See What I See: Enabling User-Centric Robotic Assistance Using First-Person Demonstrations

Yeping Wang
wyeping1@jhu.edu
Johns Hopkins University
Baltimore, MD 21218, USA

Gopika Ajaykumar
gopika@cs.jhu.edu
Johns Hopkins University
Baltimore, MD 21218, USA

Chien-Ming Huang
cmhuang@cs.jhu.edu
Johns Hopkins University
Baltimore, MD 21218, USA

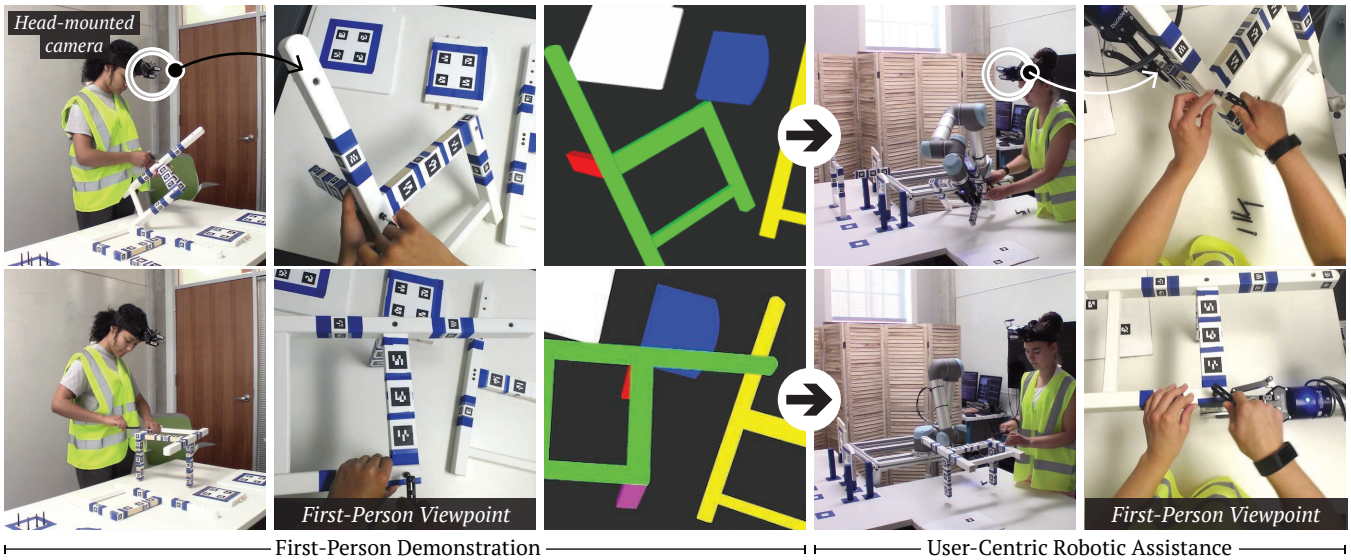

Figure 1: We explore how first-person demonstrations may capture natural behavioral preferences for task performance and how they can be utilized to enable user-centric robotic assistance in human-robot collaborative assembly tasks.

Abstract

We explore *first-person demonstration* as an intuitive way of producing task demonstrations to facilitate user-centric robotic assistance. First-person demonstration directly captures the human experience of task performance via head-mounted cameras and naturally includes productive viewpoints for task actions. We implemented a perception system that parses natural first-person demonstrations into task models consisting of sequential task procedures, spatial configurations, and unique task viewpoints. We also developed a robotic system capable of interacting autonomously with users as it follows previously acquired task demonstrations. To evaluate the effectiveness of our robotic assistance, we conducted a user study contextualized in an assembly scenario; we sought to determine how assistance based on a first-person demonstration (*user-centric assistance*) versus that informed only by the cover image of the official assembly instruction (*standard assistance*) may shape users' behaviors and overall experience when working alongside a collaborative robot. Our results show that participants felt that their robot partner was more collaborative and considerate when it provided user-centric assistance than when it offered only standard assistance. Additionally, participants were more likely to exhibit unproductive behaviors, such as using their non-dominant hand, when performing the assembly task without user-centric assistance.

HRI '20, March 23–26, 2020, Cambridge, United Kingdom.

ACM ISBN 978-1-4503-6746-2/20/03…$15.00
https://doi.org/10.1145/3319502.3374820

CCS Concepts

• **Human-centered computing** → **Collaborative interaction**; • **Computer systems organization** → **Robotics**; • **Computing methodologies** → *Vision for robotics.*

Keywords

Human-Robot Interaction; First-Person Demonstration; Programming by Demonstration; Collaborative Robotics

ACM Reference Format:
Yeping Wang, Gopika Ajaykumar, and Chien-Ming Huang. 2020. See What I See: Enabling User-Centric Robotic Assistance Using First-Person Demonstrations. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI '20), March 23–26, 2020, Cambridge, United Kingdom..* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3319502.3374820

1 INTRODUCTION

As we continue to develop collaborative robots to assist people at work and in the home, it is important to ensure that everyday users can easily customize, or "program," their robotic assistance to meet their needs and that they can comfortably interact with their robot assistants. These design goals, if met, will lead to enhanced user experience and long-term adoption of such assistance. In this work, we explore *first-person demonstration* as an alternative way of robot programming by demonstration, in which users program a collaborative robot by simply performing the task. In this method, a user's demonstrated program is recorded by head-mounted cameras that directly capture rich task contexts and natural user behavior during task performance (e.g., how their hands are interacting with various task objects); first-person viewpoints encapsulate a wealth of subtle behavioral preferences that collaborative robots can leverage to provide more user-friendly support.

We contextualized our exploration of first-person demonstration and its use in human-robot collaboration within the domain of furniture assembly, which often involves common manipulation tasks (Figure 1). In our exploration, we first developed a perception system capable of parsing a natural first-person demonstration into an operational task model. We also developed an autonomous robotic system that can utilize an acquired task model to assist people in a user-friendly way; specifically, the robot presents assembly parts to users similarly to how they would perform the task themselves. We then conducted a user study to evaluate our systems and examine how first-person demonstration may help enable user-centric robotic assistance that can positively shape human-robot collaboration.

In the next section, we review the relevant prior research that motivates this work. We then describe our method for representing demonstrated tasks from first-person demonstrations and our implementation of an autonomous robotic system that provides user-centric assistance (Section 3). In Section 4, we describe the user study that sought to evaluate the effectiveness of user-centric robotic assistance and to explore how such assistance may shape user experience and behavior during human-robot collaborations. Finally, we conclude this paper with a discussion of our findings and the limitations of this work in Section 5.

2 RELATED WORK

We review relevant prior research from three areas: programming by demonstration, task representation, and first-person vision.

