GesNavi: Gesture-guided Outdoor Vision-and-Language Navigation

Aman Jain[12], Teruhisa Misu[3], Kentaro Yamada[2], and Hitomi Yanaka[1]

[1]The University of Tokyo
[2]Honda R&D Co.,Ltd., Tokyo, Japan
[3]Honda Research Institute USA, Inc.
jain-aman@g.ecc.u-tokyo.ac.jp, tmisu@honda-ri.com
kentaro_yamada@jp.honda, hyanaka@is.s.u-tokyo.ac.jp

Abstract

Vision-and-Language Navigation (VLN) task involves navigating mobility using linguistic commands and has application in developing interfaces for autonomous mobility. In reality, natural human communication also encompasses non-verbal cues like hand gestures and gaze. These gesture-guided instructions have been explored in Human-Robot Interaction systems for effective interaction, particularly in object-referring expressions. However, a notable gap exists in tackling gesture-based demonstrative expressions in outdoor VLN task. To address this, we introduce a novel dataset for gesture-guided outdoor VLN instructions with demonstrative expressions, designed with a focus on complex instructions requiring multi-hop reasoning between the multiple input modalities. In addition, our work also includes a comprehensive analysis of the collected data and a comparative evaluation against the existing datasets.

1 Introduction

With the recent successes of autonomous mobilities, there has been an interest in developing interfaces to interact with such systems, leading to the rise of the Vision-and-Language Navigation (VLN) task. However, all the outdoor VLN tasks still consider verbal instructions as the only interface for communicating with the mobility (Vasudevan et al., 2021; Deruyttere et al., 2019). In reality, humans communicate with each other in their daily lives by using non-verbal cues like gestures as well. To allow the freedom of using this intuitive form of communication through gestures, there have been recent efforts to create datasets incorporating pointing gestures as well as an interface for communication in Human-Robot Interaction (HRI) systems (Islam et al., 2022; Chen et al., 2021). However, these datasets are designed for indoor Referring Expression Comprehension (REC) tasks and often consist of simple instructions that do not require intricate reasoning between the pointing gesture and the linguistic instruction. Hence, there is a need for datasets incorporating gesture-guided instructions in outdoor VLN tasks. Such datasets would enable the development of intelligent mobility robots that can be navigated using an intuitive interface of gestural and free-form natural language instructions.

Figure 1: An example from our GesNavi dataset containing (1) a natural language instruction (text box), and (2) a gestured object (green bounding box) that acts as an intermediate anchor for a multi-hop reasoning instruction to navigate toward (3) the target object (red bounding box) indicated by the instruction.

In this work, we tackle a part of the aforementioned challenge by constructing a novel dataset, GesNavi, consisting of instructions with gesture-guided demonstrative expressions for an outdoor VLN task. We capture images from a simulated environment for a crowded urban neighborhood and crowdsource annotations for free-form linguistic instructions to navigate the mobility. Having a simulated environment allowed us to automatically capture images and extract ground truth object labels, resulting in a fast and scalable data collection process. These instructions are also accompanied by a gestured object specified using a bounding box, which is expected to be obtained using a non-verbal gestural modality in practice. The gestured

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, pages 290–295

object is referred through demonstratives such as *this cone*, *that tall building*, etc., and is used as an intermediate anchor for navigating to the target object. This results in instructions requiring multi-hop reasoning to be performed to locate the target object by considering (1) the relationship between the demonstrative and the scene (detecting an intermediate object) and (2) the relationship between the intermediate object and the target object, as described in Figure 1. Furthermore, we also supplement our data with images from another perspective of the same scene, as shown in Figure 2, to facilitate the development of systems that can comprehend multimodal instructions in differing perspectives. This is particularly useful in applications such as service robots where the user is accompanying the mobility robot from outside with a different visual perspective than the mobility's camera.

The key contributions of this work are listed below:

- We introduce a challenging novel task of gesture-guided interaction with mobility for outdoor VLN, with significant practical application.
- We collect a dataset, GesNavi, consisting of natural instructions accompanied by an intermediate gestured object to navigate the mobility toward a target object.
- We analyze our dataset and compare it against the existing datasets for both outdoor VLN and gesture-guided referring expressions.

