Toward Human-Like Social Robot Navigation: A Large-Scale, Multi-Modal, Social Human Navigation Dataset

Duc M. Nguyen, Mohammad Nazeri, Amirreza Payandeh, Aniket Datar, and Xuesu Xiao

***Abstract*— Humans are well-adept at navigating public spaces shared with others, where current autonomous mobile robots still struggle: while safely and efficiently reaching their goals, humans communicate their intentions and conform to unwritten social norms on a daily basis; conversely, robots become clumsy in those daily social scenarios, getting stuck in dense crowds, surprising nearby pedestrians, or even causing collisions. While recent research on robot learning has shown promises in data-driven social robot navigation, good-quality training data is still difficult to acquire through either trial and error or expert demonstrations. In this work, we propose to utilize the body of rich, widely available, social human navigation data in many natural human-inhabited public spaces for robots to learn similar, human-like, socially compliant navigation behaviors. To be specific, we design an open-source egocentric data collection sensor suite wearable by walking humans to provide multi-modal robot perception data; we collect a large-scale (∼100 km, 20 hours, 300 trials, 13 humans) dataset in a variety of public spaces which contain numerous natural social navigation interactions.**[1]

I. INTRODUCTION

Social navigation is the capability of an autonomous agent to navigate in a way such that it not only moves toward its goal but also takes other agents' objective into consideration. Most humans are proficient at such a task, smoothly navigating many public spaces shared with others on a daily basis: humans form lanes or groups among crowds, use gaze, head movement, and body posture to communicate navigation intentions, wait in line to enter congested areas, or give way to others who are in a rush. With an increasing amount of autonomous mobile robots being deployed in public spaces [1], [2], those robots are also expected to navigate among humans in a similar, human-like, socially compliant manner.

However, the autonomous navigation performance of these mobile robots is still far from satisfactory. Despite extensive robotics effort to create efficient and collision-free autonomous navigation systems, we still witness the "frozen robot" problem in dense crowds and robots moving against upcoming foot traffic or cutting too close to moving humans. Unfortunately, due to such deficiencies, there is increasing fear about the public adoption and even safety of humans around these robots [3], [4]. The current lack of safe and socially compliant navigation systems still presents a major hurdle preventing service robots being widely adopted.

All authors are with the Department of Computer Science, George Mason University {mnguy21, mnazerir, apayande, adatar, xiao}@gmu.edu

[1]Website: https://cs.gmu.edu/~xiao/Research/MuSoHu/

Fig. 1: Data collection in natural human-inhabited public spaces with the open-source sensor suite including 3D LiDAR, stereo and depth camera, IMU, microphone array, and 360° camera.

One avenue toward socially compliant robot navigation is using machine learning for robots to learn the variety of unwritten social norms, for which traditional cost functions are hard to design. For example, Reinforcement Learning (RL) [5] uses trial-and-error experiences while Imitation Learning (IL) [6] requires expert demonstrations. However, both of these learning paradigms require an extensive amount of training data, which is difficult to acquire: RL in the real world is extremely expensive due to the limited availability of robots, while RL in simulation requires a good model of social navigation interactions of humans, which are what roboticists are trying to create in the first place; IL requires demonstration datasets collected on robot platforms, mostly through expensive human teleoperation at scale [7].

Considering the goal of creating socially compliant robot navigation and the availability of many humans that excel at such a task, this work leverages the easily accessible social human navigation data in public spaces for mobile robots to learn from. To be specific, we first present an open source, first-person-view, social human navigation data collection sensor suite that can be worn on the head of a walking human and provide easy access to a large body of readily available, high-quality, natural social navigation data in the wild for robot learning, as shown in Fig. 1. Our design includes a set of different robotic sensors: a 3D Light Detection and Ranging (LiDAR) sensor, stereo and depth camera, Inertia Measurement Unit (IMU), microphone array, and 360° camera. We open-source our design and software so the sensor suite can be easily replicated and used to collect social human navigation data in different places. Second, with the new data collection suite, we introduce our Multi-modal Social Human Navigation dataset (MuSoHu):

a large-scale, egocentric, multi-modal, and context-aware dataset of human demonstrations of social navigation. At the point when this paper is written, MuSoHu contains approximately 20 hours, 300 trajectories, 100 kilometers of socially compliant navigation demonstrations collected by 13 human demonstrators that comprise multi-modal data streams from different sensors, in both indoor and outdoor environments.

II. RELATED WORK

In this section, we review related work in social robot navigation and learning from human datasets.