2.1 Robot Programming by Demonstration

Research on robot programming by demonstration (PbD) [4, 8, 9], or learning from demonstration, aims to reduce barriers to authoring custom robot skills for people with diverse backgrounds and needs. To this end, prior research has explored various authoring methods, including kinesthetic teaching [1, 20], vision-based demonstrations [15, 53, 55], teleoperative demonstrations in virtual reality [56], and behavioral instructions [24, 26, 33, 46, 47, 50], as well as accessible programming interfaces that involve visual programming [3, 17, 22, 42] and situated programming [16, 45]. In addition to the exploration of interfaces and methods for skill authoring, prior research has also investigated how demonstrated skills may be applied to a variety of task configurations [2, 15, 27, 52, 55].

This work explores an alternative method for authoring robot skills. Arguably, the most simple and effective method of task demonstration is to simply perform the task. In this work, we explore *first-person demonstration*, in which head-mounted cameras capture exactly how a human demonstrator performs a task. We note that first-person demonstration is different from showing a task to a learner via a teaching process (e.g., [25, 55]), which requires consideration and estimation of the learner's perspective during the teaching of the task. In contrast, first-person demonstration allows a learner to directly channel a teacher's perspective.

Similarly, Yu et al. sought to capture the first-person perspective by fixing a camera behind a human demonstrator [55]. However, their setup was unable to acquire dynamic information about the demonstrator's head movement, which approximates the demonstrator's attentional focus during task demonstration. The most similar work to ours is perhaps that of Lee and Ryoo [28]; in their work, a robot learned collaborative behaviors from example videos in which humans executed the same collaborative behavior from a first-person viewpoint. However, in their robotic system implementation, dynamic viewpoint changes from the human demonstration were not emphasized in the robot reproduction of the collaborative behavior. In contrast to these prior works, we focus on dynamic first-person demonstration of complex manipulation tasks.

2.2 Task Representation

To enable effective human-robot collaboration, various task representations have been explored, including Finite State Automaton [38], Hierarchical Task Network (HTN) [19], and Markov Decision Process [41]. We provide a brief discussion of the HTN due to its similarity to our proposed FEAsT model detailed in Section 3.1. An HTN represents a tree structure where a leaf node denotes a primitive action or goal and a parent node denotes the abstraction or composition of its children [35]. Prior research on human-robot collaboration has investigated variations of HTNs (e.g., [10, 19]) and employed HTNs to enable interactive learning from demonstration [35], to develop transparent task planners [43], and to generate collaborative plans [34]. In this paper, we present a similar hierarchical structure to represent an egocentric assembly task demonstration, where first-person viewpoints are an important degree of information. We additionally present an algorithm to recover viewpoint information from the hierarchical structure.

2.3 First-Person Vision

First-person vision (FPV), or egocentric vision, naturally captures the region of human attention [31]. Applications of FPV have included activity recognition [31, 32, 44, 51, 54], object detection [29, 32], activity-based salient object detection [7], motion prediction [6, 48], and joint attention estimation for social scenarios [49]. In addition to applications involving pure image analysis, FPV has also been used as a command tool for disabled users to control a robotic wheelchair [30].

We conjecture that FPV can provide unique, task-relevant information critical to complex manipulation skills. For instance, when assembling a piece of furniture, a person may need to move their

head around from time to time to find better viewpoints for part alignment and screwing actions. FPV offers the demonstrator's perspective on the various aspects of a task demonstration, including how their hands interact with objects of interest and when to pay attention to what; in contrast, fixed, static sensing may not capture key task aspects due to occlusion. Besides encapsulating key task information, we believe that FPV implicitly captures appropriate points of view that match how people naturally perform manipulation tasks. In this work, we seek to understand how a collaborative robot that presents assembly parts to people in a user-centric way (i.e., close to their first-person perspective) may shape those users' task behaviors and experiences of working with the robot.

3 System Implementation

To explore first-person demonstration as an intuitive method of robot programming and the ways by which such a demonstration may enhance human-robot collaboration, we developed a system that parses first-person demonstrations into operational task models and an autonomous robot system that utilizes our generated task model to assist users in manipulation tasks. Our implementation was grounded in the context of furniture assembly (assembling an IKEA children's chair). Below, we describe our systems in detail.

3.1 Task Model of First-Person Demonstration

Given a first-person demonstration represented as a sequence of RGB images, we generated a task model encoding (1) the sequential procedure of the task, (2) spatial configurations symbolizing how different parts fit together, and (3) first-person viewpoints corresponding to task actions (i.e., connecting and screwing).

3.1.1 Task Model We developed the First-person Experience-based Assembly Tree (FEAsT) model, which hierarchically stores task-relevant and egocentric viewpoint information for rigid assembly tasks (Figure 2). In a FEAsT, a leaf node represents a single part used in the assembly; a parent node represents a *sub-assembly*, which is an assembled portion of the overall assembly consisting of two or more parts; and an edge connecting two nodes represents the *kinematic constraint* between the two nodes, which captures the relative geometric positions and orientations between the nodes [36]. We assigned a reference frame to each part and used the position and orientation of the reference frame to represent the transformation of that part. The reference frame of a sub-assembly is at the center between its children's reference frames and follows the orientation of its first child's reference frames. In addition to spatial information stored in the connecting edges, the hierarchical tree structure implicitly encodes the sequential order of the task procedure. One particularly unique aspect of the FEAsT is that it stores key task viewpoints from the original human demonstration; this additional degree of information captures which spatial configuration a human finds most natural to view a part or set of parts from during the assembly task. Viewpoint information is key to our design of user-centric robotic assistance.

A FEAsT can be populated in real time during a human assembly task demonstration. At the beginning of the demonstration, the tree is initialized to consist of independent leaf nodes for the individual assembly parts. As the demonstration progresses, the FEAsT begins

