Language and Gesture in Virtual Reality: Is a Gesture Worth 1000 Words?

Padraig Higgins[2, 1], **Cory J. Hayes**[1], **Stephanie M. Lukin**[1], **Cynthia Matuszek**[2]

[1]DEVCOM Army Research Laboratory, Adelphi, MD 20783, USA
[2]University of Maryland, Baltimore County, Baltimore MD 21250, USA
phiggins1@umbc.edu, cory.j.hayes4.civ@army.mil, stephanie.m.lukin.civ@army.mil, cmat@umbc.edu

Abstract

Robots are increasingly incorporating multimodal information and human signals to resolve ambiguity in embodied human-robot interaction. Harnessing signals such as gestures may expedite robot exploration in large, outdoor urban environments for supporting disaster recovery operations, where speech may be unclear due to noise or the challenges of a dynamic and dangerous environment. Despite this potential, capturing human gesture and properly grounding it to crowded, outdoor environments remains a challenge. In this work, we propose a method to model human gesture and ground it to spoken language instructions given to a robot for execution in large spaces. We implement our method in virtual reality to develop a workflow for faster future data collection. We present a series of proposed experiments that compare a language-only baseline to our proposed language supplemented by gesture approach, and discuss how our approach has the potential to reinforce the human's intent and detect discrepancies in gesture and spoken instructions in these large and crowded environments.

Introduction

As robots become more ubiquitous, they have the potential to be intelligent teammates for cooperative tasks with humans. When interacting with these intelligent systems, natural language provides an intuitive interaction paradigm between humans and robots; however, it can be challenging to use *only* natural language to describe objects or locations of interest under certain conditions. In environments with a number of objects with similar physical attributes and affordances, for example, differentiating specific objects can require precise and highly specified language (Kennington and Natouf 2022). This may lead to an increased cognitive load when giving commands to a robot as they are forced to over-specify or use additional dialogue to resolve the ambiguity, which takes time in critical scenarios, such as disaster response. Furthermore, it requires the robot to understand how to properly ground and resolve the ambiguity in the first place, and enough data to train such systems, which can require extensive data collections that may not account for potentially dynamic and dangerous environments.

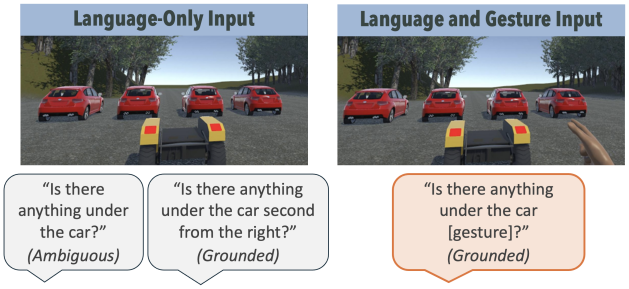

Figure 1: Interpreting human gesture alongside spoken commands can lead to grounded instruction interpretation in comparison to language-only instructions, which may be longer or left ambiguous.

Instead of relying only on language, we explore how incorporating gestures can potentially make interactions in challenging environments more intuitive and streamlined, avoiding overly specific language and long dialogues where the person has to repeatedly clarify their intent (Fig. 1). Deictic gestures are a common type of nonverbal communication (Cooperrider et al. 2021), the most common being a pointing gesture to focus attention on an object or location in a shared space. These gestures are frequently used to convey spatial information (Diessel 2006), and their use has been shown to make communicating spatial relations more natural for the speaker (Rauscher, Krauss, and Chen 1996). Deictic pointing gestures can only provide spatial information that an object being referenced is located in a certain direction relative to the gesturer, and requires complementary natural language to ground the full human intent.

Prior work in human-robot interaction has studied deictic gesture with spoken language in small-scale, typically tabletop environments. Srimal, Muthugala, and Jayasekara (2017) utilizes fuzzy logic to map language and gesture to map onto a known tabletop environment, while Matuszek et al. (2014) utilizes a data-driven approach where a classifier is trained to map hand movements to a referred object; both approaches require some prior training or knowledge. Lin et al. (2023) and Ekrekli et al. (2023) can work in a zero-shot setting, but the target of a pointing gesture is assumed to be the closest object through a geometric heuristic. In these close quarters, tabletop settings, it may be easier for people to indicate objects with greater precision than at greater

distances where there is more room for error.

