ShelfHelp: Empowering Humans to Perform Vision-Independent Manipulation Tasks with a Socially Assistive Robotic Cane

Shivendra Agrawal
University of Colorado Boulder
shivendra.agrawal@colorado.edu

Suresh Nayak
University of Colorado Boulder
suresh.nayak@colorado.edu

Ashutosh Naik
University of Colorado Boulder
ashutosh.naik@colorado.edu

Bradley Hayes
University of Colorado Boulder
bradley.hayes@colorado.edu

ABSTRACT

The ability to shop independently, especially in grocery stores, is important for maintaining a high quality of life. This can be particularly challenging for people with visual impairments (PVI). Stores carry thousands of products, with approximately 30,000 new products introduced each year in the US market alone, presenting a challenge even for modern computer vision solutions. Through this work, we present a proof-of-concept socially assistive robotic system we call ShelfHelp, and propose novel technical solutions for enhancing instrumented canes traditionally meant for navigation tasks with additional capability within the domain of shopping. ShelfHelp includes a novel visual product locator algorithm designed for use in grocery stores and a novel planner that autonomously issues verbal manipulation guidance commands to guide the user during product retrieval. Through a human subjects study, we show the system's success in locating and providing effective manipulation guidance to retrieve desired products with novice users. We compare two autonomous verbal guidance modes achieving comparable performance to a human assistance baseline and present encouraging findings that validate our system's efficiency and effectiveness and through positive subjective metrics including competence, intelligence, and ease of use.

KEYWORDS

Assistive Robotics; Computer Vision; Planner; Manipulation Guidance; Human-Robot Interaction; Markov Decision Process

ACM Reference Format:
Shivendra Agrawal, Suresh Nayak, Ashutosh Naik, and Bradley Hayes. 2023. ShelfHelp: Empowering Humans to Perform Vision-Independent Manipulation Tasks with a Socially Assistive Robotic Cane. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 10 pages.

1 INTRODUCTION

It is estimated that 295 million people have some form of vision impairment, of whom 43.3 million are blind [8]. Currently, when encountering difficulty with shopping, they can get help from sighted humans. This lack of independence has other negative connotations such as loss of privacy [3]. Some PVI shoppers have also indicated they are not willing to use store staffers for shopping for items that require discretion such as medicine and personal hygiene items [3, 20, 22]. Our work seeks to alleviate the dependence on guide availability and to mitigate the loss of privacy encountered with traditional support mechanisms. People have also expressed the need to have a system that can help them locate items in a store and even at their home [3]. Moreover, some environments present additional challenges while solely relying on tactile sensing. For example - 1) Kitchen counters may contain hot or sharp objects and may pose a safety issue. 2) Grocery stores with similar items shelved together or with an otherwise dense concentration of items contributes to poor tactile differentiability.

Figure 1: ShelfHelp includes a robotic cane equipped with RealSense D455 and T265 cameras. The system is powered through a laptop in a backpack. *Left:* The system used as a navigational device. It uses audio and haptic feedback for navigation guidance. *Right:* The system used as a manipulation device. It uses audio for manipulation guidance.

Grocery shopping primarily consists of three main subtasks: navigation, product retrieval, and product examination. This work focuses on product retrieval. Prior research on spatial cognition of PVI uses a dichotomous ontology where they distinguish the space around the person in two categories: *1) locomotor* and *2) haptic* [17]. Locomotor space is the space around the user whose exploration and access requires locomotion whereas haptic space is the immediate space around the user that can be sensed without bodily translation. Our work addresses the research problem in the haptic space of the user.

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

We propose two fine-grain manipulation guidance systems that use verbal instructions to guide the user towards retrieving the desired item. We use the grocery store domain because of the representative challenges it contains and the opportunity for immediate positive broader impact. We also introduce a novel product detection algorithm to locate desired items in a grocery setting with a hand-held camera system. In this work, we showcase the manipulation guidance system as part of a grocery shopping assistant. We define fine-grain manipulation guidance with respect to object retrieval as guidance that brings the user close to the desired object such that the object is in the haptic space of the user and is accessible to the user's grasp. Our solution aims to minimize the exhaustive search of the object in the haptic space. This is especially inefficient and a burden for PVI in a grocery store because of the high density of products that could be present even in the limited haptic space of the person.

The hardware choices for ShelfHelp are motivated by utility, cost, and extensibility. Our system extends the capability of a robotic smart cane [5] created originally for social navigation assistance (Fig.1). This area has been extensively researched [9, 11, 18, 19, 24–27, 30, 33–36, 38, 40–43, 45–48, 50], yet many important capabilities are still rich for exploration. It is practical and prudent to utilize the sensing and compute power of these existing devices to address multitude of critical tasks for a more independent lifestyle.

To summarize, this paper contributes,

- A robotic cane system that leverages computer vision to assist in independent grocery shopping for PVI.
- A modular two stage computer vision pipeline to locate desired products in a grocery setting.
- A novel fine-grain manipulation guidance system that optimizes for guide time and the number of commands without compromising legibility.
- A pilot study validating the system's success in locating and providing effective guidance to retrieve a desired object from a dense cluster of objects with novice users.
- Findings on user preferences regarding desirable planner properties.

2 RELATED WORK

2.1 Manipulation guidance

Manipulation guidance is an area that has been explored within the robotics community for over a decade. Vasquez et al.[44] showed that saliency maps could be used to find regions of interest (ROI) and directed users' hand to the ROI. They found that their verbal commands' efficacy suffered because they did not utilize a global frame of reference. Bonani et al. [7] showed promise for the concept with an experimenter-controlled teleoperated system and Bigham et al. [6] did so with a Mechanical Turk-based system, but fully autonomous implementations were outside the scope of their contributions. Their manipulation guidance guided people to the general direction of the product but the system didn't focus on fine-grain manipulation guidance which is important in many scenarios, for example a kitchen countertop or where tactile differentiability makes exhaustive search inefficient such as grocery stores which have similar items densely situated. The most popular solution in this problem domain is a human-powered service called Be My Eyes [1], but this service suffers from scalability issues due to its reliance on human availability, is not readily available in developing countries, requires an active data connection, and introduces nigh-unavoidable privacy concerns. We present a novel verbal guidance solution wherein we learn a mapping of language commands to human hand movements and map them to actions within a Markov Decision Process (MDP) that can be solved with well-established reinforcement learning techniques, informing our guidance of the user.