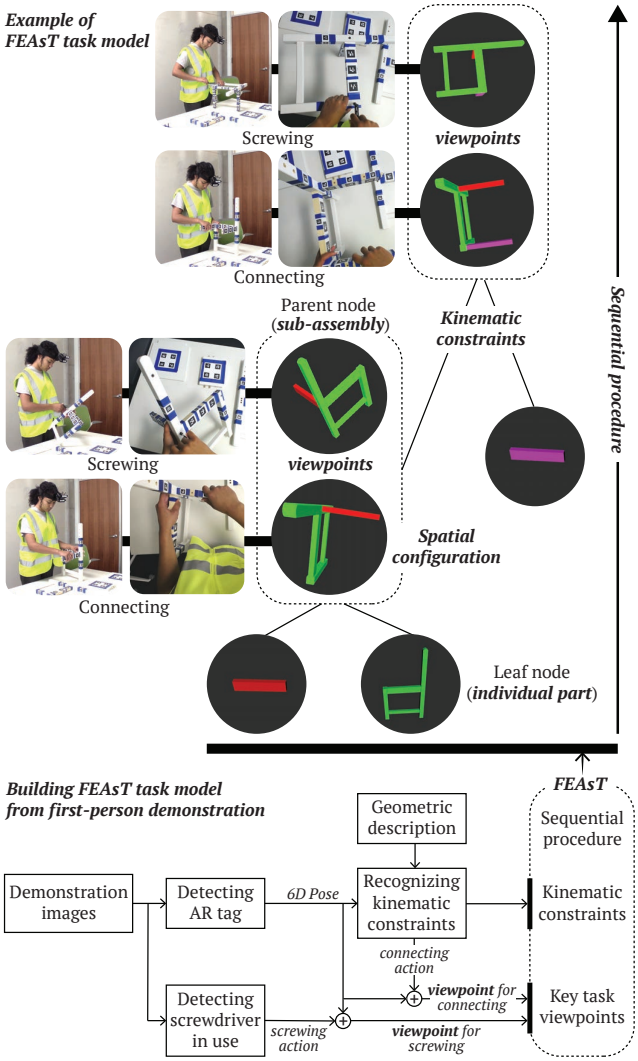

Figure 2: Top: hierarchical structure of a FEAsT: leaf nodes are *parts*; parent nodes are *sub-assemblies*; edges are *kinematic constraints* between a parent node and child nodes; and sub-assembly nodes contain relevant *first-person viewpoints*. Bottom: the process of parsing a natural first-person demonstration into a FEAsT model.

to form a hierarchical structure that encodes the interconnections between parts. When the user assembles two parts, a parent node is added to the FEAsT and is connected with edges to the part nodes. Each parent node stores the first-person viewpoints from when the human demonstrator performed an assembly action (such as connecting or screwing) on that particular sub-assembly.

3.1.2 Recognizing Kinematic Constraints Given any two parts, there may be multiple ways by which they can be assembled together (i.e., there may be multiple transformation matrices describing potential kinematic constraints). In order to recognize specific kinematic constraints from a task demonstration, our system requires the structure of assembly parts to be pre-specified in the form of *geometric descriptions*. A geometric description encodes the unique

structure of a part that allows it to connect to other components. A part with multiple connecting structures (e.g., dowel) includes a geometric description for each connecting structure, and a sub-assembly includes the geometric descriptions of each of its children. For the children's chair used in our system, the parts are connected using dowel/hole pairs. The geometric description for a connecting pair includes a *struct_name* (dowel), a *corresponding_struct* (hole) and a *pos*, which is the transformation from the part reference frame to the dowel reference frame. In this work, geometric descriptions were generated manually for each part; however, in the future, these descriptions can be obtained either from the manufacturer or through an algorithmic analysis of the part structures.

The process of kinematic constraint recognition (i.e., determining whether two parts are being connected) involves iterating through all detected parts and sub-assemblies and their geometric descriptions. At each iteration, the system checks whether detected parts and sub-assemblies have corresponding connecting structures and whether the position and orientation differences between the respective connecting structures are below empirically determined thresholds. In the dowel-hole connection case, we used a position threshold of 0.02 meters. For rotation, we obtained the similarity between two orientations, represented by quaternions, by taking the absolute value of the dot product and comparing that to a minimum threshold of 0.98. A kinematic constraint is established between the part node and sub-assembly node by adding a parent node to the FEAsT that connects them. The kinematic constraint is encoded as a transformation based on the *pos* components. We simplified detection of object pose by using an AR-tag pose-tracking package.

3.1.3 Recording Key Task Viewpoints Upon recognizing a kinematic constraint, which signifies that a part and a sub-assembly are connected, our system stores the current first-person viewpoint into the added parent node in the form of the pose (position and orientation) of the newly combined sub-assembly with respect to the first-person camera that captures the hand-object interaction. Additionally, our system records the first-person viewpoint when the human demonstrator screws the parts together. In our chair assembly context, connecting and screwing are two key task actions; people usually switch viewpoints when progressing from connecting two parts to screwing them together (Figure 2, Top). To simplify the detection of a user performing a screwing action, we defined a screwdriver "home" region. Our system employed a combination of AR tag tracking and color detection using first-person viewpoint images to determine whether or not the screwdriver was in the home region. For our purposes, a screwing action began when the screwdriver left the home region and ended when it returned. The system chose the frame in the middle of the screwing sequence as the point at which to store the first-person viewpoint.

3.1.4 Assembly Task Demonstration The hardware setup for recording first-person demonstrations involved two stacked head-mounted Logitech C930 cameras. Two cameras were used in order to ensure that the first-person view would include all task-relevant objects, with the upper camera being used to show the demonstrator's viewpoint in the form of their approximate gaze, and the lower camera being used to track their hand movements for action detection.

We conducted a data collection study to obtain natural first-person demonstrations of the assembly of an IKEA children's chair. This study involved 12 participants (7 females, 5 males), aged 18 to 46 ($M = 28.58$, $SD = 10.55$), from various educational backgrounds, including engineering, education, finance, and neuroscience. Each participant first learned the task by watching a tutorial video and then constructed the chair wearing a pair of head-mounted cameras; on average, the assembly task demonstration length was 137.96 s ($SD = 26.52\ s$). We chose one of the demonstrations to populate the FEAsT that was used in our human-robot collaborative assembly, described below. Note that the FEAsTs constructed from the participants' demonstrations are different because the participants had varying preferences about the assembly order and viewpoints. Future work will explore how to learn a common task model from multiple demonstrations.

3.2 Human-Robot Collaborative Assembly

In this section, we describe an autonomous robot system that uses a populated FEAsT to facilitate a collaborative chair assembly with a human partner. Our human-robot collaborative system uses the Universal Robots UR5 6-DOF robot manipulator, which was mounted with a two-finger Robotiq gripper.

