Safe and Efficient E-Wayfinding (SeeWay) Assistive Navigation for the Visually Impaired

Final Report

by

Dr. Bing Li
bli4@clemson.edu
(864)365-0649
Clemson University

Dr. Gurcan Comert
Benedict College

Dr. Johnell Brooks
Clemson University

Dr. Aries Arditi
Visibility Metrics LLC

October 2022

Center for Connected Multimodal Mobility (C^2M^2)

200 Lowry Hall, Clemson University
Clemson, SC 29634

DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, by the Center for Connected Multimodal Mobility (C^2M^2) (Tier 1 University Transportation Center) Grant, which is headquartered at Clemson University, Clemson, South Carolina, USA, from the U.S. Department of Transportation's University Transportation Centers Program. However, the U.S. Government assumes no liability for the contents or use thereof.

Non-exclusive rights are retained by the U.S. DOT.

ACKNOWLEDGMENT

The project is funded by the Center for Connected Multimodal Mobility (C^2M^2) (Tier 1 University Transportation Center) Grant from the U.S. Department of Transportation's University Transportation Centers Program.

Technical Report Documentation Page

1. Report No.	**2. Government Accession No.**	**3. Recipient's Catalog No.**

4. Title and Subtitle Safe and Efficient E-Wayfinding (SeeWay) Assistive Navigation for the Visually Impaired	**5. Report Date** 10/10/2022 **6. Performing Organization Code**
7. Author(s) **Bing Li**; ORCID: https://orcid.org/0000-0003-4987-6129 **Gurcan Comert**; ORCID: https://orcid.org/0000-0002-2373-5013 **Johnell Brooks**; ORCID: https://orcid.org/0000-0002-4732-8025 **Aries Arditi**; ORCID: https://orcid.org/0000-0002-1365-0788	**8. Performing Organization Report No.**
9. Performing Organization Name and Address Department of Automotive Engineering, Clemson University, Greenville, SC 29607 Department of Physics and Engineering, Benedict College, Columbia, SC 29204	**10. Work Unit No.** **11. Contract or Grant No.** 69A3551747117
12. Sponsoring Agency Name and Address Center for Connected Multimodal Mobility (C^2M^2) Clemson University 200 Lowry Hall, Clemson, SC 29634	**13. Type of Report and Period Covered** Final Report (August 2019 - July 2022) **14. Sponsoring Agency Code**

15. Supplementary Notes

16. Abstract

Despite its challenges, independent travel for blind and visually impaired (BVI) individuals is an essential component of quality of life, enabling travel to work and recreational activities. Autonomous vehicle technologies have the potential of meeting these challenges. However, efficiently and safely guiding BVI travelers between indoor environments and vehicles outdoors remains a key obstacle. In the future transportation system, assistive navigation technologies, connecting BVI travelers and vehicles, will be of extraordinary importance for BVI individuals in the context of social justice and health care/public health. Conventional research is mainly based on robotic navigation approaches through localization, mapping, and path-planning frameworks. They require heavy manual annotation of semantic information in maps and its alignment with sensor mapping. Inspired by the fact that we human beings naturally rely on language instruction inquiry and visual scene understanding to navigate in an unfamiliar environment, this study proposes a novel vision-language model-based approach for BVI navigation. It does not need heavy-labeled indoor maps and provides a Safe and Efficient E-Wayfinding (SeeWay) assistive solution for BVI individuals. The system consists of a scene-graph map construction module, a navigation path generation module for global path inference by vision-language navigation (VLN), and a navigation with obstacle avoidance module for real-time local navigation. The SeeWay system was deployed on portable iPhone devices with cloud computing assistance for the VLN model inference. The field tests show the effectiveness of the VLN global path finding and local path re-planning. Experiments and quantitative results reveal that heuristic-style instruction outperforms direction/detailed-style instructions for VLN success rate (SR), and the SR decreases as the navigation length increases.

17. Keywords The Blind or Visually Impaired, Assistive Devices, Scene-Graph Map, Vision-Language Navigation	**18. Distribution Statement** This report or any part of this report is restricted to publication until prior permission from the authors.

19. Security Classif. (of this report)	**20. Security Classif. (of this page)**	**21. No. of Pages**	**22. Price**
Unclassified	Unclassified	24	NA

Table of Contents

List of Tables

List of Figures

EXECUTIVE SUMMARY

Despite its challenges, independent travel for blind and visually impaired (BVI) individuals is an essential component of quality of life, enabling travel to work and recreational activities. Autonomous vehicle technologies have the potential of meeting these challenges. However, efficiently and safely guiding BVI travelers between indoor environments and vehicles outdoors remains a key obstacle. In the future transportation chain, assistive navigation technologies, connecting BVI travelers and vehicles, will be of extraordinary importance for BVI individuals in the context of social justice and health care/public health.

Conventional research is mainly based on robotic navigation approaches through localization, mapping, and path-planning frameworks. They require heavy manual annotation of semantic information in maps and its alignment with sensor mapping. Leveraging the state-of-the-art AI-based computer vision techniques as vision substitution for BVI travelers, we propose a Safe, Efficient and Electronic (E-) Wayfinding (*SeeWay*) assistive navigation solution for the transition between indoor-to-outdoor assistive navigation to facilitate BVI individuals to access autonomous vehicles. We aim to achieve our goal through the following research activities: To study and explore the needs, wants and concerns of BVI individuals between indoor environments (including home, school, work, etc.) and the locations where they will access future autonomous vehicles in each of those locations to allow for independent travel. Design criteria based upon a human factors analysis and the resulting models for our assistive navigation system for the BVI will be derived. To investigate multi-modal crowdsourcing-based visual simultaneous localization and mapping (SLAM) for BVI assistive navigation for the critical indoor-to-vehicle transition of the transportation chain. To design AI-based environment recognition technologies using cloud computing to detect potential traffic hazards (e.g. vehicles, motorcycles, bicycles, as well as typical traffic signs, benches, curbs, landscaping) and augment travel safety for BVI travelers.

Inspired by the fact that we human beings naturally rely on language instruction inquiry and visual scene understanding to navigate in an unfamiliar environment, this paper proposes a novel vision-language model-based approach for BVI navigation. It does not need heavy-labeled indoor maps and provides a safe and efficient assistive navigational solution for BVI individuals. The system consists of a scene-graph map construction module, a navigation path generation module for global path inference by vision-language navigation (VLN), and a navigation with obstacle avoidance module for real-time local navigation. The SeeWay system was deployed on portable iPhone devices with cloud computing assistance for the VLN model inference. The field tests show the effectiveness of the VLN global path finding and local path re-planning. Experiments and quantitative results reveal that heuristic-style instruction outperforms direction/detailed-style instructions for VLN success rate (SR), and the SR decreases as the navigation length increases.

CHAPTER 1
Introduction

According to the World Health Organization fact sheet, as of October 2021, there were 2.2 billion people have near or distant vision impairment worldwide [WHO, 2021]. Due to their visual impairments, these individuals face severe challenges in wayfinding and navigation. Although most visually impaired people are hardworking individuals, they are limited to specific job locations or are unable to maintain employment because of the challenges they face traveling independently, and BVI individuals are limited to working in jobs that are very close to available public transportation in terms of location and route schedule. As a result, this restricts the employment opportunities of BVI individuals. According to statistics published by the National Federation of the Blind, only 40.2% of the visually impaired population in the United States are employed and 90% of the world's visually impaired live in low socioeconomic conditions. The ability of visually impaired people to comfortably navigate and travel independently will enhance employment opportunities and foster personal fulfillment [WHO, 2011].

