FedHIP: Federated Learning for Privacy-Preserving Human Intention Prediction in Human-Robot Collaborative Assembly Tasks

Jiannan Cai, Ph.D.[1*], Zhidong Gao[2], Yuanxiong Guo[3], Bastian Wibranek[4], and Shuai Li[5]

[1*]School of Civil & Environmental Engineering, and Construction Management, The University of Texas at San Antonio (corresponding author). Email: jiannan.cai@utsa.edu

[2]Department of Electrical and Computer Engineering, The University of Texas at San Antonio. Email: zhidong.gao@my.utsa.edu

[3]Department of Information Systems and Cyber Security, The University of Texas at San Antonio. Email: yuanxiong.guo@utsa.edu

[4]Faculty of Architecture, Darmstadt University of Applied Sciences. Email: Bastian@wibranek.de

[5]Department of Civil and Environmental Engineering, The University of Tennessee, Knoxville. Email: sli48@utk.edu

Abstract

Human-robot collaboration is a promising solution to relieve construction workers from repetitive and physically demanding tasks, thus improving construction safety and productivity. Many studies have developed various deep learning models for human intention prediction, which will form the basis for proactive and adaptive robot planning and control to enable intelligent human-robot collaboration. However, there remain two challenges. First, most research only focuses on a single type of human intention, without a holistic understanding of multi-level intention, including both high-level intended actions and objects of interest, and low-level body movements. Second, conventional deep learning approaches train a centralized model with aggregated datasets, which requires the sharing of sensitive information (e.g., personal images and behavior data), posing broad privacy concerns in practical implementation. This study proposes a

vision-based multi-task federated learning (FL) framework, FedHIP, for multi-level human intention prediction in human-robot collaborative assembly tasks. Specifically, taking body movements and assembly components as inputs, a long short-term memory based multi-task learning model was developed to simultaneously predict multi-level human intention in assembly tasks. FL was employed to train the model in a distributed and privacy-preserving way on local clients without the need of transmitting sensitive data. The results show that the proposal FedHIP without and with pre-train can achieve an accuracy of 80.1% and 85.7% in action prediction, 97.6% and 97.8% in object prediction, and an average displacement error of 12.7 and 11.7 pixels in motion prediction, respectively. Models trained from FedHIP were also compared with those obtained from traditional centralized training and local training. It was found that FL leads to compatible accuracy with centralized training and much higher accuracy than local training while preserving data privacy.

1. Introduction

Recent advancement in robotics, artificial intelligence (AI), and sensing technologies have made it possible for robots to serve as intelligent assistants in various applications, such as smart home [1], healthcare [2], manufacturing [3], etc. In construction, human-robot collaboration (HRC) has emerged as a promising solution to relieve construction workers from repetitive and physically demanding tasks, thus improving construction safety and productivity as well as alleviating workforce crisis such as labor shortage and aging [4–7]. However, the unstructured and dynamic workspaces and the diverse and complex construction activities make it extremely difficult to apply industrial robots that are traditionally pre-programmed to conduct a single task in a fixed working environment [8]. Without eliminating construction workers, collaborative robots in construction must be designed to cognitively team up with workers, in order to assist humans in

repetitive and physically demanding tasks while leveraging human expertise to handle unexpected and complex situations. Specifically, robots should be empowered with the intelligence to perceive, understand, and adapt to human intention for proactive planning in the dynamic environment.

With a focus to facilitate intelligent HRC, different methods have been developed to predict human intention from various aspects, ranging from potential actions [9] and objects of interest [10], to the motion of human full body [11] or specific body parts [12]. Most existing studies leverage deep learning models that predict human intention from heterogenous inputs, such as imagery data [9], speech [13], EMG signals [14], eye-gaze movements [10], etc.

Despite the achievements of human intention prediction, there remain two knowledge gaps. First, most existing studies only focus on predicting a single type of human intention, i.e., either high-level intention such as objects of interests and potential actions or low-level body movement. There lacks a holistic understanding of multi-level intention (both high-level intended actions and objects of interest, and low-level body movement), which is critical to developing an intelligent robot that can figure out "when to help", "what to help", and "how to help" simultaneously and adapt to various human behavior for smooth collaboration. Second, existing methods rely on a large amount of data that is aggregated to a single dataset, to train the deep learning model in a centralized way. In the context of HRC in construction tasks, training data with different workers on various tasks across different projects is needed to ensure the generalizability of the model. The sharing of human behavior and imagery data poses significant concerns in data privacy and security, which is more critical in private and fragmented industries like construction [15]. Therefore, there is a critical need to develop a new learning-based model that can simultaneously predict multi-level human intention in a privacy-preserving way.

On the other hand, federated learning (FL) is an emerging machine learning mechanism that is trained across multiple distributed clients with local datasets and then aggregated on a centralized server. FL can cooperatively implement machine learning tasks without raw data transmissions. Therefore, it becomes the promising solution to resolve the conflicts between user privacy and data sharing. Some recent studies have adopted FL in human activity recognition [16–18] and shown promising results in achieving good recognition accuracy while eliminating the need of data sharing. However, in existing studies, FL was applied in simple classification tasks, and it remains unknown if FL could be used and how FL may perform in both classification and regression tasks in a multi-task setting to reliably predict multi-level human intention.

To this end, this study aims to develop a vision-based multi-task federated learning (FL) framework, FedHIP, for multi-level human intention prediction in human-robot collaborative assembly tasks, where multi-level human intention includes high-level intended actions and desired objects, and low-level body movements. Specifically, leveraging body movements and assembly components extracted from videos, a long short-term memory (LSTM) based MTL model was developed to simultaneously predict multi-level human intention in assembly tasks. FL was employed to train the model in a distributed and privacy-preserving way on local clients without the need of transmitting sensitive data. Furthermore, the feasibility and performance of FL in a multi-task setting with both classification and regression tasks to reliably predict multi-level human intention was analyzed and discussed.

The contribution of this study is threefold: 1) an LSTM-based MTL model is developed to predict multi-level human intention in assembly tasks including high-level intended actions and desired objects and low-level body movement, leveraging the commonality in human behavior, where involved assembly components were integrated as task contextual information to augment

the prediction capability of the model. Even though LSTM model is used as the backbone considering it being classic approach in time-series prediction with wide use in human intention prediction, our model is flexible and other state-of-the-art models, such as transformer-based models [19] and spatial-temporal graph convolutional network (ST-GCN) based models [20] could also be applied. 2) Leveraging FL, the proposed FedHIP framework allows the protection of sensitive human behavior data by training models with local data and update global model by aggregating locally trained models, without sharing and transmitting data to the central server. The proposed framework demonstrates the efficacy of FL in MTL model that consists of both classification and regression tasks, and is a pioneer study in construction research. A series of experiments were conducted to compare the performance of proposed framework with traditional centralized training and local training, and the results validated the efficacy and benefits of our approach. 3) The proposed framework could be extended to other applications in manufacturing, healthcare, etc. It enables robots to answer three important questions in HRC, i.e., "what is needed" (from object prediction), "when is needed" (from action prediction), and "where to assist" (from body movement prediction). Such knowledge enhances the intelligence of robot assistants for proactive and adaptive robot planning and control to improve safe and efficient HRC.

2. Related Studies

2.1. Applications of Human-Robot Collaboration

HRC research explores the physical and cognitive interaction between human and robot to complete a common task, investigating the feasibility to use robotic platforms (e.g., collaborative robot) to work in close proximity with human partner to achieve various joint tasks safely and efficiently [21]. Some studies focus on the cognitive HRC, where robots need to understand the current work states and human intention to determine its own task and movement, and the typical

tasks are collaborative assembly [22,23] and object handover [24,25]. For instance, in [22], robots need to understand the current step of pipe assembly and pick the corresponding components to assemble. In [23], the robot provides assistance based on the prediction of human trajectory. For object handover, it is critical for robots to understand human intention, e.g., the object desired, and pick the correct one to hand over to human partner [25]. Another type of studies focuses on the physical HRC, where robots may take more active roles and apply forces to the workpieces [21]. As a result, research emphasis has been placed to develop motion planning and control methods for robot manipulators based on the information extracted from the collaboration, with typical tasks related to collaborative manufacturing [26] and object handling [27].