To begin a collaborative assembly, our system first parses an input FEAsT to obtain the assembly procedure, following Algorithm 1. The parsing process starts by sorting the sub-assembly nodes based on their height in the tree, which represents their sequential order in the assembly task. It then iterates through the sorted nodes and through the ordered task actions (i.e., connecting and screwing) stored in that node. For each iteration, the system first recovers the current action's recorded egocentric viewpoint represented as a transformation (E_{currNode}) from the first-person camera to the current sub-assembly. It then traverses through all the descendants of the current node and recovers each descendant's transformation (E_{currDesc}) with respect to the first-person camera using the kinematic constraints (${}^{\text{descParent}}E_{\text{currDesc}}$) corresponding to edges in the FEAsT. If the current descendant is a leaf node (i.e., a part), the system adds the part's transformation into a 3D array named

Algorithm 1 First-person demonstration recovery from FEAsT

Require: FEAsT
assemblyNodes ← non-leaf nodes in the FEAsT
sort assemblyNodes by their height
partPoseArray ← empty ▷ a 3D array that stores the pose of each part for each action at each node
for all currNode ∈ assemblyNodes **do**
  for all currAction ∈ ordered task actions in currNode **do**
    E_{currNode} ← viewpoint of currNode for currAction
    for all currDesc ∈ descendants of currNode **do**
      ${}^{\text{descParent}}E_{\text{currDesc}}$ ← kinematic constraints of currDesc
      $E_{\text{currDesc}} \leftarrow E_{\text{descParent}} \times {}^{\text{descParent}}E_{\text{currDesc}}$
      if currDesc is a *part* **then**
        add E_{currDesc} in partPoseArray[currNode][currAction]
      end if
    end for
  end for
end for
return partPoseArray

partPoseArray, where the first, second, and third dimension represent the nodes (sub-assemblies) in FEAsT, the task actions in the corresponding node, and the parts that constitute the sub-assembly, respectively. The robot iterates through *partPoseArray* to obtain the appropriate positions and orientations for each part with respect to a fixed head pose, which it then uses to determine how to present parts during the collaborative assembly.

After parsing the task sequence, kinematic constraints, and viewpoint information, the robot proceeds to interact with the user. Our system follows a *turn-taking* style of interaction where the robot and the user take turns to complete the assembly task. The bottom pipeline in Figure 4 illustrates the collaborative assembly process. The process begins with the robot fetching a part (e.g., backrest) needed in the task and presenting it to the user following the viewpoint information for connecting. The user connects the presented part with the current sub-assembly; the robot then rotates the connected sub-assembly to present the parts to the user using the viewpoint information for screwing. After the user screws the parts together, the robot continues by rotating the new sub-assembly to a pose in which it is easy for the user to perform the next connecting action. This collaborative process continues until the task is completed. Due to implementation limitations and consideration of user safety, we chose a fixed head pose at an average human height as the reference frame from which the robot calculated the egocentric viewpoint transformations. In practice, a fixed head pose worked well since the participants did not move around during the experiment. Our future work will explore how to present parts adaptively according to real-time head poses. Regardless, the viewpoint information obtained from first-person demonstration was found to be sufficient to provide smooth user-centric robotic assistance.

Our robot system was designed to interact autonomously with humans to complete the assembly task. In our implementation, we used the MoveIt! Task Constructor [18] for planning the robot motions needed in the collaborative assembly. Moreover, to simplify motion planning, all task parts were given fixed locations for the robot to fetch them from. In order to achieve autonomous collaboration, our system needed to know a user's action status, which specified whether they were connecting, screwing, or holding/releasing parts. Methods for recognizing connecting and screwing actions are described in the previous section. To detect whether the user is holding a part, we generated hand and part masks and checked whether their overlap area is greater than 20 pixels for 3 frames. We adapted and trained the light-weight RefineNet [37] on the Georgia Tech Egocentric Activity (GTEA) datasets [14, 31] to generate hand masks. A holding action and a releasing action were considered mutually exclusive. Our system sometimes produced false negatives when performing user action detection during the user study; in these cases, the experimenter manually triggered the next step. The rates of experimenter intervention were 22.92%, 0.0%, 47.50%, and 20.00% for holding, connecting, the beginning of a screwing action, and the end of a screwing action, respectively.

Throughout the task assembly, the robot provided verbal instructions to the user. It told the user what to do for the next assembly step (e.g., "*Please connect the previous part to this piece*"), provided information about its actions (e.g., "*I am going to open my hand*"), and offered approval words such as "*Ok*" or "*Good work!*" to indicate that a detected user action is correct and that the system will move forward to the next stage of the procedure. Our system used the Amazon Polly service to produce verbal instructions.

4 Evaluation

Our evaluation was focused on assessing how user-centric robotic assistance may influence a user's behavior during, and their perception of, a collaboration with a robot. Our central hypothesis is that robotic assembly assistance based on a first-person demonstration will lead to better user experience than robotic assistance that does not consider first-person viewpoints.

4.1 Experimental Task, Conditions, and Design

We contextualized our evaluation in a chair assembly task by designing an experimental exercise in which a UR5 manipulator assisted participants in assembling an IKEA children's chair. As shown in Figure 3, the robot engaged in one of two types of assistance when guiding participants:

Standard Assistance (control condition): Following the top pipeline in Figure 4, the robot fetched parts and presented them to participants matching the viewpoint shown on the cover page of the official IKEA assembly manual.

User-Centric Assistance (experimental condition): As shown on the bottom pipeline in Figure 4, the robot fetched parts and presented them to users matching the first-person viewpoint extracted from one of our previously collected human demonstrations.

The robot provided verbal instructions and fetched the parts of the chair for participants in both conditions. Our user evaluation followed a within-participants design, with each participant working with the robot in both conditions. We counterbalanced the order in which the conditions were presented.

4.2 Experimental Procedure

The user evaluation began with the experimenter obtaining informed consent and providing an overview of the experiment to the participant. The experimenter then helped the participant put on a pair of head-mounted cameras and performed a camera calibration procedure. After setting up the experimental apparatus, the experimenter told the participant to follow the robot's instructions to complete the chair assembly task. Upon finishing the task, the participant filled out a questionnaire regarding their experience working with the robot. This procedure was then repeated for the other condition. After completing the assembly task in both conditions, the participant provided demographic information and was

Figure 3: Our experimental conditions and setup.

