2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Hierarchical Intention Tracking for Robust Human-Robot Collaboration in Industrial Assembly Tasks

Zhe Huang*, Ye-Ji Mun*, Xiang Li†, Yiqing Xie†, Ninghan Zhong†,
Weihang Liang, Junyi Geng, Tan Chen, and Katherine Driggs-Campbell

***Abstract*— Collaborative robots require effective human intention estimation to safely and smoothly work with humans in less structured tasks such as industrial assembly, where human intention continuously changes. We propose the concept of intention tracking and introduce a collaborative robot system that concurrently tracks intentions at hierarchical levels. The high-level intention is tracked to estimate human's interaction pattern and enable robot to (1) avoid collision with human to minimize interruption and (2) assist human to correct failure. The low-level intention estimate provides robot with task-related information. We implement the system on a UR5e robot and demonstrate robust, seamless and ergonomic human-robot collaboration in an ablative pilot study of an assembly use case.**

I. Introduction

Collaborative robot solutions are being actively developed in industrial tasks including part sorting, tool delivery, precise positioning, and cooperative transportation [1]–[4]. Teaming up humans and robots boosts production efficiency by combining cognition and dexterity from humans with repeatability and load carrying capacity from robots [5]. One key challenge for these solutions is human uncertainty [6]. We argue that reliable human intention estimation is a critical component of safe and seamless human-robot collaboration.

Many works have incorporated human intention estimation in highly structured settings [1], [7]. We observe critical limitations when previous methods are applied to less structured tasks. Take a prototypical collaborative assembly task in Figure 1 as an example, where a human-robot team assembles four pairs of male and female parts at designated regions[1]. The human is responsible for part alignment which requires delicate manipulation skills, and the robot is responsible for pushing male parts into female parts which requires a large force. In this application, the prior methods will fail due to two key factors: dynamic intention and intention hierarchy.

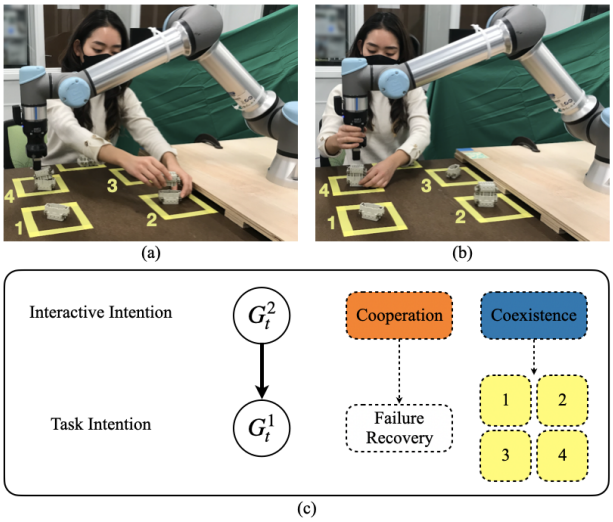

Fig. 1: A person works with a robot to assemble parts in her preferred order. The robot must keep track of which part is her likely goal G_t^1, while estimating her interactive intention G_t^2. (a) Coexistence mode: the robot performs pushing action to the parts she aligned while she is aligning other parts. (b) Cooperation mode: the robot is manually guided to recover the failed pushing attempt. (c) The hierarchy of human intention during collaboration.

First, human intention changes at different stages of the task. The robot needs to estimate and follow the sequence by which the human aligns the parts. Prior works typically define human intention as a single random variable that does not evolve over time [8]. This definition transforms a multi-step task into multiple single-step tasks, and intention estimation can converge prematurely in one task before the next begins. Alternatively, the intention is inferred by single shot classification [9], but inference capabilities are limited with partial historical information.

Second, human intention is often composed of a multi-layer hierarchy. Previous works are focused on single-layer cases [10], [11], which can be inadequate to address unstructured settings. In the proposed example, a two-layer intention hierarchy is needed to entail efficiency and robustness. The high level includes *coexistence interactive intention*, meaning the human intends to work without interruption from the robot, and *cooperation interactive intention*, meaning the human intends to physically guide the robot. The low level includes which part the human intends to work on, and assembly failure correction.

* denotes equal contribution as the first author. † denotes equal contribution as the second author.

Z. Huang, Y. Mun, X. Li, Y. Xie, W. Liang, and K. Driggs-Campbell are with the Department of Electrical and Computer Engineering at the University of Illinois at Urbana-Champaign. emails: {zheh4, yejimun2, xiangl5, yiqingx2, weihang2, krdc}@illinois.edu

N. Zhong is with the Department of Computer Science at the University of Illinois at Urbana-Champaign. email: ninghan2@illinois.edu

J. Geng is with the Department of Aerospace Engineering at Pennsylvania State University. email: jgeng@psu.edu

T. Chen is with the Department of Electrical and Computer Engineering at Michigan Technological University. email: tanchen@mtu.edu

This work was supported by Foxconn Interconnect Technology through the UIUC Center for Networked Intelligent Components and Environments (C-NICE).

[1]This task is a less structured analogy to BMW/MINI Crash Can Assembly Task demonstrated at `https://youtu.be/keh99z1M5LI`, where human aligns rivets and robot performs rivet installation.

