Find it like a dog: Using gesture to improve object search?

Madeline H. Pelgrim [1](**madeline_pelgrim@brown.edu**) **Ivy Xiao He** [2] (**xiao_he@brown.edu**)
Kyle Lee [2] (kyle_k_lee@brown.edu) **Falak Pabari** [2] (falak_pabari@brown.edu) **Stefanie Tellex** [2] (stefie10@cs.brown.edu)
Thao Nguyen [2] (thao_nguyen3@brown.edu) **Daphna Buchsbaum** [1] (daphna_buchsbaum@brown.edu)
[1]Department of Cognitive, Linguistic, & Psychological Sciences, Brown University
[2] Department of Computer Science, Brown University

Abstract

Pointing is an intuitive and commonplace communication modality. In human-robot collaborative tasks, human pointing has been modeled using a variety of approaches, such as the forearm vector or the vector from eye to hand. However, models of the human pointing vector have not been uniformly or comprehensively evaluated. We performed a user study to compare five different representations of the pointing vector and their accuracies in identifying the human's intended target in an object selection task. We also compare the vectors' performances to that of domestic dogs to assess a non-human baseline known to be successful at following human points. Additionally, we developed an observation model to transform the vector into a probability map for object search. We implemented our system on our robot, enabling it to locate and fetch the user's desired objects efficiently and accurately.

Keywords: Robotics; Comparative Studies; Gesture Analysis; Human-Animal Interaction

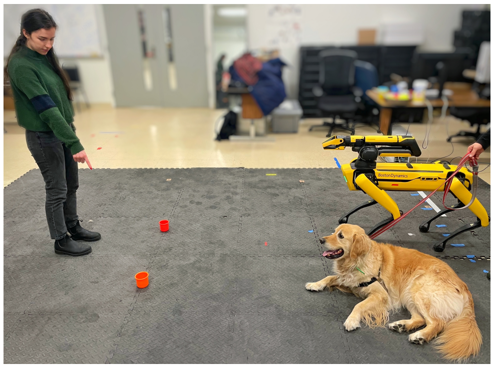

Figure 1: Our system enables a robot to locate objects using information from a person's unscripted gestures. We compare our robot's performance to that of the domestic dog.

Introduction

People must communicate locations for various tasks and often use pointing gestures. When pointing, a person uses their head, body, hand, and arm to refer to an object or location in the environment. Using a deictic gesture such as pointing is intuitive for a person and directly communicates spatial information in the form of a 3D vector through space. Existing literature has shown that people can interpret points from others from infancy (e.g., Butterworth (1998)) and are highly accurate at interpreting the specific target of human pointing gestures (Bertenthal et al., 2014; Wnuczko & Kennedy, 2011). Point following is not limited to human beings; other species, in particular dogs, are able to follow human pointing gestures to locate hidden objects (Agnetta et al., 2000; Hare et al., 2002; Miklösi et al., 1998; Soproni et al., 2001) with little or no training, and from a very young age (e.g., Bray et al. (2021); Riedel et al. (2008)). Dogs also have a unique social relationship with human beings, partnering with us as companions and in a variety of service roles. Their ability to respond to our social cues makes them particularly promising for the exploration of human gesture comprehension by a non-human agent, and an interesting model for human-robot interactions (Krueger et al., 2021; Byrne et al., 2020). Further, dog's close relationship with humans means many robots are embodied in dog-like structures, for example Sony's AIBO or Boston Dynamic's Spot (Kerepesi et al., 2006; Faragó et al., 2014).

Existing work on robotic following of human pointing gestures has used a variety of methods to obtain the 3D vector through space corresponding to the point. Previous works (Constantin et al., 2022; Ekrekli et al., 2023; Obo et al., 2018; Whitney et al., 2017) have demonstrated effective human-robot collaboration on non-search tasks by incorporating pointing gestures and speech to relay task-relevant information to a robot. Such existing approaches rely solely on social feedback and gestures to help identify the target object the human is pointing to. However, they fail to consider that objects can be hidden from view from the robot or the human's perspective, or be of different distances away from the person, so that the target of the point is ambiguous depending on how the pointing vector is identified.

Prior work in both the robotics and cognitive science communities has used a range of vectors, such as the vector from the person's eyes to their hand (Abidi et al., 2013; Azari et al., 2019; Taylor & McCloskey, 1988; Whitney et al., 2017), the forearm vector (Herbort & Kunde, 2016; Hu et al., 2022; Tölgyessy et al., 2017; Whitney et al., 2016), as well as other non-pointing vectors such as eye gaze (Nickel & Stiefelhagen, 2003; Mayer et al., 2018; Perez-Osorio et al., 2015). Other work has looked to integrate these vectors to help the point viewer localize the target, or integrate linguistic information (Kranstedt et al., 2006) However, there has been no systematic study that measures which approach most accurately enables a robot to resolve pointing gestures to spatial

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

object locations or best corresponds to what vector other entities, such as dogs, use to follow points.

Our work addresses this gap by presenting a mathematical framework for incorporating human pointing gestures into robotic object search. We present five algorithms for resolving a pointing gesture to a 3D vector in space and then transform that vector into a probability map in the physical world using a generative observation model. Using this probability map, the robot can incorporate information from the pointing vector to efficiently find objects in the environment in collaboration with the person. To our knowledge, no previous work has used pointing gestures to give a robot information for object search.