2.2 Product identification

Existing techniques with a fixed number of output classes work with a limited database of products in a grocery store, for example, Feng et al. [13] trained a system for classifying 1329 products. These techniques may be impractical for the scale required of a solution for this domain because of the sheer amount of products [29] available and the slight variations they come in. It is infeasible to require creation and maintenance of labeled data and the training of an object classifier. Existing technology that is reliable in product identification uses bar code scanning [2, 15]. This often needs internet connectivity and can only identify a product once it is in the person's possession, making it inefficient for search over the grocery shelves. Our work is inspired by the body of work in *instance retrieval*, where the task is to find a target image in the scene [10]. Our proposed product identification system does not require re-training and works offline, making it scalable and practical.

2.3 Grocery assistant systems

Recent work on grocery assistant systems focuses largely on navigation inside the store [15, 20–22, 28], also known as the locomotor space of the user. Solutions that rely on environmental augmentation, such as the addition of RFID tags and barcodes, introduce barriers to adoption as they are inapplicable in uninstrumented domains. Barcodes alone can not reliably be used to locate a product from a dense cluster of products on a grocery shelf, as the shopper can typically only scan a barcode once they have retrieved the product. Prior research [15] has also focused on text-based product selection as an input method. Researchers have also focused on other issues around shopping, including identifying products that users are running out of and organizing newly purchased products at home [49], as well as "the last few meters" way-finding problem [35] and solutions for people with low vision [51]. We focus on an unsolved research area that primarily considers the haptic space of the user.

To summarize, our solution addresses 1) the problem of locating a desired product and 2) the challenge of providing effective fine-grain verbal guidance to reach and grasp the product.

3 SYSTEM DESIGN

ShelfHelp extends the capabilities of an existing robotic cane originally developed as a navigational assistance device. This was a design choice to re-purpose an existing hardware system since many of the navigational assistance prototypes have the necessary sensing and compute (Figure 1). The system is capable of locating desired products and verbally guiding the user to retrieve them.

The software system follows a serial architecture composed of three main components - *perception, planning,* and *guidance.*

3.1 Hardware System

The hardware (Fig 1) consists of a cane-mounted RealSense D455 for RGB-Depth sensing and RealSense T265 for odometry. The sensors are connected to a Dell G15 laptop with an RTX 3060 GPU carried in a backpack worn by the user. As a navigational assistance device, the system is held with a standard cane grip but as a manipulation assistance device, a different, less collision prone grasp is used behind the sensors.

3.2 Software System

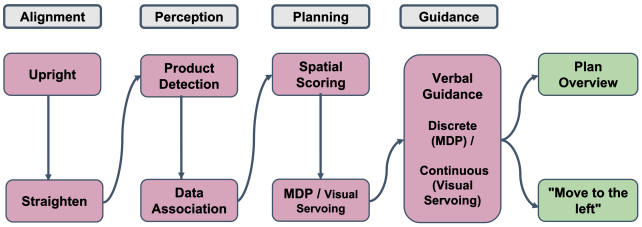

Figure 2: System Diagram. Alignment, perception, planning, and verbal conveyance are executed on a backpack-worn laptop, while all the sensing is mounted on the cane.

Our work can best be partitioned into sections regarding innovations in perception, solutions to the data association problem for maintaining a consistent and persistent mapping of product detections over time, product selection, fine-grain manipulation planning to reach the selected product, and methods for conveying this manipulation plan to the human user to complete the task.

3.2.1 Alignment. In order for the system to issue commands in the user's egocentric frame of reference, our system issues verbal commands to align the user and the system with respect to the shelf. We use the T265 camera's IMU to guide the user to point the system straight at the shelf in an upright way. We also align the user so that they directly oppose the shelf normal. This is done by locating the shelf after camera upright alignment via Hough transform. The system issues commands to turn in place until the largest horizontal line in the Hough transform is almost completely horizontal in the camera frame.

3.2.2 Product Detection. The entire product detection pipeline works in realtime making it appropriate for real world usage. To use this system, we require that the user has only a single image of the product that they want to find. This image can be acquired in a number of ways, for example by taking a picture the first time it is purchased or by downloading an image from the internet. We have developed a novel two-stage product search system (Algorithm 1).

- In the first stage, our method proposes regions in form of bounding boxes that are most likely to contain *any* product (line 1 of Algorithm 1). We train the YoloV5 network on the SKU-110K dataset [16] to create a product detector. This network is robust and generates region proposals with 0.91-precision and 0.77-recall upon cross-validation with the SKU-110K dataset. Figure 3 (upper-right) shows a sample result from a real grocery store.
- In the second stage, an encoder is used as a feature extractor (line 2 and 5) that matches the features of the proposed

Algorithm 1: Product Detection Procedure

Input: scene, target_product, camera_pose
Output: best_product
Parameters: `similarity_threshold`

1 $boundingbox\_list \leftarrow region\_proposal(scene)$
2 $target\_feature \leftarrow feature\_extractor(target\_product)$
3 $similar\_instances \leftarrow []$
4 **foreach** $bb \in boundingbox\_list$ **do**
5 $\quad bb\_feature \leftarrow feature\_extractor(bb)$
6 $\quad bb\_score \leftarrow similarity(target\_feature, bb\_feature)$
7 $\quad$ /* Scene has RGB and depth information
8 $\quad bb\_pose \leftarrow camera\_to\_world(bb, scene, camera\_pose)$
9 $\quad$ **if** $bb\_score \geq$ `similarity_threshold` **then**
10 $\quad\quad similar\_instances.append(bb\_pose, bb\_score)$
11 $similar\_products \leftarrow data\_association(similar\_instances)$
12 $best\_product \leftarrow spatial\_scoring(similar\_products)$
13 **return** $best\_product$

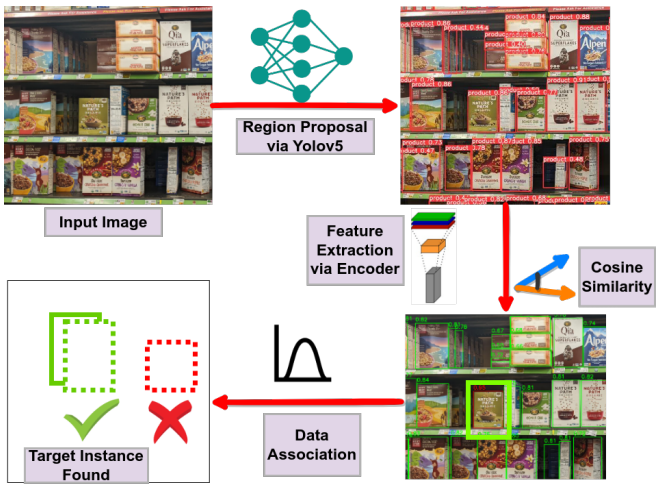

Figure 3: Our product search algorithm can reliably locate desired products on a grocery shelf. Regions with a high likelihood of containing any product are proposed in the first stage. The features of these regions are then compared against the target product image. Our data association solution is used to identify whether detections from incoming camera frames are new or re-detections of existing products. The above image shows our algorithm operating within an actual grocery store, where the product classification aspect of this work has been tested and validated. The data association and manipulation assistance components were validated within a lab-based study.

