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

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

I. 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. In this work we present our work-in-progress investigating technical solutions for enhancing instrumented canes traditionally meant for navigation tasks with capability within the domain of shopping. Our system includes a novel visual product search algorithm designed for use in the wild and a novel planner that autonomously issues verbal commands to guide the user in a reaching task to acquire them.

II. Introduction and related work

It is estimated that 295 million people have some form of vision impairment, of whom 43.3 million are blind. Currently, when encountering difficulty with shopping, they can get help from sighted humans. 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 [1], [2]. Our work seeks to alleviate the dependence on guide availability and to mitigate the loss of privacy encountered with traditional support mechanisms. Our system extends the capability of a robotic smart cane [3] created originally for social navigation assistance (Fig.1). This area has been extensively researched [4]–[26], 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. Recent work on grocery assistant systems focuses largely on navigation inside the store [1], [2], [27]–[29], 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 significant barriers to adoption as they are inapplicable in uninstrumented domains. Some researchers have focused on other issues around shopping, including identifying products that users are running out of and organizing newly purchased products at home [30], as well as "the last few meters" way-finding problem [4] and solutions for people with low vision [31]. We focus on an unsolved research area that primarily considers the haptic space of the user. In other words, our solution addresses 1) the problem of locating a desired product and 2) the challenge of providing effective verbal guidance to reach and grasp the product.

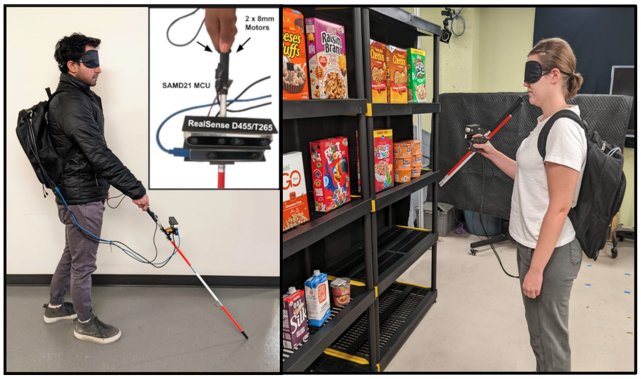

Fig. 1: Our system includes a robotic cane equipped with RealSense D455 and T265 cameras. The system is powered through a laptop in the backpack. *Left:* The system used as a navigational device. It used audio and haptic feedback for navigation guidance. *Right:* The system used as a manipulation device.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

Assistive manipulation guidance is an area that has been explored within the robotics community for over a decade. Vasques et al. [32] 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. [33] showed promise for the concept with an experimenter-controlled teleoperated system and Bigham et al. [34] did so with a mechanical turk-based system, but fully autonomous implementations were outside the scope of their contributions. The most popular solution in this problem domain is a human-powered service called Be My Eyes [35], but this service suffers from scalability issues due to its reliance on available humans, is not readily available in developing countries, 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 use that to formulate the problem as a Markov Decision Process (MDP) that can be solved with well established reinforcement learning techniques to inform our guidance of the user.

III. Current work

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

1) Perception: We have developed a novel two-stage product search system. This problem falls under the category of *instance retrieval*, where the task is to find a target image in the scene [36]. Existing techniques with a fixed number of output classes perform poorly [37] on products because of the sheer amount of products [38] available and the slight variations they come in as it is infeasible to create an object classifier and keep it up-to-date.

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. In the first stage, our method proposes regions in form of bounding boxes that are most likely to contain *any* product. We train the YoloV5 network on the SKU-110K dataset [39] to create a product detector. In the second stage, an encoder is used as a feature extractor that matches the features of the proposed regions and the target image, finding the best possible match (Fig 2). We do this by training an autoencoder on MS-COCO color images [40] and then utilizing the encoder portion as the feature extractor. This method doesn't require any retraining and works in real-time. The encoder transforms images (the proposed regions and the target image) to vectors in a latent space where we use *cosine similarity* to find closer vectors. We empirically determine a similarity score threshold (0.5), that captures satisfactory performance across real-world environments, but this value can easily be fine-tuned in case there is significant distribution shift between the evaluation and deployment environments. 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 Simulataneous Localization and Mapping (SLAM) algorithm and fuse it with the depth information obtained from a cane-mounted RealSense D455. 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.

2) Data Association: We use data association techniques to identify each instance of the same product uniquely across subsequent frames of camera capture. 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 pose 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. 2 - lower left). The IMU data from the T265 sensor helps us to calculate the object's pose in a global frame of reference. This is necessary for data association and it also helps in creating a "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 [32]. 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).

