Original Article

Institution of MECHANICAL ENGINEERS

Journal of MECHANICAL ENGINEERING SCIENCE

Proc IMechE Part C:
J Mechanical Engineering Science
0(0) 1–14
© IMechE 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0954406220931547
journals.sagepub.com/home/pic
SAGE

# Fusing and refining convolutional neural network models for assembly action recognition in smart manufacturing

Md. Al-Amin[1], Ruwen Qin[1], Wenjin Tao[2], David Doell[1], Ravon Lingard[1], Zhaozheng Yin[3] and Ming C Leu[2]

**Abstract**

Assembly carries paramount importance in manufacturing. Being able to support workers in real time to maximize their positive contributions to assembly is a tremendous interest of manufacturers. Human action recognition has been a way to automatically analyze and understand worker actions to support real-time assistance for workers and facilitate worker–machine collaboration. Assembly actions are distinct from activities that have been well studied in the action recognition literature. Actions taken by assembly workers are intricate, variable, and may involve very fine motions. Therefore, recognizing assembly actions remains a challenging task. This paper proposes to simply use only two wearable devices that respectively capture the inertial measurement unit data of each hand of workers. Then, two convolutional neural network models with an identical architecture are independently trained using the two sources of inertial measurement unit data to respectively recognize the right-hand and the left-hand actions of an assembly worker. Classification results of the two convolutional neural network models are fused to yield a final action recognition result because the two hands often collaborate in assembling operations. Transfer learning is implemented to adapt the action recognition models to subjects whose data have not been included in dataset for training the models. One operation in assembling a Bukito three-dimensional printer, which is composed of seven actions, is used to demonstrate the implementation and assessment of the proposed method. Results from the study have demonstrated that the proposed approach effectively improves the prediction accuracy at both the action level and the subject level. Work of the paper builds a foundation for building advanced action recognition systems such as multimodal sensor-based action recognition.

**Keywords**

Manufacturing assembly, action recognition, inertial measurement unit sensor, model fusion, transfer learning

Date received: 31 August 2019; accepted: 8 May 2020

## Introduction

Fluctuating market demand, increasing needs for customized products, and recent unprecedented technological advancements are leading the manufacturing industry to adopt advanced practices of smart manufacturing.[1] By means of this paradigm shift, workplaces of future manufacturing are being transformed from task-centric to worker-centric. Consequently, the role of workers is expected to be more important than ever.[2] In human-centric manufacturing tasks, like assembly, recognizing actions that a worker is taking provides just-in-time information on mistakes or difficulties the worker may have, which allows for addressing those in a near real-time manner.

An action can be defined as a pattern of motions performed by individuals, which usually lasts for a short period of time with an intention.[3] Action recognition (AR) involves automatically detecting and recognizing human motions from sensor data.[4] Analyzing motions or actions has a long history and it has been an important research topic of various disciplines ranging from psychology to computer

[1]Department of Engineering Management and Systems Engineering, Missouri University of Science and Technology, Rolla, MO, USA
[2]Department of Mechanical and Aerospace Engineering, Missouri University of Science and Technology, Rolla, MO, USA
[3]Department of Computer Science, Department of Biomedical Informatics, Stony Brook University, Stony Brook, NY, USA

**Corresponding author:**
Ruwen Qin, Department of Engineering Management and Systems Engineering, Missouri University of Science and Technology, Rolla, MO 65409, USA.
Email: qinr@mst.edu

vision.[5] AR research is motivated by numerous applications such as surveillance,[6] fall detection,[7,8] assisted living,[7,9–11] human–machine interaction,[6,9,12] healthcare[6,8–10] to name a few.

Assembly carries paramount importance in manufacturing. Assembly accounts for over 50% of total production time and 20% of total production cost.[13] One step of assembly usually is composed of multiple actions with each having a clear objective. Failing to deliver the intermediate outcome from each action in good quality may impede the progress of assembly. Therefore, being able to support assembly workers in real time to maximize their positive contributions to assembly is a tremendous interest of manufacturers. However, AR in the worker-intensive manufacturing assembly is nontrivial. Actions that workers take in assembly are distinct from other types of activities widely studied in the AR literature. They are intricate, variable, and may involve very fine motions. Moreover, complex operations are continuously introduced to assembly to meet the growing demand for customized products. Consequently, the boundary of assembly actions is open, making the AR in assembly harder than ever. Despite the importance of recognizing assembly actions, limited efforts have been dedicated to this endeavor.[14–18]

Sensor data predominantly used by AR include video data captured from conventional RGB cameras, depth and skeleton data collected from depth cameras, and orientation, motion, and electromyography data from wearable sensors. Among these, wearable sensors have their unique advantages, such as easy to carry, inexpensive, sensitive, wireless, privacy preserved, and low in computational cost.[7,9,19] Due to these advantages, wearable sensors are particularly suitable for capturing intrinsic or fine motions of assembly workers. Inertial measurement unit (IMU) is often embedded in various wearable devices, which includes an accelerometer, a gyroscope, a magnetometer, or a combination of these. Because an IMU is able to capture the orientation and motion data of a human body part where it is attached, it is widely used for AR. However, data captured from a single unit are less reliable for capturing all relevant information of actions. For example, an action taken by a worker may involve the collaboration of multiple body parts. Assembly workers often use both the left and right hands in operations. Therefore, data from multiple IMUs complement each other and may enhance the reliability of recognition.[4]

Substantial efforts have been made for fusing multiple sensors attached to different parts of the human body to improve AR performance. Some novel ideas of this endeavor have been explored. For example, Guo et al. proposed a multisensor multiclassifier hierarchical fusion algorithm based on entropy weights and discussed the implication of feature dimension in classifying ten gym activities using five wearable inertial sensors.[20] Banos et al. developed a sensor weighting hierarchical classifier by combining the classification capability at the class level and the source level.[21] The work showed the impact of feature dimension on the ability to recognize nine daily living activities using five bi-axial accelerometers. Guo et al. designed adaptive weighted logarithmic opinion pools to classify 13 daily living activities.[22] This study demonstrated the significance of sensor modality using five pairs of tri-axial accelerometer and bio-axial gyroscope. Other researchers used straightforward yet effective sensor fusion techniques and examined several crucial issues like sensor placement, sensor degradation, interconnection failures, jitter, and so on. As for instance, Zappi et al. applied the majority voting and naïve Bayesian sensor fusion schemes to demonstrate the implication of sensor scalability and robustness in recognizing 10 activities of quality inspection in a car assembly line using 19 body-worn accelerometer sensors.[23] Zhu and Wang also adopted the majority voting in recognizing 13 daily activities using two inertial sensors and discussed the capability of sensor fusion in classifying fine and coarse grain actions.[24] Yet, in the real world of manufacturing, letting workers wear less sensors is highly desired. A recent study showed that the fusion of two deep learning classifiers, which were independently developed using IMU data from two different wearable devices, was a cost-effective way to largely improve the prediction accuracy of assembly actions.[25] Determining the ability of using minimum number of IMUs to recognize assembly actions would build a knowledge foundation for creating multimodal sensor based AR.

While both recurrent neural networks (RNN) and convolutional neural networks (CNN) have been used in the literature for human AR, this study proposes the latter for its two advantages over the former: signal dependency and scale invariance.[26] Signal dependency means sensor signals are likely to be correlated. This study arranges sensor signals as images fed to CNN, making it straightforward to discover both the temporal correlation of individual signal series and the between-series correlation. This, however, is more difficult for RNN. The scale invariant property refers to the robustness of the CNN algorithm in handling different sampling rates or frequencies. Yet, this could be an issue for RNN. Furthermore, the inference time of a RNN model such as a long short term memory network generally is longer than a CNN model, thus being challenging in the real-time AR and prediction.[27]

Annotated large datasets of assembly actions are not publicly available for multiple reasons. Yet, workers are heterogeneous in work habit, job efficiency, learning ability, and sensitivity to pressure. Assembling processes are diverse. A practical way is to, firstly, train AR models in a focused setting. For instance, training an AR model for workers who perform the same operation but in different shifts on a workstation of an assembly line. Then, scaling up the implementation of the model through adapting the trained model to new workers or transferring the model to other operations of assembly.