regions and the target image, finding the best possible match (Fig 3). We do this by training an autoencoder on MS-COCO color images [23] and then utilizing the encoder portion as the feature extractor (Fig 4). This method doesn't require any retraining and works in real-time. For each new image added to the database, we pass it through the frozen encoder and save the generated feature vector on disk. This avoids requiring any retraining of the autoencoder. The encoder finds the latent space representation of the target product image and each of the proposed regions. We compare the representation vectors using *cosine similarity* to find closer vectors (line 6). We empirically determine a similarity score threshold (0.6) that captures satisfactory performance across real-world environments, but this value can easily be fine-tuned in case there is a significant distribution shift between the evaluation and deployment environments.

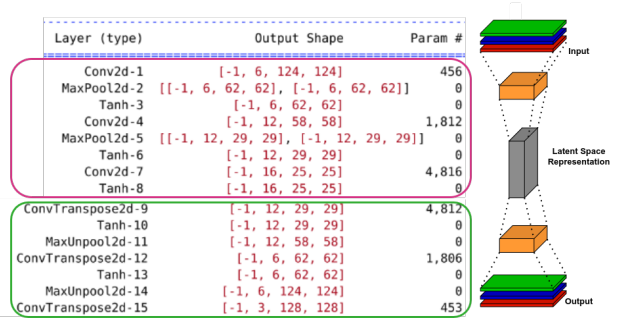

Figure 4: The architecture of the *autoencoder* trained for feature extraction. ShelfHelp uses the *encoder* portion (top layers outlined in red) as the feature extractor that creates a latent space representation of the images.

To transform the information from the camera frame to a fixed global frame, we use pose information obtained by a cane-mounted RealSense T265 (which has minimal drift in indoor settings) running an onboard Simultaneous Localization and Mapping (SLAM) algorithm, and fuse it with the depth information obtained from a cane-mounted RealSense D455 (line 8). We use a Gaussian Mixture Model (GMM) to refine detections and distinguish between the foreground and background depth information, as the bounding boxes can contain significant background pixels in case a product and its bounding box are not overlapping significantly.

3.2.3 Data Association. We use data association techniques to identify each instance of the same product uniquely across subsequent frames of camera capture (line 11). This is particularly challenging because similar products exist in groups. We do this by defining each product instance as a multivariate Gaussian defined by the tuple $p = \{x_g, y_g, z_g, w, h\}$ where x_g, y_g, z_g is the 3D location of the product in a global frame and w, h are width and height in meters. Accounting for the width and the height helps us discard some incorrect matches, as incorrectly proposed regions with our method not only have to have similar features but also similar shape to be incorrectly labeled (Fig. 3 - lower left). The IMU data from the T265 sensor helps to calculate the object's pose in a global frame of reference and helps to create a persistent "map" of the product location. This way we can align the verbal directional commands with respect to the current hand pose and avoid the drawback of formulating verbal commands generated with targets located only in the camera frame of view, which is sensitive to hand movements [44]. Product instance information is updated using a rolling mean over associated detections. We also employ a *lazy deletion strategy* to delete instances that have not been seen a sufficient number of times (sparse detections) or recently (old detections).

3.2.4 Spatial Scoring. ShelfHelp then scores each detected product instance of the target product and picks the one with the highest score as its planning goal (line 12). It considers the rolling similarity with the target image and the spatial information. This allows the system to utilize important information that is absent without an explicit physics model, namely selecting an instance from the top level of stacked items to minimize the risk of toppling (Fig 5).

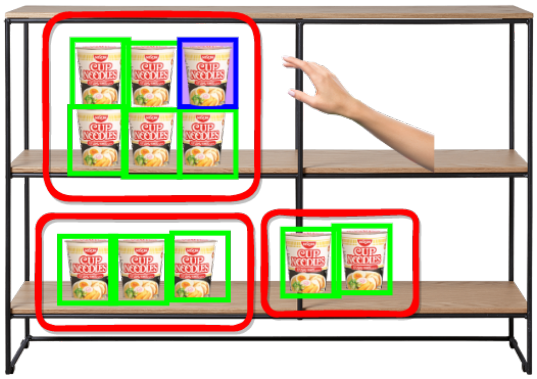

Figure 5: The spatial scoring system clusters all the found instances spatially and gives preference to the closest cluster to the current hand pose. Ties are broken arbitrarily.

3.2.5 Planning. We have developed two different guidance mechanisms to provide verbal instruction once the target has been located: *1) continuous*, for example, *"keep on going right"..."stop"* and *2) discrete*, for example, *"move 6 inches to the right"*. The decision to create a discrete guidance mode was inspired by study results showing that PVI have been known to perceive length units better than sighted people [4] and the criteria of minimizing verbal feedback for the task as research has shown that PVI prefer not being provided with excessive information [39].

3.2.6 Continuous Planner. The continuous guidance operates by calculating the relative position of the target and the device (Fig. 1), providing continuous cues along each individual axis of movement until the next is aligned. It generates the commands with the following template: *"keep on going {direction}"* where direction can be *{left, right, up, down, forward, backward}*. The system issues the command *"keep on going"* if it notices the hand slow down sufficiently in anticipation of the next command. Once the error on the current guidance axis is lower than a threshold the system issues the command *"stop"*.

3.2.7 Dataset from Human Demonstrations: To develop the discrete guidance method, we collected a dataset mapping verbal movement commands from a fixed command-set, recording participants' net hand movements upon reacting to that command (Fig 6). We formed 36 discrete commands and issued 1250 instances of the commands in total to 25 volunteers (50 commands per person) while they were blindfolded. The hand movement data were recorded using an OptiTrack motion capture system. Figure 6 shows a sample from this dataset illustrating commands pertaining to *left* movement. We can see that the hand movement caused by these commands is close to a Gaussian distribution and the *mean* increases for the verbal commands that convey a larger movement. We fitted Gaussians to characterize the movement caused by each command as

$$X_c \sim \mathcal{N}(\mu_c, \sigma_c^2)$$

where μ_c and σ_c are the mean and standard deviation of the movement caused by command c.

