Predicting Human Intentions in Human–Robot Hand-Over Tasks Through Multimodal Learning

Weitian Wang, *Member, IEEE*, Rui Li, Yi Chen, Yi Sun, and Yunyi Jia

***Abstract*— In human–robot shared manufacturing contexts, product parts or tools hand-over between the robot and the human is an important collaborative task. Facilitating the robot to figure out and predict human hand-over intentions correctly to improve the task efficiency in human–robot collaboration is therefore a necessary issue to be addressed. In this study, a teaching-learning-prediction (TLP) framework is proposed for the robot to learn from its human partner's multimodal demonstrations and predict human hand-over intentions. In this approach, the robot can be programmed by the human through demonstrations utilizing natural language and wearable sensors according to task requirements and the human's working preferences. Then the robot learns from human hand-over demonstrations online via extreme learning machine (ELM) algorithms to update its cognition capacity, allowing the robot to use its learned policy to predict human intentions actively and assist its human companion in hand-over tasks. Experimental results and evaluations suggest that the human may program the robot easily by the proposed approach when the task changes, as the robot can effectively predict hand-over intentions with competitive accuracy to complete the hand-over tasks.**

***Note to Practitioners*—This article is motivated by human–robot hand-over problems in smart manufacturing contexts. Product parts or tools delivery in worker–robot partnerships is an important collaborative task. We develop a teaching-learning-prediction (TLP) framework for the robot to learn from its human partner's multimodal demonstrations and predict human hand-over intentions. The robot can be taught by human through natural language and wearable sensing information. The extreme learning machine (ELM) approach is employed for the robot to build its cognition capacity to predict human intentions actively and assist its human companion in hand-over tasks. We demonstrate that the proposed approach presents distinct and effective advantages to facilitate human–robot hand-over tasks in collaborative manufacturing contexts.**

***Index Terms*— Extreme learning machine (ELM), human–robot hand-over, intention prediction, learning from demonstrations, natural language, wearable sensors.**

Manuscript received October 5, 2020; revised January 22, 2021; accepted March 19, 2021. This article was recommended for publication by Associate Editor Q. Xu and Editor C. Seatzu upon evaluation of the reviewers' comments. This work was supported by National Science Foundation under Grant IIS-1845779. *(Corresponding author: Yunyi Jia.)*

Weitian Wang and Rui Li are with the Department of Computer Science, Montclair State University, Montclair, NJ 07043 USA.

Yi Chen, Yi Sun, and Yunyi Jia are with the Department of Automotive Engineering, Clemson University, Greenville, SC 29607 USA (e-mail: yunyij@clemson.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TASE.2021.3074873.

Digital Object Identifier 10.1109/TASE.2021.3074873

I. Introduction

ROBOTICS technology has served humans in numerous application fields [1]–[3], especially product assembly in manufacturing contexts. Assembly tasks can benefit from both sides of the robot and the human in a proper human–robot collaboration. Basically, the human and the robot have to interact with each other in order to successfully co-conduct the assembly of products when they are in the shared workspace [4]. During the human–robot collaboration process, handing over product parts to each other is an inevitable and fundamental task. Similar to human–human interaction in object delivery, human–robot hand-over refers to an action triggered by the human or its robot companion to meet the goal of delivering a part to each other.

In recent years, robots have been increasingly used in collaborative assembly works, conducting human-to-robot or robot-to-human hand-over tasks is gradually necessary and significant. In manufacturing contexts, most of collaborative robots are generally programmed by off-line operating devices, such as FlexPendants and workstations [5], which require a lot of human effort when the tasks update and need to be recorded. Robots mimicking human demonstrations to achieve human–robot collaborative tasks is a promising and effective method [6], by which the robots can be programmed in real time and serve as assistants in human–robot hand-over even when the collaborative tasks are changed. Meanwhile, human workers are not required to have expertise and professional programming skills but just utilize demonstrations to code robots to conduct new tasks. In addition, some effective robot learning control methods such as adaptive fuzzy full-state and output-feedback control [7] and admittance-based control [8] have been proposed for human–robot collaborative tasks.

Apart from improving the robot's bootstrapping in product parts hand-over by learning from human, another significant issue to augment human–robot hand-over tasks is empowering the robot to be capable of predicting human intention in the hand-over process [9]. This implies that, in order to make the interaction more intuitive, the robot not only needs to be aware of the ongoing hand-over behaviors, but also annotates and predicts the imminent hand-over intentions of its partner. Efficient human–robot cooperation requires harmonious decisions and intentions of both sides in each step of a shared task. As a consequence, predicting the intentions of product parts hand-over by humans is a critical link for enabling the robots to effectively act and interact with their human partners in collaborative working surroundings.

Therefore, to improve the collaboration efficiency and decrease the efforts of humans in human–robot co-assembly tasks, we hope that robots and human workers can achieve the product parts hand-over seamlessly in a natural and easy-to-implement way by using multimodal information. It implies that the robot should possess high-level cognition capacities of human action understanding and intention prediction just like the human counterpart in human–robot hand-over processes.

To this end, we propose a teaching-learning-prediction (TLP) framework for the robot to learn from its human partners and predict hand-over intentions in human–robot collaborative tasks. The natural multimodal information (natural language and natural wearable sensing information) of the human is employed as the human–robot interactive input in hand-over processes. By taking advantage of all sides (human speech instructions, muscle activities, and forearm rotations) of the natural multimodal information, the human and the robot are able to communicate with each other more efficiently than single modality, allowing human–robot collaborative tasks to be productively executed. The extreme learning machine (ELM) is utilized in the proposed TLP framework. ELM was developed for generalized single-hidden-layer feedforward neural networks in which the hidden layer need not to be alike [10]. By means of its remarkable advantages, we employ the ELM for the robot to learn and predict human hand-over intentions in our approach. In what follows, several related studies of human intention understanding and prediction in human–robot collaboration are presented in Section II. The TLP framework and its working procedures are described in Section III. The detailed TLP modeling methodologies including human teaching, robot learning, and human intention prediction are expounded in Sections IV and V. We carry out a series of experiments in real-world human–robot collaborative contexts to testify and evaluate the efficiency and accuracy of the proposed approach in Sections VI and VII. The conclusions of research findings and the future work of this study are provided in Section VIII.

The contributions of this work can be summarized as follows.

1) We investigate and propose a novel solution to facilitate human-to-robot/robot-to-human hand-over tasks using natural multimodal information.
2) We develop a natural and intuitive human–robot interactive interface using natural wearable sensing and natural language, based on which the robot could learn and predict human hand-over intentions in collaborative tasks.
3) We extend the application of robots learning from human demonstrations and propose a systematic TLP framework for human–robot partnerships to co-accomplish hand-over tasks, which can reduce human manual-programming effort and enable human–robot collaboration productivity.

II. Related Work

In human–robot hand-over processes, it is critical to have the robot understand human intentions. Several related works have been conducted in various applications recently.

Human–robot hand-over intentions were usually assumed to be known in most human–robot collaboration research works. Among these studies to recognize the human hand-over intentions, there are two typical approaches: the vision-based approach and the physical contact-based approach, which are used to detect and control the interaction process in human–robot hand-over. For the first approach, Grigore *et al.* [11] demonstrated that integrating joint action understanding in human–human interaction and human–robot collaboration can enhance the performance of robot-to-human hand-over processes using the VICON motion capture system. Aleotti *et al.* [12] presented an approach for robot–human object hand-over by taking user comfort into account via a Kinect. For the second approach, Cakmak *et al.* [13] conducted a study to acquire human preference information for hand-over tasks. The human–robot interaction process was sensed by the robot's arm stiffness. Nagata *et al.* [14] solved the human–robot delivery problem by using the gripper force–torque sensing method to detect and trigger the interaction process.

However, the vision-based human hand-over intention recognition approach usually rests with working surroundings. The recognition results are easily affected by the dynamic backgrounds and limited working areas. For example, on automotive assembly lines, the automotive equipment and transport vehicles keep moving in the assembly process and the human workers sometimes need to operate the parts in the areas that may not be covered by the vision system. Therefore, such factors will have an impact on the recognition performance of the vision-based approach. In the second approach, although the robot build-in sensing devices can be used to detect the human–robot hand-over process, they require a physical and significant force contact between the human and the robot, which limits the working ranges and may cause safety concerns for the workers.

As we mentioned above, the method of learning from human demonstrations has been explored for empowering robots to be capable of imitating human behaviors. By this way, the human workers can program robots via demonstrations to have them perform the tasks actively [7]. Likewise, the physical touch approach (using robot build-in sensing devices or joysticks for human–robot interaction) and nonphysical touch approach (using wearable sensors or vision systems for human–robot interaction) are mainly developed. In the first approach, the joystick teaching [15] and kinesthetic teaching [16] are usually employed by the human to perform the work demonstrations to the robot. Additionally, the nonphysical touch approach facilitated the human–robot collaboration a lot with high-level methods such as wearable sensors-based learning [17], vision system-based learning [18], and natural language-based learning [19], which augmented the robot's capacity for handling more complex works. Upon receiving the demonstration information, the robot is able to construct its action strategies through the embedded learning algorithms [20], [21]. Then it imitated the human using its learned policy to conduct tasks in the working environment.

However, these approaches in human–robot teaching and learning processes mainly try to make the robot repeat human

behaviors, which are not sufficient for several human–robot collaborative contexts such as product parts hand-over in automotive assembly tasks. Because not only do we need the robots to mimic humans, but we also need the robot to predict human intentions and collaborate with its human partner.

Human intention prediction in human–robot interaction is an available solution to such co-conduct problems such as product parts or tools hand-over in human–robot teams. Through the Markov model, Tanaka *et al.* [22] proposed an approach for the human action trajectory prediction in the discrete workspace and developed a robot motion planner in human–robot hand-over tasks. Cohen *et al.* [23] utilized the Kinect sensor to develop a "metaphor-free" interface and proposed a human intention prediction method via the cognitive science computational model. From the image sequences in human–object interaction, Song *et al.* [24] presented a probabilistic graphical model to predict human manipulation intentions. Hawkins *et al.* [25] proposed a graphical model to predict human intentions by a probabilistic manner in the context that the robot assisted its human partner to perform different tasks.

