Find It Like a Dog: Using Gesture to Improve Robot Object Search

Ivy Xiao He, Madeline H. Pelgrim, Kyle Lee, Falak Pabari, Daphna Buchsbaum, Stefanie Tellex, Thao Nguyen
{xiao_he, madeline_pelgrim, kyle_k_lee, falak_pabari}@brown.edu,
daphna_buchsbaum@brown.edu, stefie10@cs.brown.edu, thaonguyen@brown.edu

***Abstract*—Pointing is an intuitive and commonplace communication modality that humans have with each other, and with non-human entities such as dogs. Previous work has modeled the target of human pointing gestures for various human-robot collaboration tasks using many approaches such as the forearm vector or the eye-to-hand vector. However, models of the human users' pointing vector have not been uniform across the literature nor 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 found that the gaze-only vector performs the worst, while other vectors perform similarly well. We also compare the vectors' performances to that of domestic dogs, in order to assess a non-human baseline that is already known to be successful at following human points in a search task. We implemented our system on our robot, enabling it to efficiently and accurately locate and fetch the user's desired objects.**[1]

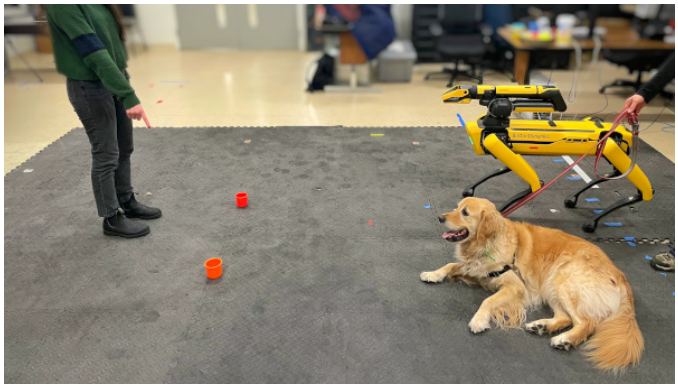

Fig. 1: Our system enables a robot to locate objects using information from a person's unscripted gestures.

I. Introduction

People need to communicate locations for a wide variety of tasks, and often use pointing gestures to do it. When pointing, a person uses their head, eyes, 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.

Existing literature has shown that people can interpret points from others from infancy (e.g., [1]), and are highly accurate at interpreting the specific target of human pointing gestures [2, 3]. Point following is not limited to human beings, other species, in particular dogs, are able to follow human pointing gestures to locate hidden objects [4, 5, 6, 7] with little or no training, and from a very young age (e.g., [8, 9]).

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 [10, 11, 12, 13, 14] have demonstrated effective human-robot collaboration on non-search tasks through the incorporation of pointing gestures along with 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, but without considering that objects can be hidden from view or be of different distances away from the robot or the human's perspective, 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 [15, 16, 11, 17], the forearm vector [10, 18, 19, 20], as well as other non-pointing vectors such as eye gaze [21, 22, 23] and pointing cone [24]. However, there has been no systematic study that measures which approach most accurately enables a robot to resolve pointing gestures to spatial 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 person's pointing gesture to a 3D vector in space, then calculate the vector's intersection point with the environment as the pointing target. The robot can use the pointing target to efficiently find objects in collaboration with the person. To our knowledge, no previous work has used pointing gestures for giving a robot information for object search. We evaluate five methods for converting body pose information into a 3D vector: eye-to-wrist, nose-to-wrist, elbow-to-wrist, shoulder-to-wrist, and eye gaze vector. We compare our method to the ability of domestic dogs to follow human pointing gestures for collaborative object search [4, 6]. Our results show that the gaze vector performs the worst, while the other vectors are similarly effective in identifying the pointed object.

