01 Aug 2019

Action Recognition in Manufacturing Assembly using Multimodal Sensor Fusion

Md. Al-Amin

Wenjin Tao

David Doell

Ravon Lingard

et. al. For a complete list of authors, see https://scholarsmine.mst.edu/comsci_facwork/968

Recommended Citation

M. Al-Amin et al., "Action Recognition in Manufacturing Assembly using Multimodal Sensor Fusion," *Procedia Manufacturing*, vol. 39, pp. 158-167, Elsevier B.V., Aug 2019.
The definitive version is available at https://doi.org/10.1016/j.promfg.2020.01.288

Available online at www.sciencedirect.com

ScienceDirect

Procedia Manufacturing 39 (2019) 158–167

Procedia MANUFACTURING

www.elsevier.com/locate/procedia

25th International Conference on Production Research Manufacturing Innovation: Cyber Physical Manufacturing
August 9–14, 2019 | Chicago, Illinois (USA)

Action Recognition in Manufacturing Assembly using Multimodal Sensor Fusion

Md. Al-Amin[a], Wenjin Tao[b], David Doell[a], Ravon Lingard[a], Zhaozheng Yin[c], Ming C. Leu[b], Ruwen Qin[a,*]

[a]*Department of Engineering Management and Systems Engineering, Missouri University of Science and Technology, Rolla 65409, USA*
[b]*Department of Mechanical and Aerospace Engineering, Missouri University of Science and Technology, Rolla 65409, USA*
[c]*Computer Science Department, Missouri University of Science and Technology, Rolla 65409, USA*

Abstract

Production innovations are occurring faster than ever. Manufacturing workers thus need to frequently learn new methods and skills. In fast changing, largely uncertain production systems, manufacturers with the ability to comprehend workers' behavior and assess their operation performance in near real-time will achieve better performance than peers. Action recognition can serve this purpose. Despite that human action recognition has been an active field of study in machine learning, limited work has been done for recognizing worker actions in performing manufacturing tasks that involve complex, intricate operations. Using data captured by one sensor or a single type of sensor to recognize those actions lacks reliability. The limitation can be surpassed by sensor fusion at data, feature, and decision levels. This paper presents a study that developed a multimodal sensor system and used sensor fusion methods to enhance the reliability of action recognition. One step in assembling a Bukito 3D printer, which composed of a sequence of 7 actions, was used to illustrate and assess the proposed method. Two wearable sensors namely Myo-armband captured both Inertial Measurement Unit (IMU) and electromyography (EMG) signals of assembly workers. Microsoft Kinect, a vision based sensor, simultaneously tracked predefined skeleton joints of them. The collected IMU, EMG, and skeleton data were respectively used to train five individual Convolutional Neural Network (CNN) models. Then, various fusion methods were implemented to integrate the prediction results of independent models to yield the final prediction. Reasons for achieving better performance using sensor fusion were identified from this study.

Peer-review under responsibility of the scientific committee of the ICPR25 International Scientific & Advisory and Organizing committee members

Keywords: Manufacturing Assembly; Multimodal Sensor System; Sensor Fusion; Deep Learning; Action Recognition

* Corresponding author. Tel.: +1-573-341-4493.
E-mail address: qinr@mst.edu

Peer-review under responsibility of the scientific committee of the ICPR25 International Scientific & Advisory and Organizing committee members
10.1016/j.promfg.2020.01.288

1. Introduction

Action recognition involves automatically detecting and recognizing purposeful motions by analyzing relevant human subject data obtained from multitudinous sensors [1]. Along with vigorous applications of action recognition in assisted living [2, 3, 4, 5] , human machine interaction [3, 6, 7], and healthcare [3, 4, 6, 8], it is also increasingly being used in manufacturing for quantifying and evaluating worker performance [9], understanding worker's operational behavior [10], enabling adaptive and incessant support to workers in assembly lines [11], and executing maintenance work [12]. Though a robust and reliable system of action recognition is greatly needed by various industrial applications, limited research has been dedicated to this endeavor [13]. Particularly, on the way to Smart Manufacturing, worker-centric assembly is becoming an inevitable part of production system [14]. About 40% of cost and 70% of production time involved with assembly still require the manual operation from workers [11]. Action recognition, if it can be reliably and timely achieved, would provide an ability to quickly identify workers' needs for assistance or on the job training, thus helping reduce production time and cost [11]. Henceforth, a highly reliable system of action recognition is of paramount interest in assembly operation.

Recognizing actions that workers take in performing complex, labor-intensive assembly tasks is not trivial due to a variety of reasons, such as varying time length of actions, kinematic complexity, anthropomorphic variation, viewpoint variation, occlusion, cluttered background, execution rate, and camera motion [15]. Following the increasing trend of demands for neoteric and customized products [16], more intricate and complex operations have been introduced to assembly processes, which magnifies the challenge of action recognition in assembly [17]. Data obtained by one sensor or a single type of sensors are becoming less reliable for recognizing those actions. Data from multiple types of sensors may complement each other, making sensor fusion a way of creating synergy [1, 18]. The fusion of meta classifiers constructed from multiple sources of sensor data has been found to increase the efficiency and accuracy of recognition system to a great extent [19]. Yet a fundamental understanding of fusion mechanisms is needed to better guide sensor fusion in practices.