We evaluate five different approaches for converting information from the human's body pose into a 3D vector in space, including the vector from eye-to-wrist, nose-to-wrist, elbow-to-wrist, shoulder-to-wrist, and the eye gaze vector. To our knowledge, we are the first work to evaluate these different approaches systematically; other works have used these approaches individually but have not directly compared them against each other to see what is most accurate. We compare our approach to the non-human baseline of the domestic dog, which is well-known to be able to follow human pointing gestures (Agnetta et al., 2000), and to engage in collaborative object search and other cooperative tasks with humans (Hare et al., 2002). Our results demonstrate quantitatively that the vector from the gaze only performs the worst, while our other four candidate vectors display comparable levels of performance at identifying the object the person is pointing at. Finally we demonstrate an end-to-end object search system running on a real robot, showing that the robot can incorporate information from the person's gesture to find objects effectively.

Related Work

Different humans may have unique ways of expressing pointing gestures, so it is imperative to understand the motor and perceptual processes used to generate pointing gestures. Past work with infants has found that from an early age, they understand that points are intended to direct another person's attention toward an object or location in the environment and that points are intentional so that people will not point at things they do not know about and cannot see (Sodian & Thoermer, 2004; Liebal et al., 2009; Woodward & Guajardo, 2002). But how are points produced and interpreted by adults? Point production and following are ubiquitous in daily life. Under normal pointing conditions (meaning full visual access to the target), pointers tend to use an eye-to-hand vector. When blindfolded, however, pointers gesture with their arm alone (Wnuczko & Kennedy, 2011). There are also differences in how far the item being indicated is from the two vectors (eye to the hand vs. down the arm), with arm-only points consistently overshooting the target. This error in production is also mirrored by errors in comprehension. While humans are generally quite accurate at producing points for others, past work has revealed minor but systemic errors in how the viewer perceives the targets of points (Herbort & Kunde, 2016; Herbort et al., 2021). Much of this has to do with errors in perspective taking, with researchers suggesting that the pointer fails to account for the different viewing angles of the viewer. While there has been important work on how people understand and produce points, there has yet to be a systemic investigation of naturalistic point production. Further, we suspect that humans point differently when they point for other humans versus non-human entities such as dogs or robots, but this has not been explored.

Dynamic gesture recognition commonly employs non-vision-based and vision-based approaches. While the former often requires external devices, the latter, within the realm of Computer Vision, has seen considerable efforts aimed at quantifying and comprehending human pointing gestures, emphasizing automated end-to-end gesture detection. (Jaiswal et al., 2018; Köpüklü et al., 2019; Yu et al., 2022; Nakamura et al., 2023). Jaiswal et al. (2018) trained a deep convolutional neural network to estimate the direction of finger pointing gestures. The estimation only used the position of the person's elbow and wrist, disregarding many other relevant key points on the human body. Nakamura et al. (2023) introduced a large-scale dataset and model for pointing recognition and direction estimation. However, their data always include the person's entire body, which a robot might not have access to from its point of view. An unsupervised learning approach (Jirak et al., 2021) has been constructed to model the variation of pointing gestures without scaling computationally with the number of gestures and objects being pointed at. Such research also distinguishes between arbitrary hand movements and meaningful gestures. Our work employs Google's MediaPipe Pose Landmarker (Bazarevsky et al., 2020), a CNN model for human pose estimation, to detect key points on the human user's body and explicitly compute 3D pointing vectors.

A number of human-computer and human-robot interaction (HRI) papers discuss how to interpret a pointing gesture. Human-computer interaction works (Mayer et al., 2018, 2015) usually require that people wear a headset and use a clicker to get visual feedback, which can be costly and difficult to use. We prefer the interaction to be as natural and as comfortable as possible for human users.

Other approaches study how to enable a robot to generate a pointing gesture. Fang et al. (2015) described incorporating pointing gestures with language where the robot points to objects in order to specify them to a person. Williams et al. (2013) studied how humans interpret robot pointing behavior, finding that the robot's head and neck were important in understanding pointing references. They studied three hypotheses: the arm vector, the line of sight from the robot's head to the end of the gripper, and the direction of the robot's head. They found that people interpret the pointing behavior of robots differently from that of people, using the robot's head gaze more than people do.

There have been many previous works on robot object search. Model-free approaches (Faust et al., 2018; Niroui, Zhang, Kashino, & Nejat, 2019; Chaplot, Gandhi, Gupta, & Salakhutdinov, 2020; Gadre, Wortsman, Ilharco, Schmidt, & Song, 2022) leverage deep neural networks to learn a policy end-to-end, and thus are data hungry and have limited generalization capabilities. Model-based works frequently employ the Partially Observable Markov Decision Processes (POMDPs) (Kaelbling, Littman, & Cassandra, 1998) framework to define and solve the problem of object search. Wandzel et al. (2019) introduce Object-Oriented POMDP (OO-POMDP) to factorize the robot's belief into independent object distributions, enabling the size of the belief to scale linearly in the number of objects, and employ it for efficient multi-object search. Zheng et al. (2023) extend OO-POMDP for efficient multi-object search in 3D space. We build off of their framework, termed GenMOS, for our robot object search demonstrations.

