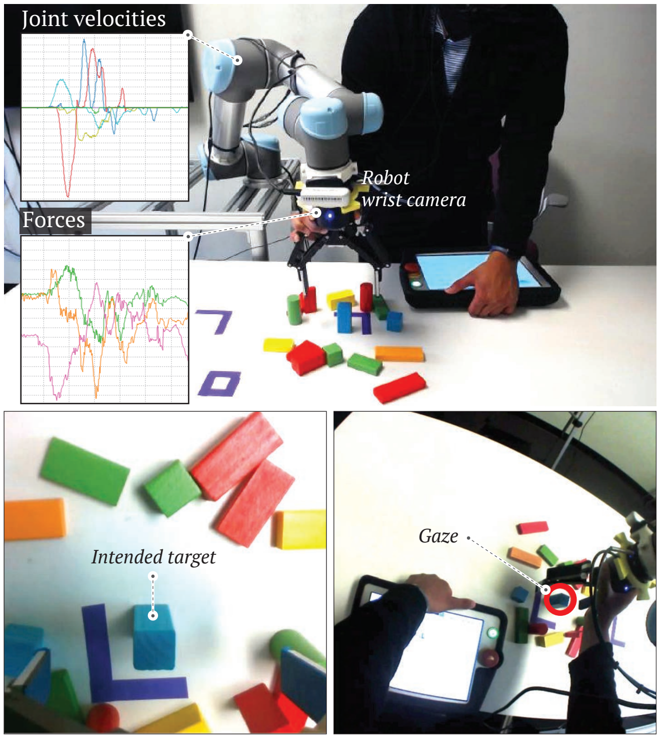
Multimodal Robot Programming by Demonstration: A Preliminary Exploration

Gopika Ajaykumar and Chien-Ming Huang
Department of Computer Science
Johns Hopkins University
{gopika,cmhuang}@cs.jhu.edu



Robot first-person view

Human first-person view

Fig. 1. We explore multimodal data from programmers' kinesthetic demonstrations and argue the importance of incorporating natural multimodal cues from human teachers to enable effective robot learning and assisted end-user robot programming.

I. Introduction

Recent years have seen a growth in the number of industrial robots working closely with end-users such as factory workers [14]. This growing use of collaborative robots has been enabled in part due to the availability of end-user robot programming methods that allow users who are not robot programmers to teach robots task actions [2] . *Programming by Demonstration* (PbD) is one such end-user programming method that enables users to bypass the complexities of specifying robot motions using programming languages by instead demonstrating the desired robot behavior [21, 8]. Demonstrations are often provided by physically guiding the robot through the motions required for a task action in a process known as *kinesthetic teaching*.

Kinesthetic teaching enables users to directly demonstrate task behaviors in the robot's configuration space, making it a popular end-user robot programming method for collaborative robots known for its low cognitive burden [6, 19, 15]. However, because kinesthetic teaching restricts the programmer's teaching to motion demonstrations, it fails to leverage information from other modalities that humans naturally use when providing physical task demonstrations to one other, such as gaze and speech (e.g., [16]). Incorporating multimodal information into the traditional kinesthetic programming workflow has the potential to enhance robot learning by highlighting critical aspects of a program [16], reducing ambiguity [22], and improving situational awareness [18] for the robot learner and can provide insight into the human programmer's intent and difficulties [22]. In this extended abstract, we describe a preliminary study on multimodal kinesthetic demonstrations and future directions for using multimodal demonstrations to enhance robot learning and user programming experiences.

II. Preliminary Exploration of Multimodal Robot Programming by Demonstration

We conducted a preliminary exploration of multimodal robot programming by demonstration with 11 users, which included users with and without prior robot programming experience. Participants were instructed to demonstrate six tasks to the robot, which variously involved pick-and-place actions, stacking, pouring, and insertion (Fig. 2), through kinesthetic teaching and by speaking aloud. For each demonstration, we recorded the robot's first-person view collected from its wrist camera and the positions, velocities, and forces of its joints (Fig. 1). In addition, we recorded the user's gaze and first-person view collected using a Pupil Invisible gaze tracker, their speech, and their input forces on the robot's end effector using a force-torque sensor. We also recorded a third-person view of the programming process using a stationary webcam. Below, we describe key observations from our initial exploration of multimodal PbD, with a focus on programmers' narrations during kinesthetic teaching.