In this work, we propose *hierarchical intention tracking* to take these factors into account. We introduce a Hierarchical Intention Tracking (HIT) based human-robot collaboration system to simultaneously track both high-level interactive intention and low-level task intention. When the high-level estimate is coexistence, the robot performs collision avoidance while reaching for the task goal estimated at the low level. When the high-level estimate is cooperation, the robot approaches to the human and provides admittance control for failure recovery. We present two main contributions:

(1) We derive *intention tracking* based on a generic graphical model of intention-evolving human-robot collaboration by treating intention as a Markov process, and extend intention tracking to a multi-layer intention hierarchy.
(2) We develop a seamless, robust, and ergonomic human-robot collaboration system based on hierarchical intention tracking, which is demonstrated in a multi-step assembly task through ablative pilot study.

II. Related Work

A. Automation Modes in Human-Robot Collaboration

A variety of industrial applications involve human-robot collaboration such as tool handover [1], [2], heavy object lifting [12], [13], surface polishing [14], welding [15], and assembly [16], [17]. Human-robot collaboration has three major automation modes: safety, coexistence, and cooperation [7], [18]. The safety mode enforces human and robot to not move at the same time in shared work space. Robot motion is paused when human is detected and is resumed after human leaves [19], [20]. The coexistence mode allows human and robot to work simultaneously in close proximity with no physical contact. Human and robot execute their own tasks, and prefer no interruption from the partner, so robot performs human avoidance [21]–[23]. The cooperation mode offers physical coordination. Manual guidance and shared control are used to address situations which are beyond robot capabilities and require human intervention [8], [10], [11]. To achieve robust and efficient close-proximity human-robot collaboration in complicated free-form assembly tasks, our work implements a coexistence module for concurrent operation and a cooperation module for failure recovery.

B. Human Intention Estimation

The research on human-centered autonomy has extensively investigated the concept of human intention [24]–[26]. To estimate human intention, various types of observation on human behavior are used as input, including human trajectories [1], [25], [26], gesture [16], [27], gaze [28], speech [29], facial expressions [30], and force-torque measurements [8], [24]. Many works study how human intention influences human behavior and benchmark approaches on human datasets [1], [25]. Other works consider the mutual influence between human and robot and take both human and robot states as input [8], [31], [32].

Human intention is typically represented by one fixed random variable. Thus, intention estimation is formulated either as intention recognition given a fixed length of observations [8], [25], [26], or recursive intention estimation, which produces an online maximum a posteriori over the same intention variable given tracked sequence [31]–[34]. Our previous work proposes intention mutation mechanism [35], but the mechanism is developed in the context of pedestrian trajectory prediction, where only human behavior is considered and changing intention is regarded as anomaly. To resolve the premature convergence issue in recursive intention estimation approaches, in this work we generalize intention mutation mechanism to intention transition dynamics by treating intention as a state variable, and formally derive intention tracking in the context of human-robot collaboration.

Intention hierarchy is often discussed when human intention is defined in terms of commands. A command can be interpreted as a pyramid of a goal, sub goals, and primitives [36]. Hierarchical Task Network is used to exploit the hierarchical structure by generating candidates of flattened primitive sequences, and intention estimation still works at a single level to distinguish among candidates [37]. Hierarchical Hidden Markov Model in [38] recognizes intention sequences by incorporating two levels of Hidden Markov Models, but its higher level is used to refine results from the lower level through context awareness, where the intention is only defined at one level. Our work defines intention hierarchy in terms of abstraction levels which are not necessarily constrained to command or task related goals. Though implementation is based on the two-layer case in Figure 1(c), our hierarchical intention tracking framework can be applied to an intention hierarchy with an arbitrary number of levels. In contrast to [36], we apply Dynamic Bayesian Networks to study hierarchical intention tracking [39], and Bayesian inference is performed both vertically (hierarchically) and horizontally (temporally).

III. Hierarchical Intention Tracking

A human-robot team is assigned m task goals. The human leads the team to accomplish all goals in his/her desired task sequence unknown to the robot. Human-robot team dynamics are formulated as the graphical model presented in Figure 2(a). The human intention G_t is a Markov process, because human has different ground truth goals at different stages of the task plan. Ideal interaction between latent human states Z_t^h and latent robot states Z_t^r would lead to seamless human-robot collaboration, where robot motion always matches human intention G_t. To achieve seamless collaboration, G_t must be effectively tracked given observed state history of both the human and the robot $X_{1:t}^{h,r}$, e.g., positions of human skeleton and robot end-effector. First, we study intention tracking when G_t is a random variable. Second, we study hierarchical intention tracking when G_t represents an intention hierarchy with an arbitrary number of layers, and apply to the two-level use case in Figure 1(c).

A. Intention Tracking in Human-Robot Collaboration

The human intention variable G_t at each time step is discrete and the cardinality of its sample space is m. Bayesian

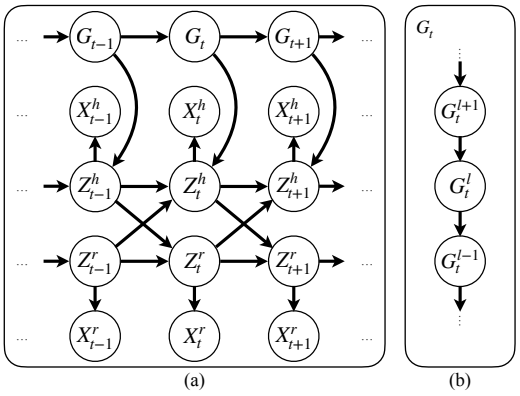

Fig. 2: (a) A graphical model of intention-evolving human-robot collaboration. We denote observed human and robot states as X^h_t and X^r_t. Latent human and robot states are represented by Z^h_t and Z^r_t. The human intention is represented by G_t. (b) A graphical model of hierarchical intentions, which is essentially a chain of intentions. The lowest-level intention G^1_t is a parent of Z^h_t in (a).