Leveraging advanced computer vision technologies, machine perception has shown great potential in augmenting assistive navigation. Our previous work developed mobile device-based assistive navigation for BVI individuals using visual mapping and conventional path planning [Li et al., 2018, 2016; Muñoz et al., 2017]. Motivated by the fact that human beings naturally rely on language instruction inquiry and visual scene understanding for navigation, we came up with the idea to bridge the gap between vision and language in assistive navigation. We conducted online interviews with ten BVI subjects (six females and four males, from a local community-based group of BVI adults) in the past year [Brooks et al., 2018] and discovered a need for a mobile phone with a user-friendly interface for language interaction.

In this study, we present our new BVI assistive navigation system *SeeWay*, navigating users by taking language instruction with visual perception as input for a machine learning model. It aims to apply language instruction as effective information to assist BVI navigation in real-world scenarios. As shown in Figure 1, the SeeWay system contains multiple functions in a portable device, such as scene-graph map construction, language interaction, visual localization, and local navigation, all of which contribute to the BVI navigation. To the best of our knowledge, this project:
(1) is the first BVI assistive navigation system that is able to take human-spoken instruction as input to navigate from place to place without a labeled semantic map, by integrating visual-language information with learning-based models. (2) newly presents the hypotheses of the correlation between navigational instruction styles and path-finding success rates. Our experiments revealed that heuristic-based instruction works best among all types. (3) proposes an effective global instruction-based path planner to improve the success rate of VLN assistive navigation with a local obstacle avoidance method for augmenting travel safety.

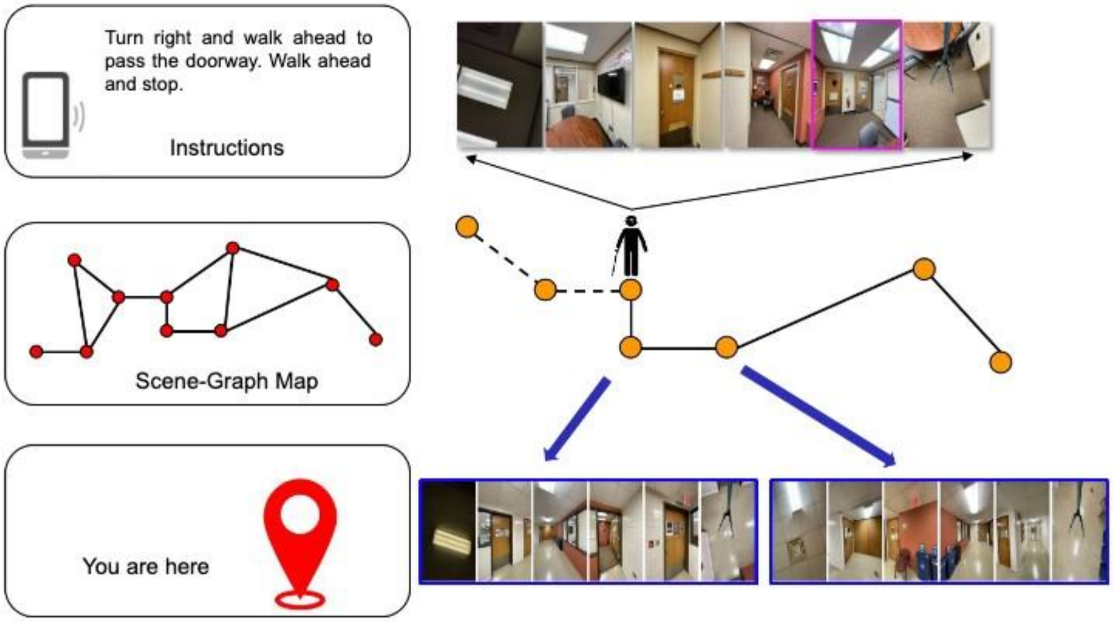

Turn right and walk ahead to pass the doorway. Walk ahead and stop.
Instructions
Scene-Graph Map
You are here

Figure 1: The vision-language assistive navigation vision for assisting BVI individuals. With voice instructions, a scenario scene-graph map and a visual localization system, it aims to search the path and navigate the BVI in complex indoor environments.

CHAPTER 2
Literature Review

2.1 Vision-Based Assistive Navigation

With the development of advanced 3D computer vision and deep learning, indoor localization and mapping techniques have been recently explored to promote the interpretation of visual perception and assistive navigation. There have been a variety of vision-based assistive navigation systems [Ahmed et al., 2019; Ali and Abou Ali, 2017; Balata et al., 2018; Caraiman et al., 2017; Cheng et al., 2018; Gomes et al., 2018; Islam et al., 2018; Kaul et al., 2021; Velázquez et al., 2018]. Ali [Ali and Abou Ali, 2017] designed a navigation system using a windowing-based means on a Microsoft Kinect camera. Ahmed [Ahmed et al., 2019] proposed a vision-based indoor navigation system for BVI people with machine learning algorithms for obstacle avoidance and object recognition. However, all of these navigation solutions are based on metric map searching, and essentially require building a global occupancy map.

2.2 Mobile Assistive Navigation

Mobile devices have been used for BVI assistive navigation in recent research such as [Bai et al., 2019; Kuriakose et al., 2021; Li et al., 2016; Oh et al., 2017; Zhang et al., 2019]. Our previous work [Li et al., 2016] designed an assistive navigation system on a mobile tablet. Zhang [Zhang et al., 2019] proposed an ARCore based user-centric navigation system with vision-based localization. These works motivate us to further our SeeWay system by leveraging the state-of-the- art mobile phone device with the RGB-D camera.

2.3 Vision-and-Language Navigation (VLN)

Unlike conventional approaches that use path planners to calculate paths on a reconstructed map, VLN models like [Anderson et al., 2018; Hochreiter and Schmidhuber, 1997; Ku et al., 2020] do not require any heavy-labeled indoor maps. Anderson [Anderson et al., 2018] introduced the Room-to-Room (R2R) task with the first indoor VLN solution. Based on the R2R dataset, Ku [Ku et al., 2020] further presented Room-Across-Room (RxR), a new VLN dataset that incorporates more languages and longer paths with instructions. These studies show the potential of designing and incorporating learning-based VLN models in BVI assistive navigation systems.

2.4 Robotic Assistive Navigation Platforms

Smart (robotic) wheelchairs have, in recent decades, become the subject of international research and development. Multiple groups are developing prototypes to test the latest input method or implement a challenging operating mode algorithm. See [Leaman and La, 2017] for a review of the smart wheelchair research between 2005 and 2015. See [Leaman and La., 2015] for a description of the intelligent power wheelchair we proposed in 2015. In 2017 researchers from the developing nation of Bangladesh built their Smart Wheelchair with windshield wiper motors, and sonar with an accelerometer for 2D mapping [Shahnaz et al., 2017]. Their prototype was able to follow a path that it had previously followed and recorded.

Also in 2018, researchers in Germany developed a SW with a GPS sensor, an absolute position sensor and medical sensors to monitor the user's ECG and blood pressure [Bumuller and Skerl, 2018]. By 2019 engineers in Lebanon had put together a fully autonomous SW using infrared computer vision [Alkhatib et al., 2019]. That same year a second team in Sri Lanka came out with a SW autonomous navigation as well as health monitoring sensors [Jayakody et al., 2019]. In 2020 a team of scientists in India made a SW that would autonomously follow a guide [Baiju et al., 2020]. That same year collaborators from Saudi Arabia, New York and California developed a vision based SW for hands-free mobility [Kutbi et al., 2020]. In 2021 a team from Romania used fuzzy control of the robotic arm for a smart electric wheelchair to assist people with movement disabilities [Pana et al., 2021].

