted to compare the accuracy of angular perception based on these two types of directional cues (see chapter 6 for details).

3.5.4. Comparison of Spatial Behavior Experiment

The comparison of observation modes experiment, comparison of description strategies experiment and the direction estimation experiment were conducted to analyze the effect of important linguistic elements found in indoor scene descriptions. Even if we optimize indoor scene descriptions based on the analysis of the experiments mentioned in this section, it is important to evaluate whether they are able to support continuous spatial behavior within an indoor scene.

Let us consider an indoor scene description of an office room, with an assumption that the listener is standing at the doorway of that room. The listener might need to plan a navigation task from desk, while he had access to the spatial description narrating the scene from the viewpoint at the door. If the listener had to plan a navigation task from the door, then his perspective coincides with the indoor scene description (updated description condition), while in this case it is not as he is standing at the desk and learned the room from the viewpoint at the doorway. In this condition the listener had to plan the navigation task based on a description that was not updated depending on his position and orientation (static description condition). When the perspectives of the listener and the indoor scene description are the same (that is, while using updated descriptions), there is no need for the listener to perform mental rotation and spatial updating processes to understand the scene. But when the perspectives are different (that is, while using static descriptions), then the listener has to mentally rotate and update his position and orientation as they move to new locations in the scene.

The presence or absence of spatial updating and mental rotation required by an indoor survey description might affect the cognitive spatial apprehension and navigation accuracy of the end user. A behavioral study was conducted to compare and analyze the navigating and updating performance of subjects when using updated and static descriptions. The interest here is to see whether dynamic descriptions (updated) lead to more accurate performance than static descriptions. Another goal of this experiment is about testing real spatial behaviors based on the descriptions, as the previous

experiments did not test actual navigation behavior based on the developed cognitive map.

Another experimental condition included in the study, in which the subject was asked to determine the spatial location of objects followed by an exploration task of the room without visual cues and any external aid, was called as self-exploration mode (amodal learning). This self-exploration mode serves as the proof of concept that the non-visual scene description system DISc, based on information from the other experiments, works and supports spatial behavior (refer to chapter 7 for more details). This is a control condition that provides no descriptive information about the room. The interest is to show that DISC actually provides useful information compared to this control condition in terms of learning time, cognitive map development, as assessed by the navigation tasks.

3.6. Chapter Summary

This chapter began by discussing various reasons to support our decision to use natural language as an interface in our proposed non-visual indoor scene description system (DISc). Following that, the chapter reviewed research projects discussing current natural language generation systems. From the discussion, it can be seen that indoor scene descriptions have not been considered in detail by any of the existing research literature. Also to my knowledge, based on the literature search in chapter 2, only few research in this domain analyzed listener's spatial apprehension of an indoor scene in order to evaluate linguistic elements of spatial descriptions. Hence, I described how the

research work presented in this thesis uses a different approach from the traditional means of evaluating spatial descriptions and why this approach is important. Finally, the chapter discussed the behavioral experiments that were conducted as a part of this thesis.

4 Comparing Observation Modes

In the previous chapters, I have reviewed the advantages of using natural language as an intuitive medium to be used by blind people for gaining access to information about their environment, e.g. indoor scenes. Also, I discussed the reasons for using smartphones as a core interface in a scene description system. In our proposed natural language Description system - DISc, photos that are taken using smartphones will be used as the basis for acquiring spatial information about objects in an indoor scene. That is, visually impaired users will be using these photos as a substitute for their vision to obtain spatial information of key objects and landmarks present in the scene via natural language descriptions. The primary concern I had about using photos as a source for generating natural language descriptions is whether functionally equivalent scene descriptions could be elicited between direct and photographic observations of an indoor scene with respect to the amount of information being collected.

Hence this chapter, describes the first behavioral experiment which compared scene apprehension performance obtained as a result of natural language descriptions generated by photographic and direct observation of an indoor scene.

4.1. Motivation

I have reviewed the advantage of visual cues in spatial perception in section 2.1. In this section, I will review the literature related to the pros and cons of direct observation and photographic observation of an indoor scene and the motivation for comparing these two observation modes.

4.1.1. Color Sensitivity

Human vision is remarkably good in constructing visual representations based on the dynamic range of colors; also, spectral variation in illumination will not affect the visual representation developed by human vision (Cornsweet, 1970). On the other hand, when photographing a scene using a camera, the clarity of information could be lost as a result of the influences of shadows. There is also a possibility that chromatic details could be lost or vary from direct observation because of the distance between the light source and the presence of the camera. This loss in chromatic information could also happen if the dynamic range of the recording medium (i.e., camera sensor) is not sufficiently efficient to capture the dynamic range of the scene being recorded (Jobson, Rahman, & Woodell, 1997). These factors could, in theory, reduce the information content when a scene is recorded as a color composite photograph.

4.1.2. Stereoscopic Vision

Another advantage of direct observation of an indoor scene is the presence of stereo vision while perceiving the scene. Humans are equipped with a pair of eyes located side-by-side, so that the retina in each eye records the same scene from two slightly different angles. This results in stereoscopic vision which helps in understanding the 3D information of the objects being perceived in the scene (Eimer, 1996). On the other hand, images captured with the smartphone cameras use only a single sensor to capture the scene resulting in a 2D image. This reduced dimensionality will affect depth perception and may result in poor distance estimation of objects that are present in the

indoor scene. This difference in dimensionality from photographic observation of a scene could affect the cognitive model being perceived.

4.1.3. Field of View

Field of view is the area of examination captured by the sensor and it is generally determined by the focal length of the optical system. Humans have an average focal length of 17 mm (Serway & Jewett, 2000), while advanced smartphone cameras generally take pictures with the focal length in the range of 35 mm (Apple, 2012) . It is important to remember that focal length and field of view are indirectly proportional. Therefore, the field of view covered by a smartphone camera is significantly less than the field of view covered by the human eye.

When the focal lengths of the human eye and smartphone sensors are converted to their field of views, they are approximately 200 degrees and 60 degrees respectively. In order to match this difference in field of view, a panoramic photograph of the indoor scenes used in the experiment were made by stitching 3 images together. Although the Field of Views were matched, the look and feel of the panoramic photograph is not as natural as regular photographs. This is mainly because with direct observation the overall structure of the room is obtained by moving the head and eyes, while the panoramic image shows a composite of multiple images in a single view. This also introduced aberrations on the geometric properties of objects. Apart from that there were image aberrations, which were created as a result of stitching the images together in order to match the field of view of the human eye.

In summary, the difference in color sensitivity, stereoscopic properties of vision, and field of view restriction could influence the difference between a photographic observation and a direct observation of indoor scenes. Acknowledging these limitations, the results of an informal analysis of the panoramic photographs of the indoor scenes used in these studies showed that accurate spatial information about its constituent objects could be readily extracted and that the information available in these photographs was subjectively equivalent to what could be directly perceived.

As DISc is proposed to use photographs to collect spatial information for generating indoor scene descriptions, it is important to further investigate whether the spatial information obtained from photographs is qualitatively on par with the spatial information obtained from direct perception. Hence it is important to compare the functional equivalence of spatial information obtained from photographic and direct observations of the same indoor scenes via formal empirical investigations.

Thus, a behavioral experiment was conducted to investigate the relation of these observation modes.

4.2. Pilot Study

A pilot study was conducted to analyze the effectiveness and practicality of the experimental design used to evaluate the functional equivalence of scene perception between the observation modes in terms of how accurately the scenes could be verbally described and subsequently re-created on the basis of this description.

4.2.1. Method

The pilot experiment was divided into two parts: a scene description phase and a scene recreation phase. Two separate groups of subjects participated in the pilot study, one for each phase.

4.2.1.1. Pilot Description Study

The experiment was designed in a way such that it motivates the subject to think of a situation in which he/she is describing an indoor scene to a blind friend over the phone. The description was recorded using a Zoom H2 Handy Recorder (www.zoom.co.jp). The recorded descriptions were subsequently transcribed to text for analysis.

4.2.1.2. Pilot Recreation Study

The indoor scene descriptions transcribed from the first phase of the study were given to a new group of subjects who had never seen the indoor scene that was being verbally described. They were then asked to recreate the indoor scene using 3D architectural rendering software called RoomArranger (See section 4.2.3 for more information about the software).

4.2.2. Participants

As mentioned in the last section, two different groups of subjects participated in the pilot experiment. 2 sighted native English speakers participated in the scene description phase (2 male subjects, mean age = 21.5) and 8 sighted native English speakers participated in the scene recreation phase of the study (4 male and 4 female subjects,

mean age = 25.5). All participants had normal or corrected to normal vision (with visual acuity of 20/20) in both eyes.

The study was approved by the Institutional Review Board (IRB) of the University of Maine and on average, each subject took between 30 & 45 minutes to complete the task. The participation of all subjects in the experiment was a voluntary decision made by them and every participant signed informed consent forms stating this right. The subjects were monetarily compensated for their time and effort to participate in the experiment.

4.2.3. Materials and Apparatus

All the experiments mentioned in the research work carried out for this thesis were conducted in indoor office scenes. Prior to setting up prototypical office rooms, I visited a number of different indoor scenes, such as doctor's offices, professor's offices, Student's offices, Hotel Lobbies, and administrative assistant's offices, and made a list of common objects that were present in each scene. Followed by this cataloging of prototypic scenes, I setup 2 different office rooms with the objects that were commonly present in all the office rooms I had canvased. Both of the experimental office rooms included objects of the same kind and number, but that were arranged in a different spatial configuration. That is, in each office room, there were 2 bookshelves, 2 file cabinets, 3 chairs, 3 tables, 1 trashcan, a door and a window. Only the spatial arrangement between the objects varied between these two rooms (total = 13 objects in each room).

I took photographs of both the office scenes with the help of a Nikon D3100 camera (with a focal length of 30 mm). The camera was mounted on a tripod, and the setup was placed at the center of the doorway of the office room. I also set the camera sensor at a height of 170 mm from the floor by adjusting the height of the tripod. This height matches the average height of males and females in the united states (Ogden & Flegal, 2008). Three images were taken for each room and then stitched together to form a panoramic view of the room in order to match the regular field of view perceived from direct visual observation from the same perspective.

For the initial scene description study, I used a Zoom H2 Handy Recorder to record the audio descriptions provided by each participant. The audio descriptions were recorded in a 2 channel recording mode and all the subjects held the microphone approximately 0.5 feet from their mouth while describing the two indoor scenes.

For the subsequent scene recreation study, I used a 3D architectural rendering software called Room Arranger (Version 3.2) (see (www.roomarranger.com) for more information about the software used in this recreation study). It is an easy-to-use room arranging software primarily used to design and view a room and is mainly used by architects. Every subject who participated in the experiment watched an introductory tutorial video about the software from which they learned how to add and manipulate objects in the virtual room using the application.

4.2.4. Procedure

4.2.4.1. Pilot Scene Description Study

There were two observation modes, namely, a direct observation mode and a photographic observation mode. In the direct observation mode, the subject observed the office scene while standing at the doorway of the room. In the photographic observation mode, the subject observed the room from a photograph.

I read aloud the following experimental instructions to each subject, ensuring that every subject received the same explanation and information content.

For the direct observation mode: "*You will be taken to an office room. Your task is to describe the office space as clearly and accurately as possible. In your description, you should include the objects that you think are important and describe the relationship between those objects. If you see two objects of the same kind, for example, if you see two tables, make sure that you use a clear way to address the two tables in a distinct manner. Also, please remember that you need not describe every single object in the room. For example, you need not describe the number of books or the number of shelves in a bookshelf, but you should describe the spatial location of the bookshelf in the room or its relation to other objects. If your description is not clear or ambiguous, I might stop you in the middle to clarify your description in order to improve its accuracy*".

For the photographic observation mode, the first sentence of the direct observation mode of the script was modified as *"You will be shown a photograph of an office room",* while the other instructions remained the same. It is important to mention that the

instructions did not specify the granularity with which the room should be described and at the same time, not give the subjects a path that they should follow while describing the scene. The subject was completely independent to choose the description strategy.

As mentioned in section 4.2.3, I used 2 different office rooms for the study. The design of the study was completely within subjects, where each subject described both the rooms based on both the observation modes. The subject distribution was counterbalanced across the two rooms and observation modes.

There were only 2 subjects in this pilot study. So at the end of the description study, 4 descriptions in total were collected, 2 per room and 2 per observation mode.

4.2.4.2. Pilot Scene Recreation Study

The audio descriptions collected in the afore mentioned description study were transcribed to textual descriptions. Only those sentences that were complete were transcribed. The design of the study was within subjects, where a subject recreated 2 descriptions, one for each room (room 1 or room 2) and one for each observation mode (direct observation or photographic observation) and these were counterbalanced.

Each subject watched the tutorial video in order to learn how to use the “Room Arranger” software. They were asked to practice with the software by recreating the room in which the experiment was conducted. Once the subject was confident about using the software, they were given the first description to recreate.

Experimental instructions were read aloud to each subject in order to ensure that every subject received the same instruction to perform the task. "I will give you a description of an indoor office scene. Your task is to read that description at your own pace. Once you feel that you are ready, you should give the description back to me and then start recreating the indoor scene that was described in the text from your memory".

4.2.5. Results

The recreations of the rooms made by subjects were saved using the (.rap) file format. Later, for each object in the recreation, I exported the information about the x and y co-ordinates, length, width, height and orientation (i.e. the angle to which the object is rotated) to an excel file.

I recreated both the office rooms based on the actual dimensions of the objects using the room arranger software. Then I extracted the x and y co-ordinates, length, width, height and orientation data of every object in the room that was recreated with the actual dimensions of those objects and exported the data to an excel file.

4.2.5.1. Scaling Error

Scaling error was calculated using the formula mentioned below.

Scaling error = ((Abs (Actual length-Recreated length)/Actual Length) + (Abs (Actual Width–Recreated Width)/Actual Width)/2)

Scaling error was calculated for the entire 13 recreated objects that were present in the room. There were 8 subjects who recreated the room, based on 2 observation modes.

Hence there should have been 208 data points in total (104 for photographic observation mode and 104 for direct observation mode). But there were 8 objects (4% of the data), which were not recreated, by subjects and these were removed from the analyzed data set. Hence the dataset was composed of 200 data points. Also, the scaling errors of 4 objects (2% of the data) were more than 2.5 times the standard deviations from the mean and were thus categorized as outliers. The outliers were then replaced with the corresponding subject's mean scaling error.

The arithmetic difference between scaling error for both direct observation and photographic observation modes were normally distributed, as assessed by visual inspection of a Normal Q-Q Plot.

The scaling error of the objects calculated by Euclidean distances was higher for recreations based on the direct observation mode (M = 0.29 feet, SD = 0.14) when compared with the recreations based on the photographic observation mode (M = 0.31 feet, SD = 0.14). Also, the direct observation mode increases the positional scaling error by 0.01 feet on average, with 95% CI [-0.01, 0.46]

Although there were numerical differences between the two observation modes, the scaling errors were quite small and the differences between observation modes were not statistically significant, $t(95) = 0.9$, $p = 0.2$ (>0.05). Thus, results from the pilot data suggest that the type of observation mode (photographic vs. direct perception) does not have any influence in the listener's representation of object dimensions that were built up from learning using verbal scene descriptions.

4.2.5.2. Positioning Error

Positioning error was calculated as the Euclidean distance between an object's actual position in the room and its recreated position by the participant. The formula used for calculating the Euclidean distance of an object is as follows,

Euclidean distance = sqrt((Actual x-coordinate of an object – Recreated x co-ordinate of an object)^2+(Actual y-coordinate of an object – Recreated y co-ordinate of an object)^2)

The Euclidean distance difference was calculated for every object in both the rooms across both the observation modes. It is important to note here that the coordinate system of the ideal recreation of the room is the same as that of the coordinate system of the room which was recreated by subjects.

There were 8 objects (4% of the data) which were not recreated by subjects and were thus removed from the data set, leaving 200 data points that were analyzed. Also, the positioning errors of 3 objects (1.5% of the data) were more than 2.5 times the standard deviations from the mean and were categorized as outliers. The outliers were then replaced with the corresponding subject's mean positioning error. The Euclidean distance differences between the observation modes were normally distributed, as verified by visually inspecting the Normal Q-Q plot.

The positioning error of the objects calculated by Euclidean distances was higher for recreations based on the direct observation mode (M=4.6 feet, SD = 3.7) when compared with the recreations based on the photographic observation mode (M = 3.8

feet, SD = 3.3). Thus, the direct observation mode increases the positional error by 0.802 feet on average, with 95% CI [-0.8, 1.6]

Although there are numerical differences, the positioning errors were very small between the observation modes and this difference was not statistically significant, $t(95) = 1.7$, $p = 0.07$ (>0.05). Thus, results from the pilot data suggest that the type of observation mode does not have any influence in the listener's understanding about the spatial representation of object locations in an indoor scene, which was built up from learning using scene description.

4.2.5.3. Orientation Error

The Orientation of an object is the direction in which the object is facing. Orientation error is determined by the angular difference between the actual object's orientation and the recreated object's orientation and is calculated as follows.

Orientation Error = Abs (Actual Orientation – Recreated Orientation)

There were 8 objects (4% of the data) which were not recreated by subjects and were removed from the data set. The dataset was supposed to have 200 data points. But the trashcans in both the office scenes were circular and hence there could be no orientation information that could be extracted. Hence, there were only 184 data points analyzed. Also, the orientation errors of 7 objects (3.5% of the data) were more than 2.5 times the standard deviations from the mean and were categorized as outliers. The outliers were then replaced with the corresponding subject's mean orientation error.

The orientation error differences between the observation modes were normally distributed, as verified by visually inspecting the Normal Q-Q plot.

The average orientation errors of the two observation modes differ by 4 degrees (direct observation mode (M=15.9 degrees, SD = 8.44) and photographic observation mode (M = 11.9 degrees, SD = 6.4)). Although there are numerical differences, the orientation errors were quite small and the differences were not statistically significant between the direct observation mode and photographic observation mode, $t(91) = 1.7$, $p = 0.3$ (>0.05). Thus, results from the pilot data suggest that the photographic and direct observation modes do not have any influence in the listener's representation of an object's orientation in an indoor scene.

4.2.5.4. Object Retention error

Object retention represents the number of objects that were recreated in each room. It was found from the data that recreations based on the direct observation mode had 12.75 objects per room on an average (where there should actually be 13 objects per room), while photographic observation of the room showed an average of 12.25 objects per room. This difference in object retention was not statistically significant based on the paired t-test results ($t(96) = 0.7$, $p = 0.5$ (>0.05)). Hence we can conclude that the results from pilot data suggest that observation modes do not have any influence in the listener's object retention ability.

4.2.5.5. Topology Score

The recreations were also analyzed based on their resulting topological structure. First I divided the actual room into 2-foot virtual square boxes (which constitutes peripersonal space – the space which could be accessed by the limb of a person when reaching) , and made a note of the objects that were present in each of those square boxes. For example, I made a note that Box 1 contained Table 1 and Chair 1. In the same way, I divided the recreated rooms to 2-foot square boxes and made a note of the objects that were contained in each of those boxes. Then I scored each box in the recreated room by comparing it with its corresponding box in the actual room. For example, if Box 1 of a recreated room has Table 1 and Bookshelf 1, then it received the score of 0.5 (because box 1 was supposed to have 2 objects and one of them (Table 1) was correct, while the other object (Chair 1) was not present in that box. Hence it received the score of 0.5. It is important to note here that if an object that is supposed to be present in a box is not there, and then the points are reduced for that box, while if an additional object is present in the box then the score for the box is not affected.