In construction sector, HRC has attracted increasing attention as a promising solution to relieve human workers from repetitive and physically demanding tasks and improve construction safety and productivity, with potential applications in various construction tasks, such as bricklaying [28], object handover [29], wood assembly [30], etc. Many review studies have been conducted to discuss the state-of-the-art approaches as well as challenges and future directions of HRC in construction. For instance, [7] summarizes the evolution of HRC and identified robot learning and human-multirobot collaboration as potential future directions for research. [8] reviews state-of-the-art methods for construction robots to learn workplace skills which could be categorized as two components, i.e., activity understanding for task planning (e.g., [30]), and motion learning from demonstration via different machine learning algorithms, such as reinforcement learning [31] and imitation learning [32]. [33] proposes a multidimensional taxonomy to characterize HRC in construction field based on the team, task, and environment characteristics of specific studies. [34] conducted a systematic review to discuss HRC for on-site

construction and identified three critical areas of research, including adaptive robot programming, human-robot interface, and safety issues in HRC.

In HRC, it is well acknowledged that the capability to understand ongoing work progress and interpret human intention is critical for robots to adaptively collaborate and execute desired tasks. Machine learning models (e.g., HMM, LSTM, CNN) are commonly used to provide cognitive capabilities for robots to understand human behavior and correspondingly plan their motion. Specifically, high-level human intention (e.g., objects of interest and goals), human pose, or movement of specific body part (e.g., hand) are widely studied via a range of sensing inputs, such as vision, accelerometry, muscular activity and brain activity (e.g., [13,35–37]). In construction, methods have been developed to recognize multidimensional human states from different sensors (e.g., haptic sensors [29], cameras [38], EEG [28]) to guide robot response and task execution. This line of research is most relevant to the present study and is discussed in more detail in the next section.

2.2. Human Intention Prediction in Human-Robot Collaboration

Human intention prediction has become an active research area in HRC, with applications in manufacturing, health care, smart home, etc. Many studies focused on predicting high-level human intention such as objects of interest and potential types of activities/actions. For instance, in applications of manufacturing operation, [35] developed a Hidden Markov model to predict pre-defined groups of human actions. [37] created a convolutional neural network (CNN)-LSTM framework to predict potential actions and tools of interest in assembly tasks. [36] developed a brain-computer interface that detect the focus of human's overt attention via EEG signal to control robot motion for safe and efficient HRC. [9] devised a multimodal transfer learning-based model to classify anticipated human action from scene images and human skeleton positions in human-

robot collaborative assembly. [20] employed a spatial temporal graph convolutional network (ST-GCN) and a FIFO queue-based model to predict assembly procedures from human skeletons obtained via camera, which will enable robots to understand and generate their tasks correspondingly. [39] uses ST-GCN and YOLOX models to predict assembly actions and assembly objects, respectively from video data.

For general human-robot handover tasks, [13] developed a framework to predict human intention as a group of commands (e.g., stop, continue, slow down, etc.), from natural language and EMG and IMU sensors using extreme learning machine. To advance the performance of patient assistive devices, [10] proposes a framework that uses spatial-temporal patterns of gaze movement with deep learning models to predict human's objects of interest in daily life.

In construction, [29] proposes a human-adaptative framework for object handover in construction, where grip states are estimated using semi-supervised methods from haptic sensors and used for robot to determine when to release the objects. [38] deployed a lightweight CNN network for hand gesture recognition using thermal images for robot control in construction. [28] developed methods to assess worker workload based on EEG signals and adjusted robot movements based on worker's status.

Other studies focused on human motion prediction, ranging from full-body movement to the motion of specific body parts. For instance, [11] created a deep learning model that leverages short-term human dynamics and object affordances in the environment to predict full-body movement in grasping and placing movements in daily life. [14] proposes a new feature extraction network to predict human motion intention from sEMG signals for exoskeleton control system. [12] developed a probabilistic dynamic movement primitive model to predict human hand motion in tabletop manufacturing task. In construction sector, [40,41] proposed a deep learning framework

to cluster gaze-hand relationship of different people and predict their hand motion from gaze trajectories in pipe skid maintenance task and safe construction robot teleoperation. In addition, [42] reviewed existing studies and associated methods on human motion prediction with a focus on human trajectory prediction and body movement prediction.

Despite the great achievements of existing studies on human intention prediction from diverse human behavior data (e.g., images, EMG signals, body motion data, etc.), most of them focus on predicting a single type of human intention, i.e., either high-level intention such as objects of interests and potential actions or low-level body movement. There lacks studies to holistically infer multi-level intention, which is critical to developing an intelligent robot that can figure out "when to help", "what to help", and "how to help" simultaneously and adapt to various human behavior for smooth collaboration. Furthermore, existing studies require collecting behavior data from different people and form a global dataset to train deep learning models in a centralized way. From practical implementation perspective, to ensure that the trained modal achieves good prediction accuracy and generalizability in desired tasks, data should be collected from realistic scenarios and reflect behavior pattern of heterogeneous people. However, the sharing of human behavior data from different workspaces/parties to form a centralized dataset could pose significant concerns in data privacy and security, which is more critical in private and fragmented industries like construction [15]. Therefore, there is a critical need to develop a new learning-based model that can simultaneously predict multi-level human intention in a privacy-preserving way.

2.3. Federated Learning

Federated learning (FL) [43] is a novel distributed machine learning paradigm, originally proposed by Google. In an FL system, a machine learning model is trained across multiple distributed clients with local datasets and then aggregated on a centralized server. FL is able to

cooperatively implement machine learning tasks without raw data transmissions, thereby promoting clients' data privacy [44]. Besides, transmitting the model update saves communication bandwidth than transmitting the raw data, as model update usually are smaller than raw data. FL has been applied to various data-sensitive scenarios, such as smart health care [45], E-commerce [46], edge computing [47,48], etc.

Recently, FL has been applied in human activity recognition. For instance, [16] employed FL to deep neural network-based human activity classifier and found that FL could produce models with slightly worse, but acceptable, accuracy compared to centralized models. [17] designed a new feature extractor network for each user in FL, which achieves better activity recognition results compared to existing FL systems. [18] focused on developing new FL algorithms with different model updating and learning mechanisms to improve model performance. On the other hand, FL has yet to be explored in construction domain. Only one study [15], to the best of authors' knowledge, applied FL in worker images for operators' facial fatigue recognition. Furthermore, all of the above studies employ FL in simple classification tasks. It is unknown how FL will perform in both classification and regression tasks in a MTL setting. Therefore, this study aims to develop a MTL model for human intention prediction, and exploit the feasibility and performance of FL in both classification and regression tasks.

3. Methodology

The proposed framework consists of three modules, as shown in Figure 1. First, given videos, human body movements were tracked as time-series skeleton positions using a deep learning-based pose tracking model. Second, an LSTM-based MTL model was created to predict multi-level human intention, including anticipated actions, objects of interest, and human body movement, by integrating observed body movements and involved assembly objects as contextual

information. The proposed LSTM-MTL model is based on our previous study [49]. Finally, FL was employed to train the above model on local data of individual participants and update the global model without the need of transmitting sensitive data.

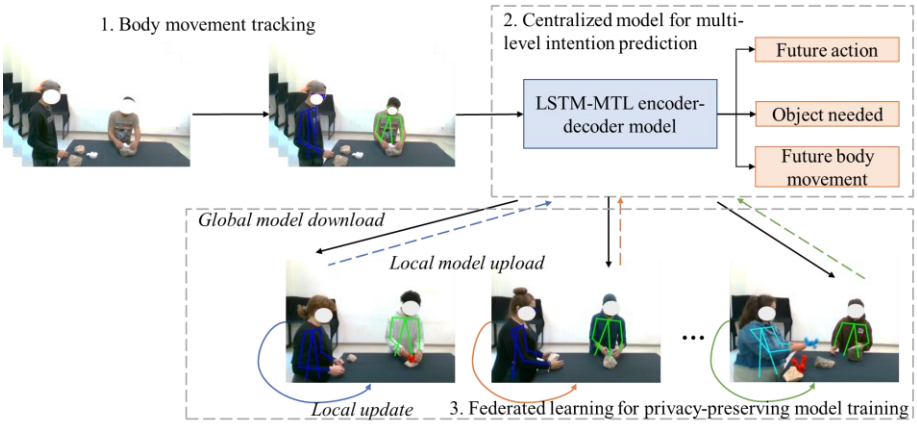


Figure 1. Overall workflow

3.1.Skeleton-based Pose Tracking

In this study, low-level body movement is represented as time-series locations of human skeleton key points, such as head, neck, shoulders, elbows, ankles, etc. A deep learning-based pose tracking algorithm, Pose Flow, developed by [50], was adopted in this study to track key point movements because of its computational efficiency and good performance in multi-person scenarios. First, skeleton pose in each image frame was extracted using multi-person pose estimation developed by Fang et al. [51]. Second, poses of the same person cross different frames were associated based on the similarity between different poses, forming multiple pose flows. Finally, non-maximum suppression mechanism was used to remove redundant pose flows and re-link temporal disjoint ones based on the confidence score of each pose flows and the distance between multiple flows. More details could be found in Xiu et al. [50]. It should be noted that the original Pose Flow algorithm tracks full-body movement with 17 key points, whereas only 13 key points of upper body (including hips), which were mainly involved in the assembly tasks, were considered in this study as lower body was occluded by the workbench.