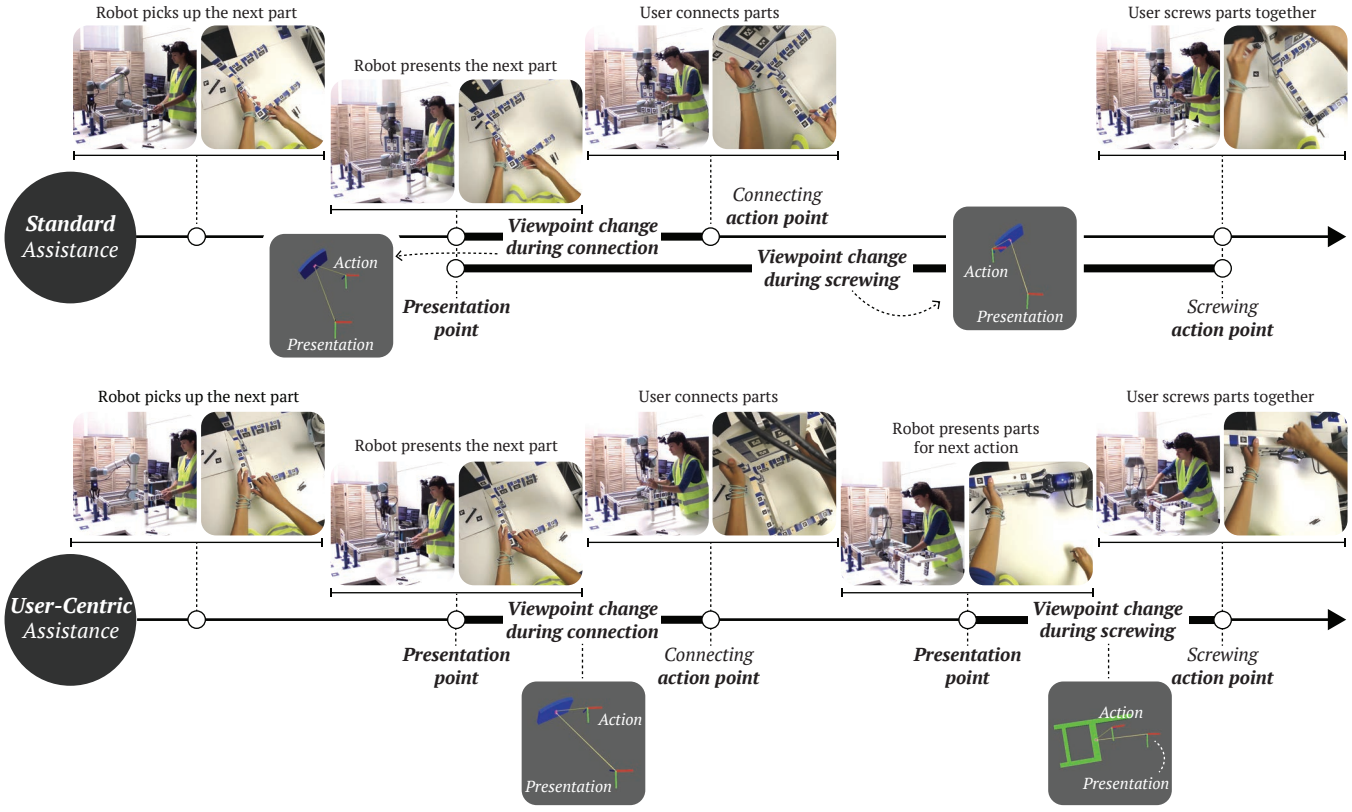

Figure 4: The pipelines the robot followed for the standard and user-centric assistance conditions used in our user evaluation. This figure also illustrates how we measured viewpoint changes during task actions.

then interviewed for feedback about their experience. Each participant was compensated with $5 USD for their participation in the experiment, which lasted approximately 30 minutes.

4.3 Measures

We employed a combination of objective, behavioral, and subjective measures to assess participants' performance and user experience.

4.3.1 Objective Measures We measured how much head movement (i.e., viewpoint change) was involved when the participant performed the assembly task. The premise behind these measures is that if a part is presented in an unnatural pose that is awkward for the user to interact with, then the user will modify their viewpoint to perform the task action. As shown in Figure 4, for each action, we defined (1) a *presentation* point, the moment when the robot presents a part to the user, and (2) an *action* point, the moment when the user conducts either a connecting (inserting dowels into holes) or screwing action. Since screwing is a continuous action, we chose a frame that was close to the middle of the frame sequence and included clearly presented AR tags. We then measured the head movement from the *presentation* point to the *action* point in terms of distance moved and viewpoint rotated, as defined below:

Distance moved (meters): Euclidean distance between the head position at the presentation point and the head position at the action point.

Viewpoint rotated (radians, $[0, \pi]$): Angle of rotation along the shortest path between the head orientation at the presentation point and the head orientation at the action point.

4.3.2 Behavioral Measures In addition to measuring head movement during task performance, we wanted to determine how the type of assistance influenced participants' behavior during the assembly task. In particular, we measured the number of times the participants used their non-dominant hands or switched between hands when performing screwing actions. In addition, we measured the number of times participants dropped the screwdriver when performing a screwing action *(tool drops)*.

4.3.3 Subjective Measures Our subjective measures involved *perceived teamwork* and *perceived consideration*. The *teamwork* scale consisted of seven items (Cronbach's $\alpha = 0.84$) and sought to measure how cooperative and how good of a teammate participants perceived the robot to be during the interaction. The *consideration* scale consisted of six items (Cronbach's $\alpha = 0.87$) and aimed to assess participants' perceptions of how considerate the robot was of the user's actions, tasks, and comfort during the interaction. In addition to these two scales, we included two task-specific questions about whether or not it was easy for the participants to *align* and *screw* the chair pieces together.