3.2.8 Discrete Planner. Through the human-subjects demonstrations, we find that the hand movement following the commands was not deterministic. Moreover, a greedy approach would not guarantee an optimal solution, akin to the 0-1 Knapsack problem. We solve this sequential decision-making problem by modeling it as an MDP formulated with (S, A, T, R) where S is the set of states in the MDP, A is the set of actions, T is a stochastic transition function describing the action-based state transition dynamics of the model, and R is a reward function.

- S is the tuple $(\Delta x, \Delta y, \Delta z, axis)$ where the first three terms are the difference in distance of the target and the hand pose, and *axis* defines the axis of motion for the previous command which could be any of three values corresponding to the X (horizontal), Y (vertical), or Z (depth) axis and a direction *{left, right, up, down, forward, backward}*. This formulation is dependent on the relative distance between the target and the hand and thus it finds a plan for all the potential states that can be encountered. We discretized the states at 5 cm resolution and considered a cuboid region of 0.8m as the operational space for the human hand. The system assumes the hand to be at the boundary of our cuboid in case it was outside of the region. In practice, a region bigger than this produced the same policy for regions that were further away.
- A is the set of discrete verbal commands such as *"Move 6 inches to the left"*.
- T is calculated from X_c as the movement caused by each command is not deterministic. We use the Euclidean difference between the states and Gaussian X_c associated with the command c to find out the probability of the transition between those states when command c is issued.
- The reward function R encourages reaching the target and discourages issuing superfluous commands. It also discourages a sequence of commands that could be illegible or frustrating by penalizing axis changes.
  Table 1 shows the reward values and their description. The *axis_order_reward* rewards adherence to a prescribed ordering of guidance with respect to the axes. For example, we set the initial guidance axis as *vertical* and the planner rewards transitioning from the *vertical* axis to the *horizontal* axis and not the *depth* axis. This helps to hone in on the product on the XY plane and then reach out for the product in the *depth* axis. The *living_penalty* penalizes excessive number of commands. The *interleaving_penalty* penalizes excessive interleaving of axes that could make the guidance illegible. It is necessary and suitable to transition once the distance to the target on the current axis of guidance is reduced and short enough. We ensure this by setting *necessary_transition_reward* inversely proportional to the distance remaining (error) on the current guidance axis.

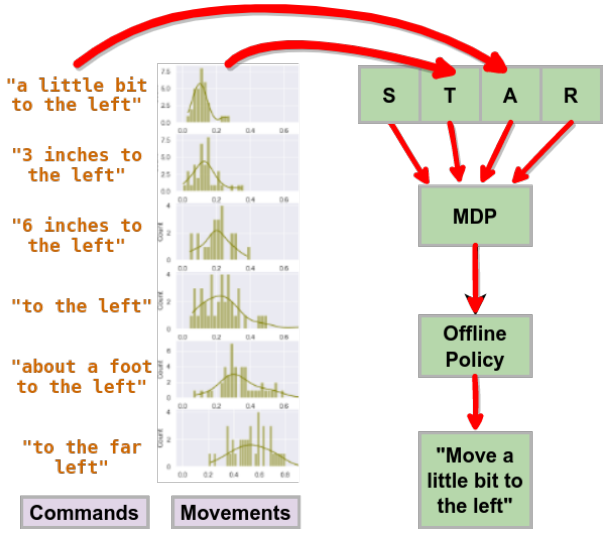

Figure 6: *(Left to right)* A sample of discrete commands. The movement (in meters) each command caused. MDP and solution definition. We train a model of human hand movement from demonstrations that inform the transition probabilities T. S defines the state space, A defines the discrete set of verbal actions, and R is the reward function. A policy is learned offline that can be used across reaching tasks.

Table 1: Outline of MDP Reward values

Reward	Description
$goal\_state\_reward = +10000$	reward for reaching the goal
$living\_penalty = -10$	penalty for using a command
$interleave\_penalty = -100$	penalty for changing axis of command
$axis\_order\_reward = +100$	reward for transitioning from vertical to horizontal axis
$necessary\_transition\_reward = (0.001 + \text{current axis error})^{-1}$	reward for axis transition when current axis error is removed

The MDP is then solved using value iteration to generate a general reaching policy that can be queried online to guide the user toward arbitrary target locations.

3.2.9 Guidance. The verbal guidance has two components. It begins with a plan overview that is followed by hand movement instructions.

1) Plan Overview: The actual guidance is preceded by a plan overview which serves to set the expectations for the general direction of the forthcoming instructions. We do so by computing the initial heading direction to the target. We project and discretize the heading direction to the target onto the plane of the shelf and use clock-based direction. The overview has the following template: *"I found the product at about { } o'clock direction".*

2) Hand Movement Instructions: The plan overview is followed by guidance instructions. The instructions to convey to the user (actions) are computed online when using the continuous guidance mode and queried online from the policy learned in the discrete guidance mode. Based on the relative position of the target and the hand, a command is formulated (or retrieved from the policy) and issued aloud from a speaker that is part of the robotic cane system. In an effort to reduce frustration, the system issues new commands only when the user's hand has slowed down sufficiently to show that they are ready for the next command. The user is asked to grasp the target object with their non-occupied hand if they are close to it with the system.

4 PILOT STUDY

We conducted an IRB-approved study with novice users (n=15) for system testing and validation. Participants were all sighted students who were blindfolded for the testing. At this stage of design and technical readiness, we follow prior work in performing preliminary validations using blindfolded users [12, 14, 31, 32, 40, 45] as a precursor to engaging with the PVI community.

4.1 Hypotheses

We test the following hypotheses in our comparison of the two planner modes.

- **H1**: Participants will retrieve products with a lesser number of commands in the discrete guidance mode compared to the continuous guidance mode.
- **H2**: Participants will retrieve the products in less time with the discrete guidance mode as compared to the continuous guidance mode, once the products are located.
- **H3**: Participants will retrieve the products with less net hand movement with the discrete guidance mode compared to the continuous guidance mode.