However, extending human-robot interaction to larger outdoor environments presents several challenges to both language and gesture. Cameras in the environment may not include both the gesturer and target in frame, requiring an alternate method to reconcile the gesture signal with the objects in the space. Field robots may also not have access to models requiring high computational resources, which can limit the implementation. While recent advances in Large Language Models and Vision Language Models have shown the ability to recognize objects based on natural language descriptions, they may not be a suitable out-of-the-box solution. These models can struggle with spatial relations (Chen et al. 2025), and furthermore, be potentially difficult to use on robots in the field with unreliable communications and limited power budgets. Additionally, large outdoor environments may be dynamic with moving objects and numerous obstacles requiring disambiguation.

The computation costs and environment dynamics make running user studies challenging, and would require a fully-working robot in potentially dangerous scenarios with still untested algorithms. To mitigate this, we use Virtual Reality as an initial development environment, allowing for quicker data collection and evaluation of our developed approaches, and mitigate the transition to real-world robotics platforms and systems using Robot Operating System (ROS). We discuss our approach and implementation of the robot system and our novel gesture tracking in the following Section. From this foundation, we outline how we plan to combine the captured language and gesture signals for the robot system. We then discuss our planned experimental scenarios and evaluation, which seek to investigate whether the results from small-scale experiments still hold in large outdoor settings. We hypothesize that incorporating gesture allows for faster communication in these settings, including faster task completion, fewer dialogue turns, and fewer words required. Finally, we conclude with a brief discussion of the predicted feasibility of our approach to real-world settings, based on the state of the art of human-robot communication, particularly in the domain of Urban Search and Rescue (USAR).

Gesture and Language Pipeline

As a motivating vignette, we consider a mixed human-robot team engaged in search and navigation tasks. The role of the robot teammate is to explore and evaluate a static environment based on spoken instructions from a human team leader. Instructions from the human teammate might refer to objects in the environment (e.g., "Is there anything under the awning there"[points]) or actions to be taken (e.g., "Open the green door"). To establish common ground with minimal friction, the robot will attempt to understand deictic gestures and referential language, seeking clarification only in cases where confidence in its understanding is low. Our approach modifies an existing Natural Language Processing pipeline and an integrated ROS robot autonomy stack that includes a custom world model. Fig. 2 shows a modular abstraction of the pipeline with our contribution of the 'Gesture Recognition' module as a simultaneous signal for interpreting human intent. We provide a description of our selected implementation of the Speech Recognition, Language Understanding, Dialogue Management, World Model, and our novel Gesture Recognition components (green boxes in Fig. 2). As our primary goal is the preliminary integration of language and gesture signals, we defer for future work the subsequent integration of the robot navigation and behaviors. We assume, for now, that once a well-formed spoken command and gesture is issued, interpreted, and grounded in the world model, the action will be carried out by a robot behavior module.

Language

Speech recognition in our pipeline is performed by linking three ROS nodes written in Python. The first is a push-to-talk interface, where a user inputs a simultaneous combination of keypresses to toggle audio capture. The second node creates a GStreamer[1] pipeline to capture raw microphone audio with timestamps to assist in post-interaction analysis. It then creates an intermediary audio representation for the final process: Automated Speech Recognition (ASR). We utilize OpenAI's Whisper ASR library (Radford et al. 2023), which transcribes the audio and publishes it as text over a ROS topic. We use the default "openai/whisper-medium.en" pre-trained model for transcription, with future plans to incorporate fine-tuned models that account for environment noise, as explored in Ojha, Gervits, and Espy-Wilson (2025).