With the immense advancements in microelectromechanical technology, micro sensors, and wireless communication technology, action recognition using wearable sensors (e.g., accelerometer, gyroscope, and magnetometer) has become a hot research topic [3, 7]. Their popularity is getting increased because of various attractive features such as easy to carry, inexpensive, wireless, privacy preservation, and low computational cost [2, 3, 20]. Wearable sensors based on different technologies are not equally good in capturing all human actions because each sensor type is specifically designed for capturing certain data. For example, a 3-axial accelerometer measures the acceleration along 3 orthogonal axes. A 3-axial gyroscope provides the orientation and rotation movement of sensed object with pitch, roll and yaw angles. A magnetometer measures the local Earth magnetic field vector. Sensor fusion is a straightforward way to take advantage of these technologies to advance action recognition.

Recognition of assembly actions was mostly done using a single classifier. For example, a hidden Markov model (HMM) was used to classify 9 activities involved in a wood workshop [12], and 21 gestures of bicycle repairing tasks [21]. Stiefmeier et al. [13] proposed a string-matching-based segmentation and classification method to identify 46 quality checking activities in car assembly. A k-nearest neighbor (k-NN) algorithm was used to classify 4 basic assembly tasks [22]. Tao et al. [9] collected data on 6 assembly actions and trained a Convolutional Neural Network (CNN) for action classification. Some exceptions have been noticed, which integrated results from multiple classifiers. For example, hierarchical with equal weights [23], hierarchical with variable weights upon classification performance [18, 24], majority voting [25], naive Bayesian [25], and the combination of both hierarchical and majority voting [19] have been proposed for the decision level fusion. Sensor fusion comes into effect at not only the decision level, but the data level and feature level. For example, features from both time and frequency domains were concatenated to recognize 11 child activities [26].

This paper presents a study of developing a sensor fusion based system for recognizing human actions in performing assembly tasks. The study aimed to advance the fundamental understanding of sensor fusion through exploiting opportunities of sensor fusion at all levels (data, feature, and score level) and discovering fusion mechanisms that can effectively enhance the ability of action recognition. Therefore, the rest of the paper is organized as the following. Section 2 presents the approach to creating the proposed system of action recognition based on sensor fusion, followed by Section 3 that presents an example for illustrating the implementation and assessment of the proposed approach. Findings from the study and needed future work are summarized at the end, in Section 4.

Nomenclature

ADS	action-level data segment
EMG	electromyography
IMU	inertial measurement unit
LOO	leave one out
TTS	train test split
i	index of body joints
j	index of joint angles
k	index of actions
m	index of models/sensor units
n	index of workers
$l/u/v$	indices of ADSs in the overall dataset, training dataset, and testing dataset
A_j	features of joint angle j
$D_{m,n}$	sensor data of worker n obtained by sensor unit m
I_{joint}	index set of body joints
I_{angle}	index set of joint angles
I_{action}	index set of actions
I_{model}	index set of models
I_{worker}	index set of workers
J_i	coordinates of joint i
L_i	distance feature of joint i
M_m	CNN model trained using the data obtained by sensor unit m
N_{action}	the number of actions
N_{worker}	the number of workers
$N_{m,n}$	the number of ADSs extracted from dataset $D_{m,n}$
$R_{m,k}$	the accuracy of model m on predicting action k
S_m	the set of ADSs using sensor unit m
$S_m^{tn/ts}$	the ADSs dataset for training/testing model M_m
$\boldsymbol{x}_{m,l/u/v}$	ADS obtained from sensor unit m
$y_{m,l/u/v}$	the ground truth of ADSs obtained from sensor unit m

2. The Approach

The approach to creating the multimodal sensor system and using sensor fusion to achieve reliable recognition of assembly actions is presented below. The proposed method of this paper is not restricted to a specific set of actions in a particular assembly process. Therefore, the method can be generalized for recognizing assembly actions at any workstation of an assembly line wherein the involvement of workers is an inevitable part of the process.

2.1. The multimodal sensor system

This study used two Myo armbands developed by the Thalmic Labs (https://developerblog.myo.com/) and one Kinect developed by Microsoft (https://developer.microsoft.com/en-us/windows/kinect) to form the multimodal sensor system. The Myo armband is a wearable device that consists of 8 surface electromyography (EMG) sensors and a 9-axis inertial measurement unit (IMU) that consists of a 3-axis gyroscope, a 3-axis accelerometer, and a 3-axis magnetometer. The IMU of Myo armband provides the spatial data (i.e., orientation and motion) of the armband in 13 channels: orientation (7 channels, including quaternions and Euler angles), angular velocities (3 channels), and accelerations (3 channels). Sampling frequencies of the EMG sensors and IMU are 200 Hz and 50 Hz, respectively. Kinect, an infrared light range-sensing sensor, contains an RGB camera (640 × 480 pixels @ 30 Hz), a 3D depth

sensor, and a four-microphone array. This study used a Kinect to track the 3D Cartesian coordinates of 17 skeletal joints of workers, as Fig. 1(a) shows. The sampling frequency of Kinect is 30 Hz. As Fig. 1(b) illustrates, the two