4.2 Experiment Design

We use the experimental setup shown in Figure 7 to test the system. While we test the system as a whole, we do so with a specific focus on the two proposed guidance planners - 1) Continuous and 2) Discrete. As a baseline, we also compare our system against a human caller (university research staff separate from the experimenters) guiding over a video call. The caller gave freeform verbal guidance with the aim of helping the participant reach the goal on the shelf. The users performed each of these 3 conditions (continuous, discrete, and human) 5 times. We had a set of 5 different products for each of the conditions and the same set of 5 products with the same spatial configuration was used for each condition.

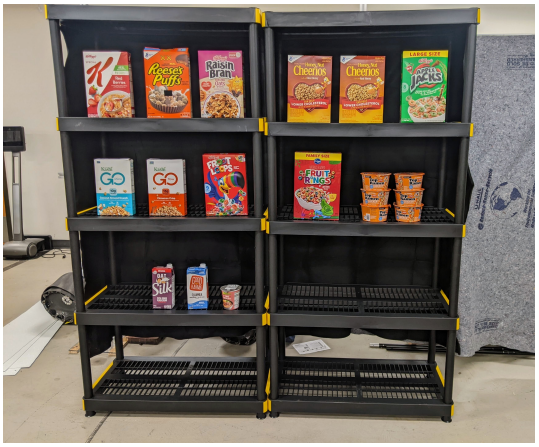

Figure 7: The experimental setup approximating a grocery store shelf, used for evaluating the efficacy of our manipulation guidance system.

The products were picked from the top two rows. We randomly shuffled the 3 conditions, counterbalancing as needed to achieve parity of orderings to avoid training/familiarity biases. We ensured that the users have the shelf in the frame at the starting pose and they were approximately 0.9m to 1.5 away from the target product in all the runs of the experiment. Each user was oriented on how to use the system, how to hold it, and how many runs they would perform. We asked them to keep their locomotion to a minimum and primarily move their hand. We also informed them about the alignment process and how to interpret the alignment guidance instructions that precede the actual guidance. The orientation process took about 3-4 minutes on average for each user.

4.3 Results

4.3.1 System Success Rate. Each participant retrieved 5 different products in each condition *(continuous, discrete, human)*. Perhaps unsurprisingly, the human condition was successful **75/75** times. We observed a few scenarios where participants picked up the wrong product initially but were corrected by the human caller. The product locator system was the same for both the *continuous* and the *discrete* modes. The product detection failed **21/150** times in locating the desired product in the participant's first attempted scan of the shelf. This mostly happened when the desired product was not in the camera frame. If similar enough products are in frame (above threshold), the system selects the most visually similar product it can find. In some cases, when the desired target product is not facing straight or is occluded, the system might locate a visually similar product that may be incorrect (Figure 9). While inconvenient, these errors are ultimately correctable using existing work (e.g., barcode based methods). The system showed promise in dealing with overexposure and blur due to the data association as shown in the example (Figure 8).

Both of our planners guided the user to the desired product **150/150** (100%) times. In the continuous mode, the participant picked up the adjacent item **8/75** times, and in the discrete mode that happened **6/75** times. It is important to note that the guidance algorithm took the participant close to the product, but since the

Figure 8: A sample product detection result that was robust to blurring and overexposure with the help of *data association*.

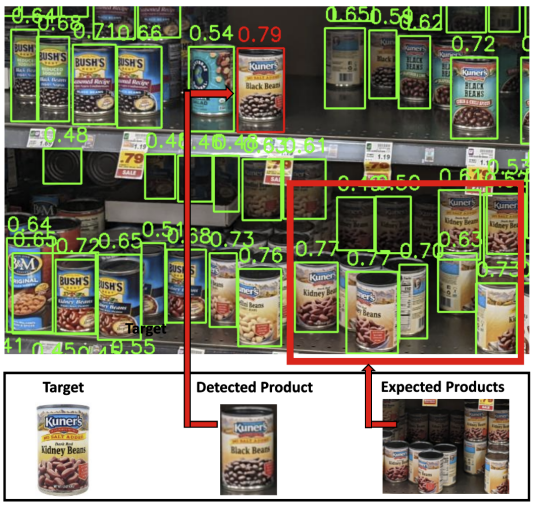

Figure 9: The product detection algorithm in a real grocery store. The system detects an incorrect product which is visually similar to the target. The expected products were either not facing straight or occluded partially by price tags hampering their visual similarity with the target.

users used their non-dominant hand to pick up the product they picked up a product immediately adjacent to the desired one (which was usually just 1-2 inches away). We include these data in our subsequent analysis, as the high level guidance was still successful in minimizing the exhaustive search required in the haptic space of the participant.

4.3.2 Hypotheses Testing. Post-hoc comparisons using Tukey's HSD test revealed that participants retrieved the products with significantly fewer commands in the discrete mode as compared to the continuous mode, **confirming H1** (Figure 10). In fact, the number of commands was similar for the human caller and the discrete mode. Using the same test we found that the participants could retrieve the products significantly faster in the discrete mode compared to the continuous mode, thus **confirming H2** (Figure 11). The guidance time for the discrete was also similar to the human caller. We ran a TOST (two one-sided test) which *confirms distributional equivalence for the number of commands and guidance time between the discrete planner and human caller.* Although the same doesn't hold true if we consider the total time which includes: alignment, searching for the product, reporting the plan, and guiding. We can see (Figure 13) that both proposed modes incur some time in aligning and reporting the plan. We did not classify all of these activities for the human caller because they would interleave these instructions into their guidance, so we classify the time prior to providing actual movement instructions as 'search time'. We did not find any statistical difference between the net hand movement caused by either planner, thus we **could not confirm H3**. To our surprise, we see that the human caller had significantly higher net movement (Figure 12). We think that this is due to the fact that the human caller relied on tracking the dominant hand of the participant in order to give them instructions with respect to the dominant hand position. The dominant hand would often go out of the cell phone camera frame and the human caller would have to issue instructions in order to gauge where the hand is causing these superfluous movements.

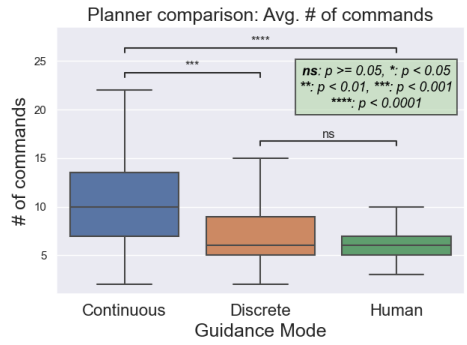

Figure 10: Average number of commands used by the different planners.