Language Understanding Our pipeline affords a semantic understanding of spoken commands to inform the robot of the user's desired behavior. In this way, a user can issue syntactically diverse instructions in a way most comfortable to them, while the pipeline translates them to a specific, well-defined format for robot execution. To achieve this, we leverage Abstract Meaning Representation (AMR), a semantic representation language where natural language sentences are converted to labeled graphs to better represent underlying sentence meaning (Banarescu et al. 2013), e.g., the spoken commands "turn left 90 degrees" and "make a left turn" are normalized to the AMR concept "turn-01." This sentence-to-graph conversion (and vice-versa) is performed by the spaCy parser (Honnibal et al. 2020) through its open-source Python library. The normalized AMR form of the spoken command is passed to a command classifier to match the AMR graph to the closest class of predetermined robot behaviors that can be executed. For example, spoken commands that map to the "turn-01" AMR concept further map to a "turn" robot action. Information required for the action, such as object name, waypoint, or distance to move/rotate, is parsed from the AMR graph.

Dialogue Management The Dialogue Management (DM) node acts as a gatekeeper to handle system communication between the user and the pipeline components that dictate a robot's autonomous behavior. For example, if the spoken command lacks required information and prevents the robot from executing the behavior, e.g., "rotate" is not given a number of degrees or direction to turn, the DM will prompt the user to provide more information, e.g., "How far should

[1]https://gstreamer.freedesktop.org

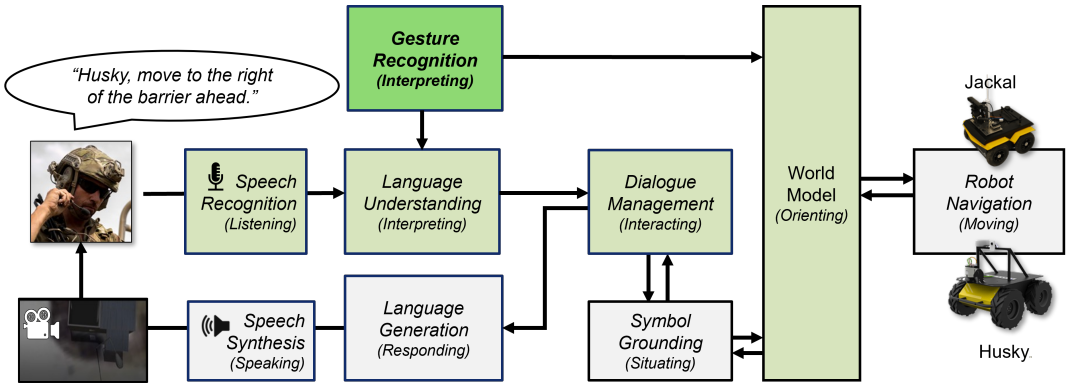

Figure 2: Abstracted representation of our Language and Gesture pipeline for cooperative robot exploration. We introduce our novel 'Gesture Recognition' component for improved language understanding, which in turn may affect the speed and accuracy of the robot planning, execution, and response. Figure modified from the pipeline presented in Bonial et al. (2024).

I turn?" Our DM implementation is based on Bonial et al. (2024)'s statistical matching for command responses based on the SCOUT dataset of human-robot dialogue (Lukin et al. 2024). Commands with all required information are forwarded to robot behavior modules, as well as action updates.

World Model

As the robot explores the environment, it builds a world model of its surroundings. During navigation, the robot utilizes YOLOv8[2] to detect objects, including labels and bounding boxes. Visual and text CLIP embeddings (Radford et al. 2021) are extracted from these labels and cropped images, and added to the world model. For this work, the world model is structured as an unsorted list of unique IDs, labels, positions, and CLIP embeddings of the detected object.

Gesture

We support the interpretation of deictic pointing gestures through 3D human pose estimation to estimate where a person is pointing. We first extract 3D object positions by combining depth images from RGB-D cameras with YOLO for the detection of 2D image keypoints. Then, the shoulder, elbow, and wrist keypoints are used to create a ray approximating the direction that each arm is pointing. The angular distance from this ray to every detected object is then calculated, and objects exceeding a threshold distance are culled. This threshold is initially set to 30 degrees and will be a parameter to be updated from data collected experimentally, as described later. A Bayesian approach similar to Li et al. (2021) is used to mitigate uncertainty in both the possible positions of the detected objects and the direction of the pointing gesture. We determine the probability that an object is being pointed to given a pose

$$P(target|pose) = \frac{P(pose|target)P(target)}{P(pose)}$$

where the probability $P(target|pose)$ is approximated as a Gaussian distribution, as objects with less angular distance are more likely to be the target.