Although human intentions are predicted in these studies, they usually employ additional sensing (e.g., vision systems) to annotate ParaFirstLine-Indhuman data manually [26], which may cost a lot of human efforts. However, few investigations have developed an online TLP approach in human–robot collaboration, especially the dexterous hand-over tasks. Additionally, different human workers usually present diverse assembly or hand-over preferences, which require the robot to be rapidly programmed according to different individuals or new tasks. Consequently, developing a method which can enable the robot to understand human intentions effectively as well as facilitate the human–robot hand-over to be performed is a considerable issue to explore and address.

To solve this, a TLP framework is developed by means of natural multimodal human information based on ELM algorithms. In human–robot collaboration, product parts or tools hand-over processes are dynamically conducted in real time by the human and the robot via online natural language and natural wearable sensing information. By taking advantage of the proposed method, the human could program/teach the robot easily employing partial demonstrations through natural multimodal information in line with the new task requirements and his/her working preferences. After that, the robot is able to autonomously build its cognitive competence utilizing the TLP approach to understand and predict human hand-over intentions. Moreover, the robot can make its action planning decisions actively to deliver, pick up, or reject the parts using its learned strategies in human–robot hand-over processes. By leveraging the developed framework on practical experiments of human–robot collaborative tasks, we demonstrate that the product parts hand-over processes between the human and the robot are conducted accurately and effectively.

III. TLP Framework

The overarching vision of this work is to empower the robot to learn from partial human demonstrations online and predict human hand-over intentions as well as help the human in hand-over tasks flexibly. Fig. 1 shows the TLP framework, which includes: 1) human teaching via multimodal information; 2) robot learning from human demonstrations; and 3) human–robot hand-over in collaborative tasks.

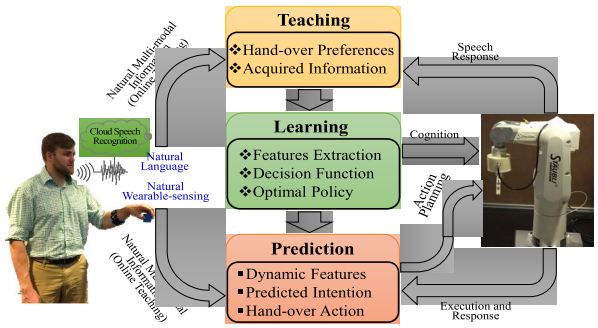


Fig. 1. TLP framework.

The robot must be able to understand human intentions in order to deliver/pick up a part to/from the human. To this end, before human–robot collaboration in a robot-unknown task, the human presents partial demonstrations of each hand-over intention to the robot online by natural language instructions and a wearable sensory system according to his/her personalized working preferences and the product parts or tools needed to be handed over. Likewise, the robot also responds to the human through speech once it gets the intention information. In the teaching process, the quantitative elements of human intentions are extracted from the multimodal information to be employed in the learning algorithm to update the robot's hand-over cognition.

In the robot learning process, the extracted human hand-over intention information is further parameterized online by different sets of features correspondingly. The processed natural language instructions are learning objectives, and the wearable sensing information works as knowledge sets for the robot. After that, these features are utilized by the robot as inputs to the TLP model based on the ELM algorithm to construct its cognition capacity of understanding different human hand-over intentions. The robot will learn from the human, just like the child learns skills from the adult, to construct its customized task cognition. Additionally, it can be obviously concluded that the robot's cognition capacities are different on account of diverse individuals' hand-over preferences and task requirements. In this process, the TLP model can decrease human coding efforts to a great extent compared to the traditional offline robot programming methods.

Based on the learned policy, the robot can employ the human intention prediction algorithm in the TLP model to make hand-over decisions. In the prediction process, the human just uses natural wearable sensing information to work with the robot in hand-over tasks, which will reduce human manual-programming efforts as well as help the human get rid of repetitive speech. The optimal hand-over policy is generated for the robot according to the online dynamic input information from the wearable sensory system. After that, the robot utilizes the prediction results to cooperate with the human to finish the parts or tools hand-over by

executing different hand-over actions accordingly. Once the human changes or the task is updated, the human may employ this proposed easy-to-use TLP model to teach the robot to join the new task quickly.

IV. Teaching Robots Using Natural Multimodal Information

A. Multimodal-Based Hand-Over Intentions

In human–robot hand-over tasks, human partners' intentions, which are known as what the human wants to do for the robot or what the human wants the robot to do, are essentially significant for the collaboration. For example, "I need an object from you" or "I want to give you an object." These intentions can be detected and represented via numerous methods, such as by vision systems or force sensors. To make it more robust than a single modal human–robot interface for hand-over tasks, we employ natural multimodal information, including natural language and natural gesture information, of the human as human–robot interactive input in hand-over processes.

The natural language instructions in human–robot hand-over tasks are utilized to describe human intentions. Synchronously, the gesture information is applied to parameterize the intentions expressed by human motions, which contain forearm posture and muscle activity. As shown in Fig. 1, the natural language commands are decoded based on the Google Cloud speech-to-text (StT) platform [27]. Meanwhile, the human forearm action information is acquired and figured out via a wearable sensory system. Therefore, each human intention I_{H} can be formulized by the target factors $\mathbf{I_{NL}}$ extracted from natural language instructions and the features $\mathbf{I_{WS}}$ extracted from human forearm rotations and electromyography (EMG) signals which are collected by the wearable sensory system

$$I_{\mathrm{H}} = \{(\mathbf{i_{NL}}, \mathbf{i_{WS}}) | \mathbf{i_{NL}} \in \mathbf{I_{NL}},\ \mathbf{i_{WS}} \in \mathbf{I_{WS}}\}. \tag{1}$$

It can be seen that, for different human workers, they usually have their own preferred hand-over manners to teach the robot. This information is able to be included in $\mathbf{I_{WS}}$. Moreover, in human–robot collaboration, the I_{H}, which contains the robot's learning objectives and knowledge sets, is employed in the TLP model to teach the robot to construct its cognition capacity that can be utilized to predict human intentions and handle the hand-over process.

Generally, different humans have individual working preferences, which can be known as the manners that humans employed for interacting with robots in this study, to execute their hand-over actions, respectively, so it is difficult to code a universal program to cover their hand-over features. As described in (1), via the I_{H} in the proposed TLP model, different individuals are able to teach the robot online with their working preferences. However, for a given hand-over intention, the human cannot perform all the possible observation information to the robot in the teaching process. Therefore, we just collect partial observation data of each hand-over intention to online feed into the TLP model. In addition, it is more effective to use multimodal information than single modality in human–robot hand-over processes. For instance, if the human plans to give an object to the robot, his/her palm may face downward. The hidden object will not be recognized by the top-down camera, which is used in many industrial applications. However, if more modalities such as EMG signals from human forearm muscle activities are added in this process, the object can be detected based on the features of EMG signals. When the human wants to get an object from the robot, it will be more intuitive and efficient for him/her to tell the robot about the requested object using speech instructions than other coexisting body languages. Therefore, the multimodal information is significantly necessary to facilitate human–robot collaborative tasks. Section IV-B and IV-C present how to teach the robot via the multimodal information.

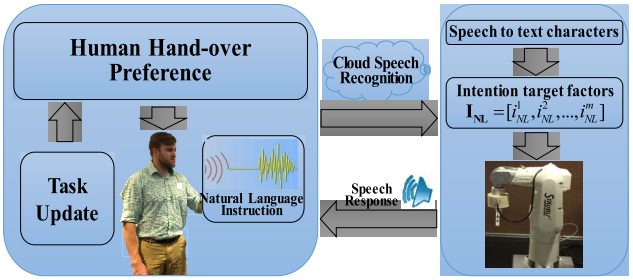


Fig. 2. Natural language-based human teaching in human–robot hand-over tasks.

B. Natural Language-Based Teaching

Human hand-over intentions are described and demonstrated by speech instructions in the teaching process. A natural language collection application (APP) is developed by us to acquire human speech information. We employ a headset for the human to wear to decrease the noise disturbance when the human performs speech instructions to the robot. After being collected, the speech information is transferred to the Google Cloud StT platform [27] to be decoded as text embodiment. Then it is organized as available arguments for the robot learning algorithms based on our previous studies [28], [29], in which a human natural language utterance representation framework is developed. Additionally, as shown in Fig. 2, a voice synthesis module in the control system is employed for the robot to reply to its human partner with corresponding hand-over instructions.

If the tasks are changed, the human will perform new hand-over demonstrations to the robot by his/her working preferences according to task requirements. For a given collaboration task, there are normally a set of hand-over intentions. In this work, we employ a target factor $\mathbf{I_{NL}}$ to parameterize the hand-over intention demonstrated by natural language instructions. Therefore, the target factors $\mathbf{I_{NL}}$ extracted from this teaching process can be expressed as

$$\mathbf{I_{NL}} = \left[\mathbf{i^1_{NL}}, \mathbf{i^2_{NL}}, \ldots, \mathbf{i^m_{NL}}\right]^{\mathrm{T}}, \quad m \in [1,\ M] \tag{2}$$

where M denotes the number of types of human hand-over intentions such as "Give a part" and "Need a part," $\mathbf{i^m_{NL}}$ denotes the mth hand-over intention target factor from M. $\mathbf{I_{NL}}$ is employed as the learning objective for the robot in our TLP model. Since this study is mainly focusing on the robot

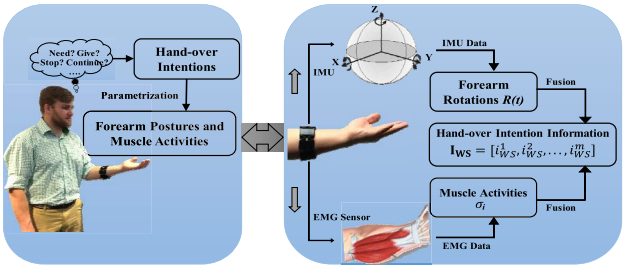


Fig. 3. Natural wearable sensing-based teaching in human–robot hand-over tasks.

learning and prediction approach development and evaluation, we assume that the human natural language recognition accuracy from the Google Cloud StT platform is 100%.

C. Natural Wearable Sensing-Based Teaching

A wearable sensory system is used when the human teaches the robot via natural language in hand-over tasks. By the wearable sensors, the human forearm action information (muscle activity signals and postures) can be collected. The sensory system that we employ is Myo [30], which contains eight channels of EMG sensors [31] and an inertial measurement unit (IMU) [32] and can be easily worn on the human forearm. As presented in Fig. 3, the human forearm motion information can be calculated via the IMU when the hand-over intentions are shown to the robot [33]. Simultaneously, the muscle activity signals generated by the forearm are able to be quantified through EMG sensors. These muscle activity signals generally correspond to human finger activities that can be extracted to estimate human hand behaviors such as expressing different finger motions and holding different kinds of parts.