filtering is recursively applied every T_p time steps to obtain the posteriori over G_t conditioned on observed states $X^{h,r}_{1:t}$. For simplicity, we assume the intention outcomes $g_{t+1:t+T_p}$ are consistent. This assumption is reasonable in our implementation because T_p time steps are less than 0.2 second. The prediction step of Bayesian filtering is as follows.

$$P(g_{t+T_p}|x^{h,r}_{1:t}) = \sum_{g_t} P(g_{t+T_p}|g_t)P(g_t|x^{h,r}_{1:t}) \tag{1}$$

We define the transition model for intention dynamics $P(g_{t+T_p}|g_t)$ as a time-invariant matrix

$$P(g_{t+T_p}|g_t) = \begin{cases} \alpha, & \text{if } g_{t+T_p} = g_t; \\ (1-\alpha)/(m-1), & \text{if } g_{t+T_p} \neq g_t. \end{cases} \tag{2}$$

where a large α indicates the human is more likely to keep the current intention. The probability of the human shifting to another intention is equivalent. We preserve non-zero probability for intentions representing tasks already performed by the human, because the human may perform the same task again due to failed robot attempts. The update step of Bayesian filtering is as follows.

$$P(g_{t+T_p}|x^{h,r}_{1:t+T_p}) \propto P(x^{h,r}_{t+1:t+T_p}|x^{h,r}_{1:t}, g_{t+T_p})P(g_{t+T_p}|x^{h,r}_{1:t}) \tag{3}$$

We derive the prediction model $P(x^{h,r}_{t+1:t+T_p}|x^{h,r}_{1:t}, g_{t+T_p})$ by the recursive expression below.

$$\begin{aligned} &P(x^{h,r}_{t+1:t+\tau}|x^{h,r}_{1:t}, g_{t+T_p}) \\ &= P(x^h_{t+\tau}|x^{h,r}_{1:t+\tau-1}, x^r_{t+\tau}, g_{t+T_p})P(x^r_{t+\tau}|x^{h,r}_{1:t+\tau-1}, g_{t+T_p}) \\ &\quad P(x^{h,r}_{t+1:t+\tau-1}|x^{h,r}_{1:t}, g_{t+T_p}) \\ &= P(x^h_{t+\tau}|x^{h,r}_{1:t+\tau-1}, x^r_{t+\tau}, g_{t+\tau})P(x^r_{t+\tau}|x^{h,r}_{1:t+\tau-1}, g_{t+\tau-1}) \\ &\quad P(x^{h,r}_{t+1:t+\tau-1}|x^{h,r}_{1:t}, g_{t+T_p}) \\ &= P(x^h_{t+\tau}|x^{h,r}_{1:t+\tau-1}, x^r_{t+\tau}, g_{t+\tau})P(x^{h,r}_{t+1:t+\tau-1}|x^{h,r}_{1:t}, g_{t+T_p}) \end{aligned} \tag{4}$$

The second equality in Equation 4 is because $g_{t+1:t+T_p}$ are assumed the same. The third equality in Equation 4 is because we use the maximum likelihood estimate of the intention as input to the downstream robot control pipeline, which leads to a deterministic mapping from the observed state history to the next observed robot state. We apply the recursive expression and get the prediction model.

$$P(x^{h,r}_{t+1:t+T_p}|x^{h,r}_{1:t}, g_{t+T_p}) = \prod_{\tau=1}^{T_p} P(x^h_{t+\tau}|x^{h,r}_{1:t+\tau-1}, x^r_{t+\tau}, g_{t+\tau}) \tag{5}$$

B. Extension to Hierarchical Intentions

Consider G^l_t and G^{l+1}_t of the hierarchical intention structure in Figure 2(b). The intention transition models follow Equation 2. We expect α^{l+1} is larger than α^l because a higher-level intention is less frequently changed. We derive the prediction model for G^{l+1}_t in terms of the counterpart for G^l_t, where $P(g^l_{t+T_p}|x^{h,r}_{1:t}, g^{l+1}_{t+T_p})$ is related to $P(g^l_{t+T_p}|x^{h,r}_{1:t})$ and prior knowledge of relations between $g^l_{t+T_p}$ and $g^{l+1}_{t+T_p}$.

$$\begin{aligned} &P(x^{h,r}_{t+1:t+T_p}|x^{h,r}_{1:t}, g^{l+1}_{t+T_p}) \\ &= \sum_{g^l_{t+T_p}} P(x^{h,r}_{t+1:t+T_p}|x^{h,r}_{1:t}, g^l_{t+T_p})P(g^l_{t+T_p}|x^{h,r}_{1:t}, g^{l+1}_{t+T_p}) \end{aligned} \tag{6}$$

The procedure of Hierarchical Intention Tracking is as follows. First, perform intention tracking at a lower level conditioned on a fixed higher level intention to get $P(g^l_{t+T_p}|x^{h,r}_{1:t}, g^{l+1}_{t+T_p})$. Second, derive the prediction model at the higher level $P(x^{h,r}_{t+1:t+T_p}|x^{h,r}_{1:t}, g^{l+1}_{t+T_p})$ by Equation 6. Repeat the first and the second steps until we reach the highest level L, where we can perform Bayesian filtering without conditioning on any other intentions to get $P(g^L_{t+T_p}|x^{h,r}_{1:t+T_p})$. We can then get the posteriors from higher levels to lower levels.