Hence each box can get a maximum score of 1 (when all objects that are meant to be in the box are present) and a minimum score of 0 (when none of the objects that are meant to be in the box are present). And the scores of every box in a recreation were summed up to get the overall topological score of the room.

It was found from the data that recreations based on the direct observation mode had an average score of 28.59 (where the ideal score for each recreation is 32, as there were

32 boxes), while photographic observation of the room had an average score of 26.1. This difference in topological score was not statistically significant based on the paired t-test results ($t(7) = -2.42$, $p = 0.6$ (>0.05)). Hence we can conclude that the results from pilot data suggest that observation modes do not have any influence in the user's apprehension of topology of an indoor scene during a reconstruction task.

4.2.6. Discussion

The data collected from the recreation study was analyzed by calculating the scaling, position, orientation, and object retention error of the recreated objects, and the overall topology scores of recreated indoor scenes. The errors were then analyzed across the photographic and direct observation modes used to describe the indoor scenes.

It can be seen from the null results of this pilot study (e.g., no reliable differences between observation modes) that the type of observation modes used to generate indoor scene descriptions likely do not affect the listener's spatial apprehension and subsequent mental representation of the environments. Hence, the results from this study support the use of photographic images as an input to DISc, which are then used as the basis of generating indoor scene descriptions.

But, I was concerned regarding the experimental design used in this pilot study. In the scene recreation task, subjects recreated the room from memory. Hence the recreations were based on memorial representations of the scene descriptions. However, I was

interested in the spatial perception based on their apprehension and learning from the descriptions, not how well they were remembered.

It is important to note here that memory is an important factor, which could affect spatial cognition, and this has been discussed a lot in the scientific literature. For instance, spatial relations recreated from memory are reported to violate the Euclidean geometry and this violation is mainly influenced by the nature of landmarks included in the memorial representation of the spatial relation. For example, in Euclidean geometry, the distance from point A to point B is the same as the distance from point B to point A. But it is been shown that subjects tend to violate this Euclidean concept in estimating distances when the estimation has to be done from memory (McNamara & Diwadkar, 1997; Newcombe, Sandberg, & Johnson, 1999; Sadalla, Burroughs, & Staplin, 1980). Also when recreating from memory, a bias in the distance is often seen when the objects are functionally related to each other (Hirtle & Mascolo, 1986). Also the results of their study showed that participants often added objects that were not present in the actual scene. This is because of the background knowledge of the participant about the object and the nature of objects with which it is spotted frequently (Bartlett, 1932). An example of participant's background knowledge affecting the spatial perception is clearly seen in the current recreation data as participants added a chair in front of all the 3 desks that was present in the room, where only 2 desks in the room actually had a chair in front of them. At this point, I also was worried that memory would have affected the distance information extracted from the data for calculating the positioning error and scaling error of recreations. Another important reason to not rely on the

results suggested by the data of this pilot experiment is the small sample size of participants.

Hence, another experiment (as described in the next section of this thesis) was conducted, with a few key changes introduced in the experimental design to overcome the problems just discussed with the pilot description and recreation study.

4.3. Experiment 1 – Scene Description Study

In this study, I dealt with the concerns of the pilot study by recruiting a larger subject sample.

4.3.1. Experiment design

The experiment design was the same as that of the previous pilot description study and it was designed in a way such that it motivates the subject to think of a situation in which he/she is describing an indoor scene to a blind friend over the phone. The descriptions were recorded using a digital recorder and were subsequently transcribed into text.

4.3.2. Participants

12 sighted native English speakers participated in this study (6 male subjects and 6 female subjects, mean age = 20.7). The study was approved by the Institutional Review Board (IRB) of the University of Maine and on average each subject took 30 to 45 minutes to complete the task. The participation of all subjects in the experiment was a voluntary decision made by them and every participant signed informed consent stating

this right. The subjects were monetarily compensated for their time and effort to participate in the experiment.

4.3.3. Materials and Apparatus

The materials and apparatus were the same as were used in the pilot description study. (refer to section 4.2.3 for more details). Following are the panoramic pictures of the office scenes that were set up for this study.

Figure 4.1 Panoramic photograph of room 1

Figure 4.2 Panoramic photograph of room 2

4.3.4. Procedure

The procedure of this study was the same as the pilot description study described in section 4.2.4, where the participants were asked to describe an indoor scene based on two observation modes namely a Direct observation mode and a Photographic observation mode. The subject distribution for this study followed a within-subjects design and the order of observation modes was counterbalanced.

4.3.5. Results

In total, 24 descriptions were collected from the 12 participants,– six each for both observation modes and office rooms. These descriptions were later used in Experiment 2, where their syntactic and semantic correctness with respect to the actual spatial parameters of the physical indoor scenes were evaluated.

4.4. Experiment 2 -Inter Subject Rating Study

From the scene description study elaborated in section 4.3, 24 audio descriptions were recorded and then transcribed to textual descriptions. Only those sentences that were complete were transcribed. These descriptions were then used for the scene recreation study mentioned in section 4.5 (similar to the pilot recreation study discussed in section 4.2.4)

But logistical design issues arise if we have to recruit subjects to recreate 24 descriptions. Since a participant can only recreate 2 descriptions (1 description per room), in order to avoid any learning effects about the rooms, approximately 50 subjects should be recruited for the recreation study in order to get a data set based on all possible recreations.

Besides the time issue, there is another problem as we only have 2 rooms and 12 descriptions for each room. Hence it would be impossible if we want to use a within subjects design, which is the preference as this model is able to reduce error based on individual differences. On the other hand, besides increased subject error, if we are using a between-subjects design we would be introducing another factor besides the 2 observation modes, as the descriptions themselves were different. So the recreation would also depend on individual differences found in the descriptions, in addition to the two observation modes. The only way to resolve this issue is to use 2 optimal descriptions for each room, one each for each observation mode. The question then is how to pick which descriptions generated from the previous experiment were the best?

That is why I conducted this study in which 6 independent judges were asked to rate each description.

4.4.1. Method

24 descriptions collected from Experiment 1 were given to 6 judges, who rated each description based on their syntactic and semantic correctness with respect to the actual spatial parameters of the physical indoor scenes being described. Before giving the descriptions to the judges for rating, the descriptions were morphed in order to maintain consistency in naming the objects. That is, after morphing, the names of the objects were made to be consistent across the descriptions in order to avoid confusion. For example, if there were two tables present in the room, they were called table 1 and table 2 across all descriptions, whereas in the original descriptions they were referred to by different names in each description; for example, table a and table b or red table and brown table, etc.

The rating for each description collected from the judges was then analyzed for their reliability. An inter rater reliability analysis is generally performed using the kappa statistic to determine consistency among raters (Cohen, 1960). This is considered to be the most robust way to calculate the inter rater reliability when the ratings are categorical in nature, since it considers the agreement that happened by chance between the judges.

The judges of this experiment were asked to rate the descriptions based on the accuracy of scaling, positioning, and orientation information that was conveyed in the

description. Hence the subject ratings are continuous data and not categorical. So the kappa statistic could not be used in this study. Instead, I calculated the inter rater reliability using intra-class correlation, which in this case describes how strongly the ratings given by individual judges for a description resemble each other (McGraw & Wong, 1996)

4.4.2. Participants

6 native English speakers (3 male and 3 female, Mean age =21.7) were recruited as independent judges for this study. The study was approved by the Institutional Review Board (IRB) of the University of Maine and on average each subject took 60 minutes to complete the task. The participation of all subjects in the experiment was a voluntary decision made by them and every participant signed informed consent forms stating this right. The subjects were monetarily compensated for their time and effort to participate in the experiment.

4.4.3. Procedure

24 descriptions were collected as a result of Experiment 1 – six each for both observation modes and office room. Hence there were 4 categories. Each subject received 6 descriptions of the same category at a time and they were requested to rank those descriptions completely before moving on to the next category of six descriptions. While doing the task, the judges sat at a computer in which the photograph of the room being described in the experimental description was displayed (the photographs used in this experiment are the same as shown in figure 4.2 and 4.3).

The task for them was to read the descriptions thoroughly and to compare the spatial information of the room presented in the description to the actual room as shown in the photograph. They were given a response sheet in which the description numbers were pre-printed, so that they could write the rank they determined for each description in that response sheet.

When using independent raters, having clear instructions and a consistent protocol is critical to ensure inter-rater consistency in the process. Hence I read the following experimental instructions aloud to each subject in order to ensure that every subject got the same instruction to do the task. Also, this script briefly explains the methodology of the experiment and provides the judge guidance on how to consistently rank each description.

"You will be given 6 descriptions and a photograph of an office room. All 6 descriptions describe the same office room as is depicted in the photograph given to you. Your goal is to rank order the descriptions from best to worst, so we can identify the single most accurate description to associate with the picture. Note that in a future study, we will ask people to read the description and try to recreate the office space based on this information using a 3D modeling program. Thus, identifying the description that most accurately reflects the contents of the picture is important. Your task is to read each description and evaluate its accuracy based on three factors including positional accuracy, orientation accuracy and object retention accuracy. Each is described below.

Positional accuracy is the accuracy with which the object's original location is described in the photograph. For example, the description has a sentence "A file cabinet is next to the table and 2 feet to the left of the window". The positional accuracy of the description of the file cabinet is 100% if this description clearly describes the location of a file cabinet as seen in the picture and it might be 50% if this description describes the object's position but there is less clarity in determining its position in the room. You would score 0% if the file cabinet is missing from the description. This is only one example but try to consider the position of all the items in the description with respect to the picture when coming up with your total position accuracy score.

Orientation accuracy is the accuracy with which the description describes the orientation of objects. For example, the sentence "The chair is facing the table" in a description will have 100% orientation accuracy if the corresponding chair in the photograph was completely facing the table. This could be 50% if this orientation was not explicitly described, but you were able to deduce this information from the description. The Orientation accuracy will be 0% if the orientation of an object could not be deduced from the description at all. This is only one example but try to consider the position of all the items in the description with respect to the picture when coming up with your total position accuracy score

Object retention accuracy depends on the objects that are being described in the description against the total number of objects present in the room as shown in the photograph. It would be 100% if all objects in the room are described, 50 % if only half of

the objects in the room were described, and 0% if no object in the room was included in the description.

The most important factor is that you judge all descriptions consistently, using the same scoring criteria for each description.

There are also two columns in this sheet. The first column gives the name of the descriptions given to you. As I said before, your task now is to read the 6 descriptions given to you and evaluate each description based on the positional accuracy, orientation accuracy, and object retention accuracy and to rank each description on a scale of 10. Your scores can fall anywhere along a scale from 0 to 10, with 0 being the worst and 10 being the best. Before scoring, it might help to look through all of the descriptions to get an idea of the range of accuracies across the 6 descriptions.

Once you determine your scoring criterion, judging the positional, orientation and object retention accuracies will be easier. Keep in mind that descriptions rated 9 should look more accurate to you than those rated 8. Likewise, the descriptions rated 8 on a 10 point scale, should be twice as accurate as one you gave a rating of 4. Thus, if you rank order the maps by the numbers you assign, their accuracy should progressively improve as your scores increase. Doing so will allow you to validate your scores.

Please do remember that your ranking should be an overall rank for the description based on the above mentioned accuracies."

I printed a copy of this instruction and gave it to the independent judges before they started their task, so that they could refer to the instructions if they wanted to refresh their memory.

A similar procedure was employed by (Giudice, 2004) for rating maps and I believe that the extension to scenes is a natural extension to the earlier work.

4.4.4. Results

The ratings for each description were collected and an average rating for each description was calculated based on the individual ratings given by 6 independent judges. Following is the table that shows the average ratings for each description.

Condition	Description Number	Average Rating
Room 1 Direct Observation mode	1	7.783333333
	2	7.55
	3	8.166666667
	4	6.1
	5	7.5
	6	5.55
Room 2 Direct Observation mode	7	7.05
	8	6.766666667
	9	7.666666667
	10	6.55
	11	4.45
	12	7
Room 2 Photo Observation mode	13	7.783333333
	14	7.05
	15	7.466666667
	16	4.533333333
	17	7.5
	18	5.383333333
Room 1 Photo Observation mode	19	7.716666667
	20	7.5
	21	7.45
	22	7.833333333
	23	6.283333333
	24	6.916666667

Table 4.1 Average rating for each description

The average rating obtained by each description could not be the sole deciding factor to select the best representative description. It is important to find if there is a correlation between judges in giving their ratings to each description.

The degree of consistency in the rating for each description is determined by intra-class correlation. Hence, the intra-class correlation coefficient (ICC) was calculated in order to understand the reliability in rating given by the judges. In this ranking data, the average

measure ICC was 0.638, 95% CI [0.345,0.822]. Generally if the value of ICC is between 0.6 – 0.8, it indicates a strong correlation. Hence, based on the ICC of this ranking data, we can say that there existed a strong correlation between the six judges in ranking the descriptions (see (Landis & Koch, 1977) for more information about ICC).

The description with the highest average rating in each category was selected as the representative description for that category, because of the fair correlation existing between judges. Based on these data, 4 representative descriptions were selected for each category and were used in experiment 3, where the scenes narrated in the description were recreated. The recreations were evaluated later to see the functional equivalence of photographic and direct observation modes in terms of the conveyed spatial information about indoor scenes.

4.5. Experiment 3 – Scene Recreation Study

In this study, I analyzed the spatial apprehension of the listener of an indoor scene description in order to examine the effect of observation modes used on generating the descriptions.

4.5.1. Experiment Design

The 4 representative descriptions, selected based on the results from the inter subject rating experiment (refer to section 4.4 for more details about the experiment) were used in this study. Participants were asked to recreate indoor scenes based on the descriptions using a 3D architectural rendering software called RoomArranger (See section 4.2.3 for more information about the software).

4.5.2. Participants

12 sighted native English speakers participated in the study (6 male subjects and 6 female subjects, mean age = 21.4) The study was approved by the Institutional Review Board (IRB) of the University of Maine and on average each subject took 60 to 75 minutes to complete the task. The participation of all subjects in the experiment was a voluntary decision made by them and every participant signed informed consent forms stating this right. The subjects were monetarily compensated for their time and effort to participate in the experiment.

4.5.3. Material and Apparatus

The materials and apparatus were the same as were used in the pilot scene recreation study. (Refer to section 4.2.3 for more details).

4.5.4. Procedure

The procedure of this study was the same as the pilot scene recreation study described in section 4.2.4, where the participants were asked to recreate an indoor scene based on the descriptions given to them. The subject distribution for this study followed a within-subjects design and the order of recreating observation modes was counterbalanced.

4.5.5. Results

The recreations of the rooms made by subjects were saved. Later, for each object in the recreation, I extracted and exported the information about the x and y coordinates,

length, width, height and orientation (that is, the angle to which the object was rotated during recreation) to an excel file.

In order to evaluate the accuracy of subject's recreations, I recreated both the office rooms based on the actual physical dimensions of all objects using the room arranger software. I extracted and imported x and y coordinates, length, width, height and orientation data of every object in the room with the original dimensions. Later, the accuracy of subject's recreation was evaluated by comparing it against the recreation of the room with the actual dimensions.

The accuracy was evaluated based on 4 factors namely the scaling error, positioning error, orientation error and object retention error.

4.5.5.1. Scaling Error

Scaling error was calculated for each recreated object in the room based on the formula mentioned in section 4.2.5. There were 11 objects for which the percentage scaling error was calculated and there were 12 subjects who recreated the room, based on 2 observation modes. Hence there should have been 264 data points (object recreations) in total (132 for the photographic observation mode and 132 for the direct observation mode).

But there were 5 objects (3.7 % of the data), which were not recreated by subjects, and they were removed from the data set. Hence the dataset had 259 data points.

The differences between scaling error for direct observation and photographic observation modes were normally distributed, as assessed by visual inspection of a Normal Q-Q Plot.

The scaling error of the objects recreated by participants using descriptions that were generated based on the photographic observation mode (M = 0.28 feet, SD = 0.14) were numerically higher than the scaling error of the objects recreated by participants using description that were generated based on the direct observation mode (M = 0.23 feet, SD = 0.28).

Although there are numerical differences, the scaling errors were quite small and not statistically significant between the direct observation mode and photographic observation mode, $t(126) = 0.9$, $p = 0.2$ (>0.05). Thus, results from this scene recreation study suggest that observation modes do not have any influence in the listener's perception of object dimensions.

4.5.5.2. Positioning Error

Positioning error was calculated as the Euclidean distance between an object's actual position in the room and its recreated position, based on the formula mentioned in section 4.2.5. Euclidean distance was calculated for every object in both the rooms across both the observation modes.

There were 5 objects (3.7 % of the data), which were not recreated by subjects, and they were removed from the data set. Hence the dataset had 259 data points. Also, the positioning errors of 11 objects (8.6% of the data) were more than 2.5 times the

standard deviations from the mean and they were categorized as outliers. The outliers were then replaced with the corresponding subject's mean positioning error.

The positioning error of the objects calculated by Euclidean distance was higher for recreations based on the direct observation mode (M=5.6 feet, SD = 4.3) when compared with the recreations based on the photographic observation mode (M = 4.2 feet, SD = 2.9).

Although there are numerical differences, the small difference in positioning errors was not statistically significant between the direct observation mode and photographic observation mode, $t(126) = 0.98$, $p = 0.08$ (>0.05). Thus, results from this experiment suggest that observation modes do not have any influence in the listener's perception of object locations in an indoor scene.

4.5.5.3. Orientation Error

Orientation error was calculated for each recreated object in the room based on the formula mentioned in section 4.2.5. The trashcan was circular and hence there could be no orientation information that could be extracted for this trashcan. Hence there should only be 240 orientation errors in the dataset. In addition, 5 objects (2.1% of the data) were not recreated by the subjects and they were removed from the data set. Also, there were 5 outliers detected (2.1% of the data set), that were more than 2.5 box-lengths from the edge of the box in a boxplot data. Those outliers were replaced with the average Euclidean distance with the corresponding subject's other objects being recreated.

The orientation error difference between the observation modes was normally distributed and this normality was verified by visually inspecting the Normal Q-Q plot.

The orientation error of the objects calculated by Euclidean distance is higher for recreations based on the photographic observation mode (M=9.2 degrees, SD = 6.2) when compared with the recreations based on the direct observation mode (M = 6.3 degrees, SD = 4.1).

Although there are small numerical differences, the orientation errors are not statistically significant between the direct observation mode and photographic observation mode, $t(114) = 2.12$, $p = 0.62$ (>0.05). Thus, results from this scene recreation study suggest that the use of direct versus photographic observation modes do not have any influence in the listener's perception of object orientation.

4.5.5.4. Object Retention error

Object retention is calculated by counting the number of objects being recreated in each room by the participants. The data shows that recreations based on the photographic observation mode constituted 131 objects, with an average of 10.91 objects per room (where actually there should be 132 object recreations, since 12 subjects were recreating rooms with 11 objects each based on two observation modes). On the other hand, photographic observation of the room resulted in the recreation of 128 objects, with an average of 10.66 objects per room (where actually there should be 132 object recreations). But this difference in object retention was not statistically significant between the photographic and direct observation modes of the room, ($t(126) = 0.7$, $p =$

0.5 (>0.05)). Hence we can conclude that the results from the experiment suggest that the two observation modes do not have any influence in the listener's object retention ability.

4.5.5.5. Topology Score

The topological scores for each recreation are evaluated as per the discussion in section (4.2.5). It was found from the data that recreations based on the direct observation mode had an average score of 26.5 (where the ideal score for each recreation is 32), while photographic observation of the room had an average score of 24.8. This difference in topological score was not statistically significant based on paired t-test results ($t(11) = 7.45$, $p = 0.09$ (>0.05)). Hence we can conclude that observation modes do not have any influence in the listener's apprehension of topology during a scene reconstruction task.