As illustrated in Fig. 3, when the human prepares to deliver/need a part to/from the robot in the collaborative task, his forearm posture information can be acquired and quantified by the IMU, which includes the raw three-axis angular velocity data and the raw three-axis acceleration data. Based on these signals, Euler angles, including roll, pitch, and yaw, are used to parameterize the forearm motions in the 3-D workspace [33]. The roll–pitch–yaw angles are described as

$$\mathbf{R(t)} = [\phi(t), \theta(t), \psi(t)]^{\mathrm{T}} \tag{3}$$

where t is the sampling time, which is 0.02 s in this work.

Meanwhile, the EMG signals from the human forearm's muscles, which would be used to measure human finger activities are acquired by the EMG sensors. They can be described as

$$\mathbf{E(t)} = [e_1(t),\ e_2(t), \ldots, e_n(t)]^{\mathrm{T}} \tag{4}$$

where t is the sampling time of the EMG sensor, $e(t)$ is each EMG sensor's output, and n is the number of EMG channels, which is 8 in our device. In this work, the EMG data are sampled at the frequency of 50 Hz instead of the inherent frequency of 200 Hz in sync with the IMU.

Along with the finger activities, the raw EMG signals are different sets of discrete points with positive and negative elements [33]. Moreover, we find that the electric potentials formed by muscle cells have evident impacts on the dispersion of EMG signals. Therefore, to take advantage of the EMG data accurately, we employ the standard deviation (StD) of the EMG signals to extract the features of finger activities [33]. The StD can be evaluated as

$$\sigma_i = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(e_i(k) - \frac{1}{K}\sum_{k=1}^{K} e_i(k)\right)^2} \tag{5}$$

where $e_i(k),\ k = 1, 2, \ldots, K$ is a set of EMG signals, and K is the window size for determining the number of EMG signals to calculate the standard deviation. K is selected as 200 in this work.

Furthermore, the hand-over intention information extracted from the natural wearable sensing-based teaching can be characterized by

$$\mathbf{I_{WS}} = \left[\mathbf{i^1_{WS}}, \mathbf{i^2_{WS}}, \ldots, \mathbf{i^m_{WS}}\right], \quad m \in [1, M] \tag{6}$$

where $\mathbf{i^m_{WS}} = [i^m_\phi, i^m_\theta, i^m_\psi, i^m_{\sigma 1}, i^m_{\sigma 2}, \ldots, i^m_{\sigma 8}]$ is an eleven-element vector and denotes the hand-over intentions indicated by the forearm rotations and EMG signals, M denotes the number of types of human hand-over intentions, $\mathbf{i^m_{WS}}$ is the mth hand-over intention information from M. $\mathbf{I_{WS}}$ is utilized as the knowledge set for the robot in our TLP model.

D. Human Intentions Normalization

Since the extracted natural gesture information in human hand-over intentions has no consistent dimension and order of magnitude, they need to be normalized to decrease the computational complexity and augment the result accuracy for the TLP model. In this study, the gesture information is normalized by

$$i^*_{\mathrm{ws}} = \frac{i - i_{\mathrm{mean}}}{i_{\max} - i_{\min}} \tag{7}$$

where i_{mean}, $i_{\max}$, and $i_{\min}$ are the average, the maximum element, and minimum element of each set of hand-over intention information acquired by the wearable sensory system.

Additionally, for the natural language-based teaching, target factors are parameterized as a set of sequence numbers, which can be denoted by $\mathbf{i^{m*}_{NL}}$ and where $\mathbf{i^{m*}_{NL}} \in \mathbf{I^*_{NL}}$. Therefore, according to (1), (2), and (6), the processed human intention information can be represented by

$$I^*_{\mathrm{H}} = \left\{\left(\mathbf{i^{m*}_{NL}}, \mathbf{i^{m*}_{WS}}\right) \middle| \mathbf{i^{m*}_{NL}} \in \mathbf{I^*_{NL}},\ \mathbf{i^{m*}_{WS}} \in \mathbf{I^*_{WS}}\right\}. \tag{8}$$

Similarly, it can be observed that the learning objective and knowledge set in (8) are $\mathbf{i^{m*}_{NL}}$ and $\mathbf{i^{m*}_{WS}}$, respectively.

Equation (8) describes a generalized representation of human intentions that are parameterized by natural language information and gesture information. For example, for the intention "Give a part," it can be characterized by three specific rotation angles and eight specific EMG signals. The values of these parameters are not predefined, they are acquired online depending on the human worker's personalized working preferences and the task requirements. That is to say, for different human workers, these parameters might be different. It also means that the kinds of human intention denoted by these parameters are not predefined.

V. Learning and Predicting Human Intentions Based on ELM

A. Extreme Learning Machine

The ELM is employed for the robot to learn and predict human hand-over intentions in our TLP model. In the robot learning process, the extracted hand-over intention information $\mathbf{I}^{*}_{\mathbf{WS}}$ works as the ELM input. The hidden layer outputs serve as mapping results by the activation function about the input features of hand-over intentions. Given a d-dimensional hand-over intention feature x, the feature mapping from hidden nodes can be described by

$$\mathbf{h}(\mathbf{x}) = [g(a_1, b_1, x), g(a_2, b_2, x), \ldots, g(a_L, b_L, x)] \quad (9)$$

where L is the number of hidden nodes in the ELM, $g(\mathbf{a}, b, \mathbf{x})$ is a nonlinear piecewise continuous function meeting ELM universal approximation capability theorems [34]–[36], $\mathbf{a}$ is the input weight vector, b is the hidden node bias, and $\{(\mathbf{a}_i,\ b_i)\}_{i=1}^{L}$ are stochastically generated in accordance with any continuous probability distribution [37]. In ELM algorithms, $\mathbf{h}(\mathbf{x})$ maps the data from the d-dimensional feature space to the L-dimensional hidden-layer feature space $\mathbf{H}$. For the output layer, the output function for generalized single-hidden-layer feedforward neural networks can be expressed as

$$f_L(\mathbf{x}) = \sum_{i=1}^{L} \beta_i h_i(\mathbf{x}) = \mathbf{h}(\mathbf{x})\boldsymbol{\beta} \quad (10)$$

where $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_L]^{\mathrm{T}}$ is the output weight vector between the hidden layer of L nodes to the M output nodes. In this work, the input feature sample $\mathbf{x} = [\mathbf{i}^{\mathbf{m}*}_{\mathbf{WS}}]$, $\mathbf{i}^{\mathbf{m}*}_{\mathbf{WS}} \in \mathbf{I}^{*}_{\mathbf{WS}}$. Accordingly, it can be seen that each ELM output node $f_j(\mathbf{x})$, $j \in [1, M]$ denotes a kind of hand-over intention in human–robot hand-over processes.

B. Human Intentions Learning

In human–robot collaboration, the target of the hand-over intention learning by the robot is to figure out the optimal strategy to understand what the human demonstrated to the robot in the teaching process. In other words, the robot is able to employ datasets collected by the wearable sensory system to optimize its cognition to minimize the deviation with learning objectives demonstrated by human speech instructions. Compared to the traditional machine learning, the ELM can reach the minimal training error as well as the minimal norm of output weights [37]. Therefore, by means of the ELM, the robot learning goal is to minimize the approximation error and the norm of the output weights

$$\mathrm{Min}_{\beta \in \mathbf{R}^{L\times m}} \ \|\mathbf{H}\boldsymbol{\beta} - \mathbf{T}\|^2 \text{ and } \|\boldsymbol{\beta}\| \quad (11)$$

where $\mathbf{H}\boldsymbol{\beta}$ is the actual output of the ELM, $\mathbf{H}$ is the hidden-layer output matrix

$$\mathbf{H} = \begin{bmatrix} \mathbf{h}(\mathbf{x}_1) \\ \vdots \\ \mathbf{h}(\mathbf{x}_N) \end{bmatrix} = \begin{bmatrix} g(\mathbf{a}_1, b_1, \mathbf{x}_1) & \cdots & g(\mathbf{a}_L, b_L, \mathbf{x}_1) \\ \vdots & \vdots & \vdots \\ g(\mathbf{a}_1, b_1, \mathbf{x}_N) & \cdots & g(\mathbf{a}_L, b_L, \mathbf{x}_N) \end{bmatrix}$$

and $\mathbf{T}$ is the robot learning target matrix

$$\mathbf{T} = \begin{bmatrix} \mathbf{t}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{t}_N^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} \mathbf{t}_{11} & \cdots & \mathbf{t}_{1m} \\ \vdots & \vdots & \vdots \\ \mathbf{t}_{N1} & \cdots & \mathbf{t}_{Nm} \end{bmatrix}$$

where N is the number of input human intention feature samples. In this work, each learning target $\mathbf{t}_{Nm} = [\mathbf{i}^{\mathbf{m}*}_{\mathbf{NL}}]$, $\mathbf{i}^{\mathbf{m}*}_{\mathbf{NL}} \in \mathbf{I}^{*}_{\mathbf{NL}}$. Furthermore, the robot learning process can be characterized as a constrained-optimization issue with multioutput nodes [37]–[39]

$$\begin{aligned} &\mathrm{Min}_{\beta \in R^{L\times m}}: \ L_{\mathrm{ELM}} = \frac{1}{2}\left(\|\boldsymbol{\beta}\|^2 + C\sum_{i=1}^{N}\|\xi_i\|^2\right) \\ &\text{S.T. } \mathbf{h}(\mathbf{x}_i)\boldsymbol{\beta} = \mathbf{t}_i^{\mathrm{T}} - \xi_i^{\mathrm{T}}, \quad i \in [1,\ N] \end{aligned} \quad (12)$$

where $\xi_i = [\xi_{i,1}, \xi_{i,2}, \ldots, \xi_{i,m}]^{\mathrm{T}}$ is the approximation error vector of the m output nodes with respect to the input human intention feature sample in the robot learning process, and C denotes the regularization factor.