Technical Approach

Our approach enables a Boston Dynamics Spot robot to interpret a person's pointing gesture to find objects in the environment. To perform this task, we need to estimate the person's body pose and then use information from the pose to interpret pointing gestures. We explore different approaches to converting the body pose into a vector in the world. Finally, we define an observation model to convert this vector to a Gaussian expectation to enable the robot to use information from the pointing gesture to search for objects.

Human Body Pose Estimation

We record the person's body pose with an Intel RealSense camera. We used a camera on a tripod pointed at the scene for our experiments. For the on-board experiments, we used an RGB-D camera that is part of the Spot robot's on-board sensing. Given a camera image, we need to estimate the human body pose. We use Google's MediaPipe Pose Landmarker (Bazarevsky et al., 2020) to process input RGB images and detect key points on the human body. We then employ the depth information to transform the relevant key points' coordinates from 2D into 3D space. Example input RGB-D images and MediaPipe's output are shown in Figure 3. We then ray-casted vectors from various key points with the user's wrist as the endpoint and calculated the vectors' intersection points with the environment as the pointing targets. Figure 2 presents visualizations of the five pointing vectors.

We assume the person is already in the robot's field of view. We also need to make sure the camera is calibrated relative to the position of the robot in order to situate the pointing vector correctly in the robot's frame of reference. This requires a camera calibration step for an off-board camera, which we perform using April tags (Wang & Olson, 2016).

Converting Human Body Pose to Pointing Vectors

Given the body pose of a person, we need to extract a vector according to the pointing gesture. We explore five different algorithms for computing a vector from the person's body pose. Given this vector, we re-cast it into the environment and calculate the vectors' intersection points with the environment as the pointing targets. We explore two different high-level approaches: the vector from the head to the hand and the vector from the arm. We use the person's wrist position as a proxy for their hand, as fingers are much smaller and thus more difficult to detect. We also formed a corresponding gaze vector that represents the general direction the user is looking at. To distinguish meaningful gestures from arbitrary noise, such as from crossed arms, we included an angle threshold so that vectors could be filtered out. Visualizations of the five pointing vectors are shown in Figure 2.

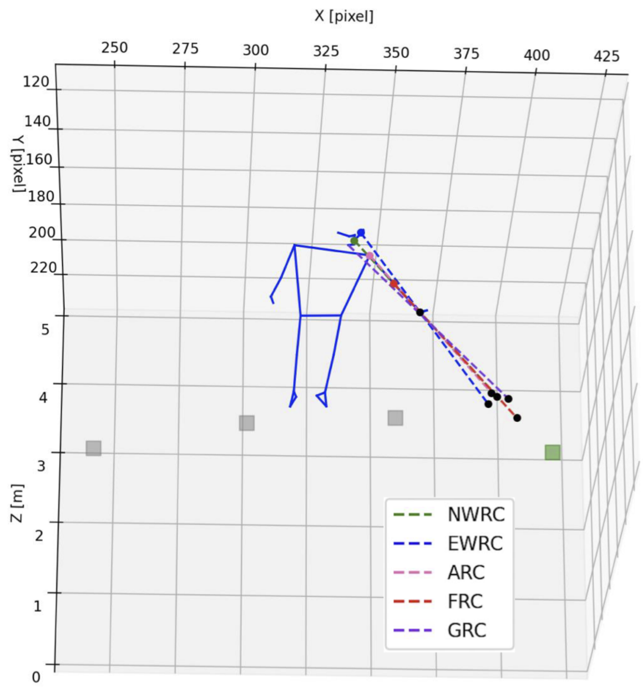

Figure 2: Five pointing vectors on a sample image: eye-to-wrist, nose-to-wrist, shoulder-to-wrist, elbow-to-wrist, and eye gaze. The left wrist is used as the frame of reference.

Our work uses five pointing vectors: First, we use the **Eye-to-wrist ray-cast (EWRC)**, defined by a vector connecting the eye and wrist of the pointing arm. Next, we use the **Nose-to-wrist ray-cast (NWRC)**, which is defined by a vector connecting the nose and wrist of the pointing arm. Third, we use the **Arm ray-cast (ARC)**, a ray-cast defined by a vector connecting the shoulder and wrist of the pointing arm. Fourth, we use the **Forearm ray-cast (FRC)**, a ray-cast defined by a vector connecting the elbow and wrist of the pointing arm. Finally, we use the **Gaze ray-cast (GRC)**, a ray-cast that establishes a corresponding gaze vector representing the general direction the user is looking at. To find the GRC, we computed the normal vector to the plane passing through their left eye, right eye, and center of the mouth.

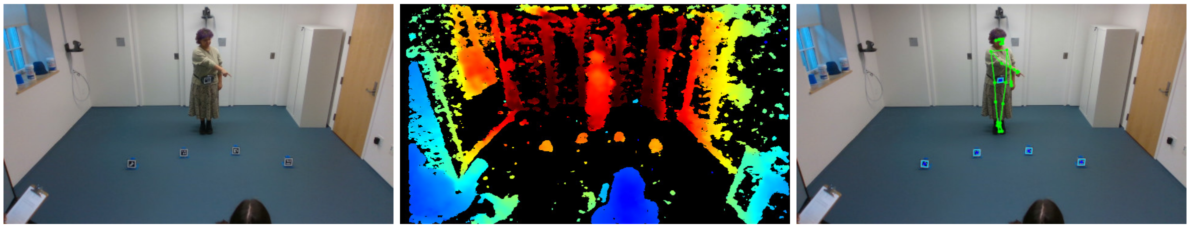

Figure 3: An example RBG image (Left), depth image (Middle) and MediaPipe's keypoint detection output (Right).