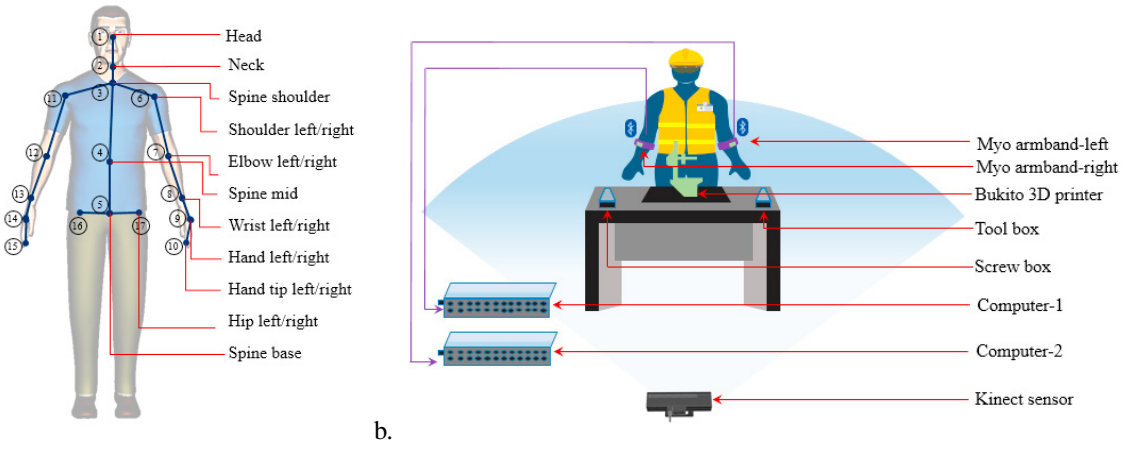


Fig. 1. The multimodal sensor system: (a) tracked skeletal joints and (b) setup of the sensor system.

Myo armbands were worn at the left and right forearm regions of a worker, respectively. The EMG and IMU signals of each armband were transmitted to a laptop via the Bluetooth communication unit. The Kinect was used to monitor the assembly operation by the worker, and stored the RGB images and the skeletal joints data of the worker.

2.2. Data preparation

Workers sensed by the multimodal sensor system are indexed by n and the total number of workers is N_{worker}. Sensor data of these workers were collected for either training action recognition models or testing the models. Sensor data were pre-processed and turned into data with a structure suitable for the convolutional neural network (CNN) model chosen by this study. The data preparation involves four procedures, which are presented below.

2.2.1. Sensor data fusion

The IMU signals of each armband were collected and saved as a time series dataset with 13 channels (3 for acceleration, 3 for velocity, and 7 for orientation), as Fig. 2(a) illustrates. The dataset obtained from the armband worn on the right arm of any worker n, denoted by $D_{R-IMU,n}$, provides information on the orientation and motion of the worker's right hand. Information of the left hand is given by the the other dataset, $D_{L-IMU,n}$, obtained from the left hand armband.

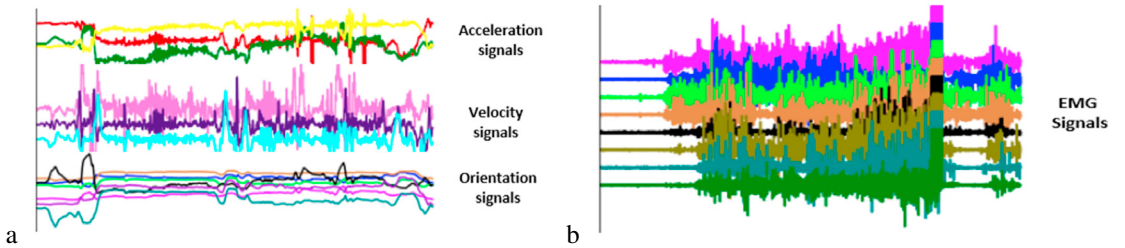


Fig. 2. Data Fusion: (a) IMU time series data of 13 channels, $D_{L/R-IMU,n}$; (b) EMG time series data of 8 channels, $D_{L/R-EMG,n}$.

The EMG signals of each armband on worker n, collected by eight independent EMG sensors, were similarly turned into a 8-channel time series dataset, illustrated in Fig. 2(b). With two armbands, two datasets were obtained, $D_{R-EMG,n}$ and $D_{L-EMG,n}$, which capture muscle activities of the right arm and left arm, respectively.

2.2.2. Skeletal feature calculation and feature fusion

Coordinates of skeletal joints are spatial data containing the information of human configuration and motion [10]. This study turned the coordinates of joints into two types of skeletal features and used the time series of the features to model worker assembly actions.