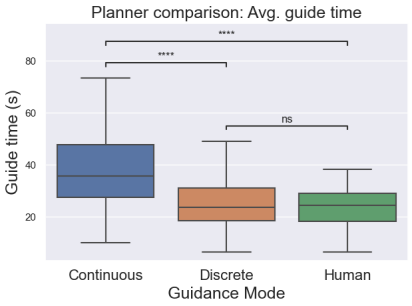

Figure 11: Average guide time used by the different planners.

4.3.3 Subjective Evaluation. We administered a survey after each condition. Participants rated both of our planners high on metrics concerning: human-like, interactive, competent, and intelligent (Figure 14). The two planners and the human performance were not statistically different on the competence and intelligence metrics. Both of the planners were rated less humanlike than our human baseline. The discrete planner was rated slightly lower in the interactive metric which could be because the discrete planner did not provide any affirmations, which is an area of future investigation. The participants rated the plan overview's helpfulness as 4.2 (std

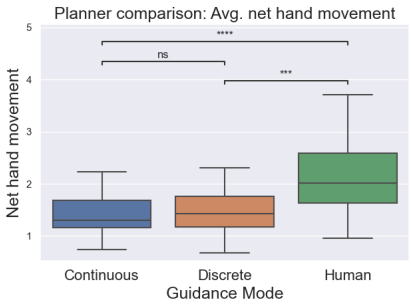

Figure 12: Average net hand movement caused by the different planners.

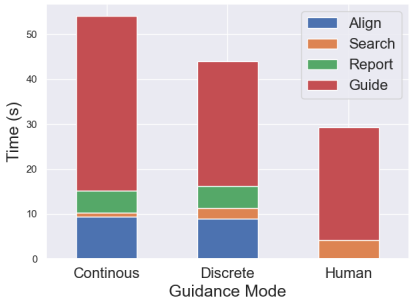

Figure 13: Average time taken by each planner. The alignment, search, and report durations are similar for Continuous and Discrete but the guidance time is lower for Discrete, comparable to human levels.

1.01) out of 5, indicating that this component was valuable to them. The large standard deviation is attributed to two participants rating this feature very low and in the exit interview they mentioned that it was hard and confusing for them to understand the overview template. We noticed participants mentioning a clear preference for one or the other planner. Some participants who liked the continuous mode mentioned that they liked that it provided some affirmation in the form of *"Stop"* command. We also noticed the human caller using some kind of affirmation along with the movement instructions, for example, *"Alright"*, *"Good"*, *"That one"*. Participants who liked the discrete planner more mentioned that they liked getting precise movement instructions and not relying on the system to stop them without much certainty about how much they would have to move. Participants also rated ShelfHelp favorably on metrics such as *ease of use, confidence, mental demand, temporal demand, frustration.*

5 DISCUSSION AND FUTURE WORK

We caution the reader to exercise care to avoid drawing strong conclusions about device readiness based solely on pilot tests with blindfolded, non-BVI participants, as these are not necessarily a reliable proxy for the (eventual) target population of this device [12, 37], but we find the preliminary results promising and indicative of value in further investigation. With this pilot study, we have shown a proof-of-concept with regard to the hardware and software features of our proposed self-contained system (i.e., not dependent on external compute or data connection), alongside a validation

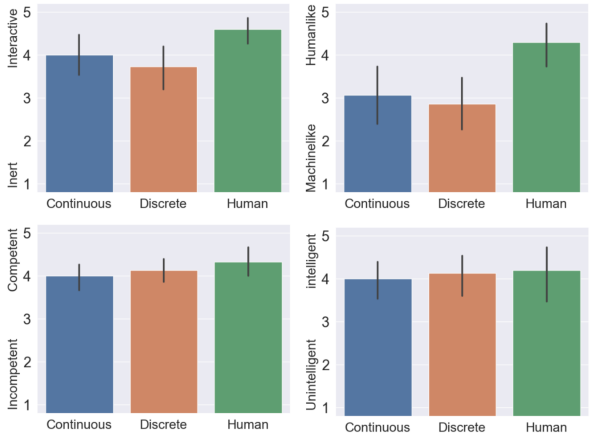

Figure 14: Average rating on subjective metrics: *(clockwise from upper-left) Interactive, Humanlike, Intelligent, Competent.* Both planners performed well on all the metrics compared to the human guide except for *Humanlike.*

of a novel product locator system coupled with a novel fine-grain manipulation guidance system, demonstrating that the system is able to locate a target product, plan to reach it, and effectively guide retrieval in real-time. The discrete planner performs quantitatively better than the continuous and is on par with the human in our study. The continuous planner elicited a positive response as well, in part due to the affirmations it provides by using the *"Stop"* command. It could be beneficial to add this feature to the discrete planner, effectively grounding the user during the execution of the plan. The product locator system relies on visual features which do not effectively leverage all of the available semantic information on product packaging, and a future direction would be real-time incorporation of semantic information as well.

6 CONCLUSION

In this work, we present our system ShelfHelp which enhances an instrumented cane meant for navigational assistance with the addition of manipulation assistance capabilities. It includes a novel, maintenance-free product locator system that does not need regular re-training as new products are introduced, and works offline making it scalable and practical. We present a novel fine-grain manipulation guidance planner that effectively guides manipulation-based reachability in the haptic space of the user. We tested aspects of the system using a pilot study with novice users that provided an initial validation for ShelfHelp and show that our discrete guidance planner optimizes for guide time and the total number of commands without compromising legibility. We tested two guidance systems with a quantitative and qualitative comparison against each other and a human baseline. The discrete guidance mode was on par with the human caller in terms of guide time, the number of commands, perceived competence, and intelligence but lacked in terms of interactiveness in form of affirmation. The study also surfaced positive feedback for qualitative metrics such as ease of use, confidence, mental demand, temporal demand, frustration, intelligence, and competence for the system with both of the guidance planners.

REFERENCES

[1] 2022. *Be My Eyes.* https://www.bemyeyes.com/

[2] 2022. *Lookout.* https://play.google.com/store/apps/details?id=com.google.android.apps.accessibility.reveal&hl=en_US&gl=US

[3] Tousif Ahmed, Roberto Hoyle, Kay Connelly, David Crandall, and Apu Kapadia. 2015. Privacy concerns and behaviors of people with visual impairments. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 3523–3532.

[4] Yiannoula Andreou and Konstantinos T Kotsis. 2005. The estimation of length, surface area, and volume by blind and sighted children. In *International Congress Series*, Vol. 1282. Elsevier, 780–784.