A team in Japan developed a power wheelchair for the physically weak, who can still use their feet to control an accelerator and brake pedal [Woo et al., 2021]. Spanish researchers are working on virtual reality simulations of robotic wheelchair [Ortiz et al., 2021], and teams in Saudi Arabia and Bangladesh continued making strides toward the ultimate hands-free interface, namely brain control [Mohammad Monirujjaman Khan Shamsun Nahar Safa, 2021]. The most complete autonomous robotic wheelchair prototype of 2021 came from scientists in Korea [Ryu et al., 2022].

CHAPTER 3
Methodology

As depicted in Figure 2, the SeeWay system introduces a novel approach to bridge the VLN based global path searching and the local path re-planning for obstacle avoidance based on language instruction. It has the capability to run on a portable device, providing BVI navigation in a complex indoor environment without heavy labeling and 3D indoor model reconstruction. Some heavy tasks like global path searching are deployed on the cloud side.

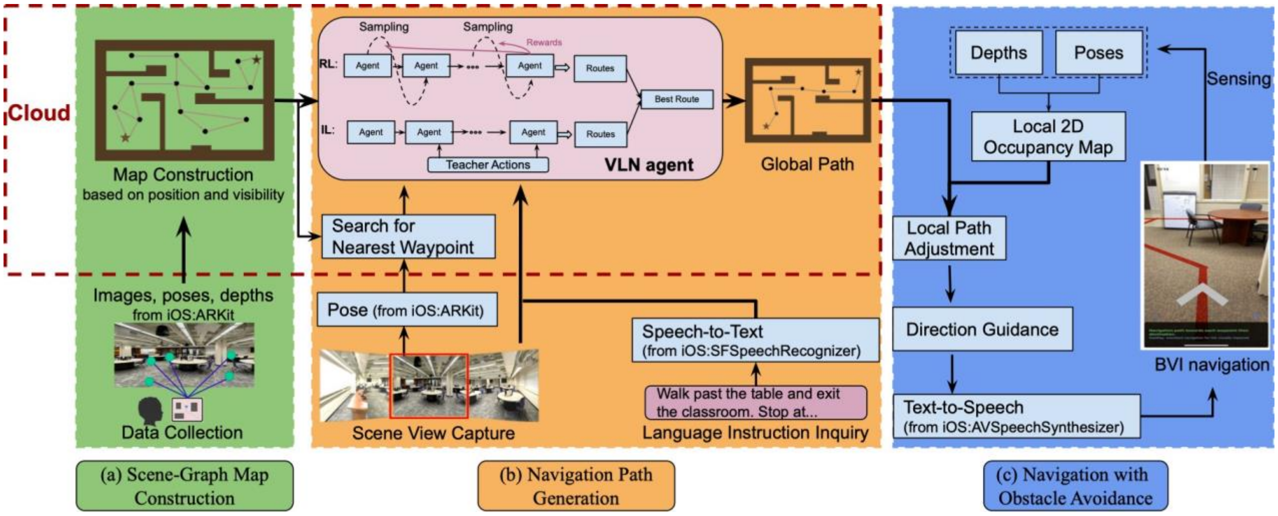


Figure 2: Overview of SeeWay system. (a) Data is collected on the edge side and transmitted to the cloud side for graph map construction. (b) Based on language instruction, current viewpoint and graph map, the VLN agent generates a global path for navigation. (c) Adjusts local path by the real scenario's occupancy map and generates BVI-applicable speech guidance.

The SeeWay system solution is based on three modules:

(a) a scene-graph map construction module that generates a 2D graph, where each node contains 6 poses (obtained from Visual-SLAM) and corresponding panoramic images (sky-box views);

(b) a navigation path generation module that infers the global path using language instruction, which is obtained from an inbuilt speech recognition system;

(c) the navigation with obstacle avoidance module that incorporates the searched path as a prior and performs online local obstacle avoidance for BVI navigation.

3.1 Scene-Graph Map Construction

As shown in Figure 2(a), the scene-graph map construction module consists of two parts. The first part is the data collection unit in which visual SLAM is utilized to obtain pose information, color panoramic images, and depth images from each viewpoint in the scenario. Each viewpoint can be represented by three features (6 poses, 6 sky-box views' color image, and depth image). The second part is the scene graph map construction unit that is based on the viewpoints' relative position and the `visibility` among viewpoints. The collected data (images, poses, depths) will be transmitted from the edge to the cloud to generate the scene-graph map. Moreover, the generated map with viewpoint features is loaded into a VLN simulation environment such as Matterport3D simulator [Anderson et al., 2018] for navigation purposes.

Visual-inertial SLAM: Apple's iOS:ARKit framework [“ARKit,” n.d.] estimates camera pose based on VI-SLAM [Mourikis and Roumeliotis, 2007], which combines the merits of an inertial measurement unit (IMU) and a visual camera to estimate the real-time 6-DoF pose. In this study, we use ARKit to provide the pose information to allow navigation in an environment, which is able to provide a real-time per-frame pose estimation.

Scene-graph Map Construction: In our system, we use an iPhone 12 Pro Max to collect the 6 sky-box images (up, front, right, back, left, bottom) and their corresponding poses via ARKit. The pose coupled color frames are represented by scenario viewpoint using the sky-box image of the viewpoint and the pose for the corresponding color image. We then pass each color image to a ResNet to extract the feature representation of each viewpoint. In order to make sure there is no view block between any two viewpoints or collision with the environment, we manually connect the viewpoints to generate the graph $G = \{V, E\}$, where V indicates vertices and E are the edges.

3.2 Navigation Path Generation

With a scene graph map constructed, we have a graph that contains viewpoints encompassing position and 6 skybox images. When navigating online, the SeeWay system first asks the BVI user to obtain language instruction to the target destination from people around. Once the literal instruction is converted to text by iOS:SFSpeechRecognizer [“Recognizing Speech in Live Audio,” n.d.] and transmitted to the cloud side along with current scene view and pose, SeeWay searches the nearest viewpoint in the graph map as the starting viewpoint by comparing the distance between its current position with each viewpoint.