Observation Model for Pointing Vectors

A second technical contribution of this paper is a formal observation model for object search defined in terms of a pointing vector established from the person's body in Section . Following (Zheng et al., 2023), we formalize the object search problem as an OO-POMDP. The robot state consists of its pose. Actions consist of moving through the environment, looking with its sensor at a particular location, and marking the object as found. After correctly marking the object as found, the robot receives a reward.

We define an observation model for pointing for GenMOS as object search as:

$$\Pr(o|s,a). \tag{1}$$

In previous work, the observation consisted of sensor input from the robot's sensor, as well as natural language input from the person. However, to our knowledge, no previous work has used pointing gestures to give a robot information for object search. Our work assumes the vector can be parameterized by two angles, corresponding to the "pitch" and "yaw" of the pointing vector, which we term α and β, defined in Section . We assume the vector is parameterized by the person's location, which is known, height, which is directly observed, and two angles, α and β, giving us:

$$\Pr(o|s,a) = \Pr(\alpha,\beta|s,a) \tag{2}$$

We project this distribution forward to the ground as a vector as shown in Figure 4. The noise of this vector will result in a Gaussian shaped like an ellipse because moving the "pitch" of the arm will move the target farther away, creating more uncertainty about where the object is; whereas moving the "yaw" will keep the same distance. Thus, uniform variance in both of these two angles will result in a different distribution on where the point intersects the plane.

$$\Pr(\alpha,\beta|X) = \mathcal{N}(\mu,\Sigma) \tag{3}$$

In our evaluation, we collect a dataset of people pointing out objects. For each item in the dataset, n, we have information about the true pose of the object, X, as well as the pointing ray α and β, observed from depth measurements. Given this information, we can compute the perplexity of the model over the dataset N as follows:

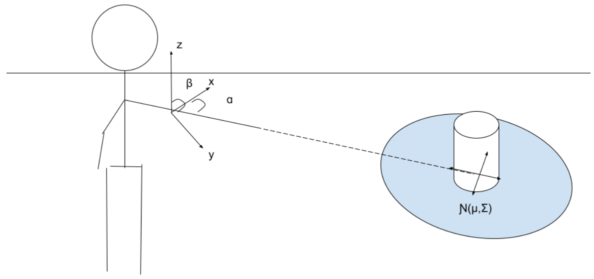

Figure 4: A diagram showing our observation model as a Gaussian projected onto the ground plane. For simplicity, we show this vector as the arm vector (method ARC); in reality, our evaluation assesses several different vectors.

$$Perplexity(N) = exp\left\{-\frac{1}{|N|}\sum_{n\in N}\log\Pr(\alpha,\beta|X)\right\} \tag{4}$$

In the special case where we know the object is in one of a predefined set of locations t_i and the distance from the pointing ray intersection location to each target d_n, we can compute the perplexity as a multinomial over the true location as follows:

$$\mathcal{L}(t_i|d_1,\ldots,d_n) \propto \frac{d_i^{-1}}{\sum_{j=1}^{n} d_j^{-1}} \tag{5}$$

$$Perplexity(N) = exp\left\{-\frac{1}{|N|}\sum_{n\in N}\log\mathcal{L}(t_i|d_1,\ldots,d_n)\right\} \tag{6}$$

Equation 6 is how we compute the perplexity scores in our experiments in Section .

Evaluation

The aim of our evaluation is to measure the effectiveness of different vectors for enabling a robot to accurately and efficiently resolve human pointing gestures to find objects. We collect a new dataset of humans pointing to a non-human partner, the domestic dog. We hypothesize that human-dog interaction is similar to human-robot interaction. We contrast this with humans, pointing for another human to see if there are differences in behavior. The robot we use for interpreting the pointing gesture is a quadruped robot, the Boston Dynamics Spot robot. We use this dataset to evaluate the performance of our five different approaches for resolving pointing gestures based on human body pose and also compare our algorithm's performance to that of the dogs.

Finally, we perform an end-to-end demonstration on the real robot, demonstrating our algorithm's use to enable a robot to resolve pointing gestures.

Experimental Setup

To assess the natural interaction between humans and dogs through deictic gestures, we brought dog-guardian pairs into the lab to observe both how guardians naturally point for their dogs and how their dogs behave.

Participants Six human-dog pairs participated in the pointing tasks. Dog owners were all adults (over 18 years of age) who acted as the primary caretaker for their dog. The dogs were 5.2 years old on average, and three of the six dogs were female.

Materials The experimental setup, illustrated in Figure 5, comprises four cups placed equidistant in front of the dog. To minimize external device interference in the dogs' decision-making process, we utilized the Intel RealSense D435 camera to capture depth and RGB image. Dog treats were used to motivate dogs to search.