The study presented in this paper aims to meet the need for an AR model in assembly, which can be developed with a minimum number of wearable sensors and is transferable to new workers or new operations. For this purpose, the study first builds AR models for different body parts that collaborate in assembly operations. For example, if an operation is mainly performed by hands, two IMU are respectively attached to the two arms of each worker. The collected data are used to independently create two IMU-based AR models, one for the left hand and the other for the right hand. Then, the study investigates the effectiveness of model fusion in depth, at both the action level and the subject level. Beyond that, AR models are refined to adapt to new workers.

The remainder of the paper is organized as following. The next section presents the proposed approach to creating, refining and fusing IMU-based AR models. Then, experiments are designed for obtaining required data of study, followed by the assessment of the proposed approach. Findings from this study and important extensions of the current work are summarized at the end.

## The methodology of model fusion and calibration

This study set up two wearable sensors to respectively capture IMU data of two hands for training two deep neural networks (NN) for recognizing assembly actions. Then the two models were further fused to provide more reliable performance of AR. Transfer learning was implemented to adapt the AR models to new workers whose data were not included in the original training dataset. Details of the proposed methodology are presented below.

### Armbands with IMU sensor for data collection

Wearable devices used by this study are two Myo armbands developed by the Thalmic Labs,[28] worn on a worker's left and right forearms, respectively. The IMU in the armband is a nine-axis device that consists of a three-axis gyroscope, a three-axis accelerometer, and a three-axis magnetometer. It provides the data of the armband in 13 columns including the orientation (seven columns, in both quaternions and Euler angles), velocities (three columns), and accelerations (three columns), all at the sampling frequency of 50 Hz. Figure 1 illustrates a sample of the IMU time series data.

The two armbands were used to collect data for respectively training and testing two AR models: the left-hand IMU-based model (LH-IMU), $M_L$, and the right-hand IMU-based model (RH-IMU), $M_R$.

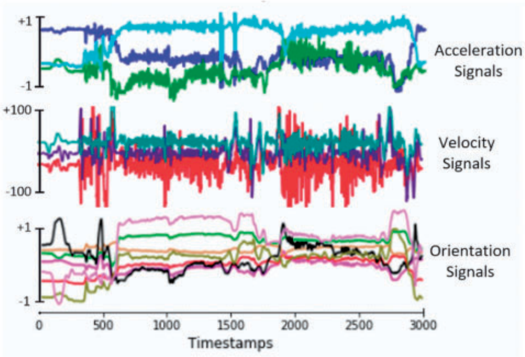


**Figure 1.** Thirteen-column time series data collected by the 9-axis IMU.

Data were repeatedly collected from $N_{wk}$ workers, indexed by $n$, who were performing the same assembly operation. Workers took a sequence of $N_{ac}$ actions, indexed by $k$, to complete the operation. Let $D_{L,n}$ and $D_{R,n}$ denote the IMU time series data collected from the left-hand and the right-hand armbands of worker $n$, respectively. In the remainder of the paper, the index of armbands is denoted by $m$, for $m \in \{L, R\}$.

### Data preparation

A sliding window technique was used to extract meaningful segments of time series data from the long time series of signals captured by armbands. In this paper, the extracted segments are named action-level signal images (ASI) that are used to train and test CNN AR models. ASIs are in the size of $50\Delta t \times 13 \times 1$, and the three dimensions are width, height, and color channel of ASIs, respectively. 50 is the sampling frequency of the Myo armbands, and $\Delta t$ denotes the time span of the sliding window in second (e.g. when $\Delta t = 2$, the width of ASIs is 100 frames). 13 is the number of signal series captured from each armband, and 1 means ASIs have only one color channel just like grayscale images. The ASIs obtained from the armband $m$ were divided into the training dataset, $S_m^{tr}$, and the testing dataset, $S_m^{ts}$, which are mutually exclusive.

The window length in second, $\Delta t$, needs to be appropriately selected to be able to capture distinct features of individual actions. Successive windows are overlapped to better handle the transition from one action to another.[25] The selection of these two parameters are largely dependent of the actions to be studied.

### CNN AR models and the model fusion

Two independent CNN models, $M_L$ and $M_R$, were trained with the training datasets $S_L^{tr}$ and $S_R^{tr}$, respectively. Figure 2 illustrates the proposed CNN architecture. For each ASI fed to the CNN, the first two convolutional layers filter it with kernels of size

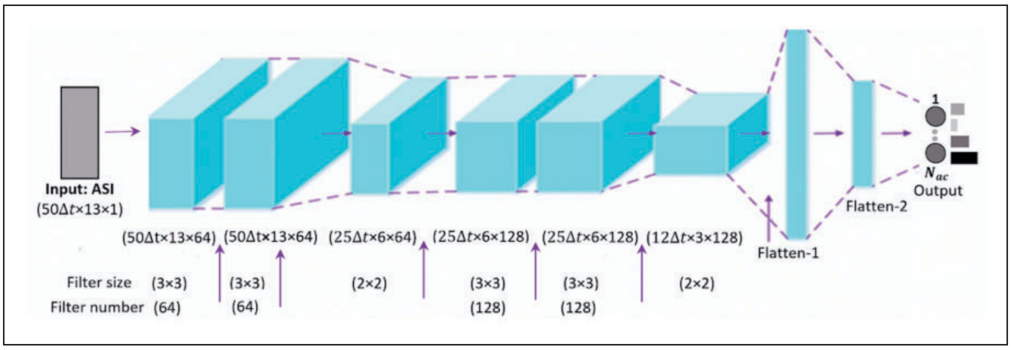


**Figure 2.** The proposed CNN architecture.

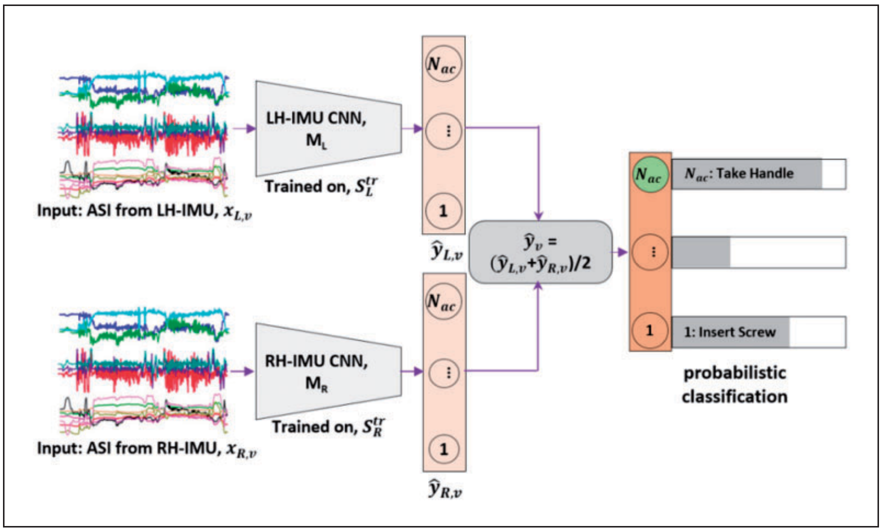


**Figure 3.** Schematic diagram of the fusion of right-hand and left-hand AR models.

$3 \times 3$ and then the $2 \times 2$ downsampling is performed by the max pooling layer. A feature map is generated using the ReLU function from each of these three layers. This procedure is repeated one more time. Then, the obtained feature map is flattened, densified, and converted into a feature vector of size $N_{ac}$. This feature vector is converted into a probability distribution on the $N_{ac}$ actions, becoming a probabilistic classification of the input ASI. The proposed CNN models have a relatively shallow architecture, which is suitable for analyzing the signal data of this study and effectively mitigates the issue of vanishing gradient. Interested readers may refer to Krizhevsky et al.[29] for details of CNN.

To train the CNN models, Adam, an adaptive learning rate optimizer coupled with the cross entropy cost function was used. The learning rate of Adam dynamically decreases with the iterations. The study used the default setting of the algorithm.[30] The L2 and dropout regularization techniques were also used to avoid the issue of overfitting.