3.2.LSTM-MTL Model for Multi-Level Intention Prediction

To enable prediction of multi-level human intention, including high-level intentions of intended action and object of interest, and low-level intention of body movement, MTL mechanism was adopted considering its ability to capture both commonality and difference among different aspects of intention for efficient learning. It could also handle the situation where specific classes have limited training data [52,53]. Specifically, three tasks were formulated for the above three intentions, respectively, where Task 1 action prediction and Task 2 object prediction are multi-class classification problems, and Task 3 movement prediction is a regression problem. Furthermore, an LSTM encoder-decoder model was used as a backbone for multi-level human intention prediction, considering it being a classic approach for time-series prediction with wide adoption and good performance in human intention prediction [37,54]. As a result, an LSTM based MTL model with encoder-decoder architecture was created for multi-level human intention prediction, as illustrated in Figure 2. In recent studies, other modern models, such as transformer-based models [19] and ST-GCN based models [20] were used to recognize and predict human body movements and other assembly intentions, and could also be potentially used as the backbone in the proposed model.

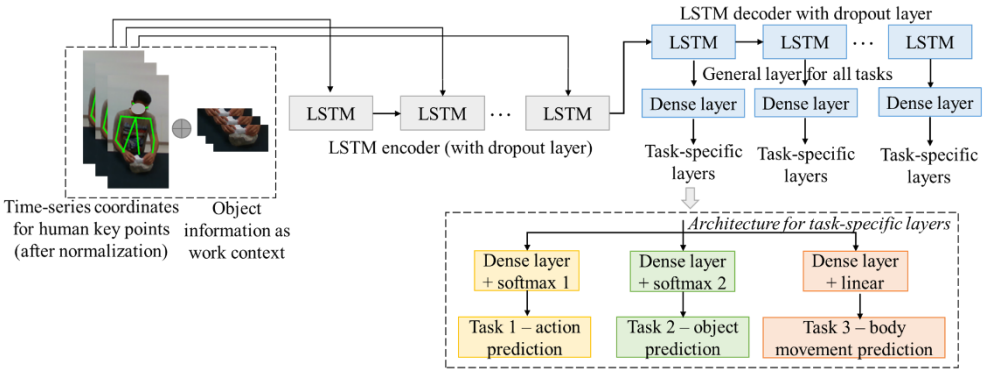


Figure 2. Proposed LSTM-MTL model for multi-level human intention prediction

In the proposed model, the inputs of LSTM encoder consider the observed human movement and corresponding objects of interest over a past period. Specifically, at time step *t*, the input is a 27-dimentional feature vector, denoted as $\boldsymbol{X}_t = [x_t^1, y_t^1, x_t^2, y_t^2, \dots, x_t^{13}, y_t^{13}, o_t]$, where x_t^i and y_t^i ($i \in [1,13]$) are the normalized *x* and *y* coordinates of 13 human key points obtained from pose tracking in Section 3.1. o_t is the type of object associated with current time step and was incorporated as task contextual information, represented as a categorical variable. The time-series features are then constructed by chaining a series of time-variant feature vectors over a time period, denoted by $\{\boldsymbol{X}_t, \boldsymbol{X}_{t+\Delta t}, \boldsymbol{X}_{t+\Delta 2t}, \dots, \boldsymbol{X}_{t+T}\}$, where *t* is the starting time, Δt is the temporal resolution that determined by data frequency, and *T* is the observation duration. Consequently, the LSTM network acts as an encoder that takes above time-series features as inputs to capture the temporal dependencies of human movements and work context. The architecture and principle of the LSTM network used in this study could be found in [55], and a dropout layer was applied in the states of LSTM encoder to mitigate overfitting issue.

The encoder outputs an encoded vector that is the hidden state of the last LSTM cell. The encoded vector encapsulates the information from observed movement and corresponding work context. In the proposed MTL mechanism, the encoded vector was shared among different tasks and captured the common representation, which was then fed into a second LSTM decoder followed by a dense layer. Then, task-specific dense layers were used to learn the uniqueness of individual tasks, leading to separate outputs for each task. Benefit from the encoder-decoder architecture, the model can generate predicted human intention over multiple time steps, i.e., one prediction for each LSTM decoder cell, denoted as $\{\boldsymbol{Y}_{obs+\Delta t}, \boldsymbol{Y}_{obs+2\Delta t}, \dots, \boldsymbol{Y}_{obs+n\Delta t}\}$, where *obs* is the observation duration, and $n\Delta t$ indicates the prediction duration. Furthermore, for each predicted result, it consists of three types of human intention, i.e., $\boldsymbol{Y}_t =$

$[a_t, o_t, (x_t^1, y_t^1, x_t^2, y_t^2, \dots, x_t^{13}, y_t^{13})\ ]$, where a_t is the intended action for Task 1, o_t is the desired object for Task 2, and $(x_t^1, y_t^1, x_t^2, y_t^2, \dots, x_t^{13}, y_t^{13})$ is the future body movement for Task 3.

The model was trained end-to-end to minimize a joint loss function that is the weighted combination of losses for each task. The joint loss function is formulated as $L(\theta) = w_1 L_1(\theta) + w_2 L_2(\theta) + w_3 L_3(\theta)$, where L_1 and L_2 are categorical cross-entropy loss and L_3 is mean squared error (MSE) loss. The weights could be set based on the relatively importance of each task and the scale of each loss function. In this study, a series of trial experiments were conducted, and the weights are set as $w_1 : w_2 : w_3 = 1 : 1 : 5$ for better performance.

3.3. FedHIP – Federated Learning for Privacy Preserving Model Training

In the proposed FedHIP framework, federated learning mechanism was deployed to train the above multi-level intention prediction model collectively among multiple participants without sharing individual data (see Figure 3). Specifically, in the considered FL system, there is a central server which orchestrates the training process. Each client is a person that performs assembly tasks, whose behavior data are continuously collected via cameras. Due to privacy concerns, the data is stored locally and never disclosed to third parties. The training process of FedHIP can be summarized as follow:

In each communication round,

1. The server maintains a global model, and the server broadcasts the global model to all clients.
2. The client receives the global model and uses it to initialize its local model parameter.
3. The client updates the local model on its local training data with several local epochs.
4. The client uploads the local model to the cloud server.
5. The server receives the local models from all clients and uses them to compute the latest global model (through different aggregation algorithms discussed below).

In the first communication round, two approaches were considered for global model initialization in Step 1: 1) The global model is randomly initialized to consider the scenario where there is no existing model or public dataset for the intended task (i.e., multi-level human intention prediction in this study), and the model is trained solely relying on data collected from individual clients via FL mechanism. 2) The initial global model is pre-trained using data from partial clients that are used to simulate potential public dataset or data collected from initial experiments without privacy concern. After that, the pre-trained global model is updated via FL training process following steps 2-5. In this study, the number of clients used to pre-train initial global model is varied to examine the impact of pre-train data size on the model performance.

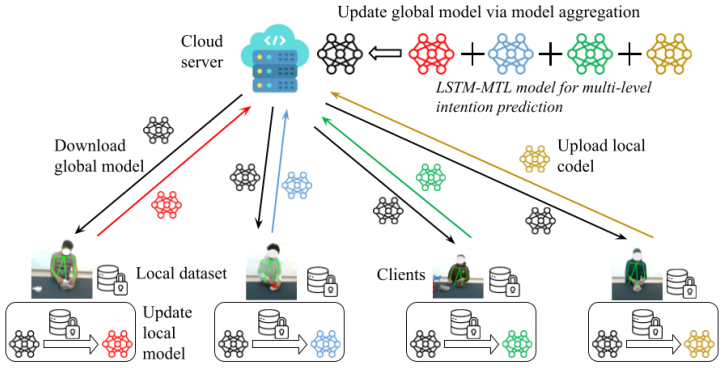


Figure 3. Architecture of proposed FedHIP

In FL, the critical process is the update of global model from the aggregation of local models in Step 5, and model aggregation algorithms will directly influence the performance of global model. This study will compare two commonly used aggregation algorithms in FL, i.e., FedAvg [43] and FedProx [56]. FedAvg is the first aggregation algorithm in FL and demonstrates empirical success in different settings. In FedAvg, the weights of different local models are averaged to obtain the weights of global model. Despite the simplicity, FedAvg may suffer from high heterogeneity of local data. On the other hand, FedProx is developed as a generalization of FedAvg by allowing variable amounts of work performed by different clients. Thus, it provides

improved robustness and stability for heterogeneous federated networks. Considering the potential heterogeneity of behavior data for different people, both aggregation algorithms were adopted and compared.