Procedure Before pointing, dog-human pairs completed two warm-up activities. First, in the initial familiarization phase, dogs observed their guardians place a treat under a cup and then were released to touch the cup, constituting a choice. This familiarized dogs first with touching a target to reveal the hidden treat and was repeated four times. Next, during the hidden familiarization phase, dogs got to practice leaving the room and returning to locate a treat under the one hidden target. For our critical test trials, as in hidden familiarization at the start of each trial, the dog was led out of the room so the guardian could place a concealed treat beneath one of the targets as instructed by an experimenter (4 targets used, order semi-randomized). The dog was returned to the room, and guardians were instructed to point their dogs to the hidden treat. Dogs were then allowed to search exhaustively. This procedure was repeated for 12 trials and 72 recorded trials across dog-human pairs.

After pointing to their dogs, 3 of the 6 guardians were recorded pointing the human experimenter to the cups in a semi-randomized order. At the start of each human pointing trial, dog guardians were asked to point to one of the four cups. The experimenter then waited approximately 2 seconds before following the point visually with their eyes and then instructing the dog guardian to point to the next cup. The room setup was the same as in Figure 3, and this was repeated for 12 trials, for a total of 36 recorded trials. We evaluated the 5 vectors' performance on this data and report the results along with the 95% confidence interval in Table 2.

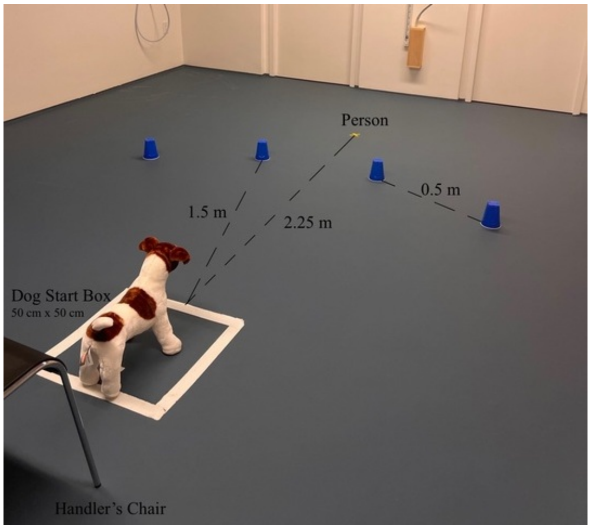

Figure 5: Annotated image of the room setup

Evaluation Metrics

We manually annotate the frame of the image used to evaluate human pointing results. We use three metrics to evaluate average object selection performance. First, we used the Euclidean distance offset to the target object, where lower values are better. Second, we evaluated the weighted accuracy, where higher values indicate better performance. Where A is 1 if the correct target was selected and 0 otherwise, n is the number of selections made until the target is selected, and w is the probability of a target being selected—calculated using the normalized inverse Euclidean distance.

$$acc = \frac{\sum_{i=1}^{n} w_i A_i}{\sum_{i=1}^{n} w_i}$$

Finally, we evaluate perplexity (PP), measuring a model's uncertainty where a lower value is better. following equation 6, the target can be 1 of 4 cups, and we use the Euclidean distance from the pointing ray intersection to each cup to compute the perplexity score.

Human-Dog Pointing Results

Even under naturalistic pointing conditions, dogs sometimes had difficulty following the human pointing gesture. Dogs were allowed to search exhaustively, and on their first choice dogs chose the pointed location on 37% of trials (chance being 1/4 or 25%, and on 42% of trials dogs chose correct location as their second choice. This is fairly consistent with past work with dogs when four search locations are used (Lakatos et al., 2012). The two locations closer to the human pointer tend to be chosen more frequently than those on the periphery. In our sample, consistent with past work, dogs were highly accurate at choosing the correct side of the indicated cup, going to the correct side (to the pointer's Left or Right) on the first trial 76% of the time. Most errors dogs made involved choosing the cup closer to the guardian, rather than the one further from the guardian on the same side. The proximity of the cup to the guardian may make it more attractive, as the proximity of a person is a cue that dogs can use to find hidden food (Hare & Tomasello, 1999). It is also possible that dogs were seeking attention from their guardians and were thus attracted to the closer locations or that their past reward history with their guardians (meaning they have received lots of rewards directly from their guardians) causes dogs to prefer to search nearer to their guardians. We leave a full evaluation of these results to a future paper, as the primary focus of this paper is the performance of our autonomous pointing algorithms.

Ray-cast Performance

Table 1 shows the performance of our five different vectors for resolving pointing gestures along with 95% confidence intervals. Our primary result is that most methods perform similarly, with the lowest-performing method using gaze alone, which performs significantly worse than other baselines. It is interesting to see that the eye-to-wrist vector has higher accuracy, but the shoulder-to-wrist has the lowest PP. Given the confidence intervals, there may not be much of a significant difference between which vector to use. The perplexity differs from accuracy as it is more resistant to noise in the data: a small change in the distance from the vector's intersection location to the cups can result in a large change in the accuracy but not the perplexity score.

Table 1: Performance from humans pointing for their dogs

	Distance(m)↓	Accuracy(%)↑	PP↓
EWRC	0.516 *(0.071)*	**96.9 *(3.4)***	3.213 *(0.135)*
NWRC	**0.514 *(0.065)***	95.7 *(3.8)*	3.128 *(0.112)*
ARC	0.565 *(0.065)*	94.0 *(4.2)*	**3.111 *(0.135)***
FRC	0.868 *(0.272)*	92.5 *(4.2)*	3.372 *(0.131)*
GRC	2.711 *(0.158)*	51.8 *(8.4)*	3.581 *(0.120)*

Human-Human Pointing Experiment

There appears to be consistent performance between nose, eye, and shoulder-to-wrist vectors. The accuracy does not differ much in human-to-dog versus human-to-human, but PP is better in the human-to-human case (PP_dog_baseline = 4). While conclusions should be limited at this time given the reduced sample size, it is interesting that, as observed in the human-dog pointing data, the gaze-only vector is a much worse fit, while all other vectors perform exceptionally well.