In human–robot hand-over tasks, generally, the number of input human intention feature samples N is far more than the number of hidden nodes L. Based on the Karush–Kuhn–Tucker (KKT) theorem and Lagrange multiplier method [37], the optimal learned policy $\boldsymbol{\beta}$ from human demonstrations can be got by

$$\boldsymbol{\beta}^{*} = \left(\frac{\mathbf{I}}{C} + \mathbf{H}^{\mathrm{T}}\mathbf{H}\right)^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{T} \quad (13)$$

where $\mathbf{I}$ is an identity matrix of dimension L. In this work, we select the number of input feature samples as 5000 for the robot to learn.

C. Human Intentions Prediction

Based on the learned hand-over policy, the robot can work online to cooperate with its human companion for collaborative tasks. If the human performs hand-over actions, the new acquired intention information $\mathbf{x}$ can be fed to the TLP model online. As a consequence, by the ELM algorithms, the prediction results of each hand-over intention are able to be evaluated as

$$\mathbf{f}(\mathbf{x}) = \mathbf{h}(\mathbf{x})\boldsymbol{\beta} = \mathbf{h}(\mathbf{x})\left(\frac{\mathbf{I}}{C} + \mathbf{H}^{\mathrm{T}}\mathbf{H}\right)^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{T}. \quad (14)$$

In order to reduce the online computational complexity of the TLP model, we utilize a kernel function for the ELM based on Mercer's conditions in robot learning and prediction process. By using the kernel function, we do not need to consider if the feature mapping $\mathbf{h}(\mathbf{x})$ is known or not. In addition, sometimes the input features, which cannot be well linearly partitioned in a low dimensional space by the simple ELM, can be transferred into a higher dimensional space and better linearly partitioned by the kernel ELM.

The radial basis function (RBF) kernel function is employed in this work

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right) \quad (15)$$

where σ is the spread factor. Therefore, the online outputs of the ELM algorithms are able to be expressed by

$$\mathbf{f}(\mathbf{x}) = \mathbf{h}(\mathbf{x})\left(\frac{\mathbf{I}}{C} + \mathbf{H}^{\mathrm{T}}\mathbf{H}\right)^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{T} = \left(\frac{\mathbf{I}}{C} + \mathbf{\Omega}\right)^{-1}\begin{bmatrix} K(\mathbf{x}, \mathbf{x}_1) \\ \vdots \\ K(\mathbf{x}, \mathbf{x}_N) \end{bmatrix}^{\mathrm{T}}\mathbf{T} \tag{16}$$

where

$$\mathbf{\Omega} = \begin{bmatrix} K(\mathbf{x}_1, \mathbf{x}_1) & \cdots & K(\mathbf{x}_1, \mathbf{x}_N) \\ \vdots & \vdots & \vdots \\ K(\mathbf{x}_N, \mathbf{x}_1) & \cdots & K(\mathbf{x}_N, \mathbf{x}_N) \end{bmatrix}.$$

Moreover, based on the evaluated possible human intentions, which are denoted by the vector $\mathbf{f}(\mathbf{x})$ in (14), we can get that the predicted hand-over intention of the online input features maps to the index number of the output node with maximum value. Consequently, the online predicted hand-over intention is

$$I_p = \underset{i \in [1,\ M]}{\arg\max}\ f_i(\mathbf{x}) \tag{17}$$

where $f_i(\mathbf{x}) \in \mathbf{f}(\mathbf{x})$.

Therefore, by taking advantage of (17), the robot is able to predict human intentions online and further to collaborate with its human partner in collaborative hand-over tasks.

VI. Experimental Setup

A. Experimental Platform

As shown in Fig. 4, the proposed approach is applied on a multimodal-based collaborative robotics research platform (MCRRP), which is built by our laboratory for the studies on human–robot collaboration. The MCRRP contains a six DOF robot, an operator station, an engineer station, and several multimodal human–robot interactive interfaces, including the wearable sensing system, the natural language processing system, and the 3-D vision system. As described in Section IV, we employ a wearable sensory system Myo and a natural language processing system, including text-to-speech (TtS) and StT, as the human–robot interface. The operator station (configured with 16-GB RAM and Intel Core i7-5500U CPU) is employed for fusing human–robot interactive information and running the TLP model.

B. Typical Hand-Over Intentions in Collaborative Tasks

In this work, ten participants are recruited to conduct a hand-over task by picking up/delivering objects from/to a robot in realistic human–robot collaborative contexts. By studying and investigating their hand-over manners, we focus on three types of commonly used human hand-over intentions: "Give," "Need," and "Mode Adjustment" in this study. The "Give" intention means that the human wants to deliver something to the robot; the "Need" intention denotes that the human wants something from the robot; the "Mode Adjustment" intention means that the human intends to adjust the robot action modes in the hand-over process, such as "Stop" (the human wants the robot to stop in the hand-over process), "Continue" (the human wants the robot to resume the ongoing task in the hand-over process), "Speed Up" (the human wants the robot to move faster in the hand-over process), and "Slow Down" (the human wants the robot to move slower in the hand-over process). The overall framework in this study is generalized and not limited to these three types of human hand-over intentions.

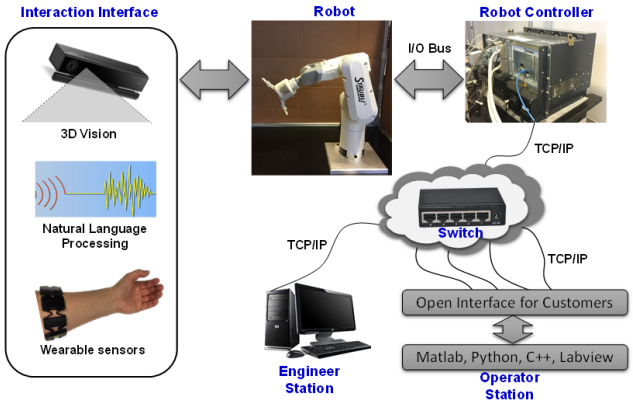


Fig. 4. MCRRP.

C. Task Description

We conducted a set of typical human–robot hand-over experiments based on practical collaborative tasks to verify the efficiency and advantages of the proposed approach. Before performing the experiments, the participants are presented with a one-minute instruction, in which the usage and steps of the proposed TLP approach are included. In order to correctly collect human hand-over intention information in the teaching process, the participants are asked to perform forearm gestures first, then perform each corresponding speech instruction to the robot. For each participant, 5000 sets of hand-over intention features are collected. The human and the robot work on the hand-over tasks in a laboratory-based realistic manufacturing context. During the experiment, there are some random noise from outside and some random moving people from other projects. In the teaching process, the human can teach the robot using partial demonstrations via the multimodal information of each hand-over intention according to his working preferences. In addition, the human randomly picks up some parts from the parts container, and then expresses the "Give" intention to the robot. The given parts are with different weights and sizes: large and heavy parts, large parts, heavy parts, medium parts, and small parts. Because of the limited payload of the robot gripper in the experiment, only the large part, the medium part, and the small part can be picked up by the robot. Therefore, the robot will make a prediction to accept the part or not based on its learned policies from the human. In the third experiment, we verify the robot's prediction capacity via four "Mode Adjustment" intentions, which include "Stop," "Continue," "Speed up," and "Slow down" in the robot working process. Likewise, the "Need" intention is also performed to the robot in the fourth experiment to demonstrate if the robot could predict its partner's request. Since the factors of

personalization and customization are involved in the TLP model, different individuals are able to teach the robot with diverse working habits or preferences for other different tasks using the proposed approach. All parts with different weights and sizes in this task are supposed to be picked up by one human hand. In this work, the sampling frequency of the wearable sensory system is 50 Hz.

D. Parameters Tuning

As mentioned before, in robot learning and prediction processes, the ELM randomly generates the hidden node parameters $(\mathbf{a}, b)$, which are independent of the input intention information. Additionally, the feature mapping matrix in the ELM is irrelevant to the target. Therefore, it can be observed that only the regularization factor C and the kernel function parameter σ need to be tuned in the proposed TLP model.

In this study, 5000 sets of samples are collected from the ten participants and ten kinds of hand-over intentions (500 sets per participant and 50 set per intention) to tune these parameters, in which 4000 sets are randomly selected for offline learning to train the ELM and the rest sets are used for offline prediction. First, the kernel function parameter σ is set as 1, 10, 100, and 1000 to tune the regularization factor C, respectively. However, we find that the outputs of the relationships among the learning accuracy, the prediction accuracy, and the regularization factor are very similar at different values. Fig. 5 presents the variation of the learning accuracy and prediction accuracy with the regularization factor C at $\sigma = 1$, $\sigma = 10$, and $\sigma = 1000$, separately. It can be seen that along with the increase of C, the learning accuracy and prediction accuracy gradually go up as well until a point which starts a constant at $C = 5600$. In addition, by employing the tuned regularization factor, we get the variation of the learning accuracy and prediction accuracy with the kernel function parameter σ, as shown in Fig. 6. It can be observed that the learning accuracy is declined when the σ rises. Meanwhile, the prediction accuracy is increased gradually but becomes decreasing at $\sigma = 1.2$. As a result, we set $C = 5600$ and $\sigma = 1.2$ in this work.

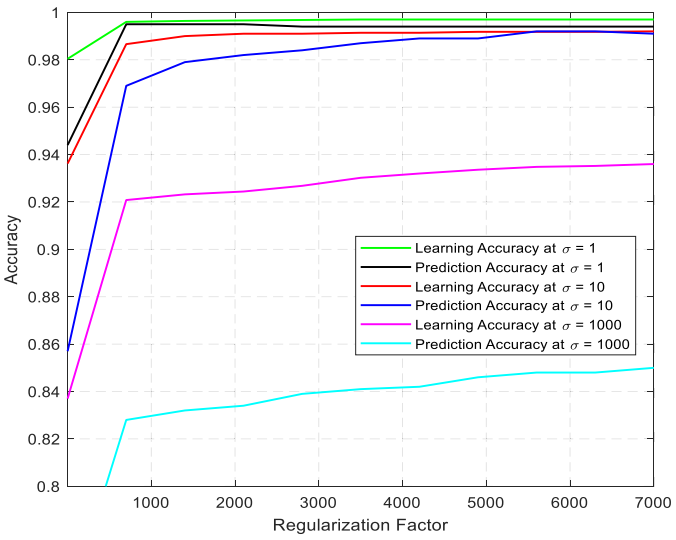


Fig. 5. Variation of the learning accuracy and prediction accuracy with the regularization factor C at $\sigma = 1$, $\sigma = 10$, and $\sigma = 1000$.