II. Related Work

Humans express pointing gestures in various ways, such as head nodding, chin pointing, index finger, whole hand, eye gaze, nose, elbow, shoulder, thumb, and foot gestures, making it crucial to understand the motor and perceptual processes behind them. From an early age, infants can understand points

[1]The full paper is submitted to CogSci2024

are intended to direct another's attention towards an object or location in the environment, and people will not point at things they do not know about and cannot see [25, 26, 27]. But how are points produced and interpreted by adults? Point production and following is ubiquitous in daily life. Under pointing conditions with full visual access to the target, pointers tend to use an eye-to-hand vector. When blindfolded, however, pointers gesture with their arm alone [2]. There are also differences in how far the item being indicated is from the two vectors (eye to the hand vs. arm-only), with arm-only points consistently overshooting the target. This error in production is also mirrored with errors in comprehension. While in general, humans are quite accurate at producing points for others, past work has revealed that there are minor but systemic errors in how the viewer perceives the targets of points [18, 28]. Much of this has to do with errors in perspective taking, with researchers suggesting that the pointer fails to account for the different viewing angle of the viewer. While there has been important work in 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.

A number of human-robot interaction (HRI) papers discuss how to interpret a pointing gesture. Gesture tracking work [23, 29] usually require that people wear a headset and use a clicker to get visual feedback, which can be costly and difficult to use. We also want the interaction to be as natural and as comfortable as possible for human users. Nickel and Stiefelhagen [21] characterizes three approaches for estimating pointing direction: the line of sight between head and hand, forearm (elbow-to-wrist), and head orientation. In previous HRI work, one or another of these models was arbitrarily chosen as the "obvious" interpretation of pointing gestures: the eye-to-hand vector[11, 17], elbow-to-wrist vector[19, 20], eye gaze vector[23, 30], shoulder-to-wrist vector[30]. However, to our knowledge, there has not been a systematic evaluation of the different approaches to interpret a pointing gesture.

III. Technical Approach

Our approach enables a robot to interpret a person's pointing gesture and find objects in the environment. We first estimate the person's body pose, and then use the pose information to interpret pointing gestures. We explore different approaches to converting the body pose into a 3D vector.

A. Human Body Pose Estimation

RGB-D data is collected for human pose estimation. We use Google's MediaPipe Pose Landmarker [31],a deep-learning model for human pose estimation, to process input RGB images and detect keypoints on the human body. We then employ the depth information to transform the relevant keypoints' coordinates into 3D space.

We assume the person is already in view, and the camera is calibrated relative to the person's position to situate the pointing vector correctly in the camera's frame of reference using April tags [32].

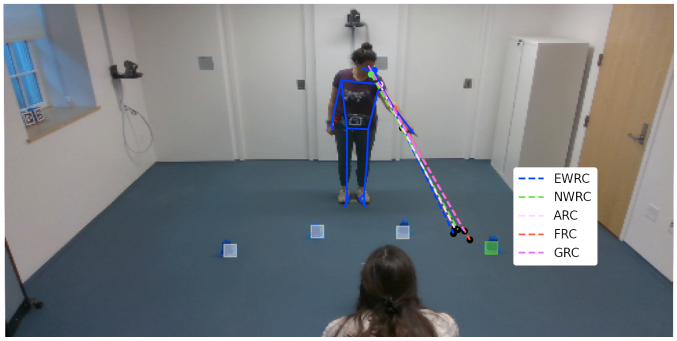

Fig. 2: Five pointing vectors on a sample image: eye-to-wrist, nose-to-wrist, shoulder-to-wrist, elbow-to-wrist, and eye gaze. The left wrist is used as the frame of reference.

B. Converting Human Body Pose to Pointing Vectors

Given the body pose of a person, we explore five different algorithms for computing a vector from the person's body pose. We 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.

Our work uses five pointing vectors, visualizations of which are shown in Figure 2:

1) Eye-to-wrist ray-cast (EWRC): Defined by a vector connecting the eye and wrist of the pointing arm.

2) Nose-to-wrist ray-cast (NWRC): Defined by a vector connecting the nose and wrist of the pointing arm.

3) Arm ray-cast (ARC): A ray-cast defined by a vector connecting the shoulder and wrist of the pointing arm.