Let $\boldsymbol{x}_{m,v}$ be an ASI that is indexed by $v$ and fed to the CNN model $M_m$. Then, $y_{m,v}$ and $\hat{\boldsymbol{y}}_{m,v}$ denote the ground truth and the probabilistic classification of $\boldsymbol{x}_{m,v}$, respectively. Figure 3 illustrates the approach to model fusion. An action lasting $\Delta t$ is simultaneously captured as two ASIs, $\boldsymbol{x}_{L,v}$ and $\boldsymbol{x}_{R,v}$. Then, the LH-IMU and RH-IMU models provide two independent probabilistic classifications, $\hat{\boldsymbol{y}}_{L,v}$ and $\hat{\boldsymbol{y}}_{R,v}$, respectively. This study averages the two classification results as the final probabilistic prediction of the action

$$\hat{y}_v = (\hat{\boldsymbol{y}}_{L,v} + \hat{\boldsymbol{y}}_{R,v})/2 \tag{1}$$

### Transfer learning for model calibration

The models $M_L$ and $M_R$ may also be used to recognize the actions of inference subjects who are workers not in the group of subjects sampled for training the models. When this happens, the AR models may not perform well, probably because of worker heterogeneity. This study used transfer learning[31] to effectively adapt the trained AR models to new workers. Specifically, a small set of ASIs collected from new workers were used to fine-tune the trained AR models by training only the last few layers of the CNNs to capture the unique features of new workers. Let $\tilde{\boldsymbol{y}}_{m,v}$ be the probabilistic classification made by the calibrated model $m$, then the fusion of calibrated models yields the prediction, $(\tilde{y}_{L,v} + \tilde{y}_{R,v})/2$.

## Experimental studies

To illustrate the proposed methodology of AR in assembly and assess the performance of it, a workstation for assembling Bukito 3D printers was set up in a lab, as shown in Figure 4. Material and tools to be used in the assembly were set on the workstation. Following the method delineated in the subsection "Data preparation", time series sensor data of workers were obtained. IMU data collected from the two armbands were transmitted to two separate computers via the Bluetooth units of the armbands. Then, ASIs were extracted from the time series of signals using a sliding window that can cover 2 s of data. Any two successive ASIs have a 50% overlap. Based on numerical experiments, the window size of 2 s and the 50% overlap were found to be an appropriate setting of the sliding window technique for extracting ASIs in this study.

### Seven assembly actions

The study used one step in assembling Bukito 3D printers, named "putting on the handle" in the assembly manual, as an example. This step of assembly consists of seven actions, as described in Figure 5. From the description of actions, it can be seen that the worker dominantly uses his left hand to perform action-1 and his right hand to perform action 2. The remaining actions require the collaboration of two hands, but roles of the two hands vary among these actions. For example, in action-4 the worker uses his left hand to hold the handle and his right hand to rotate the screw to manually tighten it. That is, the left hand mainly facilitates the right hand that dominantly performs the assembly. Yet in action-5, the left hand of the worker grabs the Allen key set and uses the right hand to pick the desired tool from the set. That is, the two hands have near equal importance in this action for preparing the next action. This assembly step contains a variety of actions with either a single hand or the collaboration of two hands, thus being a good example for illustrating and evaluating the effectiveness of model fusion.

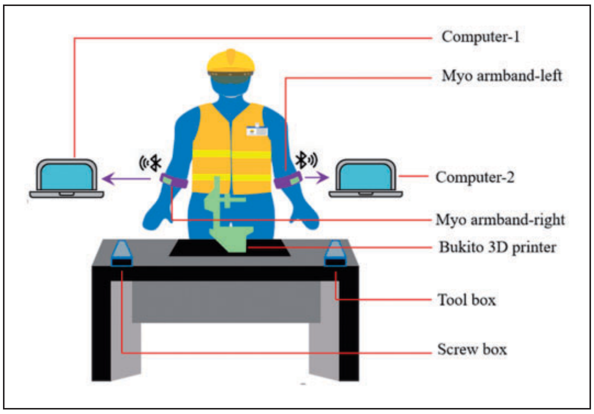


**Figure 4.** Experimental setup.

### The group size of subjects

While the smart manufacturing community is using new technologies such as IoT and wearable devices to study human, there is no publicly available large benchmark datasets of assembly actions. Publicly available datasets on daily living activities, which have been used for the purpose of CNN-based AR, mostly include 4–14 subjects.[26] Therefore, this study invited a group of 11 volunteers to participate in the experiments to capture between-subject variability (i.e. worker heterogeneity). To count the within-subject variability (the randomness of human actions), the 11 subjects repeated the assembly step for 10 times. This group is called the 11-subject group in the remainder of this paper. A subset of the 11-subject group, which is composed of only five subjects and named the 5-subject group, was defined as well. The study used these two groups of subjects to demonstrate the impact of worker heterogeneity to both model development and model implementation.

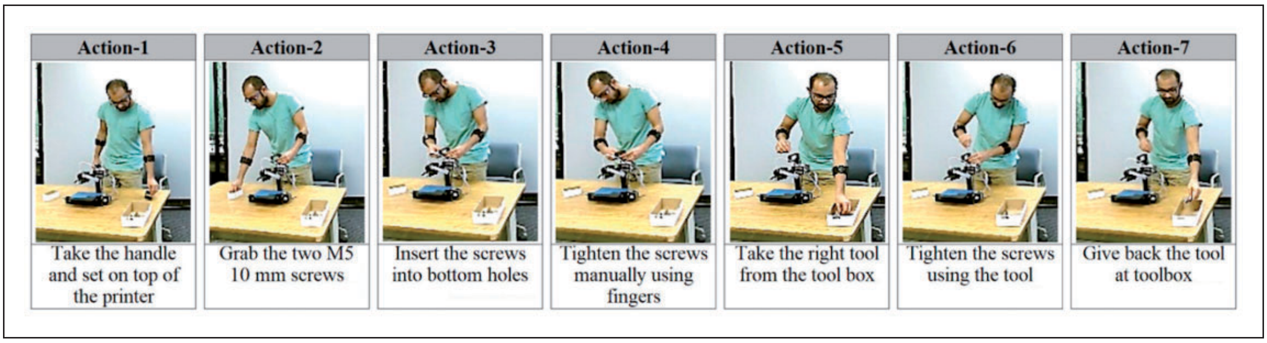

Action-1
Take the handle and set on top of the printer
Action-2
Grab the two M5 10 mm screws
Action-3
Insert the screws into bottom holes
Action-4
Tighten the screws manually using fingers
Action-5
Take the right tool from the tool box
Action-6
Tighten the screws using the tool
Action-7
Give back the tool at toolbox

**Figure 5.** The seven assembly actions.

### Dataset details

The assembly step in this study lasts for approximately 120 s. For each subject, nearly 60,000 samples (10 repetitions × 120 s × 50 Hz) of IMU data were acquired. Therefore, the sample size for the 5-subject group is 300,000, and sample size for the 11-subject group is 660,000. From these samples of IMU data, 2230 ASIs were created for the 5-subject group and 4934 ASIs for the 11-subject group. The distributions of ASIs across actions are further summarized in Table 1.

The study used two cross-validation methods: training-testing split (TTS) and leave one out (LOO). In the implementation of TTS, 80% of ASIs were used for training and 20% for testing. When implementing LOO for the 5-subject dataset, about 80% of images were used for training and 20% for testing. For the 11-subject dataset, about 91% of images were for training and 9% for testing. Table 1 further summarizes the sizes of training dataset and testing dataset, respectively, in details.

### Evaluation methods

This study primarily used the TTS method to evaluate the effectiveness of model fusion. Then, it further used the LOO method to assess the effectiveness of model calibration. Figure 6 illustrates how the collected data from a group of subjects were split into a training dataset and a testing dataset. When the TTS method was used, 80% of the data from each subject were randomly sampled to train the AR models and the remaining 20% of the data were used for testing the models. The TTS evaluation was repeated 15 times to construct an interval estimate for the classification accuracy.

**Table 1.** Sample size (# signal images) of action classes.

| | 5-subject group | 11-subject group |
|---|---|---|
| Action-1 | 136 | 286 |
| Action-2 | 164 | 376 |
| Action-3 | 268 | 606 |
| Action-4 | 362 | 788 |
| Action-5 | 278 | 666 |
| Action-6 | 740 | 1560 |
| Action-7 | 282 | 652 |
| Total | 2230 | 4934 |
| TTS | | |
| Training | 1784 | 3948 |
| Testing | 446 | 986 |
| LOO | | |
| Training | 1970–1646 | 4674–4350 |
| Testing | 260–584 | 260–584 |

TTS: training-testing split; LOO: leave one out.

In use of the LOO method, the data of $N_{ac} - 1$ subjects were used for training the AR models and the data of the remaining one subject were used for testing. The LOO evaluation was performed $N_{ac}$ times and each time a different worker from the group was tested.