$$\begin{aligned} &P(g^l_{t+T_p}|x^{h,r}_{1:t+T_p}) \\ &= \sum_{g^{l+1}_{t+T_p}} P(g^l_{t+T_p}|x^{h,r}_{1:t}, g^{l+1}_{t+T_p})P(g^{l+1}_{t+T_p}|x^{h,r}_{1:t+T_p}) \end{aligned} \tag{7}$$

We apply hierarchical intention tracking to our use case as in Figure 1(c), which is a two-layer intention hierarchy composed of a low-level task intention G^1_t and a high-level interactive intention G^2_t. The sample space of G^1_t comprises four task-related goals ($i = 1, 2, 3, 4$) and failure recovery (FR). The sample space of G^2_t comprises cooperation (CO) and coexistence (CE) interactive intentions. Intention transition at both levels are controlled by α^1 and α^2. In our use case, $P(g^1_{t+T_p}|x^{h,r}_{1:t}, g^2_{t+T_p})$ is as follows.

$$\begin{aligned} P(FR|x^{h,r}_{1:t}, CO) &= 1 \\ P(i|x^{h,r}_{1:t}, CO) &= 0 \\ P(FR|x^{h,r}_{1:t}, CE) &= 0 \\ P(i|x^{h,r}_{1:t}, CE) &= \frac{P(i|x^{h,r}_{1:t})}{\sum_{j=1}^4 P(j|x^{h,r}_{1:t})} \end{aligned} \tag{8}$$

In addition, when conditioned on the coexistence interactive intention, we can simplify the prediction model by

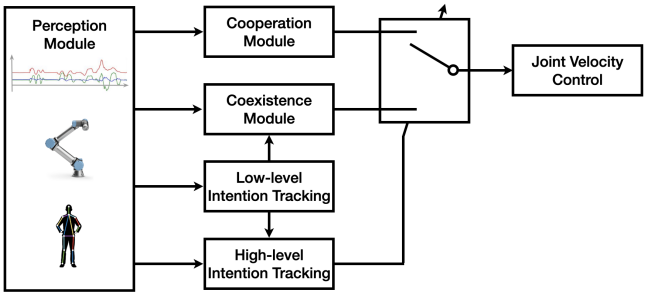

Fig. 3: The architecture of Hierarchical Intention Tracking (HIT) based human-robot collaboration system. Perception module includes robot wrist force/torque measurements, robot proprioception, and human skeleton tracking. Low-level intention tracking module takes human wrist positions as input to track task intention. High-level intention tracking module takes human wrist positions, robot end-effector positions, and low-level intention tracking outputs to track interactive intention. Based on the tracked interactive intention, coexistence or cooperation module is selected for motion planning. Coexistence module plans motion according to the task plan generated from the tracked task intentions. Cooperation module plans motion by following human's guidance.

removing the effect of the robot on human motion.

$$\begin{aligned} & P(x^{h,r}_{t+1:t+T_p} | x^{h,r}_{1:t}, G^2_{t+T_p} = CE) \\ &= \prod_{\tau=1}^{T_p} P(x^h_{t+\tau} | x^h_{1:t+\tau-1}, G^2_{t+\tau} = CE) \\ &= P(x^h_{t+1:t+T_p} | x^h_{1:t}, G^2_{t+T_p} = CE) \end{aligned} \tag{9}$$

IV. Collaborative Robot Architecture

We introduce Hierarchical Intention Tracking (HIT) based human-robot collaboration system as presented in Figure 3.

A. Experimental Setup for Collaborative Assembly

A human and a robot work together on an assembly task in close proximity. The task involves assembling four pairs of Misumi Waterproof E-Model Crimp Wire Connectors [40]. The connectors are in asymmetrical shapes and have tight clearances. At the beginning of one experiment trial, female parts are separately placed in four square regions denoting task intentions part $1, 2, 3$, and 4. All male parts are initially placed in a rectangular region denoting preparation area. The human picks up a male part from the preparation area and aligns it to a female part within the corresponding region. The aligned parts are left at the same region, and the robot reaches to them and perform the pushing action. The human decides the sequence of task intentions for executing part alignment. The robot knows locations of task intention regions, but not the task sequence. The aligned male part may fall off due to table shaking from robot motion. The robot may not push on an ideal position and fail the assembly. The collaborative robot system should robustly handle these cases.

B. Robot and Perception Setup

The robot is a UR5e arm equipped with a Robotiq Hand-E Gripper. An embedded sensor measures force and torque on the end-effector at 500 Hz. The robot is controlled by joint velocity commands and operated at a reduced speed (30%).

Fig. 4: Human skeleton detection by OpenPose on frames from a side-view and a top-view camera. Multiple cameras are used to address occlusion which happens frequently in close-proximity human-robot collaboration.

Intel RealSense RGBD Cameras provide a top-down and a side view of the shared space to alleviate occlusion. We run OpenPose [41] to detect human skeleton positions from both views. We apply Kalman Filter to these detections and track 3D positions of human right wrist at 30 Hz. All modules are executed using Robot Operating System (ROS).

C. Low-level Intention Tracking

Task intentions among four parts and the preparation area are tracked at 30 Hz. We adapt Mutable Intention Filter (MIF) with Intention-aware Linear Model (ILM) as the prediction model [35] to the assembly task. MIF is a particle filtering variant for single-layer intention tracking when human is not affected by robot. The inputs are observed 3D human wrist trajectories and potential task intention regions, and the output is the probability distribution over the task intentions.