[2]https://yolov8.com/

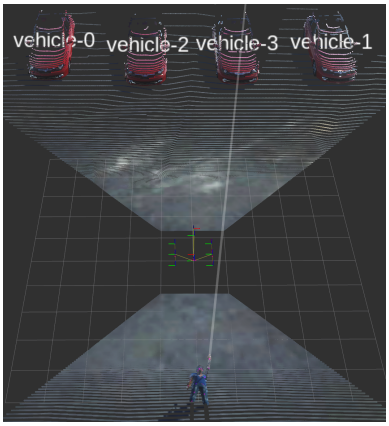

Figure 3: Visualization of the point clouds from the front and rear RGBD navigation cameras, with the detected object locations and pointing ray overlayed.

	D (deg)	P
vehicle-3	16.7	0.355
vehicle-1	19.4	0.334
vehicle-2	22.3	0.311
vehicle-0	31.9	0.0

Table 1: The angular distance (degrees) between the pointing ray and each detected object, and the estimated probability that the person was pointing at that object.

This probability is used to generate 'interest', similar to Li et al. (2021). Interest is calculated as $P(target|pose) * dt$, where dt is the time elapsed between two pointing poses. Over the course of an utterance, if sufficient interest accumulates, the human is likely to be pointing at an object. Fig. 3 and Table 1 show a sample single instance of a person's pose. Here, the person is pointing at "vehicle-2". "Vehicle-0" would be assigned no interest as it is over 30°, but would not give significantly more interest to "vehicle-3" compared to "vehicle-1" and "vehicle-2".

Combining Language and Gesture Signals (Proposed) To combine a well-formed spoken command with an inter-

preted gesture, we propose the following two-phase process. When a spoken utterance begins, interest accumulates for all objects. If any object has sufficient interest (we propose > 0.3 seconds) during the speaking period, we posit that the person has pointed at or otherwise indicated an object. If not, we rely solely on the spoken language.

When sufficient interest from pointing is detected, we will leverage the gesture signal to cull objects without sufficient interest; if someone is indicating in a direction, objects far from the ray are likely not of interest. The gesture then serves as a reinforcing signal. When attempting to ground the spoken instruction to the objects in the world model, objects with sufficient interest will be used, leaving objects with lower scores to be considered if no matches are found. Additionally, we will explore how gesture will be used to detect discrepancies, e.g., pointing to a car on the right while speaking the command "Go to the car on the left." A mismatch in the detected gesture object with the verbal disambiguation will trigger the DM to request further clarification.

Experiment and Evaluation Plans

With our gesture and language understanding components nearly completed, we outline our plans for evaluation. Pilot user studies will validate the pipeline, gather data to tune parameters, and compare the pose of the virtual avatar to the participants' real-world pose. Beyond this, we envison conducting real-world experiments to evaluate both the simulation and our approach to joining language and gesture.

User Study Scenarios Virtual Reality affords a clean environment that minimizes some real-world challenges of skeleton tracking and gesture recognition. We use this to our benefit to focus on the combined effects of speech and gesture for efficient autonomous robot navigation. To build our simulated virtual environment, we utilize the Unity game engine and SteamVR, which connects to the Robots in Virtual Reality system (Higgins et al. 2021) to interface with ROS. We use the Meta Quest 2 VR headsets, which allow for more natural tracking using the Quest's camera-based inside-out tracking, capturing head pose and hand pose, including fingers. We create a human avatar in the scene and rig it to the tracking information received from the Quest headset, with the avatar's hand and fingers designed to track the positions of the participant's hand and fingers.

We are designing a between-subjects study where we vary whether the robot is capable of interpreting language only, or both language and gesture. The goal is to vary the levels of reliance on the gesture signal and the difficulty of interpretation. We hypothesize that this increased load would lead people to more frequently use gestures, and that the incorporation of gestures would allow for better performance, lower cognitive load, and a more pleasant interaction.