To calibrate the AR models and make them adapt to the subject of testing in LOO, 20% of the data from that subject were used to fine-tune the models, and the remaining 80% of the data were used to test the performance of the fine-tuned models.

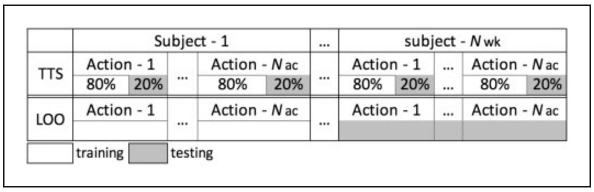
## Result analysis

The study assessed the fusion of LH-IMU and RH-IMU AR models using the TTS method, which is applicable to the situation that the group of workers does not change from the model development to implementation. Then, the assessment using the LOO method was further performed to evaluate the effectiveness of model calibration for the scenario where the AR models are applied to new workers whose data originally were not included in the dataset for training the AR models.

### Effectiveness of model fusion (assessed using the TTS method)

*The overall effectiveness.* Figure 7 compares the 90% interval estimates of the prediction accuracy of the LH-IMU model, the RH-IMU model, and the fusion of the two models. It can be seen that the RH-IMU model performs better than the LH-IMU model, with a 3.5% increase for the 5-subject group, and a 5.6% increase for the 11-subject group. Fusion of the two models further adds 4.6% and 4.9% increases for the two groups, respectively. To gain more insights into the observation, the study further assessed the model performance at the action level.

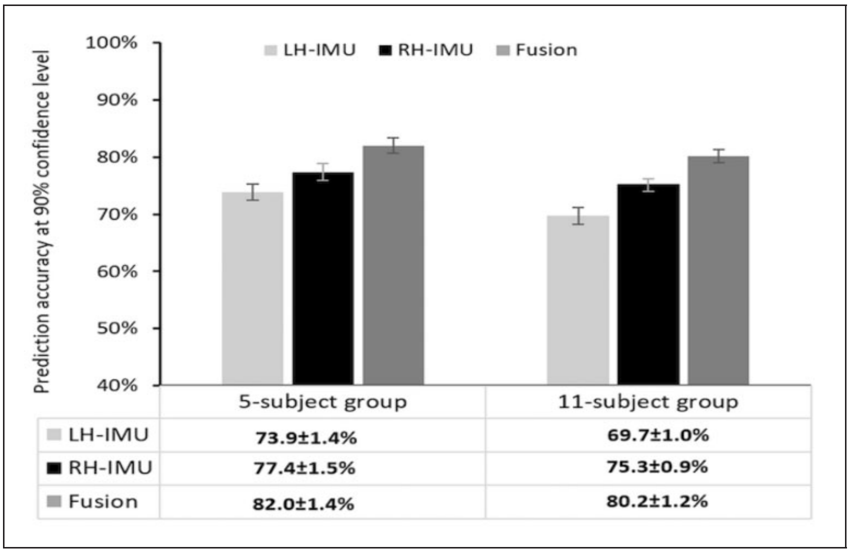
*At the action level.* Using the 11-subject group as an example, Figure 8 compares the interval estimates of the prediction accuracy of LH-IMU, RH-IMU, and the fusion of them in predicting each of the seven actions. The RH-IMU model performs better than the LH-IMU model in recognizing 6 out of 7 actions (except for action-1). This is because the right hand plays either a dominant role or an indispensable role in almost all actions (except for action-1), whereas the



| | Subject - 1 | | | ... | subject - $N_{wk}$ | | |
|---|---|---|---|---|---|---|---|
| TTS | Action - 1<br>80% 20% | ... | Action - $N_{ac}$<br>80% 20% | ... | Action - 1<br>80% 20% | ...<br>... | Action - $N_{ac}$<br>80% 20% |
| LOO | Action - 1 | ... | Action - $N_{ac}$ | ... | Action - 1 | ... | Action - $N_{ac}$ |

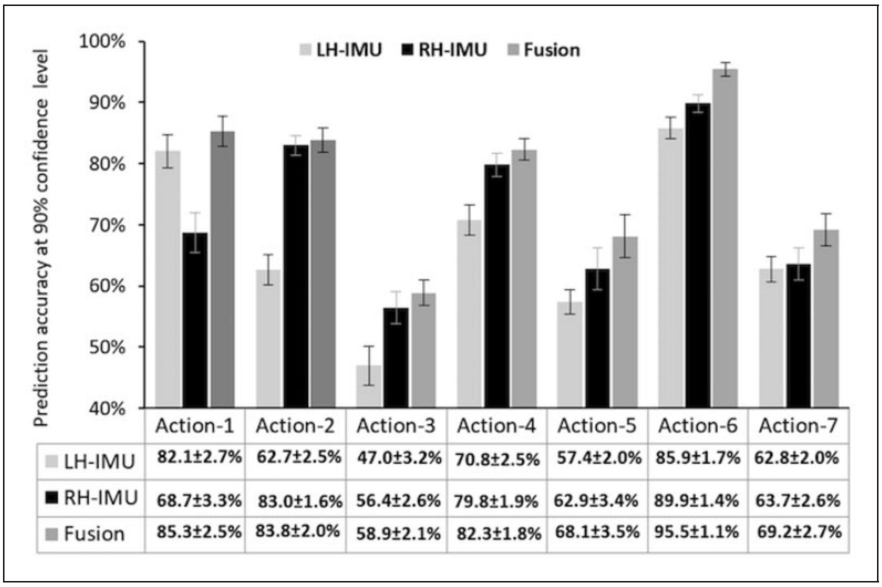
□ training ▨ testing

**Figure 6.** TTS and LOO evaluation methods.



| | 5-subject group | 11-subject group |
|---|---|---|
| LH-IMU | 73.9±1.4% | 69.7±1.0% |
| RH-IMU | 77.4±1.5% | 75.3±0.9% |
| Fusion | 82.0±1.4% | 80.2±1.2% |

**Figure 7.** The overall effectiveness of model fusion (TTS evaluation).



| | Action-1 | Action-2 | Action-3 | Action-4 | Action-5 | Action-6 | Action-7 |
|---|---|---|---|---|---|---|---|
| LH-IMU | 82.1±2.7% | 62.7±2.5% | 47.0±3.2% | 70.8±2.5% | 57.4±2.0% | 85.9±1.7% | 62.8±2.0% |
| RH-IMU | 68.7±3.3% | 83.0±1.6% | 56.4±2.6% | 79.8±1.9% | 62.9±3.4% | 89.9±1.4% | 63.7±2.6% |
| Fusion | 85.3±2.5% | 83.8±2.0% | 58.9±2.1% | 82.3±1.8% | 68.1±3.5% | 95.5±1.1% | 69.2±2.7% |

**Figure 8.** The effectiveness of AR model fusion (by actions) for the 11-subject group.

left hand has such importance in less number of actions (action-5, -6, and -7). Therefore, the RH-IMU model provides more useful features for AR than does the LH-IMU model.

The difference between the average prediction accuracy of the RH-IMU model and that of the LH-IMU model is relatively large in predicting action-1 and -2, moderate in predicting action-3 and -4, and small in action-5, -6, and -7. That is, the seven actions can be grouped as three clusters. This observation is supported by the relationship of the two hands in assembly, explained in the following.

- In performing action-1 and -2, workers mainly used one hand while the other hand was idle or near-idle.
- In performing action-3 and -4, workers dominantly used one hand but required a little bit of help from the other hand.
- In performing the remaining three actions, two hands either closely collaborated with each other (such as action-6) or the two hands respectively had relatively independent operations (such as action-5 and -7).

Figure 8 shows that the model fusion increases the accuracy in predicting action-5, -6, and -7 than the best individual AR mode for them at a 90% confidence level. Therefore, the model fusion particularly improves the prediction accuracy on actions that require a collaboration of two hands with near equal importance (i.e. no hand is idle or just facilitates the other hand). Figure 8 further shows that the model fusion has higher average accuracy than both AR models in recognizing any of the seven actions although the increase may not be at the 90% confidence level. As it can provide consistently better performance than any individual model in recognizing any of the seven actions, the model fusion is more reliable than individual AR models in assembly AR.