Altogether the data collected from the scene recreation study was analyzed by calculating the scaling, position, orientation and object retention error of the recreated objects and the overall topology scores of recreated rooms. The errors were then analyzed across the two observation modes used to describe the indoor scene.

It can be seen that the results of the scene recreation study and pilot recreation study (as discussed in section 4.2.5) converge to demonstrate that the observation modes (direct observation mode and the photographic observation mode) used to generate indoor scene descriptions likely do not affect the listener's spatial apprehension of the scenes. It is important to note that humans interpreted the spatial information from

photographs. However it might be possible in the future for machines or robots to profit from the depth cues as discussed in (O'Shaughnessy, 2012) to interpret spatial information from photographs. This would open the door for the design of a fully automated DISc system as being functionally feasible. But this automation of DISc is out of the scope of this thesis. In the future, an experiment comparing observation modes could be used for establishing guidelines to compare different spatial information acquiring systems against the direct observation of scenes.

4.6. Summary and Discussion

This chapter started by reviewing the literature discussing various characteristics of photographs as a result of design restrictions like sensor sensitivity, absence of stereovision and limited field of view. These restrictions could possibly affect the information content of a photograph, resulting from the photos as a limited information medium. This limitation in information was expected to influence the difference between photographic observation and direct observation of an indoor scene.

As photographs are considered as an information source for generating natural language descriptions in DISc, the question was is there any difference between the listener's spatial apprehension of an indoor scene developed based on the natural language descriptions generated by observing an indoor scene in a photograph and the listener's spatial apprehension of the same indoor scene developed based on the natural language descriptions generated by directly observing that scene. Hence, I conducted 3 experiments as described in sections 4.3, 4.4 and 4.5.

The results from experiment 1, 2 and 3 suggest that both the observation modes result in functionally equivalent indoor scene descriptions which will help its listener's to apprehend the space with the same level of scaling, positioning, orientation, object retention errors and scene topology scores. Overall, it can be seen from the results that the type of observation mode does not have any influence on the amount of information being communicated through the natural language Descriptions of the room.

Hence irrespective of the nature of photographs, which are a limited information media, we can proceed with the idea of using them as an information source to generate natural language descriptions in our non-visual scene description system – DISc.

It is important to note here that the photographs used in these experiments were digitally stitched to match the field of view of direct observation. Thus, in an ideal situation, these results are true, but it is not clear if it would hold for a much more limited photo with a reduced FOV. Hence in the future this research could be extended to investigate the minimal field of view required to obtain spatial information to be on par with the direct observation.

5 Comparing Linearization Strategies in Indoor Scene Descriptions

In an indoor scene description, the spatial relation between objects is described one by one, following a specific strategy, but there can be multiple ways to order the spatial description of the objects. For example in experiment 1, 24 unique descriptions about 2 rooms (12 for each room) were collected as a result of the scene description study (See section 4.3 for more information about the experimental design used to collect descriptions). Although the descriptions were collected with the intention to evaluate the functional equivalence between the natural language descriptions generated based on different observation modes (e.g., direct perception vs. from a photograph) , this chapter investigates global strategies in which the objects were described. This analysis identified 6 different strategies used by 12 different participants. This global order of describing the objects in a room seemed interesting with respect to specifying description logic, and this investigation considers if the strategy used affects the listener's accuracy of spatial apprehension and coherence. Hence in this experiment, I compared various description strategies by evaluating the accuracy of listener's indoor spatial apprehension that was developed based on the natural language description of indoor scenes.

5.1. Motivation

Levelt studyied the global structure of ordering spatial information of objects in a spatial description, early in 1981. He introduced the term "linearization" relating to the strategy in which the spatial information of objects is organized in a description. According to

him, linearization is a process in which spatial patterns, involving two or three dimensions, are converted to a linear one dimensional spatial description (Levelt, 1981).

While Levelt was interested to see if the linearization process is a function of the modality in which language is used, Ehrich and Koster (later in 1983) were interested to understand the different ways of linearizing spatial properties of objects present in an indoor scene. Hence they conducted a behavioral study in which they showed a miniature room to the participants and later asked them to describe the room from memory (Ehrich & Koster, 1983).

They analyzed the descriptions and found that their participants employed linearization at two levels. Higher level linearization is used to organize the spatial description of object clusters that are present in the room, while the lower level linearization is used to organize the spatial description of individual objects present within each object cluster. In the higher level technique, two types of linearization were commonly observed, Parallel-Line and Round-About (the difference between these two linearization types is the presence and absence of spatial discontinuity while jumping from one wall to the other while describing object clusters respectively). In lower level linearization, two types of sub-linearization techniques were found, namely Sequencing and Grouping. Sequencing is used in connected descriptions, where the first object is used as the base object for referring to the spatial location of the second object, the second object for referring to the third object and so on. Grouping is used in disconnected descriptions,

where one object is used to describe the spatial location of other objects present in that object cluster.

Although Ehrich and Koster (1987) tried to understand the linearization process in indoor scene descriptions, they did not explore if the linearization process affects the accuracy of listener's spatial apprehension of that indoor scene. Since the motive of any spatial description is to impart spatial knowledge about the scene to the listener, it is important to evaluate if the linearization strategy used in the scene description influences the accuracy with which the listener apprehends the global spatial representation of the scene.

5.2. Types of Linearization techniques

For this experiment I considered the linearization strategies identified in the descriptions that I collected from Experiment 1. These included 24 unique descriptions for 2 different office scenes (12 descriptions for each scene), given by 12 participants. For each description, the order in which the objects were described was evaluated and categorized according to Ehrich and Koster's (1983) linearization schemes. As a result of this analysis, I identified 2 major linearization strategies as proposed by Ehrich and Koster, namely the Round – About linearization and Parallel – Line Linearization techniques.

In the following section, I will discuss the linearization strategies identified from my corpus of indoor scene descriptions.

It is important to note here that the indoor scenes used in behavioral experiments conducted for this thesis do not have a complex spatial configuration with object clusters, such as a couch with a center table, dining table with chairs, etc. As a result, the lower-level sub linearization strategy involving sequencing and grouping were not considered and only higher-level linearization strategies, namely Round-About and Parallel Line approaches, are considered in this chapter."

5.2.1. Round-About Linearization

The name of the linearization strategy by itself suggests that the describers narrated the spatial location of the first object and then used that object as the base object for referring to the spatial location of the second object, and used the second object to describe the third object and so on. Hence in this type of linearization method there is a spatial continuity in the order in which the objects were described. The arrows in figure 5.1 show the order in which the objects were described in a Round-About linearization method.

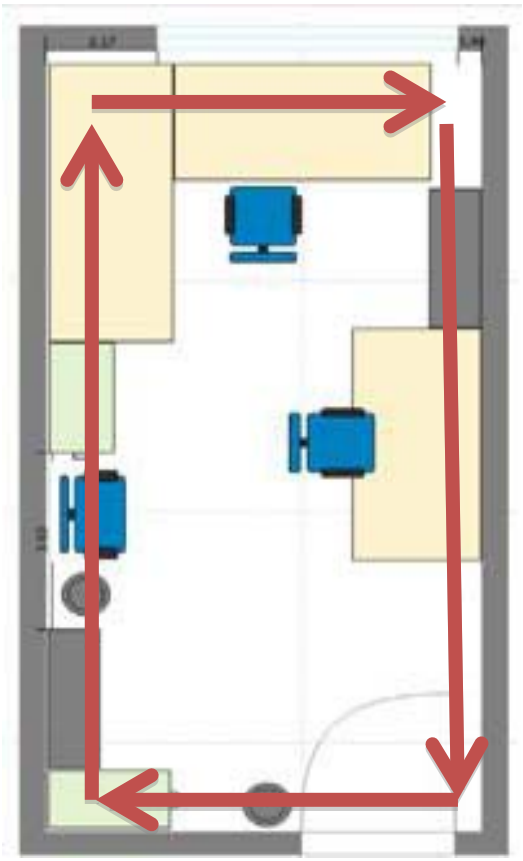

Figure 5.1 Example of a round-about linearization strategy

This linearization strategy is further sub-classified into two types based on the location of the first object being described in the description. They are discussed below.

5.2.1.1. Cyclic Linearization

In this linearization method, participants described the closest object either on their right or left side as the first object in their room description and then continued describing the room by using the Round-About linearization strategy discussed earlier.

5.2.1.2. Center – Cyclic Linearization

In this linearization method, participants described objects located against the far wall of the room from their position and continued describing the room by using the Round-About linearization strategy.

5.2.2. Parallel-Line Linearization

In the Parallel-Line Linearization technique, subjects described objects that were located against a wall (it could be the right, the left or the far wall of the room). First they started by describing objects that were found in one corner of the room and continued describing objects while maintaining spatial continuity in their description until they reached the end of the wall. After describing every object that was located against that wall they jumped to another wall, thereby breaking the spatial continuity in the description when switching walls. The path marked by red colored arrowheads in figure 5.2 shows an example of the parallel line linearization strategy.

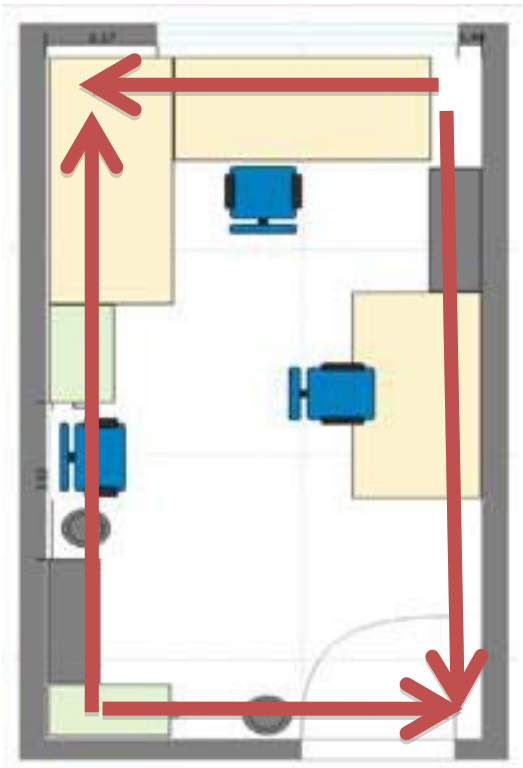

Figure 5.2 Example of the parallel line linearization strategy

Based on the spatial location of the first object that was described, Parallel Line linearization is further divided into two types.

5.2.2.1. Side-Side Linearization

In this type of linearization method, participants first described objects that were either located against the right wall or left wall of the room and then followed the Parallel-Line linearization strategy in their description.

5.2.2.2. Center-Side Linearization

In this type of linearization method, participants first described objects that were located against the far wall of the room and then followed a Parallel-Line linearization strategy in their description.

5.2.3. Functional Linearization

In this type of linearization strategy, participants grouped the objects based on their functionality and described the spatial location of every individual object that was present in the group before proceeding to the next group. For example, subjects described the spatial location of every table that was present in the office scene, and then described the spatial location of every chair and so on.

5.2.4. Random Linearization

There were a few descriptions in the corpus, which could not be classified into any of the above-mentioned linearization categories. For these descriptions, the participants described the objects in the indoor scenes without any meaningful linearization pattern, except for describing the objects that caught their attention while observing the scene.

5.3. Distribution of Linearization Order of Descriptions

Among 24 descriptions collected, 6 used cyclic linearization, 7 used Center-Cyclic linearization, 5 used Side-Side linearization, 3 used Center-Sides linearization, 2 used Functionality based linearization and 2 used random linearization.

The pie chart in figure 5.3 shows the distribution of linearization strategies found in the corpus of office scene descriptions that were collected in Experiment 1. It can be seen from the distribution that the Round-About linearization method (13 out of 24 descriptions) was the dominant linearization strategy adopted by the participants, followed by the Parallel-Line linearization strategy (8 out of 24 descriptions). By contrast, the functionality based linearization strategy (2 out of 24 descriptions) and

random linearization strategy (2 out of 24 descriptions) were not commonly used by our participants.

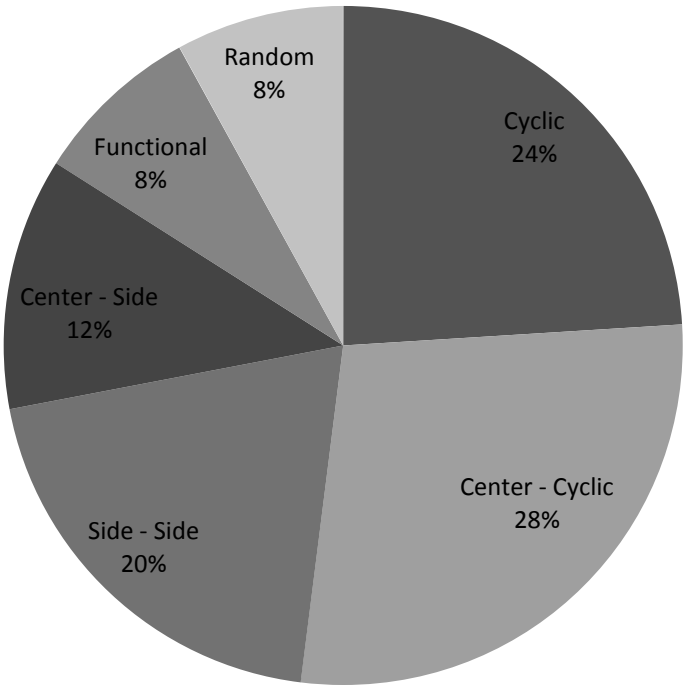

Figure 5.3 Distribution of linearization strategies used in our office scene description corpus

Functionality based linearization and random linearization were not the dominant form of describing a space and apparently not found in the corpus of scene descriptions collected by Ehrich's team. Owing to the absence of discussion about these linearization strategies in the literature and considering their low counts in the corpus of my indoor scene descriptions, the functionality based linearization strategy and random linearization strategy were not included for further analysis and interpretation in this experiment.

5.4. Experiment 4 – Comparing Linearization strategies

In this experiment, participants were given written textual descriptions of indoor scenes and were then asked to recreate the scene being described. Descriptions of 2 different office scenes were used, both with the same number and same kind of objects. These were the same scenes whose descriptions were collected in experiment 1, see figures 4.2 and 4.3 for photographs of the indoor scenes.

The experiment followed a between – subjects design, where each subject recreated 2 descriptions, one from each linearization type (that is, Round – About linearization and Parallel – Line linearization) and one from each sub-linearization method (Center-Cyclic or Center-Sides and Cyclic or Side-Side) and they were counterbalanced. The subjects were not given any information about the linearization strategy they were exposed to.

5.5. Participants

16 sighted native English speakers participated in this study (8 male subjects and 8 female subjects, mean age = 20.8)

The Institutional Review Board (IRB) of the University of Maine approved the study and on average each subject took 60 minutes to complete the task. The participation of all subjects in the experiment was a voluntary decision made by them and every participant signed informed consent forms stating this right. The subjects were monetarily compensated for their time and effort to participate in the experiment.

5.6. Software used

I used 3D architectural rendering software called Room Arranger (Version 3.3). It is the same as the software used in Experiment 3 (refer to section 4.2.3 for more information about the software specifications).

5.7. Descriptions Used

We manually authored 8 indoor scene descriptions for 2 different office scenes (presented in figure 4.2 and 4.3), 4 for each room following the 4 linearization strategies mentioned in section 5.2.1 and 5.2.2.

The descriptions were carefully authored based on the terminology and structure of the previous participant-generated descriptions from Experiment 1. Also, while authoring the descriptions, we made sure that the spatial location of every object in the room was referred to with the same amount of references and same kind of spatial language to avoid any potential noise or bias in the recreation data developed as a result of the difference in language used. Likewise, all 4 descriptions were alike with respect to the content and spatial language used, except for the order in which the content was organized/presented in the description.

5.8. Procedure

Each subject watched the tutorial video in order to learn how to use the “Room Arranger” software. They were asked to practice with the software by recreating a sample description given to them. The imaginary room described in the sample description had a totally different layout and object set than was used in the actual

office scenes. While recreating the sample description the participants were allowed to get help from the experimenter regarding the software and the sample description. Once the subject was confident about using the software, they were given the first description to recreate.

I read the following script aloud to each subject to ensure that every subject received the same instructions about the task. "I will give you a description of an indoor office scene. Your task is to read that description at your own pace and recreate the indoor scene described using the Room Arranger software."

In this study, the subjects were allowed to keep the descriptions with them while recreating the room, so that they can go back and forth while recreating. This method eliminates the potential noise in the data related to the use of memory, as described in section 4.2.6.

Each subject received 2 descriptions for recreation and they were not given any time limit to complete the task but the average time taken for recreating each description was approximately 15 minutes. Once they finished their scene creation, the descriptions were removed and the next description to be recreated was given to them. Each subject recreated 2 descriptions based on 2 different linearization strategies (as described below).

We authored indoor scene descriptions only for 2 rooms. Hence in order to avoid a learning effect about the rooms, participants were asked to recreate only 2 descriptions, one for each room. The descriptions included an example from each higher level

linearization strategy (Round-About and Parallel-Line linearization) and one from the subclass of each higher level linearization strategy (cyclic or center – cyclic and side – side or center – side (refer to section 5.2 for more information about the higher and lower-level linearization strategies)). Hence the design of the study was within-subjects in order to evaluate the higher-level description strategies and between-subjects design to evaluate lower-level description strategies. The sequence of recreating descriptions based on linearization strategies was counterbalanced.

5.9. Results

The recreations of the rooms made by all subjects were saved and subsequently, for each object in the recreation, I extracted and exported the information about the x and y coordinates, length, width, height and orientation (that is, the angle to which the object was rotated during recreation) to an excel file.

In order to evaluate the accuracy of participant's recreation, I recreated both the office rooms based on the actual physical dimensions of the objects using the room arranger software. I then extracted and imported x and y coordinates, length, width, height and orientation data of every object in the room with the original dimensions. Later, the accuracy of participant's recreation was evaluated by comparing it with the accurate recreation of the room based on the actual dimensions.

The accuracy of participant recreations were evaluated based on 3 factors, namely the scaling error, positioning error, and orientation error for each recreated object. There

were 16 participants in this experiment and each recreated 2 descriptions. Each description had 11 objects. Hence there were 352 (16*11*2) data points in total.

The scaling errors of 7 objects, positioning errors of 11 objects, and orientation errors of 4 objects were more than 2.5 times their corresponding standard deviations from the group's mean and were categorized as outliers. The outliers were then replaced with the corresponding subject's mean scaling error.

5.9.1. Round – About Versus Parallel Line Strategy

The recreations of indoor scenes were classified based on the linearization type used in the indoor scene description with which the recreations were made. In order to investigate the combined effect of the multiple variables: Scaling, Positioning and Orientation errors on two different linearization strategies, a one-way Multivariate Analysis of Variance (one-way MANOVA) was performed on the data. The results suggest that there is a statistically significant difference between linearization strategies in obtaining spatial knowledge, F (2, 28) = 10.08, $P < .05$, Wilk's $\lambda = 0.481$, partial $\varepsilon^2 = 0.519$. Power to detect the effect was 0.995, with the Round – About Linearization strategy showing better recreation accuracy.

Given the significance of the overall omnibus test, subsequent post hoc paired t-tests were performed on the data to see the effects of linearization strategy in scaling, positioning and orientation accuracies individually. The results suggest that there is a statistically significant difference between the Round – About and Parallel line linearization strategies in terms of positioning accuracy ($t(176) = 1.53$, $p = 0.04$) and

orientation accuracy (t(176) = 3.75, p = 0.01). Also, the paired t-test results suggest that there is no significant difference between Round – About and Parallel line linearization strategies in terms of scaling accuracy (t (176 = 3.44, p =0.6).