Let the spine shoulder (joint #3 in Fig. 1(a)) be the reference point or the origin of human skeleton. The distance from each joint to the origin was calculated as a distance feature. Yet distance features may vary with workers who are different in the height, size, arm length, and the location from the camera. To make distance features invariant to aforesaid factors, distance features were further normalized by dividing them by the vertical distance between spine shoulder (joint #3) and spine base (joint #5). Let J_i be the location coordinate of the ith joint sensed by the Kinect, $i \in I_{joint} = \{1, \ldots, 17\}$. The normalized distance feature L_i is computed as:

$$L_i = \frac{\|J_i - J_3\|_2}{\|J_5 - J_3\|_2}, \quad \forall \ i \in I_{joint} \backslash \{3, 5\}. \tag{1}$$

Eq. (1) indicates L_3 and L_5 were excluded, which are equal to 0 and 1, respectively.

Joint coordinates were further used to calculate joint angles between any two connected limbs [27]. Provided with the 17 joints, there are 16 joint angles in total, indexed by j, $j \in I_{angle} = \{1, \ldots, 16\}$. Let b_j and b'_j be the orientations of the two limbs forming the jth skeletal angle, the angle feature A_j is calculated as

$$A_j = \arccos \frac{{b_j}^T b'_j}{\|b_j\|_2 \|b'_j\|_2}, \quad \forall j \in I_{angle}. \tag{2}$$

The orientation of any limb, such as b_j in (2), can be calculated from the coordinates of joints on the two ends of the limb.

Skeletal distance features and joint angle features of any worker n were further fused, becoming a skeletal feature vector with 31 channels, $[L_1, L_2, L_4, L_6, \ldots, L_{17}, A_1, \ldots, A_{16}]$. The time series data of this 31-channel feature vector is the Kinect dataset, $D_{Kinect,n}$, illustrated in Fig. 3.

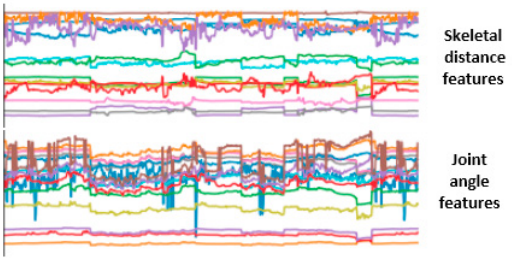


Fig. 3. Time series of skeletal features including both distance features and joint angle features.

2.2.3. Segmentation of action-level time series data

Consider an assembly operation that requires a worker to sequentially take N_{action} actions indexed by k. Throughout the operation, the multimodal sensor system continuously captured five sets of time series data aforementioned, $D_{m,n}$, for $m \in I_{model}$ = {L-EMG, R-EMG, L-IMU, R-IMU, Kinect} and for any $n \in I_{worker} = \{1, \ldots, N_{worker}\}$. Each dataset $D_{m,n}$ is long time series data covering N_{action} actions. In this study, a sliding window technique was applied to extract meaningful short segments of time series data from the long time series datasets $\{D_{m,n}\}$. In the remainder of the paper,

the extracted segments are named action-level time series data segments (ADSs). These ADSs were used to train and test CNN models for action recognition. The selection of a window length is vital. ADSs of very short time span do not contain sufficient information to fully recognize the performed actions. Those of very long time span may contain data of more than one action [28]. Successive windows were overlapped to better handle the transition from one action to another. The length of sliding window and the overlapping degree largely depend on sensor data and actions to be recognized [28]. Given the appropriately selected window length and overlapping degree, in total $N_{m,n}$ ADSs were extracted from $D_{m,n}$, indexed by l. Let $\boldsymbol{x}_{n,m,l}$ denote an ADS extracted from dataset $D_{m,n}$, and $y_{n,m,l}$ be the ground truth that identifies the action of worker n corresponding to this ADS.

2.2.4. Oversampling to balance the sample sizes among actions

ADSs extracted from the dataset $D_{m,n}$ were samples of multiple actions. Among the $N_{m,n}$ ADSs, $N_{m,n,k}$ ADSs were samples of action k. $N_{m,n,k}$ can dramatically vary from one action to another due to the fact that different actions possess different cycle times in a manufacturing assembly. For instance, grabbing a tool might take much less time compared to tightening a screw using an Allen key. The very imbalanced sample sizes across different actions would imapct the performance of classification algorithms [29]. To overcome this issue, this study adopted an oversampling technique to balance the sample distribution among actions through a random replication of minority class instances [30]. After applying the oversampling technique, all the actions a worker took have the same number of ADSs, equal to $\max_{k\in I_{action}}\{N_{m,n,k}\}$. Consequently, the total number of ADSs for any action k, collected from the N_{worker} workers, is $\sum_{n\in I_{worker}} \max_{k\in I_{action}}\{N_{m,n,k}\}$, $m \in I_m$.