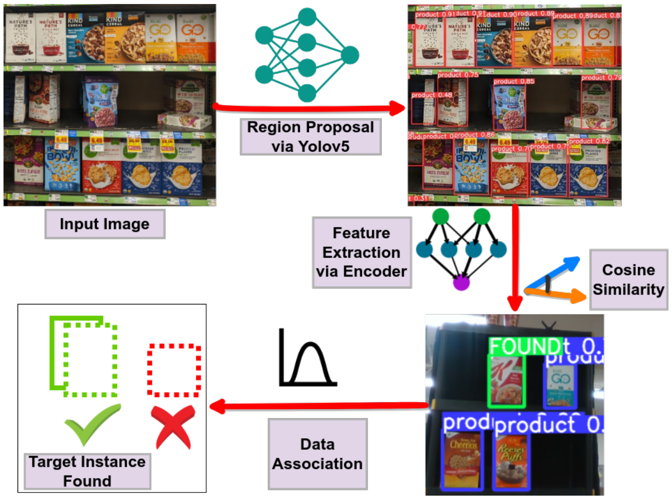

Fig. 2: Our product search algorithm can reliably locate desired products in the wild. Regions with a high likelihood of containing any products are proposed in the first stage. The features of these regions are then compared against the target product image. Our data association solution is used to identify whether detections from incoming camera frames are new or re-detections of existing products. The product classification aspect of this work has been tested and validated in actual grocery stores, whereas the data association and manipulation assistance components are currently being validated within a lab-based study.

3) Scoring: Our system then scores each detected product instance of the target product and picks the one with the highest score as its planning goal. It considers the rolling similarity with the target image and the spatial information. This allows the system to exercise some 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 3).

4) Planning: We have developed two different guidance mechanisms to provide verbal instruction once the target has been located (continuous versus discrete guidance), for

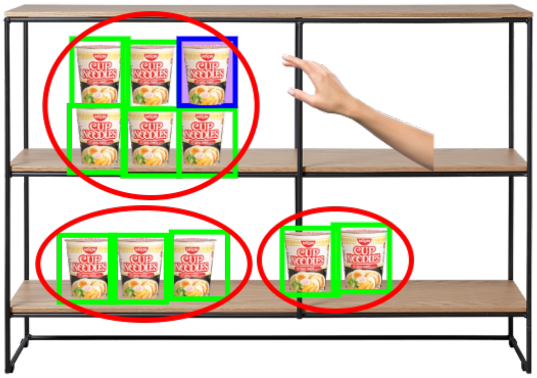

Fig. 3: The spatial scoring system clusters all the found instances spatially and gives preference to the closest cluster to the current hand pose. Ties are broken arbitrarily.

example, “keep on going right”...“stop” or “move 6 inches to the right”. 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.

The decision to create a discrete guidance mode was inspired by study results showing that PVI have been known to perceive length units even better than sighted people [41] and the criteria of minimizing verbal feedback for the task. 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 4). We formed 36 discrete commands and issued 1220 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 4 shows a sample from this dataset illustrating commands pertaining to *left* 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.

This information is used to formulate this problem as an MDP (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 the previous command which could be any of six values corresponding to the X, Y, or Z 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 can solve for all the potential states that can be encountered. We discretized the states at 10 cm resolution and considered a cuboid region of 1.5m as the operational space for the human hand. A is the set of discrete verbal commands. T is calculated from X_c as the movement caused by each command is not deterministic. 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. 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.

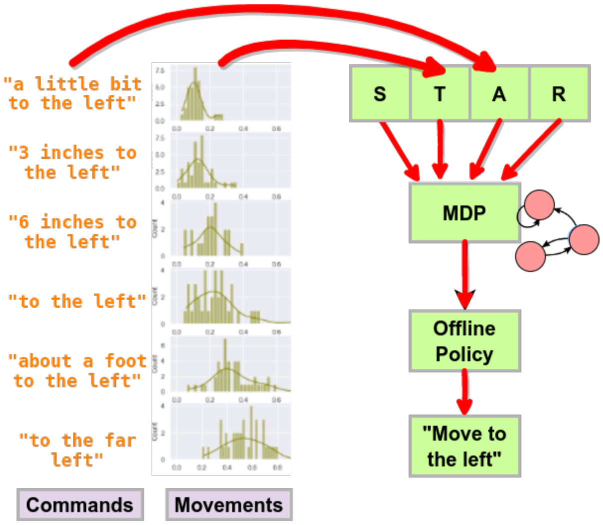

Fig. 4: *From left to right:* A sample of discrete commands. The movement (in meters) each command caused. We learned a model of human hand movement from demonstrations that gave us the transition probabilities T. S defines the state space, A defines the discrete set of verbal actions, and R is the reward function. A policy is learned offline that can be used across reaching tasks.