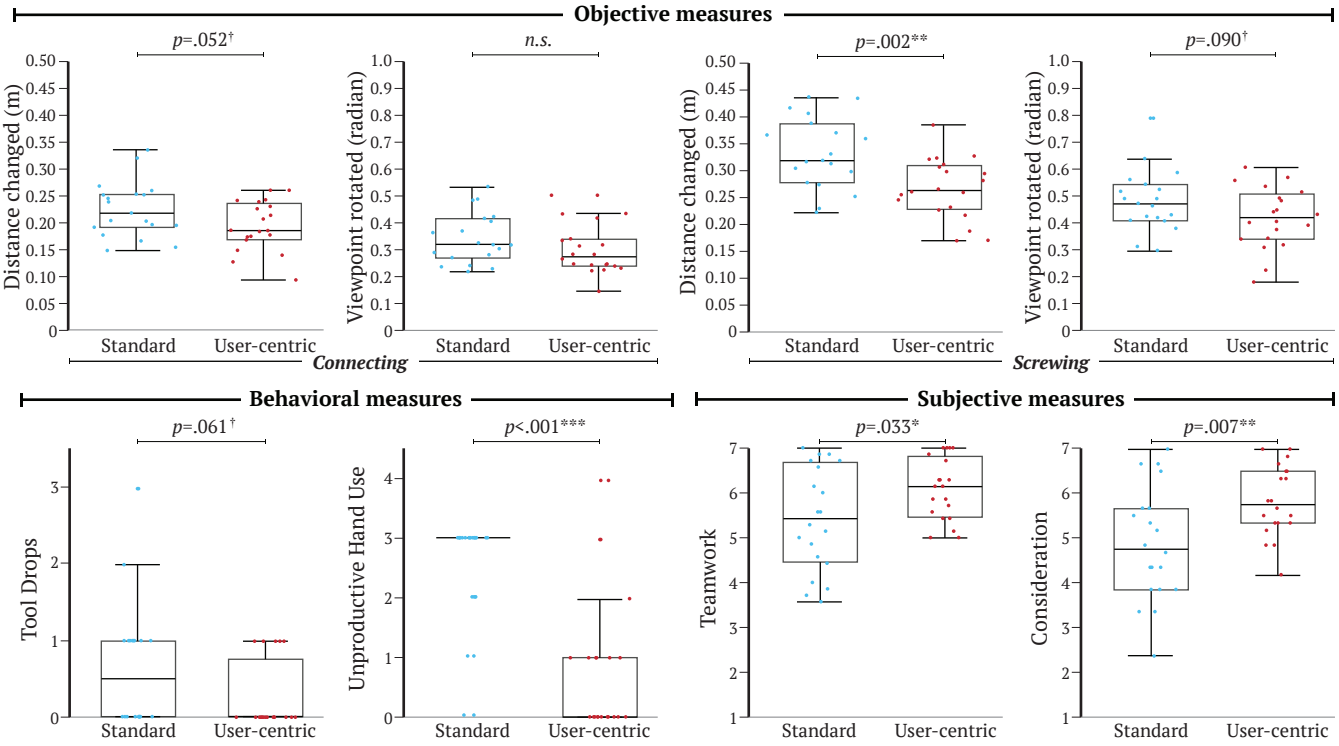

Figure 5: Box and whisker plots of data on the objective measures of head movement, behavioral measures of unproductive behavior, and subjective measures of user experience. The top and bottom of each box represent the first and third quartiles, and the line inside each box is the statistical median of the data. The length of the box is defined as the interquartile range (IQR). The ends of the whiskers are the first quartile minus 1.5 IQR and the third quartile plus 1.5 IQR. For the subjective measures, a higher value close to 7 indicates positive perception of teamwork or consideration.

4.4 Participants

A total of 21 participants were recruited from the local community. One participant was excluded from our data analysis due to a system failure during the experiment. The resulting 20 participants (10 females, 10 males) were aged 18 to 65 (M = 27.70, SD = 13.86) and had a variety of educational backgrounds, including engineering, applied math, biology, music, and international studies. Nineteen out of 20 participants reported themselves as right-handed, with one participant being left-handed.

4.5 Results

Figure 5 and Table 1 summarize our results based on a one-way repeated-measures analysis of variance (ANOVA), in which the type of assistance—either standard or user-centric—was set as a fixed effect, and the participant was set as a random effect.

4.5.1 Objective Measures Our data indicated that when collaborating with the robot providing user-centric assistance, participants had less distance changed (p = .002) and moderately less rotation (p = .090) of their head movements when conducting screwing actions than in the standard assistance collaboration. At the same time, our results showed only a marginal difference (p = .052) in distance changed and no significant difference in head rotation between the two conditions when the participants performed connecting actions. When reviewing the selected first-person demonstration employed in our study, we noticed that the IKEA manual viewpoint was quite similar to the demonstrated viewpoint for connecting actions; therefore, we speculate that the similarity in the viewpoints resulted in no substantial differences observed in these measures.

4.5.2 Behavioral Measures Our results revealed that participants were more likely to use their non-dominant hand or switch hands ($p < .001$) during the assembly task when working with the robot providing standard assistance compared to the robot providing user-centric assistance. When reviewing the video recordings of the interactions, we observed that the fixed—and sometimes awkward—presentation of the chair parts in the control condition prompted participants to engage in unproductive hand use. In contrast, participants were able to more often use their dominant hand to carry out task actions when provided with user-centric assistance.

Moreover, we observed that the participants tended to drop the screwdriver more often (p = .061) when working with the robot providing standard assistance than when working with the user-centric robot. For reference, the number of tool drops averaged over the 12 participants from whom we collected human demonstrations in Section 3.1 was 0.42 (SD = 0.90). While we were unable to identify a direct relationship between non-dominant hand use and tool drops, we speculate that these two behaviors are possibly associated and might have influenced each other.

Table 1: Statistical results of our measures.

Objective measures	Control	Experimental	Statistical test results
Connecting action			
Distance changed	0.23 (*SD*=0.05)	0.19 (*SD*=0.05)	$F(1,37)=4.03$, $p=.052^{\dagger}$, $\eta_p^2=0.101$
Viewpoint rotated	0.34 (*SD*=0.09)	0.30 (*SD*=0.09)	$F(1,37)=2.30$, $p=.138$, $\eta_p^2=0.060$
Screwing action			
Distance changed	0.33 (*SD*=0.07)	0.27 (*SD*=0.06)	$F(1,37)=11.15$, $p=.002^{**}$, $\eta_p^2=0.234$
Viewpoint rotated	0.48 (*SD*=0.12)	0.42 (*SD*=0.11)	$F(1,37)=3.04$, $p=.090^{\dagger}$, $\eta_p^2=0.076$
Behavioral measures	Control	Experimental	Statistical test results
Tool drops	0.65 (*SD*=0.81)	0.25 (*SD*=0.44)	$F(1,38)=3.73$, $p=.061^{\dagger}$, $\eta_p^2=0.089$
Unproductive hand use	2.65 (*SD*=0.81)	0.75 (*SD*=1.12)	$F(1,38)=37.79$, $p<.001^{***}$, $\eta_p^2=0.499$
Subjective measures	Control	Experimental	Statistical test results
Teamwork	5.42 (*SD*=1.15)	6.09 (*SD*=0.69)	$F(1,38)=4.88$, $p=.033^{*}$, $\eta_p^2=0.114$
Consideration	4.86 (*SD*=1.28)	5.83 (*SD*=0.80)	$F(1,38)=8.27$, $p=.007^{**}$, $\eta_p^2=0.179$
Easy to align	5.75 (*SD*=1.16)	6.55 (*SD*=0.69)	$F(1,38)=7.01$, $p=.012^{*}$, $\eta_p^2=0.156$
Easy to screw	5.15 (*SD*=1.69)	6.15 (*SD*=1.14)	$F(1,38)=4.80$, $p=.035^{*}$, $\eta_p^2=0.112$

4.5.3 Subjective Measures Participants reported significantly greater perceived teamwork ($p = .033$) with the user-centric robot compared to the standard robot. They also perceived the user-centric robot to be more considerate ($p = .007$) in terms of accommodating how they preferred to perform the task; for example, P15 commented that *"[The user-centric robot] could actually feel when I was uncomfortable, like I couldn't actually put the pieces together in a really comfortable way, so he just kind of helped me do that, and it was really nice because I felt like we were working together."*

Additionally, the participants agreed that the user-centric robotic assistance allowed them to align pieces and screw them together more easily than in the standard assistance condition—which did not consider user task viewpoints—as described by P2 : *"The [user-centric] robot did a slight nice movement into the position where I was kind of facing it, so it was easier for me to screw."*

5 DISCUSSION

We explored how first-person demonstration may be used to generate user-centric assistance in human-robot collaborative assembly. Specifically, the viewpoints from a first-person demonstration enable a robot to deliver an assembly part to a human partner in a user-friendly way. Our user evaluation shows that, when working with a robot offering user-centric assistance, participants felt that the robot was a more considerate collaborator and were less likely to engage in unproductive behavior, such as using their non-dominant hand or switching which hand they used during the screwing actions. Below, we discuss additional findings, implications for collaborative robotics, and the limitations of this work.