4. Implementation

4.1.Design of Collaborative Assembly Tasks

To validate the efficacy and evaluate the performance of the proposed framework, six collaborative assembly tasks were designed, which involve different actions (e.g., pick, carry, assemble, etc.) and components (e.g., main structure, connector). 54 experiments were conducted and recorded where each of nine participants performed all six tasks.

In this study, column-like stone structures were designed where unprocessed stones were connected by 3D-printed connectors for dry-joined assembly without any adhesives. Due to the perfect fit of dry-stacked unprocessed stones and the digitally designed and fabricated connectors, these structures required high precision during assembly [57]. For each structure, stones were aligned along the vertical thrust line with their center of mass. To accommodate unique geometry of each stone, the stones 3D scanned and reconstructed via photogrammetry techniques and connectors were digitally designed and 3D-printed based on the arrangement of the stones. Six structures were designed with different levels of complexity, where tasks 1-3 consistent of three stones and two connectors for more complex scenarios, and tasks 4-6 include two stones and one connector for simpler scenarios (see Figure 4).

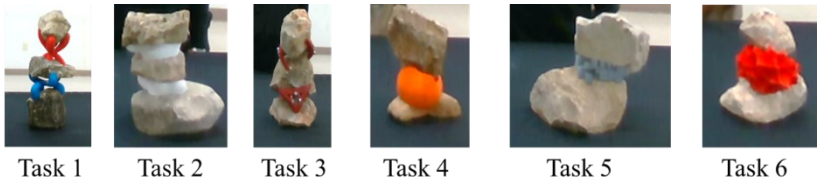
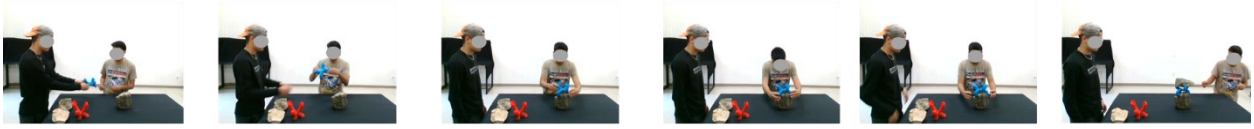


Figure 4. Assemble tasks (Tasks 1-3 are more complex than Tasks 4-6)

353 The designed assembly tasks require the participants to adjust relative positions and
354 orientations between stones and connectors to precisely align with the design, so that the structure
355 is stable without the need of extra fixation, such as glues, nails, and screws. Therefore, it could
356 simulate complex assembly tasks in construction that require human dexterity and experience,
357 where robots could serve as assistants to hand over components and tools following human needs.
358 Furthermore, the study on 3D printing and robotic-assisted assembly could also facilitate the
359 development of contemporary architecture construction using irregular materials [57].

360 In HRC research, studies have shown that robot could learn from conventions in human-
361 human interactions [58], thus, human behavior collected during human-human collaboration in
362 similar settings could be used to train models for human intention prediction. Such configuration
363 makes experiments and data collection easier and more achievable. Therefore, in this study, the
364 above tasks were performed collaboratively by nine groups of participants. In each group, one
365 person simulates the robot assistant to pass stones and connectors to another person who acts like
366 the human partner in HRC and only focuses on the assembly tasks. There are three object classes
367 considered, i.e., stone, connector, and stone with connector. In addition, the designed tasks involve
368 six types of actions, i.e., pick, carry, assemble, adjust, inspect, and release (see Figure 5). The
369 duration and order of each action varies across different participants and different tasks. Some
370 actions may be even absent from certain experiments, for instance, some participants may not
371 “inspect” and/or “adjust” the components when assembling simple structures. Such variability
372 leads to heterogeneity in the data distribution generated by different participants in different tasks
373 and highlights the importance of generalization capability for deep learning-based human intention
374 prediction.



(a) Pick (b) Carry (c) Assemble (d) Adjust (e) Release (e) Inspect

Figure 5. Samples actions involved in assembly tasks

4.2.Data Collection and Pre-processing

A total of 54 videos were collected at 15 frames per second (fps) where 9 participants were recorded to perform all of six assembly tasks, resulting in a total of 27,314 image frames. The movements of upper human body (with 13 key points) were first extracted from the videos using the Pose Flow algorithm [50]. Besides, the actions and objects involved in each frame were manually annotated. Based on relevant studies in motion prediction, e.g., [59], the observation duration was set to 400ms (i.e., 6 frames) and the prediction duration was set to 400ms (i.e., 6 frames) for short-term prediction.

In the proposed FedHIP, each person is treated as a client with their own data forming local datasets that do not share with others to protect privacy. Then, the local dataset for each person was randomly partitioned by 80% and 20% for training and testing, respectively, forming local training and test datasets. Global training and testing datasets were composed by pooling all people's training and testing data, respectively. In FedHIP, the global model is evaluated on the global testing dataset at every communication round. Table 1 lists sample size for each person (i.e., client). Figure 6 illustrates the heterogeneity in the distribution of different action classes across different people. It necessitates the adoption of aggregation algorithms designed for heterogeneous dataset like FedProx.

Table 1. Sample size by person

ID	1	2	3	4	5	6	7	8	9
Train	4380	4920	4086	4128	3018	4020	5304	5928	6906
Test	1092	1230	1020	1026	750	1002	1326	1482	1722

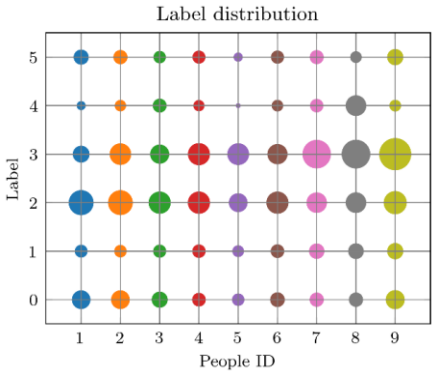


Figure 6. Distribution of different action classes across different people. X-axis shows person ID; Y-axis is the label for six different actions; Circle size represents the data fraction, computed as percentage of specific action class over entire data for specific person.

4.3.Experiment Settings

Inspired by relevant studies on FL in human activity recognition [16,60], five series of experiments with different training and testing configurations were conducted to analyze the efficacy and benefits of proposed FedHIP framework, compared to conventional centralized training and local training.

1. Centralized training was used as a benchmark, where the proposed LSTM-MTL model was trained on global training dataset. Such a mechanism is expected to achieve high prediction accuracy considering the use of all available training data. However, it poses significant privacy concerns by aggregating all local data for centralized training. Furthermore, to demonstrate the effectiveness of MTL mechanism for multi-level intention prediction, the results were compared with those obtained via conventional single task learning (STL), where three separate LSTM models were trained and tested for three types of intentions respectively.
2. Proposed FedHIP framework was trained from scratch, where the model is trained using local data and then used to update the global model (with global model randomly initialized in the first communication round). The resulting global model trained via FL paradigm is tested on

global dataset for accuracy evaluation. This approach may result in slightly lower prediction accuracy, but can preserve user privacy by eliminating data sharing and keeping data local.

3. The training and testing processes are the same as in the second scenario, expect that the initial global model was pre-trained by a public dataset (composed of training data from partial clients), and the final global model was evaluated on the test dataset consisting of test data from remaining clients. This scenario is to examine the influence of pre-trained global model on the performance of FedHIP. For experiments 2 and 3, both FedAvg and FedProx were used as model aggregation algorithm for comparison.
4. The model was trained and tested on local training and test dataset for each client, respectively. This setting is to simulate the training of personalized model, which may achieve highest performance on its local dataset and can eliminate privacy concern. However, the generalization ability may be poor and cannot work well in unseen data.
5. The model was trained on local training dataset while tested on global dataset. This evaluation provides valuable insights on how the model can be generalized on unseen data, which is crucial for human intention prediction considering the heterogeneity in performed tasks and human behavior.

Table 2 summarizes the settings and difference of five experiments. For all experiments, hyper-parameters were set as follows, learning rate: 0.0002; batch size: 10; optimizer: Adam; and dropout: 0.2; local epoch (for FL): 1; the weights of loss functions for the three tasks were set as 1:1:5. The learning rate was decayed by half at 50% and 75% of total training rounds. For FedProx, the coefficient of regularization was set as 0.005.