A. Social Robot Navigation

To organically integrate service robots into the fabric of our society, these robots must be capable of moving in human-inhabited spaces in a socially compliant manner. One difficulty in creating such socially compliant navigation systems is to hand-craft appropriate rules or cost functions to cope with unwritten social norms in public spaces [8]. Therefore, researchers have sought help from machine learning and aimed at *learning* socially compliant navigation behaviors in a data-driven manner [9], [10].

RL has shown success in learning a variety of behaviors from simulated trial-and-error experiences [11]–[13]. However, the high fidelity of simulated social interactions required by RL for social navigation poses its own challenges and requires a good understanding and then analytical representation of the unwritten social norms to create such simulated interactions, which is the difficulty in social navigation in the first place. Additionally, the reward function in RL needs to be carefully-designed but can still be brittle [14].

To address such issues, IL [15], [16] utilizes expert demonstrations to learn socially compliant navigation behaviors [17], [18]. Kretzschmar et al. [19] has proposed Inverse Reinforcement Learning (IRL) to learn the reward function from demonstrations for social navigation policies. Behavior Cloning [15], [16] has treated the social navigation problem as supervised learning and regressed to an end-to-end motion policy that maps from perception to actions. However, to facilitate IL, a large corpus of socially compliant navigation demonstration data is essential. For example, the Socially Compliant Navigation Dataset (SCAND) [7] is a recent effort to provide social robot navigation behaviors demonstrated by human teleoperation.

B. Learning from Human Datasets

SCAND [7] is a recent dataset that aims at tackling the challenges of socially compliant robot navigation. SCAND includes socially compliant, human teleoperated robot navigation demonstrations in indoor and outdoor environments on The University of Texas at Austin campus. Using SCAND, researchers have shown that IL policies can be trained end-to-end for socially-aware global and local planners for robot navigation. However, SCAND requires a significant amount of cost and effort to set up and deploy the robot platforms in the wild and to collect large-scale human-teleoperated robot navigation demonstrations to cover the plethora of interesting social interactions in public spaces. Furthermore, how people react differently to a teleoperated mobile robot followed by a human operator is also unclear.

Considering the difficulty in acquiring large-scale real-world data, researchers have also looked into utilizing recorded videos of human activities in the wild. For example Ego4D [20] is an egocentric video dataset, which offers daily-life activity video of different scenarios (household, outdoor, workplace, leisure, etc.) captured by different humans wearing cameras from different locations worldwide. Ego4D offers a solution to the scalability of datasets by introducing a standard and wearable design so many people can collect data in real-world, daily settings from different parts of the world. However, Ego4D is not specifically designed for robotics (hence the lack of common robot sensors and perception like LiDAR, depth camera, IMU, and odometry), so it is difficult for mobile robots to directly learn socially compliant navigation behaviors from the raw video feed in Ego4D.

Inspired by the pros and cons of both SCAND and Ego4D, we introduce a wearable data collection sensor suite specifically designed to provide data to enable social robot navigation. It allows us to collect social human navigation data from the perspective of a suite of multi-modal robotic sensors in our daily life with a small setup overhead (i.e., with a wearable helmet). We provide a large-scale social human navigation dataset, which can be easily extended in the future by robotics researchers all around the world, show that human-like social robot navigation behaviors can be learned through such a dataset, and point out future research directions and anticipated use cases of our dataset. For other robot navigation datasets which are less relevant to our work compared to SCAND and Ego4D, we refer the readers to Table I in the SCAND paper [7].

III. SENSOR SUITE

We design and make publicly available a data collection device, which is wearable by a human walking in public spaces and provides multi-modal perceptual streams that are commonly available on mobile robot platforms.[2] We also process the raw data to extract human navigation behaviors, i.e., the paths and actions taken by the human demonstrator to navigate through social spaces.

To be specific, our data collection sensor suite is equipped with a 3D LiDAR, a stereo and depth camera with built-in IMU, a microphone array to provide ambient sound, and a 360° camera that offers spherical view of the environment. All the sensors are mounted to a helmet via open-sourced hardware to capture egocentric data of the demonstrator during social navigation. To stream and store real-time social human navigation data, all sensors are connected to a laptop carried by the demonstrator with wired connections (Fig. 1 middle).