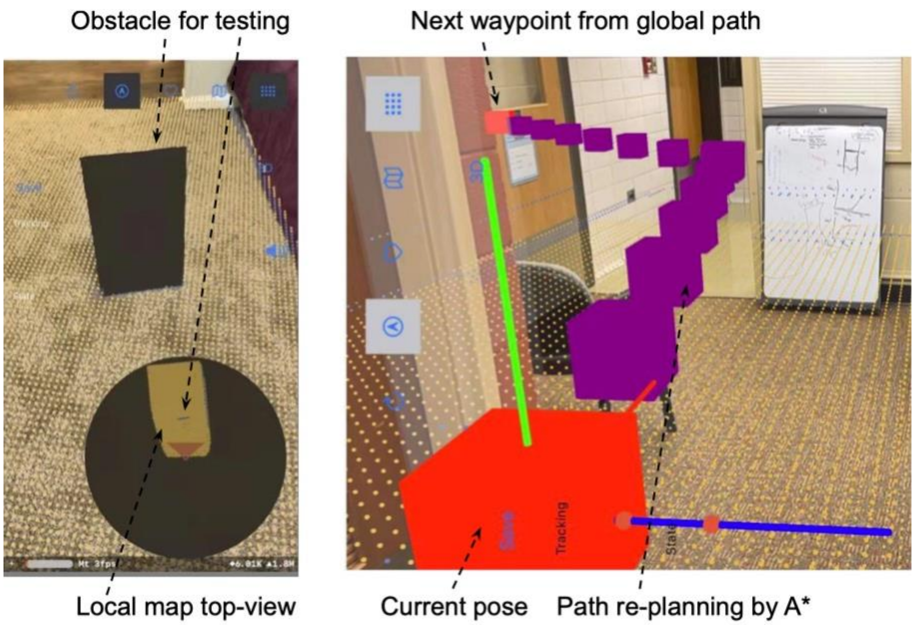
Then the VLN agent is activated to infer the global path based on the scene-graph map, starting viewpoint, and language instruction, see Figure 2(b). This study takes advantage of EnvDrop [Tan et al., 2019] to perform path exploration since it is capable of exploring in unseen environments with competitive leading performance at the challenging competition [“Leaderboard - EvalAI,” n.d.].

3.3 Navigation with Obstacle Avoidance

SeeWay navigates the BVI by taking the piece-wise linear path $\mathrm{P} = \{\mathrm{p}_0, \ldots, \mathrm{p}_{\mathrm{n}-1}\}$ generated by VLN as a global prior. For each part of the global path (p_i, p_{i+1}), we conduct the adaptive local navigation. The SeeWay system takes the obstacle mask M and back-projects based on depth,

$$R_{obs} = \cup \, [R, \, t] \cdot K^{-1} \cdot \left[u_{m_i}, v_{m_i}, 1\right]' \cdot d_{m_i} \qquad (1)$$

where $\mathrm{R}_{\mathrm{obs}} \in \mathrm{R}^3$ is the obstacle area, R, t are rotation and translation, K is camera intrinsic parameter, and $\mathrm{d}_{\mathrm{m_i}}$ is the depth value of a pixel $\left(u_{m_i}, v_{m_i}\right)$. Then we project the obstacles on top of the top-down view map to generate a local 2D occupancy map, by which the obstacles are highlighted along with path piece $(\mathrm{p_i}, \mathrm{p_{i+1}})$.



(a) Local 2D Occupancy Map (b) Local Obstacle Avoidance

Figure 3: SeeWay App GUI: (a) Local map example. (b) Local navigation example with path re-planning when encountering an obstacle (the chair) that conflicts with the navigability towards the next waypoint. Granularity is 0.3 m for (b).

Ideally, SeeWay will use the direct line-segment from viewpoint $\mathrm{p_i}$ to $\mathrm{p_{i+1}}$ to conduct navigation if no obstacles are detected on the track. Otherwise, it will execute the path re-planning method A* [Hart et al., 1968] to avoid the local obstacles. It is worth mentioning that the waypoints along $(\mathrm{p_i}, \mathrm{p_{i+1}})$ will guarantee a minimum safe radius r to the nearest obstacle, and this is implemented by setting a large-enough granularity for the local 2D occupancy map.

Figure 3(a) depicts the top 2D view of a local occupancy map through 3D depth perception and map projection. Figure 3(b) shows an example in which the local straight line segment collides with a chair and SeeWay is able to avoid a collision by re-planning a new path. Finally, we generate the navigation instruction considering the piece-wise path direction and the user's pose. For online BVI navigation (see Figure 2(c), we convert the navigation instruction to auditory output via iOS:AVSpeechSynthesizer [“Speech Synthesis,” n.d.].

3.4 Vision-Language Navigation

VLN focuses on the problem of taking the panoramic images and the language instruction as input to search for a feasible route from the start pose to the literal destination. In this study, we employ the EnvDrop model [Tan et al., 2019] to search a global path. Extending from the Speaker-Follower model [Fried et al., 2018], the EnvDrop model introduces two strategies to improve the model performance in unseen environments.

The first strategy is the mixture training of imitation learning and reinforcement learning. The second strategy is data augmentation, in which the model applies back-translation [Sennrich et al., 2015] to create new instructions based on extra routes shown in the new environment generated by the effective environmental dropout method.

VLN Agent Model: The VLN agent model uses a general encoder-decoder structure in which the encoder is a Bi-LSTM [Graves and Schmidhuber, 2005] using an embedding layer for instruction feature encoding.

The decoder is an attentive LSTM with an attention mechanism [Bahdanau et al., 2014] to predict the probabilities of all possible actions $a_{\&,Z}$ at time step t, that is, the next direction and action.

Given an instruction X, containing n words, i.e., $\mathrm{X} = \mathrm{x}_0,\ \mathrm{x}_1,\ \dots,\ \mathrm{x}_{n-1}$. The encoder computes an instruction feature representation of X with the following equation:

$$\mathrm{w} = \mathrm{w}_0,\ \mathrm{w}_1,\ \dots,\ \mathrm{w}_{n-1} = \mathrm{Bi} - \mathrm{LSTM}(\widehat{\mathrm{x}_0},\ \widehat{\mathrm{x}_1},\ \dots,\ \widehat{\mathrm{x}_{n-1}}) \qquad (2)$$

where $\widehat{\mathrm{x}_\mathrm{i}}$ is the embedding feature of x_i, w_i is the corresponding encoded instruction feature.

At each decoding step t, the decoder takes the attentive view feature and previous action a_{t-1}'s feature embedding as input, then applies the attention mechanism on the instruction features to generate the attentive instruction feature and predict the next step's action.

The attentive view features $\widehat{\mathrm{V}_\mathrm{t}}$ and attentive instruction feature $\widehat{\mathrm{w}_\mathrm{t}}$ can be calculated as follows:

$$\widehat{\mathrm{V}_\mathrm{t}} = \backslash \mathrm{cA}\left(\mathrm{V}_\mathrm{t}, \widehat{\mathrm{h}_{\mathrm{t}-1}}\right) = \sum_\mathrm{k} \mathrm{V}_{\mathrm{k,t}} \cdot \sigma_\mathrm{k}\left(\mathrm{V}_{\mathrm{k,t}}^\mathrm{T} \cdot \mathrm{W}_\mathrm{V} \cdot \widehat{\mathrm{h}_{\mathrm{t}-1}}\right) \qquad (3)$$

$$\widehat{\mathrm{w}_\mathrm{t}} = \backslash \mathrm{cA}(\mathrm{w}, \mathrm{h}_{\mathrm{t}-1}) = \sum_\mathrm{i} \mathrm{w}_\mathrm{i} \cdot \sigma_\mathrm{i}\left(\mathrm{w}_\mathrm{i}^\mathrm{T} \cdot \mathrm{W}_\mathrm{w} \cdot \mathrm{h}_{\mathrm{t}-1}\right) \qquad (4)$$

where σ is the softmax function. $V_t = V_{0,t},\ V_{1,t},\ V_{2,t},\ \dots$ represents the view features at time step t. h_{t-1} means the hidden output of decoder's LSTM at time step $t-1$. $\widehat{h_{t-1}}$ is the instruction-based hidden output at time step $\mathrm{t}-1$ generated by $\widehat{w_{t-1}}$ and h_{t-1}.