*Impact of worker heterogeneity to AR.* From Figure 7 it can be observed that the two AR models and the fusion of them all have a higher average prediction accuracy, but a wider 90% confidence interval, in recognizing the actions of the 5-subject group than of the 11-subject group. This observation is associated with the fact that the larger size group usually has greater worker heterogeneity than does the smaller size group. In this study, the 11 subjects provided more ASIs for model testing than did 5 subjects and thus 90% confidence intervals of the prediction accuracy for the 11-subject groups is narrower. Since large worker heterogeneity would deteriorate the performance of AR, it would be helpful to develop AR models for smaller groups of workers with similarity than to create a model for any workers.

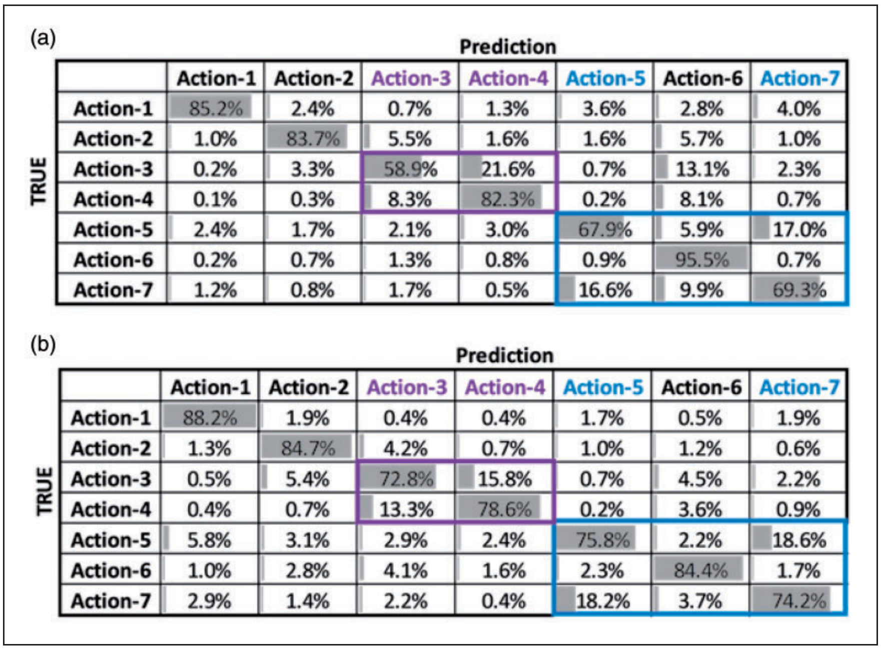
*Impact of action similarity to AR.* Some actions that workers take in assembly may share some similarity, making it difficult to correctly classify them. Using the 11-subject group as an example, the study computed the recall matrix and precision matrix for the classification result from model fusion, shown in Figure 9, to verify that action similarity is a reason for misclassification.

- Recall: the ratio of correctly predicted classes to the total number of ground-truths.
- Precision: the ratio of correctly predicted classes to the total number of predictions.

In both matrices, rows are true actions and columns are predicted actions. Therefore, the element at the intersection of row *i* and column *j* in the recall matrix counts the percentage of true action *i* recognized as action *j*. In the precision matrix this element computes the percentage of instances classified as action *j* are action *i*. Diagonal elements in the two matrices represent correct classifications.

From both the recall and precision matrices, it can be seen that the recognition performance for action-3 was the lowest. A significant portion of action-3 (21.6% in the recall matrix and 15.8% in the precision matrix) is misclassified as action-4, and vice-versa. This is because the left hand holds the handle in both actions, and the right hand inserts screws into holes in action-3 and manually tightens screws using fingers in action-4. That is, these actions involve with the same left hand activity and slightly different right hand activities. As a result, the misclassification happens at a relatively higher rate in recognizing these two actions.

It has also been found that a significant portion of action-5 (17.0% and 18.6% in the recall and precision matrices, respectively) is misclassified as action-7, and vice versa. This is due to the high similarity between these two actions. Action-5 is pulling out the desired



(a)

Prediction

| TRUE | Action-1 | Action-2 | Action-3 | Action-4 | Action-5 | Action-6 | Action-7 |
|---|---|---|---|---|---|---|---|
| Action-1 | 85.2% | 2.4% | 0.7% | 1.3% | 3.6% | 2.8% | 4.0% |
| Action-2 | 1.0% | 83.7% | 5.5% | 1.6% | 1.6% | 5.7% | 1.0% |
| Action-3 | 0.2% | 3.3% | 58.9% | 21.6% | 0.7% | 13.1% | 2.3% |
| Action-4 | 0.1% | 0.3% | 8.3% | 82.3% | 0.2% | 8.1% | 0.7% |
| Action-5 | 2.4% | 1.7% | 2.1% | 3.0% | 67.9% | 5.9% | 17.0% |
| Action-6 | 0.2% | 0.7% | 1.3% | 0.8% | 0.9% | 95.5% | 0.7% |
| Action-7 | 1.2% | 0.8% | 1.7% | 0.5% | 16.6% | 9.9% | 69.3% |

(b)

Prediction

| TRUE | Action-1 | Action-2 | Action-3 | Action-4 | Action-5 | Action-6 | Action-7 |
|---|---|---|---|---|---|---|---|
| Action-1 | 88.2% | 1.9% | 0.4% | 0.4% | 1.7% | 0.5% | 1.9% |
| Action-2 | 1.3% | 84.7% | 4.2% | 0.7% | 1.0% | 1.2% | 0.6% |
| Action-3 | 0.5% | 5.4% | 72.8% | 15.8% | 0.7% | 4.5% | 2.2% |
| Action-4 | 0.4% | 0.7% | 13.3% | 78.6% | 0.2% | 3.6% | 0.9% |
| Action-5 | 5.8% | 3.1% | 2.9% | 2.4% | 75.8% | 2.2% | 18.6% |
| Action-6 | 1.0% | 2.8% | 4.1% | 1.6% | 2.3% | 84.4% | 1.7% |
| Action-7 | 2.9% | 1.4% | 2.2% | 0.4% | 18.2% | 3.7% | 74.2% |

**Figure 9.** Classification of actions for the 11-subject group using action fusion: (a) recall matrix and (b) precision matrix.

tool from the tool set and action-7 is the opposite to action-5.

Action-6 has been found to have the highest recognition accuracy (95.5% recall). An important contributor to this high accuracy is that action-6 takes the longest time to complete and thus had the largest number of ASIs in this experimental study. Therefore, the AR models learn the pattern of this action very well, thus subsequently improving the recognition accuracy of this action. It has been also observed that many actions were misclassified as action-6. This is because this action covers almost all the motions involved with other actions of this assembly step, for example, finger movement, wrist movement, and holding the handle.

## *Effectiveness of model calibration (assessed using the LOO method)*

Worker heterogeneity is also a challenging issue in the model implementation. When inference subjects are different than the subjects sampled for training models, the prediction ability of the AR models may be less satisfied due to worker heterogeneity. This study used transfer learning to address this issue and illustrated the effectiveness of this approach using the LOO cross-validation. As described in the subsection "Evaluation methods", to use the AR models to recognize actions of a new worker who was not included in the group of subjects for model training, a calibration was performed: the new worker was asked to perform the assembly step twice and this small set of data was used to fine-tune the AR models by only training the two Flatten layers and the Dense layer of the CNN in Figure 2; that is, all the layers before the Flatten-1 layer were frozen during the model refinement.

*Overall effectiveness of model calibration.* Figure 10 compares the interval estimates of the prediction accuracy of LH-RMU, RH-RMU, and the fusion of them across the following three cases. The comparison based on the 5-subject group is presented in Figure 10(a) and 10(b) is the 11-subject group.

- Case (i), 80% Train – 20% Test: inference subjects are the same as the subjects sampled for training the AR models;
- Case (ii), LOO-Be. cali: inference subjects are different than the subjects sampled for training the AR models and no model calibration is made;
- Case (iii), LOO-Af. cali: inference subjects are different than the subjects sampled for training the AR models and model calibration is made.

Figure 10(a) shows the effectiveness of model calibration and fusion in improving the recognition accuracy for 5-subject groups in LOO evaluation. For the LH-IMU model, the model calibration increases the prediction accuracy from 59.2% to 71.0%, and with the model fusion the accuracy reaches 79.1%, resulting an overall improvement of 19.9% (=79.1–59.2%). Similarly, for the RH-IMU model, the implementation of aforesaid two techniques yields an improvement of 13.5% (=79.1–65.6%). Figure 10(b) shows similar improvements can also be achieved for the 11-subject group. An improvement of 18.6% (=77.1–58.5%) is achieved for the LH-IMU model and 9.0% (=77.1–68.1%) for the RH-IMU model. In short, this study broke down the overall improvement into two elements and allocated them to the two contributors: model calibration and model fusion, which are summarized in Table 2 and discussed in additional details in the following.