[2] https://github.com/RobotiXX/MuSoHu-data-collection

a) 3D LiDAR: As most mobile robots use LiDARs as a reliable sensor to acquire accurate and robust geometric information about the environment, we include a 3D LiDAR to capture such information around the human demonstrator. Considering the different heights of the mounting locations (on robot vs. on our helmet), we use a 3D LiDAR to collect 3D point clouds, which can be converted to 2D scans at different heights if necessary. We choose a Velodyne Puck VLP-16 for our sensor suite, which has a range of 100 meters and generates up to 600,000 points/second, across a 360° horizontal and 30° vertical field of view. The 3D LiDAR is mounted on the top of the helmet to record spatial measurements of the surrounding.

b) Stereo and Depth Camera: RGB cameras provide visual and semantic information of the environment. In addition to the geometric information provided by the LiDAR, semantics also plays a vital role in social navigation interactions. For example, humans use gesture, gaze, and body posture to explicitly or implicitly convey navigational intentions and facilitate interactions. Those behaviors can be used to understand the intentions of other people sharing the same space but are difficult to capture with 3D LiDAR alone. For our sensor suite, we choose Stereolabs ZED 2, a stereo camera with depth sensing and a built-in IMU (see below for more details), considering its compact form factor and efficient power consumption (in contrast to other RGB-D cameras that require a separate power supply, ZED 2 can be efficiently powered by the same USB cable for data transmission). The camera is positioned in the front of the helmet, with the optical axis pointing forward. The wide 120° field of view captures interesting social interactions happening in front of and from the sides of the human demonstrator.

c) IMU: Many mobile robots are also equipped with IMUs to measure linear accelerations and rotational speeds. Therefore, we also utilize the built-in IMU from the ZED 2 camera and record their raw measurements. It is worth to note that due to the difference between walking humans and wheeled or tracked robots that drive, the IMU readings collected in our dataset may be significantly different than those from such types of mobile robots, especially the acceleration along the vertical axis. We posit that to leverage the IMU data in MuSoHu, special techniques such as transfer learning [21] may be necessary.

d) Odometry / Actions: Similar to SCAND, we collect visual-inertia odometry provided by the ZED 2 camera. Such positional odometry provides learning data of navigation path and can be utilized to learn robot global planners. Different than SCAND, in which the robot navigation actions can be directly recorded as teleoperation commands, our data collection hardware does not have access to such actions, i.e., how the human demonstrator walks. Therefore, we extract linear and angular velocities from the positional odometry using the difference between two consecutive odometry frames.

e) 360° Camera: In addition to the forward facing stereo and depth camera, we also collect 360° RGB video to provide better situational awareness of the surrounding and include all possible sensory information that can be provided by active pan-tilt cameras onboard many mobile robot platforms. We use a Kodak Pixpro Orbit360 4K VR Camera to collect 360° images. The camera has a very compact form factor with two lenses integrated in one camera body to provide spherical 360° view. Note that due to software limitations the camera's webcam mode does not allow both lenses to stream live video to the laptop, so we save the spherical 360° view from both lenses to an SD card in the camera.

f) Microphone Array: Although not commonly used for navigation tasks, microphones are available on many mobile robot platforms, e.g., for verbal communications. Furthermore, recent research has started to investigate using sound for navigation [22]. Considering the extra information provided by this different perception modality, we also include a microphone array, a Seeed Studio ReSpeaker Mic Array v2.0, to collect ambient sound during social human navigation.

IV. DATASET

The sensor suite described in Sec. III is designed to be easily replicable by any research group and to collect data worldwide. But we collect an initial Multi-modal Social Human navigation dataset (MuSoHu) on the George Mason University campus and in the Washington DC metropolitan area (Fig. 2).[3]

A. Data Collection Procedure

To collect multi-modal, socially compliant, human-level navigation demonstrations to learn future robot navigation, seven human demonstrators wear the sensor suite helmet and navigate to predefined goals in public spaces in a socially compliant manner. We choose navigation scenarios with frequent social interactions in various indoor and outdoor environments at different time periods (e.g., after class or during weekends). The sensor suite's superior portability (i.e., only a helmet and a laptop) also allows us to record portions of MuSoHu in other settings in the Washington DC Metropolitan Area, including Fairfax, Arlington, and Springfield in Virginia and the National Mall in DC. Notably, for a trajectory at a certain location at the same time period, in many cases, we record three trials to capture three navigation contexts, i.e., *casual*, *neutral*, and *rush*, in which walking speed and safety distance from others may vary, in order to encourage different social navigation interactions based on different contexts. We intend such context awareness in MuSoHu to be useful for future studies on context-aware social navigation, e.g., social compliance when facing someone who is about to be late for a class is different than that when facing someone who is taking a casual afternoon stroll in the park.