5.1 A Closer Look at User Experience

While our results indicated that user-centric assistance enhanced user experience overall, there seems to be a "sweet spot" for how considerate a robot should be. One participant commented, *"The [user-centric] robot was trying to do almost too much. So by trying to make it easier for me, it was making it… not easier…"* (P13). The perception of *trying too hard* put the participant off and could possibly lead to negative interaction outcomes [23]. Moreover, it is important to consider how the robot's motion should be accounted for when designing user-centric assistance. One participant stated, *"I guess the discomfort [of working with the user-centric robot] was just kind of not knowing where … how it would move the chair."* (P17). Future work should explore the integration of legible motions [11, 12, 40] into user-centric assistance. Beyond user experience, user-centric assistance could help aid in task performance. One user mentioned that the user-centric viewpoints could have helped them spot a task mistake: *"In the [standard condition], I forgot to screw [one screw] and I didn't realize. And [in the user-centric condition], it's easier to see like, 'Oh! There is a screw missing for me!'"* (P17).

5.2 Implications for Collaborative Robotics

First-person vision offers unique advantages to collaborative robotics. It approximates a person's attentional focus, and therefore can be used to estimate task intent [21] (e.g., what part the user might need next) and identify errors in collaboration [5] (e.g., the user focuses on the wrong part). Moreover, for real-world applications in the wild, fixed external sensing is generally infeasible; FPV serves as an alternative sensing modality and can therefore bring human-robot collaboration into a wider range of applicable domains, including search and rescue and in-home assistance. Finally, the common robot programming method used for collaborative robots in manufacturing is a hybrid use of kinesthetic teaching and visual programming; however, the unfamiliarity of a robot's kinematics and programming interface can pose challenges to everyday users when authoring robot programs. We believe that "programming by doing" via first-person demonstration presents new opportunities for a wider array of users to reskill robots.

5.3 Limitations & Future Work

While our results suggest the potential of first-person demonstration in creating more productive and natural human-robot collaboration, the limitations of the present work also highlight directions for future research. Our current human-robot collaborative system operates on a selected FEAsT constructed from a human demonstration. Future work should investigate how to abstract task information from multiple first-person demonstrations; these investigations will involve exploring different task representations (e.g., [10, 19]) and learning approaches (e.g., [13, 39]). Moreover, in our human-robot collaboration study, our system simply transformed previously acquired viewpoint information onto a fixed, predetermined head pose; this decision was mostly due to limitations in motion planning and consideration of user safety. However, our future work will focus on enabling real-time adjustment of robot motions and task-level execution to achieve adaptive collaboration.

Acknowledgments

We would like to thank Jaimie Patterson, Maia Stiber, Annie Mao, and Yuxiang Gao for their help with, and Johns Hopkins University for its support of, this work.

References

[1] Baris Akgun, Maya Cakmak, Jae Wook Yoo, and Andrea Lockerd Thomaz. 2012. Trajectories and keyframes for kinesthetic teaching: A human-robot interaction perspective. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. ACM, 391–398.

[2] Jacopo Aleotti and Stefano Caselli. 2006. Robust trajectory learning and approximation for robot programming by demonstration. *Robotics and Autonomous Systems* 54, 5 (2006), 409–413.

[3] Sonya Alexandrova, Zachary Tatlock, and Maya Cakmak. 2015. Roboflow: A flow-based visual programming language for mobile manipulation tasks. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 5537–5544.

[4] Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. 2009. A survey of robot learning from demonstration. *Robotics and autonomous systems* 57, 5 (2009), 469–483.

[5] Reuben M Aronson and Henny Admoni. 2018. Gaze for Error Detection During Human-Robot Shared Manipulation. In *Fundamentals of Joint Action workshop, Robotics: Science and Systems*.

[6] Gedas Bertasius, Aaron Chan, and Jianbo Shi. 2018. Egocentric basketball motion planning from a single first-person image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5889–5898.

[7] Gedas Bertasius, Hyun Soo Park, Stella X Yu, and Jianbo Shi. 2016. First person action-object detection with egonet. In *Proceedings of Robotics: Science and Systems (RSS)*.

[8] Aude Billard, Sylvain Calinon, Ruediger Dillmann, and Stefan Schaal. 2008. Robot programming by demonstration. In *Springer handbook of robotics*. Springer, 1371–1394.

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

[10] LS Homem De Mello and Arthur C Sanderson. 1990. AND/OR graph representation of assembly plans. *IEEE Transactions on robotics and automation* 6, 2 (1990), 188–199.

[11] Anca Dragan and Siddhartha Srinivasa. 2013. Generating legible motion. (2013).

[12] Anca D Dragan, Shira Bauman, Jodi Forlizzi, and Siddhartha S Srinivasa. 2015. Effects of robot motion on human-robot collaboration. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 51–58.

[13] Staffan Ekvall and Danica Kragic. 2006. Learning task models from multiple human demonstrations. In *ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 358–363.

[14] Alireza Fathi, Xiaofeng Ren, and James M Rehg. 2011. Learning to recognize objects in egocentric activities. In *CVPR 2011*. IEEE, 3281–3288.

[15] Chelsea Finn, Tianhe Yu, Tianhao Zhang, Pieter Abbeel, and Sergey Levine. 2017. One-Shot Visual Imitation Learning via Meta-Learning. In *Conference on Robot Learning*. 357–368.

[16] Yuxiang Gao and Chien-Ming Huang. 2019. PATI: a projection-based augmented table-top interface for robot programming. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, 345–355.

[17] Dylan F Glas, Takayuki Kanda, and Hiroshi Ishiguro. 2016. Human-robot interaction design using Interaction Composer eight years of lessons learned. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 303–310.

[18] Michael Görner, Robert Haschke, Helge Ritter, and Jianwei Zhang. 2019. MoveIt! Task Constructor for task-level motion planning. (2019).

[19] Bradley Hayes and Brian Scassellati. 2016. Autonomously constructing hierarchical task networks for planning and human-robot collaboration. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 5469–5476.

[20] Micha Hersch, Florent Guenter, Sylvain Calinon, and Aude Billard. 2008. Dynamical system modulation for robot learning via kinesthetic demonstrations. *IEEE Transactions on Robotics* 24, 6 (2008), 1463–1467.

[21] Chien-Ming Huang and Bilge Mutlu. 2016. Anticipatory robot control for efficient human-robot collaboration. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*. IEEE Press, 83–90.

[22] Justin Huang and Maya Cakmak. 2017. Code3: A system for end-to-end programming of mobile manipulator robots for novices and experts. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 453–462.