The probability of the next possible action $a_{k,t}$ at time step t can be obtained by using:

$$h_t = LSTM\left(\left[\widehat{V_t};\ \widehat{a_{t-1}}\right], \widehat{h_{t-1}}\right)$$

$$\widehat{h_t} = tanh(W \cdot [\widehat{w_t};\ h_t]) \qquad (5)$$

$$Prob\left(a_{k,t}\,\middle|\,t\right) = \sigma_k\left(G_{k,t}^T \cdot W_G \cdot \widehat{h_t}\right)$$

where $\widehat{a_{t-1}}$ means the embedding of the agent's previous action at time step $t-1$ which is generated in a similar way to instruction embedding $\widehat{x_\iota}, G_{k,t}$ means the view feature of the next k-th navigable viewpoint.

Loss Design: In this study, we adopt the mixed-loss design as introduced in [Tan et al., 2019], which utilizes both off-policy and on-policy optimization [Poole and Mackworth, 2010].

The loss is a combination of imitation learning [Bojarski et al., 2016] (off-policy) and reinforcement learning [Mnih et al., 2016] (on-policy), and is defined as:

$$\begin{gathered} Loss_{mix} = Loss_{RL} + \lambda Loss_{IL} \\ Loss_{RL} = \left(r(\hat{y}) - r(y^s)\right) sum_t(\log\{p(y_t^s | y_1^s, \dots . y_{t-1}^s, x)\}) \\ Loss_{IL} = sum_t(-\log p_t(a_t^*)) \end{gathered} \qquad (6)$$

here for $Loss_{RL}$, the $r(*)$ represents the reward function, $\hat{y}$ is the baseline output calculated by maximizing the output probability distribution, and y^s is the sampled output from distribution $p(y_t^s|y_1^s, \dots, y_{t-1}^s, x\}$. $Loss_{IL}$ tries to minimize the negative log probability of the imitated teacher's action a_t^*.

CHAPTER 4
Experiments and Results

In this section, we evaluate our proposed method and the SeeWay system with the Matterport3D public dataset as well as our field-collected Clemson-VLN dataset. In addition, the performance of VLN global path generation was evaluated quantitatively in an comparision study, and qualitatively with field tests [Yang et al., 2022].

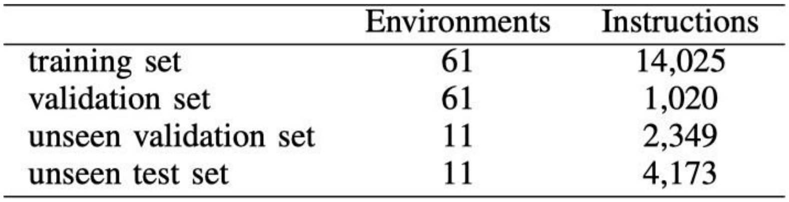
User Interviews: Ten subjects (six females and four males) have been recruited from a local community-based group of BVI adults [Brooks et al., 2022]. Their ages are between 32 and 77 with an average of 57.5. BVI user interviews were conducted online via Zoom for guiding our system design of choosing a 3D perception-enabled mobile platform (i.e., iPhone 12 Pro Max) with a speech-auditory interface.

4.1 System Setup and Data Statistics

The iPhone 12 Pro Max device equipped with the state-of-the-art mobile-inbuilt 3D LiDAR sensor allows us to perceive the 3D structure of the environment with pose estimation. We select AWS cloud as the cloud computing platform on which we use the AWS Lambda as the computing unit, AWS S3 bucket as the storage unit, and AWS Kinesis as the data transmission unit.

The SeeWay system is implemented by iOS/Swift and AWS Amplify to smooth the interaction between the edge device (iPhone 12) and the cloud (AWS). The VLN model embedded in the system was firstly trained with the R2R dataset in the visual RL environment--Matterport3D simulator. Then, the VLN model predicted the navigation path for both the Matterport3D and Clemson-VLN dataset based on the language instructions.

Table 1: R2R dataset statistics



	Environments	Instructions
training set	61	14,025
validation set	61	1,020
unseen validation set	11	2,349
unseen test set	11	4,173

Matterport3D contains a holistic RGB-D dataset for indoor scene understanding and 3D modeling, in which there are 194,400 RGB-D frames obtained from 90 building-scale scenes, with a total of 10,800 panoramic views. The R2R dataset is constructed by combining Matterport3D with 21,567 navigation instructions. During the VLN model training process, the R2R dataset (data statistic is given in Table 1 is split into the training set, seen validation set, unseen validation set, and unseen test set. The unseen sets contain environments that are never used to train the model.

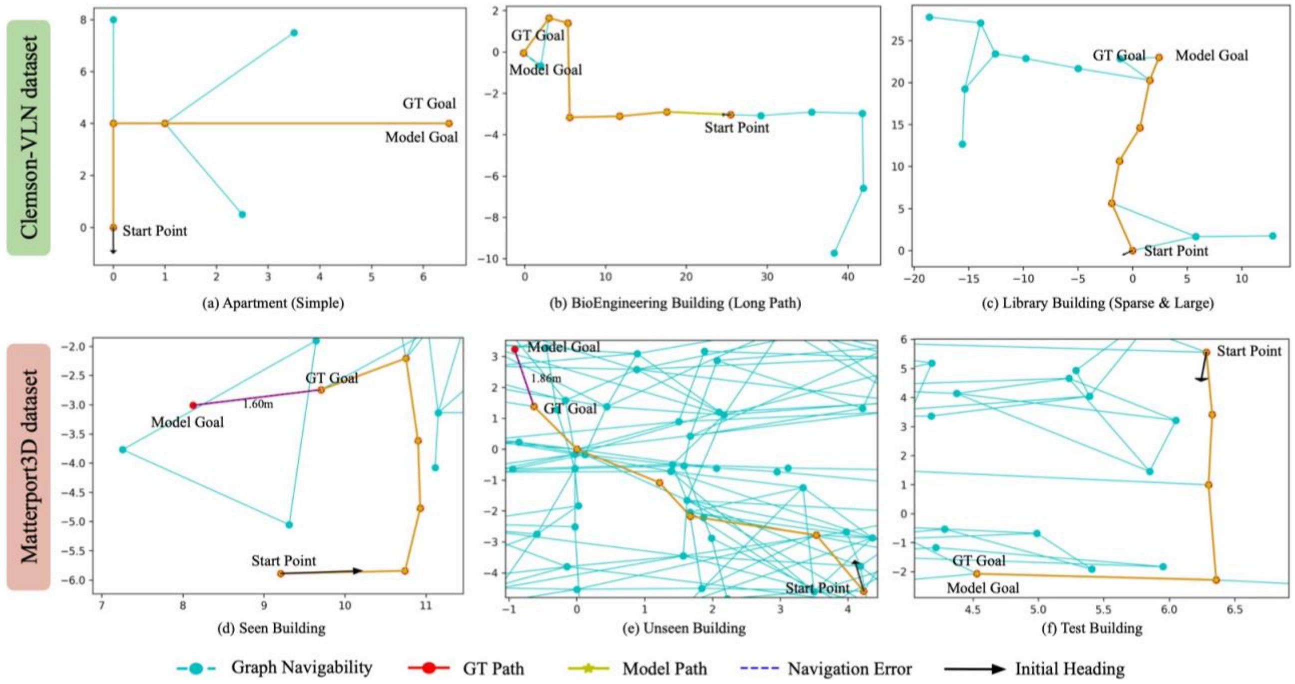
The Clemson-VLN dataset selected several environments as typical travel scenes, such as the student apartment for simple and small scenes, the 2nd floor of the Clemson BioEngineering Building (The Rhodes Research Center) for relatively long path scene, and the 2nd floor of the

Clemson Cooper Library as an open and large scene.