In one set of experiments, we are designing scenarios that alter the complexity and potential load on the person between low complexity (6 objects, all in view, spread over $90°$), medium complexity (12 objects, spread over $180°$), and high complexity (24 objects, spread over $270°$). In a second set of experiments, we will allow for the spoken commands to specify locations, rather than only objects, and to modify the robot's behaviors and actions. Beyond this, we will seek to better understand when and how someone is gesturing within an instruction, e.g., (Nakagawa et al. 2024), within dynamic scenarios where the person and robots are moving around, as might be expected in disaster response.

Measures We will measure the effect of utilizing gestures alongside spoken language on the system performance and the user experience. In measuring the *system performance*, we will calculate the accuracy in selecting the correct target initially, how long it takes the system to select the correct target, the number of dialogue turns and words used, and how long the system takes to make a selection from the end of an utterance to a target selection. We will compare against a language-only baseline and on different robot systems to determine the computational and speed trade-offs.

In measuring the *user experience*, we will use subjective measures for perceived workload, e.g., NASA-TLX (Hart and Staveland 1988), and how intuitive, pleasant, and natural the user found the experience. Objective measures, such as how long it takes participants to give a command or response, will also be recorded, i.e., how long it takes for the human to start giving a clarifying statement after the robot makes an incorrect choice.

We will first verify in simulation that the system performance and user experience meet the expectations of the scenario. As the Language and Gesture pipeline matures, we will endeavour to address the challenges required in transitioning from simulation to a real-world, physical platform. In addition to bridging the gap in Sim-to-Real, we intend to explore how our pipeline could someday support first responders under time and resource constraints. In a recent focus group report of emergency responders using robotic technology (Mackanin et al. 2025), the top priorities included Maneuverability and Autonomous Capabilities, with User Interface a higher priority, and audible, two-way communication rated as only "somewhat important". We intend to explore the impact of our completed system, which would leverage both text and text-to-speech for UI, and add to the limited human-robot communication research for language-guided exploration (Howard et al. 2022; Tellex et al. 2020), particularly in USAR, where communication concerns are primarily inter-robot (Gielis, Shankar, and Prorok 2022).

Conclusion

Gestures are a natural component of human-human communication that can enhance the meaning of spoken language. We seek to leverage its benefits by focusing on deictic gestures to reduce ambiguity when grounding spoken language to objects and locations within an environment. We have presented our work in progress for a gesture and language pipeline for mixed human-robot teams in an exploration task. We have outlined our future experiments and evaluations, and plans to advance and combine these human signals in increasingly complex scenarios.

Acknowledgments

This work was sponsored by the National Science Foundation, award numbers 2435593 and 2346667

References

Banarescu, L.; Bonial, C.; Cai, S.; Georgescu, M.; Griffitt, K.; Hermjakob, U.; Knight, K.; Koehn, P.; Palmer, M.; and Schneider, N. 2013. Abstract Meaning Representation for Sembanking. In *7th Linguistic Annotation Workshop and Interoperability with Discourse*.

Bonial, C.; Lukin, S. M.; Abrams, M.; Baker, A.; Donatelli, L.; Foots, A.; Hayes, C. J.; Henry, C.; Hudson, T.; Marge, M.; et al. 2024. Human–robot dialogue annotation for multimodal common ground. *Language Resources and Evaluation*, 1–51.

Chen, S.; Zhu, T.; Zhou, R.; Zhang, J.; Gao, S.; Niebles, J. C.; Geva, M.; He, J.; Wu, J.; and Li, M. 2025. Why Is Spatial Reasoning Hard for VLMs? An Attention Mechanism Perspective on Focus Areas. In *Forty-second International Conference on Machine Learning*.

Cooperrider, K.; Fenlon, J.; Keane, J.; Brentari, D.; and Goldin-Meadow, S. 2021. How Pointing is Integrated into Language: Evidence From Speakers and Signers. *Frontiers in Communication*.

Diessel, H. 2006. Demonstratives, Joint Attention, and the Emergence of Grammar. *Cognitive Linguistics*.

