# Paper Review: Assisted Guidance for the Blind Using the Kinect Device

Andrew DiStasi
Rochester Institute of Technology
Rochester, NY
add5980@rit.edu

## 1. PAPER & CONFERENCE OVERVIEW

For my review, I selected a paper entitled “Assisted Guidance for the Blind Using the Kinect Device.” This paper was submitted to the Seventh International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion Conference, taking place at the Universidade de Tras-os-Montes e Alto Douro (UTAD) in Vila Real, Portugal from December 1-3, 2016. I consider this to be a reputable venue as it is sponsored by the Association for Computing Machinery through the ACM’s International Conference Proceedings Series program. The ICPS program essentially offers sponsorship, cooperation, and ACM branding from the ACM to International Conferences that meet a series of reviews by ACM officials. All papers submitted to the conference are copyrighted and published by the ACM. Additionally, select papers from this conference are included in the Universal Access in the Information Society journal. Beyond that, this paper has been cited in other, related accessibility studies published under the umbrella of the ACM and was the recipient of funding from the Portuguese Foundation for Science and Technology, the national scientific funding body in Portugal. Essentially, I view this paper as coming from a reputable venue because it comes from an established, international conference that has been running for approximately a decade with sponsorship and cooperation from the ACM and its published in an official ACM proceedings journal.

This paper explores a real-time system that provides location based guidance with obstacle avoidance to blind users in an indoors environment. The system makes use of visual recognition of markers and obstacle detection and classification with Microsoft Kinect sensors to acquire RGB-D images. The user’s pose and location are estimated by combining marker information with GIS data [1]. This paper is relevant to my proposed research for two reasons. First, at its core, both topics focus on creating software that addresses accessibility issues for users, with their study addressing GIS guidance for blind users and my proposed study addressing speech technology for users who cannot fully use a traditional screen, mouse, and keyboard input system for computers. Second, both topics utilize Text-to-Speech systems as the primary means of User Interaction. Given that this study makes use of a Text to Speech system to relay all information to the user, both studies clearly share an interest in the utilizing speech to convey information to users in the most efficient manner possible. Additionally, the studies share a target set of users because both projects address similar accessibility issues with related solutions.

## 2. OVERVIEW OF THE WORK

The paper begins by discussing the current landscape of tools used to aid the 285 million people estimated to be visually impaired worldwide. The motivation for this work is simple: with such a large number of individuals adversely affected by vision impairment, aiding these users in more comfortably and safely navigating their surroundings becomes a very practical and worthwhile challenge. One of the guiding maxims of accessibility research is that we need to strive to help all users be able to interact with their surroundings or with applications in as typical of a manner as possible. Creating an application to assist vision impaired users in navigating their environment certainly falls within the bounds of these challenges. Traditionally, white canes and guide dogs have been the main tool aiding mobility among blind people, but there are several assistive technologies that have recently emerged as potential solutions to address the issue of providing interactive information about the surrounding environment to blind users. Historically, several different guidance systems for blind and visually impaired users have been proposed. Most of these systems were based on GPS technology, offering wayfinding and obstacle avoidance to blind individuals in outdoor environments. However, indoor environments drastically degrade the quality of the GPS signal, effectively making this method unsuitable for indoor use.

More recently, various computer vision techniques have been utilized to offer improved indoor functionality. The NAVI project used a fuzzy clustering algorithm to process and extract information from a series of images (generated from optical sensors). This information is mapped to a stereo acoustic pattern and transferred audially to the user to inform them about obstacles in their environment [2]. Other works have utilized stereovision to obtain 3D information of the surrounding environment, using computer vision algorithms to “detect and decode square fiduciary markers in real time with off-the-shelf camera phones” [1], and using RGB-D (Red, Green, Blue, and Depth) Cameras, such as the Microsoft Kinect, to simultaneously process color and depth images to allow for faster detecting, classifying, mapping, and alerting of obstacles in a blind user’s environment. Building on all of these previous approaches, this study proposes a system that offers obstacle avoidance in addition to using location information from optical markers placed on walls to simultaneously generate three types of information: location of nearby points of interest, detection and classification of obstacles which the user should be aware of, and Navigation information to reach a desired destination [1]. A combination of this information creates a more complete representation of the environment, allowing the system to convey a more realistic connotation of a blind user’s surrounding environment.