A. Highlighting Important Aspects for Learning

Prior work on programming by demonstration has relied on approaches such as clustering multiple task demonstrations (e.g., [3]) or segmenting one-shot demonstrations (e.g., [7]) to develop skill models that capture the essential robot motions and configurations required for a task. We found that having

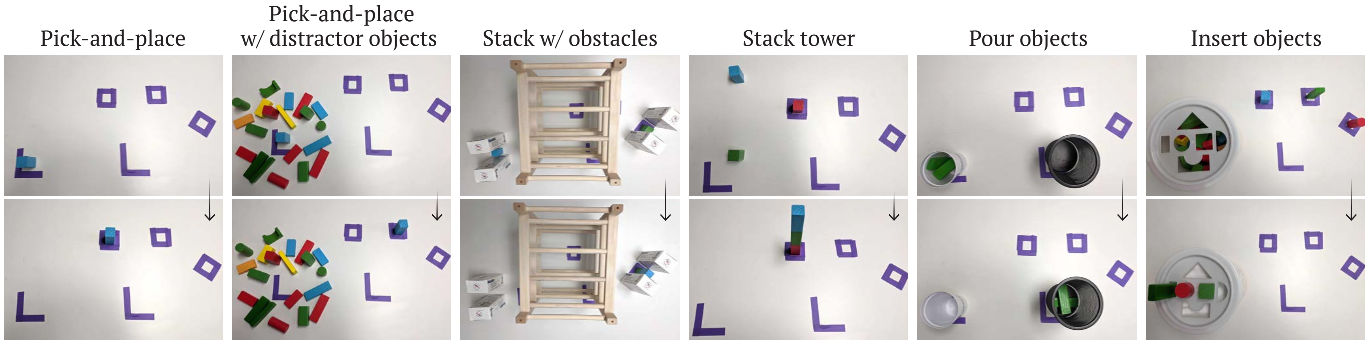
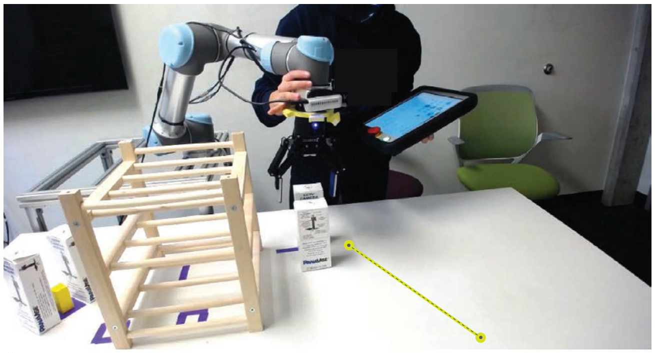


Fig. 2. Participants demonstrated six manipulation tasks in our study: pick-and-place, pick-and-place with distractor objects, stack with obstacles, stack tower, pour objects, and insert objects.

programmers speak aloud during kinesthetic teaching can help with the process of capturing critical low-level and high-level aspects of a task within a single demonstration.

Participants often verbalized how the robot should be positioned with respect to task-relevant objects (e.g., *"I'm going to move the gripper over to the block so the gripper's positioned directly above it"* (P6), *"I'm gonna move the robot over and move the gripper parallel to the table but in the other way"* (P3)), as well as stating the direction (e.g., *"So I'm moving the robot over to the right, um, taking it right above as close as I can"* (P8)) and speed (e.g., *"I'm gonna position the gripper over the block and then slowly close the gripper"* (P6)) of demonstrated motions. In addition, participants occasionally repeated particular phrases in succession to indicate the length of time a motion should continue (e.g., *"close close close close"* (P2) throughout the duration of a grip, *"Keep going. Keep going. Keep going."* (P5) throughout a continuous trajectory). Participants also indicated key robot configurations within a demonstrated task using phrases such as *"hold here"* (P1) and *"record"* (P11) and described the level of care required for particular motions (e.g., *"Lower the object very carefully"* (P11), *"I need to gently pick it up without hurting the other two paper boxes"* (P10)). Participants' verbal descriptions helped contextualize the motion trajectories and waypoints traditionally used for PbD in terms of the task environment and constraints and may be used to associate demonstration waypoints with reference frames (e.g., [4]) (Fig. 3).



"Uh okay, so I need to get the gripper to be over the object... so that when the gripper closes, it's not going to knock over the boxes."

Fig. 3. Participants described their actions in terms of objects in the environment when narrating their demonstration, which can enable demonstrated motions to be contextualized within the task environment, including in terms of constraints such as obstacles.