Table 5.1 shows the mean and standard error for scaling error, positioning error and orientation error based on the Round-About and Parallel-Line linearization modes with 95% confidence intervals.

Dependent Variable	Linearization Type	Mean (in feet)	Std. Error (in feet)	95% Confidence Interval	
				Lower Bound	Upper Bound
Scaling Error	Parallel Line	4.843	1.771	3.267	6.418
	Round About	0.564	0.784	-1.011	2.14
Positioning Error	Parallel Line	6.68	2.835	5.587	7.774
	Round About	3.979	1.581	1.886	4.073
Orientation Error	Parallel Line	7.5	2.739	1.907	13.093
	Round About	1.6	0.674	-5.593	5.593

Table 5.1 Descriptive statistics of the dependent variables based on different linearization strategies

Another sub-analysis was done on the recreation data to see if the number of spatial discontinuities in an indoor scene description affect the way in which the user of the description perceives the indoor scene that was described. In this analysis, the descriptions were classified based on the number of spatial discontinuities and the recreations based on those descriptions were evaluated for scaling error, positioning error and orientation error (as shown in table 5.2).

Dependent Variable	Linearization Type	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Scaling Error	0 Discontinuity	0.564	0.784	-1.011	2.14
	3 Discontinuities	4.445	2.535	1.91	6.98
	4 Discontinuities	5.241	1.632	3.609	6.873
Positioning Error	0 Discontinuity	3.979	1.581	1.886	4.073
	3 Discontinuities	5.67	2.684	2.986	8.354
	4 Discontinuities	7.69	1.673	6.017	9.363
Orientation Error	0 Discontinuity	1.6	0.674	-5.593	5.593
	3 Discontinuities	6.857	2.431	4.426	9.288
	4 Discontinuities	8.143	1.847	6.296	9.99

Table 5.2 Descriptive statistics of the dependent variables based on the number of spatial discontinuities found in the scene descriptions

The results suggest that increasing the number of spatial discontinuities in an indoor scene description negatively affects the accuracy of spatial apprehension of the scene based on those descriptions.

5.9.1.1. Discussion

The results of a one way MANOVA suggest that Round-About and Parallel-Line linearization methods employed in an indoor scene description significantly affect the listener's spatial apprehension of that indoor scene. A comparison on the means of scaling, positioning and orientation error of recreation based on the descriptions which employed Round-About and Parallel-Line linearization methods suggest that indoor scene descriptions which employed a Round-About linearization method helped the participant to acquire better spatial knowledge rather than the descriptions which employed the Parallel-Line linearization method. Also, the topological scores for each recreation were evaluated as per the discussion in section (4.2.5). It was found from the data that recreations based on the Round about Linearization strategy had an average score of 29.6 (where the ideal score for each recreation is 32), while Parallel-line linearization strategy of the room had an average score of 24.1. This difference in topological score was statistically significant based on the paired t-test results ($t(15) = -7.43$, $p = 0.03$ (>0.05)). Hence we can conclude that the results from experiment 4 suggest that the linearization strategy employed in a scene description affects the listener's apprehension of scene topology during reconstruction.

The results of this behavioral experiment support the earlier research work conducted by Ehrlich and Johnson-Laird in 1982. In their research, they found that the spatial comprehension is enhanced by the presence of spatial continuity in the description of spatial components (Ehrlich & Johnson-Laird, 1982). Supporting their hypothesis, the results of the current behavioral experiment (Experiment 4) suggest that the Round-

About linearization strategy which results in the descriptions with the most spatially continuous while describing objects in an indoor scene, ranked better than the Parallel-Line linearization strategy which results in descriptions with more than 2 spatial discontinuities between objects in the scene.

Considering these results with respect to our non-visual scene description system – DISc, the data suggests that employing a Round-About linearization strategy versus a Parallel-line linearization strategy to describe the spatial location of objects provides the greatest benefit for listeners of indoor scene descriptions to learn the layout and to develop more accurate global spatial representations of the scene.

5.9.2. Center – Cyclic Versus Cyclic Linearization mode

In the previous section, we discussed that the Round-About linearization strategy led to significantly more accurate performance than the Parallel-Line linearization strategy to describe indoor scenes. But as discussed in section 5.2.1, the Round-About linearization method can be further sub-divided into two types based on the location of the first object that was described, in other words the starting point of the description (refer to section 5.2.1 to understand more about the sub-types of Round-About linearization methods).

Hence it is important to analyze if the lower-level linearization method used in the indoor scene descriptions affects the accuracy of participant's spatial apprehension. In this section, I analyzed the combined effect of multiple variables like Scaling, Positioning

and Orientation errors on two different sub-linearization methods of the Round-About linearization strategy.

Results from a one-way MANOVA suggests that there are no statistically significant differences between the Center – Cyclic and the Cyclic linearization methods in obtaining spatial knowledge, F (1,13) = 0.36, P =0.7(> 0.05); Wilk's $\lambda = 0.94$, partial $\varepsilon^2 = 0.52$. Power to detect the effect was 0.096.

Subsequent post hoc Paired t-tests performed on the data to evaluate whether the presence of different starting points employed in the spatial description affect scaling, positioning, and orientation accuracies independently showed no significant results between the two strategies for any measure.

Dependent Variable	Linearization Type	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Scaling Error	Center-Cyclic	0.403	0.27	-0.177	0.982
	Cyclic	0.726	0.27	0.147	1.306
Positioning Error	Center-Cyclic	2.929	0.552	1.746	4.112
	Cyclic	3.029	0.552	1.846	4.212
Orientation Error	Center-Cyclic	0	0	0	0
	Cyclic	0	0	0	0

Table 5.3. Descriptive statistics of dependent variables based on different sub-linearization types of the round-about linearization strategy

Table 5.2 shows the mean and standard error for Scaling, Positioning and Orientation errors based on the sub categories of Round-About linearization strategy (that is, Center – Cyclic and Cyclic linearization techniques) with 95% confidence intervals.

Also the difference in topological scores between Center- cyclic (with an average topological score of 28.5) and cyclic linearization strategies (with an average topological score of 29.3) was not statistically significant based on the paired t-test results ($t(15) = -6.8$, $p = 0.8$ (>0.05)). Hence we can see that the starting point of a linearization strategy employed in a scene description does not affect the listener's apprehension of scene topology.

5.9.2.1. Discussion

The result of the one-way MANOVA suggests that the sub-categorization of a linearization method based on the spatial location of the first object described in an indoor scene description does not affect participant's spatial apprehension as long as the global linearization method has spatial continuity.

The selection of the first object to convey in a scene description depends on the perspective of that scene adopted by the describer and the perspectives are affected by various factors like the spatial setting of the scene, dimension and functionality of objects present in that scene, individual differences due to the proficiency of language use, vocabulary restrictions of a language by itself, and reference frame selection (Taylor & Tversky, 1996).

The results from this behavioral experiment suggest that the starting point of a scene description (influenced by the perspective adopted by the describer) while describing an indoor scene does not affect recreation of the scene based on this information, as long

as there is consistent spatial continuity while describing the constituent components of that scene.

5.10. Chapter Summary

In this chapter, I discussed a behavioral experiment which compared different linearization strategies employed by people describing indoor scenes (based on the descriptions generated in Experiment 1, elaborated in section 4.3). The results demonstrated that the Round-About linearization strategy, which incorporates spatial continuity in the order of objects being described, is better than the Parallel – Line linearization strategy, which incorporates spatial discontinuities in the order of describing objects. Further analysis on the sub categories of Round-About linearization suggests that the starting point used for describing an indoor scene does not affect the accuracy of spatial apprehension of the listener.

Hence, in the sample scenario discussed in section 1.1, these results suggest that the most efficient approach to be included in the non-visual scene description system – DISc would use a Round-About linearization strategy.

6. Comparing Direction Estimation using two modes of angular units

For a linguistic scene description to be on par with that derived from visual perception, it is extremely important to have an efficient and accurate method for specifying directional cues about the spatial locations of objects within the scene. As with the content included in scene descriptions, there are different ways to verbally present these directional cues to the user. The work done by (Ishikawa & Kiyomoto, 2008) suggests that people best understand directional cues when they are presented using relative directions (for example, 'to the right of', 'to the left of', 'in front of' etc.) rather than using only absolute directions (for example 'North-East', 'North-West' etc.).

Perceiving direction with the help of vision is a relatively easy task compared to doing the same in the absence of visual cues owing to various advantages of visual cues (refer to section 2.1 for discussion on the advantages of vision when compared to other sensory modalities in perceiving space). As this thesis focuses on the development of a non-visual scene description system, it is important to understand the effects of linguistic elements on direction estimation. The unit of direction is a type of linguistic element in a non-visual indoor scene description that could possibly affect the perception and representation of angular direction and spatial relation.

There is formal research interested in studying the human perception of direction. One such project was carried out by (Barber & Lederman, 1988), where the researchers were interested to characterize the direction estimates of their participants who learned the direction of targets via finger movement, imagination and vision. The results from their

study showed that direction estimation by humans was better when finger movement was used to learn directions, rather than using vision. Another research study conducted by (Zanelli, Cappa, Petrarca, & Berthoz, 2011) investigated the direction perception based on whole body rotation when their participants were standing. They found that the interconnection between body movement and the corresponding information obtained from proprioceptive systems are important to estimate directions.

Although researchers are interested in analyzing direction estimation by humans, there is no formal research investigating the effect of directional units used to present direction information. The unit of direction is one of the important linguistic elements that are used while describing the spatial relations of objects in an indoor scene. Hence, in this chapter I analyzed the effects of direction units in a human angular estimation task.

The indoor scene descriptions collected from Experiment 1 (refer to section 4.3 for more information about the experimental design) showed that describers used both degree measurements and clock face directions to present angular information when they were using a relative frame of reference. But it is not possible from the data to select one of these angular units as better than the other for incorporation in DISc. Thus, a behavioral study was conducted to investigate the effects of both forms of direction units on direction estimation performance.

We all know that both degree measurements and clock face directions are capable of specifying the spatial location of an object. For example, “a desk is at your 1 o’ clock

position" and "a desk is at 30 degrees on your right" both specify the same spatial location of the desk. But it is important to know which of these presentation methods of angular description lead to the most accurate perception of directional information. To address this question, a behavioral study was conducted to compare the accuracy of listener's angular perception based on these two types of directional cues.

6.1. Motivation

The real motivation for this behavioral experiment is the lack of empirical research done with the explicit intention of comparing metric units used for describing directional information. In the natural language research community, a considerable amount of research has been done to compare relative and absolute directions while giving information. One such example is the research work conducted by Anacta and Schwering, where they found that both men and women performed better in wayfinding tasks based on relative directional cues rather than absolute directional cues (Anacta & Schwering, 2010).

There are also research studies that have been conducted to compare the efficiency of route directions given as metric information compared to GPS route directions given as landmark anchored directions. This work has found that both approaches yielded similar navigation performance (Rehrl, Häusler, Leitinger, & Straße, 2010). However, to my knowledge, none of the literature has investigated the effect of angular units on spatial apprehension for direction estimation and production performance.

Hence, in order to describe these spatial relations, metric units are used for specifying distance and angles in a spatial description. While measuring units for distance vary depending on culture (for example, Americans use units like miles, feet and inches for referring to Euclidean distances, whereas most other countries use Kilometers, meters, centimeters and millimeters for referring to distances), By contrast, most cultures use both analog clock face units and degree measurements to refer to angular displacements.

Hence the behavioral experiment reported in this chapter was designed with an intention to see if the measuring unit of angle, which is used for referring to directions, affects listener's direction estimation. Apart from angular measurement units, it is important to evaluate how precisely the angular information could be given in a spatial description in order to still be meaningful and comprehensible for use in a scene description or for direction giving to support spatial behaviors. This makes it important to discuss the clock-face system being quantized into 12 divisions, which is 30 degrees in angular precision. In this thesis, I am evaluating whether I can push the envelope of traditional boundaries, by providing a 100% increase of precision by testing the half hour clock-face directions (e.g., 1:30), which would result in 15 degrees precision.

On the other hand, degree measurements are not limited by such quantization but degree or partial degree units are very small and might not be perceptually meaningful. This is because, we usually only talk of 45 degrees, 90 degrees, 180 degrees etc. These canonical angles are well-known and easily learned (Ivanenko, Grasso, Israël, & Berthoz,

1997) but it is not known if the range of other non-canonical angles are perceptually meaningful at 15 degree increments.

6.2. Experiment 5 – Comparing Direction Estimation using two modes of angular units

In this experiment, I evaluated blindfolded participant's accuracy in estimating angular directions at 15 degree intervals. That is, I evaluated their direction estimation for 15, 30, 45, 60, 75, 90, 105, 120, 135, 150, 165 and 180 degrees, from both the right and left 180 degree hemi-field from a 0 degrees forward facing direction. Table 6.1 shows the stimulus set of angles and their equivalent clock face directions.

Owing to the unnatural state of providing directions in terms of angles greater than 180 degrees, and to keep the complexity of both directional units (degree measurements and clock faces) the same, all measurements of angular extent were presented in terms of the right and left hemisphere. For example, 270 degrees would be represented as 90 degrees, left.

It can also be seen from Table 6.1 that I evaluated participant's direction estimation for every half hour interval.

S.No	Angle	Side	Equivalent Clock Face
1	15	Right	12:30
2	30	Right	1:00
3	45	Right	1:30
4	60	Right	2:00
5	75	Right	2:30
6	90	Right	3:00
7	105	Right	3:30
8	120	Right	4:00
9	135	Right	4:30
10	150	Right	5:00
11	165	Right	5:30
12	180	Right	6:00
13	165	Left	6:30
14	150	Left	7:00
15	135	Left	7:30
16	120	Left	8:00
17	105	Left	8:30
18	90	Left	9:00
19	75	Left	9:30
20	60	Left	10:00
21	45	Left	10:30
22	30	Left	11:00
23	15	Left	11:30

Table 6.1 Degree measurement and their equivalent clock face angles

In the experiment, a single angle is tested in each trial after resetting the subject's orientation to 0 degrees. The experiment was carried out using a mixed-factorial design, where the direction units were tested using a within-subjects design, and the individual angles were tested using a between-subjects design.

6.3. Participants

32 sighted native English speakers participated in the study (16 male subjects and 16 female subjects, mean age = 25.6).

The study was approved by the Institutional Review Board (IRB) of the University of Maine and on average each subject took 90 minutes to complete the task. The participation of all subjects in the experiment was a voluntary decision and every participant gave signed consent stating this right. The subjects were monetarily compensated for their time and effort to participate in the experiment.

6.4. Method

32 subjects were divided into two subject pools. Refer to table 6.2 for the angles used for the participants in subject pool A and subject pool B respectively.

It can be seen from these tables that the individual trials in each subject pool were balanced between the angles across right and left hemispheres and angles across front and back hemispheres. Also the angles are balanced across the units of measurement. The angles in pool A and pool B are mutually exclusive, except for cardinal angles 45, 90, 135, 180, -45, -90, and -135 degrees which were evaluated in both the pools. At the end of the experiment, there were 16 data points for every angle in both degree measurement and clock face units.

Subject Pool A			
Angle	**Angle Unit**	**Left/Right**	**Front/Back**
90	Degrees	Right	F
60	Degrees	Right	F
75	Degrees	Right	F
-90	Degrees	Left	F
-60	Degrees	Left	F
-75	Degrees	Left	F
180	Degrees	Right	B
150	Degrees	Right	B
165	Degrees	Right	B
-150	Degrees	Left	B
-165	Degrees	Left	B
3:00	Clock Position	Right	F
2:00	Clock Position	Right	F
2:30	Clock Position	Right	F
9:00	Clock Position	Left	F
10:00	Clock Position	Left	F
9:30	Clock Position	Left	F
6:00	Clock Position	NA	B
5:00	Clock Position	Right	B
5:30	Clock Position	Right	B
7:00	Clock Position	Left	B
6:30	Clock Position	Left	B

Subject Pool B			
Angle	**Angle Unit**	**Left/Right**	**Front/Back**
15	Degrees	Right	F
30	Degrees	Right	F
90	Degrees	Right	F
-15	Degrees	Left	F
-30	Degrees	Left	F
-90	Degrees	Left	F
105	Degrees	Right	B
120	Degrees	Right	B
180	Degrees	Right	B
-105	Degrees	Left	B
-120	Degrees	Left	B
12:30	Clock Position	Right	F
1:00	Clock Position	Right	F
3:00	Clock Position	Right	F
11:30	Clock Position	Left	F
11:00	Clock Position	Left	F
9:00	Clock Position	Left	F
3:30	Clock Position	Right	B
4:00	Clock Position	Right	B
6:00	Clock Position	NA	B
8:30	Clock Position	Left	B
8:00	Clock Position	Left	B

Table 6.2 Angle distribution between the subject pools

6.5. Procedure

The experimental instructions that were given to the participants give a complete overview of the procedure that was followed in the experiment. I read aloud the following experimental instructions to each subject, ensuring that every participant received the same instructions about the task.

"In this experiment, we are going to assess the accuracy of your direction estimation by asking you to turn to a particular direction when you are blindfolded. This is the setup that we will be using for our experiment (see Figure6.1 for a photograph of the experimental setup). The big white circle you are seeing here has both clock face and degree measurements on its circumference. We will be using this circle to measure the accuracy of your direction estimation. The small circular rug you see in the center of this big circle will help you to keep yourself in the center of this circle, as you will be blindfolded during the trials.

Every trial begins with you facing the zero degree or home position as you see here. As I said before, you will be blindfolded throughout the trials. So, in order to help you orient yourself towards the home position we have this small handle bar set up for you. All you have to do is extend your arms until you find this handle bar. Once you grasp the handle bar with both your hands, make sure that your feet are parallel to your hands and also try to put your feet together as close as possible. This ensures that you are facing the home direction.

When you are facing this home location, I will ask you to turn to a particular direction. For example, I will say "turn to 30 degrees on your right" or I will say "turn to the 8 o' clock position". I will repeat the turning direction twice. Then I will confirm the same by asking you to repeat the instructions given to you. In this way, we can make sure that you clearly understood the instruction that was given to you. Then I will ask you to turn. Now you can turn to face the angle or clock face that was given to you in the instruction. Once you are done with turning, you should say, "Done". So that I can measure the angle to which you have rotated.

After I am done with making my measurements, I will ask you to rotate back to the home position. Now you can extend both your arms to reach this handle bar and as I explained before you should orient yourself toward this home location. This completes a trial.

We will be doing 24 trials like this, 12 for degree measurements and 12 for clock face positions. Before beginning your actual trials we will have 8 practice trials and you have to pass our learning criterion to continue with our experimental trials".

It is important to note here that we employed bodily rotation in subject's responses when they were executing the requested angular trials since this has demonstrated to be the best method to evaluate an individual's perception of direction when compared with other traditional means like finger pointing or using a rotating dial (Haber, Haber, Penningroth, Novak, & Radgowski, 1993; D. R. Montello, Richardson, Hegarty, & Provenza, 1999)

As mentioned in the script above, 8 practice trials were included in the study to help the participant in understanding the instructions and to give them some experience with the bodily rotation. The angles in the main experiment were not used in the practice trials. In the practice trials, there was also a learning criterion that the subjects had to pass in order to proceed to the main set of trials.