The proposed system features a chest-mounted Microsoft Kinect that supplies depth images and RBG images as input to a real-time tracking algorithm that identifies “a trained set of wall-mounted optical markers strategically placed throughout the building” [1]. The depth image is analyzed to detect and classify obstacles and offers a redundancy if optical markers are not detected by the RGB camera, increasing the inherent safety of the application. The RGB image processing allows for calculating and modeling the user's location, pose, and heading with respect to the markers mounted on the wall. This dual system more completely models the surrounding area with GIS data and, as previously mentioned, offers inherent redundancies to improve the safety of the application. Markers and Points of interest are placed at strategic points around the room and on any obstacles or otherwise significant points in the room. These markers serve as a geospatial reference to the system, conveying latitude, longitude, and orientation information. Points of interest are also tagged with an information field, generally describing the point, which is used in speech synthesis to create an audio message relaying that information when a user is near the POI.

When one of these markers or points is detected, the algorithm generates a transformation matrix that is populated with information that relates the marker's position relative to the user. The direction at which the user is "viewing" the object is calculated with the Euler rotation angle and is paired with the user's position. This data is used to provide directional information to the user regarding routing, obstacles, and Points of Interest within their "field of view." This information is translated into 8 possible directions relayed as speech: "Go Ahead," "Turn Slightly to the Left," "Turn Slightly to the Right," "Turn Left," "Turn Right," "Turn Back Slightly to the Left," "Turn Back Slightly to the Right," and "Turn Back" [1]. As a complement to the RGB-D Image based tracking, depth image processing is used in obstacle detection and depth processing. A depth image is processed and fed into a trained neural network designed to classify "line profiles" (the path directly in front of the user) into one of four classes: "No obstacles in the way (free path)," "Obstacle ahead," "Upstairs ahead," and "Downstairs ahead." These profiles, like the directions, are relayed as audio messages in response to real-time processing of the user's immediate environment.

The main method of interfaced with the system, and the part of this research that interests me the most, is an audio description generated through text-to-speech synthesis. The document shows an interaction system diagram that governs the user interaction workflow. The system continuously acquires images and identifies the type of image. If it's a depth image, it either alerts the user if there is an obstacle or says nothing and acquires a new image. If the image is an RGB image, it identifies if the image contains a Marker. If it does, it determines the Marker's purpose and reacts accordingly. If the marker is a Destination, it calculates the route and informs the user that they're reached the destination or about the next node direction towards the destination. If the Marker is a Point of Interest, it determines whether or not it is an obstacle or danger, alerting the user if it is. If it is not an obstacle, it prompts the user for a voice command, which it uses to inform the user about each nearby Point of Interest.

To evaluate their system, a prototype was developed in C#, using the Microsoft Kinect SDK 1.8 ("to control depth and RGB image acquisition by the Kinect Device" [1]), the open source library FANN – Fast Artificial Neural Network, and the open source library NyARToolkit (to implement the optical tracking algorithm). The system was deployed on a laptop carried in a backpack by the user, with the Kinect mounted on their chest, oriented diagonally downwards towards the ground. This Kinect orientation ensures proper detection of optical markers mounted on walls from one to four meters away and a vertical field of view ranging from 0.8 to 4 meters in front of the user. While this is still a limited field of view, it offers a much more complete, realistic imaging of the surrounding environment than traditional visual aid technology has in the past. Two blind subjects – one male and one female – were trained with the system and then asked to follow a series of predefined routes ranging from 16 meters to 48 meters in length within the "Engenharias I" building on the UTAD campus. The users had no familiarity with the building nor the routes they were asked to navigate. For each trial, a route was chosen at random and users were told to follow any given audio instructions and to move freely when they were not receiving any feedback. Each participant navigated these routes until they had successfully completed 5 trials (that is, reached the destination within the time limit for a given route). The total time for each route was recorded, and an average and standard deviation (in seconds) were calculated for the five successful trials. Route A (16 meters) had a time limit of 120 seconds and averaged 51.4 ± 31 seconds to reach the destination. Route B (42 meters) had a time limit of 180 seconds and averaged 114.8 ± 47.6 seconds to reach the destination. Finally Route C (48 meters) had a time limit of 180 seconds and took 73.25 ± 13.6 seconds to reach the destination.