2 Related Works

2.1 Gestures in Human-Robot Interaction

The field of Human-Robot Interaction (HRI) has extensively explored integrating gestures as an additional input modality (Bolt, 1980; Ende et al., 2011; Wu et al., 2021; Sato et al., 2007; Hu et al., 2018). Jain et al. (2023) utilized a VR setup for collecting non-verbal gestural data in a simulated environment but faced challenges in scaling to a larger dataset. CAESAR (Islam et al., 2022) and YouRefIt (Chen et al., 2021) are two major datasets consisting of embodied gesture-aided expressions for the Referring Expression Comprehension (REC) task.

CAESAR, though comprehensive, is based on a fully simulated environment with auto-generated instructions and pointing gestures, lacking natural variations in human utterance. In contrast, YouRefIt features a real-world setting with natural language instructions and pointing gestures. However, it incorporates pointing gestures as optional information, leading to simpler instructions, as evidenced by the low average instruction length in Table 1. To address this limitation, our work extends these datasets to encompass more complex and free-form natural instructions, challenging multi-hop reasoning.

2.2 VLN Tasks

Our study focuses on an outdoor Vision-and-Language Navigation (VLN) task, involving a mobility agent receiving navigational instructions to locate a target position. VLN datasets encompass both indoor (Anderson et al., 2018) and outdoor (Vasudevan et al., 2021; Deruyttere et al., 2019) environments. Previous outdoor VLN approaches, such as those in Hermann et al. (2020); Chen et al. (2019), provided detailed step-by-step directional commands for mobility. Tasks like Talk2Car (Deruyttere et al., 2019) evolved this by incorporating more natural and free-form verbal instructions for autonomous vehicle control.

However, existing VLN tasks exclusively rely on verbal instructions, overlooking demonstrative cues prevalent in human speech. Our work addresses this gap by exploring the incorporation of gesture-guided instructions in outdoor VLN tasks.

3 Dataset

The data collection procedure for our task is divided into two steps. The first step consists of collecting images that capture a wide range of outdoor scenes. Then, we collect annotations for the gesture-guided instructions on these images. We will describe each of these steps in detail in the following subsections.

3.1 Collecting Images

To create a diverse image dataset for our task, we used a simulated environment replicating crowded streets in a dense Tokyo neighborhood ($70{,}000\,\mathrm{m}^2$) on the Airsim platform (Shah et al., 2017), Unreal engine. Beyond the urban elements, we strategically placed various objects (vehicles, pedestrians, trees, cones, vending machines) in diverse locations.

Using a simulator provided three key benefits in our study: (i) random image sampling from any coordinate, (ii) automated extraction of ground truth

Datasets	Task	G	N	R	P	Total samples	Mean instruction length (words)
Talk2Car (Deruyttere et al., 2019)	Outdoor VLN	✗	✓	✓	✗	11,959	11.0
CAESAR-XL (Islam et al., 2022)	REC	✓	✗	✗	✓	1,367,305	5.3
YouRefIt (Chen et al., 2021)	REC	✓	✓	✓	✗	4,195	3.7
GesNavi (Ours)	Outdoor VLN	✓	✓	✗	✓	3,100	13.1

Table 1: Comparison of datasets relevant to this work. G, N, R and P denote the use of gestures, non-templatized natural instructions, use of real-world images (versus simulated images) and multiple perspective images, respectively. The mean instruction length is used here to compare the instruction sentence complexities in the respective datasets.

Figure 2: Our dataset consists of a supplementary image for each scene to mimic a more challenging situation where the user is outside the mobility and their visual perspective (left image) is slightly different than the mobility's camera (right image).