For passing the learning criterion in the practice trials, subjects had to turn within +/-15 degrees of the angle given in the instruction. For example, if the instruction was to turn to 60 degrees to the right (or 2'o clock), the subject must turn within a 30 degree error tolerance (45 to 75 degrees) to their right in order to pass the learning criterion for this angle. Subjects were not informed about the bounding limits of the angle with which they have to turn in the practice trials in order to pass the learning criterion. If the participant failed in a given angular trial, that particular trial was repeated after completing all other trials in that set. The process was repeated until the participant passed the learning criterion for each practice trial. It is important to note that there were no limits to the attempts taken by a participant to pass the learning criterion for any trial, but on an average they took 2 attempts to pass a trial.

During the practice trials, subjects were allowed to take a look at the direction to which they had turned by taking off their blindfolds after turning to the angle mentioned in the instruction. This procedure was to let the participants learn from their mistakes, if any, and to understand the relation between their rotational body movement and the angle to which they were instructed to turn.

Once the participant passed the learning criterion for every practice trial, actual trials were conducted. So every participant did 8 practice trials based on degree / clock face instructions followed by 12 actual trials based on degree/clock face instructions, with a break after 6 trials. Then they did another set of 8 practice trials based on clock face / degree instructions followed by 12 actual trials based on clock face/degree instructions. Thus, in total, they performed 40 trials each. The actual trials (a set of 12 trials for each measuring unit) were completely randomized within each stimulus block.

6.6. Results

The primary intention of this experiment was to compare the functional equality of angular measurement units, namely clock faces and degree measurements, which were seen in our previous indoor scene natural language descriptions. Altogether, there were 1280 trials involved in this study, collected from 32 subjects including the learning phase (512 trials, excluding the attempts taken to pass each trial) and the testing phase (736 trials).

The accuracy of direction estimation based on direction units was calculated based on Absolute Angular Offset (AAO), which is a measure of error calculated as follows,

Absolute Angular Offset = Abs (Actual Angle – Observed Angle)

Where Actual Angle is the angle to which the participant was asked to turn, while Observed angle is the angle to which the participant physically turned during the trials.

Figure 6.1 shows the distribution of AAO computed for clock face and degree units, for each angle that was considered in trials and figure 6.2 shows the confidence interval of the Average Absolute Error for each angle that was being tested. From these figures we can visually see that the angles that were presented behind the participants (in the rear facing 180 degree hemisphere) were estimated with higher absolute angular offset, e.g. higher angular error, when compared with the angles that were presented in the front facing 180 degree hemisphere.

After collapsing the data across front and back hemispheres, the t-test results suggest that presenting direction cues using clock face units yielded a statistically significant lower Absolute Angular error (9.71 ± 8.1 degrees) when compared with presenting direction cues using degree measurements (10.33 ± 8.87 degrees) [$t(368) = -2.49$, $p = 0.01$].

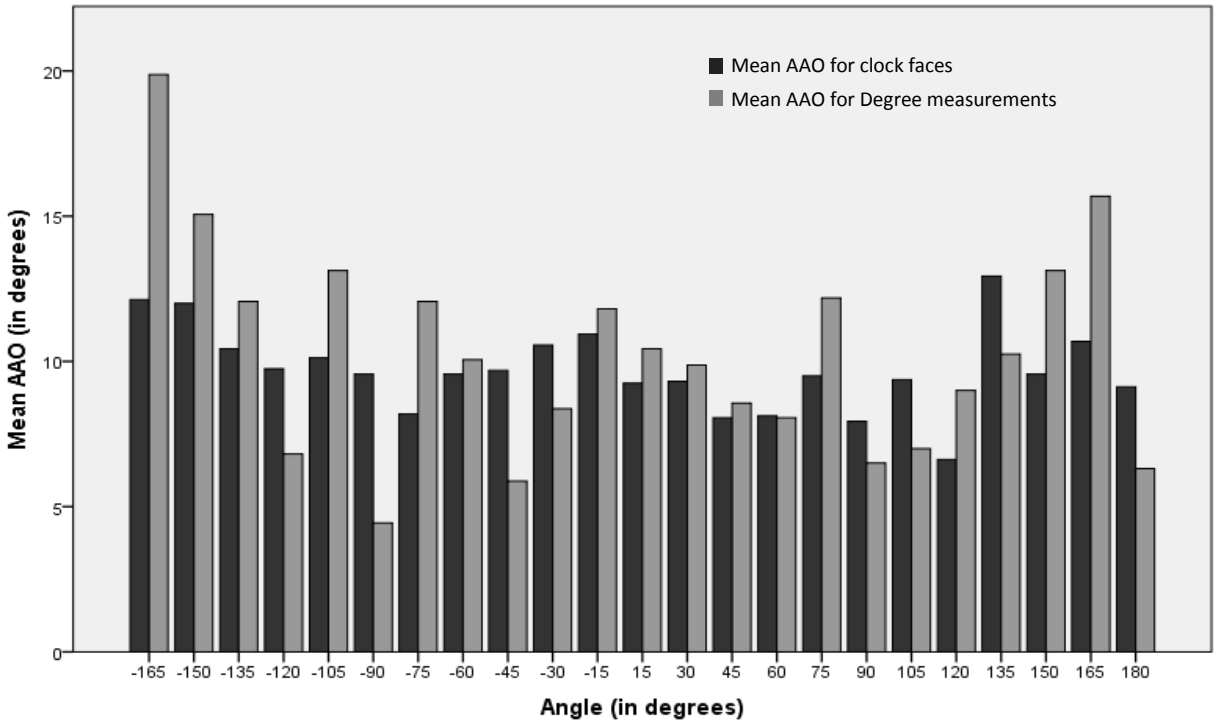

Figure 6.1 Average absolute errors based on units of angular measurements

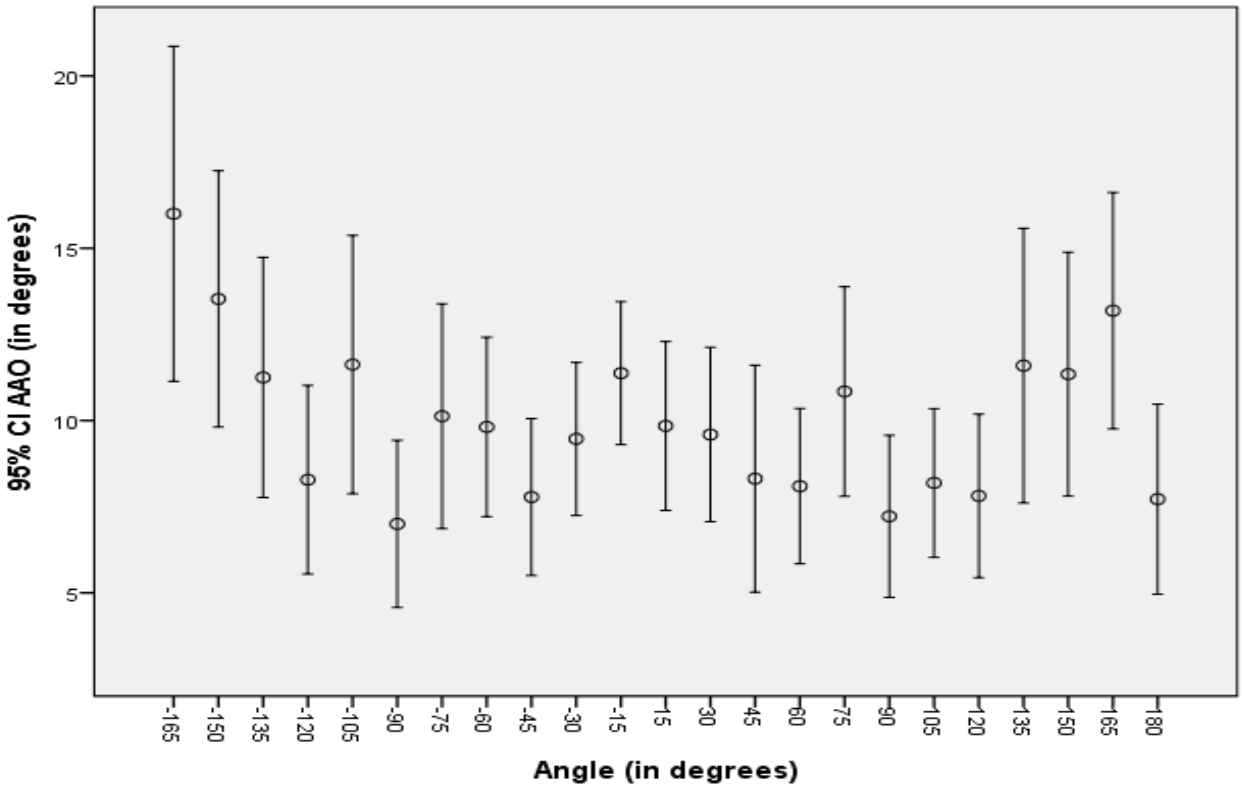

Figure 6.2 Confidence interval graph of absolute angular offset

A paired-samples t test was performed on the data to see the effects of direction units on the estimation of just the canonical angles, such as 30, 45, 60, 90, 120, 150 and 180 degrees (for both right and left 180 degree hemispheres). The results suggest that there is no statistically significant difference between clock face units and degree measurements in the estimation of canonical directions [$t(208) = 1.61$, $p = 0.1$].

Another paired t test suggests that presenting direction cues as clock faces had statistically significant lower Absolute Angular Offset (10 ± 8 degrees) in non-canonical angles, such as 15, 75, 105, 135, and 165 degrees (for both the right and left hand side) when compared with presenting direction cues as degree measurements (12 ± 10 degrees) [$t(160) = -5.537$, $p = 0$].

A One-way between-subjects ANOVA was conducted to compare the effect of the hemisphere (Front or Back hemisphere) on Average Angular Offset when directional cues were presented in clock faces and degrees across the full stimulus set. The results revealed that there was a statistically significant difference between groups, $F(3, 736) = 3.67$, $p = 0.01$. A Tukey post-hoc test revealed that there was a significant difference in AAO error between front and back hemispheres when the directional cues were present in degrees, while the angles in front and back were estimated with the same AAO (that is, with the same level of accuracy) when presented in clock face units.

A One-way between-subjects ANOVA was conducted to compare the effect of angle units on the cardinality of the angle (half-hour angles like 15, 45 degrees etc., versus full hour angles like 30, 60 degrees etc.) being measured based on the Average Angular

Offset recorded for each angle. The results suggested that there is no statistical difference between participant's estimation of half-hour angles and full-hour angles (15 vs. 30 degree increments) whether the instruction was presented in clock-face units or degree measurements, $F(3, 736) = 2.65$, $p = 0.85$.

Another One-way between-subjects ANOVA was conducted to compare the effect of angle units between the angles that are present in the right and left hemisphere. The results suggested that there is no statistical difference between participant's hemispheric estimation of angles whether the instruction was presented in clock-face units or degree measurements, $F(3, 736) = 4.19$, $p = 0.32$

As a part of the experiment, participants were asked to complete a questionnaire in which they had to select their preference for the angular measurement unit after the completion of all trials. Based on the graph shown in Figure 6.3 we can see that 25 participants (78% of the total sample) voted for use of Clock face units while the remaining 7 participants voted for Degree Measurements as their preferred metric of angular extent.

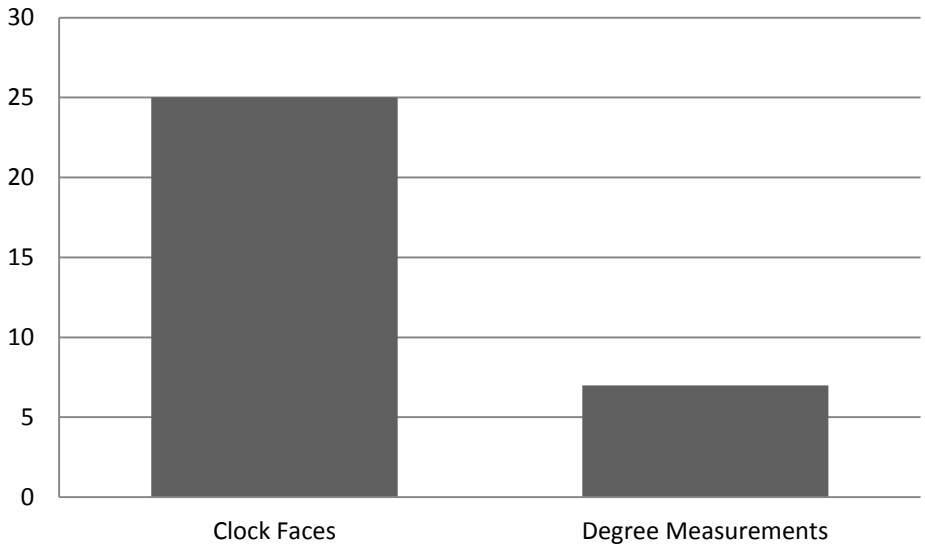

Figure 6.3 Participants preference for angular measurement units

6.7. Discussion

In this experiment, I evaluated participant's perception of direction when the directional cues were presented as clock-face values and degree measurements. The preliminary analyses showed that the unit of directional cues used to describe angular information significantly affects an individual's estimation of direction. Further analyses showed that the units of angular information do not have any significant effect in perceiving canonical angles, while the unit used to specify angular extent significantly affected the angular error of non-canonical angles. That is, the results suggest that using clock face units to describe non-canonical angles will significantly improve an individual's estimation accuracy of the direction being specified.

It was also seen from the results that the angles in both front and back hemispheres were perceived with the same level of accuracy when they were presented in clock face units. However, this was not the case with degrees, where the estimation of angles in

the rear hemisphere was significantly poorer when compared with the estimation of angles in the front facing hemisphere.

Combining the results suggested by the data from this experiment and also considering the participant preference rating, it can be concluded that using clock face units in a non-visual scene description system like DISc, will help the end users to perceive angular information with better precision when compared with presenting angular information in degree measurements. Also it can be seen from the results that with little training, participants found it meaningful and efficient when the clock-face measurements were presented in half-hour intervals (with 15 degree precision), thus providing a 100% increase to the traditional angular values being employed in linguistic descriptions and those tested in behavioral experiments. Hence these findings demonstrate that DISc and other linguistic description systems could describe space with increased angular granularity than is currently being done.

6.8. Chapter Summary

I began this chapter by reviewing literature related to direction perception and estimation and pointed out that there were no research projects that are focused on finding how an individual's direction perception is affected by the unit in which the angular information was presented. In this chapter, I discussed a behavioral experiment which compared the effects of clock-face information and degree measurements in direction estimation and found that providing directional cues as clock face units will

help the end users of DISc to perceive angular information with more precision than using degree values.

These findings will be useful for guiding both future development of DISC, as well as to other researchers who are interested in natural language for specifying spatial information.

7. Comparing Spatial Updating performance based on Indoor Scene descriptions

In the sample scenario discussed in section 1.1, Jack who is standing at the doorway of his tax-consulting firm and was expected to acquire indoor scene knowledge with the help of scene descriptions generated by DISc. In the previous chapters of this thesis, I compared linguistic elements of indoor scene descriptions by conducting behavioral experiments. The results suggested by those behavioral experiments would be used to design the future scene description component of DISc to generate indoor scene descriptions that are based on good spatio-cognitive principles, meaning that they are easy to apprehend and use.

Hence, Jack was expected to get a global spatial overview of the scene from a scene description, which will empower him to plan a route from his current position to the information desk in the lobby of his tax-consulting firm. As long as Jack's listening perspective is oriented with the intrinsic viewpoint of the indoor scene description, it is expected to be relatively easy for him to develop a cognitive map of other objects in the scene. However, after reaching the information desk, Jack will have a different viewpoint on the scene then was conveyed in the original description from the perspective at the door. The difference in viewpoints necessitates him to mentally rotate the orientation of objects that he learned from the original description to match his current orientation in the environment. This viewpoint offset, owing to his movement in the space, introduces cognitive load as mental rotations are known to be slow and to introduce error (Shepherd & Metzler, 1971).

Although the act of mentally rotating a scene and spatial updating are time consuming, these processes can be done using non-visual cues (Dodds, 1983; Marmor & Zaback, 1976). However, doing so without visual feedback is a slower and more error prone cognitive process, rather than being an automatic perceptual process as occurs with vision (Rieser, 1982; 1986). Therefore, Jack would likely be able to navigate in the lobby of his tax consulting firm based on the original description by mentally rotating and updating the scene based on his location, albeit it may be an effortful process. My interest however was to investigate whether the navigation process was affected by using descriptions that matched his changing viewpoint as he moved through the lobby. That is, is he better able to learn and navigate to locations when using an updated description, based on his current perspective versus using a fixed static description, given only from the doorway?

By definition, the updated scene description accounts for a user's current position and orientation. So I was interested to investigate whether an updated description from the information desk, or wherever he was in the space, would then help Jack to better understand the global structure of the scene and especially the egocentric relations between objects, which is necessary for performing route-finding tasks.

Hence I conducted a behavioral experiment to study the effects of using an updated scene description to acquire indoor scene knowledge and to plan a route-finding task as compared to planning the same route-finding task based on the indoor scene

knowledge acquired from a static (non-updated) scene description from a fixed viewpoint.

7.1. Motivation

Hollins and Kelley studied spatial updating abilities in blind and sighted people in the mid 1980s. The results of their behavioral experiments suggested that both sighted and non-sighted people demonstrated good spatial updating abilities (Hollins & Kelley, 1988). It is also known that after developing the internal representation of an object's location, spatial updating is independent of the modality in which the spatial knowledge was acquired (Loomis, Lippa, Klatzky, & Golledge, 2002).

Rieser and his colleagues emphasized the importance of spatial updating in navigation in 1982. In this work, they showed that a spatial updating task helped non-sighted people to understand the global layout of a scene, but even with this knowledge, non-sighted people found it difficult to accurately use global spatial information in terms of navigation from one point to the other within that scene (Rieser, Guth, & Hill, 1982). Later, an extensive follow-up study was carried out by the same authors and its results also suggested the same (Rieser, Guth, & Hill, 1986). In their behavioral studies, they guided their participants to different target objects from a fixed starting point and later the subjects performed a pointing task to different target objects while standing at the starting point and also while standing at one of the target objects.

In their case, the modality of spatial learning was discontinuous, where they learned the spatial location of each target object by means of separate guided walks from a starting

station. Also in their case, the mode of learning the global overview of the space was based on proprioceptive information obtained from bodily movement, rather than explicit verbal descriptions or other spatial cues about the overall configuration.

In this thesis, I was interested to investigate if providing participants with indoor scene descriptions will help them to understand the global structure of a scene in order to plan and perform navigation tasks. Also I was interested to see if updating the perspective adopted by the scene description, based on the user's current location and orientation, improves navigation performance compared to non-updated descriptions.

7.2. Experiment 6 – Comparing navigation performance based on static and updated indoor scene descriptions

In this experiment, I analyzed the efficiency of a participant's spatial performance (i.e., navigation) based on 3 conditions as described below.

7.2.1. Static Description condition

In this condition, participants were asked to navigate in a virtual indoor scene (see section 7.4 for more information about the virtual indoor scenes) based on a static description provided to them from a fixed point of view at the door of that scene.

7.2.2. Updated Description Condition

In this condition, participants were asked to navigate in a virtual indoor scene based on a verbal description that was updated with respect to their position and orientation in the scene.

7.2.3. Manual Exploration Condition

This is a control condition in which the participants were asked to manually explore a virtual indoor scene without any verbal description of the space to aid their navigation. After learning in all conditions, they were asked to navigate between multiple object pairs in the scene.

7.3. Participants

12 sighted native English speakers participated in this study (6 male subjects and 6 female subjects, mean age = 21.4).

The Institutional Review Board (IRB) of the University of Maine approved the study and on average, each subject took 90 minutes to complete the task. The participation of all subjects in the experiment was a voluntary decision made by them and every participant signed informed consent forms stating this right. The subjects were monetarily compensated for their time and effort to participate in the experiment.