D. High-level Intention Tracking

Interactive intentions are tracked at 5 Hz. The low-level intention tracking module feeds $P(g^1_{t+T_p} | x^{h,r}_{1:t}, G^2_{t+T_p} = CE)$ to Equation 6, but a probabilistic prediction model for G^1_t is still required to compute the prediction model for G^2_t. Thus, we use a Gaussian variant of ILM

$$x^h_{t+1} = x^h_t + \frac{\tilde{d}_t}{||g^1_t - x^h_t||} \left(g^1_t - x^h_t\right) + w_t \tag{10}$$

by which human wrist moves to the goal region g^1_t at an average speed of $\tilde{d}_t$ during the observation window, with a Gaussian process noise w_t. Note for g^1_t as failure recovery, the goal region is defined as a neighboring region around the robot end-effector position which would move through time.

E. Planning and Control

Coexistence Module is developed for concurrent task execution without interruption. A task set of all parts is initialized. If the most likely task intention belongs to the task set, and its probability stays above 80% over 1.5 second, the human is detected aligning parts and the intention is

TABLE I: Results of the ablative pilot study. (*) Completion time of Cooperation Mode Baseline is for reference but does not provide fair comparison against other systems. The robot is implemented at a consistently low speed in Coexistence Mode Baseline or when controlled by the coexistence module of HIT system, in order to enforce safety of human subjects. An effective speed and separation monitoring module [19] would significantly improve the completion time for Coexistence Mode Baseline and HIT System to match the efficiency of Cooperation Mode Baseline without affecting performance in other metrics. We will leave the development for future work.

System	Completion Time (sec)	Automated Path (m)	Guided Path (m)	Human Force (N)	Human Energy (J)	Number of Failures
Coexistence Mode	104.2±17.6	3.80±0.65	N/A	N/A	N/A	1.2
Cooperation Mode	35.2±3.4$^{(*)}$	N/A	2.36±0.17	5.51±0.39	21.04±2.52	0.0
HIT System	97.2±15.5	3.40±0.51	0.65±0.37	0.48±0.24	5.71±3.30	0.0

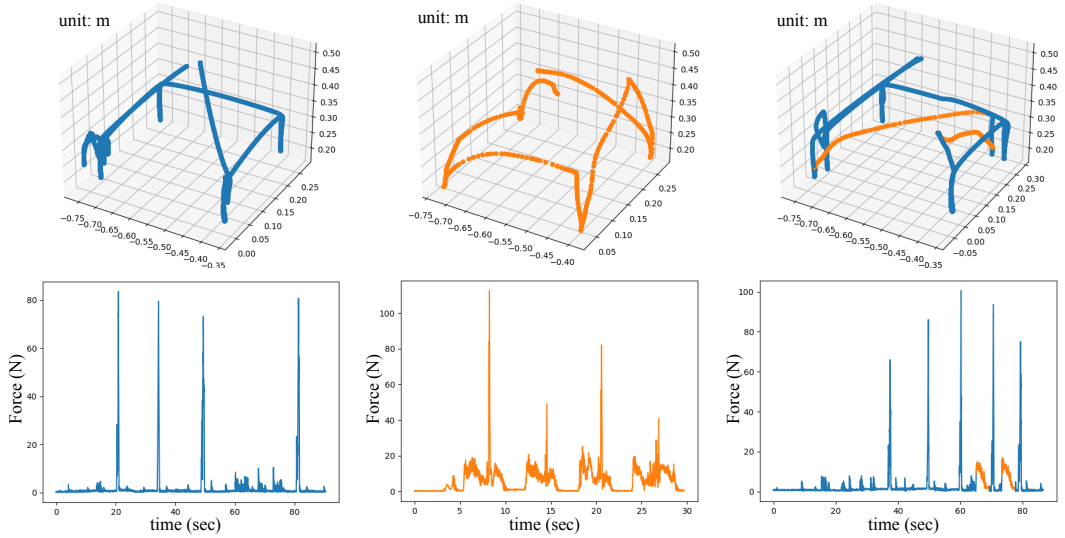

Fig. 5: Trajectory and force visualization in (a) Coexistence Mode Baseline; (b) Cooperation Mode Baseline; (c) Hierarchical Intention Tracking system. Blue denotes coexistence mode and orange denotes cooperation mode. Note (c) shows a trial where the robot failed in assembly steps, the human recovered the failure, and the robot performed assembly steps again. The robot did more than 4 assembly steps so there are more than 4 peaks in force visualization in (c).

moved from the task set to an ongoing task queue. An intention is popped out of the ongoing task queue and pushed into a ready task queue when its probability drops below 25%, which indicates the alignment in the corresponding region is finished. The robot plans trajectories to reach to and perform the pushing action on the part corresponding to the task intention popped out of the ready task queue. The robot moves repulsively from the human wrist by Artificial Potential Field to avoid collision [42]. Force control is implemented to detect the completion of pushing action. Note a small uniform random noise is added to the goal position of the robot, which is intended to emphasize the robustness by controlling the failure rate of push.

Cooperation Module is developed for manual guidance to handle failure recovery. If the probability of cooperation interactive intention is above 90% over 0.5 sec, the robot is switched to the cooperation module. An attractive Artificial Potential Field is implemented so the robot end-effector approaches the human wrist. Admittance control is activated after intended contact is detected, and the human starts guiding the robot to the desired ready-to-push position. Detection on the end of human intervention by force measurement switches the robot back to the coexistence module, and the robot immediately executes the pushing action to recover the assembly failure. A running mean of joint velocities is used to avoid discontinuous control input when switching between coexistence and cooperation modules [8].