4) Forearm ray-cast (FRC): A ray-cast defined by a vector connecting the elbow and wrist of the pointing arm.

5) Gaze ray-cast (GRC): To establish a corresponding gaze vector representing the general direction the user is looking at, we computed the normal vector to the plane passing through their left eye, right eye, and center of the mouth.

C. Evaluation Metrics

We evaluate the performance of our algorithms at resolving the pointing gestures that people produced, as well as the dog's performance as measured by touching the object. We manually annotate the frames with pointing gesture. Three metrics are used to evaluate average object selection performance:

1) Euclidean distance(lower is better): The Euclidean distance measures how closely the pointing ray intersects with the plane of objects, offset by the distance to the target object. This metric is exclusively used to assess human pointing accuracy.

2) Weighted accuracy(higher is better):

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

where A_i 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_i is the probability of a target being selected—calculated using the normalized inverse Euclidean distance.

3) Perplexity(PP)(lower is better): Perplexity quantifies how well the model performs, with a lower perplexity score indicating better predictive performance and less surprise at the actual object location. This metric is particularly useful for assessing the model's ability to interpret pointing gestures and determine the intended target among multiple predefined locations, thereby providing insight into the model's reliability and precision in practical applications.

For each item in the dataset, n, the object is in one of k predefined locations t_i and the distance from the pointing intersection location to each target d_j. Thus, we can compute the perplexity as a multinomial over the true location:

$$\mathcal{L}(t_i|d_1, ..., d_k) = \frac{d_i^{-1}}{\sum_{j=1}^{k} d_j^{-1}} \tag{2}$$

$$Perplexity(N) = exp\left\{-\frac{1}{|N|}\sum_{n\in N} \log \mathcal{L}(t_n|d_1, ..., d_k)\right\} \tag{3}$$

IV. 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 for 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 humans 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 at enabling a robot to resolve pointing gestures.

A. 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.

a) 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. Among the dogs, five were mixed breeds and one was a Golden Retriever.

b) Materials: The experimental setup comprises four upside-down cups placed equidistant in front of the dog. This setup helps minimize the likelihood of the dog approaching the closer target first. To minimize external device interference in the dogs' decision-making process, we utilized the Intel RealSense D435 camera to capture RGB-D image. Dog treats were used to motivate dogs to search. To address the potential confounding effect of the treat's smell, the cups were rubbed with treats to standardize this variable. Additionally, previous research [33] indicates that dogs are generally poor at localizing treats by smell alone in similar tasks.

c) Procedure: Before the pointing task, dog-human pairs completed two warm-up activities. First, dogs watched their guardians place a treat under a cup and retrieved it by touching the cup, repeated four times. Next, dogs practiced leaving and re-entering the room to find a hidden treat under a cup.

For the test trials, dogs were led out while the guardian hid a treat under one of four semi-randomized targets. The dog returned, and the guardian pointed to the hidden treat, allowing the dog to search. This was repeated for 12 trials per pair, totaling 72 recorded trials.

Afterward, 3 guardians pointed to the cups for a human experimenter. The setup was the same as in the human-dog trials, and this was repeated for 12 trials, for a total of 36 recorded trials.

B. 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, 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 [34]. The two locations closer to the human pointer tend to be chosen more frequently than those on the periphery. In our sample, 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 made by dogs 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 [35]. 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 guardian (meaning they have received lots of rewards directly from their guardian) causes dogs to prefer to search nearer to their guardian. 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.

C. Ray-cast Performance

Table Ia shows the performance with 95% confidence interval of our five different vectors for resolving pointing gestures. Our primary result is that most vectors perform

TABLE I: Performance on object selection task

(a) Performance on humans pointing for their dogs.

	Euclidean 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)*

(b) Performance on data of humans pointing for other humans.