In addition to pinpointing aspects of the motion that are important for learning, participants' speech also focused on higher level characteristics of the demonstration related to task order, demonstration goals, and error avoidance. Participants indicated the sequence of actions required for the task (e.g., *"Let it go in a little bit, and then drop it"* (P9)), along with preconditions necessary for any of the task steps (e.g., *"Once it's about in line with the hole I'm going to click open and then try again"* (P8)). Some participants began their demonstration with a summary of the overall task and goal (e.g., *"We're gonna put the figures, the cubes into its corresponding shape, starting with the blue one, and then the yellow, and then the red"* (P7)). Similarly, participants' gaze occasionally "summarized" the task in terms of relevant objects at the beginning of the demonstration (Fig. 4). Participants also provided warnings about potential errors to avoid throughout their demonstrations, especially related to physical obstacles (e.g., *"Make sure that when the gripper closes it's not gonna hit the table"* (P6), *"Bring the robot back up slowly to avoid colliding with the cardboard boxes"* (P1)). Participants' gaze cues also focused on potential obstacles, which was in line with findings from previous work on using human gaze for

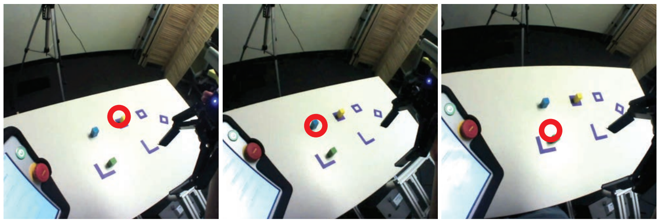


Fig. 4. Participants looked at all relevant task objects at the beginning of the demonstration, providing a "summary" of the object targets of the demonstrated task. The center of the red circle indicates the participant's gaze.

assisted teleoperation [5]. Multimodal kinesthetic teaching enables identification of relevant task targets, obstacles, and actions for each portion of a demonstration.

B. Pinpointing Program Suboptimalities

While some demonstration errors correspond with clear signals, such as the robot's emergency stop triggering, others may be more difficult to identify. Multimodal cues can reveal which parts of a programmer's demonstration are suboptimal or erroneous in terms of task efficiency or probability of success. For example, participants often indicated when the robot reached a joint limit during the demonstration (e.g., *"The gripper can't really go down any further"* (P6), *"Moving the elbow is going to be hard. Okay, I think I'm at the max of the joint."* (P4)), when they can't grip an object from the optimal approach (e.g., *"I'll have to pick it up from, uh, like at an angle"* (P4), *"I will probably just try a bad angle like this"* (P5)), or when they didn't get a good grip on the object with the end effector (e.g., *"Now I'm gonna use the close gripper [command] to grab the block, ah that didn't work so I'm gonna open it."* (P1), *"Oop, bad grip."* (P4)). These motion suboptimalities could also be identified by discrepancies between the force input by the user and the resulting positional change of the robot's joints (Fig. 5). Suboptimal actions in a demonstration often corresponded with participants expressing uncertainty of the actions' success (e.g., *"It's at an angle so I'm not sure if it'll go in."* (P3), *"If it doesn't stumble, I think it will work now"* (P4)) or laughing. In many cases, participants directly indicated when their demonstration failed or when an error occurred in the course of a demonstration (e.g., *"Oops, I put the arm too far down so I'm gonna move it up a bit"* (P6), *"The angle was a little bit too much so it bounced off position"* (P7)). On the other hand, participants also provided positive feedback when the robot was at a good configuration during the demonstration (e.g., *"That looks like a good position to grab"* (P7)) or when the demonstration was going well by giving quick responses such as "good" or "perfect" after a task step. Overall, multimodal cues helped distinguish which portions of a demonstration indicated what the robot should do and which did not.

C. Revealing User Challenges and Intent

Multimodal cues can indicate challenges users are facing during kinesthetic teaching. Participants frequently indicated when they were having trouble with an aspect of the kinesthetic demonstration (e.g., *"Little hard to find the right angle"* (P6), *"Struggling a little bit with how far the robot can go"* (P4)). Participants also asked questions, directed at themselves, while narrating the demonstration when they encountered a challenge (e.g., *"How do you even grab this?"* (P9), *"Can I open it? I should open it, right?"* (P4)). When encountering difficulties with a motion, such as reaching the correct alignment or grip, participants' gaze tended to shift rapidly between targets, such as the robot's joint and the task object (Fig. 6), which aligns with previous work indicating that users tend to look at problematic joints when experiencing challenges in moving a robot [5]. Behavioral cues indicating user frustration and fatigue included sighing and frequent shifts of body posture (e.g., bending, changing contact point with robot) (Fig. 5).