Spot Demonstration

We demonstrate the effectiveness of resolving pointing vectors on the Spot robot. First we assess the accuracy of pointing by directly using the vector to resolve the object reference to the object closest to the pointing vector intersection. Then we used the Spot API to direct the robot to pick up that object. The results are demonstrated in Figure 6. Spot was able to follow the human pointer to correctly approach and select the indicated object from a set of four candidate objects.

Table 2: Performance on pointing for another person

	Distance(m)↓	Accuracy(%)↑	PP↓
EWRC	0.607 *(0.117)*	**100.0 *(0)***	3.228 *(0.162)*
NWRC	**0.591 *(0.108)***	**100.0 *(0)***	3.199 *(0.177)*
ARC	0.593 *(0.123)*	**100.0 *(0)***	**3.066 *(0.213)***
FRC	0.742 *(0.170)*	98.6 *(2.8)*	3.265 *(0.187)*
GRC	2.947 *(0.305)*	57.0 *(10.0)*	3.986 *(0.002)*

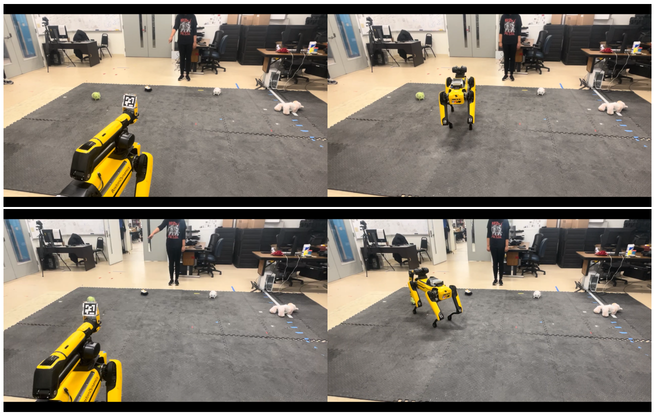

Figure 6: Our system enables the robot to correctly fetch the object the human user is pointing at, such as the penguin plush (left) and green cat (right).

Conclusion

In this paper we presented an evaluation of different pointing vectors to resolve human pointing gestures to locations in the environment. We present a probabilistic observation model for how this vector can be used for object search, the primary reason humans point for other humans. We evaluated our system on a new dataset of humans pointing for their domestic dog, as well as humans pointing for other humans, and compared the performance of our autonomous algorithms to that of dogs.

Future work can consider using timecourse data and pointing information from videos rather than still images. Anecdotally, many pointers first aligned their gaze with the target, then moved their gaze back to the point viewer when initiating arm movement. This could help to explain why, at the moment of pointing, the gaze-only vector had such poor accuracy. Further, this approach can be used on more complex applied object search tasks, such as those requiring cooperation between human and dog or human and robot.

Acknowledgments

Thank you to the dogs and their guardians who participated in this work. Thanks to Olivia Williamson, Brooke Jordan, and the whole Brown Dog Lab team. This work is supported by the Brown Data Science Institute, Natural Sciences and Engineering Research Council of Canada (NSERC) Postgraduate Scholarship - Doctoral under award number 567607, Office of Naval Research under grant number N00014-22-1-2592, Air Force Office of Scientific Research under grant number FA9550-21-1-0214, and with support from Amazon Robotics.

References

Abidi, S., Williams, M., & Johnston, B. (2013). Human pointing as a robot directive. In *8th acm/ieee international conference on human-robot interaction (hri)* (pp. 67–68).

Agnetta, B., Hare, B., & Tomasello, M. (2000). Cues to food location that domestic dogs (Canis familiaris) of different ages do and do not use. *Animal Cognition*, *3*(2), 107–112. doi: 10.1007/s100710000070

Azari, B., Lim, A., & Vaughan, R. (2019). Commodifying pointing in hri: simple and fast pointing gesture detection from rgb-d images. In *16th Conference on Computer and Robot Vision (CRV)* (pp. 174–180).

Bazarevsky, V., Grishchenko, I., Raveendran, K., Zhu, T., Zhang, F., & Grundmann, M. (2020). *Blazepose: On-device real-time body pose tracking.*

Bertenthal, B. I., Boyer, T. W., & Harding, S. (2014). When do infants begin to follow a point? *Developmental Psychology*, *50*(8), 2036–2048. doi: 10.1037/a0037152

Bray, E. E., Gnanadesikan, G. E., Horschler, D. J., Levy, K. M., Kennedy, B. S., Famula, T. R., & MacLean, E. L. (2021, July). Early-emerging and highly heritable sensitivity to human communication in dogs. *Current Biology*, *31*(14), 3132–3136.e5. doi: 10.1016/j.cub.2021.04.055

Butterworth, G. (1998). What is special about pointing in babies? In *The development of sensory, motor and cognitive capacities in early infancy: From perception to cognition* (pp. 171–190). Hove, England: Psychology Press/Erlbaum (UK) Taylor & Francis.