[5] Anonymized. In press. A Novel Perceptive Robotic Cane with Haptic Navigation for Enabling Vision-Independent Participation in the Social Dynamics of Seat Choice. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. -

[6] Jeffrey P Bigham, Chandrika Jayant, Andrew Miller, Brandyn White, and Tom Yeh. 2010. VizWiz:: LocateIt-enabling blind people to locate objects in their environment. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE, 65–72.

[7] Mayara Bonani, Raquel Oliveira, Filipa Correia, André Rodrigues, Tiago Guerreiro, and Ana Paiva. 2018. What my eyes can't see, a robot can show me: Exploring the collaboration between blind people and robots. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*. 15–27.

[8] Rupert Bourne, Jaimie D Steinmetz, Seth Flaxman, Paul Svitil Briant, Hugh R Taylor, Serge Resnikoff, Robert James Casson, Amir Abdoli, Eman Abu-Gharbieh, Ashkan Afshin, et al. 2021. Trends in prevalence of blindness and distance and near vision impairment over 30 years: an analysis for the Global Burden of Disease Study. *The Lancet global health* 9, 2 (2021), e130–e143.

[9] Qingtian Chen, Muhammad Khan, Christina Tsangouri, Christopher Yang, Bing Li, Jizhong Xiao, and Zhigang Zhu. [n.d.]. CCNY Smart Cane. In *2017 IEEE 7th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems*. https://doi.org/10.1109/CYBER.2017.8446303

[10] Wei Chen, Yu Liu, Weiping Wang, Erwin Bakker, Theodoros Georgiou, Paul Fieguth, Li Liu, and Michael S Lew. 2021. Deep Learning for Instance Retrieval: A Survey. *arXiv preprint arXiv:2101.11282* (2021).

[11] Il Yong Chung, Sanghag Kim, and Kang Hyeon Rhee. 2014. The smart cane utilizing a smart phone for the visually impaired person. In *2014 IEEE 3rd Global Conference on Consumer Electronics (GCCE)*. 106–107. https://doi.org/10.1109/GCCE.2014.7031333

[12] Aline Darc dos Santos, Fausto Orsi Medola, Milton José Cinelli, Alejandro Rafael Garcia Ramirez, and Frode Eika Sandnes. 2020. Are electronic white canes better than traditional canes? A comparative study with blind and blindfolded participants. *Universal Access in the Information Society* 20, 1 (2020). https://doi.org/10.1007/s10209-020-00712-z

[13] Chia-Hui Feng, Ju-Yen Hsieh, Yu-Hsiu Hung, Chung-Jen Chen, and Cheng-Hung Chen. 2020. Research on the Visually Impaired Individuals Shopping with Artificial Intelligence Image Recognition Assistance. In *International Conference on Human-Computer Interaction*. Springer, 518–531.

[14] A Jin Fukasawa and Kazusihge Magatani. [n.d.]. A navigation system for the visually impaired an intelligent white cane. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*.

[15] Chaitanya P Gharpure and Vladimir A Kulyukin. 2008. Robot-assisted shopping for the blind: issues in spatial cognition and product selection. *Intelligent Service Robotics* 1, 3 (2008), 237–251.

[16] Eran Goldman, Roei Herzig, Aviv Eisenschtat, Jacob Goldberger, and Tal Hassner. 2019. Precise Detection in Densely Packed Scenes. In *Proc. Conf. Comput. Vision Pattern Recognition (CVPR)*.

[17] Reginald G Golledge, Roberta L Klatzky, and Jack M Loomis. 1996. Cognitive mapping and wayfinding by adults without vision. In *The construction of cognitive maps*. Springer, 215–246.

[18] João Guerreiro, Daisuke Sato, Saki Asakawa, Huixu Dong, Kris M. Kitani, and Chieko Asakawa. 2019. CaBot: Designing and Evaluating an Autonomous Navigation Robot for Blind People. In *The 21st ACM SIGACCESS Conference on Computers and Accessibility* (Pittsburgh, PA, USA). NY, USA. https://doi.org/10.1145/3308561.3353771

[19] Robert K. Katzschmann, Brandon Araki, and Daniela Rus. 2018. Safe Local Navigation for Visually Impaired Users With a Time-of-Flight and Haptic Feedback Device. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 26, 3 (2018), 583–593. https://doi.org/10.1109/TNSRE.2018.2800665

[20] Vladimir Kulyukin, Chaitanya Gharpure, and John Nicholson. 2005. Robocart: Toward robot-assisted navigation of grocery stores by the visually impaired. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2845–2850.

[21] Vladimir Kulyukin and Aliasgar Kutiyanawala. 2010. Accessible shopping systems for blind and visually impaired individuals: Design requirements and the state of the art. *The Open Rehabilitation Journal* 3, 1 (2010).

[22] Vladimir A Kulyukin and Chaitanya Gharpure. 2006. Ergonomics-for-one in a robotic shopping cart for the blind. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*. 142–149.

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[24] Rajesh Kannan Megalingam, Aparna Nambissan, Anu Thambi, Anjali Gopinath, and Megha Nandakumar. 2014. Sound and touch based smart cane: Better walking experience for visually challenged. In *2014 IEEE Canada International Humanitarian Technology Conference - (IHTC)*. 1–4. https://doi.org/10.1109/IHTC.2014.7147543

[25] Vidula V. Meshram, Kailas Patil, Vishal A. Meshram, and Felix Che Shu. 2019. An Astute Assistive Device for Mobility and Object Recognition for Visually Impaired People. *IEEE Trans. on Human-Machine Systems* (2019). https://doi.org/10.1109/THMS.2019.2931745

[26] Sharada Murali, R Shrivatsan, V Sreenivas, Srihaarika Vijjappu, S Joseph Gladwin, and R Rajavel. 2016. Smart walking cane for the visually challenged. In *IEEE Region 10 Humanitarian Technology Conference*. 1–4. https://doi.org/10.1109/R10-HTC.2016.7906791

[27] Arshad Nasser, Kai-Ning Keng, and Kening Zhu. 2020. ThermalCane: Exploring Thermotactile Directional Cues on Cane-Grip for Non-Visual Navigation. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, Greece). New York, NY, USA, Article 20, 12 pages. https://doi.org/10.1145/3373625.3417004

[28] John Nicholson, Vladimir Kulyukin, and Daniel Coster. 2009. ShopTalk: independent blind shopping through verbal route directions and barcode scans. *The Open Rehabilitation Journal* 2, 1 (2009).