4.2 Scene-Graph Map Results

Our scene graph map was created based on the visual SLAM positioning and connectivity between viewpoints. One of the most advantageous features of this map is that it does not require any semantic label or map coordinate alignment, while the conventional robotic navigation approach uses global occupancy for path planning and needs semantic labeling and coordinate alignment between global occupancy and semantic info [Li et al., 2016].

4.3 Navigation Path Evaluation



(a) Apartment (Simple)

(b) BioEngineering Building (Long Path)

(c) Library Building (Sparse & Large)

(d) Seen Building

(e) Unseen Building

(f) Test Building

Figure 4: The effectiveness of the VLN model on the Clemson-VLN and Matterport3D datasets.

Navigation Path: Once the BVI uses the SeeWay application running on the iPhone 12 Pro Max in a real-world scenario, the SeeWay system will automatically generate a navigation path after language instruction inquiry. Figure 4(d) shows an example of a single navigation path. The cyan line indicates the overall scene-graph map of this scenario. The red line and yellow line indicate the ground-truth path and the VLN agent path, respectively. The dashed blue line reveals the navigation error between the ground-truth path and the VLN agent path. The black arrow indicates the initial heading of the BVI. As shown in the figure, the ground truth path and VLN agent path do not completely overlap. Considering the navigation error between their destination nodes is only 1.60 m, we can treat the VLN agent path as a success path.

Effectiveness of VLN Agent: Now we further validate its performance in the real-world scenarios (Clemson-VLN dataset) and the Matterport3D dataset with customized instructions. Figure 4 shows the VLN agent's results on both Clemson-VLN and Matterport3D datasets. For the Clemson-VLN dataset, all three typical buildings (Apartment, Bioengineering Building, Clemson Cooper Library Building) are selected for testing. For the Matterport3D dataset, model effectiveness is demonstrated in both seen building (seen in training), unseen building (used in validation), and test building (used in the leaderboard). The model navigation path mostly overlapped with the ground truth path.

Time Cost: In SeeWay module (b), the average time for global path inference is 0.2 s. In SeeWay module (c), the average refresh rate for local path adjustment is 3 Hz, which ensures BVI navigation needs. With a 2.5 Mb/s internet speed, all data transmission tasks between edge and cloud will be completed within 80 ms for the module (b) and (c).

4.4 Ablation Experiments

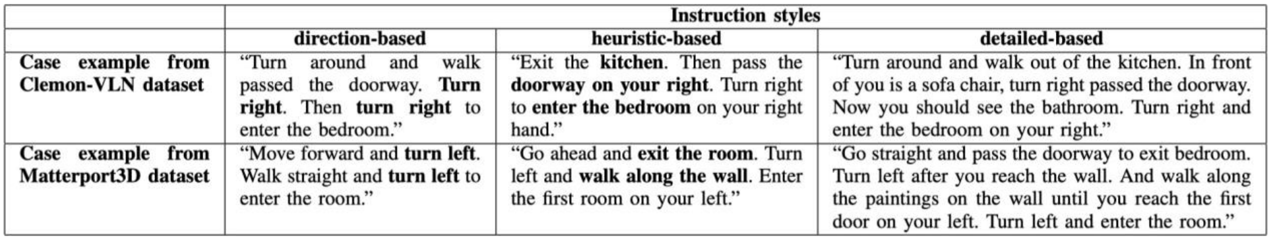
According to experiment results with different language instruction styles and different navigation paths, instruction style and the navigation length were found to have significant effects on the navigation result.

Evaluation Metrics: To quantitatively evaluate the VLN agent's navigation performance, we use multiple metrics such as Success Rate (SR), Navigation Error (NE), Navigation Length (NL), etc. NE relates to the distance between destinations of ground truth path and the agent's predicted path. According to [Anderson et al., 2018], the agent's navigation path is considered a success when NE is less than 3 meters.

In terms of *instruction styles*, when predicting navigation path with the trained VLN agent, the agent was affected by object phrase listed the language instruction, which may guide the agent to a false destination if that destination has some similar object features to the target location. To clearly understand how this affects the VLN agent's performance, we designated three specific instruction styles to test the VLN agent. They are direction-based instruction, heuristic-based instruction, and detail-based instruction.

Direction-based instruction mainly contains concise direction-related features but few object features. The heuristic-based instruction is intuitive and described according to human understanding. The detail-based instruction includes both directional features and object features of each scene point.

Table 2: Examples of different instruction styles.

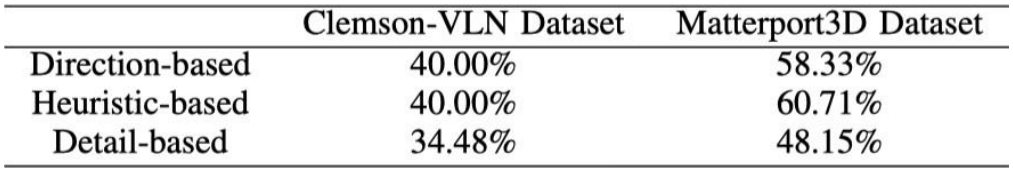


	Instruction styles		
	direction-based	**heuristic-based**	**detailed-based**
Case example from Clemon-VLN dataset	"Turn around and walk passed the doorway. **Turn right**. Then **turn right** to enter the bedroom."	"Exit the **kitchen**. Then pass the **doorway on your right**. Turn right to **enter the bedroom** on your right hand."	"Turn around and walk out of the kitchen. In front of you is a sofa chair, turn right passed the doorway. Now you should see the bathroom. Turn right and enter the bedroom on your right."
Case example from Matterport3D dataset	"Move forward and **turn left**. Walk straight and **turn left** to enter the room."	"Go ahead and **exit the room**. Turn left and **walk along the wall**. Enter the first room on your left."	"Go straight and pass the doorway to exit bedroom. Turn left after you reach the wall. And walk along the paintings on the wall until you reach the first door on your left. Turn left and enter the room."

Table 2 shows examples of different instruction styles for both the Clemson-VLN and Matterport3D datasets. The performance of the VLN agent was evaluated regarding these three types of instruction styles.

The testing results in Table 3 reveal that heuristic-based instruction performs better and is more robust in multiple datasets than other instruction types. It is due to the fact that VLN leverages the path searching problem by combining direction features and object features.

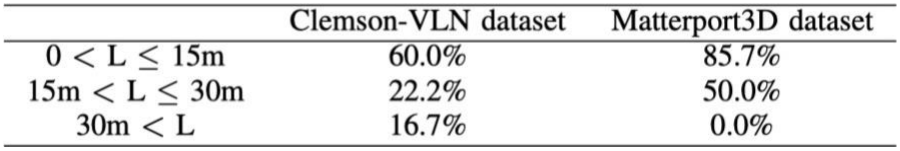
Table 3: Success Rate (SR) using different instruction style.



	Clemson-VLN Dataset	Matterport3D Dataset
Direction-based	40.00%	58.33%
Heuristic-based	40.00%	60.71%
Detail-based	34.48%	48.15%

In terms of *navigation length distance*, we also noticed that the VLN model performance can be affected by the navigation length (from the current position to the destination). In order to quantitatively reveal this influence, multiple sets of experiments with different navigation length ranges are conducted for both Clemson-VLN and Matterport3D datasets.

Table 4: Success Rate (SR) under different navigation length (unit: meter).