Byrne, C., Logas, J., Freil, L., Allen, C., Baltrusaitis, M., Nguyen, V., ... Jackson, M. M. (2020, January). Dog Driven Robot: Towards Quantifying Problem-Solving Abilities in Dogs. In *Proceedings of the Sixth International Conference on Animal-Computer Interaction* (pp. 1–5). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3371049.3371063

Chaplot, D. S., Gandhi, D. P., Gupta, A., & Salakhutdinov, R. R. (2020). Object Goal Navigation using Goal-Oriented Semantic Exploration. *Advances in Neural Information Processing Systems*, *33*, 4247–4258.

Constantin, S., Eyiokur, F. I., Yaman, D., Bärmann, L., & Waibel, A. (2022). Interactive Multimodal Robot Dialog Using Pointing Gesture Recognition. In *European conference on computer vision* (pp. 640–657).

Ekrekli, A., Angleraud, A., Sharma, G., & Pieters, R. (2023). Co-speech gestures for human-robot collaboration. *arXiv preprint arXiv:2311.18285.*

Fang, R., Doering, M., & Chai, J. Y. (2015). Embodied collaborative referring expression generation in situated human-robot interaction. In *Proceedings of the tenth annual acm/ieee international conference on human-robot interaction* (pp. 271–278).

Faragó, T., Gácsi, M., Korcsok, B., & Miklósi, (2014, August). Why is a dog-behaviour-inspired social robot not a doggy-robot? *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems*, *15*(2), 224–232. doi: 10.1075/is.15.2.11far

Faust, A., Oslund, K., Ramirez, O., Francis, A., Tapia, L., Fiser, M., & Davidson, J. (2018). PRM-RL: Long-range Robotic Navigation Tasks by Combining Reinforcement Learning and Sampling-based Planning. In *Proceedings of the international conference on robotics and automation* (pp. 5113–5120).

Gadre, S. Y., Wortsman, M., Ilharco, G., Schmidt, L., & Song, S. (2022). CLIP on Wheels: Zero-Shot Object Navigation as Object Localization and Exploration. *arXiv preprint arXiv:2203.10421.*

Hare, B., Brown, M., Williamson, C., & Tomasello, M. (2002). The Domestication of Social Cognition in Dogs. *Science*, *298*(5598), 1634. doi: 10.1126/science.1072702

Hare, B., & Tomasello, M. (1999). Domestic dogs (Canis familiaris) use human and conspecific social cues to locate hidden food. *Journal of Comparative Psychology*, *113*(2), 173–177. (Place: US Publisher: American Psychological Association) doi: 10.1037/0735-7036.113.2.173

Herbort, O., Krause, L.-M., & Kunde, W. (2021, April). Perspective determines the production and interpretation of pointing gestures. *Psychonomic Bulletin & Review*, *28*(2), 641–648. doi: 10.3758/s13423-020-01823-7

Herbort, O., & Kunde, W. (2016). Spatial (mis-)interpretation of pointing gestures to distal referents. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(1), 78–89. doi: 10.1037/xhp0000126

Hu, Z., Xu, Y., Lin, W., Wang, Z., & Sun, Z. (2022). Augmented Pointing Gesture Estimation for Human-Robot Interaction. In *Ieee international conference on robotics and automation* (pp. 6416–6422).

Jaiswal, S., Mishra, P., & Nandi, G. (2018). Deep learning based command pointing direction estimation using a single rgb camera. In *5th ieee uttar pradesh section international conference on electrical, electronics and computer engineering (upcon)* (pp. 1–6).

Jirak, D., Biertimpel, D., Kerzel, M., & Wermter, S. (2021). Solving visual object ambiguities when pointing: an unsupervised learning approach. *Neural Computing and Applications*, *33*, 2297–2319.

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, *101*(1-2), 99–134.

Kerepesi, A., Kubinyi, E., Jonsson, G. K., Magnusson, M. S., & Miklósi, (2006, July). Behavioural comparison of human–animal (dog) and human–robot (AIBO) interactions. *Behavioural Processes*, *73*(1), 92–99. doi: 10.1016/j.beproc.2006.04.001

Köpüklü, O., Gunduz, A., Kose, N., & Rigoll, G. (2019). Real-time hand gesture detection and classification using convolutional neural networks. In *14th ieee international conference on automatic face & gesture recognition (fg 2019)* (pp. 1–8).

Kranstedt, A., Lücking, A., Pfeiffer, T., Rieser, H., & Wachsmuth, I. (2006, January). Deixis: How to Determine Demonstrated Objects Using a Pointing Cone.. doi: 10.1007/11678816$_3$4

Krueger, F., Mitchell, K. C., Deshpande, G., & Katz, J. S. (2021, January). Human–dog relationships as a working framework for exploring human–robot attachment: a multidisciplinary review. *Animal Cognition*. doi: 10.1007/s10071-021-01472-w

Lakatos, G., Gácsi, M., Topál, J., & Miklósi, Á. (2012, March). Comprehension and utilisation of pointing gestures and gazing in dog–human communication in relatively complex situations. *Animal Cognition*, *15*(2), 201–213. doi: 10.1007/s10071-011-0446-x

Liebal, K., Behne, T., Carpenter, M., & Tomasello, M. (2009, March). Infants use shared experience to interpret pointing gestures. *Developmental Science*, *12*(2), 264–271. doi: 10.1111/j.1467-7687.2008.00758.x

Mayer, S., Schwind, V., Schweigert, R., & Henze, N. (2018). The Effect of Offset Correction and Cursor on Mid-Air Pointing in Real and Virtual Environments. In *Proceedings of the 2018 chi conference on human factors in computing systems* (pp. 1–13).