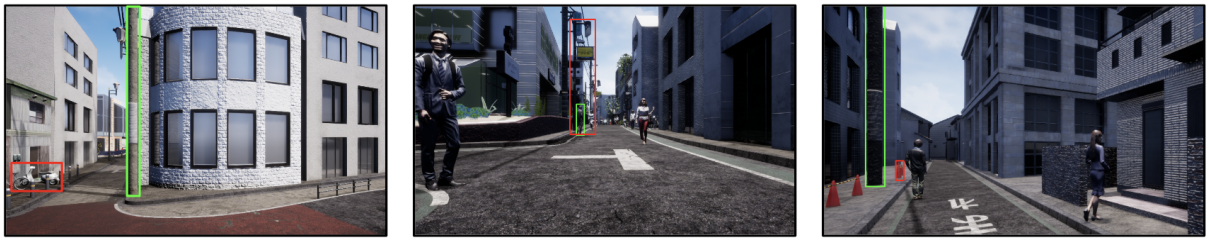
object labels, and (iii) algorithmic computation of 2D/3D bounding boxes using mesh coordinates for rendering. While simulated images may lack the natural features and imperfections found in real-world objects and scenes, they offer a controlled method for generating data. The ability to control the diversity in objects and scenes facilitates the creation of a challenging multi-hop reasoning task.

To capture the visual data automatically, we developed a function navigating a virtual camera along simulator roads. Varied parameters captured data under different lighting conditions (morning, afternoon, evening). Each captured data includes two images taken from a few meters apart with a relative angle of 45 degrees — representing slightly differing perspectives of the same scene, as depicted in Figure 2. One image is used for annotating navigational instructions, while the other serves as a supplementary image for another visual perspective to facilitate research for comprehending outdoor VLN instructions in applications like service robots. The captured data also includes depth maps, ground truth object classes, and positions relative to the mobility robot. A human annotator monitored the image capture process to ensure diverse scenes with minimal duplicates and unnatural scenes.

3.2 Annotating gesture-guided linguistic instructions

Upon acquiring all the images and their associated ground truth data, the next phase involves gathering gesture-guided linguistic instructions for an outdoor VLN task. Obtaining hand gesture annotations in a simulated environment typically involves the use of a virtual reality (VR) setup, as demonstrated by Jain et al. (2023). This setup utilizes a VR headset and hand controllers to capture head and hand movements while performing pointing gestures. However, it is crucial to acknowledge that such a configuration is not only expensive but also time-intensive. To address these challenges, we have adopted a more straightforward approach of annotating the gestured object by enclosing it within a bounding box. In practice, we expect that the gestured object can be determined by leveraging the existing research in recognizing non-verbal cues, such as gestures, from visual input (Nickel and Stiefelhagen, 2003; Stiefelhagen et al., 2004). While this simplification results in the loss of some raw features related to hand motion during the pointing gesture, it enables us to collect a larger dataset for this task.

To crowd-source annotations, we used Amazon Mechanical Turk (MTurk). Our guidelines instructed annotators to assume human-like mobility controlled by linguistic and gestural instructions. The annotators chose any target object of their liking and formulated navigational instructions to guide the mobility robot to that target which were collected in the form of text. In addition, the annotators were asked to imagine the use of hand gestures like pointing, annotate the gestured object with a bounding box, and use it as an *intermediate anchor* to create a multi-hop instruction based on its relation to the target object. Finally, annotators were required to label the target object with a tightly drawn bounding box.



My scooter is across from that pole, please take me there.

Find and stop at the pole with the yellow sign. It's to the left of that man in a grey jacket and khaki pants.

After you pass this tall pole, stop at the first construction cone.

Figure 3: Example gesture-guided instructions in our GesNavi dataset, with a wide variety of syntactic and semantic structures. The gestured and target objects are annotated with green and red bounding boxes, respectively.