V. Ablative Pilot Study

We conduct a pilot study of HIT system against two human-robot collaboration baselines: Coexistence Mode Baseline and Cooperation Mode Baseline. In the Coexistence Mode Baseline, the robot executes low-level intention tracking. The robot is able to perform concurrent task execution with the human but not able to recover the failure. In the Cooperation Mode Baseline, admittance control is executed and the robot is passively compliant. Human guidance is required all the time to move the robot and perform pushing actions. These uni-modal systems are chosen as baselines because coexistence-only or cooperation-only robot are prevalent solutions in industry [7], [43].

Five human subjects participate in the pilot study. Each subject do ten trials of the collaborative assembly task on each system. Evaluation metrics include (1) completion time, (2) length of the automated path executed by the coexistence module, (3) length of the path when robot is guided by the human, (4) average force applied by the human throughout the trial, (5) total energy the human spent throughout the trial, and (6) average number of assembly failures.

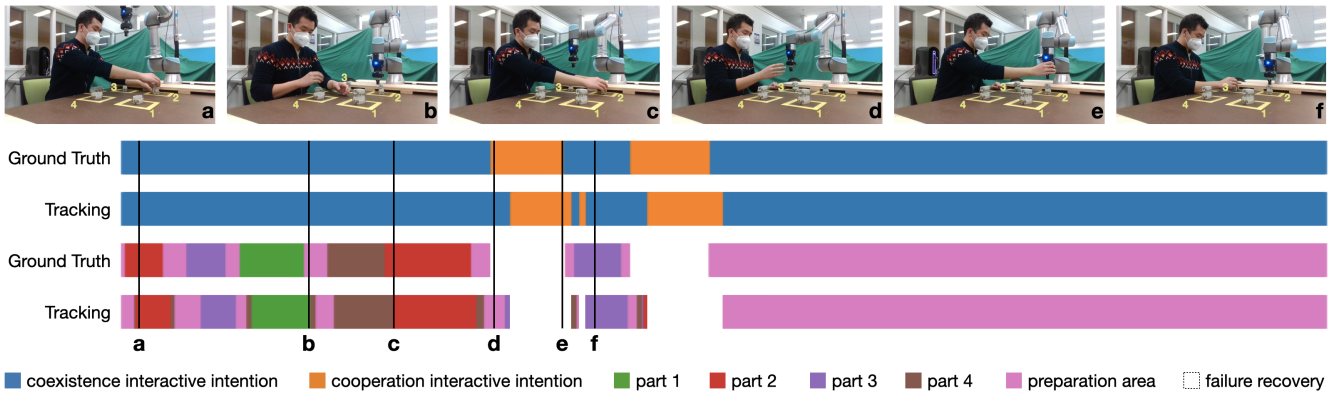

Fig. 6: High-level and low-level intention tracking in a trial of the collaborative assembly task. The tracking bar shows the most likely intention tracked at each time step. Snapshots show a failure recovery example. (a) The human aligns parts. (b) The robot fails to assemble the parts. (c) The human realigns the part. (d) The human shows cooperation interactive intention by reaching to the end-effector. (e) The human guides the robot to the appropriate ready-to-push position. (f) The robot recovers the failure by pushing the parts again. The frame-wise accuracy is 90.4% for low-level intention and 94.5% for high-level intention.

The quantitative evaluation is presented in Table I. Coexistence Mode Baseline has limited efficiency in terms of the completion time and the automated path due to two reasons. First, the robot is required to work at a low speed with human in close proximity during the coexistence mode. Second, the robot may not be able to reach an appropriate ready-to-push position for the aligned parts. To prevent assembly failure, the human has to adjust the position of aligned parts multiple times. The robot can get trapped in a local minimum due to human avoidance and goal-directed motion when the human is doing adjustment. There are in average 4.6 times of adjustments in one trial, but there are still 1.2 pairs of parts failed to be assembled in the Coexistence Mode Baseline. Cooperation Mode Baseline has the shortest completion time at the cost of excessive human effort. As demonstrated in Figure 5(b), the human guides the robot all the time in order to reach aligned parts and perform the pushing actions. Continuous guidance for the goal-reaching motion consumes much human energy and leads to human fatigue, while large impact during the pushing action exhibits poor ergonomics which may cause workplace injuries.

In contrast to the baselines, HIT system effectively balances coexistence and cooperation to combine advantages from both sides. HIT system has a shorter length of the guided path and significantly decreased human force and energy than the Cooperation Mode Baseline, because human only needs to guide the robot when failure recovery is needed. Figure 5(c) demonstrates that the cooperation mode is active only for two short periods, while most goal-reaching motion and all pushing actions are executed by the coexistence module. HIT system also has a shorter length of the automated path compared against Coexistence Mode Baseline, since the human does not have to readjust the part positions with the capability to recover the failure. Though the total length of path is longer than the Coexistence Mode Baseline, HIT system still has a shorter completion time. The robot is allowed to move faster than the coexistence mode when manually guided.