For each trajectory, all sensor data are collected using the Robot Operating System (ROS) Bag functionality, except the 360° camera, which does not allow data streaming of

[3] https://dataverse.orc.gmu.edu/dataset.xhtml?persistentId=doi:10.13021/orc2020/HZI4LJ

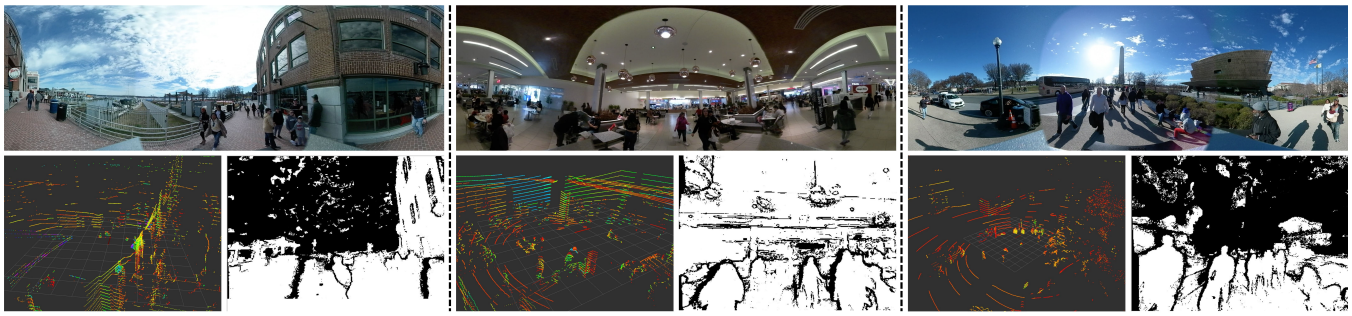

Fig. 2: Three example data frames in Old Town Alexandria, VA, Springfield Towncenter, VA, and National Mall, Washington DC. 360° view (top), 3D LiDAR point cloud (bottom left), and depth image (bottom right) are shown for each data frame.

Fig. 3: Learned Obstacle Avoidance Behavior from MuSoHu.

both built-in cameras to provide spherical 360° view to ROS. Therefore, we store the 360° video on an SD card and provide synchronization using a movie clapboard.

B. Dataset Analyses

1) Labeled Annotations of Social Interactions: MuSoHu includes a list of textual tags for each trajectory that describe the different social interactions that occur along the path. We expand beyond the tags from SCAND and the full list of 17 predefined labels can be found in Table I (with five new tags in bold font).

2) Proof-of-Concept Usage: We use a small subset of MuSoHu data (ten navigation trials) to train a Behavior Cloning policy that maps from raw LiDAR input to linear and angular velocity. The learned policy is deployed on two physical robots, an AgileX Hunter SE (an Ackermann steering wheeled vehicle) and a Unitree Go1 (a quadruped robot), both of which exhibit collision avoidance behavior learned from MuSoHu (Fig. 3).

V. CONCLUSIONS

We present a large-scale, multi-modal, social human navigation dataset, MuSoHu, to allow robots to learn human-like, socially compliant navigation. Our open-sourced design allows our portable sensor suite to be easily replicated and used to collect data in a variety of public spaces worldwide. Such an easy access to a variety of natural social navigation interactions in human-inhabited public spaces in the wild is shown in our preliminary experiments to be useful to learn social robot navigation.

TABLE I: Descriptions of Label Tags Contained in MuSoHu.

Tag	Description	# Tags
Against Traffic	Navigating against oncoming traffic	210
With Traffic	Navigating with oncoming traffic	170
Street Crossing	Crossing across a street	120
Overtaking	Overtaking a person or groups of people	100
Sidewalk	Navigating on a sidewalk	160
Passing Conversational Groups	Navigating past a group of 2 or more people that are talking amongst themselves	94
Blind Corner	Navigating past a corner where the human cannot see the other side	90
Narrow Doorway	Navigating through a doorway where the human opens or waits for others to open the door	45
Crossing Stationary Queue	Walking across a line of people	50
Stairs	Walking up and/or down stairs	30
Vehicle Interaction	Navigating around a vehicle	26
Navigating Through Large Crowds	Navigating among large unstructured crowds	45
Elevator Ride	Navigating to, waiting inside, and exiting an elevator	15
Escalator Ride	Navigating to and riding an escalator	6
Waiting in Line	Waiting in Line to enter congested areas	5
Time: Day	Navigation during day time	150
Time: Night	Navigation during night time	40

REFERENCES

[1] Amazon, "Meet scout," www.aboutamazon.com/news/transportation/meet-scout, 2022.