From Figure 10(a) it can be seen that the prediction accuracy of individual AR models in case (ii) is at least 11.8% lower than that in case (i) for the 5-subject group. This demonstrates that worker heterogeneity, reflected by the change in subjects from model training to implementation, is an issue affecting the AR accuracy. The model calibration effectively increases the prediction accuracy of LH-IMU by 11.8% and 7.8% for RH-IMU. Although the prediction accuracy of individual AR models in case (iii) is still lower than that in case (i), the gap is no more than 4.0%, confirming that the model calibration can effectively lower the impact of worker heterogeneity at the implementation stage. Similar observations are in Figure 10(b), and the model calibration increases the prediction accuracy of LH-IMU by 8.1% and 6.1% for RH-IMU.

Figure 10 further shows that the fusion of calibrated AR models further improves the accuracy above the model calibration. For the 5-subject group, the model fusion achieves an accuracy of 79.1%, 8.1% higher than the accuracy of the calibrated LH-IMU model (71.0%) and 5.7% above that of the calibrated RH-IMU model (73.4%). For the 11-subject group, the model fusion yields an accuracy of 77.1%, 10.5% higher than the accuracy of the calibrated LH-IMU model (66.6%) and 2.9% above that of the calibrated RH-IMU model (74.2%).

*At the action level.* The developed AR models have varied ability in recognizing different actions. Therefore, the study further examined the effectiveness of model calibration and the fusion of calibrated models at the action level. Results for the 5-subject group and the 11-subject group are summarized in Tables 3 and 4, respectively.

The calibration of the LH-IMU model for the 5-subject group improves the accuracy in predicting almost every individual action except for action-2. But the reduced accuracy is only −0.3%, which can be ignored. The calibration of the RH-IMU model improves the accuracy for all actions. The fusion of

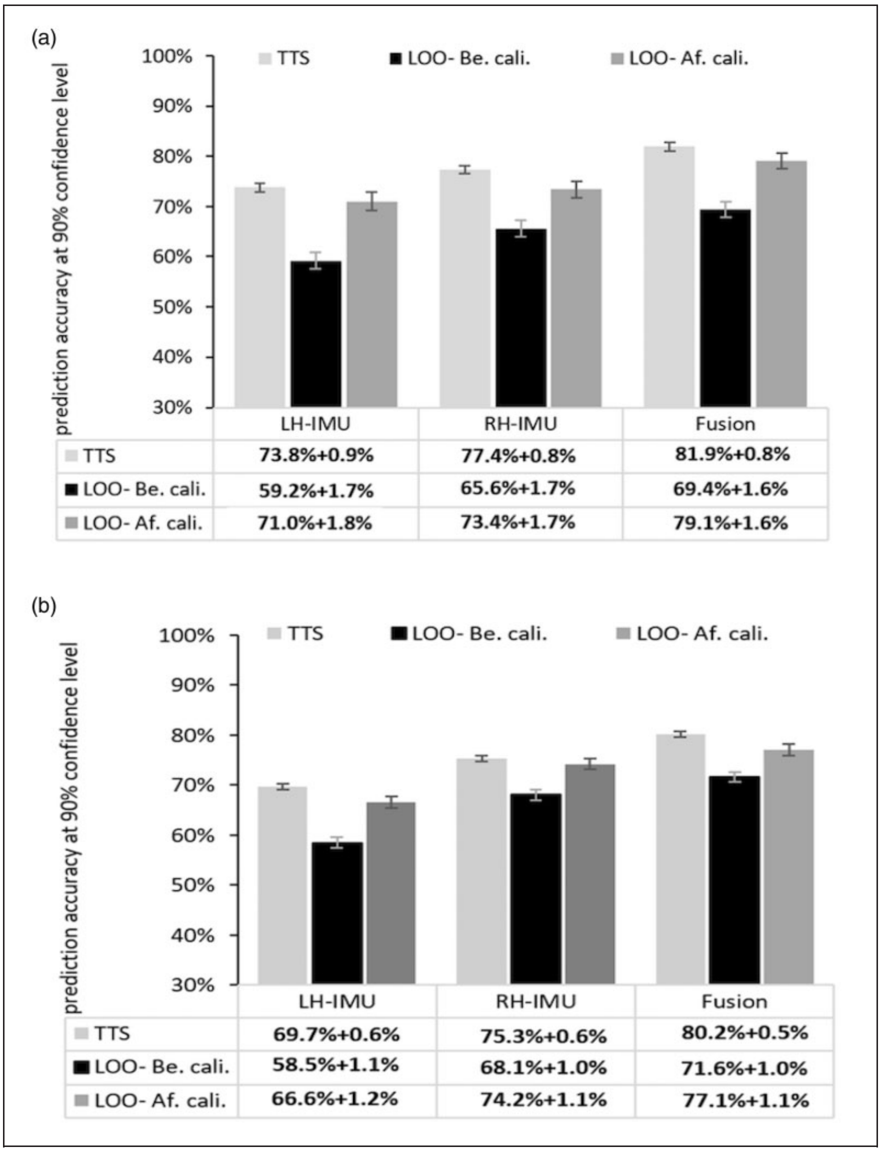


**Figure 10.** Overall improvement by model calibration: (a) 5-subject group, (b) 11-subject group.

**Table 2.** Improvements by calibration and fusion of AR models.

| | 5-subject group | | 11-subject group | |
|---|---|---|---|---|
| | LH-IMU | RH-IMU | LH-IMU | RH-IMU |
| Cal. | 11.8% | 7.8% | 8.1% | 6.1% |
| Fus. | 8.1% | 5.7% | 10.5% | 2.9% |
| Both | 19.9% | 13.5% | 18.6% | 9.0% |

LH-IMU: left-hand inertial measurement unit; RH-IMU: right-hand inertial measurement unit.

the calibrated models has better accuracy than the best individual model in predicting almost every action except for action-1. The reason that the fusion of models has lower accuracy than the calibrated LH-IMU model in predicting action-1 is due to the fact that a 50% weight was put on the RH-IMU model that was not very capable in predicting the left-hand dominated action-1 (44.4% accuracy after calibration).

On the 11-subject group, calibration of the LH-IMU model improves the accuracy in predicting every individual action. The calibration of the RH-IMU model also improves the accuracy in predicting every individual action. The fusion of calibrated models has a higher prediction accuracy than the best individual model in predicting 5 out of 7 actions (except for action-2 and -5).

The analysis above indicates that the model calibration and fusion might not work well for actions dominantly performed by a single hand such as actions-1 and -2. But statistically speaking, the calibration of AR models improves the accuracy in predicting individual actions at a 90% confidence level; the fusion of the calibrated models can improve the accuracy in predicting individual actions, yet it is not

**Table 3.** Prediction accuracy (%) at the action level: before vs after calibration (5-subject group).

| | Action-1 | Action-2 | Action-3 | Action-4 | Action-5 | Action-6 | Action-7 | Avg. | m.e. |
|---|---|---|---|---|---|---|---|---|---|
| LH-IMU | | | | | | | | | |
| Before | 55.1 | 61.6 | 49.6 | 54.7 | 38.8 | 75.4 | 52.5 | 55.4 | 8.2 |
| After | 80.6 | 61.3 | 55.7 | 66.7 | 58.8 | 88.7 | 57.3 | 67.0 | 9.4 |
| Impr | 25.5 | −0.3 | 6.1 | 12.0 | 20.0 | 13.3 | 4.8 | 11.6 | 6.6 |
| RH-IMU | | | | | | | | | |
| Before | 29.4 | 73.8 | 53.4 | 72.9 | 51.4 | 83.5 | 47.5 | 58.8 | 13.8 |
| After | 44.4 | 82.3 | 61.0 | 74.7 | 69.0 | 86.7 | 61.6 | 68.5 | 10.6 |
| Impr | 15.0 | 8.5 | 7.6 | 1.8 | 17.6 | 3.2 | 14.1 | 9.7 | 4.5 |
| Fusion | | | | | | | | | |
| Before | 45.6 | 80.5 | 57.1 | 69.3 | 51.8 | 87.7 | 55.6 | 63.9 | 11.5 |
| After | 71.3 | 86.3 | 63.3 | 79.0 | 69.0 | 93.6 | 65.9 | 75.5 | 8.3 |
| Impr | −9.3 | 4.0 | 2.3 | 4.3 | 0.0 | 4.9 | 4.3 | 1.5 | 3.7 |

LH-IMU: left-hand inertial measurement unit; RH-IMU: right-hand inertial measurement unit.