7.4. Virtual Rooms

3 virtual indoor scenes (an office scene, a kitchen, and a hotel lobby) were designed for this experiment and each scene was composed of 5 different objects. The dimensions of the rooms were all the same and the spatial arrangement of objects was constant across all three rooms. Figure 7.1 shows the floor plan of the virtual rooms used in this study. For office scene A, B, C, and D shown in figure 7.1 was represented by the objects Trashcan, Desk, Chair and Bookshelf. For the Kitchen scene, the objects were Food

Processor, Refrigerator, Microwave and Stove. For the hotel lobby scene, the objects were Flower Pot, Fountain, Lamp and Television.

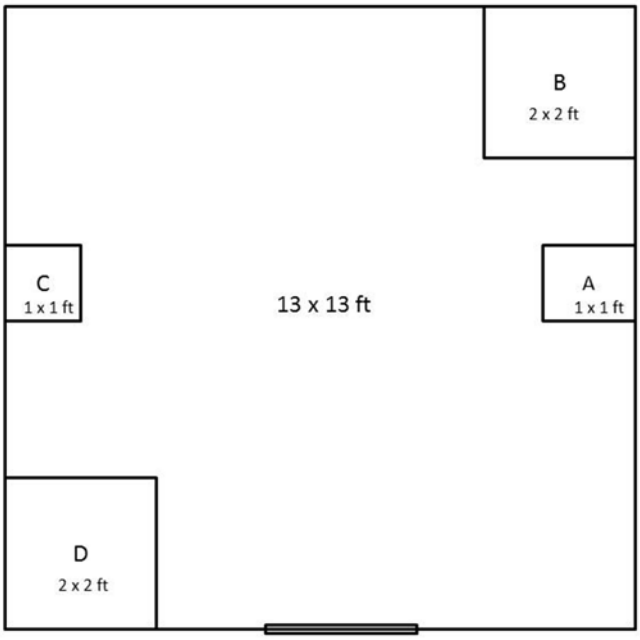

Figure 7.1 Floor plan of virtual rooms

7.5. Indoor Scene Descriptions

In order to describe the scenes, we need to linearize the 3D scene to a one dimensional natural language Description. Based on the results suggested by experiment 4 of this thesis, we know that the Center-Cyclic strategy helps to acquire an accurate global overview of the scene. Hence, for the purpose of this experiment we manually authored 5 indoor scene descriptions for each virtual scene following the center-cyclic linearization method (see chapter 5 for more information about Center – Cyclic

linearization mode). Also, the linguistic elements of the description were developed from the corpus of indoor scene descriptions collected in Experiment 1 (see section 4.3 of this thesis for more information about the corpus).

Out of 5 indoor scene descriptions, 1 description depicted the indoor scene from the standpoint of the door (used in the static description mode), while the other 4 descriptions described the indoor scene from the perspective of each of the five target locations in the room (used in the updated description mode).

The descriptions were carefully authored such that the spatial location of every object in the room was referred to with the same amount of references and the same kind of spatial language to maintain uniformity of information presented across all descriptions. In other words, the content and structure of the two description modes were equivalent but for the user's perspective adopted in the scene description,.

7.6. Apparatus Used

Creative HS-1200 (Creative Technology Ltd. USA) wireless headphones were used by the participants during the study in order to listen to the indoor scene descriptions and other instructions.

A PPT - Precise Position Tracker (Worldviz.Inc) was used to track the movement of subjects through the Virtual Environments used in the experiment, which was done in the VEMI lab (www.vemilab.org). A battery powered Light Emitting Diode was placed on the wireless headphones to track participant's motion using the PPT system. I used

python 2.4 (Python Software Foundation) in vizard 3.13 to render the virtual rooms described in section 7.5.

The Wiimote developed by Nintendo systems (Nintendo.Inc) was used by the participants to mark their response. I used a second Wiimote for myself while conducting the study in order to control the flow of the experimental sequences.

Before commencing the experiment, every subject was shown the Wiimote and the specific buttons in the Wiimote that they would use during the course of the experiment. I used 3D architectural rendering software called Room Arranger (Version 3.3). See section 4.2.3 for more information about the software.

7.7. Method

This experiment used a within-subjects design, where every subject participated in 3 different modes of the study. The 3 virtual indoor scenes (office scene, kitchen scene and a hotel lobby scene) were balanced across the three modes described in section 7.2.

In each mode, the participants had to walk to 5 different target locations in a room based on the instruction given in a pseudo random order and hence there were 5 trials in each mode. Two walking trajectories were designed in order to avoid learning effects across the conditions and these were balanced across the observation modes.

I also informed the participants that for each trial I would compute the thinking time (time taken by a participant to plan a walking route to the target object based on the instruction) and positional coordinates of their responses. The procedure for logging

these measures was the same across the three modes (Refer to the static description mode in section 7.7.1 for information about the procedure used for collecting thinking time and positional coordinate responses).

After completing all 5 trials in each condition, participants were asked to recreate the room they learned with the help of the Room Arranging Software (the protocol for this task was the same as discussed in full detail in section 4.2.). Also for each mode, participants were asked to fill out a NASA task load questionnaire and a user preference survey (refer to section 7.8 for more discussion about these two tasks).

7.7.1. Procedure

The procedures that were followed during each observation mode are discussed below.

7.7.1.1. Static Description Mode

In this mode, participants heard a scene description from a fixed perspective at the door of the scene. The description was always given from this perspective, e.g. was not updated with respect to the participant's location and orientation in the room. A sample scenario of this mode is described below.

A blindfolded sighted participant heard an indoor scene description when he was standing with his back against the door (point A) of a virtual indoor scene (see section 7.4 for more details about the structure of virtual scenes used). The spatial locations of 4 target objects that were present in the virtual indoor scene were conveyed in the scene description. At this point, the participant's location and orientation was in sync with the viewpoint provided in the description. When the participant finished listening

to the description, I played an audio instruction in a pseudo-randomized order. The audio instruction asked the participant to walk to one of the target objects described to him. For instance assume that the participant received an instruction “Walk to Desk”. After listening to this instruction, they mentally planned the route that they had to follow in order to reach the desk. They pressed the response button (Button A) in their Wiimote once they had imagined the requested destination and then started walking to the desk. The programming script automatically calculated the thinking time for planning the route by computing the time gap between the moment the participant got the instruction and till he pressed the response button in his wiimote. After reaching the place where the participant recalled the desk to be located, they hit the button a second time to indicate their response. This completed the first trial of the experiment.

From desk (point B) I walked them back to the starting point, i.e. door (point A) of the room to begin the second trial. From this point, they again listened to the same indoor scene description that they heard in their previous trial. Once they finished listening to the description, I walked them back to point B (Desk) and oriented them in a fixed orientation and position and verbally informed them of this information (i.e. there back was against the virtual object and they were facing the room).

They were then instructed to walk to the third target object (point C, which might be a bookshelf) from point B (which is desk), while I reminded them about the protocol for thinking and using the Wiimote [button A] as with the previous trial. It is important to note here that, the participant was instructed to walk to point C from point B, while the

indoor scene description provided spatial information of the object arrays only from point A. Thus, to perform this task, the participant needed to infer the travel route from their cognitive map, as the direct object relations were never specified from the description. There were 5 trials in this mode, each following the same procedure.

It can be seen from the experiment design that the indoor scene descriptions were static and were not spatially updated at the point from where the participant had to do his navigation planning (point B) in order to walk to the target location (point C).

7.7.1.2. Updated Description Mode

In the updated description mode, the indoor scene description used by the participant was always updated based on their location and orientation in the scene. A sample scenario is described below.

The first trial in the updated description condition was the same as the first trial in the static description mode, as described in the previous section. Let us assume that the target object in the first trial was desk. The subject was instructed to follow the same protocol for thinking and walking the route to the desk. After their response to mark the desk's location, they were walked to the actual location of the virtual desk (done if they made a mistake in indicating its location). Once at the target, all participants were rotated such that their back was against the virtual object and they were facing the room from this perspective. Therefore, this correction ensured that all participants were standing at a fixed position and orientation at the target before beginning the second

trial. They were also given verbal confirmation of their position and orientation at this target.

From this perspective, I played a description that was updated based on the location and orientation of the participant. That is, the point of view conveyed in the description matches with the point of view of the participant's location. In this case, the description narrated the scene with a viewpoint from the desk. Then they were instructed to walk to the third target object (point C, which might be a bookshelf) from point B (which is desk), while I reminded them about the protocol for thinking and using the Wiimote like the previous trial. It is important to note here that the participant was instructed to walk to point C from point B, while the indoor scene description was updated to the perspective of each starting target. This differs from the static condition, where descriptions were always provided from the point A position at the door. In this way they did not need to infer the inter-object relations, as the updated description specified the information directly from their current location. There were 5 trials in the updated description condition.

It can be seen from this experiment design that unlike the static description condition, the indoor scene descriptions were spatially updated at the point from where the participant had to do his navigation planning (point B) in order to walk to the target location (point C).

7.7.1.3. Manual Exploration Mode

This is a control condition in which the participants were asked to manually explore a virtual indoor scene without any assistance from verbal descriptions to learn the space or specify object position. This is analogous to Jack needing to learn the space with no navigation aid. Like the static and updated description conditions, there were five target objects in the scene including the door of the scene. It is important to emphasize here that I am talking about the virtual indoor scenes described in section 7.4 where all the objects were virtual in nature.

In the learning phase, the participants were asked to learn the indoor scene by physically walking in the virtual room without any spatial information about the scene being initially provided from external aids (e.g. verbal descriptions). For instance, in the case of the virtual office scene, the learning phase began when the blindfolded sighted participants were standing with their back against the door, while facing the scene. From that position, I informed the participant about their location and orientation. I also informed them about the size of the room that they will be manually exploring and the names of the objects they will encounter while exploring the scene (but no information about object position was given).

For the learning phase, I wrote a programming script that uses the PPT system to track the location of the participant in terms of their x and y coordinates in the virtual scene. The script also had pre defined location information about the virtual objects that were present in the scene. So whenever the participant intersected any virtual object's x-y

location, they were notified of the name of the object. For example, when the participant walked into the region where the virtual desk was located, the wiimote in his hand rumbled to indicate that he is standing at a location in which a virtual object was present and he heard an auditory signal giving its name through the headphones, e.g. “desk”.

It is important to clarify here that I am not interested in simulating mobility, e.g. obstacle avoidance or detection of obstructions, as would be done with a long cane or guide dog, but the learning of object locations, inter-object spatial relations, and a global representation of the space, e.g. the cognitive map. As such, I walked with the participants to serve as the “mobility aid” ensuring that they did not run into anything in the lab and letting them know if they walked out of the boundary of the virtual room.

There was no time limit to the learning phase. But on average participants took 20 minutes to learn the spatial location of 4 different target objects. Also, the participants were encouraged to walk back and forth between the targets in order to gain knowledge of the inter-object spatial relations. However, no specific strategies were given or hint as to where they were in the room or their relation to target locations. The learning phase was completed after they found the location of every object in the scene. They were given 20 minutes for exploring the scene, but it was not explicitly mentioned to them.

The testing phase began after all targets were learned. To perform, I guided the participants to the door of the virtual indoor scene and oriented them so that their back

was against the door. When the first trial began, I played an auditory instruction, for example “Walk to Desk”. I reminded them about the protocol for thinking and using the Wiimote like in the previous modes. Note that the audio label was not triggered at the target location, so they needed to indicate its location from memory based on the learning phase. A trial was completed once the participants mark their response. Participants performed 5 trials in this manual exploration mode.

Before beginning the second trial, I guided the participants to a fixed position and orientation at the desk and verbally informed them about their location. It can be seen from this experiment design that in contrast to the static and updated description conditions, the participants learned the object relations and global overview of the scene without any external aid (e.g., verbal scene description) and relied solely on proprioceptive information from body movement to gather information.

7.8. Results

There were 12 subjects who participated in this experiment and each subject performed 15 trials (5 for each condition). Hence there were 180 trials in total (60 trials for each condition). The data collected from each trial yielded the following information.

7.8.1. Thinking time

For each condition, the programming script calculated the time taken by the subjects to mentally plan the route to the target object. Out of the complete dataset, there were 180 thinking times (60 for each condition), 6 of these data points were more than 2.5 times the standard deviation of the overall mean thinking time of each condition

(approximately 3% of the data) and these were categorized as outliers. The outliers were then replaced with the corresponding subject's mean thinking time.

A One-way within subjects ANOVA showed that there was a statistically significant difference between the thinking times taken by participants in the three conditions, $F(2,22) = 2.32$, $P = 0.03$. A Tukey Post hoc test revealed that the thinking times of each condition significantly varied from each other, with the manual exploration requiring a reliably longer thinking time to plan routes to targets (M = 82 ± 6 seconds), while the updated description condition required the shortest thinking time to plan routes (M = 13±8 seconds).

A t-test was performed to compare the effects of the controlled condition (manual exploration mode) versus both the static and updated description modes for thinking time. The results showed that participants reliably outperformed the control with both description modes when compared to the manual condition (Manual Vs. Static mode: $t(60) = 8.31$, $p = 0.01$), Manual Vs. Updated description mode: $t(60) = 2.10$, $p = 0.01$)). When comparing the two description modes, the paired t-test results revealed that participants showed better performance with shorter thinking times when using the updated description condition (M = 13± 8 seconds), when compared with the static description condition (M = 37 ± 6 seconds), ($t(60) = 3.46$, $p = 0.02$).

7.8.2. Euclidean Distance Error

For each trial, the x and y coordinates of participant's target responses were recorded using the PPT system. The actual x and y coordinates of each object was determined

precisely using the PPT system and physical measurements. Based on this data, the Euclidean distance error for each trial was calculated using the formula given below.

Euclidean Distance error= sqrt((Actual x-coordinate of an object – Estimated x coordinate of an object)^2+(Actual y-coordinate of an object – Estimated y coordinate of an object)^2)

Hence there were 60 Euclidean distance errors calculated for each condition. Out of 180 Euclidean distance errors, 14 of them were more than 2.5 times the standard deviations from the mean Euclidean distance error of each condition (approximately 8% of the data) and these were categorized as outliers. The outliers were then replaced with the corresponding subject's mean Euclidean distance error value.

A One-way within subjects ANOVA showed that there was a statistically significant difference between the Euclidean distance errors calculated for the three conditions, $F(2,22) = 4.52$, $P = 0.01$. A Tukey Post hoc test revealed that the Euclidean distance errors of each condition significantly varied from each other, with manual exploration resulting in the greatest error ($M = 6.3 \pm 0.3$ feet), while the updated description condition showed the best results with the least error ($M = 1.7 \pm 0.5$ feet).

A t-test performed to evaluate the effects of the controlled condition versus both the static and updated description modes based on the Euclidean distance error showed that participants reliably outperformed the control with both the description modes when compared to the manual exploration mode (Manual Vs. Static Description mode: $t(60) = -8.32$, $p = 0.02$), Manual Vs. Updated Description Mode: $t(60) = -6.38$, $p = 0.01$)). When comparing the two description modes, the paired t-test results showed that

participants yielded the best performance with the least Euclidean distance error when using the updated description condition (M = 1.7 ± 0.5 feet), when compared with the static description condition (M = 3.9 ± 0.2 feet), (t (60) = 9.32, p = 0.03)).

7.8.3. Evaluation of Global Overview of the Scene

As mentioned in section 7.7.1, after each condition participants were assessed on their ability to acquire a global overview (that is, a mental map) of the indoor scene. In order to evaluate this acquired global overview, participants were asked to recreate the indoor scene they learned with the help of a room arranging software (described in section 4.2.3). The accuracy of the mental map, assessed through their physical re-creation, was evaluated by calculating the scaling, positioning, orientation and object retention errors of their recreations (refer to section 4.2.5 for more details about the evaluation of mental maps).

In order to find the combined effect of multiple variables like Scaling, Positioning, and Orientation errors on three different learning conditions, a one-way Multivariate Analysis of Variance (one-way MANOVA) was performed on the data. The results suggest that there was a statistically significant difference between learning conditions in obtaining a global overview of the scene, $F(2, 22) = 8.34$, $P < .05$, Wilk's $\lambda = 0.32$, partial $\varepsilon 2 = 0.83$. Power to detect the effect was 0.73.

A Tukey Post hoc test revealed that the accuracy of the mental map acquired as a result of all learning conditions significantly varied from each other, with manual exploration showing the least capability to help the participants acquire a global overview of the

scene, while the static description condition most helped the participants to acquire accurate mental maps.

A t-test was performed to assess the effects of the controlled condition versus both the static and updated description modes in terms of the knowledge gained about the global overview of the scene. The results showed that participants reliably outperformed the control with both the description modes (Manual Vs. Static mode: $t(12) = 5.19$, $p = 0.02$), Manual Vs. Updated description mode: $t(12) = 3.85$, $p = 0.03$)). When comparing the two description modes, the paired t-test results showed that participants developed a significantly better understanding about an indoor scene if they learned it using a static scene description compared with the updated description condition ($t(12) = 11.43$, $p = 0.01$).

The topological scores of each recreation also support this result. A t-test was performed on topological scores to assess the effects of the controlled condition versus both of the static and updated description modes in terms of the overall topological knowledge gained by the participants. The results showed that participants reliably outperformed the control with both the description modes (Manual Vs. Static mode: $t(11) = 5.8$, $p = 0.02$), Manual Vs. Updated description mode: $t(11) = -4.9$, $p = 0.01$)). When comparing the two description modes, the paired t-test results showed that participants demonstrated better understanding about an indoor scene topology if they learned it using a static scene description when compared with the updated description condition ($t(11) = 6.7$, $p = 0.01$).

7.8.4. Cognitive Load Estimates

The NASA Task Load Index (NASA-TLX) was used to procure the workload estimates of each subject while learning the room based on the different observation modes (Hart & Staveland, 1988). From the NASA Task Load Index sheet, subjective information about the mental demand, physical demand, temporal demand, performance, and effort and frustration while learning the indoor scene was calculated and this data is presented in the following chart.