Given these results, the authors feel that their system offers a reasonable degree of assistance to vision impaired users. They note that obstacles are almost always correctly classified and typically detected 2 meters ahead of the user's trajectory. However, the authors indicate that the participants felt that the text-to-speech system was not sufficient in conveying information and have considered a haptic interface to provide my comprehensive information in future development, before acknowledging that a test of larger scale is needed to effectively evaluate any proposed improvements to the system.

## 3. ANALYSIS

I have several initial questions regarding the work completed in this study. First, I would like to find out exactly how they conveyed the speech information to the user. The inclusion of an analysis on what type of language they used to convey the information (and specifically what phrases were generated by the system). Were all generated utterances relative to the user's position or where concrete distances used? Was distance conveyed in any way or was it a series of instructions without reference to distance (i.e. "Keep Going Straight" repeated)? I would also like to have seen the feedback from the users regarding the appropriateness of speech as a feedback method for the system. Knowing exactly what they felt was lacking or struggled with can offer valuable insight into what types of information and ways of conveying data are most useful for vision impaired users. With regards to the experiment participants, I would like to know what criteria they used for selecting users. There are drastically different levels of vision impairment, but the research made no reference to differentiating or selecting participants by the level of their vision impairment. To me, these are the main weaknesses of the paper. There seems to be an insufficient amount of testing performed and I feel some design motivations weren't fully fleshed out. I think the validity of the results suffers from the small dataset presented and that components of the application (such as the speech-to-text interface) could have been more strongly motivated through relevant stakeholder analyses or literature review.

I think a lot of these questions posed also constitute worthwhile future work to be done on this experiment. Most immediately, a greater number of trials with more routes over significantly more users would offers substantially more valuable data. A more convenient and portable version of the application, perhaps deployed on a smart phone or tablet could be implemented to measure the feasibility of utilizing the application on a device more appropriate for carrying around. However, the extension of this study I would most like to see is a study involving the most appropriate way to convey this geospatial data to the user. Given that the users indicated that the current text-to-speech implementation was insufficient, I am interested in seeing both what made it insufficient and what techniques can be used to better convey the data in a more meaningful way. I would like to see a comparison of user comfort between the haptic feedback system mentioned in the future work section and a more robust speech system, potentially some form of Intelligent Personal Assistant. A study of the appropriate data type and modality of interface to convey this kind of information could greatly enhance the usability of this application.

## 4. RELATED WORKS COMPARISON

This paper fits nicely into the established research in this field. Examining the core components of the knowledge offered by this study – a new computer vision application that dynamically processes multiple types of image data in parallel – shows that the paper has made use of relatively new technology to make a novel contribution to its field, despite the weaknesses in its testing. Most of the cited and related works are references to studies that have previously been done on similar applications exploring different types of computer vision or GIS applications. None of the related works I found dealt with a topic exactly like this, so it’s hard to say that this or any other work did it “better,” rather, it fit in as an initial exploration into a new subset of a relevant field. I think this paper serves as a strong starting point for research into indoor geospatial assistive technology. A strong theoretical foundation was established, and a reasonably functional prototype was demoed and tested. There is certainly room for more thorough testing and a refinement of the needs specific to this study, but I feel this paper offers a solid foundation to build upon in this field.

## 5. REFERENCES

[1] Filipe, V., et al., *Assisted guidance for the blind using the Kinect device*, in *7th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion (DSAI 2016)*. 2016: Vila Real, Portugal.

[2] Sainarayanan, G., Nagarajanand R., Yaacob, S. 2007. Fuzzy image processing scheme for autonomous navigation of human blind. *Applied Soft Computing*, 7(1), 257-264.