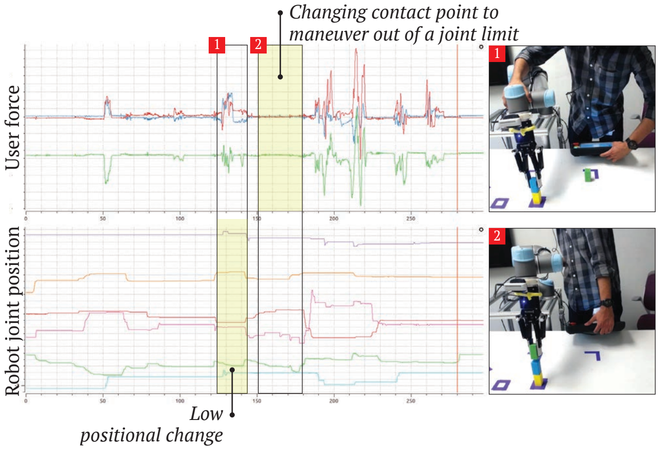


Fig. 5. Common motion suboptimalities and difficulties encountered by participants included: (1) encountering a joint limit, which resulted in low positional change in response to user input force and (2) having to change their contact point away from the robot's wrist to maneuver out of a joint limit.

In addition to difficulties, users also described the reasoning behind their actions during kinesthetic teaching (e.g., *"Um, so if I were to do it right now, they would fall off to the side, so I want to move the arm a little bit closer to itself."* (P7), *"So I'm first moving the robot over and twisting it so that I can close on the edge of the cup"* (P8)). Participants also clarified which portions of a demonstration were allocated towards trial and error or brainstorming purposes (e.g., *"Okay, let's see how we can move these things. Okay, this moves like that. This thing should move this way. Okay, that doesn't tip"* (P9), *"I'm taking a pause to find out the right angle."* (P4)). Multimodal cues provide insight into programmers' intent and challenges and can signal which parts of a demonstration are intended for human learning (e.g., becoming familiar with the robot's capabilities and constraints) rather than for robot learning (e.g., functional motions).

III. Implications of Multimodal Robot Programming by Demonstration

For robots to fully leverage the rich set of information available in a human teacher's demonstration, they must move beyond solely considering motion aspects of a demonstration. Demonstrations that include the multimodal cues that humans naturally use in teaching, such as gaze and speech, can provide a range of information, from task sequence and goals to programming difficulties, and can pave the way for improved robot learning and easier robot programming.

A. Robot Learning with Multimodal PbD

Just as humans make use of multiple modalities such as vision, speech, and sound to obtain a coherent understanding of their environment [9], robots may obtain a more complete

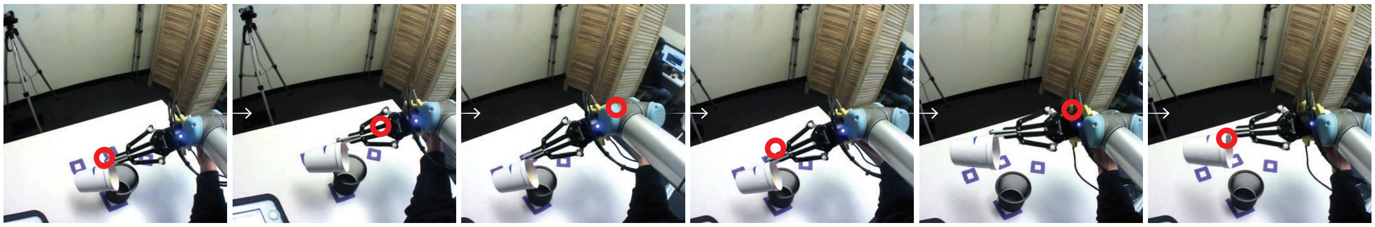


Fig. 6. Rapid gaze shifts from robot joints to target objects tended to indicate difficulties with a particular motion such as reaching an optimal end effector alignment.

understanding of how a demonstration is situated within a task environment by taking into account multiple modalities in a demonstration, including the human teacher's speech, which can provide semantic information on task objects, and gaze, which can highlight which environmental factors are relevant for the task at hand. Multimodal cues from a human teacher's demonstration can also help robots obtain an understanding of the principal steps involved in a task. In line with prior work indicating humans teach robots using structured processes [20], our study participants frequently stated task steps in terms of action preconditions and effects, which can be used by robots for developing task plans automatically based on human demonstrations (e.g., [12, 17]).