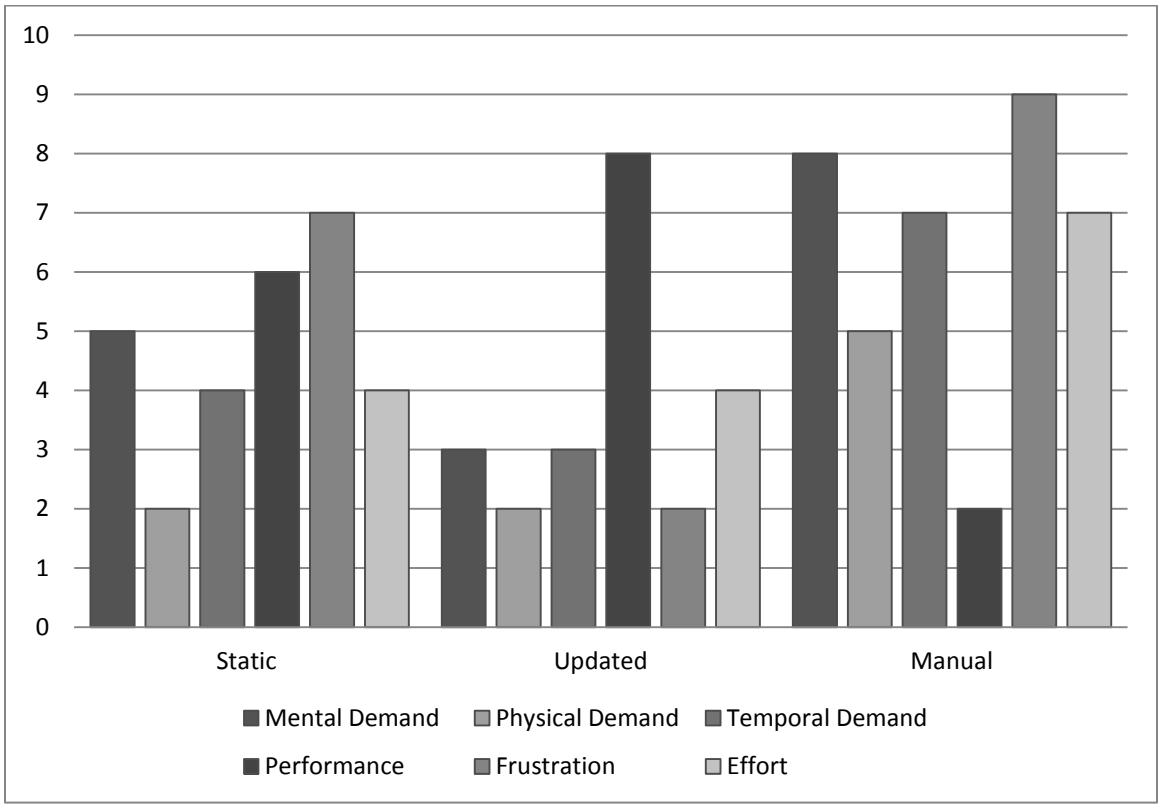

Figure 7.2 NASA task load Index for three observation conditions for scene learning

From the results of the Task load index we can see that the manual exploration mode requires the highest cognitive load while showing the poorest performance measures. In

general, we can see that the Updated description mode is the best and that the static description mode is intermediate, while manually exploring the scene is the worst method for learning and navigating an unfamiliar indoor scene.

7.8.5. User Preference Survey

At the end of the study, every participant was given a questionnaire in which they were asked to rank the modes (static description mode, updated description mode and manual exploration mode) for learning and exploring an indoor scene based on their experience during the course of the experiment. Out of 12 participants, 7 of them voted for the updated description mode and 5 of them voted for static description mode, while none of them voted for the manual description mode, as shown in the graph seen in figure 7.4.

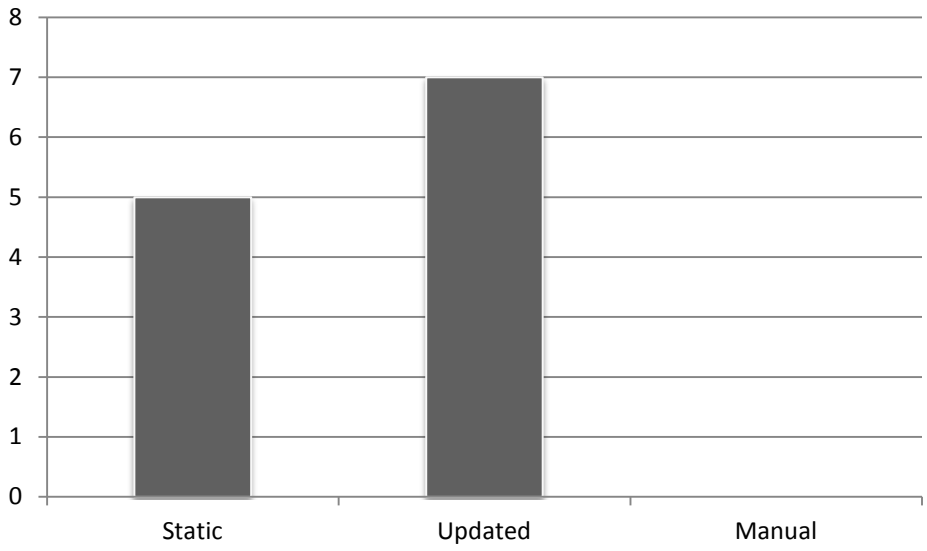

Figure 7.3 User preference survey for the 3 observation conditions used at learning

7.9. Discussion

Based on the results revealed by the data for thinking time, Euclidean distance error, estimated cognitive load, user preference surveys, and the accuracy of indoor scene recreations, the following interpretation are made about the description modes considered in this experiment.

7.9.1. Manual Exploration Versus Description Assisted Learning

One of the primary interests for conducting this experiment was to investigate whether the use of scene descriptions improved spatial learning and navigation of indoor scenes when compared with the manual exploration of the scene based only on body movement with no external aid. By analyzing the data about thinking time, it was clearly shown in section 7.8 that participants took significantly longer durations to plan a route from point A to point B when they manually explored an indoor scene, when compared with the other modes tested in this experiment. This finding suggests that manual exploration hindered participants ability to learn the spatial relations of objects and to gain an understanding of the global configuration of the scenes. This could have resulted in longer thinking times when the participants were asked to perform navigation process. That is longer thinking times generally mean more cognitive effort. This can be for many reasons. Sometimes, it may take longer but subsequent performance is okay, simply suggesting that the task is hard. But in this case, the subsequent navigation task was also worst for the manual condition, indicating that the longer times reflected increased cognitive effort due to an underlying lack of accurate spatial learning. Similarly, the fastest times for the updated condition led to the best

subsequent performance, supporting this correlation. This interpretation is also supported by the NASA load data, suggesting increased cognitive load/effort for the manual condition, which is corroborated by the user preference surveys.

These results indicate that participants did not gain much knowledge about the spatial location of objects using the manual exploration mode, although they walked and explored the spatial location of every object in the room before they were tested. This demonstrates the problem is not about encountering the objects but in encoding and learning them in a way that led to an accurate cognitive map that was accessible to subsequent navigation tasks and spatial behaviors.

These results provide unequivocal and positive answers to one of our research questions, namely whether using scene descriptions enhances the spatial behavior compared to using traditional, unaided techniques. The answer is clearly yes, as both description modes led to better performance than the unaided manual mode across all measures. These results affirm the efficacy and validity of further development of the DISc system.

7.9.2. Static Description Mode Versus Updated Description Mode

In the previous section (section 7.9.1) we saw that scene descriptions certainly assisted our participants to learn the spatial relations of objects and to gain a global spatial overview of an unfamiliar indoor scene when they have to navigate using non-visual conditions. But an important question addressed by this study was to assess which of the description modes led to the best performance and provided the most assistance.

By analyzing the data about thinking time, we can see that people planned the routes with significantly less time based on the learning supported by updated descriptions compared with the static descriptions. This result is likely because participants had to mentally update the spatial relations learned from a fixed point in the room with the static descriptions, whereas the updated descriptions explicitly conveyed the relevant spatial information, meaning that route planning was more direct and less computationally challenging

The additional mental rotation task mandated by the learning based on the static description mode also has the possibility to introduce errors. This was shown in the results of the positioning errors, which are a measure for target localization, and it was significantly larger for static descriptions when compared with the updated description conditions.

Also, the task load index showed that the static description mode required greater cognitive load when compared with the updated description mode and this result is likely due to the additional mental rotation task needed to perform the navigation task. So far, the updated description mode led to the best performance, least effort, and greatest help for the participants compared with the static description mode. This advantage is clearly verified by the results of the user preference survey, where more than 50% of the participants voted in favor of using the updated description mode to learn and navigate in unfamiliar indoor scenes.

So far the use of updated descriptions has showed to be better than using static descriptions to learn the spatial layout of an indoor scene. But the results from scene recreation in section 7.8.5 showed that static descriptions led to reliably better global recreation accuracy of indoor scenes when compared with the Updated description mode. It is important to note here that with the static descriptions, participants heard a single description of the scene (from the perspective of the door) for 5 times, while in the updated description mode, participants heard a new description every time whenever they reached a new target. This repetition in descriptions likely caused a bias in the resulting data for the static description condition as it was over-learned from one perspective; indeed, the same perspective that it was also re-created from.

Considering the clear advantages of using updated descriptions in terms of reduced cognitive load and better spatial performance, and given the uncertainty about the results from the scene recreation task, we argue that the updated description mode is the clear best choice for supporting the most intuitive and accurate indoor spatial behavior.

On the whole, based on the results discussed in this chapter, it is proposed that DISc should use updated descriptions so that its end users, like Jack, will be able to acquire the best global overview of the scene with the least cognitive effort.

7.10. Chapter Summary

I began this chapter by referring back to the sample scenario discussed in chapter 1. In that sample scenario I focused on the problem of spatial updating and described a

behavioral experiment conducted to analyze the importance of providing spatially updated descriptions. The results suggest the development of DISc would definitely help people to learn about the global spatial layout of an indoor scene in order to enhance the indoor spatial behavior in non-visual conditions. It is also seen from the data that updating scene descriptions based on the end user's location and orientation will help them to create a precise mental map of the indoor scene with the least cognitive effort.

8. General Discussion

In this chapter I will summarize the major research findings of this thesis and will then describe their importance with respect to current natural language Generation Systems. To help instantiate the benefits of applying the findings of this research in a real-world situation, I will refer to the sample scenario discussed in Chapter 1 again. But this time, Jack – the blind person who was standing at the doorway of his tax consulting firm's lobby in the sample scenario is now assumed to use an application called DISc, which is installed in his smartphone to learn and explore the lobby. It is important to remind the reader that the concept of DISc is still theoretical and the physical design and development of the system is beyond the scope of this thesis. This work is meant to show the efficacy of using natural language as the basis of such a system and to help guide the content and structure of the verbal information that would be used. For the sake of the current discussion, it is assumed that DISc is commercially available as a software application.

8.1. Jack with DISc

As mentioned throughout this thesis, DISc is a non-visual indoor scene description system that is used to obtain global scene knowledge. As it is a stand-alone system by design, photographs are the only information source to obtain spatial relations between objects. Compared to direct perception when observing a scene, photographs have their own limitations for conveying spatial information, as discussed in section 4.1 of this thesis. Despite these limitations, it is shown in experiments 1,2 and 3 of this thesis that

use of photographs resulted in equivalent performance in the ability to apprehend and use spatial information as did direct apprehension of the scene. These results provide compelling support for the use of photographs as the primary information source in DISc.

I also conducted a behavioral experiment to investigate whether the order of describing objects present in an indoor scene (e.g., description strategy) affected the end user's scene knowledge acquisition (refer to experiment 4 discussed in chapter 5 of this thesis). The results shown by this experiment were used to design the natural language Generation component of DISc. The results suggested that DISc should implement a Round-About strategy (as discussed in section 5.2.1 of this thesis) to describe the spatial location of objects (collected by analyzing the photographs of that scene) in the lobby of Jack's consulting firm, as it is demonstrated to be the best description order to assist indoor scene knowledge acquisition.

Also results from the direction estimation experiment (experiment 5 discussed in chapter 6 of this thesis) showed that people understood angular directions better when they were given as clock face units rather than degree measurements. The results of the same experiment suggested that with little training, participants found it meaningful and efficient when the clock-face measurements were presented in half-hour intervals (with 15 degree precision), thus providing a 100% increase to the traditional angular values being employed in linguistic descriptions and those tested in behavioral experiments.

Finally, I was interested to evaluate whether using DISc would really benefit people in a situation like Jack. So I conducted another behavioral experiment as discussed in chapter 7 of this thesis. In that experiment, I had a control condition of manually exploring the scene and the results supported the fact that the use of scene descriptions allowed the participants to navigate and acquire knowledge about the scene reliably more easily and efficiently than with the manual (unaided) mode. Hence DISc would certainly help Jack to navigate in the lobby, rather than trying to explore the scene manually by himself.

After confirming that the scene descriptions would help indoor spatial behavior in non-visual conditions, I analyzed two different modes to describe an indoor scene: a Static description mode and an Updated description mode (see section 7.2 for more discussion about these modes). Results from comparing these modes showed that the end users of DISc will benefit most when Updated descriptions are used for acquiring knowledge about the global spatial overview of indoor scenes. So for Jack, DISc updated the lobby scene description based on his position and orientation in order to support a more efficient navigation process.

8.2. Future Directions

As discussed in this thesis, the goal of my research was to optimize a natural language Scene description system by analyzing its user's performance for supporting indoor scene knowledge acquisition and subsequent spatial behavior. As natural language is flexible and versatile, there are a wide range of issues that could be considered for

refining the scene descriptions like sentence grouping, vocabulary, reference object selection etc., In this thesis, I have studied only a subset of elements associated with verbal scene descriptions, such as the description strategies used, precision of angular units described, and the effect of static vs. updated descriptions on several performance metrics. In the future, this research could be extended by analyzing factors associated with natural language scene descriptions like analyzing the sentence structure, range of vocabulary to be used, determining the content of the description, and so on in order to enhance user performance.

Also, I only considered indoor scenes with limited variations in size and complexities throughout all the behavioral experiments discussed in this thesis. It will be worthwhile to extend the research to analyze if we could generalize the results obtained from this thesis to other prototypic indoor scenes like kitchens, bedrooms, etc., It could be that the actual navigation performance might vary with the size and complexity of the environment.

natural language Generation architecture involves a procedural and formal way of arranging raw spatial information that must then be converted to a natural language (Reiter & Dale, 2000). To support this process, it is important to represent spatial information of indoor scenes in a formal structure, e.g. as an indoor scene ontology. Although we have ontologies available for characterizing indoor spaces in terms of corridors and pathways (Worboys, 2011), there are currently no ontologies available to represent indoor scenes. It is worth arguing for the importance of constructing an

indoor scene ontology that represents human described scene information. The primary goal of this ontology could formally reflect and represent the ways in which humans perceive space (from the above mentioned behavioral experiments) and to structure the relevant information into a robust and flexible natural language description. For example, the envisaged indoor scene ontology could involve a saliency rating of the objects that are typically present in an indoor scene. It should also be related with the existing linguistic ontology of space as proposed by (Bateman, Hois, Ross, & Tenbrink, 2010) in order to fill the gap between human perception and formal linguistic procedures used in NL generation.

8.3. Conclusion

We envisage the findings of this thesis as a first step towards an attempt to include indoor scenes in the current world of navigation systems. The research discussed in this thesis clearly showed that, when optimized, natural language is an effective communication medium to convey spatial information to support non-visual spatial learning and navigation in indoor scenes.

BIBLIOGRAPHY

A.veltman, M. (2010). *A framework for extracting spatial relation information from natural language spatial descriptions*.

Anacta, V. J. A., & Schwering, A. (2010). Men to the East and Women to the Right Wayfinding with Verbal Route Instructions. *Spatial Cognition VII Lecture Notes in Computer Science* (pp. 70–84).

Apple. (2012). Apple - iPhone 5 - Compare specifications between iPhone models. *www.apple.com*.

Ariadne, G. (2012). Ariadne GPS- mobility and map exploration for all. *iTunes Store*. Retrieved from http://www.ariadnegps.eu/

Avraamides, M. N., Loomis, J. M., Klatzky, R. L., & Golledge, R. G. (2004). Functional Equivalence of Spatial Representations Derived From Vision and Language Evidence From Allocentric Judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(4), 801–814. doi:10.1037/0278-7393.30.4.801

Barber, P. O., & Lederman, S. J. (1988). Encoding Direction in Manipulatory Space and the role of Visual Experience. *Journal of Visual Impairment & Blindness*, *82*(3), 99–106.

Bartlett, F. . (1995). *Remembering: A study in experimental and social psychology*. Cambridge University Press. Retrieved from http:// books.google.com /books?hl=en&lr=&id=WG5ZcHGTrm4C&oi=fnd&pg=PR9&dq=Remembering+:+A+Study+in+Experimental+and+Social+Psychology&ots=BBgYlAHqgJ&sig=VgiBijzK3xdIaIKNe94x98Ijjs

Bateman, J. a., Hois, J., Ross, R., & Tenbrink, T. (2010). A linguistic ontology of space for natural language processing. *Artificial Intelligence*, *174*(14), 1027–1071. doi:10.1016/j.artint.2010.05.008

Beard, K., Giudice, N. A., Latecki, L., Moratz, R., & Daniilidis, K. (2010). Perception of indoor scene layout by machines and visuall impaired users.

Bergen, B. K., Lindsay, S., Matlock, T., & Narayanan, S. (2007). Spatial and linguistic aspects of visual imagery in sentence comprehension. *Cognitive science*, *31*(5), 733–64. doi:10.1080/03640210701530748

Borenstein, J. (1997). The Guidecane - A Computerized Travel Aid for the active guidance of blind perdestrians. *Proceedings of the 1997 IEEE Intemational Conference on Robotics and Automation* (pp. 1283–1288).

Bowditch, N. (1802). CHAPTER 1 INTRODUCTION TO MARINE NAVIGATION. *The American Practical Navigator* (pp. 1–14). Deffense Mapping Agency Hydrographic/Topographic Center, Maryland.

Buchanan, B. G., Moore, J. D., Forsythe, D. E., Carenini, G., Ohlsson, S., & Banks, G. (1995). An intelligent interactive system for delivering individualized information to patients. *Artificial intelligence in medicine*, *7*(2), 117–154. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7647838

Casey, S. . (1978). Cognitive mapping by the blind. *Journal of Visual Impairment & Blindness*, *72*(8), 297–301.

Cheok, A. D., & Yue, L. (2011). A Novel Light-Sensor-Based Information Transmission System for Indoor Positioning and Navigation. *IEEE Transactions on Instrumentation and Measurement*, *60*(1), 290–299. doi:10.1109/TIM.2010.2047304

Chumkamon, S., Tuvaphanthaphiphat, P., & Keeratiwintakorn, P. (2008). A blind navigation system using RFID for indoor environments. *5th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*, *2*, 765–768. doi:10.1109/ECTICON.2008.4600543

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37–46.

Cornsweet, T. N. (1970). *Visual Perception*.

Dale, R., Geldof, S., & Prost, J. (2005). Using Natural Language Generation in Automatic Route Description. *Journal of Research and Practice in Information Technology*, *37*(1), 89–105.

Debnath, N., Hailani, Z. A., Jamaludin, S., Syed, I., & Kader, A. (2001). An electronically guided walking stick for the blind. *Proceeding of the 23rd annual EMBS Internation Conference,* (pp. 1377–1379).

Dekel, A., & Schiller, E. (2010). DRExploring Indoor Navigation with an Un - Augmented Smart Phone. *Proceedings of the 12th international conference on Human computer interaction with mobile devices and services* (pp. 393–394).

Denis, M. (1997). The description of routes: A cognitive approach to the production of spatial discourse. *Cahiers de psychologie cognitive*, *16*(4), 409–458. Retrieved from http://cat.inist.fr/?aModele=afficheN&cpsidt=2773586

Dodds, A. G. (1983). Mental Rotation and Visual Imagery. *Journal of Visual Impairment & Blindness1*, *77*(1), 16–18.

Ehrich, V., & Koster, C. (1983). Discourse organization and sentence form : The structure of room descriptions in Dutch. *Discourse Processes*, *6*(2), 169–195.

Ehrlich, K., & Johnson-Laird, P. N. (1982). Spatial Descriptions and Referential Continuity. *Journal of Verbal Learning Ad Verbal Behavior*, *21*, 296–306.

Eimer, M. (1996). Stereoscopic depth perception. *Handbook of Perception and Action* (pp. 71–102).

Faria, J., Lopes, S., Martins, H., & Barroso, J. (2010). ELECTRONIC WHITE CANE FOR BLIND PEOPLE NAVIGATION ASSISTANCE. *World Automation Conference* (Vol. 36, pp. 1–7).

Geomagic.Inc. (2012). PHANTOM OMNI - Sensable. *http://www.sensable.com/company-contact.htm*.

Giudice, N. A. (2004). *Navigating Novel Environments: A Comparison of Verbal and Visual Learning*. University of Minnesota, Twin Cities, MN.