Table 2 Experiment configuration

Experiment	Training mechanism	Global model initialization	Training data	Testing data

1	Centralized learning	N/A	Global dataset	Global dataset
2	Federated learning	Randomized	Local dataset	Global dataset
3	Federated learning	Pre-trained	Local dataset	Global dataset
4	Local learning	N/A	Local dataset	Local dataset
5	Local learning	N/A	Local dataset	Global dataset

4.4.Evaluation Metrics

In this study, the performance of action and object prediction was primarily evaluated using accuracy, computed as $Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \times 100\%$, where *TP* is true positive; *TN* is true negative; *FP* is false positive; and *FN* is false negative. It measures the percentage of images being correctly classified for action and object classes. For motion prediction, two metrics were used, i.e., average displacement error (ADE) and Final displacement error (FDE). ADE measures the average distance between predicted positions of all key points and the ground-truth positions across all predicted time steps, calculated as $\frac{1}{N \times T} \sum_{i=1}^{N} \sum_{t=0}^{t=T} \left\| \hat{\boldsymbol{y}}_t^i - \boldsymbol{y}_t^i \right\|$, where N is sample size, $\hat{\boldsymbol{y}}_t^i$ is predicted positions of i^{th} data at time t, $\boldsymbol{y}_t^i$ is ground-truth position of i^{th} data at time t, and T is prediction duration. FDE measures distance between final predicted key point positions and ground-truth positions, computed as $\frac{1}{N} \sum_{i=1}^{N} \left\| \hat{\boldsymbol{y}}_T^i - \boldsymbol{y}_T^i \right\|$. It is noted that to make the results comparison consistent across different experiments, the evaluation metrics was computed as the average performance among all clients, e.g., the reported accuracy for Task 1 action prediction is the average accuracy on each client's testing dataset.

5. Results and Discussion

5.1.Performance of Centralized Model for Multi-level Intention Prediction

The proposed multi-task LSTM model is effective for multi-level human intention prediction, and achieves an accuracy of 87.7% and 97.3% for action and object prediction, respectively, and an ADE of 10.2 pixels and an FDE of 11.7 pixels in movement prediction. Table

3 lists the comparison between MTL and STL regarding both prediction performance and processing time. From Table 3, MTL achieves significant improvement in computational efficiency, with 59.0% and 55.6% reduction in training and test time, respectively, compared to STL. It shows a major advantage in MTL, especially in the application of HRC where computing resources are limited and real-time performance is required. Regarding prediction performance, MTL achieves better accuracy in action prediction and almost the same high accuracy in objective prediction while higher ADE and FDE, compared to STL. A potential reason for the improvement in action prediction could be that the prediction of body motion facilitates the prediction of intended action, as human's potential action is reflected by their body movement. On the other hand, future body movement is primarily predicted based on the observed body movement, the incorporation of other task may not improve the performance. Depending on the importance of each task, the weights in loss function (Section 3.2) could be further fine-tuned to improve the performance of corresponding tasks in MTL.

Table 3 Comparison between multi-task learning and single-task learning

Model	Training time (s)	Test time (s)	Task 1 – Action prediction	Task 2 – Object prediction	Task 3 – Body motion prediction	
			Accuracy (%)	Accuracy (%)	ADE (pixel)	FDE (pixel)
MTL	9694	0.0016	87.7	97.3	10.2	11.7
STL	23626	0.0036	86.5	97.9	4.7	6.6

Table 4 and Table 5 further lists the F1 score of each class in Task 1 and Task 2, where $F_1 = \frac{2TP}{2TP+FP+FN}$, which shows that MTL and traditional STL perform similar on each class. Figure 7 illustrates the confusion matrix for action and object prediction in MTL, which shows the classification performance of each class. For action prediction, the accuracy of "release" and "inspect" was relatively low, which could potentially be caused by the smaller sample size.

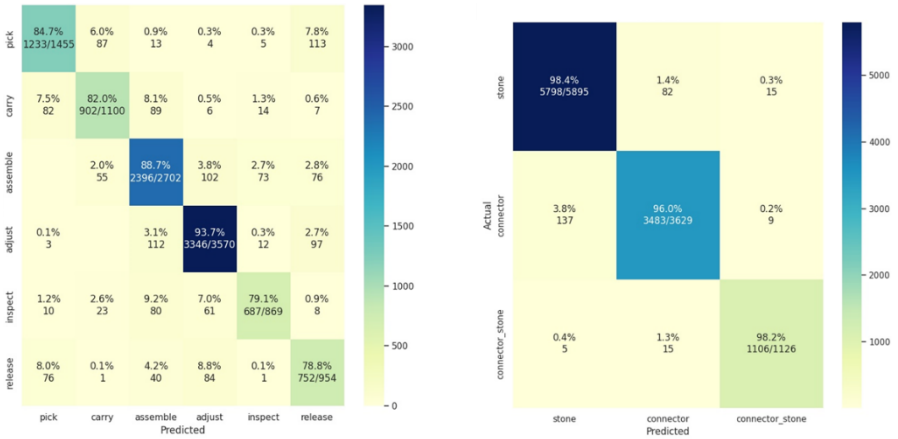
Moreover, misclassifications are more likely to occur in similar actions that appear alternatively during the task, such as “adjust” and “release”, and “adjust” and “assemble”. In this case, additional cues (e.g., visual attention, scene images) may be needed to better differentiate them. For object prediction, the accuracy is very high considering the limited number of classes as well as the inclusion of observed objects as contextual information. Besides, Figure 8 shows a sample prediction result for body movement prediction.

Table 4 F1 score for Task 1-action prediction

F1 score	Pick	Carry	Assemble	Adjust	Inspect	Release
MTL	0.86	0.82	0.88	0.94	0.83	0.76
STL	0.86	0.83	0.87	0.93	0.81	0.74

Table 5 F1 score for Task 2-object prediction

F1 score	Stone	Connector	Connector_stone
MTL	0.98	0.97	0.98
STL	0.98	0.97	0.99



(a) Action prediction (b) Object prediction

Figure 7. Confusion matrix for action and object prediction

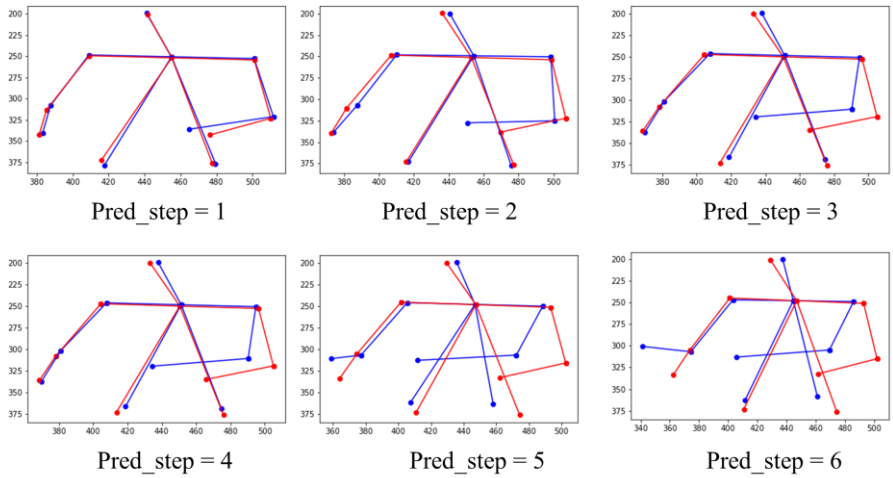


Figure 8. Demonstration of predicted body movements over six prediction steps (Blue lines indicate ground truth poses, and red lines indicate predicted poses).

5.2.Comparison of Centralized Learning and FedHIP

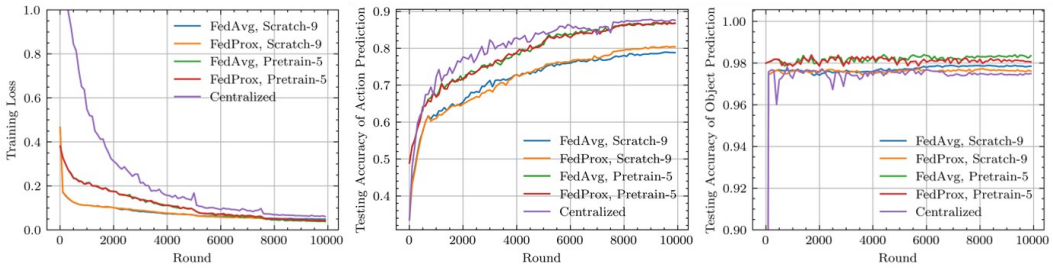
Table 6 lists the accuracy of all three tasks for multi-level human intention prediction using both centralized learning and different settings of FedHIP (i.e., Experiments 1-3 in Table 2). Compared to centralized learning, FL (especially pre-trained FL) can achieve compatible accuracy, despite a slight decrease in accuracy of action prediction and body motion prediction. It is reasonable because centralized model used the entire dataset for training, while the advantage of FL is to use local dataset to protect data privacy, while still maintaining compatible accuracy. In comparison of two aggregation algorithms, FedProx achieves better performance than FedAvg, especially when the global model is randomly initiated, resulting in higher level of heterogeneity of local training data.