	Euclidean 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)*

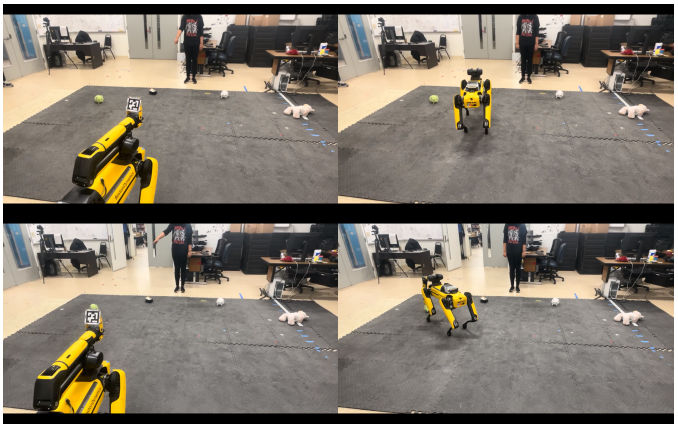

Fig. 3: Our system enables the robot to correctly fetch the object the human user is pointing at, such as the penguin plush (left) and green cat (right).

similarly, with the lowest-performing vector using gaze alone, which performs significantly worse than the other vectors. All methods significantly outperform dogs as measured by accuracy and perplexity, probably because dogs preferred cups nearer to their guardian. It is interesting to see that the eye-to-wrist vector has higher accuracy, but the shoulder-to-wrist vector (arm ray-cast) has the lowest PP. Given the confidence intervals, there might 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.

D. Human-Human Pointing Experiment

Table Ib shows the vectors' performance on the humans pointing for other humans data. As the human experimenter told the participants which cups to point to, there is no human performance on pointing gesture resolution to report on this data. There appears to be consistent performance between nose, eye, and shoulder-to-wrist vectors. The weighted 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, determined through a logarithmic base-2 approach and assuming a uniform distribution). 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. When pointing for dogs, humans use additional gestures like joint attention, exaggeration, repetition, and touching, less common in human-to-human pointing. The fatigue effect could also contribute to the result, as the data was collected after the participants have completed the pointing trials with their dog.

E. Spot Demonstration

We tested the shoulder-to-wrist vector on Spot[2] as it has the lowest perplexity score as shown in Figure 3. 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. Spot was able to follow the human pointer to correctly approach and select the indicated object from a set of four candidates.

V. Conclusion

In this paper, we evaluated various vectors for interpreting human pointing gestures to locations in the environment. We introduced a probabilistic observation model to utilize this vector for object search, a primary reason for human pointing to other entities. Our system was tested on a new dataset of humans pointing for their dogs and for other humans, comparing the performance of our autonomous algorithms to that of dogs. We intend to refine our experimental design to better distinguish between the performance of different vectors in resolving pointing gestures.

Future work can also consider using timecourse data of pointing information and Bayesian inference to predict the pointed objects. 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. In addition, we assumed that the person and their pointing gesture is within the robot's field of view. This assumption can be relaxed by employing additional cameras in the environment following Sprute et al. [36] or human detection and tracking methods [37, 38].

References

[1] G. Butterworth, "What is special about pointing in babies?" in *The development of sensory, motor and cognitive capacities in early infancy: From perception to*

[2]https://github.com/boston-dynamics/spot-sdk

cognition. Hove, England: Psychology Press/Erlbaum (UK) Taylor & Francis, 1998, pp. 171–190.

[2] M. Wnuczko and J. M. Kennedy, "Pivots for pointing: Visually-monitored pointing has higher arm elevations than pointing blindfolded." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 37, no. 5, p. 1485, 2011.

[3] B. I. Bertenthal, T. W. Boyer, and S. Harding, "When do infants begin to follow a point?" *Developmental Psychology*, vol. 50, no. 8, pp. 2036–2048, 2014.

[4] B. Agnetta, B. Hare, and M. Tomasello, "Cues to food location that domestic dogs (Canis familiaris) of different ages do and do not use," *Animal Cognition*, vol. 3, no. 2, pp. 107–112, 2000.