2.3. CNN models for action recognition and model fusion

Let $S_m = \{\boldsymbol{x}_{m,l}\}$ denote the set of ADSs (indexed by l) that were collected from the N_{worker} workers and covered all actions they took. S_m was split into two mutually exclusive subsets: $S_m^{tn} = \{\boldsymbol{x}_{m,u}\}$ and $S_m^{ts}\{\boldsymbol{x}_{m,v}\}$, where u and v are indices of ADSs in these two datasets, respectively. A CNN model M_m was trained using S_m^{tn}, and S_m^{ts} was used to test the model. In total five models were trained: M_{L-EMG}, M_{R-EMG}, M_{L-IMU}, M_{R-IMU}, and M_{Kinect}.

For any v, there were five ADSs obtained by the five different sensor units, respectively. Therefore, worker actions can be predicted by five independent models as well as by a combination of these models. For an ADS $\boldsymbol{x}_{m,v}$, the prediction made by model M_m is a probability distribution on I_{action}, denoted by $P_{m,v}$ and $\|P_{m,v}\|_1 = 1$. Model fusion was used in this study to attempt to improve the prediction performance. Table 1 lists four methods of model fusion considered in this study. AF1, AF2, and AF3 in Table 1 are based the average fusion method that averages the predictions of various independent models. WF is a weighted average of predictions, which is discussed below.

Table 1. Methods of model fusion.

Fusion Model Name	Index Set of Models for Fusion	Fusion Method
AF1	$I_{AF1} = \{L\text{–}IMU, R\text{–}IMU\}$	$\frac{1}{2}\sum_{m\in I_{AF1}} P_{m,v}$
AF2	$I_{AF2} = \{L\text{–}IMU, R\text{–}IMU, Kinect\}$	$\frac{1}{3}\sum_{m\in I_{AF2}} P_{m,v}$
AF3	$I_{AF3} = \{L\text{–}IMU, R\text{–}IMU, L\text{–}EMG, R\text{–}EMG, Kinect\}$	$\frac{1}{5}\sum_{m\in I_{AF3}} P_{m,v}$
WF	$I_{WF} = I_{AF3}$	$\frac{1}{5}\sum_{m\in I_{WF}} w_m \cdot P_{m,v}$

The five CNN models may not be equally good at predicting all the actions of assembly operation. For example, tightening a screw using fingers may be predicted better by M_{R-EMG} or M_{R-IMU} than does M_{Kinect} because EMG and IMU data should be more capable than skeletal features in capturing fine motions. A weighted fusion method emphasizing the strength of each model in recognizing specific actions was proposed. In this study a modification factor assigned to the prediction of action k made by model m, $w_{m,k}$, was determined using the corresponding recall, $R_{m,k}$. For any action k, the five CNN models were divided into two tiers according to their recall values: models with the top two recall values were in the first tier and their predictions were amplified by multiplying the modification factor $w_{m,k}$ that is greater than 1. The remaining three were in the second tier and their predictions are remained unchanged; that is, the modification factor for tier models is 1. Let $w_m = [w_{m,1}, \ldots, w_{m,N_{action}}]$ be the vector of the modification factors for model m. In the last line of Table 1 $w_m \cdot P_{m,v}$ is the modified prediction of model M_m, obtained by performing the element-wise multiplication between w_m and $P_{w,v}$.

3. An Illustrative Example

3.1. The experiment setup

The workstation for assembling a Bukito 3D printer was set up in a lab. The study used one step of the assembly process, named "putting on the handle" in the product assembly instruction manual, as an example to illustrate and assess the proposed approach to creating the action recognition system. This step of assembly includes seven actions shown in Fig. 4. To capture the between-subjects variability, five subjects were recruited to perform the assembly. Their ages were ranged from 18 to 55 years. The multimodal sensor system for sensing workers in this assembly task was set up in the way shown in Fig. 1(b). The Allen key set, screw box, tool box, and Bukito kit were tools and material involved in this assembly. To count the randomness of human actions (i.e., the within-subject variability), the five subjects were asked to repeat the assembly for 10 times.

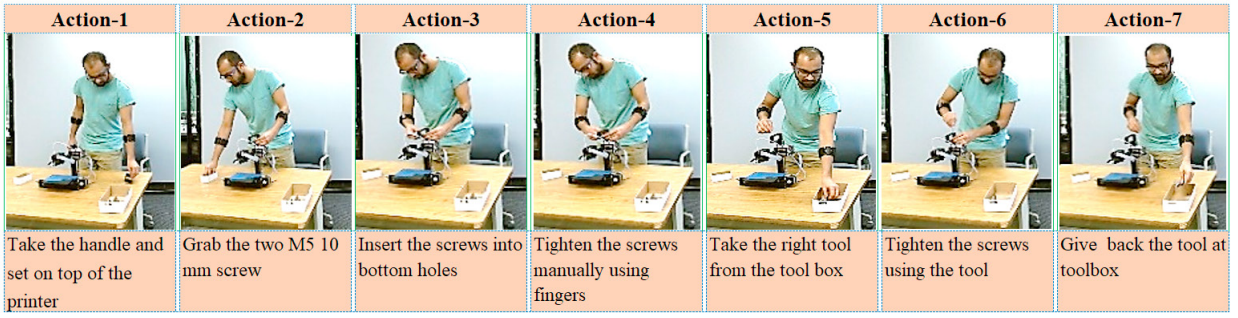


Fig. 4. Sequential actions for "putting on the handle" in assembling a Bukito 3D printer.