Table 6 Prediction accuracy of centralized learning and FedHIP

Model	Task 1 – Action prediction	Task 2 – Object prediction	Task 3 – Body motion prediction	
	Accuracy (%)	Accuracy (%)	ADE (pixel)	FDE (pixel)
Centralized model	87.7	97.3	10.2	11.7

FedHIP_Avg (w/o pre-train)	78.5	97.9	13.7	15.1
FedHIP_Prox (w/o pre-train)	80.1	97.6	12.7	14.0
FedHIP_Avg (pre-train with 5 clients)	86.0	98.0	11.8	13.1
FedHIP_Prox (pre-train with 5 clients)	85.7	97.8	11.7	13.0

In the case of pre-trained FedHIP, the results using two algorithms are almost the same expect for a slight improvement of body motion prediction via FedProx. This might be because that the global model is pre-trained with clients' data performing the same assembly task, thus the remaining local training data is less heterogeneous compared to the case without pretrain. Furthermore, pretrained FedHIP leads to significant improvements in prediction accuracy compared to its counterparts without pretrain. Figure 9 further shows the training loss and prediction accuracy of each model across communication rounds. It could be seen that FedHIP achieves compatible convergence as centralized learning.



(a) Training loss (b) Accuracy of action prediction (c) Accuracy of object prediction

Figure 9 Training loss and accuracy across communication rounds

5.3.Influence of Pretrain Data on FedHIP Prediction Accuracy

In the pretrained FedHIP experiments, the number of clients used for pretrain was varied (i.e., n = 3, 5, and 7) to evaluate the influence of pretrain data on FedHIP performance (Figure 11). Specifically, for each scenario, a certain number of clients were randomly selected to pretrain the global model, and the data from remaining clients were used to train and test FedHIP. Considering

the various combinations of clients used for pretrain, each scenario was repeated for three trials and the average accuracy was computed. From Figure 10, in general, the increase in number of clients used for pre-train can improve the accuracy. It could be considered to combine the advantages of both centralized learning and FL, which improves the model generalizability leveraging potential public dataset that does not have privacy concern. It was also noted that the accuracy for action prediction and object prediction was slightly decreased when 7 clients were used to pretrain the global model, which could be potentially caused by the large variation in the testing data of the remaining clients.

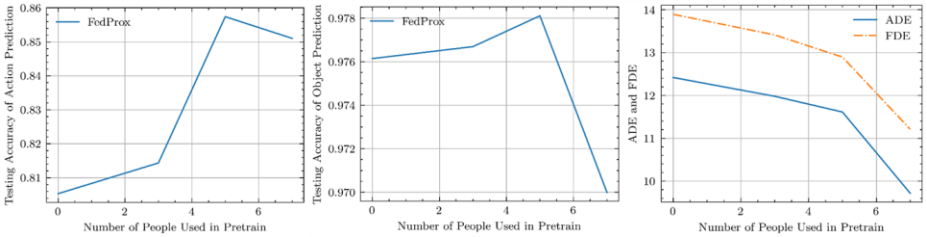


Figure 10. Change of prediction accuracy based on number of clients used in FL pre-train

5.4.Comparison of Local Training and FedHIP

Table 7 lists the comparison of prediction accuracy of local training and FedHIP (i.e., Experiments 2-5 in Table 2). It was found that training and testing using only local data achieves the highest accuracy. It is expected because in this study, local dataset is relatively large and sufficient to train the HIP model. Furthermore, local datasets exhibit much less heterogeneity in contrast to global datasets, which is not sufficient to prove the generalizability of the model. As a more realistic setting, models trained using local data were tested on global dataset (i.e., unseen data distribution), which led to much lower accuracy for all three tasks. Regarding the proposed FedHIP framework, the pre-trained version achieves compatible accuracy as local training (test on local data) for action and object prediction. Both FedHIP frameworks (with and without pretrain) achieve much higher accuracy than local training (test on global data) in all three tasks. This proves

the benefits of FL in human intention prediction – through collaborative training on multiple individual clients, it can achieve high accuracy while preserving data privacy.

Table 7 Prediction accuracy of local learning and FedHIP

Model	Task 1 – Action prediction	Task 2 – Object prediction	Task 3 – Body motion prediction	
	Accuracy (%)	Accuracy (%)	ADE (pixel)	FDE (pixel)
Local training (test on local data)	87.0	97.5	7.5	8.7
Local training (test on global data)	42.6	90.7	20.7	23.7
FedHIP_Avg (w/o pre-train)	78.5	97.9	13.7	15.1
FedHIP_Prox (w/o pre-train)	80.1	97.6	12.7	14.0
FedHIP_Avg (pre-train with 5 clients)	86.0	98.0	11.8	13.1
FedHIP_Prox (pre-train with 5 clients)	85.7	97.8	11.7	13.0

Conclusion

HRC is a promising solution to relieve construction workers from repetitive and physically demanding tasks, thus improving construction safety and productivity, and overcoming the challenges posed by labor shortage and workforce aging. To enable efficient and safe HRC, it is critical for robots to understand and predict human intention and thus proactively and adaptively planning their own tasks and motions. This study proposes a vision-based FedHIP framework for privacy-preserving multi-level human intention prediction in human-robot collaborative assembly tasks. Specifically, taking body movements and assembly components as inputs, an LSTM based MTL model was developed to simultaneously predict multi-level human intention, including high-level actions and objects, and low-level body movements. FL was employed to train the model in a distributed and privacy-preserving way on local clients without transmitting sensitive data.

The proposed framework was validated using video data from 6 assembly tasks conducted by 9 people. First, the efficacy of proposed LSTM-MTL model for multi-level human intention prediction was demonstrated via conventional centralized training. It achieves an accuracy of 87.7% and 97.3% for action and object prediction, respectively, and an ADE of 10.2 pixels and an FDE of 11.7 pixels in movement prediction. Second, the proposed FedHIP was implemented. The results show that the proposed FedHIP without and with pre-train can achieve an accuracy of 80.1% and 85.7% in action prediction, 97.6% and 97.8% in object prediction, an ADE of 12.7 and 11.7 pixels and an FDE of 14 and 13 pixels of in motion prediction, respectively. It was found that pre-trained global model in FL could significantly increase the prediction accuracy, which could be realized using public datasets and/or preliminary experiments of similar tasks. Third, the proposed framework was compared with conventional centralized training and local training. It was found that FL leads to compatible prediction accuracy with centralized training and much higher accuracy than local training while preserving data privacy.

There remain some limitations that deserve future research. First, 2D skeleton positions were used in this study, which could be expanded to 3D motion obtained from other sources, such as stereo vision and motion capturing system. Second, types of objects involved in the current work status were manually annotated, and the process could be automated in future study by integrating with object detection module. Third, human behavior data were collected in human-human interaction setting, with the premise that HRC could learn from human-human interaction conventions. In our ongoing study, data was collected in HRC assembly tasks, which will be used to evaluate the proposed framework. Fourth, small-scale stone assembly tasks in the controlled environment were designed for data collection and to train and test the proposed FedHIP framework. It is expected simulate complex assembly tasks in construction that require human

dexterity and experience, where robots could serve as assistants to hand over components and tools following human needs. In future study, experiments of different construction tasks in real-world settings should be conducted to collect more diverse data to test the performance of proposed method. Finally, considering the heterogeneity of different people performing different tasks, personalized FL framework will be developed in future research to further improve the performance. Moreover, other modern models, such as transformer-based models and ST-GCN based models could be used as backbone in the model to test the usability and performance in multi-level human intention prediction.

Acknowledgements

This research was funded by the U.S. National Science Foundation (NSF) via Grants 2138514, 2222670, and 2222810. The authors gratefully acknowledge NSF's support. Any opinions, findings, recommendations, and conclusions in this paper are those of the authors, and do not necessarily reflect the views of NSF, The University of Texas at San Antonio, The University of Tennessee, Knoxville, and Darmstadt University of Applied Sciences.

References

[1] G. Wilson, C. Pereyda, N. Raghunath, G. de la Cruz, S. Goel, S. Nesaei, B. Minor, M. Schmitter-Edgecombe, M.E. Taylor, D.J. Cook, Robot-enabled support of daily activities in smart home environments, Cognitive Systems Research. 54 (2019) 258–272. https://doi.org/10.1016/j.cogsys.2018.10.032.