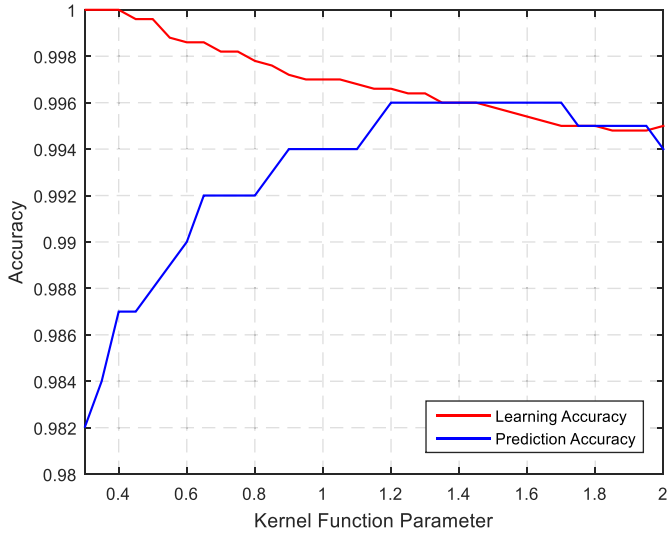


Fig. 6. Variation of the learning accuracy and prediction accuracy with the kernel function parameter σ at $C = 5600$.

VII. Results and Evaluations

A. Learning From Hand-Over Demonstrations

The robot learning from hand-over demonstrations of one of the participants is shown in Fig. 7. Based on the developed TLP model, the human presents the robot with different hand-over intentions using the natural language and natural gesture information according to his preference. As mentioned above, the human cannot demonstrate all the possible features of each hand-over intention in the teaching process, hence it can be considered that only partial intention information is observed and employed to predict the human potential future intentions.

In this work, the hand-over intentions include ten subclasses, which are "Give the large and heavy part," "Give the large part," "Give the heavy part," "Give the medium part," "Give the small part," "Stop the work," "Continue the work," "Speed up," "Slow down," and "Need a part." As described in Section VI-C, in the "Give" intentions, only the large part, the medium part, and the small part can be picked up because of the payload of the robot gripper. Consequently, in human–robot hand-over processes, the robot not only needs to predict the "Give" intentions, but also could actively distinguish if the delivered part should be accepted or not. After getting the demonstrations in each intention, the robot repeats them to its human partner by a synthesized voice. In the learning process, 5000 sets (500 sets per intention) of hand-over intention features are online acquired from the demonstrations for the robot to build its cognition capacity based on the ELM algorithms. The human spends about 100 s teaching the robot using the described ten hand-over intentions. Each case costs 10 s and all the cases go through with an order from (a) to (j) described in Fig. 7.

B. Human-to-Robot Hand-Over

The scenario procedure of the online human-to-robot hand-over is shown in Fig. 8. In this study, we define an

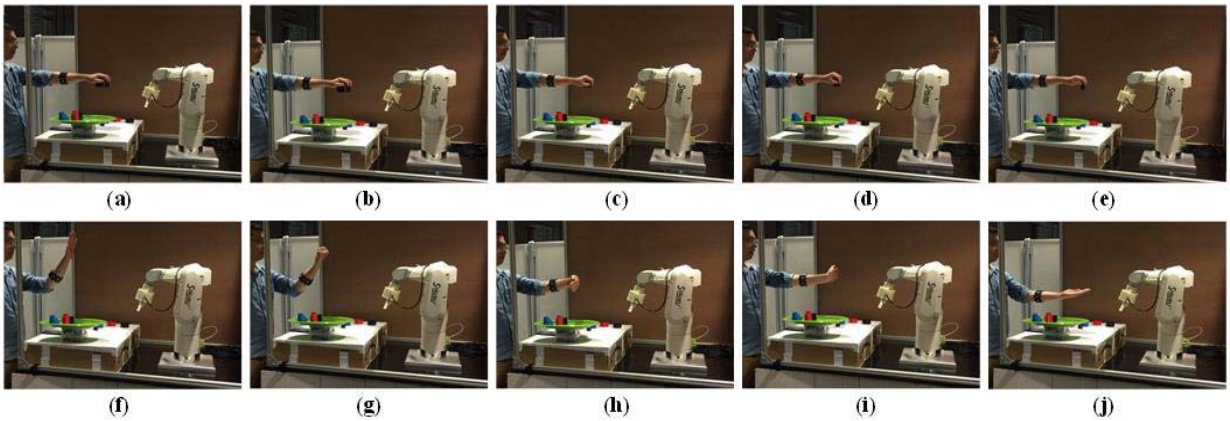


Fig. 7. Robot learning from hand-over demonstrations. (a) Give the large and heavy part. (b) Give the large part. (c) Give the heavy part. (d) Give the medium part. (e) Give the small part. (f) Stop the work. (g) Continue the work. (h) Speed up. (i) Slow down. (j) Need a part.

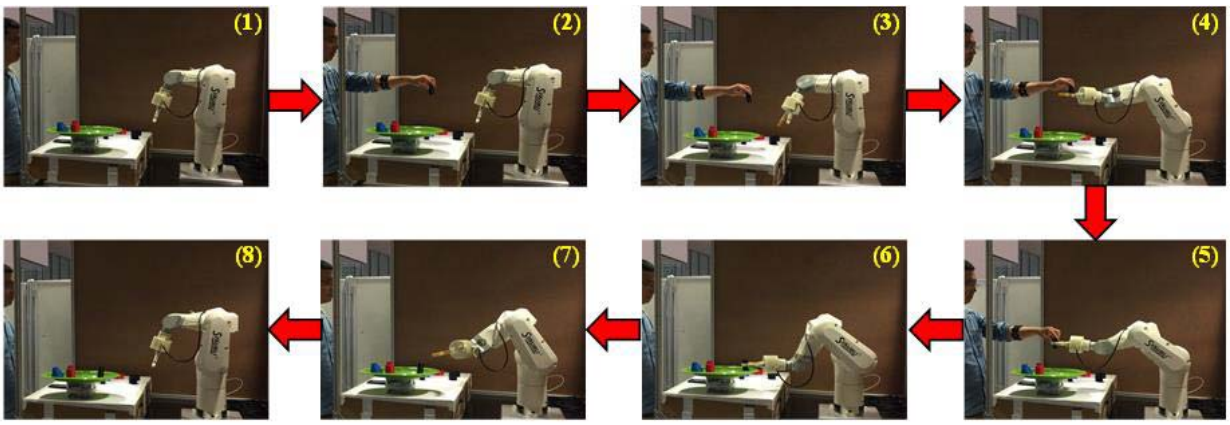


Fig. 8. Procedure of the human-to-robot hand-over intention prediction (accept the small part).

action set for the robot to plan corresponding motions to map the learned human behavior in human–robot collaboration. The action mapping will be activated by the human intention prediction results then sends homologous control signals to plan motions for the robot to interact with its human partner. For example, if the prediction result indicates that the weight of the delivered part exceeds the payload of the robot gripper, the robot will give a refusal action to the human, or it will move to the human and pick up the part. The human's "Give" intention and the part's weight in Fig. 8 can be predicted by using the multimodal information including forearm rotations and EMG signals acquired by the wearable sensory system instead of a single top-down camera since it cannot "see" the object at this pose. As shown in Fig. 8, the first two clips present that the human prepares to give a small part to the robot. Based on the TLP model, the robot is able to make a prediction about its human partner's intention. As presented in Fig. 8(3)–(5), the robot starts to pick up the delivered part and responds to the human by the speech with "Small part." Afterward, the robot puts the part into the desired container and goes back to the initial position in Fig. 8(6)–(8).

However, the human gives a large heavy part to the robot in Fig. 9(1). After getting the online intention based on the multimodal information, as depicted in Fig. 9(2)–(4), the robot actively shakes its wrist to present a refusing to its human partner by speaking "Large heavy part." From these processes, it can be seen that, by taking advantage of the multimodal information of the human, the robot successfully predicts the "Give the small part" intention and "Give the large heavy part" intention to make active decisions to pick up or reject the delivered part from the human through the TLP model.

C. Mode Adjustment

During the robot's working process, the human performs the "Stop," "Continue," "Slow down," and "Speed up" intentions online, separately. By employing the TLP model, the scenarios of the robot's predictions are shown in Fig. 10. It can be observed that, once the human presents the "Stop the work" intention, the robot is stopped in Fig. 10(a-1) and (a-2). As shown in Fig. 10(a-3) and (a-4), when the human performs the "Continue the work" intention, the robot resumes its task actively. Fig. 10(b-1) and (b-2) presents that the robot reduces its speed when the human presents the "Slow down" intention. Oppositely, the robot accelerates its working process when the human performs the "Speed up" intention, which is shown in Fig. 10(b-3) and (b-4). Therefore, it is apparent that, based on the TLP model, the multimodal information collected

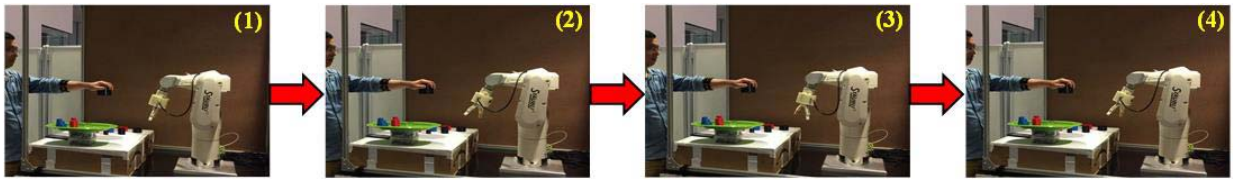


Fig. 9. Procedure of the human-to-robot hand-over intention prediction (reject the large heavy part).