**Table 4.** Prediction accuracy (%) at the action level: before vs after calibration (11-subject group).

| | Action-1 | Action-2 | Action-3 | Action-4 | Action-5 | Action-6 | Action-7 | Avg. | m.e. |
|---|---|---|---|---|---|---|---|---|---|
| LH-IMU | | | | | | | | | |
| Before | 63.3 | 41.5 | 28.7 | 46.8 | 51.8 | 78.5 | 42.4 | 50.4 | 11.9 |
| After | 67.8 | 60.0 | 43.7 | 68.4 | 55.1 | 85.1 | 56.5 | 62.4 | 9.6 |
| Impr | 4.5 | 18.5 | 15.0 | 21.6 | 3.3 | 6.6 | 14.1 | 11.9 | 5.3 |
| RH-IMU | | | | | | | | | |
| Before | 54.5 | 78.7 | 48.8 | 68.1 | 58.9 | 84.4 | 50.0 | 63.4 | 10.3 |
| After | 57.0 | 81.0 | 60.0 | 77.8 | 64.7 | 88.2 | 62.8 | 70.2 | 8.8 |
| Impr | 2.5 | 2.3 | 11.2 | 9.7 | 5.8 | 3.8 | 12.8 | 6.9 | 3.2 |
| Fusion | | | | | | | | | |
| Before | 72.4 | 75.0 | 42.2 | 69.9 | 61.7 | 92.1 | 49.2 | 66.1 | 12.3 |
| After | 74.3 | 78.3 | 60.8 | 80.6 | 63.8 | 93.2 | 63.8 | 73.5 | 8.5 |
| Impr | 6.5 | −2.7 | 0.8 | 2.8 | −0.9 | 5.0 | 1.0 | 1.8 | 2.4 |

LH-IMU: left-hand inertial measurement unit; RH-IMU: right-hand inertial measurement unit.

at the 90% confidence level. The average improvement by the calibration of LH-IMU is the largest, followed by the calibration of RH-IMU. The improvement by the fusion of the calibrated models is smaller than the model calibration.

*At the subject level.* The study also examined the effectiveness of model calibration and fusion at the subject level. Results for the 5-subject group and the 11-subject group are summarized in Tables 5 and 6, respectively. For both groups, the model calibration improved the prediction accuracy for all subjects. The average improvement for the LH-IMU is higher than that for the RH-IMU model.

The fusion of calibrated models has a higher accuracy than the best calibrated model for all subjects in the 5-subject group and for most subjects in the 11-subject group (except for subjects-4, -6, and -7). Yet statistically speaking, the fusion of calibrated

**Table 5.** Prediction accuracy (%) at the subject level: before vs after calibration (5-subject group).

| | Sub-1 | Sub-2 | Sub-3 | Sub-4 | Sub-5 | Avg. | m.e. |
|---|---|---|---|---|---|---|---|
| LH-IMU | | | | | | | |
| Before | 61.9 | 63.2 | 69.6 | 53.9 | 54.6 | 60.6 | 6.2 |
| After | 74.5 | 70.5 | 83.0 | 70.7 | 64.0 | 72.5 | 6.6 |
| Impr | 12.6 | 7.3 | 13.4 | 16.8 | 9.4 | 11.9 | 3.5 |
| RH-IMU | | | | | | | |
| Before | 69.8 | 66.4 | 78.5 | 61.0 | 60.6 | 67.3 | 7.0 |
| After | 74.8 | 71.6 | 89.3 | 70.9 | 68.9 | 75.1 | 7.8 |
| Impr | 5.0 | 5.2 | 10.8 | 9.9 | 8.3 | 7.8 | 2.5 |
| Fusion | | | | | | | |
| Before | 73.0 | 72.2 | 84.2 | 61.7 | 65.2 | 71.3 | 8.2 |
| After | 81.1 | 77.0 | 90.8 | 78.8 | 74.7 | 80.5 | 5.9 |
| Impr | 6.3 | 5.4 | 1.5 | 7.9 | 5.8 | 5.4 | 2.3 |

LH-IMU: left-hand inertial measurement unit; RH-IMU: right-hand inertial measurement unit.

**Table 6.** Prediction accuracy (%) at the subject level: before vs. after calibration (11-subject group).

| | Sub-1 | Sub-2 | Sub-3 | Sub-4 | Sub-5 | Sub-6 | Sub-7 | Sub-8 | Sub-9 | Sub-10 | Sub-11 | Avg. | m.e. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| LH-IMU | | | | | | | | | | | | | |
| Before | 53.7 | 62.3 | 57.6 | 61.5 | 35.1 | 46.6 | 55.9 | 61.7 | 54.5 | 66.2 | 51.5 | 55.1 | 4.7 |
| After | 73.0 | 70.8 | 64.9 | 75.2 | 63.4 | 65.6 | 59.7 | 73.7 | 61.9 | 67.8 | 62.0 | 67.1 | 2.9 |
| Impr | 19.3 | 8.5 | 7.3 | 13.7 | 28.3 | 19.0 | 3.8 | 12.0 | 7.4 | 1.6 | 10.5 | 11.9 | 4.2 |
| RH-IMU | | | | | | | | | | | | | |
| Before | 63.4 | 72.4 | 58.0 | 87.3 | 64.4 | 76.8 | 46.8 | 65.2 | 72.6 | 72.7 | 65.8 | 67.8 | 5.8 |
| After | 75.8 | 78.1 | 77.9 | 93.2 | 67.0 | 82.1 | 51.5 | 74.4 | 78.1 | 78.1 | 70.9 | 75.2 | 5.6 |
| Impr | 12.4 | 5.7 | 19.9 | 5.9 | 2.6 | 5.3 | 4.7 | 9.2 | 5.5 | 5.4 | 5.1 | 7.4 | 2.7 |
| Fusion | | | | | | | | | | | | | |
| Before | 78.5 | 74.4 | 66.0 | 89.2 | 62.0 | 66.2 | 56.7 | 74.2 | 76.7 | 77.3 | 57.7 | 70.8 | 5.4 |
| After | 80.4 | 81.5 | 81.1 | 91.3 | 71.1 | 78.5 | 57.2 | 78.8 | 81.9 | 82.4 | 72.3 | 77.9 | 4.7 |
| Impr | 4.6 | 3.4 | 3.2 | −1.9 | 4.1 | −3.6 | −2.5 | 4.4 | 3.8 | 4.3 | 1.4 | 1.9 | 1.7 |

LH-IMU: left-hand inertial measurement unit; RH-IMU: right-hand inertial measurement unit.

models improves the prediction accuracy for individual subjects at a 90% confidence level. The average improvement by the fusion of calibrated models in the larger group is 1.9%, less than the average improvement of 5.4% in the smaller group. This might be an evidence of the challenge facing a larger group of subjects. The fusion of the calibrated models makes a smaller contribution to the prediction accuracy than does the calibration of individual models.

## Conclusions and future work

This paper proposed a method that uses two devices worn on the two hands of workers to collect IMU data when they are assembling products. Two CNN models for AR, a left-hand model, and a right-hand model, were independently developed using the respective two sources of IMU data. The two CNN models were fused together to yield a better AR result. Transfer learning was applied to refine the AR models and make them adapt to new workers who were not in the original group of subjects sampled for training the AR models. A small-scale assembly of Bukito 3D printers was performed in a lab setting to illustrate the implementation and assessment of the proposed method. Improvements made by the model fusion and refining have been consistently achieved in recognizing various actions of different assembly workers.