Mayer, S., Wolf, K., Schneegass, S., & Henze, N. (2015). Modeling distant pointing for compensating systematic displacements. In *Proceedings of the 33rd annual acm conference on human factors in computing systems* (pp. 4165–4168).

Miklösi, A., Polgárdi, R., Topál, J., & Csányi, V. (1998). Use of experimenter-given cues in dogs. *Animal Cognition*, *1*(2), 113–121. doi: 10.1007/s100710050016

Nakamura, S., Kawanishi, Y., Nobuhara, S., & Nishino, K. (2023). DeePoint: Visual Pointing Recognition and Direction Estimation. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 20577–20587).

Nickel, K., & Stiefelhagen, R. (2003). Pointing gesture recognition based on 3d-tracking of face, hands and head orientation. In *Proceedings of the 5th international conference on multimodal interfaces* (pp. 140–146).

Niroui, F., Zhang, K., Kashino, Z., & Nejat, G. (2019). Deep Reinforcement Learning Robot for Search and Rescue Applications: Exploration in Unknown Cluttered Environments. *IEEE Robotics and Automation Letters*, *4*(2), 610–617.

Obo, T., Kawabata, R., & Kubota, N. (2018). Cooperative human-robot interaction based on pointing gesture in informationally structured space. In *World automation congress (wac)* (pp. 1–5).

Perez-Osorio, J., Müller, H. J., Wiese, E., & Wykowska, A. (2015). Gaze Following Is Modulated by Expectations Regarding Others' Action Goals. *PLOS ONE*, *10*, e0143614. doi: 10.1371/journal.pone.0143614

Riedel, J., Schumann, K., Kaminski, J., Call, J., & Tomasello, M. (2008). The early ontogeny of human-dog communication. *Animal Behaviour*, *75*(3), 1003–1014. (Place: Netherlands Publisher: Elsevier Science) doi: 10.1016/j.anbehav.2007.08.010

Sodian, B., & Thoermer, C. (2004). Infants' Understanding of Looking, Pointing, and Reaching as Cues to Goal-Directed Action. *Journal of Cognition and Development*, *5*(3), 289–316. doi: 10.1207/s15327647jcd0503$_1$

Soproni, K., Miklósi, A., Topál, J., & Csányi, V. (2001, June). Comprehension of human communicative signs in pet dogs (Canis familiaris). *Journal of Comparative Psychology*, *115*(2), 122–126. doi: 10.1037/0735-7036.115.2.122

Taylor, J. L., & McCloskey, D. I. (1988). Pointing. *Behavioural Brain Research*, *29*, 1–5. doi: 10.1016/0166-4328(88)90046-0

Tölgyessy, M., Dekan, M., Duchoň, F., Rodina, J., Hubinskỳ, P., & Chovanec, L. (2017). Foundations of visual linear human–robot interaction via pointing gesture navigation. *International Journal of Social Robotics*, *9*, 509–523.

Wandzel, A., Oh, Y., Fishman, M., Kumar, N., LS, W. L., & Tellex, S. (2019). Multi-Object Search using Object-Oriented POMDPs. In *Proceedings of the international conference on robotics and automation* (pp. 7194–7200).

Wang, J., & Olson, E. (2016). AprilTag 2: Efficient and robust fiducial detection. In *2016 ieee/rsj international conference on intelligent robots and systems (iros)* (pp. 4193–4198).

Whitney, D., Eldon, M., Oberlin, J., & Tellex, S. (2016). Interpreting Multimodal Referring Expressions in Real Time. In *IEEE International Conference on Robotics and Automation.*

Whitney, D., Rosen, E., MacGlashan, J., Wong, L. L., & Tellex, S. (2017). Reducing Errors in Object-Fetching Interactions through Social Feedback. In *Ieee international conference on robotics and automation* (pp. 1006–1013).

Williams, M.-A., Abidi, S., Gärdenfors, P., Wang, X., Kuipers, B., & Johnston, B. (2013). Interpreting robot pointing behavior. In *Social robotics: 5th international conference, icsr 2013, bristol, uk, october 27-29, 2013, proceedings 5* (pp. 148–159).

Wnuczko, M., & Kennedy, J. M. (2011). Pivots for pointing: Visually-monitored pointing has higher arm elevations than pointing blindfolded. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(5), 1485.

Woodward, A. L., & Guajardo, J. J. (2002). Infants' understanding of the point gesture as an object-directed action. *Cognitive Development*, *17*(1), 1061–1084. doi:

10.1016/S0885-2014(02)00074-6

Yu, J., Qin, M., & Zhou, S. (2022). Dynamic gesture recognition based on 2D convolutional neural network and feature fusion. *Scientific Reports*, *12*(1), 4345.

Zheng, K., Paul, A., & Tellex, S. (2023). A System for Generalized 3D Multi-Object Search. In *Ieee international conference on robotics and automation (icra).*