Multimodal PbD can help robots utilize information more effectively within a single demonstration. Multimodal cues can indicate the human demonstrator's focus, which can in turn reveal which aspects of a demonstration are critical to the task being learned (e.g., speed, force, particular configurations, action preconditions, obstacles) and which aspects the robot may have more freedom to stray from (e.g., approach angle for gripping an object). Because multimodal cues can reveal programmer intent and include real-time feedback on a demonstration, they can help robots distinguish between good and failed or suboptimal demonstrations and effectively use a demonstration in its entirety by using portions of a demonstration that correspond to positive teacher feedback to learn what to do and portions of a demonstration that corresponded to negative teacher feedback, program suboptimalities, or programming difficulties to learn what not to do (e.g., [10]). By taking advantage of additional information on how to successfully perform a task and avoid erroneous behaviors from a single demonstration, effective robot learning may occur with a smaller quantity of demonstrations from a human teacher.

B. Assisted Robot Programming with Multimodal PbD

Multimodal cues can signal when programmers are encountering difficulties during robot programming. Because interaction modalities such as gaze and speech precede motion during kinesthetic teaching, multimodal behavioral signatures indicating programming difficulties can be identified and difficulties can be prevented early on in a demonstration. Developing programming assistance triggered by multimodal cues, such as autocompletion for pick-and-place tasks or help maneuvering away from a joint limit, may improve the programmer's experience during robot programming and may reduce programmer challenges stemming from the high physical workload involved in kinesthetic teaching (e.g., [19, 1]). Assistance triggered by the programmer's multimodal data may also help optimize users' demonstrations by reducing the amount of robot motion unrelated to the task at hand within a demonstration (e.g., alignment motions or wrangling the robot into a specific configuration).

IV. Conclusion and Future Work

In this abstract, we presented initial observations into the range of information that narration, gaze, motion, and force data can provide for task learning and assisted robot programming. While prior work has investigated multimodal PbD that takes into account users' motion demonstrations and speech (e.g., [13]), we believe a multimodal learning approach that takes into account the programmer's gaze and speech and considers motion and force cues indicating user challenges can enable better robot understanding of human demonstrations and more personalized assistance to facilitate easier robot programming by demonstration. Our future work will involve developing computational models that predict when the programmer is experiencing difficulties, such as challenges in maneuvering the gripper to be in line with the target object, based on behavioral signatures such as those observed in our initial study (e.g., Figs. 5 and 6). By drawing off of natural multimodal human teaching processes, we aim to minimize the burden of the human teacher in providing optimal demonstrations while expanding the available resources for the robot to understand a teacher's demonstration.

Ethical Impact Statement

Modeling users' multimodal cues to improve robot learning and provide online assistance during kinesthetic teaching can benefit users by reducing the user burden in providing large quantities of high-quality demonstrations to the robot learner and in performing complicated maneuvering during kinesthetic teaching. This can help further lower the barriers in robot programming for everyday users of collaborative robots. However, such an approach may also involve risks in the long term. Prior work has suggested that end-users without professional programming experience may be more likely to rely on shortcuts and workarounds during programming [11]. Multimodal data-driven online programming assistance and multimodal robot learning that is robust to program suboptimalities may encourage programmer overreliance on robot learning algorithms and system assistance, possibly encouraging the practice of sloppy programming behaviors from end-users in the long term. Furthermore, online programming assistance based on multimodal cues will involve shifting control over program specification from the human-teacher to the robot learner. While such an assisted programming approach may result in more optimal and robust programs, it could also overly disrupt users' programming workflows or their mental models on how their program works.

Our goal with this work is to use users' multimodal cues as a means to improve users' experiences with robot programming by minimizing difficulties in kinesthetic teaching. However, we acknowledge that data-driven programming assistance may not always be warranted or desired and could lead to programmer overtrust and automation bias in the long-term. We encourage further work into understanding end-users' perceptions of personalized assistance in real-world scenarios. In particular, future work that builds off of multimodal data-driven robot programming should examine whether users are able to recover from failed online assistance based on misconstrued multimodal cues and whether multimodal programming by demonstration encourages suboptimal teaching from users in the long term.

Acknowledgments

This work was supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1746891 and the Nursing-Engineering Joint Fellowship from Johns Hopkins University.