[2] Dilligent Robotics, "Dilligent robotics," www.diligentrobots.com/, 2022.

[3] The Seattle Times, "Amazon robots scooting along sidewalks? Hold on, Kirkland says," www.seattletimes.com/seattle-news/eastside/for-now-kirkland-says-no-to-small-amazon-robots-scooting-along-streets-and-sidewalks/, 2022.

[4] Bloomberg, "The sidewalk robot resistance begins in san francisco," www.bloomberg.com/news/articles/2017-05-19/the-sidewalk-robot-resistance-begins-in-san-francisco, 2017.

[5] H. Karnan, G. Warnell, X. Xiao, and P. Stone, "Voila: Visual-observation-only imitation learning for autonomous navigation," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 2497–2503.

[6] L. Tai, J. Zhang, M. Liu, and W. Burgard, "Socially compliant navigation through raw depth inputs with generative adversarial imitation learning," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 1111–1117.

[7] H. Karnan, A. Nair, X. Xiao, G. Warnell, S. Pirk, A. Toshev, J. Hart, J. Biswas, and P. Stone, "Socially compliant navigation dataset (scand): A large-scale dataset of demonstrations for social navigation," *IEEE Robotics and Automation Letters*, 2022.

[8] R. Mirsky, X. Xiao, J. Hart, and P. Stone, "Prevention and resolution of conflicts in social navigation–a survey," *arXiv preprint arXiv:2106.12113*, 2021.

[9] X. Xiao, B. Liu, G. Warnell, and P. Stone, "Motion planning and control for mobile robot navigation using machine learning: a survey," *Autonomous Robots*, vol. 46, no. 5, pp. 569–597, 2022.

[10] K. T. Baghaei, A. Payandeh, P. Fayyazsanavi, S. Rahimi, Z. Chen, and S. B. Ramezani, "Deep representation learning: Fundamentals, perspectives, applications, and open challenges," *arXiv preprint arXiv:2211.14732*, 2022.

[11] Y. F. Chen, M. Everett, M. Liu, and J. P. How, "Socially aware motion planning with deep reinforcement learning," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 1343–1350.

[12] Z. Xu, G. Dhamankar, A. Nair, X. Xiao, G. Warnell, B. Liu, Z. Wang, and P. Stone, "Applr: Adaptive planner parameter learning from reinforcement," in *2021 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2021, pp. 6086–6092.

[13] Z. Xu, B. Liu, X. Xiao, A. Nair, and P. Stone, "Benchmarking reinforcement learning techniques for autonomous navigation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023.

[14] W. B. Knox, A. Allievi, H. Banzhaf, F. Schmitt, and P. Stone, "Reward (mis) design for autonomous driving," *Artificial Intelligence*, vol. 316, p. 103829, 2023.

[15] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang *et al.*, "End to end learning for self-driving cars," *arXiv preprint arXiv:1604.07316*, 2016.

[16] M. H. Nazeri and M. Bohlouli, "Exploring reflective limitation of behavior cloning in autonomous vehicles," in *2021 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2021, pp. 1252–1257.

[17] X. Xiao, T. Zhang, K. M. Choromanski, T.-W. E. Lee, A. Francis, J. Varley, S. Tu, S. Singh, P. Xu, F. Xia, S. M. Persson, L. Takayama, R. Frostig, J. Tan, C. Parada, and V. Sindhwani, "Learning model predictive controllers with real-time attention for real-world navigation," in *Conference on robot learning*. PMLR, 2022.

[18] X. Xiao, B. Liu, G. Warnell, J. Fink, and P. Stone, "Appld: Adaptive planner parameter learning from demonstration," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4541–4547, 2020.

[19] H. Kretzschmar, M. Spies, C. Sprunk, and W. Burgard, "Socially compliant mobile robot navigation via inverse reinforcement learning," *The International Journal of Robotics Research*, vol. 35, no. 11, pp. 1289–1307, 2016.

[20] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu *et al.*, "Ego4d: Around the world in 3,000 hours of egocentric video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 995–19 012.

[21] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big data*, vol. 3, no. 1, pp. 1–40, 2016.

[22] Z. Chen, X. Hu, and A. Owens, "Structure from silence: Learning scene structure from ambient sound," *arXiv preprint arXiv:2111.05846*, 2021.