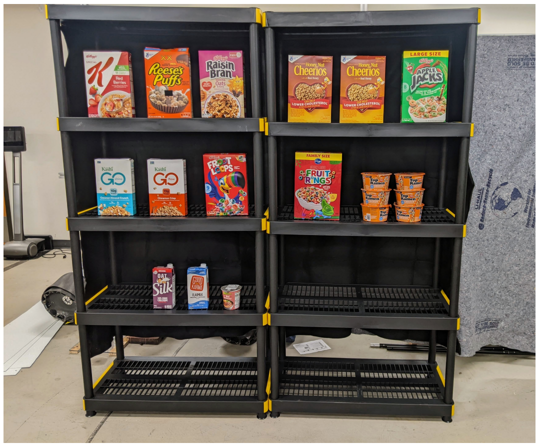

Fig. 5: An experimental setup approximating a grocery store shelf, used for evaluating the efficacy of our proposed manipulation guidance system.

5) Conveyance: The commands 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.

References

[1] V. Kulyukin, C. Gharpure, and J. Nicholson, "Robocart: Toward robot-assisted navigation of grocery stores by the visually impaired," in *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2005, pp. 2845–2850.

[2] V. A. Kulyukin and C. Gharpure, "Ergonomics-for-one in a robotic shopping cart for the blind," in *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, 2006, pp. 142–149.

[3] S. Agrawal, M. West, and B. Hayes, "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, In press. [Online]. Available: http://www.cairo-lab.com/papers/iros22.pdf

[4] M. Saha, A. J. Fiannaca, M. Kneisel, E. Cutrell, and M. R. Morris, "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*, 2019, pp. 222–235.

[5] S. Real and A. Araujo, "Navigation systems for the blind and visually impaired: Past work, challenges, and open problems," *Sensors*, 2019.

[6] H. Takizawa, S. Yamaguchi, M. Aoyagi, N. Ezaki, and S. Mizuno, "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)*, 2012, pp. 740–745.

[7] Q. Chen, M. Khan, C. Tsangouri, C. Yang, B. Li, J. Xiao, and Z. Zhu, "Ccny smart cane," in *2017 IEEE 7th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems*.

[8] H. Zhang and C. Ye, "An indoor wayfinding system based on geometric features aided graph slam for the visually impaired," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 9, pp. 1592–1604, 2017.

[9] M. Saaid, A. Mohammad, and M. Megat Ali, "Smart cane with range notification for blind people," in *2016 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS)*, 2016.

[10] Y. Niitsu, T. Taniguchi, and K. Kawashima, "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)*, 2014, pp. 68–69.

[11] S. Murali, R. Shrivatsan, V. Sreenivas, S. Vijjappu, S. J. Gladwin, and R. Rajavel, "Smart walking cane for the visually challenged," in *IEEE Region 10 Humanitarian Technology Conference*, 2016, pp. 1–4.

[12] H.-C. Wang, R. K. Katzschmann, S. Teng, B. Araki, L. Giarré, and D. Rus, "Enabling independent navigation for visually impaired people through a wearable vision-based feedback system," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2017.

[13] J. Sakhardande, P. Pattanayak, and M. Bhowmick, "Smart cane assisted mobility for the visually impaired," *International Journal of Electrical and Computer Engineering*, vol. 6, no. 10, pp. 1262 – 1265, 2012.

[14] M. H. A. Wahab, A. A. Talib, H. A. Kadir, A. Johari, A. Noraziah, R. M. Sidek, and A. A. Mutalib, "Smart cane: Assistive cane for visually-impaired people," 2011.

[15] R. K. Megalingam, A. Nambissan, A. Thambi, A. Gopinath, and M. Nandakumar, "Sound and touch based smart cane: Better walking experience for visually challenged," in *2014 IEEE Canada International Humanitarian Technology Conference - (IHTC)*, 2014, pp. 1–4.

[16] B. Singh and M. Kapoor, "Assistive cane for visually impaired persons for uneven surface detection with orientation restraint sensing," *Sensor Review*, vol. 40, no. 6, p. 687–698, 2020.

[17] V. V. Meshram, K. Patil, V. A. Meshram, and F. C. Shu, "An astute assistive device for mobility and object recognition for visually impaired people," *IEEE Trans. on Human-Machine Systems*, 2019.