3.2. Data preparation

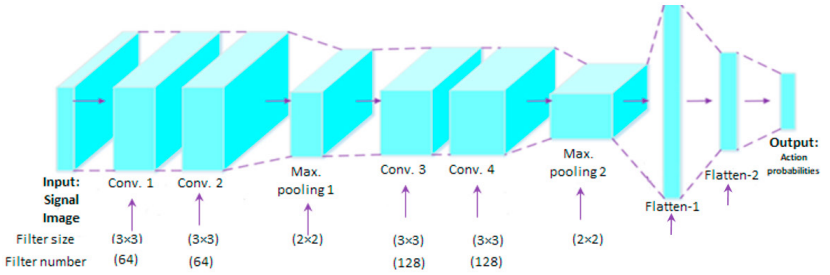
Time series sensor datasets were created by following the method described in section 2.2. Then, ADSs were segmented from these datasets using a sliding window that can cover 2 seconds of time series data. Any two successive ADSs have a 50% ovelap. The time to complete an action varies significantly among the 7 actions so that the number of ADSs varied largely from one action to another, as Table 2 illustrates. The oversampling method was applied to create evenly distributed ADSs across the seven actions.

Table 2. The sample size of ADSs (before/after oversampling)

Subjects	Action-1	Action-2	Action-3	Action-4	Action-5	Action-6	Action-7
Subject-1	26/122	46/122	46/122	70/122	44/122	122/122	50/122
Subject-2	30/128	24/128	66/128	72/128	62/128	128/128	64/128
Subject-3	10/100	18/100	44/100	24/100	36/100	100/100	30/100
Subject-4	10/78	16/78	32/78	64/78	30/78	78/78	30/78
Subject-5	22/134	30/134	60/134	74/134	68/134	134/134	56/134

3.3. Training CNN models for action recognition

Five CNN models $\{M_m | m \in I_{model}\}$ for classifying actions were independently trained. The model M_m was trained to automatically extract discriminative features of worker actions from the dataset S_m^{st}. ADSs were normalized to be within the range $[-1, 1]$ before being fed to the network. The proposed CNN architecture is illustrated in Fig. 5. Parameters for individual CNNs are also summarized in this figure. The Leave One Out (LOO) validation method

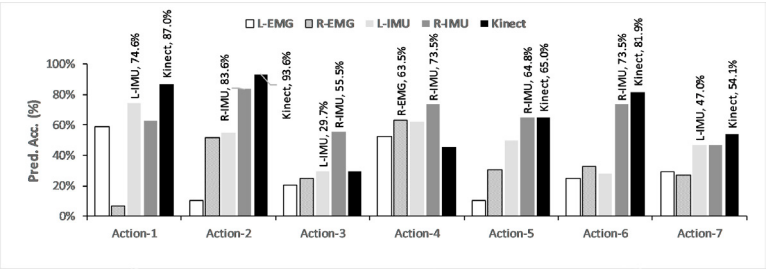


Models			
Layers	L/R-EMG	L/R-IMU	kinect
Input	(400,8,1)	(100,13,1)	(32,31,1)
Conv-1	(400,8,64)	(100,13,64)	(32,31,64)
Conv-2	(400,8,64)	(100,13,64)	(32,31,64)
Max.Pooling-1	(200,4,64)	(50,6,64)	(16,15,64)
Conv-3	(200,4,128)	(50,6,128)	(16,15,128)
Conv-4	(200,4,128)	(50,6,128)	(16,15,128)
Max.Pooling-2	(100,2,128)	(25,3,128)	(8,7,128)
Flatten-1	25,600	9,600	7,168
Flatten-2	256	256	256
Dense	7	7	7

Fig. 5. The proposed CNN architecture.

was performed to evaluate the strength of individual CNN models in recognizing human actions in assembly. Each time, data of four subjects were used for training the CNN models and data of the remaining one subject were used to assess the prediction accuracy of the models. The LOO validation was performed five times and each time the subject for testing was rotated to a different worker. Consequently, all the ADSs were tested. Given the testing result, the prediction accuracy of each model on each task was calculated, shown in Fig. 6(a). For each action, the two models of the first tier are labeled with their accuracy values in Fig. 6. Accordingly, the vector of prediction modification factors, w_m, was determined for each model m, shown in Fig. 6(b).

a



b

Models, m	Actions, k						
	1	2	3	4	5	6	7
L-EMG	1	1	1	1	1	1	1
R-EMG	1	1	1	1.4	1	1	1
L-IMU	1.4	1	1.4	1	1	1	1
R-IMU	1	1.4	1.4	1.4	1.4	1.4	1.4
Kinect	1.4	1.4	1	1	1.4	1.4	1.4

Fig. 6. Evaluation of model strength: (a) prediction accuracy, and (b) prediction modification factors $w_{m,k}$.