References

[1] Gopika Ajaykumar and Chien-Ming Huang. User needs and design opportunities in end-user robot programming. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 93–95, 2020.

[2] Gopika Ajaykumar, Maureen Steele, and Chien-Ming Huang. A survey on end-user robot programming. *arXiv preprint arXiv:2105.01757*, 2021.

[3] Baris Akgun, Maya Cakmak, Karl Jiang, and Andrea L Thomaz. Keyframe-based learning from demonstration. *International Journal of Social Robotics*, 4(4):343–355, 2012.

[4] Sonya Alexandrova, Maya Cakmak, Kaijen Hsiao, and Leila Takayama. Robot programming by demonstration with interactive action visualizations. In *Robotics: science and systems*. Citeseer, 2014.

[5] Reuben M Aronson and Henny Admoni. Eye gaze for assistive manipulation. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 552–554, 2020.

[6] Daniel Bambušsek, Zdeněk Materna, Michal Kapinus, Vítězslav Beran, and Pavel Smrž. Combining interactive spatial augmented reality with head-mounted display for end-user collaborative robot programming. In *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1–8. IEEE, 2019.

[7] Riccardo Caccavale, Matteo Saveriano, Alberto Finzi, and Dongheui Lee. Kinesthetic teaching and attentional supervision of structured tasks in human–robot interaction. *Autonomous Robots*, 43(6):1291–1307, 2019.

[8] Sonia Chernova and Andrea L Thomaz. Robot learning from human teachers. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 8(3):1–121, 2014.

[9] Marc O Ernst and Heinrich H Bülthoff. Merging the senses into a robust percept. *Trends in cognitive sciences*, 8(4):162–169, 2004.

[10] Daniel H Grollman and Aude G Billard. Robot learning from failed demonstrations. *International Journal of Social Robotics*, 4(4):331–342, 2012.

[11] Warren Harrison. From the editor: The dangers of end-user programming. *IEEE software*, 21(4):5–7, 2004.

[12] Ying Siu Liang, Damien Pellier, Humbert Fiorino, and Sylvie Pesty. A framework for robot programming in cobotic environments: first user experiments. In *Proceedings of the 3rd International Conference on Mechatronics and Robotics Engineering*, pages 30–35, 2017.

[13] Anahita Mohseni-Kabir, Changshuo Li, Victoria Wu, Daniel Miller, Benjamin Hylak, Sonia Chernova, Dmitry Berenson, Candace Sidner, and Charles Rich. Simultaneous learning of hierarchy and primitives for complex robot tasks. *Autonomous Robots*, 43(4):859–874, 2019.

[14] International Federation of Robotics. World robotics report 2020, 2020.

[15] SK Ong, AWW Yew, NK Thanigaivel, and AYC Nee. Augmented reality-assisted robot programming system for industrial applications. *Robotics and Computer-Integrated Manufacturing*, 61:101820, 2020.

[16] Julia Oppenheim, Jindan Huang, Isabel Won, and Chien-Ming Huang. Mental synchronization in human task demonstration: Implications for robot teaching and learning. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, pages 470–474, 2021.

[17] Mikkel Rath Pedersen and Volker Krüger. Gesture-based extraction of robot skill parameters for intuitive robot programming. *Journal of Intelligent & Robotic Systems*, 80(1):149–163, 2015.

[18] Svetlin Penkov, Alejandro Bordallo, and Subramanian Ramamoorthy. Physical symbol grounding and instance learning through demonstration and eye tracking. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5921–5928. IEEE, 2017.

[19] Camilo Perez Quintero, Sarah Li, Matthew KXJ Pan, Wesley P Chan, HF Machiel Van der Loos, and Elizabeth Croft. Robot programming through augmented trajectories in augmented reality. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1838–1844. IEEE, 2018.

[20] Preeti Ramaraj, Charles L Ortiz Jr, Matthew Klenk, and Shiwali Mohan. Unpacking human teachers' intentions for natural interactive task learning. *arXiv preprint arXiv:2102.06755*, 2021.

[21] Harish Ravichandar, Athanasios S Polydoros, Sonia Chernova, and Aude Billard. Recent advances in robot learning from demonstration. *Annual Review of Control, Robotics, and Autonomous Systems*, 3:297–330, 2020.

[22] Akanksha Saran, Elaine Schaertl Short, Andrea Thomaz, and Scott Niekum. Understanding teacher gaze patterns for robot learning. In *Conference on Robot Learning*, pages 1247–1258. PMLR, 2020.