[5] A. Miklösi, R. Polgárdi, J. Topál, and V. Csányi, "Use of experimenter-given cues in dogs," *Animal Cognition*, vol. 1, no. 2, pp. 113–121, 1998.

[6] B. Hare, M. Brown, C. Williamson, and M. Tomasello, "The Domestication of Social Cognition in Dogs," *Science*, vol. 298, no. 5598, p. 1634, 2002.

[7] K. Soproni, A. Miklósi, J. Topál, and V. Csányi, "Comprehension of human communicative signs in pet dogs (Canis familiaris)," *Journal of Comparative Psychology*, vol. 115, no. 2, pp. 122–126, Jun. 2001.

[8] E. E. Bray, G. E. Gnanadesikan, D. J. Horschler, K. M. Levy, B. S. Kennedy, T. R. Famula, and E. L. MacLean, "Early-emerging and highly heritable sensitivity to human communication in dogs," *Current Biology*, vol. 31, no. 14, pp. 3132–3136.e5, Jul. 2021.

[9] J. Riedel, K. Schumann, J. Kaminski, J. Call, and M. Tomasello, "The early ontogeny of human-dog communication," *Animal Behaviour*, vol. 75, no. 3, pp. 1003–1014, 2008, place: Netherlands Publisher: Elsevier Science.

[10] D. Whitney, M. Eldon, J. Oberlin, and S. Tellex, "Interpreting Multimodal Referring Expressions in Real Time," in *IEEE International Conference on Robotics and Automation*. ICRA, 2016.

[11] D. Whitney, E. Rosen, J. MacGlashan, L. L. Wong, and S. Tellex, "Reducing Errors in Object-Fetching Interactions through Social Feedback," in *IEEE International Conference on Robotics and Automation*. ICRA, 2017, pp. 1006–1013.

[12] T. Obo, R. Kawabata, and N. Kubota, "Cooperative human-robot interaction based on pointing gesture in informationally structured space," in *World Automation Congress (WAC)*. IEEE, 2018, pp. 1–5.

[13] A. Ekrekli, A. Angleraud, G. Sharma, and R. Pieters, "Co-speech gestures for human-robot collaboration," *arXiv preprint arXiv:2311.18285*, 2023.

[14] S. Constantin, F. I. Eyiokur, D. Yaman, L. Bärmann, and A. Waibel, "Interactive Multimodal Robot Dialog Using Pointing Gesture Recognition," in *European Conference on Computer Vision*. Springer, 2022, pp. 640–657.

[15] J. L. Taylor and D. I. McCloskey, "Pointing," *Behavioural Brain Research*, vol. 29, pp. 1–5, 1988.

[16] S. Abidi, M. Williams, and B. Johnston, "Human pointing as a robot directive," in *8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2013, pp. 67–68.

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

[18] O. Herbort and W. Kunde, "Spatial (mis-)interpretation of pointing gestures to distal referents," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 42, no. 1, pp. 78–89, 2016.

[19] M. Tölgyessy, M. Dekan, F. Duchoň, J. Rodina, P. Hubinskỳ, and L. Chovanec, "Foundations of visual linear human–robot interaction via pointing gesture navigation," *International Journal of Social Robotics*, vol. 9, pp. 509–523, 2017.

[20] Z. Hu, Y. Xu, W. Lin, Z. Wang, and Z. Sun, "Augmented Pointing Gesture Estimation for Human-Robot Interaction," in *IEEE International Conference on Robotics and Automation*. ICRA, 2022, pp. 6416–6422.

[21] K. Nickel and R. Stiefelhagen, "Pointing gesture recognition based on 3d-tracking of face, hands and head orientation," in *Proceedings of the 5th international conference on Multimodal interfaces*, 2003, pp. 140–146.

[22] J. Perez-Osorio, H. J. Müller, E. Wiese, and A. Wykowska, "Gaze Following Is Modulated by Expectations Regarding Others' Action Goals," *PLOS ONE*, vol. 10, p. e0143614, 2015.