Ekrekli, A.; Angleraud, A.; Sharma, G.; and Pieters, R. 2023. Co-speech Gestures for Human-Robot Collaboration. In *IEEE International Conference on Robotics Computing*.

Gielis, J.; Shankar, A.; and Prorok, A. 2022. A Critical Review of Communications in Multi-Robot Systems. *Current Robotics Reports*.

Hart, S. G.; and Staveland, L. E. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Advances in psychology*, 52: 139–183.

Higgins, P.; Kebe, G. Y.; Berlier, A.; Darvish, K.; Engel, D.; Ferraro, F.; and Matuszek, C. 2021. Towards making virtual human-robot interaction a reality. In *Proc. of the 3rd International Workshop on Virtual, Augmented, and Mixed-Reality for Human-Robot Interactions (VAM-HRI)*.

Honnibal, M.; Montani, I.; Landeghem, S. V.; and Boyd, A. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Howard, T. M.; Stump, E.; Fink, J.; Arkin, J.; Paul, R.; Park, D.; Roy, S.; Barber, D.; Bendell, R.; Schmeckpeper, K.; Tian, J.; Oh, J.; Wigness, M.; Quang, L.; Rothrock, B.; Nash, J.; Walter, M. R.; Jentsch, F.; and Roy, N. 2022. An Intelligence Architecture for Grounded Language Communication with Field Robots. *Journal of Field Robotics*.

Kennington, C.; and Natouf, O. 2022. The Symbol Grounding Problem Re-framed as Concreteness-Abstractness Learned through Spoken Interaction. In *26th Workshop on the Semantics and Pragmatics of Dialogue*.

Li, Z.; Zhao, M.; Wang, Y.; Rashidian, S.; Baig, F.; Liu, R.; Liu, W.; Beaudouin-Lafon, M.; Ellison, B.; Wang, F.; Ramakrishnan, I.; and Bi, X. 2021. BayesGaze: A Bayesian Approach to Eye-Gaze Based Target Selection. In *Graphics Interface 2021*.

Lin, L.-H.; Cui, Y.; Hao, Y.; Xia, F.; and Sadigh, D. 2023. Gesture-Informed Robot Assistance via Foundation Models. In *7th Annual Conference on Robot Learning*.

Lukin, S.; Bonial, C.; Marge, M.; Hudson, T. A.; Hayes, C.; Pollard, K.; Baker, A.; Foots, A. N.; Artstein, R.; Gervits, F.; et al. 2024. SCOUT: A Situated and Multi-Modal Human-Robot Dialogue Corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 14445–14458.

Mackanin, T.; Posner, H.; Crouse, U.; Ellis, J.; Ortega, S.; and Bartholomew, R. A. 2025. Multifunctional Unmanned Ground Vehicles for Emergency Response. Technical report, National Urban Security Technology Laboratory.

Matuszek, C.; Bo, L.; Zettlemoyer, L.; and Fox, D. 2014. Learning from Unscripted Deictic Gesture and Language for Human-Robot Interactions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 28(1).

Nakagawa, H.; Hasegawa, S.; Hagiwara, Y.; Taniguchi, A.; and Taniguchi, T. 2024. Pointing Frame Estimation With Audio-Visual Time Series Data for Daily Life Service Robots. In *2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2949–2956.

Ojha, S.; Gervits, F.; and Espy-Wilson, C. 2025. Speaking with Robots in Noisy Environments. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 1057–1061. IEEE.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of the 40th International Conference on Machine Learning*.

Rauscher, F. H.; Krauss, R. M.; and Chen, Y. 1996. Gesture, speech, and lexical access: The role of lexical movements in speech production. *Psychological Science*, 7.

Srimal, A. S.; Muthugala, V. J.; and Jayasekara, B. P. 2017. Deictic Gesture Enhanced Fuzzy Spatial Relation Grounding in Natural Language. In *2017 IEEE International Conference on Fuzzy Systems*. IEEE.

Tellex, S.; Gopalan, N.; Kress-Gazit, H.; and Matuszek, C. 2020. Robots That Use Language. *Annual Review of Control, Robotics, and Autonomous Systems*.