3.4. Assessment of sensor fusion methods

The five CNN models and the four fusion methods based on these models were assessed using two cross validation methods: the LOO and the Train-Test Split (TTS). In use of the TTS method, 50% of ADSs of every subject were used for training and the remaining 50% were used for testing. Fig. 7 shows the prediction accuracy of the five individual CNN models and the four fusion methods evaluated by both LOO and TTS methods. Among the five individual models, the prediction accuracy of M_{R-IMU} was similar to that of M_{Kinect}, and they outperformed the other three models. $AF1$, the average fusion of M_{L-IMU} and M_{R-IMU}, had higher accuracy than did each individual model. $AF2$, which adds M_{Kinect} to $AF1$, further increased the accuracy by 6.7% and 8.2% in the LOO and TTS evaluations, respectively. Compared to $AF2$, the average fusion of all five CNN models, $AF3$, only improved the accuracy by 0.6% in the LOO evaluation. The proposed weighted average fusion, WF, outperformed $AF3$ by increasing the accuracy by 1.5% in the LOO evaluation and 2.7% in the TTS evaluation. It was noticed that $AF3$ in the TTS evaluation had

a lower accuracy (81.9%) than $AF2$ (84.3%). The weighted average fusion WF overcame the limitation of $AF3$ and achieved the highest accuracy (84.6%).

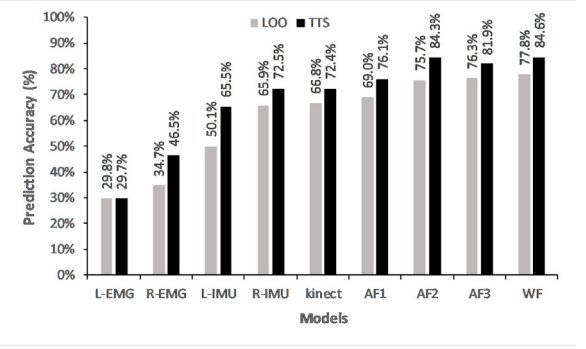


Fig. 7. Evaluation of prediction accuracy.

4. Conclusions and Future Work

In this study, we propose a multimodal sensor system and weighted fusion method in developing a robust and reliable action recognition system. The study shows that, EMG and IMU data perform well in recognizing fine actions involved with finger motion i.e., insert screw, tighten screw manually. On the other hand skeletal data does better in recognizing coarse actions involved with arm motion i.e., take the handle, give back the tool. Moreover, we have shown that reliability of the recognition system can be enhanced by fusing the aforesaid information at data, feature and score levels to complement the information and applying oversampling technique to overcome the limitation of unbalanced dataset. Furthermore, we also proposed a new weighted fusion method, that emphasizes each model's strength in recognizing specific actions and put modification factor to the predictions accordingly in decision making which improves the result. The effectiveness of our proposed sensor system and weighted fusion method has been verified with an illustrative example of assembling a Bukito 3D printer in lab setting.

The study of this paper builds a foundation for important future work. For instance, exploring other fusion methods to further improve the accuracy; scaling up the current work by applying the proposed approach to the recognition of assembly actions on various workstations of an assembly line.

Acknowledgements

This work was supported by NSF grant CMMI-1646162 on cyber-physical systems. 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. The authors also thank the project team members Md Moniruzzaman and Ze-Hao Lai for their insightful comments and suggestions on this study.

References

[1] C. Chen, R. Jafari, N. Kehtarnavaz, A survey of depth and inertial sensor fusion for human action recognition, Multimedia Tools and Applications 76 (3) (2017) 4405–4425.

[2] A. Wang, G. Chen, J. Yang, S. Zhao, C.-Y. Chang, A comparative study on human activity recognition using inertial sensors in a smartphone, IEEE Sensors Journal 16 (11) (2016) 4566–4578.

[3] H. Junker, O. Amft, P. Lukowicz, G. Tröster, Gesture spotting with body-worn inertial sensors to detect user activities, Pattern Recognition 41 (6) (2008) 2010–2024.

[4] F. Attal, S. Mohammed, M. Dedabrishvili, F. Chamroukhi, L. Oukhellou, Y. Amirat, Physical human activity recognition using wearable sensors, Sensors 15 (12) (2015) 31314–31338.