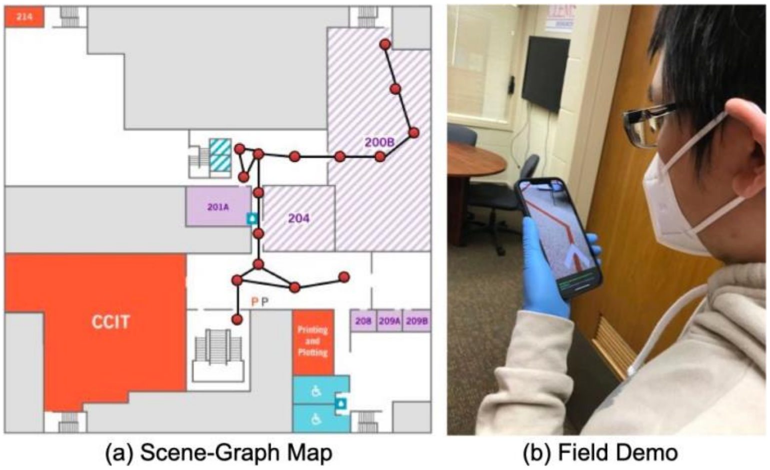


	Clemson-VLN dataset	Matterport3D dataset
$0 < L \leq 15m$	60.0%	85.7%
$15m < L \leq 30m$	22.2%	50.0%
$30m < L$	16.7%	0.0%

As shown in Table 4, the SR of the VLN model degrades as the navigation length increases. This inspired us to introduce an effective instruction re-inquiry process to remind BVI users to ask for possible new instructions in order to guarantee the highest SR.

4.5 Field Evaluation

Without BVI-subject field test approval, we conducted our field evaluation by a sighted subject, in the Rhodes Research Center at Clemson University, as illustrated in Figure 5(b). A sample graph map is illustrated in Figure 5(a). The red nodes are locations representing navigation viewpoints where the sky-box view color, depth, and pose are collected. By obtaining the language instruction through the inbuilt speech recognition system, SeeWay can infer the global navigation path online from the VLN model. The red polyline shows the virtual rendering of the path results in the current camera view with local path adjustment in case of obstacles in front.



(a) Scene-Graph Map (b) Field Demo

Figure 5: SeeWay field test: (a) Scene-graph map of viewpoints (red nodes) on Clemson University Cooper Library's first floor-map. (b) Field demonstration of the SeeWay system.

CHAPTER 5
Conclusions

To address the wayfinding guidance gap for BVI travelers between indoor environments and an automated vehicle, this study explored the solution for challenges to further the PIs' previous research. First, given the unavailability and inaccuracy of GPS localization near buildings (due to GPS signal reflections from walls), and the inconsistency of sun light illumination outdoors (or even in the night), an affordable highly-sensitive vision sensor or multi-modal sensor fusion solution in a portable device are prototyped as a sensory substitution for BVI travelers for more robust visual perception. Second, based on state-of-the-art artificial intelligence (AI), especially deep convolutional neural network (CNN) techniques, effectively processing the sensor or multi-modal sensor data for visual localization and environment recognition and understanding are employed to augment the travel safety of BVI individuals.

This study introduced a novel visual-language navigation (VLN) assistive wayfinding system for BVI people by taking human-spoken literal instruction as input. The VLN path generation system relies on visual features and literal instructions to search a global path from the user's current position to the literal destination. Comparative studies demonstrated the effectiveness of the VLN agent on global navigation and the local navigation strategy of obstacle avoidance. We further discussed the impact of three different literal instruction styles and navigation length on the success rate of navigation, which indicated that heuristic instruction combined with the re-inquiry strategy guaranteed the highest navigation success rate.

REFERENCES

Ahmed, M.U., Altarabichi, M.G., Begum, S., Ginsberg, F., Glaes, R., Östgren, M., Rahman, H., Sorensen, M., 2019. A vision-based indoor navigation system for individuals with visual impairment. Int. J. Artif. Intell. 17, 188–201.

Ali, A., Abou Ali, M., 2017. Blind navigation system for visually impaired using windowing-based mean on Microsoft Kinect camera, in: 2017 Fourth International Conference on Advances in Biomedical Engineering (ICABME). pp. 1–4.

Alkhatib, R., Swaidan, A., Marzouk, J., Sabbah, M., Berjaoui, S., O.Diab, M., 2019. Smart Autonomous Wheelchair, in: 2019 3rd International Conference on Bio-Engineering for Smart Technologies (BioSMART). pp. 1–5. https://doi.org/10.1109/BIOSMART.2019.8734264

Anderson, P., Chang, A., Chaplot, D.S., Dosovitskiy, A., Gupta, S., Koltun, V., Kosecka, J., Malik, J., Mottaghi, R., Savva, M., others, 2018. On evaluation of embodied navigation agents. arXiv Prepr. arXiv1807.06757.

ARKit, n.d. . Apple Dev. Doc.

Bahdanau, D., Cho, K., Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. arXiv Prepr. arXiv1409.0473.

Bai, J., Liu, Z., Lin, Y., Li, Y., Lian, S., Liu, D., 2019. Wearable travel aid for environment perception and navigation of visually impaired people. Electronics 8, 697.

Baiju, P.V., Varghese, K., Alapatt, J.M., Joju, S.J., Sagayam, K.M., 2020. Smart Wheelchair for Physically Challenged People, in: 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS). pp. 828–831. https://doi.org/10.1109/ICACCS48705.2020.9074188

Balata, J., Mikovec, Z., Slavik, P., 2018. Landmark-enhanced route itineraries for navigation of blind pedestrians in urban environment. J. Multimodal User Interfaces 12, 181–198.

Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L.D., Monfort, M., Muller, U., Zhang, J., others, 2016. End to end learning for self-driving cars. arXiv Prepr. arXiv1604.07316.

Brooks, J., Li, B., Jenkins, C., Dylgjeri, L., Sarathkrishna, Ajayan, A., Ghodekar, M., Nikhal, S., Rana, A., Yang, Z., 2022. Transportation Preferences, Challenges & Opportunities for Visually Impaired Users. J. Vis. Impair. Blind. Under Rev.

Brooks, J.O., Mims, L., Jenkins, C., Lucaciu, D., Denman, P., 2018. A User-Centered Design Exploration of Fully Autonomous Vehicles's Passenger Compartments for At-Risk Populations, in: SAE Technical Paper.

Bumuller, A., Skerl, K., 2018. Development of a modular smart wheelchair, in: 2018 International IEEE Conference and Workshop in Îbuda on Electrical and Power Engineering (CANDO-EPE). pp. 49–54. https://doi.org/10.1109/CANDO-EPE.2018.8601155

Caraiman, S., Morar, A., Owczarek, M., Burlacu, A., Rzeszotarski, D., Botezatu, N., Herghelegiu, P., Moldoveanu, F., Strumillo, P., Moldoveanu, A., 2017. Computer vision for the visually impaired: the sound of vision system, in: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 1480–1489.

Cheng, R., Wang, K., Lin, S., 2018. Intersection navigation for people with visual impairment, in: International Conference on Computers Helping People with Special Needs. pp. 78–85.

Fried, D., Hu, R., Cirik, V., Rohrbach, A., Andreas, J., Morency, L.-P., Berg-Kirkpatrick, T., Saenko, K., Klein, D., Darrell, T., 2018. Speaker-follower models for vision-and-language navigation. arXiv Prepr. arXiv1806.02724.