[29] NielsenIQ. 2019. *Bursting with new products, there's never been a better time for breakthrough innovation.* https://nielseniq.com/global/en/insights/analysis/2019/bursting-with-new-products-theres-never-been-a-better-time-for-breakthrough-innovation

[30] Yoshihiro Niitsu, Toshiki Taniguchi, and Kouichiro Kawashima. 2014. Detection and notification of dangerous obstacles and places for visually impaired persons using a smart cane. In *2014 Seventh International Conference on Mobile Computing and Ubiquitous Networking (ICMU)*. 68–69. https://doi.org/10.1109/ICMU.2014.6799060

[31] Ryuta Okazaki and Hiroyuki Kajimoto. 2014. Perceived distance from hitting with a stick is altered by overlapping vibration to holding hand. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems*. 1903–1908.

[32] Emily E O'Brien, Aaron A Mohtar, Laura E Diment, and Karen J Reynolds. 2014. A detachable electronic device for use with a long white cane to assist with mobility. *Assistive Technology* 26, 4 (2014), 219–226.

[33] Santiago Real and Alvaro Araujo. 2019. Navigation Systems for the Blind and Visually Impaired: Past Work, Challenges, and Open Problems. *Sensors* (2019). https://doi.org/10.3390/s19153404

[34] M.F. Saaid, A.M. Mohammad, and M.S.A. Megat Ali. 2016. Smart cane with range notification for blind people. In *2016 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS)*. https://doi.org/10.1109/I2CACIS.2016.7885319

[35] Manaswi Saha, Alexander J Fiannaca, Melanie Kneisel, Edward Cutrell, and Meredith Ringel Morris. 2019. Closing the gap: Designing for the last-few-meters wayfinding problem for people with visual impairments. In *The 21st international acm sigaccess conference on computers and accessibility*. 222–235.

[36] Jayant Sakhardande, Pratik Pattanayak, and Mita Bhowmick. 2012. Smart Cane Assisted Mobility for the Visually Impaired. *International Journal of Electrical and Computer Engineering* 6, 10 (2012), 1262 – 1265.

[37] Arielle M. Silverman, Jason D. Gwinn, and Leaf Van Boven. 2015. Stumbling in Their Shoes: Disability Simulations Reduce Judged Capabilities of Disabled People. *Social Psychological and Personality Science* 6, 4 (2015), 464–471. https://doi.org/10.1177/1948550614559650 arXiv:https://doi.org/10.1177/1948550614559650

[38] Bhupendra Singh and Monit Kapoor. 2020. Assistive cane for visually impaired persons for uneven surface detection with orientation restraint sensing. *Sensor Review* 40, 6 (2020), 687–698. https://doi.org/10.1108/sr-04-2020-0097

[39] Milo Marsfeldt Skovfoged, Alexander Schiller Rasmussen, Lucie Kalova, Laura-Dora Daczo, and Hendrik Knoche. 2022. " Is it There or Not?" Why Augmented White Canes Do Not Need to Provide Detailed Feedback about Obstacles. In *Adjunct Proceedings of the 2022 Nordic Human-Computer Interaction Conference*. 1–5.

[40] Patrick Slade, Arjun Tambe, and Mykel J Kochenderfer. 2021. Multimodal sensing and intuitive steering assistance improve navigation and mobility for people with impaired vision. *Science Robotics* 6, 59 (2021).

[41] Hotaka Takizawa, Shotaro Yamaguchi, Mayumi Aoyagi, Nobuo Ezaki, and Shinji Mizuno. 2012. Kinect cane: An assistive system for the visually impaired based on three-dimensional object recognition. In *2012 IEEE/SICE International Symposium on System Integration (SII)*. 740–745. https://doi.org/10.1109/SII.2012.6426936

[42] I. Ulrich and J. Borenstein. 2001. The GuideCane-applying mobile robot technologies to assist the visually impaired. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* (2001). https://doi.org/10.1109/3468.911370

[43] Melvin Varghese, Shreeprasad S. Manohar, Kean Rodrigues, Vinayak Kodkani, and Shantanu Pendse. 2015. The smart guide cane: An enhanced walking cane for assisting the visually challenged. In *2015 International Conference on Technologies for Sustainable Development (ICTSD)*. 1–5. https://doi.org/10.1109/ICTSD.2015.7095907

[44] Marynel Vázquez and Aaron Steinfeld. 2014. An assisted photography framework to help visually impaired users properly aim a camera. *ACM Transactions on Computer-Human Interaction (TOCHI)* 21, 5 (2014), 1–29.

[45] Andreas Wachaja, Pratik Agarwal, Mathias Zink, Miguel Reyes Adame, Knut Möller, and Wolfram Burgard. 2017. Navigating blind people with walking impairments using a smart walker. *Autonomous Robots* 41, 3 (2017).

[46] Mohd Helmy Abd Wahab, Amirul A. Talib, Herdawatie A. Kadir, Ayob Johari, A. Noraziah, Roslina M. Sidek, and Ariffin A. Mutalib. 2011. Smart Cane: Assistive Cane for Visually-impaired People. arXiv:1110.5156 [cs.SY]

[47] Hsueh-Cheng Wang, Robert K. Katzschmann, Santani Teng, Brandon Araki, Laura Giarré, and Daniela Rus. 2017. Enabling independent navigation for visually impaired people through a wearable vision-based feedback system. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*. https://doi.org/10.1109/ICRA.2017.7989772

[48] Anxing Xiao, Wenzhe Tong, Lizhi Yang, Jun Zeng, Zhongyu Li, and Koushil Sreenath. 2021. Robotic Guide Dog: Leading a Human with Leash-Guided Hybrid Physical Interaction. arXiv:2103.14300 [cs.RO]

[49] Chien Wen Yuan, Benjamin V Hanrahan, Sooyeon Lee, Mary Beth Rosson, and John M Carroll. 2019. Constructing a holistic view of shopping with people with visual impairment: a participatory design approach. *Universal Access in the Information Society* 18, 1 (2019), 127–140.

[50] He Zhang and Cang Ye. 2017. An indoor wayfinding system based on geometric features aided graph SLAM for the visually impaired. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25, 9 (2017), 1592–1604.

[51] Yuhang Zhao, Sarit Szpiro, Jonathan Knighten, and Shiri Azenkot. 2016. CueSee: exploring visual cues for people with low vision to facilitate a visual search task. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 73–84.