[18] M. Varghese, S. S. Manohar, K. Rodrigues, V. Kodkani, and S. Pendse, "The smart guide cane: An enhanced walking cane for assisting the visually challenged," in *2015 International Conference on Technologies for Sustainable Development (ICTSD)*, 2015, pp. 1–5.

[19] A. Nasser, K.-N. Keng, and K. Zhu, "Thermalcane: Exploring thermotactile directional cues on cane-grip for non-visual navigation," in *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, New York, NY, USA, 2020.

[20] I. Ulrich and J. Borenstein, "The guidecane-applying mobile robot technologies to assist the visually impaired," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 2001.

[21] I. Y. Chung, S. Kim, and K. H. Rhee, "The smart cane utilizing a smart phone for the visually impaired person," in *2014 IEEE 3rd Global Conference on Consumer Electronics (GCCE)*, 2014, pp. 106–107.

[22] J. a. Guerreiro, D. Sato, S. Asakawa, H. Dong, K. M. Kitani, and C. Asakawa, "Cabot: Designing and evaluating an autonomous navigation robot for blind people," in *The 21st ACM SIGACCESS Conference on Computers and Accessibility*, NY, USA, 2019.

[23] A. Xiao, W. Tong, L. Yang, J. Zeng, Z. Li, and K. Sreenath, "Robotic guide dog: Leading a human with leash-guided hybrid physical interaction," 2021.

[24] R. K. Katzschmann, B. Araki, and D. Rus, "Safe local navigation for visually impaired users with a time-of-flight and haptic feedback device," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 3, pp. 583–593, 2018.

[25] P. Slade, A. Tambe, and M. J. Kochenderfer, "Multimodal sensing and intuitive steering assistance improve navigation and mobility for people with impaired vision," *Science Robotics*, vol. 6, no. 59, 2021.

[26] A. Wachaja, P. Agarwal, M. Zink, M. R. Adame, K. Möller, and W. Burgard, "Navigating blind people with walking impairments using a smart walker," *Autonomous Robots*, vol. 41, no. 3, 2017.

[27] C. P. Gharpure and V. A. Kulyukin, "Robot-assisted shopping for the blind: issues in spatial cognition and product selection," *Intelligent Service Robotics*, vol. 1, no. 3, pp. 237–251, 2008.

[28] V. Kulyukin and A. Kutiyanawala, "Accessible shopping systems for blind and visually impaired individuals: Design requirements and the state of the art," *The Open Rehabilitation Journal*, vol. 3, no. 1, 2010.

[29] J. Nicholson, V. Kulyukin, and D. Coster, "Shoptalk: independent blind shopping through verbal route directions and barcode scans," *The Open Rehabilitation Journal*, vol. 2, no. 1, 2009.

[30] C. W. Yuan, B. V. Hanrahan, S. Lee, M. B. Rosson, and J. M. Carroll, "Constructing a holistic view of shopping with people with visual impairment: a participatory design approach," *Universal Access in the Information Society*, vol. 18, no. 1, pp. 127–140, 2019.

[31] Y. Zhao, S. Szpiro, J. Knighten, and S. Azenkot, "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*, 2016, pp. 73–84.

[32] M. Vázquez and A. Steinfeld, "An assisted photography framework to help visually impaired users properly aim a camera," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 21, no. 5, pp. 1–29, 2014.

[33] M. Bonani, R. Oliveira, F. Correia, A. Rodrigues, T. Guerreiro, and A. Paiva, "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*, 2018, pp. 15–27.

[34] J. P. Bigham, C. Jayant, A. Miller, B. White, and T. Yeh, "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, 2010, pp. 65–72.

[35] Be my eyes. [Online]. Available: https://www.bemyeyes.com/

[36] W. Chen, Y. Liu, W. Wang, E. Bakker, T. Georgiou, P. Fieguth, L. Liu, and M. S. Lew, "Deep learning for instance retrieval: A survey," *arXiv preprint arXiv:2101.11282*, 2021.

[37] C.-H. Feng, J.-Y. Hsieh, Y.-H. Hung, C.-J. Chen, and C.-H. Chen, "Research on the visually impaired individuals shopping with artificial intelligence image recognition assistance," in *International Conference on Human-Computer Interaction*. Springer, 2020, pp. 518–531.

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

[39] E. Goldman, R. Herzig, A. Eisenschtat, J. Goldberger, and T. Hassner,

"Precise detection in densely packed scenes," in *Proc. Conf. Comput. Vision Pattern Recognition (CVPR)*, 2019.

[40] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[41] Y. Andreou and K. T. Kotsis, "The estimation of length, surface area, and volume by blind and sighted children," in *International Congress Series*, vol. 1282. Elsevier, 2005, pp. 780–784.