[2] V. Vasco, A.G.P. Antunes, V. Tikhanoff, U. Pattacini, L. Natale, V. Gower, M. Maggiali, HR1 Robot: An Assistant for Healthcare Applications, Frontiers in Robotics and AI. 9 (2022). https://doi.org/10.3389/frobt.2022.813843.

[3] E. Matheson, R. Minto, E.G.G. Zampieri, M. Faccio, G. Rosati, Human-robot collaboration

in manufacturing applications: A review, Robotics. 8 (2019). https://doi.org/10.3390/robotics8040100.

[4] S. Park, H. Yu, C.C. Menassa, V.R. Kamat, A Comprehensive Evaluation of Factors Influencing Acceptance of Robotic Assistants in Field Construction Work, Journal of Management in Engineering. 39 (2023). https://doi.org/10.1061/jmenea.meeng-5227.

[5] K.S. Saidi, T. Bock, C. Georgoulas, Robotics in construction, in: Springer Handbook of Robotics, 2016: pp. 1493–1519. https://doi.org/10.1007/978-3-319-32552-1_57.

[6] J. Cai, A. Du, X. Liang, S. Li, Prediction-Based Path Planning for Safe and Efficient Human–Robot Collaboration in Construction via Deep Reinforcement Learning, Journal of Computing in Civil Engineering. 37 (2023). https://doi.org/10.1061/(asce)cp.1943-5487.0001056.

[7] C.-J. Liang, X. Wang, V.R. Kamat, C.C. Menassa, Human–Robot Collaboration in Construction: Classification and Research Trends, Journal of Construction Engineering and Management. 147 (2021). https://doi.org/10.1061/(asce)co.1943-7862.0002154.

[8] H. Wu, H. Li, X. Fang, X. Luo, A survey on teaching workplace skills to construction robots, Expert Systems with Applications. 205 (2022). https://doi.org/10.1016/j.eswa.2022.117658.

[9] S. Li, P. Zheng, J. Fan, L. Wang, Toward Proactive Human-Robot Collaborative Assembly: A Multimodal Transfer-Learning-Enabled Action Prediction Approach, IEEE Transactions on Industrial Electronics. 69 (2022) 8579–8588. https://doi.org/10.1109/TIE.2021.3105977.

[10] F. Koochaki, L. Najafizadeh, A Data-Driven Framework for Intention Prediction via Eye Movement with Applications to Assistive Systems, IEEE Transactions on Neural Systems and Rehabilitation Engineering. 29 (2021) 974–984. https://doi.org/10.1109/TNSRE.2021.3083815.

[11] P. Kratzer, N.B. Midlagajni, M. Toussaint, J. Mainprice, Anticipating Human Intention for Full-Body Motion Prediction in Object Grasping and Placing Tasks, in: 29th IEEE International Conference on Robot and Human Interactive Communication, RO-MAN 2020, 2020: pp. 1157–1163. https://doi.org/10.1109/RO-MAN47096.2020.9223547.

[12] R.C. Luo, L. Mai, Human Intention Inference and On-Line Human Hand Motion Prediction for Human-Robot Collaboration, in: IEEE International Conference on Intelligent Robots and Systems, 2019: pp. 5958–5964. https://doi.org/10.1109/IROS40897.2019.8968192.

[13] W. Wang, R. Li, Y. Chen, Y. Sun, Y. Jia, Predicting Human Intentions in Human-Robot Hand-Over Tasks Through Multimodal Learning, IEEE Transactions on Automation Science and Engineering. 19 (2022) 2339–2353. https://doi.org/10.1109/TASE.2021.3074873.

[14] Z. Ding, C. Yang, Z. Wang, X. Yin, F. Jiang, Online adaptive prediction of human motion intention based on semg, Sensors. 21 (2021). https://doi.org/10.3390/s21082882.

[15] X. Li, H. lin Chi, W. Lu, F. Xue, J. Zeng, C.Z. Li, Federated transfer learning enabled smart work packaging for preserving personal image information of construction worker, Automation in Construction. 128 (2021). https://doi.org/10.1016/j.autcon.2021.103738.

[16] K. Sozinov, V. Vlassov, S. Girdzijauskas, Human activity recognition using federated learning, in: 2018 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications (ISPA/IUCC/BDCloud/SocialCom/SustainCom), 2019: pp. 1103–1111. https://doi.org/10.1109/BDCloud.2018.00164.

[17] Z. Xiao, X. Xu, H. Xing, F. Song, X. Wang, B. Zhao, A federated learning system with

enhanced feature extraction for human activity recognition, Knowledge-Based Systems. 229 (2021). https://doi.org/10.1016/j.knosys.2021.107338.

[18] X. Ouyang, Z. Xie, J. Zhou, G. Xing, J. Huang, ClusterFL: A Clustering-based Federated Learning System for Human Activity Recognition, ACM Transactions on Sensor Networks. 19 (2022). https://doi.org/10.1145/3554980.

[19] T. Jiang, N.C. Camgoz, R. Bowden, Skeletor: Skeletal transformers for robust body-pose estimation, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2021: pp. 3389–3397. https://doi.org/10.1109/CVPRW53098.2021.00378.

[20] Z. Liu, Q. Liu, W. Xu, L. Wang, Z. Ji, Adaptive real-time similar repetitive manual procedure prediction and robotic procedure generation for human-robot collaboration, Advanced Engineering Informatics. 58 (2023). https://doi.org/10.1016/j.aei.2023.102129.

[21] F. Semeraro, A. Griffiths, A. Cangelosi, Human–robot collaboration and machine learning: A systematic review of recent research, Robotics and Computer-Integrated Manufacturing. 79 (2022). https://doi.org/10.1016/j.rcim.2022.102432.

[22] A. Cunha, F. Ferreira, E. Sousa, L. Louro, P. Vicente, S. Monteiro, W. Erlhagen, E. Bicho, Towards collaborative robots as intelligent co-workers in human-robot joint tasks: What to do and who does it?, in: 52nd International Symposium on Robotics, ISR 2020, 2020: pp. 141–148.

[23] E.C. Grigore, A. Roncone, O. Mangin, B. Scassellati, Preference-Based Assistance Prediction for Human-Robot Collaboration Tasks, in: IEEE International Conference on Intelligent Robots and Systems, 2018: pp. 4441–4448. https://doi.org/10.1109/IROS.2018.8593716.

[24] S. Choi, K. Lee, H.A. Park, S. Oh, A Nonparametric Motion Flow Model for Human Robot Cooperation, in: Proceedings - IEEE International Conference on Robotics and Automation, 2018: pp. 7211–7218. https://doi.org/10.1109/ICRA.2018.8463201.

[25] D. Shukla, O. Erkent, J. Piater, Learning semantics of gestural instructions for human-robot collaboration, Frontiers in Neurorobotics. 12 (2018). https://doi.org/10.3389/fnbot.2018.00007.

[26] Q.X. Long, X.J. Tang, Q.L. Shi, Q. Li, H.J. Deng, J. Yuan, J.L. Hu, W. Xu, Y. Zhang, F.J. Lv, K. Su, F. Zhang, J. Gong, B. Wu, X.M. Liu, J.J. Li, J.F. Qiu, J. Chen, A.L. Huang, Clinical and immunological assessment of asymptomatic SARS-CoV-2 infections, Nature Medicine. 26 (2020) 1200–1204. https://doi.org/10.1038/s41591-020-0965-6.

[27] L. Roveda, S. Haghshenas, M. Caimmi, N. Pedrocchi, L.M. Tosatti, Assisting operators in heavy industrial tasks: On the design of an optimized cooperative impedance fuzzy-controller with embedded safety rules, Frontiers in Robotics and AI. 6 (2019). https://doi.org/10.3389/frobt.2019.00075.

[28] Y. Liu, M. Habibnezhad, H. Jebelli, Brainwave-driven human-robot collaboration in construction, Automation in Construction. 124 (2021). https://doi.org/10.1016/j.autcon.2021.103556.

[29] H. Yu, V.R. Kamat, C.C. Menassa, W. McGee, Y. Guo, H. Lee, Mutual physical state-aware object handover in full-contact collaborative human-robot construction work, Automation in Construction. 150 (2023). https://doi.org/10.1016/j.autcon.2023.104829.

[30] X. Wang, S. Wang, C.C. Menassa, V.R. Kamat, W. McGee, Automatic high-level motion sequencing methods for enabling multi-tasking construction robots, Automation in Construction. 155 (2023). https://doi.org/10.1016/j.autcon.2023.105071.