Giudice, N. A., & Legge, G. E. (2008). Blind navigation and the role of technology. *Handbook of Smart Technology for Aging, Disability, and Independence* (pp. 479–500). Retrieved from http:// onlinelibrary.wiley.com/ doi/10.1002/ 97804 70379 424.ch25/ summary

Giudice, N. A., & Li, H. (2012). The Effects of Visual Granularity on Indoor Spatial Learning Assisted by Mobile 3D Information Displays. *Spatial Cognition VIII, Lecture Notes in Computer Science* (pp. 163–172).

Giudice, N. A., Walton, L. A., & Worboys, M. . (2010). The informatics of indoor and outdoor space: a research agenda. *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Indoor Spatial Awareness*, (November), 47–53. Retrieved from http://dl.acm.org/citation.cfm?id=1865897

Goldberg, E., & Driedger, N. (1994). Using natural-language processing to produce weather forecasts. *IEEE Expert*, *9*(2), 45–53.

Golding, A. R., & Leash, N. (1999). Indoor navigation using a diverse set of cheap, wearable sensors. *The Third International Symposium on Wearable Computers, 1999. Digest of Papers.* (pp. 29–36).

Golledge, R. G. (1991). Tactual Strip maps as Navigational Aids. *Journal of Visual Impairment & Blindness*, *85*(7), 296–301.

Golledge, R. G., Klatzky, R. L., & Loomis, J. M. (1996). Cognitive mapping and wayfinding by adults wihtout vision. *The construction of cognitive maps* (pp. 215–246).

Golledge, R. G., Marston, J. R., Loomis, J. M., & Klatzky, R. L. (2004). Stated Preference for Components of a Personal Guidance System for Nonvisual Navigation. *Journal of Visual Impairment & Blindness*, *98*(3), 135–147.

Grow, S. La. (1999). The use of sonic pathfinded as a secondary mobility aid for travel in business environments: A single-subject design. *Journal of rehabiliation research and development*, *36*(4), 333–340.

Guide dogs for the blind.inc. (2012). Fact Sheet : The Skills of a Guide Dog.

Haber, L., Haber, R. N., Penningroth, S., Novak, K., & Radgowski, H. (1993). Comparison of nine methods of indicating the direction to objects: data from blind adults. *Perception*, *22*(1), 35–47.

Hansen, S., Richter, K., & Klippel, A. (2006). Landmarks in OpenLS — A Data Structure for Cognitive Ergonomic Route Directions. *Geographic Information Science, Lecture Notes in Computer Science* (pp. 128–144).

Herzog, G., Sung, C., Andr, E., Enkelmann, W., Nagel, H., Rist, T., Wahlster, W., et al. (1989). *Incremental Natural Language Description of Dynamic Imagery* (Vol. 1).

Hirtle, S. C., & Mascolo, M. F. (1986). Effect of semantic clustering on the memory of spatial locations. *Journal of experimental psychology. Learning, memory, and cognition*, *12*(2), 182–189. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2939175

http://www.senderogroup.com/products/GPS/allgps.htm. (2013). Accessible GPS Products. *Sendero Group*.

Iordanskaja, L., Kim, M., Kittredge, R., Lavoie, B., & Polgu, A. (1992). Generation of Extended Bilingual Statistical Reports Interlingual Representation Realizer Design. *Proceedings of the 14th conference on Computational linguistics* (pp. 1019–1023).

Ishikawa, T., & Kiyomoto, M. (2008). Turn to the Left or to the West: Verbal Navigational. *Proceedings of the 5th international conference, GIScience*, *5266*, 119–132.

Ivanenko, Y., Grasso, R., Israël, I., & Berthoz, A. (1997). Spatial orientation in humans: perception of angular whole-body displacements in two-dimensional trajectories. *Experimental brain research.*, *117*(3), 419–27. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9438709

Jobson, D. J., Rahman, Z., & Woodell, G. a. (1997). A multiscale retinex for bridging the gap between color images and the human observation of scenes. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, *6*(7), 965–76. doi:10.1109/83.597272

Johannsen, K., Swadzba, A., Aiegler, L., Wachsmuth, S., & Ruiter, J. de. (2013). A Computational Model for Reference Object Selection in Spatial Relations. *Lecture Notes in Computer Science* (p. (in press)).

Klatzky, R. L., Marston, J. R., Giudice, N. a, Golledge, R. G., & Loomis, J. M. (2006). Cognitive load of navigating without vision when guided by virtual sound versus spatial language. *Journal of experimental psychology, Applied*, *12*(4), 223–32. doi:10.1037/1076-898X.12.4.223

Klippel, A., Hansen, S., Davies, J., & Winter, S. (2005). A HIGH-LEVEL COGNITIVE FRAMEWORK FOR ROUTE DIRECTIONS. *Proceedings of SSC 2005 Spatial Intelligence, Innovation and Praxis: The national biennial Conference of the Spatial Science Institute,*.

Klippel, A., & Winter, S. (2005). Structural Salience of Landmarks for Route Directions. *Spatial Information Theory, Lecture Notes in Computer Science* (pp. 347–362).

Kostopoulos, K., Moustakas, K., Tzovaras, D., Nikolakis, G., Thillou, C., & Gosselin, B. (2007). HAPTIC ACCESS TO CONVENTIONAL 2D MAPS FOR THE VISUALLY IMPAIRED. *Journal on multimodal user interfaces*, *1*(2), 13–19.

Laboratory, Z. S. (2013). Zoom H2 Handy recorder.

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174.

Lawson, G. D., & Wiener, W. R. (2010). Improving the use of hearing for orientation and mobility. In W. Wiener, R. Welsh, & B. Blasch (Eds.), *Foundations of orientation and mobility: Instructional strategies and practical applications* (3rd ed., pp. 91–117).

Levelt, W. J. M. (1981). The Speaker's Linearization Problem. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *295*(1077), 305–315. doi:10.1098/rstb.1981.0142

Linde, C., & Labov, W. (1975). Spatial Networks as a Site for the Study of Language and Thought. *Language*, *51*(4), 924–939.

Liu, H., Yang, X., & Jan, L. (2012). Dense Neighborhoods on Affinity Graph. *INTERNATIONAL JOURNAL OF COMPUTER VISION*, *98*(1), 65–82. doi:10.1007/s11263-011-0496-1

Long, R. ., & Giudice, N. . (2010). Establishing and maintaining orientation for mobility. *Foundations of orientation and mobility* (3rd ed., Vol. 1, pp. 45–62). Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Establishing+and+Maintaining+Orientation+for+Mobility#0

Loomis, J M, Klatzky, R. L., Golledge, R. G., Cicinelli, J. G., Pellegrino, J. W., & Fry, P. a. (1993). Nonvisual navigation by blind and sighted: assessment of path integration ability. *Journal of experimental psychology. General*, *122*(1), 73–91. doi:10.1037/0096-3445.122.1.73

Loomis, J., Lippa, Y., Klatzky, R. L., & Golledge, R. G. (2002). Spatial Updating of location specified by 3-D sound and spatial language. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(2), 335–345.

Loomis, Jack M, & J.Lederman, S. (1994). Tactual Perception. *Handbook of perception and human performance* (p. Chapter 31).

Loomis, Jack M, Marston, J. R., Golledge, R. G., & Klatzky, R. L. (2005). Personal Guidance System for People with Visual Impairment: A Comparison of Spatial Displays for Route Guidance. *Journal of visual impairment & blindness*, *99*(4), 219–232. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2801896&tool=pmcentrez&rendertype=abstract

Lovelace, K. L., Hegarty, M., & Montello, D. R. (1999). Elements of Good Route Directions in Familiar and Unfamiliar Environments. *Spatial Information Theory. Cognitive and Computational Foundations of Geographic Information Science - Lecture Notes in Computer Science Volume 1661* (pp. 65–82).

Lukianto, C., Christian, H., & Sternberg, H. (2010). Pedestrian Smartphone-Based Indoor Navigation Using Ultra Portable Sensory Equipment MTi-G. *International Conference on Indoor Positioning and Indoor Navigation (IPIN)* (pp. 1–5).

Ma, T., Latecki, L. J., & Sciences, I. (2011). From Partial Shape Matching through Local Deformation to Robust Global Shape Similarity for Object Detection. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Vol. 1, pp. 1441–1448).

Marmor, G. S., & Zaback, L. a. (1976). Mental rotation by the blind: does mental rotation depend on visual imagery? *Journal of experimental psychology. Human perception and performance*, *2*(4), 515–21. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1011000

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46. doi:10.1037//1082-989X.1.1.30

McNamara, T. P., & Diwadkar, V. a. (1997). Symmetry and asymmetry of human spatial memory. *Cognitive psychology*, *34*(2), 160–190. doi:10.1006/cogp.1997.0669

Millar, S. (1994). *Understanding and representing space: Theory and evidence from studies with blind and sighted children*. Clarendon Press. Oxford.

Montello, D., & Freundschuh, S. (1995). Sources of spatial knowledge and their implications for GIS: An introduction. *Geographical Systems*, *2*, 169 – 176. Retrieved from http://geog.ucsb.edu/~montello/pubs/gisintro.pdf

Montello, D. R., Richardson, a E., Hegarty, M., & Provenza, M. (1999). A comparison of methods for estimating directions in egocentric space. *Perception*, *28*(8), 981–1000. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10664749

Moratz, R. (2006). Representing Relative Direction as a Binary Relation of Oriented Points, 407–411.

Newcombe, N., Sandberg, E., & Johnson, S. (1999). What Do Misestimations and Asymmetries in Spatial Judgment Indicate About Spatial Representation? *Journal of Experimental Psychology: Learning, Memory & Cognition*, *25*(4), 986–996.

Nintendo.Inc. (n.d.). Wii.

O'Shaughnessy, B. (2012). *3-D Geometric Sensor-based Object Detection in Indoor Environments*.

Ogden, C. L., & Flegal, K. M. (2008). *Anthropometric Reference Data for Children and Adults : United States , 2003 – 2006* (pp. 1–45).

Ohkugo, H., Kamakura, K., Kitakaze, S., Fujishima, Y., Watanabe, N., & Kamata, M. (2005). Integrated wayfinding/guidance system using GPS/IR/RF/RFID with mobile device. *20th Annual CSUN Int Conf Technology and Persons with Disabilities, LA, USA*.

Passini, R., & Proulx, G. (1988). Wayfinding without Vision: An Experiment with Congenitally Totally Blind People. *Environment and Behavior*, *20*(2), 227–252. doi:10.1177/0013916588202006

Pei, L., Chen, R., Chen, Y., Leppäkoski, H., & Perttula, A. (2009). Indoor / Outdoor Seamless Positioning Technologies Integrated on Smart Phone. *First International Conference on Advances in Satellite and Space Communications, SPACOMM* (pp. 393–394). doi:10.1109/.11

Python Software Foundation. (n.d.).*Python Software Foundation*.

Raja, M. K. (2012). *The development and validation of a new smartphone based non-visual spatial interface for learning indoor layouts*.

Ran, L., Helal, S., & Moore, S. (2004). Drishti: an integrated indoor/outdoor blind navigation system and service. *Proceedings of the Second IEEE Annual Conference on Pervasive Computing and Communications*, 23–30. doi:10.1109/PERCOM.2004.1276842

Ratnaparkhi, A. (2000). Trainable Methods for Surface Natural Language Generation. *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference* (pp. 194–201).

Rehrl, K., Häusler, E., Leitinger, S., & Straße, J. H. (2010). Comparing the Effectiveness of GPS-Enhanced Voice Guidance for Pedestrians with Metric- and Landmark-Based Instruction Sets. *Geographic Information Science, Lecture Notes in Computer Science* (pp. 189–203).

Reiter, E., & Dale, R. (2000). *Building applied natural language generation systems*. *Natural Language Engineering* (Vol. 3). doi:10.1017/S1351324997001502

Rieser, J J, Guth, D. a, & Hill, E. W. (1986a). Sensitivity to perspective structure while walking without vision. *Perception*, *15*(2), 173–88.

Rieser, J J, Guth, D. a, & Hill, E. W. (1986b). Sensitivity to perspective structure while walking without vision. *Perception*, *15*(2), 173–88. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/3774488

Rieser, John J, Guth, D. A., & Hill, E. W. (1982a). Mental Processes Mediating Independent Travel: Implication for Orientation and Mobility. *Journal of Visual Impairment & Blindness*, *76*(6), 213–218.

Rieser, John J, Guth, D. A., & Hill, E. W. (1982b). Mental Processes Mediating Independent Travel: Implication for Orientation and Mobility. *Journal of Visual Impairment & Blindness*, *76*(6), 213–218.

Room Arranger. (n.d.).*www.roomarranger.com*.

Roush, W. (2010). The Story of Siri, from Birth at SRI to Acquisition by Apple---Virtual Personal Assistants Go Mobile. *Xconomy*. Retrieved from http://www.xconomy.com/san-francisco/2010/06/14/the-story-of-siri-from-birth-at-sri-to-acquisition-by-apple-virtual-personal-assistants-go-mobile/

Rowell, J., & Ungar, S. (2003a). The world of touch: an international survey of tactile maps. Part 1: production. *British Journal of Visual Impairment*, *21*(3), 98–104. doi:10.1177/026461960302100303

Rowell, J., & Ungar, S. (2003b). The world of touch: an international survey of tactile maps. Part 2: design. *British Journal of Visual Impairment*, *21*(3), 105–110. doi:10.1177/026461960302100304

Rowell, J., & Ungar, S. (2003c). The World of Touch: Results of an International Survey of Tactile Maps and Symbols. *Cartographic Journal, The*, *40*(3), 259–263. doi:10.1179/000870403225012961

Rowell, Jonathan, & Ungar, S. (2005). Feeling our way: Tactile map user requirement - A survey. *International cartographic conference* (p. retrieved from www.icaci.org).

Sadalla, E. K., Burroughs, W. J., & Staplin, L. J. (1980). Reference points in spatial cognition. *Journal of experimental psychology. Human learning and memory*, *6*(5), 516–528. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7430967

Serra, A., Carboni, D., & Marotto, V. (2010). Indoor Pedestrian Navigation System Using a Modern Smartphone. *Proceedings of the 12th international conference on Human computer interaction with mobile devices and services* (pp. 397–398).

Serway, R. A., & Jewett, J. W. (2000). *Physics for Scientists and Engineers with Modern Physics*.

Shepherd, R. N., & Metzler, J. (1971). Mental Rotation of three-Dimensional Objects.pdf. *Science*, *171*(3972), 701–703.

Shingledecker, C, A. (1983). Measuring the mental effort of blind mobility. *Journal of Visual Impairment & Blindness*, *77*(7), 334–339.

Struiksma, M. E., Noordzij, M. L., & Postma, A. (2009). What is the link between language and spatial images? Behavioral and neural findings in blind and sighted individuals. *Acta psychologica*, *132*(2), 145–56. doi:10.1016/j.actpsy.2009.04.002

Su, J., Rosenzweig, A., Goel, A., Lara, E. De, & N.Truong, K. (2010). Timbremap : Enabling the Visually-Impaired to Use Maps on Touch-Enabled Devices. *Proceedings of the 12th international conference on Human computer interaction with mobile devices and services* (pp. 17–26).

Taylor, H. A., & Tversky, B. (1996). Perspective in Spatial Descriptions. *Journal of Memory & Language*, *391*(3), 371–391.

Tenbrink, T., & Coventry, K. (2011). Spatial Strategies in the Description of Complex Configurations. *Discourse Processes: A multi disciplinary journal*, *48*(4), 237–266. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/0163853X.2010.549452

Thinus-Blanc, C., & Gaunet, F. (1997). Representation of space in blind persons: Vision as a spatial sense? *Psychological Bulletin*, *121*(1), 20–42. doi:10.1037/0033-2909.121.1.20

Tomko, M., & Winter, S. (2009). Pragmatic Construction of Destination Descriptions for Urban Environments. *Spatial Cognition & Computation*, *9*(1), 1–29. doi:10.1080/13875860802427775

Tversky, B. (1993). Cognitive Maps, Cognitive Collages, and Spatial Mental Models. *Spatial Information Theory A Theoretical Basis for GIS Lecture Notes in Computer Science* (pp. 14–24).

Ungar, S. (2000). Cognitive Mapping without Visual Experience. *Cognitive Mapping: Past Present and Future* (pp. 221–248).

Wang, J., Bai, X., You, X., Liu, W., & Latecki, L. J. (2012). Shape matching and classification using height functions. *Pattern Recognition Letters*, *33*(2), 134–143. doi:10.1016/j.patrec.2011.09.042

Wang, Xin, Matsakis, P., Trick, L., Nonnecke, B., & Veltman, M. (2008). A Study on how Humans Describe Relative Positions of Image Objects. *Headway in Spatial Data Handling-13th International Symposium on Spatial Data Handling* (pp. 1–18).

Wang, Xinggang, Bai, X., Liu, W., & Jan, L. (2011). Feature Context for Image Classification and Object Detection. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 961–968).

Wiener, W., Welsh, R., & Blasch, B. (2010). *Foundations of Orientation and Mobility: Instructional Strategies and Practical Applications* (3rd ed., p. 833). American Foundation for the blind. Retrieved from http://books.google.com/books?hl=en&lr=&id=hso50ocsEpsC&oi=fnd&pg=PR7&dq=Foundations+of+Orientation+and+Mobility:+Instructional+Strategies+and+Practical+Applications&ots=V_RwRnnbrS&sig=0jG6U-DDdfY72obVElbyy_pJ2oI

Williams, S., & Reiter, E. (2003). A corpus analysis of discourse relations for Natural Language Generation Conference Item. *Proceedings of the Corpus Linguistics 2003 conference*. Retrieved from http://oro.open.ac.uk/id/eprint/12456

Worboys, M. (2011). Modeling indoor space. *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Indoor Spatial Awareness - ISA '11*, 1–6. doi:10.1145/2077357.2077358

Worldviz.Inc. (n.d.). PPT X WorldViz.

www.vemilab.org. (n.d.). Vemilab.

www.viewplus.com. (n.d.). Braille Printers (Braille Embossers) and Assistive Technology Products.

Yang, X., Adluru, N., & Latecki, L. J. (2011). Particle filter with state permutations for solving image jigsaw puzzles. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2873–2880). Ieee. doi:10.1109/CVPR.2011.5995535

Yatani, K., Banovic, N., & Truong, K. N. (2012). SpaceSense Representing Geographical Information to Visually Impaired People Using Spatial Tactile Feedback. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 415–424).

Yu, J., & Habel, C. (2012). A Haptic-Audio Interface for Acquiring Spatial Knowledge about Apartments. In D. Szymczak & S. Brewster (Eds.), *Haptic and Audio Interaction Design, Lecture Notes in Computer Science* (Vol. 1, pp. 21–30).

Zanelli, G., Cappa, P., Petrarca, M., & Berthoz, A. (2011). Vestibular and proprioceptive estimation of imposed rotation and spatial updating in standing subjects. *Gait & posture*, *33*(4), 582–7. doi:10.1016/j.gaitpost.2011.01.013

BIOGRAPHY OF THE AUTHOR

Saranya Kesavan was born in Sulur, India on 14th June, 1987. She was raised in Coimbatore and she graduated from TKSM higher secondary school in 2005. She then attended College of Engineering, Guindy in Chennai, which is one of the four constituent colleges of Anna University, Chennai, India and graduated in 2009 with a Bachelor of Engineering in Geo Informatics. She worked as a Programmer Analyst in Cognizant Technological Solutions in India from December 2010 to June 2011. Saranya enrolled as a Master of Science graduate student in the Department of Spatial Information Science and Engineering at The University of Maine in Fall 2010. Saranya Kesavan is a candidate for the Master of Science degree in Spatial Information Science and Engineering from The University of Maine in August 2013.