[5] S. Dernbach, B. Das, N. C. Krishnan, B. L. Thomas, D. J. Cook, Simple and complex activity recognition through smart phones, in: 2012 8th International Conference on Intelligent Environments, IEEE, 2012, pp. 214–221.
[6] M. Ramanathan, W.-Y. Yau, E. K. Teoh, Human action recognition with video data: research and evaluation challenges, IEEE Transactions on Human-Machine Systems 44 (5) (2014) 650–663.
[7] T. T. Ngo, Y. Makihara, H. Nagahara, Y. Mukaigawa, Y. Yagi, Similar gait action recognition using an inertial sensor, Pattern Recognition 48 (4) (2015) 1289–1301.
[8] K. Altun, B. Barshan, Human activity recognition using inertial/magnetic sensor units, in: International Workshop on Human Behavior Understanding, Springer, 2010, pp. 38–51.
[9] W. Tao, Z.-H. Lai, M. C. Leu, Z. Yin, Worker activity recognition in smart manufacturing using IMU and sEMG signals with convolutional neural networks, Procedia Manufacturing 26 (2018) 1159–1166.
[10] M. Al-Amin, R. Qin, W. Tao, M. C. Leu, Sensor data based models for workforce management in smart manufacturing, in: Proceedings of The 2018 Industrial and Systems Engineering Research Conference (ISERC'18), Orlando, FL, 2018, pp. 481–486.
[11] M. Aehnelt, E. Gutzeit, B. Urban, Using activity recognition for the tracking of assembly processes: Challenges and requirements, in: Proceedings of the Workshop on Sensor-Based Activity Recognition, 2014.
[12] J. A. Ward, P. Lukowicz, G. Troster, T. E. Starner, Activity recognition of assembly tasks using body-worn microphones and accelerometers, IEEE transactions on Pattern Analysis and Machine Intelligence 28 (10) (2006) 1553–1567.
[13] T. Stiefmeier, D. Roggen, G. Ogris, P. Lukowicz, G. Tröster, Wearable activity tracking in car manufacturing, IEEE Pervasive Computing 7 (2) (2008) 42–50.
[14] W. Unzeitig, M. Wifling, A. Stocker, M. Rosenberger, Industrial challenges in human-centred production, in: MOTSP 2015-lnternational Conference Management of Technology, 2015, pp. 10–12.
[15] J. Zhang, W. Li, P. O. Ogunbona, P. Wang, C. Tang, RGB-D-based action recognition datasets: A survey, Pattern Recognition 60 (2016) 86–105.
[16] C. Scheuermann, S. Verclas, B. Bruegge, Agile factory-an example of an industry 4.0 manufacturing process, in: 2015 IEEE 3rd International Conference on Cyber-Physical Systems, Networks, and Applications (CPSNA), IEEE, 2015, pp. 43–47.
[17] O. Sand, S. Büttner, V. Paelke, C. Röcker, smart. assembly–projection-based augmented reality for supporting assembly workers, in: International Conference on Virtual, Augmented and Mixed Reality, Springer, 2016, pp. 643–652.
[18] M. Guo, Z. Wang, N. Yang, Z. Li, T. An, A multisensor multiclassifier hierarchical fusion model based on entropy weight for human activity recognition using wearable inertial sensors, IEEE Transactions on Human-Machine Systems 49 (1) (2019) 105–111.
[19] O. Banos, M. Damas, H. Pomares, F. Rojas, B. Delgado-Marquez, O. Valenzuela, Human activity recognition based on a sensor weighting hierarchical classifier, Soft Computing 17 (2) (2013) 333–343.
[20] S. Gaglio, G. L. Re, M. Morana, Human activity recognition process using 3-D posture data, IEEE Transactions on Human-Machine Systems 45 (5) (2015) 586–597.
[21] T. Stiefmeier, G. Ogris, H. Junker, P. Lukowicz, G. Troster, Combining motion sensors and ultrasonic hands tracking for continuous activity recognition in a maintenance scenario, in: 2006 10th IEEE International Symposium on Wearable Computers, IEEE, 2006, pp. 97–104.
[22] S. Kaghyan, H. Sarukhanyan, Activity recognition using k-nearest neighbor algorithm on smartphone with tri-axial accelerometer, International Journal of Informatics Models and Analysis (IJIMA) 1 (2012) 146–156.
[23] C. Zhu, W. Sheng, Wearable sensor-based hand gesture and daily activity recognition for robot-assisted living, IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans 41 (3) (2011) 569–573.
[24] Y. Guo, W. He, C. Gao, Human activity recognition by fusing multiple sensor nodes in the wearable sensor systems, Journal of Mechanics in Medicine and Biology 12 (05) (2012) 1250084.
[25] P. Zappi, T. Stiefmeier, E. Farella, D. Roggen, L. Benini, G. Troster, 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, IEEE, 2007, pp. 281–286.
[26] Y. Nam, J. W. Park, Child activity recognition based on cooperative fusion model of a triaxial accelerometer and a barometric pressure sensor, IEEE Journal of Biomedical and Health Informatics 17 (2) (2013) 420–426.
[27] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, R. Bajcsy, Sequence of the most informative joints (SMIJ): A new representation for human skeletal action recognition, Journal of Visual Communication and Image Representation 25 (1) (2014) 24–38.
[28] O. D. Lara, M. A. Labrador, A survey on human activity recognition using wearable sensors, IEEE Communications Surveys & Tutorials 15 (3) (2013) 1192–1209.
[29] R. Akbani, S. Kwek, N. Japkowicz, Applying support vector machines to imbalanced datasets, in: European Conference on Machine Learning, Springer, 2004, pp. 39–50.
[30] S. Kotsiantis, D. Kanellopoulos, P. Pintelas, et al., Handling imbalanced datasets: A review, GESTS International Transactions on Computer Science and Engineering 30 (1) (2006) 25–36.