693 [31] T. Zhang, H. Mo, Reinforcement learning for robot research: A comprehensive review and
694 open issues, International Journal of Advanced Robotic Systems. 18 (2021).
695 https://doi.org/10.1177/17298814211007305.
696 [32] C.J. Liang, V.R. Kamat, C.C. Menassa, Teaching robots to perform quasi-repetitive
697 construction tasks through human demonstration, Automation in Construction. 120 (2020).
698 https://doi.org/10.1016/j.autcon.2020.103370.
699 [33] P.B. Rodrigues, R. Singh, M. Oytun, P. Adami, P.J. Woods, B. Becerik-Gerber, L.
700 Soibelman, Y. Copur-Gencturk, G.M. Lucas, A multidimensional taxonomy for human-
701 robot interaction in construction, Automation in Construction. 150 (2023).
702 https://doi.org/10.1016/j.autcon.2023.104845.
703 [34] M. Zhang, R. Xu, H. Wu, J. Pan, X. Luo, Human–robot collaboration for on-site
704 construction, Automation in Construction. 150 (2023).
705 https://doi.org/10.1016/j.autcon.2023.104812.
706 [35] H. Liu, L. Wang, Human motion prediction for human-robot collaboration, Journal of
707 Manufacturing Systems. 44 (2017) 287–294. https://doi.org/10.1016/j.jmsy.2017.04.009.
708 [36] J. Lyu, A. Maýe, M. Görner, P. Ruppel, A.K. Engel, J. Zhang, Coordinating human-robot
709 collaboration by EEG-based human intention prediction and vigilance control, Frontiers in
710 Neurorobotics. 16 (2022). https://doi.org/10.3389/fnbot.2022.1068274.
711 [37] Z. Liu, Q. Liu, W. Xu, Z. Liu, Z. Zhou, J. Chen, Deep learning-based human motion
712 prediction considering context awareness for human-robot collaboration in manufacturing,
713 Procedia CIRP. 83 (2019) 272–278. https://doi.org/10.1016/j.procir.2019.04.080.
714 [38] H. Wu, H. Li, H.L. Chi, Z. Peng, S. Chang, Y. Wu, Thermal image-based hand gesture
715 recognition for worker-robot collaboration in the construction industry: A feasible study,

Advanced Engineering Informatics. 56 (2023). https://doi.org/10.1016/j.aei.2023.101939.

[39] Y. Zhang, K. Ding, J. Hui, J. Lv, X. Zhou, P. Zheng, Human-object integrated assembly intention recognition for context-aware human-robot collaborative assembly, Advanced Engineering Informatics. 54 (2022). https://doi.org/10.1016/j.aei.2022.101792.

[40] T. Zhou, Y. Wang, Q. Zhu, J. Du, Human hand motion prediction based on feature grouping and deep learning: Pipe skid maintenance example, Automation in Construction. 138 (2022). https://doi.org/10.1016/j.autcon.2022.104232.

[41] T. Zhou, Q. Zhu, Y. Shi, J. Du, Construction Robot Teleoperation Safeguard Based on Real-Time Human Hand Motion Prediction, Journal of Construction Engineering and Management. 148 (2022) 04022040. https://doi.org/10.1061/(asce)co.1943-7862.0002289.

[42] X. Xia, T. Zhou, J. Du, N. Li, Human motion prediction for intelligent construction: A review, Automation in Construction. 142 (2022). https://doi.org/10.1016/j.autcon.2022.104497.

[43] H. Brendan McMahan, E. Moore, D. Ramage, S. Hampson, B. Agüera y Arcas, H.B. McMahan, E. Moore, D. Ramage, S. Hampson, B.A. y Arcas, Communication-efficient learning of deep networks from decentralized data, in: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 2017: pp. 63–69.

[44] C. Ma, J. Li, M. Ding, H.H. Yang, F. Shu, T.Q.S. Quek, H. Vincent Poor, On safeguarding privacy and security in the framework of federated learning, IEEE Network. 34 (2020) 242–248. https://doi.org/10.1109/MNET.001.1900506.

[45] T. Wang, Y. Du, Y. Gong, K.K. Raymond Choo, Y. Guo, Applications of Federated Learning in Mobile Health: Scoping Review, Journal of Medical Internet Research. 25 (2023). https://doi.org/10.2196/43006.

[46] C. Niu, F. Wu, S. Tang, L. Hua, R. Jia, C. Lv, Z. Wu, G. Chen, Billion-Scale Federated Learning on Mobile Clients: A Submodel Design with Tunable Privacy, in: Proceedings of the 26th Annual International Conference on Mobile Computing and Networking, 2020: pp. 1–14.

[47] Z. Zhang, Z. Gao, Y. Guo, Y. Gong, Scalable and Low-Latency Federated Learning With Cooperative Mobile Edge Networking, IEEE Transactions on Mobile Computing. (2022). https://doi.org/10.1109/TMC.2022.3216837.

[48] R. Hu, Y. Guo, Y. Gong, Energy-Efficient Distributed Machine Learning at Wireless Edge with Device-to-Device Communication, in: IEEE International Conference on Communications, 2022: pp. 5208–5213. https://doi.org/10.1109/ICC45855.2022.9838508.

[49] J. Cai, X. Liang, B. Wibranek, Y. Guo, Multi-task Deep Learning-based Human Intention Prediction for Human-Robot Collaborative Assembly, in: ASCE International Conference on Computing in Civil Engineering (I3CE 2023), 2023.

[50] Y. Xiu, J. Li, H. Wang, Y. Fang, C. Lu, Pose flow: Efficient online pose tracking, in: British Machine Vision Conference 2018, BMVC 2018, BMVA Press, 2019.

[51] H.-S.S. Fang, S. Xie, Y.-W.W. Tai, C. Lu, S. Jiao Tong University, T. YouTu, RMPE: Regional Multi-person Pose Estimation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017: pp. 2353–2362. https://doi.org/10.1109/ICCV.2017.256.

[52] J. Cai, X. Li, X. Liang, W. Wei, S. Li, Construction Worker Ergonomic Assessment via LSTM-Based Multi-Task Learning Framework, in: 2022 Construction Research Congress, 2022: pp. 215–224. https://doi.org/10.1061/9780784483961.023.

[53] J. Cai, L. Yang, Y. Zhang, S. Li, H. Cai, Multitask Learning Method for Detecting the Visual

762 Focus of Attention of Construction Workers, Journal of Construction Engineering and
763 Management. 147 (2021) 4021063. https://doi.org/10.1061/(asce)co.1943-7862.0002071.
764 [54] Z. Huang, A. Hasan, K. Shin, R. Li, K. Driggs-Campbell, Long-Term Pedestrian Trajectory
765 Prediction Using Mutable Intention Filter and Warp LSTM, IEEE Robotics and Automation
766 Letters. 6 (2021) 542–549. https://doi.org/10.1109/LRA.2020.3047731.
767 [55] J. Cai, Y. Zhang, H. Cai, Two-step long short-term memory method for identifying
768 construction activities through positional and attentional cues, Automation in Construction.
769 106 (2019) 102886. https://doi.org/10.1016/j.autcon.2019.102886.
770 [56] T. Li, A.K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, V. Smith, Federated Optimization
771 in Heterogeneous Networks, (2018).
772 [57] B. Wibranek, O. Tessmann, Digital rubble compression-only structures with irregular rock
773 and 3D printed connectors, in: IASS Symposium 2019 - 60th Anniversary Symposium of
774 the International Association for Shell and Spatial Structures; Structural Membranes 2019
775 - 9th International Conference on Textile Composites and Inflatable Structures, FORM and
776 FORCE, 2019: pp. 2488–2495.
777 [58] N.C. Krämer, A. Von Der Pütten, S. Eimler, Human-agent and human-robot interaction
778 theory: Similarities to and differences from human-human interaction, Studies in
779 Computational Intelligence. 396 (2012) 215–240. https://doi.org/10.1007/978-3-642-
780 25691-2_9.
781 [59] W. Mao, M. Liu, M. Salzmann, H. Li, Learning trajectory dependencies for human motion
782 prediction, in: Proceedings of the IEEE International Conference on Computer Vision, 2019:
783 pp. 9488–9496. https://doi.org/10.1109/ICCV.2019.00958.
784 [60] S. Ek, F. Portet, P. Lalanda, G. Vega, Evaluation of federated learning aggregation

785 algorithms: application to human activity recognition, in: Adjunct Proceedings of the 2020
786 ACM International Joint Conference on Pervasive and Ubiquitous Computing and
787 Proceedings of the 2020 ACM International Symposium on Wearable Computers, 2020: pp.
788 638–643.
789