This paper has both methodological contributions and insightful findings. Therefore, it builds a sound foundation for recognition of assembly actions. Firstly, the method proposed in this paper is kept in general. Manufacturers can easily follow the proposed approach to build the capability of AR in their own assembly lines. Secondly, the proposed method is focused on recognizing composite actions that each is clearly driven by an intermediate objective and sequentially meeting these objectives will achieve certain progress of assembly (such as one step of assembly). With the AR capability at such a detailed level, manufacturers are able to diagnose potential issues that their assembly workers have in real time and quickly address the issues with appropriate methods such as on-the-job training or personalized assistance. Thirdly, the paper has demonstrated that the fusion of two independent IMU-based AR models largely improves the prediction accuracy of actions that require an active collaboration of two hands. This provides a guidance for designing and deploying multiple sensors to capture human actions involved in the collaboration of multiple body parts. Last but not the least, this paper has demonstrated the impact of worker heterogeneity to both the model development and implementation. To address this issue in the model development stage, the paper suggests to train customized AR models for each individual group of subjects who share similarity. The paper has shown the effectiveness of using transfer learning to refine AR models to make them adapt to subjects different than those that the models were trained from.

Although the refining and fusion of only two IMU-based deep learning models have been found to be effective in creating a good capability of recognizing assembly actions, the room and opportunities for improvement exist. First of all, actions that take a short period of time to finish usually have less samples of ASI and cannot be reliably recognized. Oversampling is a way to address this uneven distribution of samples across actions. Moreover, IMU sensors may be outperformed by other sensors in recognizing some actions. For example, IMU may not have the right level of accuracy to detect and characterize subtle motions but sEMG sensors can achieve that by providing signals of muscle activities. The fusion of multiple types of sensors is another dimension of performance improvement that can be built above this paper. Time coherence is also an approach to improving the accuracy and precision of assembly AR. Assembly actions are taken in sequence with each lasting for a period of time. Incorporating the time coherence information of actions in assembly AR would help reduce the number of misclassifications. The number of subjects in the training dataset can also be increased to incorporate more human heterogeneity so that the model is more robust. The work can also be scaled up by increasing the number of actions, which can better facilitate AR in a wider range of operations in assembly. This paper had built a sound foundation for exploring these opportunities of improvement.

## Note to practitioners

The advancement of automation and smart technologies has shifted the role of manufacturing workers from tedious, time-consuming, and risky operations to knowledge-intensive tasks. For example, assembly is an important stage of manufacturing that workers have many complex actions to take. Therefore, manufacturers are interested in providing real-time assistance or on-the-job training to assure their workers have reliable and high performance in such operations and can effectively collaborate with machines, robots, and smart technologies. Difficulties that workers may have in operations are partially reflected by what they are doing and how. This paper used wearable devices to collect sensor data of workers in operations and developed a deep learning based tool to process the data for recognizing assembly actions. This paper further improved the performance of AR by calibrating multiple solutions and then integrating them into one that has a better accuracy and can reliably recognize various assembly actions of different workers. The proposed method of assembly AR is kept in general. Manufacturers can easily follow the approach in this paper to build the capability of AR in their own assembly lines. The proposed method is able to recognize assembly actions that each has a clear intermediate objective and, therefore, manufacturers can analyze the recognized actions to better assess the performance of workers. Opportunities for improving the performance of the proposed method have been identified in this paper, such as using the sequential relationship of actions to reduce mistakes in assembly AR.

### Acknowledgment

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

### Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by NSF grant CMMI-1646162 on cyber-physical systems.

### ORCID iD

Ruwen Qin https://orcid.org/0000-0003-2656-8705

### References

1. Bartodziej CJ. The concept industry 4.0. In: *The concept industry 4.0*. New York: Springer, 2017, pp.27–50.
2. May G, Taisch M, Bettoni A, et al. A new human-centric factory model. *Procedia CIRP* 2015; 26: 103–108.
3. Turaga P, Chellappa R, Subrahmanian VS, et al. Machine recognition of human activities: a survey. *IEEE Trans Circuits Syst Video Technol* 2008; 18: 1473.
4. Chen C, Jafari R and Kehtarnavaz N. A survey of depth and inertial sensor fusion for human action recognition. *Multimedia Tool Appl* 2017; 76: 4405–4425.
5. Herath S, Harandi M and Porikli F. Going deeper into action recognition: a survey. *Image Vision Comput* 2017; 60: 4–21.
6. Ramanathan M, Yau WY and Teoh EK. Human action recognition with video data: research and evaluation challenges. *IEEE Trans Hum-Mach Syst* 2014; 44: 650–663.
7. Wang A, Chen G, Yang J, et al. A comparative study on human activity recognition using inertial sensors in a smartphone. *IEEE Sensor J* 2016; 16: 4566–4578.
8. Altun K and Barshan B. Human activity recognition using inertial/magnetic sensor units. In: *International workshop on human behavior understanding*. New York: Springer, 2010, pp.38–51.
9. Junker H, Amft O, Lukowicz P, et al. Gesture spotting with body-worn inertial sensors to detect user activities. *Pattern Recogn* 2008; 41: 2010–2024.
10. Attal F, Mohammed S, Dedabrishvili M, et al. Physical human activity recognition using wearable sensors. *Sensors* 2015; 15: 31314–31338.
11. Dernbach S, Das B, Krishnan NC, et al. Simple and complex activity recognition through smart phones. In: *2012 eighth international conference on intelligent environments*, 2012, pp.214–221. New York: IEEE.
12. Ngo TT, Makihara Y, Nagahara H, et al. Similar gait action recognition using an inertial sensor. *Pattern Recogn* 2015; 48: 1289–1301.
13. ElMaraghy H and ElMaraghy W. Smart adaptable assembly systems. *Procedia CIRP* 2016; 44: 4–13.
14. Ward JA, Lukowicz P, Troster G, et al. Activity recognition of assembly tasks using body-worn microphones and accelerometers. *IEEE Trans Pattern Anal Mach Intell* 2006; 28: 1553–1567.
15. Stiefmeier T, Roggen D, Ogris G, et al. Wearable activity tracking in car manufacturing. *IEEE Pervasive Comput* 2008; 2: 42–50.
16. Kaghyan S and Sarukhanyan H. Activity recognition using k-nearest neighbor algorithm on smartphone with tri-axial accelerometer. *Int J Inform Models Anal* 2012; 1: 146–156.
17. Stiefmeier T, Ogris G, Junker H, et al. Combining motion sensors and ultrasonic hands tracking for continuous activity recognition in a maintenance scenario. In: *2006 10th IEEE international symposium on wearable computers*, 2006, pp.97–104. New York: IEEE.

18. Tao W, Lai ZH, Leu MC, et al. Worker activity recognition in smart manufacturing using IMU and SEMG signals with convolutional neural networks. *Procedia Manuf* 2018; 26: 1159–1166.
19. Gaglio S, Re GL and Morana M. Human activity recognition process using 3-d posture data. *IEEE Trans Hum-Mach Syst* 2015; 45: 586–597.
20. Guo M, Wang Z, Yang N, et al. A multisensor multiclassifier hierarchical fusion model based on entropy weight for human activity recognition using wearable inertial sensors. *IEEE Trans Hum-Mach Syst* 2019; 49: 105–111.
21. Banos O, Damas M, Pomares H, et al. Human activity recognition based on a sensor weighting hierarchical classifier. *Soft Comput* 2013; 17: 333–343.
22. Guo Y, He W and Gao C. Human activity recognition by fusing multiple sensor nodes in the wearable sensor systems. *J Mech Med Biol* 2012; 12: 1250084.
23. Zappi P, Stiefmeier T, Farella E, et al. Activity recognition from on-body sensors by classifier fusion: sensor scalability and robustness. In: *2007 3rd international conference on intelligent sensors, sensor networks and information*, 2007, pp.281–286. New York: IEEE.
24. Zhu C and Sheng W. Wearable sensor-based hand gesture and daily activity recognition for robot-assisted living. *IEEE Trans Syst Man Cybernet A* 2011; 41: 569–573.
25. Al-Amin M, Tao W, Doell D, et al. Action recognition in manufacturing assembly using multimodal sensor fusion. *Procedia Manuf* 2019; 39: 158–167.
26. Wang J, Chen Y, Hao S, et al. Deep learning for sensor-based activity recognition: a survey. *Pattern Recogn Lett* 2019; 119: 3–11.
27. Nweke HF, Teh YW, Al-Garadi MA, et al. Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: state of the art and research challenges. *Expert Syst Appl* 2018; 105: 233–261.
28. Myoband, https://developerblog.myo.com, Thalmic Labs.
29. Krizhevsky A, Sutskever I and Hinton GE. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 2012; pp. 1106–1114.
30. Kingma DP and Ba J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:14126980*, 2014.
31. Pan SJ and Yang Q. A survey on transfer learning. *IEEE Trans Knowl Data Eng* 2010; 22: 1345–1359.