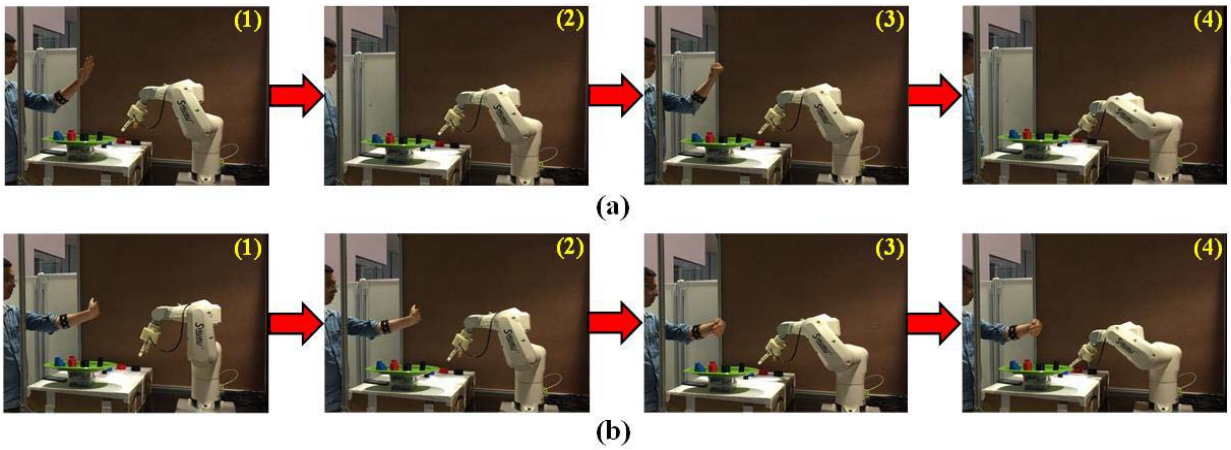


Fig. 10. Robot's responses to the human intentions. (a) Stop the work and continue the work. (b) Slow down and speed up.

from the human's forearm enables the robot to make correct predictions of the "Mode Adjustment" intentions in human–robot hand-over tasks.

D. Robot-to-Human Hand-Over

The scene process of the online robot-to-human hand-over is exhibited in Fig. 11. A 3-D vision system is developed for the robot to locate the parts in the task. The vision system is installed over the robot in a birds-eye way. The parts and their background are static in the workspace during the human–robot collaboration process. The parts' types, positions, shapes, and colors are able to be acquired by the vision system. Then the robot can plan its actions to manipulate the corresponding part according to the collaborative scenario [28].

First, the human performs the "Need a part" intention to the robot in Fig. 11(1) and (2). Via the online data acquired from the natural wearable sensory system, the TLP model is utilized to predict the human intention. As shown in Fig. 11(3) and (4), after getting the prediction output, the robot asks the human by "What kind of part do you want." Afterward, the human presents his request of a medium part by the speech in Fig. 11(5). As shown in Fig. 11(6)–(8), the robot picks up the medium part from the container and delivers it to the human. After the human gets the part, the robot moves back to its initial position. Consequently, it can be revealed that, through the efficient and multimodal human forearm rotations and speech instructions as well as human–robot dialogue, the robot accurately predicts its human partner's intention based on the proposed easy-to-implement TLP model.

E. Evaluation and Discussion of Prediction Accuracy

1) Comparison of Different Approaches: We conduct a prediction accuracy test of these ten hand-over intentions by different approaches, including our approach, basic ELM [37], support vector machine (SVM) [40], least-squares support vector machine (LS-SVM) [41], linear discriminant analysis (LDA) [42], and k-nearest neighbor (KNN) [43]. All the algorithms are tested on the same operation station.

We ask another group of five subjects (two females and three males) within an age range of 21–35 to collaborate with the robot to conduct hand-over tasks using the proposed TLP approach according to his/her working preferences. 5000 sets (500 sets per intention) of hand-over intention features are collected and employed in the accuracy evaluation. Additionally, the k-fold cross-validation method [44] is utilized for these approaches to objectively verify and evaluate their generalization ability in the hand-over intention prediction problem. Based on the empirical evidence [45], we divide the acquired data into ten equal-sized subsets so that all observations can be used for both learning and prediction. One of the ten subsets is worked as the validation (prediction) set for testing the learned model, and the other nine subsets are employed together to form a learning set. Therefore, the cross-validation process is repeated ten times, with each subset used exactly once as the validation data. As shown in Table I, we average the ten prediction results of each hand-over intention as the accuracy estimation.

The average prediction accuracy (APA) results and the standard deviation of the prediction errors (StD-E) of all hand-over intentions are presented in Fig. 12. Based on the histogram, the APA results of these approaches are 99.7%, 98.94%, 94%,

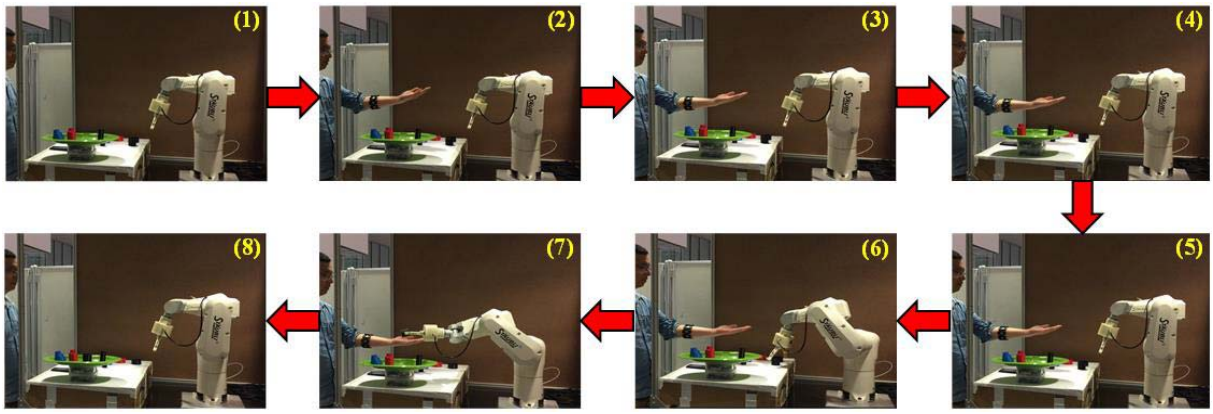


Fig. 11. Procedure of the robot-to-human hand-over intention prediction (deliver a medium part).

TABLE I
COMPARISONS OF EACH INTENTION'S APA BY DIFFERENT APPROACHES

Algorithm	Give the LHP	Give the LP	Give the HP	Give the MP	Give the SP	Stop	Continue	Speed Up	Slow Down	Need
Our Approach	98.00%	99.00%	100%	100%	100%	100%	100%	100%	100%	100%
Basic ELM	95.10%	97.00%	98.30%	100%	100%	99.00%	100%	100%	100%	100%
SVM	94.00%	100%	94.00%	100%	80.00%	100%	100%	100%	100%	100%
LS-SVM	91.00%	99.00%	100%	95.00%	100%	100%	100%	100%	100%	100%
LDA	36.00%	48.00%	59.00%	55.00%	86.00%	100%	100%	100%	98.00%	63.00%
KNN	90.00%	99.00%	95.00%	100%	100%	100%	100%	100%	100%	100%

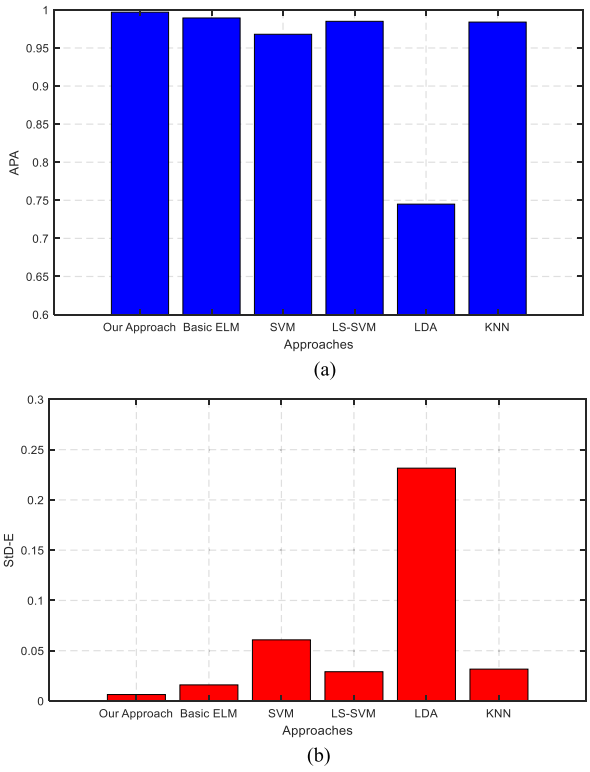


Fig. 12. Comparison results of different approaches in hand-over intentions prediction. (a) APA. (b) StD-E.

98.5%, 74.5%, and 98.4%, respectively. It can be seen that the proposed TLP model can successfully and sensitively predict all the hand-over intentions in the human–robot collaboration with a higher accuracy than other approaches. Furthermore, Fig. 12 shows that the StD-E generated by the TLP model is about 0.0064, which is much smaller than the StD-Es of basic ELM (0.01602), SVM (0.0608), LS-SVM (0.0291), LDA (0.2361), and KNN (0.0317). Therefore, it can be concluded that our approach is more stable in the hand-over intention prediction.

2) Comparison to Other Works: The performance of our method is compared to that of some previous studies, which investigated on different applications to human–robot interaction, including human intention understanding and prediction [11], [22], [23], [46], [47], human motion prediction [48]–[50], human gesture recognition [51]–[53], and human behavior detection and prediction [54], [55]. Kelley *et al.* [46] proposed the vision system-based human intention understanding approach using hidden Markov models (HMMs). Wang *et al.* [47] developed a probabilistic movement model to infer human intentions based on the vision system as well in [47]. For the studies [48]–[50], the authors predicted human motions in human–robot interaction based on the vision system, EMG sensors, and simulator, separately. Human gesture recognition in human–robot interaction was carried out by Georgi *et al.* [51] (based on IMU and EMG sensors), Chen *et al.* [52] (based on the vision system), and Gu *et al.* [53] (based on the vision system), respectively. Human behaviors are predicted by Pentland and Liu [54] and Ryoo *et al.* [55] through the simulator and vision system in [54] and [55], separately. The human–robot interface and employed algorithms of these studies are summarized in Table II. We also compare the performance of this approach to our previous studies [56], [57], which are implemented in

TABLE II
COMPARISONS OF OUR APPROACH TO PREVIOUS WORKS

Works	Human-Robot Interface	Algorithm	Average Accuracy
Our Approach	Natural language and Wearable sensing	Kernel ELM	99.7%
Georgi et al. [51]	IMU and EMG sensors	HMM	97.8%
Kelley et al. [46]	Vision system (PTZ Sony)	HMM	92.57%
Chen et al. [52]	Vision system	HMM	85%
Gu et al. [53]	Vision system (Kinect)	HMM	84.76%
Grigore et al. [11]	Vision system (VICON)	HMM	75.63%
Wang et al. [57]	IMU and EMG sensors	HMM and DAG-SVM	92.91%
Elfring et al. [48]	Vision system	GHMMs	90%
Ryoo et al. [55]	Vision system	SVM	72.9%
Song et al. [23]	Vision system (Kinect)	SOM and GMM	93.5%
Mainprice et al. [50]	Simulator	GMM	92%
Pentland et al. [54]	Simulator	MDMs	95.24%
Wang et al. [56]	Natural language	MaxEnt-IRL	95%
Bu et al. [49]	EMG sensors	BN	86.4%
Wang et al. [47]	Vision system (Prosilica GE640C)	IDDM	83.8%
Cohen et al. [22]	Simulator	MIPM	83%

the same platform and conditions as used in this work. In [56], a teaching-learning-collaboration (TLC) model was developed by us via the maximum entropy inverse reinforcement learning (MaxEnt-IRL) algorithm to have the robot learn and assist its human companion in shared tasks. Different from the TLC model, in which the robot learning process was constrained to be deterministic with specific input features of human intentions, the proposed TLP model in this study can have the robot learn from human demonstrations with unspecific input features (extracted from gesture information and not always absolutely identical even for the same human intention) based on the ELM algorithms. Wang *et al.* [57] utilized a wearable sensory system to control the human–robot object hand-over process based on HMM and directed acyclic graph-support vector machine (DAG-SVM) algorithms. Different from [57], in which the human behaviors were offline labeled manually, the proposed TLP model in this study can online annotate human intentions by natural language and natural gesture information for the robot to learn and predict human intentions.