Figure 6 shows the effective performance of both high-level and low-level intention tracking. In this visualized trial, the human aligns the parts in the order of $\{2,3,1,4\}$. The assembly of part 2 and part 3 are initially failed and then recovered. The snapshots in Figure 6 show the sequence of assembling part 2 including the initial attempt and the failure recovery. Note that high-level and low-level intention tracking modules have not only different tracking frequencies, but also different sensitivities on evolving intentions. The low-level intention tracking has a higher sensitivity and a shorter tracking delay. This short delay is critical to the accuracy of the high-level prediction model which uses the probability distribution over low-level task intentions. The high-level intention tracking has a lower sensitivity so the most likely high-level interactive intention is less likely to jump between coexistence and cooperation, which can result in the instability of robot control.

VI. Conclusions

We propose the concept of hierarchical intention tracking to take into account the continuously changing human intention and its hierarchical structure in the context of human-robot collaboration. We introduce a Hierarchical Intention Tracking (HIT) based human-robot collaboration system which effectively integrates the coexistence and cooperation modules. We demonstrate seamless interaction, robust failure recovery and enhanced ergonomics of the HIT system against baselines through real-world experiments on a collaborative assembly task. In future work, we will develop speed and separation monitoring to improve the efficiency of the coexistence module. We attempt to generalize our framework from rule-based intentions to latent human states learned from variational inference or MCMC, and compare performance with these learning-based methods [44], [45]. We plan to perform extensive human subject experiments. We would like to explore a time-varying intention transition setting, such as how to incorporate the prior knowledge of the assembly task sequence by learning from task recordings.

REFERENCES

[1] C. Pérez-D'Arpino and J. A. Shah, "Fast target prediction of human reaching motion for cooperative human-robot manipulation tasks using time series classification," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2015, pp. 6175–6182.

[2] Y. Cheng, L. Sun, C. Liu, and M. Tomizuka, "Towards efficient human-robot collaboration with robust plan recognition and trajectory prediction," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 2602–2609, 2020.

[3] T. Wojtara, M. Uchihara, H. Murayama, S. Shimoda, S. Sakai, H. Fujimoto, and H. Kimura, "Human–robot collaboration in precise positioning of a three-dimensional object," *Automatica*, vol. 45, no. 2, pp. 333–342, 2009.

[4] K. I. Alevizos, C. P. Bechlioulis, and K. J. Kyriakopoulos, "Physical human–robot cooperation based on robust motion intention estimation," *Robotica*, vol. 38, no. 10, pp. 1842–1866, 2020.

[5] Y. Wang and F. Zhang, *Trends in control and decision-making for human-robot collaboration systems*. Springer, 2017.

[6] S. Huang, M. Ishikawa, and Y. Yamakawa, "A coarse-to-fine framework for accurate positioning under uncertainties—from autonomous robot to human–robot system," *The International Journal of Advanced Manufacturing Technology*, vol. 108, no. 9, pp. 2929–2944, 2020.

[7] V. Villani, F. Pini, F. Leali, and C. Secchi, "Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications," *Mechatronics*, vol. 55, pp. 248–266, 2018.

[8] J. Cacace, A. Finzi, and V. Lippiello, "Shared admittance control for human-robot co-manipulation based on operator intention estimation." in *ICINCO (2)*, 2018, pp. 71–80.

[9] K. Driggs-Campbell and R. Bajcsy, "Identifying modes of intent from driver behaviors in dynamic environments," in *IEEE International Conference on Intelligent Transportation Systems (ITSC)*, 2015, pp. 739–744.

[10] M. Geravand, F. Flacco, and A. De Luca, "Human-robot physical interaction and collaboration using an industrial robot with a closed control architecture," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2013, pp. 4000–4007.

[11] D. Nicolis, A. M. Zanchettin, and P. Rocco, "Human intention estimation based on neural networks for enhanced collaboration with robots," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 1326–1333.

[12] S. Grahn, B. Langbeck, K. Johansen, and B. Backman, "Potential advantages using large anthropomorphic robots in human-robot collaborative, hand guided assembly," *Procedia CIRP*, vol. 44, pp. 281–286, 2016.

[13] W. Kim, J. Lee, L. Peternel, N. Tsagarakis, and A. Ajoudani, "Anticipatory robot assistance for the prevention of human static joint overloading in human–robot collaboration," *IEEE robotics and automation letters*, vol. 3, no. 1, pp. 68–75, 2017.

[14] A. D. Wilbert, B. Behrens, O. Dambon, and F. Klocke, "Robot assisted manufacturing system for high gloss finishing of steel molds," in *International Conference on Intelligent Robotics and Applications*. Springer, 2012, pp. 673–685.

[15] J. Shi, G. Jimmerson, T. Pearson, and R. Menassa, "Levels of human and robot collaboration for automotive manufacturing," in *Proceedings of the Workshop on Performance Metrics for Intelligent Systems*, 2012, pp. 95–100.

[16] P. Tsarouchi, A.-S. Matthaiakis, S. Makris, and G. Chryssolouris, "On a human-robot collaboration in an assembly cell," *International Journal of Computer Integrated Manufacturing*, vol. 30, no. 6, pp. 580–589, 2017.

[17] S. Heydaryan, J. Suaza Bedolla, and G. Belingardi, "Safety design and development of a human-robot collaboration assembly process in the automotive industry," *Applied Sciences*, vol. 8, no. 3, p. 344, 2018.

[18] A. De Luca and F. Flacco, "Integrated control for phri: Collision avoidance, detection, reaction and collaboration," in *IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, 2012, pp. 288–295.

[19] P. Svarny, M. Tesar, J. K. Behrens, and M. Hoffmann, "Safe physical hri: Toward a unified treatment of speed and separation monitoring together with power and force limiting," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 7580–7587.