[23] James Kennedy, Paul Baxter, and Tony Belpaeme. 2015. The robot who tried too hard: Social behaviour of a robot tutor can negatively affect child learning. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 67–74.

[24] Thomas Kollar, Stefanie Tellex, Deb Roy, and Nicholas Roy. 2010. Toward understanding natural language directions. In *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction*. IEEE Press, 259–266.

[25] Yasuo Kuniyoshi, Masayuki Inaba, and Hirochika Inoue. 1994. Learning by watching: Extracting reusable task knowledge from visual observation of human performance. *IEEE transactions on robotics and automation* 10, 6 (1994), 799–822.

[26] Stanislao Lauria, Guido Bugmann, Theocharis Kyriacou, and Ewan Klein. 2002. Mobile robot programming using natural language. *Robotics and Autonomous Systems* 38, 3-4 (2002), 171–181.

[27] Alex X Lee, Henry Lu, Abhishek Gupta, Sergey Levine, and Pieter Abbeel. 2015. Learning force-based manipulation of deformable objects from multiple demonstrations. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 177–184.

[28] Jangwon Lee and Michael S Ryoo. 2017. Learning robot activities from first-person human videos using convolutional future regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 1–2.

[29] Kyungjun Lee and Hernisa Kacorri. 2019. Hands Holding Clues for Object Recognition in Teachable Machines. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 336.

[30] Haoxiang Li, Mohammed Kutbi, Xin Li, Changjiang Cai, Philippos Mordohai, and Gang Hua. 2016. An egocentric computer vision based co-robot wheelchair. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 1829–1836.

[31] Yin Li, Zhefan Ye, and James M Rehg. 2015. Delving into egocentric actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 287–295.

[32] Minghuang Ma, Haoqi Fan, and Kris M Kitani. 2016. Going deeper into first-person activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1894–1903.

[33] Cynthia Matuszek, Liefeng Bo, Luke Zettlemoyer, and Dieter Fox. 2014. Learning from Unscripted Deictic Gesture and Language for Human-Robot Interactions.. In *AAAI*. 2556–2563.

[34] Grégoire Milliez, Raphaël Lallement, Michelangelo Fiore, and Rachid Alami. 2016. Using human knowledge awareness to adapt collaborative plan generation, explanation and monitoring. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*. IEEE Press, 43–50.

[35] Anahita Mohseni-Kabir, Charles Rich, Sonia Chernova, Candace L Sidner, and Daniel Miller. 2015. Interactive hierarchical task learning from a single demonstration. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 205–212.

[36] Yoan Mollard, Thibaut Munzer, Andrea Baisero, Marc Toussaint, and Manuel Lopes. 2015. Robot programming from demonstration, feedback and transfer. In *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 1825–1831.

[37] Vladimir Nekrasov, Chunhua Shen, and Ian Reid. 2018. Light-weight refinenet for real-time semantic segmentation. In *Proceedings of the British Machine Vision Conference (BMVC)*.

[38] Scott Niekum, Sachin Chitta, Andrew G Barto, Bhaskara Marthi, and Sarah Osentoski. 2013. Incremental Semantically Grounded Learning from Demonstration.. In *Robotics: Science and Systems*, Vol. 9. Berlin, Germany.

[39] Scott Niekum, Sarah Osentoski, George Konidaris, and Andrew G Barto. 2012. Learning and generalization of complex tasks from unstructured demonstrations. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 5239–5246.

[40] Stefanos Nikolaidis, Anca Dragan, and Siddharta Srinivasa. 2016. Viewpoint-Based Legibility Optimization. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*. IEEE Press, 271–278.

[41] Stefanos Nikolaidis, Ramya Ramakrishnan, Keren Gu, and Julie Shah. 2015. Efficient model learning from joint-action demonstrations for human-robot collaborative tasks. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*. ACM, 189–196.

[42] Chris Paxton, Andrew Hundt, Felix Jonathan, Kelleher Guerin, and Gregory D Hager. 2017. CoSTAR: Instructing collaborative robots with behavior trees and vision. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 564–571.

[43] Alessandro Roncone, Olivier Mangin, and Brian Scassellati. 2017. Transparent role assignment and task allocation in human robot collaboration. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 1014–1021.

[44] MS Ryoo, Thomas J Fuchs, Lu Xia, Jake K Aggarwal, and Larry Matthies. 2015. Robot-centric activity prediction from first-person videos: What will they do to me?. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 295–302.

[45] Yasaman S Sefidgar, Prerna Agarwal, and Maya Cakmak. 2017. Situated tangible robot programming. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 473–482.

[46] Lanbo She, Yu Cheng, Joyce Y Chai, Yunyi Jia, Shaohua Yang, and Ning Xi. 2014. Teaching robots new actions through natural language instructions. In *Robot and Human Interactive Communication, 2014 RO-MAN: The 23rd IEEE International Symposium on*. IEEE, 868–873.

[47] Lanbo She, Shaohua Yang, Yu Cheng, Yunyi Jia, Joyce Chai, and Ning Xi. 2014. Back to the blocks world: Learning new actions through situated human-robot dialogue. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. 89–97.

[48] Hyun Soo Park, Jyh-Jing Hwang, Yedong Niu, and Jianbo Shi. 2016. Egocentric future localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4697–4705.

[49] Hyun Soo Park and Jianbo Shi. 2015. Social saliency prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4777–4785.

[50] Maj Stenmark and Pierre Nugues. 2013. Natural language programming of industrial robots.. In *ISR*. Citeseer, 1–5.

[51] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. 2019. Lsta: Long short-term attention for egocentric action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 9954–9963.

[52] C Sylvain. 2009. *Robot programming by demonstration: A probabilistic approach*. EPFL Press.

[53] Faraz Torabi, Garrett Warnell, and Peter Stone. 2018. Behavioral cloning from observation. *arXiv preprint arXiv:1805.01954* (2018).

[54] Lu Xia, Ilaria Gori, Jake K Aggarwal, and Michael S Ryoo. 2015. Robot-centric activity recognition from first-person rgb-d videos. In *2015 IEEE Winter Conference on Applications of Computer Vision*. IEEE, 357–364.

[55] Tianhe Yu, Chelsea Finn, Annie Xie, Sudeep Dasari, Tianhao Zhang, Pieter Abbeel, and Sergey Levine. 2018. One-shot imitation from observing humans via domain-adaptive meta-learning. In *Proceedings of Robotics: Science and Systems (RSS)*.

[56] Tianhao Zhang, Zoe McCarthy, Owen Jow, Dennis Lee, Ken Goldberg, and Pieter Abbeel. 2018. Deep imitation learning for complex manipulation tasks from virtual reality teleoperation. *2018 IEEE International Conference on on Robotics and Automation (ICRA)* (2018).