As presented in Table II, even though some tasks and datasets are different in these studies, the average output accuracy (the success rate of the human–robot collaboration executions) is able to be employed as a statistical indicator to assess their performances. First, compared to the previous works [22], [50], [54], which were implemented in simulation scenarios with little interference factors, the proposed TLP approach is able to achieve a higher accuracy. Additionally, compared to the previous studies [11], [23], [46]–[49], [51]–[53], [55], it can be seen that our approach has a competitive average accuracy over them. Moreover, as shown in Table II, the performance comparisons also indicate that the TLP model is more accurate than our previous works [56], [57].

In this study, we utilize the natural language and integrated wearable sensing system as the human–robot interface that can provide a more effective interaction way and a less cost in contrast to vision-based systems, separate IMU sensors, and separate EMG sensors. In addition, the wearable sensory system incorporates all sensing capabilities into an armband, which is wireless and self-powered by an embedded battery. Humans can naturally wear it just like wearing a smartwatch without any complex setup. Human natural language can also be used as a direct and inherent manner to express and transfer human intention information. Therefore, our approach has a natural and simple configuration and can be easily used by humans.

F. Evaluation and Discussion of Learning and Prediction Time

Time cost is also an important factor to influence task quality in human–robot collaboration [58]. During the accuracy prediction process, we also collect the robot learning time and prediction time in each task by different approaches (our approach, basic ELM, SVM, LS-SVM, LDA, KNN). After that, the average time cost and total time cost of each approach in the robot learning process and prediction process are figured out, respectively. The cost time computation of all approaches is based on the CPU time on the same computer.

As shown in Table III, these approaches share different performances in time consumption. In the robot learning process, the time cost of our approach is 2.2902 s, which is less than that of basic ELM, SVM, and LS-SVM, but a little greater than that of LDA and KNN. However, the APA of LDA and KNN is lower than that of our approach according to Table I. In addition, for human intention prediction, our approach performs a competitive average prediction time (0.1594 s) over most of the methods in Table III such as SVM, LS-SVM, LDA, and KNN, whereas it takes a little longer than basic ELM. But the total time cost of our approach is 2.4496 s, which is the least in the total time consumption of all the approaches. Consequently, it can be concluded that the proposed TLP approach, which is based on kernel ELM algorithms has a higher efficiency in human–robot hand-over tasks than other methods.

To sum up, the experimental results and analysis demonstrate that by taking advantage of the proposed TLP approach, the robot can correctly predict hand-over intentions online.

TABLE III
TIME COST COMPARISONS BY DIFFERENT APPROACHES

Algorithms	Average Learning Time (s)	Average Prediction Time (s)	Total Time (s)
Our Approach	2.2902	0.1594	2.4496
Basic ELM	2.7656	0.1208	2.8864
SVM	6.5781	0.1812	6.7593
LS-SVM	7.2960	0.3188	7.6148
LDA	2.1563	0.8173	2.9736
KNN	2.0781	0.6913	2.7694

Furthermore, the evaluation results suggest that the proposed method shows robust efficiency and accuracy in human–robot hand-over tasks.

VIII. CONCLUSION

An effective and easy-to-use TLP framework has been developed for the robot to online learn from human hand-over demonstrations and then predict human hand-over intentions in human–robot collaborative tasks. Different from existing approaches, we have employed natural multimodal information (natural language and natural wearable sensing information) to online program the robot once the collaborative task is updated. The robot can build its cognition capacity of understanding human hand-over actions and predicting hand-over intentions online by learning from human demonstrations. Moreover, the robot is able to make active decisions to plan its actions to deliver, pick up or reject the parts using its learned strategies in human–robot hand-over processes. Several practical experiments have been conducted on a collaborative robot research platform for verifying and validating the performance of the TLP model. Experimental results and evaluations demonstrate that the proposed approach has a distinct advantage over traditional solutions to program robots, and the robot can predict human hand-over intentions effectively with high accuracy and robustness in human–robot collaboration. Although the performance of our approach is more competitive over that of other methods, we need to future verify it by online human–robot collaboration. It is also our future work based on this study that we will develop an ensemble framework, which integrates different kinds of approaches, for the human to teach robot online, then having the robot score the learning approaches and select the best one to predict human intentions in human–robot collaborative tasks. In addition, we will make a comprehensive comparison of computation efficiency for different approaches via other metrics such as computation complexity.

REFERENCES

[1] M. Ficocelli, J. Terao, and G. Nejat, "Promoting interactions between humans and robots using robotic emotional behavior," *IEEE Trans. Cybern.*, vol. 46, no. 12, pp. 2911–2923, Dec. 2016.

[2] D. Bourne *et al.*, "Mobile manufacturing of large structures," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2015, pp. 1565–1572.

[3] W. Wang, R. Li, Z. M. Diekel, and Y. Jia, "Robot action planning by online optimization in human–robot collaborative tasks," *Int. J. Intell. Robot. Appl.*, vol. 2, no. 2, pp. 161–179, 2018.

[4] H. Bley, G. Reinhart, G. Seliger, M. Bernardi, and T. Korne, "Appropriate human involvement in assembly and disassembly," *CIRP Ann.*, vol. 53, no. 2, pp. 487–509, 2004.

[5] J. Léger and J. Angeles, "Off-line programming of six-axis robots for optimum five-dimensional tasks," *Mechanism Mach. Theory*, vol. 100, pp. 155–169, Jun. 2016.

[6] W. Wang, Y. Chen, R. Li, Z. Zhang, V. Krovi, and Y. Jia, "Human-robot collaboration for advanced manufacturing by learning from multi-modal human demonstrations," in *Recent Advances in Industrial Robotics*. Singapore: World Scientific, 2020, pp. 87–116.

[7] X. Yu, W. He, H. Li, and J. Sun, "Adaptive fuzzy full-state and output-feedback control for uncertain robots with output constraint," *IEEE Trans. Syst., Man, Cybern. Syst.*, early access, Feb. 3, 2020, doi: 10.1109/TSMC.2019.2963072.

[8] W. He, C. Xue, X. Yu, Z. Li, and C. Yang, "Admittance-based controller design for physical human–robot interaction in the constrained task space," *IEEE Trans. Automat. Sci. Eng.*, early access, Apr. 21, 2020, doi: 10.1109/TASE.2020.2983225.

[9] W. Wang, R. Li, Y. Chen, and Y. Jia, "Human intention prediction in human-robot collaborative tasks," in *Proc. Companion ACM/IEEE Int. Conf. Human-Robot Interact.*, Mar. 2018, pp. 279–280.

[10] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, nos. 1–3, pp. 489–501, Dec. 2006.

[11] E. C. Grigore, K. Eder, A. G. Pipe, C. Melhuish, and U. Leonards, "Joint action understanding improves robot-to-human object handover," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Nov. 2013, pp. 4622–4629.

[12] J. Aleotti, V. Micelli, and S. Caselli, "Comfortable robot to human object hand-over," in *Proc. IEEE RO-MAN: 21st IEEE Int. Symp. Robot Human Interact. Commun.*, Sep. 2012, pp. 771–776.

[13] M. Cakmak, S. S. Srinivasa, M. K. Lee, J. Forlizzi, and S. Kiesler, "Human preferences for robot-human hand-over configurations," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sep. 2011, pp. 1986–1993.

[14] K. Nagata, Y. Oosaki, M. Kakikura, and H. Tsukune, "Delivery by hand between human and robot based on fingertip force-torque information," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Systems. Innov. Theory, Pract. Appl.*, Oct. 1998, pp. 750–757.

[15] J. Chen and A. Zelinsky, "Programing by demonstration: Coping with suboptimal teaching actions," *Int. J. Robot. Res.*, vol. 22, no. 5, pp. 299–319, May 2003.

[16] A. G. Billard, S. Calinon, and F. Guenter, "Discriminative and adaptive imitation in uni-manual and bi-manual tasks," *Robot. Auto. Syst.*, vol. 54, no. 5, pp. 370–384, May 2006.

[17] J. Aleotti and S. Caselli, "Learning manipulation tasks from human demonstration and 3D shape segmentation," *Adv. Robot.*, vol. 26, no. 16, pp. 1863–1884, Nov. 2012.

[18] M. Ferreira, P. Costa, L. Rocha, and A. P. Moreira, "Stereo-based real-time 6-DoF work tool tracking for robot programing by demonstration," *Int. J. Adv. Manuf. Technol.*, vol. 85, pp. 57–69, Jun. 2014.

[19] Y. Jia, N. Xi, J. Y. Chai, Y. Cheng, R. Fang, and L. She, "Perceptive feedback for natural language control of robotic operations," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2014, pp. 6673–6678.

[20] A. Yahya, A. Li, M. Kalakrishnan, Y. Chebotar, and S. Levine, "Collective robot reinforcement learning with distributed asynchronous guided policy search," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2017, pp. 79–86.