Gomes, J.P., Sousa, J.P., Cunha, C.R., Morais, E.P., 2018. An indoor navigation architecture using variable data sources for blind and visually impaired persons, in: 2018 13th Iberian Conference on Information Systems and Technologies (CISTI). pp. 1–5.

Graves, A., Schmidhuber, J., 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural networks 18, 602–610.

Hart, P., Nilsson, N., Raphael, B., 1968. A Formal Basis for the Heuristic Determination of Minimum Cost Paths. IEEE Trans. Syst. Sci. Cybern. 4, 100–107. https://doi.org/10.1109/tssc.1968.300136

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9, 1735–1780.

Islam, M.I., Raj, M.M.H., Nath, S., Rahman, M.F., Hossen, S., Imam, M.H., 2018. An indoor navigation system for visually impaired people using a path finding algorithm and a wearable cap, in: 2018 3rd International Conference for Convergence in Technology (I2CT). pp. 1–6.

Jayakody, A., Nawarathna, A., Wijesinghe, I., Liyanage, S., Dissanayake, J., 2019. Smart Wheelchair to Facilitate Disabled Individuals, in: 2019 International Conference on Advancements in Computing (ICAC). pp. 249–254. https://doi.org/10.1109/ICAC49085.2019.9103409

Kaul, O.B., Rohs, M., Mogalle, M., Simon, B., 2021. Around-the-head tactile system for supporting micro navigation of people with visual impairments. ACM Trans. Comput. Interact. 28, 1–35.

Ku, A., Anderson, P., Patel, R., Ie, E., Baldridge, J., 2020. Room-Across-Room: Multilingual Vision-and-Language Navigation with Dense Spatiotemporal Grounding. arXiv Prepr. arXiv2010.07954.

Kuriakose, B., Shrestha, R., Eika Sandnes, F., 2021. Towards Independent Navigation with Visual Impairment: A Prototype of a Deep Learning and Smartphone-based Assistant, in: The 14th PErvasive Technologies Related to Assistive Environments Conference. pp. 113–114.

Kutbi, M., Du, X., Chang, Y., Sun, B., Agadakos, N., Li, H., Hua, G., Mordohai, P., 2020. Usability Studies of an Egocentric Vision-Based Robotic Wheelchair. J. Hum.-Robot Interact. 10.

Leaderboard - EvalAI, n.d.

Leaman, J., La., H.M., 2015. iChair: Intelligent Powerchair for Severely Disabled People., in: ISSAT International Conference on Modeling of Complex Systems and Environments (MCSE). Da Nang, Vietnam.

Leaman, J., La, H.M., 2017. A Comprehensive Review of Smart Wheelchairs: Past, Present, and Future. IEEE Trans. Human-Machine Syst. 47, 486–499.

Li, B., Muñoz, J.P., Rong, X., Chen, Q., Xiao, J., Tian, Y., Arditi, A., Yousuf, M., 2018. Vision-Based Mobile Indoor Assistive Navigation Aid for Blind People. IEEE Trans. Mob. Comput. 18, 702–714. https://doi.org/10.1109/tmc.2018.2842751

Li, B., Muñoz, J.P., Rong, X., Xiao, J., Tian, Y., Arditi, A., 2016. ISANA: Wearable Context-Aware Indoor Assistive Navigation with Obstacle Avoidance for the Blind, in: Hua, G., Jégou, H. (Eds.), European Conference on Computer Vision (ECCV) Workshop. Springer International Publishing, Cham, pp. 448–462.

Mnih, V., Badia, A.P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., Kavukcuoglu, K., 2016. Asynchronous methods for deep reinforcement learning, in: International Conference on Machine Learning. pp. 1928–1937.

Mohammad Monirujjaman Khan Shamsun Nahar Safa, M.H.A.M.M.M.A.A., 2021. Research and Development of a Brain-Controlled Wheelchair for Paralyzed Patients. Intell. Autom. Soft Comput. 30, 49–64.

Mourikis, A.I., Roumeliotis, S.I., 2007. A multi-state constraint Kalman filter for vision-aided inertial navigation, in: Proceedings 2007 IEEE International Conference on Robotics and Automation. pp. 3565–3572.

Muñoz, J.P., Li, B., Rong, X., Xiao, J., Tian, Y., Arditi, A., 2017. An Assistive Indoor Navigation System for the Visually Impaired in Multi-Floor Environments, in: IEEE International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER). IEEE, pp. 7–12. https://doi.org/10.1109/CYBER.2017.8446088

Oh, Y., Kao, W.-L., Min, B.-C., 2017. Indoor navigation aid system using no positioning technique for visually impaired people, in: International Conference on Human-Computer Interaction. pp. 390–397.

Ortiz, J.S., Palacios-Navarro, G., Andaluz, V.H., Guevara, B.S., 2021. Virtual Reality-Based Framework to Simulate Control Algorithms for Robotic Assistance and Rehabilitation Tasks through a Standing Wheelchair. Sensors 21.

Pana, C.F., P?tra?cu-Pan?, D.M., Vladu, I.C., Manta, L.F., Besnea Petcu, F.-L., Cismaru, ?tefan Irinel, Tr??culescu, A.C., 2021. Fuzzy Control of the Robotic Arm for a Smart Electric Wheelchair to Assist People with Movement Disabilities, in: 2021 22nd International Carpathian Control Conference (ICCC). pp. 1–6.

Poole, D.L., Mackworth, A.K., 2010. Artificial Intelligence: foundations of computational agents. Cambridge University Press.

Recognizing Speech in Live Audio, n.d. . SFSpeechRecog.

Ryu, H.-Y., Kwon, J.-S., Lim, J.-H., Kim, A.-H., Baek, S.-J., Kim, J.-W., 2022. Development of an Autonomous Driving Smart Wheelchair for the Physically Weak. Appl. Sci. 12.

Sennrich, R., Haddow, B., Birch, A., 2015. Improving neural machine translation models with monolingual data. arXiv Prepr. arXiv1511.06709.

Shahnaz, C., Maksud, A., Fattah, S.A., Chowdhury, S.S., 2017. Low-cost smart electric wheelchair with destination mapping and intelligent control features, in: 2017 IEEE International Symposium on Technology and Society (ISTAS). pp. 1–6. https://doi.org/10.1109/ISTAS.2017.8318978

Speech Synthesis, n.d. . AVSpeechSynthesizer.

Tan, H., Yu, L., Bansal, M., 2019. Learning to navigate unseen environments: Back translation with environmental dropout. arXiv Prepr. arXiv1904.04195.

Velázquez, R., Pissaloux, E., Rodrigo, P., Carrasco, M., Giannoccaro, N.I., Lay-Ekuakille, A., 2018. An outdoor navigation system for blind pedestrians using GPS and tactile-foot feedback. Appl. Sci. 8, 578.

WHO, 2021. Blindness and vision impairment.

WHO, 2011. World Report on Disability. World Heal. Organ.

Woo, J., Yamaguchi, K., and Yasuhiro Ohyama, 2021. Development of a Control System and Interface Design Based on an Electric Wheelchair. J. Adv. Comput. Intell. Intell. Informatics 25, 655–663.

Yang, Z., Yang, L., Kong, L.K., Wei, A., Brooks, J.B., Li, B., 2022. SeeWay: Vision-Language Assistive Navigation for the Visually Impaired, in: IEEE International Conference on Systems, Man, and Cybernetics.

Zhang, X., Yao, X., Zhu, Y., Hu, F., 2019. An ARCore based user centric assistive navigation system for visually impaired people. Appl. Sci. 9, 989.