[23] S. Mayer, V. Schwind, R. Schweigert, and N. Henze, "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*, 2018, pp. 1–13.

[24] A. Kranstedt, A. Lücking, T. Pfeiffer, H. Rieser, and I. Wachsmuth, "Deixis: How to determine demonstrated objects using a pointing cone," in *Gesture in Human-Computer Interaction and Simulation. GW 2005. Lecture Notes in Computer Science*, ser. Lecture Notes in Computer Science, S. Gibet, N. Courty, and J.-F. Kamp, Eds. Springer, Berlin, Heidelberg, 2006, vol. 3881. [Online]. Available: https://doi.org/10.1007/11678816_34

[25] B. Sodian and C. Thoermer, "Infants' Understanding of Looking, Pointing, and Reaching as Cues to Goal-Directed Action," *Journal of Cognition and Development*, vol. 5, no. 3, pp. 289–316, 2004.

[26] K. Liebal, T. Behne, M. Carpenter, and M. Tomasello, "Infants use shared experience to interpret pointing gestures," *Developmental Science*, vol. 12, no. 2, pp. 264–271, Mar. 2009.

[27] A. L. Woodward and J. J. Guajardo, "Infants' understanding of the point gesture as an object-directed action," *Cognitive Development*, vol. 17, no. 1, pp. 1061–1084, 2002.

[28] O. Herbort, L.-M. Krause, and W. Kunde, "Perspective

determines the production and interpretation of pointing gestures," *Psychonomic Bulletin & Review*, vol. 28, no. 2, pp. 641–648, Apr. 2021.

[29] S. Mayer, K. Wolf, S. Schneegass, and N. Henze, "Modeling distant pointing for compensating systematic displacements," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015, pp. 4165–4168.

[30] S. Yoon, Y. Kim, C. R. Ahn, and M. Park, "Challenges in Deictic Gesture-Based Spatial Referencing for Human-Robot Interaction in Construction," in *ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction*, vol. 38. IAARC Publications, 2021, pp. 491–497.

[31] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "Blazepose: On-device real-time body pose tracking," 2020.

[32] J. Wang and E. Olson, "AprilTag 2: Efficient and robust fiducial detection," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 4193–4198.

[33] D. Alberghina, E. Bray, D. Buchsbaum, S.-E. Byosiere, J. Espinosa, G. Gnanadesikan, C.-N. A. Guran, E. Hare, D. Horschler, L. Huber, V. A. Kuhlmeier, E. MacLean, M. H. Pelgrim, B. Perez, D. Ravid-Schurr, L. Rothkoff, C. Sexton, Z. Silver, and J. R. Stevens, "Manydogs project: A big team science approach to investigating canine behavior and cognition," *Comparative Cognition and Behavior Reviews*, vol. 18, pp. 59–77, 2023. [Online]. Available: https://manydogsproject.github.io/publications.html

[34] G. Lakatos, M. Gácsi, J. Topál, and Á. Miklósi, "Comprehension and utilisation of pointing gestures and gazing in dog–human communication in relatively complex situations," *Animal Cognition*, vol. 15, no. 2, pp. 201–213, Mar. 2012.

[35] B. Hare and M. Tomasello, "Domestic dogs (Canis familiaris) use human and conspecific social cues to locate hidden food," *Journal of Comparative Psychology*, vol. 113, no. 2, pp. 173–177, 1999, place: US Publisher: American Psychological Association.

[36] D. Sprute, R. Rasch, A. Pörtner, S. Battermann, and M. König, "Gesture-based object localization for robot applications in intelligent environments," in *2018 14th International Conference on Intelligent Environments (IE)*. IEEE, 2018, pp. 48–55.

[37] J. Zhou and J. Hoang, "Real time robust human detection and tracking system," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops*. IEEE, 2005, pp. 149–149.

[38] M. J. Islam, J. Hong, and J. Sattar, "Person-following by autonomous robots: A categorical overview," *The International Journal of Robotics Research*, vol. 38, no. 14, pp. 1581–1618, 2019.