[21] G. Konidaris, S. Kuindersma, R. Grupen, and A. Barto, "Robot learning from demonstration by constructing skill trees," *Int. J. Robot. Res.*, vol. 31, no. 3, pp. 360–375, Mar. 2012.

[22] Y. Tanaka, J. Kinugawa, Y. Sugahara, and K. Kosuge, "Motion planning with worker's trajectory prediction for assembly task partner robot," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2012, pp. 1525–1532.

[23] L. Cohen, S. Haliyo, M. Chetouani, and S. Regnier, "Intention prediction approach to interact naturally with the microworld," in *Proc. IEEE/ASME Int. Conf. Adv. Intell. Mechatronics*, Jul. 2014, pp. 396–401.

[24] D. Song *et al.*, "Predicting human intention in visual observations of hand/object interactions," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2013, pp. 1608–1615.

[25] K. P. Hawkins, N. Vo, S. Bansal, and A. F. Bobick, "Probabilistic human action prediction and wait-sensitive planning for responsive human-robot collaboration," in *Proc. 13th IEEE-RAS Int. Conf. Humanoid Robots (Humanoids)*, Oct. 2013, pp. 499–506.

[26] M. Awais and D. Henrich, "Human-robot interaction in an unknown human intention scenario," in *Proc. 11th Int. Conf. Frontiers Inf. Technol.*, Dec. 2013, pp. 89–94.

[27] Google Cloud Platform. (2018). *Speech API-Speech Recognition*. Accessed: Mar. 9, 2018. [Online]. Available: https://cloud.google.com/speech/

[28] Y. Jia, L. She, Y. Cheng, J. Bao, J. Y. Chai, and N. Xi, "Program robots manufacturing tasks by natural language instructions," in *Proc. IEEE Int. Conf. Autom. Sci. Eng. (CASE)*, Aug. 2016, pp. 633–638.

[29] L. She, S. Yang, Y. Cheng, Y. Jia, J. Chai, and N. Xi, "Back to the blocks world: Learning new actions through situated human-robot dialogue," in *Proc. 15th Annu. Meeting Special Interest Group Discourse Dialogue (SIGDIAL)*, 2014, pp. 89–97.

[30] *Myo*. Accessed: May 1, 2018. [Online]. Available: https://www.myo.com/

[31] C. J. De Luca, M. Kuznetsov, L. D. Gilmore, and S. H. Roy, "Inter-electrode spacing of surface EMG sensors: Reduction of crosstalk contamination during voluntary contractions," *J. Biomech.*, vol. 45, no. 3, pp. 555–561, Feb. 2012.

[32] H. Ahmed and M. Tahir, "Improving the accuracy of human body orientation estimation with wearable IMU sensors," *IEEE Trans. Instrum. Meas.*, vol. 66, no. 3, pp. 535–542, Mar. 2017.

[33] W. Wang, R. Li, Z. M. Diekel, and Y. Jia, "Hands-free maneuvers of robotic vehicles via human intentions understanding using wearable sensing," *J. Robot.*, vol. 2018, Apr. 2018, Art. no. 4546094, doi: 10.1155/2018/4546.094.

[34] G.-B. Huang and L. Chen, "Enhanced random search based incremental extreme learning machine," *Neurocomputing*, vol. 71, nos. 16–18, pp. 3460–3468, Oct. 2008.

[35] G.-B. Huang and L. Chen, "Convex incremental extreme learning machine," *Neurocomputing*, vol. 70, nos. 16–18, pp. 3056–3062, Oct. 2007.

[36] G.-B. Huang, L. Chen, and C.-K. Siew, "Universal approximation using incremental constructive feedforward networks with random hidden nodes," *IEEE Trans. Neural Netw.*, vol. 17, no. 4, pp. 879–892, Jul. 2006.

[37] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Trans. Syst., Man, Cybern., B (Cybern.)*, vol. 42, no. 2, pp. 513–529, Apr. 2012.

[38] G. Huang, G.-B. Huang, S. Song, and K. You, "Trends in extreme learning machines: A review," *Neural Netw.*, vol. 61, pp. 32–48, Jan. 2015.

[39] J. Tang, C. Deng, and G.-B. Huang, "Extreme learning machine for multilayer perceptron," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 4, pp. 809–821, Apr. 2016.

[40] M. A. Hearst, S. T. Dumais, E. Osman, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intell. Syst. Appl.*, vol. 13, no. 4, pp. 18–28, Jul./Aug. 2008.

[41] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Process. Lett.*, vol. 9, no. 3, pp. 293–300, Jun. 1999.

[42] E. I. Altman, G. Marco, and F. Varetto, "Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience)," *J. Banking Finance*, vol. 18, no. 3, pp. 505–529, May 1994.

[43] T. Denoeux, "A k-nearest neighbor classification rule based on Dempster-Shafer theory," *IEEE Trans. Syst., Man, Cybern.*, vol. 25, no. 5, pp. 804–813, May 1995.

[44] T. Fushiki, "Estimation of prediction error by using K-fold cross-validation," *Statist. Comput.*, vol. 21, no. 2, pp. 137–146, Apr. 2011.

[45] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 1995, pp. 1137–1145.

[46] R. Kelley, A. Tavakkoli, C. King, M. Nicolescu, M. Nicolescu, and G. Bebis, "Understanding human intentions via hidden Markov models in autonomous mobile robots," in *Proc. 3rd Int. Conf. Human robot Interact. (HRI)*, 2008, pp. 367–374.

[47] Z. Wang *et al.*, "Probabilistic movement modeling for intention inference in human-robot interaction," *Int. J. Robot. Res.*, vol. 32, no. 7, pp. 841–858, 2013.

[48] J. Elfring, R. van de Molengraft, and M. Steinbuch, "Learning intentions for improved human motion prediction," *Robot. Auto. Syst.*, vol. 62, no. 4, pp. 591–602, Apr. 2014.

[49] N. Bu, M. Okamoto, and T. Tsuji, "A hybrid motion classification approach for EMG-based human–robot interfaces using Bayesian and neural networks," *IEEE Trans. Robot.*, vol. 25, no. 3, pp. 502–511, Jun. 2009.

[50] J. Mainprice and D. Berenson, "Human-robot collaborative manipulation planning using early prediction of human motion," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Nov. 2013, pp. 299–306.

[51] M. Georgi, C. Amma, and T. Schultz, "Recognizing hand and finger gestures with IMU based motion and EMG based muscle activity sensing," in *Proc. Int. Conf. Bio-Inspired Syst. Signal Process.*, 2015, pp. 99–108.

[52] F.-S. Chen, C.-M. Fu, and C.-L. Huang, "Hand gesture recognition using a real-time tracking method and hidden Markov models," *Image Vis. Comput.*, vol. 21, no. 8, pp. 745–758, Aug. 2003.

[53] Y. Gu, H. Do, Y. Ou, and W. Sheng, "Human gesture recognition through a kinect sensor," in *Proc. IEEE Int. Conf. Robot. Biomimetics (ROBIO)*, Dec. 2012, pp. 1379–1384.

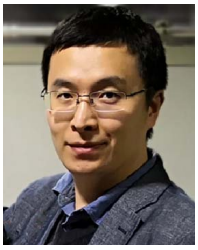
[54] A. Pentland and A. Liu, "Modeling and prediction of human behavior," *Neural Comput.*, vol. 11, no. 1, pp. 229–242, Jan. 1999.

[55] M. Ryoo, T. J. Fuchs, L. Xia, J. K. Aggarwal, and L. Matthies, "Robot-centric activity prediction from first-person videos: What will they do to me?" in *Proc. 10th ACM/IEEE Int. Conf. Hum.-Robot Interact. (HRI)*, Mar. 2015, pp. 295–302.

[56] W. Wang, R. Li, Y. Chen, Z. Diekel, and Y. Jia, "Facilitating human–robot collaborative tasks by teaching-learning-collaboration from human demonstrations," *IEEE Trans. Automat. Sci. Eng.*, vol. 16, no. 2, pp. 640–653, Apr. 2019.

[57] W. Wang *et al.*, "Controlling object hand-over in human–robot collaboration via natural wearable sensing," *IEEE Trans. Human-Mach. Syst.*, vol. 41, no. 9, pp. 59–71, Feb. 2019.

[58] D. R. Olsen and M. A. Goodrich, "Metrics for evaluating human-robot interactions," in *Proc. Perform. Metrics Intell. Syst. (PerMIS)*, 2003, pp. 1–8.



Weitian Wang (Member, IEEE) received the Ph.D. degree from the University of Science and Technology of China, Hefei, China, in 2016.

He worked as a Post-Doctoral Fellow with the Collaborative Robotics and Automation Laboratory, Clemson University, Greenville, SC, USA, from 2016 to 2019. He is currently an Assistant Professor with the Department of Computer Science, Montclair State University, Montclair, NJ, USA. His research interests include collaborative robotics, smart systems, human–robot interaction, machine learning, and sensing technology.

Rui Li received the Ph.D. degree from the University of Science and Technology of China, Hefei, China, in 2016.

She was a Research Associate with Clemson University ICAR, Greenville, SC, USA. She is currently working in the Department of Computer Science, Montclair State University, Montclair, NJ, USA. Her research interest covers virtual reality, machine learning, 3-D computer animation, 3-D visualization, robotics, and human–machine interaction.

Yi Chen received the M.S. degree in mechanical engineering from the University of Florida, Gainesville, FL, USA, in 2016. He is currently pursuing the Ph.D. degree in automotive engineering with Clemson University, Greenville, SC, USA.

His research interests include collaborative robotics, robot motion planning, and human–robot interaction in manufacturing.

Yi Sun received the master's degree from New York University, New York, NY, USA, in 2018.

He was an Assistant Researcher with the Collaborative Robotics and Automation Laboratory, Clemson University, Greenville, SC, USA. His research interests include machine learning, robotics, human–robot interaction, and computer vision.

Yunyi Jia received the B.S. degree from the National University of Defense Technology, Changsha, China, in 2005, the M.S. degree from the South China University of Technology, Guangzhou, China, in 2008, and the Ph.D. degree from Michigan State University, East Lansing, MI, USA, in 2014.

He is currently the Director of the Collaborative Robotics and Automation (CRA) Laboratory and the McQueen Quattlebaum Assistant Professor with the Department of Automotive Engineering, Clemson University, Greenville, SC, USA. His research interests include collaborative robotics, autonomous vehicles, and advanced sensing systems.