Target Objects		**Gestured Objects**	
Object	Frequency	Object	Frequency
Obstacle	334	Person	1350
Person	330	Building	191
Car	276	Obstacle	160
Bicycle	226	Car	114
Dispenser	146	Pole	95

Table 2: Top five most frequently used target and gestured objects in our dataset

Multiple tests were conducted to refine guidelines, throughout emphasizing on crafting free-form natural instructions incorporating gesture demonstratives and necessitating multi-hop reasoning. Expert MTurk workers with native English skills and track record in annotation tasks were invited to a screening test. 25 workers who correctly performed at least four out of five annotations in the screening test were selected and received individual feedback to ensure their complete understanding of the task for good annotation quality. All images were published in small batches, with simultaneous batch reviews and feedback to maintain the desired annotation quality. Each image costed $0.75 and took the workers an average of around 10.7 minutes per annotation.

4 Dataset Analyses

We collected a total of 3,100 gesture-guided VLN instructions on outdoor scenes. The instructions in our dataset comprise a vocabulary of 924 words. Since our approach did not rely on templates or impose constraints on linguistic instructions, we were able to capture the commonly used natural language instructions in navigational scenarios, as exemplified in Figure 3. The examples illustrate the wide variety of syntactic and semantic structures present in our instructions.

Our instructions vary in length from 6 to 34 words, with an average length of 13.1 words. This average length is comparable to the text-only outdoor VLN dataset, Talk2Car (Deruyttere et al., 2019), and significantly larger than other gesture-guided HRI datasets like CAESAR (Islam et al., 2022) and YouRefIt (Chen et al., 2021), which have average expression lengths of just 5.3 words and 3.7 words, respectively. The longer expressions in our dataset reflect the emphasis on free-form natural instructions requiring complex multi-hop reasoning, in contrast to these earlier works.

From the annotated bounding boxes of the gestured and target objects, we determined their labeled class by identifying the ground truth object with the highest Intersection over Union (IoU) overlap. The five most frequent objects used for gesturing and as the target objects are summarized in Table 2. Notably, a significant proportion of gestured objects are pedestrians, likely due to their prevalence in crowded street scenes, making them a convenient intermediate object for conveying instructions about the intended target object. It is also worth mentioning that the average distance between the camera and the target objects selected by annotators is 12.2 meters, which is around 6% more than the average distance of 11.5 meters for the gestured objects. Moreover, the average bounding box size for gestured objects is around 25% larger than the target objects. This observation suggests a general human tendency to use closer and larger objects for non-verbal gestural cues to navigate to more distant and smaller target objects.

5 Conclusion

This work introduces a novel dataset, GesNavi, designed for gesture-guided multimodal interaction

with mobility in the context of an outdoor VLN task. Moreover, in contrast to the prior efforts in the related field of gesture-guided REC tasks, our dataset specifically emphasizes natural free-form instructions that require complex multi-hop reasoning. This is evident from the significantly longer expressions in our dataset compared to the previous works.

In the future, this dataset can be expanded to also include the general case of using gestures to refer to multiple objects or a group of objects, rather than a single object in the current setup. Another future work includes developing methods to tackle this task, including the current state-of-the-art multimodal architectures for VLN (Yan et al., 2023; Kamath et al., 2021), and evaluating their performances on our GesNavi dataset. It is also worth evaluating how the multimodal LLMs (OpenAI, 2023; Team et al., 2023) perform in our task. It is a particularly challenging task and will require designing models that can effectively combine multimodal information and perform multi-hop reasoning to find the target object.

6 Limitations

While this work represents a significant stride in developing a valuable resource for gesture-guided outdoor VLN task, certain limitations in its design deserve consideration. Firstly, the dataset relies on simulated environment images, potentially limiting real-world applicability due to the absence of genuine environmental complexity and randomness. Secondly, assuming a single object referenced through gestures may overlook the broader potential of gestural instructions for groups or multiple objects within a single instruction. These limitations acknowledge the current scope and highlight opportunities for future enhancements in this field.

Acknowledgments

This work was partially supported by JST, PRESTO Grant Number JPMJPR21C8, Japan.

References

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian D. Reid, Stephen Gould, and Anton van den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3674–3683. IEEE Computer Society.

Richard A. Bolt. 1980. "put-that-there": Voice and gesture at the graphics interface. In *Proceedings of the 7th annual conference on Computer graphics and interactive techniques*, SIGGRAPH '80, pages 262–270, New York, NY, USA. ACM.

Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. 2019. TOUCHDOWN: natural language navigation and spatial reasoning in visual street environments. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 12538–12547. Computer Vision Foundation / IEEE.

Yixin Chen, Qing Li, Deqian Kong, Yik Lun Kei, Song-Chun Zhu, Tao Gao, Yixin Zhu, and Siyuan Huang. 2021. Yourefit: Embodied reference understanding with language and gesture. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1365–1375. IEEE.

Thierry Deruyttere, Simon Vandenhende, Dusan Grujicic, Luc Van Gool, and Marie-Francine Moens. 2019. Talk2Car: Taking control of your self-driving car. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2088–2098, Hong Kong, China. Association for Computational Linguistics.

Tobias Ende, Sami Haddadin, Sven Parusel, Tilo Wüsthoff, Marc Hassenzahl, and Alin Albu-Schäffer. 2011. A human-centered approach to robot gesture based communication within collaborative working processes. In *Proceedings of the 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3367–3374, San Francisco, California. IEEE.

Karl Moritz Hermann, Mateusz Malinowski, Piotr Mirowski, Andras Banki-Horvath, Keith Anderson, and Raia Hadsell. 2020. Learning to follow directions in street view. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 11773–11781. AAAI Press.

Jun Hu, Zhongyu Jiang, Xionghao Ding, Taijiang Mu, and Peter Hall. 2018. Vgpn: Voice-guided pointing robot navigation for humans. In *Proceedings of the 2018 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 1107–1112, Kuala Lumpur, Malaysia. IEEE.

Md Mofijul Islam, Reza Manuel Mirzaiee, Alexi Gladstone, Haley N Green, and Tariq Iqbal. 2022. CAESAR: An embodied simulator for generating multimodal referring expression datasets. In *Proceedings of the Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, pages 21001–21015, New Orleans, Los Angeles. Curran Associates, Inc.

Aman Jain, Anirudh Reddy Kondapally, and Kentaro Yamadaand Hitomi Yanaka. 2023. A neuro-symbolic approach for multimodal reference expression comprehension. In *Proceedings of the 37th Annual Conference of the Japanese Society for Artificial Intelligence*, Kumamoto, Japan. Japanese Society for Artificial Intelligence.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. 2021. MDETR - modulated detection for end-to-end multi-modal understanding. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1760–1770. IEEE.

Kai Nickel and Rainer Stiefelhagen. 2003. Pointing gesture recognition based on 3d-tracking of face, hands and head orientation. In *Proceedings of the 5th International Conference on Multimodal Interfaces*, ICMI '03, page 140–146, New York, NY, USA. Association for Computing Machinery.

OpenAI. 2023. Gpt-4v(ision) technical work and authors.

Eri Sato, Toru Yamaguchi, and Fumio Harashima. 2007. Natural interface using pointing behavior for human–robot gestural interaction. *IEEE Transactions on Industrial Electronics*, 54(2):1105–1112.

Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. 2017. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. volume abs/1705.05065.

R. Stiefelhagen, C. Fugen, R. Gieselmann, H. Holzapfel, K. Nickel, and A. Waibel. 2004. Natural human-robot interaction using speech, head pose and gestures. In *Proceedings of the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, volume 3, pages 2422–2427, Miyagi, Japan. IEEE.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *ArXiv preprint*, abs/2312.11805.

Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool. 2021. Talk2nav: Long-range vision-and-language navigation with dual attention and spatial memory. *International Journal of Computer Vision*, 129:246–266.

Qi Wu, Cheng-Ju Wu, Yixin Zhu, and Jungseock Joo. 2021. Communicative learning with natural gestures for embodied navigation agents with human-in-the-scene. In *Proceedings of the 2021 International Conference on Intelligent Robotics and Systems (IROS)*, pages 4095–4102, Prague, Czech Republic. IEEE.

Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Zehuan Yuan, Ping Luo, and Huchuan Lu. 2023. Universal instance perception as object discovery and retrieval. In *CVPR*.