[20] S. Papanastasiou, N. Kousi, P. Karagiannis, C. Gkournelos, A. Papavasileiou, K. Dimoulas, K. Baris, S. Koukas, G. Michalos, and S. Makris, "Towards seamless human robot collaboration: integrating multimodal interaction," *The International Journal of Advanced Manufacturing Technology*, vol. 105, no. 9, pp. 3881–3897, 2019.

[21] F. Flacco, T. Kröger, A. De Luca, and O. Khatib, "A depth space approach to human-robot collision avoidance," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2012, pp. 338–345.

[22] P. A. Lasota, G. F. Rossano, and J. A. Shah, "Toward safe close-proximity human-robot interaction with standard industrial robots," in *IEEE International Conference on Automation Science and Engineering (CASE)*, 2014, pp. 339–344.

[23] J.-H. Chen and K.-T. Song, "Collision-free motion planning for human-robot collaborative safety under cartesian constraint," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 4348–4354.

[24] Y. Li and S. S. Ge, "Human–robot collaboration based on motion intention estimation," *IEEE/ASME Transactions on Mechatronics*, vol. 19, no. 3, pp. 1007–1014, 2013.

[25] A. Rasouli, I. Kotseruba, T. Kunic, and J. K. Tsotsos, "PIE: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 6262–6271.

[26] K. D. Katyal, G. D. Hager, and C.-M. Huang, "Intent-aware pedestrian prediction for adaptive crowd navigation," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 3277–3283.

[27] O. Mazhar, B. Navarro, S. Ramdani, R. Passama, and A. Cherubini, "A real-time human-robot interaction framework with robust background invariant hand gesture detection," *Robotics and Computer-Integrated Manufacturing*, vol. 60, pp. 34–48, 2019.

[28] C.-M. Huang and B. Mutlu, "Anticipatory robot control for efficient human-robot collaboration," in *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2016, pp. 83–90.

[29] H. Ahn, S. Choi, N. Kim, G. Cha, and S. Oh, "Interactive text2pickup networks for natural language-based human–robot collaboration," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3308–3315, 2018.

[30] X. T. Truong and T. D. Ngo, "Social interactive intention prediction and categorization," in *ICRA Workshop on MoRobAE-Mobile Robot Assistants for the Elderly, Montreal Canada, May 20-24*, 2019.

[31] Z. Wang, K. Mülling, M. P. Deisenroth, H. Ben Amor, D. Vogt, B. Schölkopf, and J. Peters, "Probabilistic movement modeling for intention inference in human–robot interaction," *The International Journal of Robotics Research*, vol. 32, no. 7, pp. 841–858, 2013.

[32] C. Park, J. Ondřej, M. Gilbert, K. Freeman, and C. O'Sullivan, "Hi robot: Human intention-aware robot planning for safe and efficient navigation in crowds," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016, pp. 3320–3326.

[33] A. M. Zanchettin and P. Rocco, "Probabilistic inference of human arm reaching target for effective human-robot collaboration," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017, pp. 6595–6600.

[34] P. Du, Z. Huang, T. Liu, K. Xu, Q. Gao, H. Sibai, K. Driggs-Campbell, and S. Mitra, "Online monitoring for safe pedestrian-vehicle interactions," in *IEEE International Conference on Intelligent Transportation Systems (ITSC)*, 2020.

[35] Z. Huang, A. Hasan, K. Shin, R. Li, and K. Driggs-Campbell, "Long-term pedestrian trajectory prediction using mutable intention filter and warp lstm," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 542–549, 2021.

[36] J.-H. Hong, Y.-S. Song, and S.-B. Cho, "Mixed-initiative human–robot interaction using hierarchical bayesian networks," *Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 37, no. 6, pp. 1158–1164, 2007.

[37] J. Cacace, R. Caccavale, A. Finzi, and V. Lippiello, "Interactive plan execution during human-robot cooperative manipulation," *IFAC-PapersOnLine*, vol. 51, no. 22, pp. 500–505, 2018.

[38] C. Zhu, Q. Cheng, and W. Sheng, "Human intention recognition in smart assisted living systems using a hierarchical hidden markov model," in *IEEE International Conference on Automation Science and Engineering (CASE)*, 2008, pp. 253–258.

[39] K. P. Murphy, *Dynamic bayesian networks: representation, inference and learning*. University of California, Berkeley, 2002.

[40] K. Kimble, K. Van Wyk, J. Falco, E. Messina, Y. Sun, M. Shibata, W. Uemura, and Y. Yokokohji, "Benchmarking protocols for evaluating small parts robotic assembly systems," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 883–889, 2020.

[41] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 7291–7299.
[42] O. Khatib, "Real-time obstacle avoidance for manipulators and mobile robots," in *Autonomous Robot Vehicles*. Springer, 1986, pp. 396–404.
[43] G. Michalos, N. Kousi, P. Karagiannis, C. Gkournelos, K. Dimoulas, S. Koukas, K. Mparis, A. Papavasileiou, and S. Makris, "Seamless human robot collaborative assembly–an automotive case study," *Mechatronics*, vol. 55, pp. 194–211, 2018.
[44] D. P. Losey, H. J. Jeon, M. Li, K. Srinivasan, A. Mandlekar, A. Garg, J. Bohg, and D. Sadigh, "Learning latent actions to control assistive robots," *Autonomous Robots*, vol. 46, no. 1, pp. 115–147, 2022.
[45] M. Zolotas and Y. Demiris, "Disentangled sequence clustering for human intention inference," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 9814–9820.