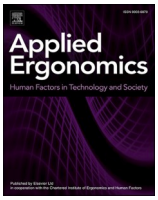
Applied Ergonomics 113 (2023) 104090

Contents lists available at ScienceDirect

Applied Ergonomics

journal homepage: www.elsevier.com/locate/apergo



Human activity recognition in an end-of-life consumer electronics disassembly task

Yuhao Chen [a,1], Hao-Yu Liao [b,1], Sara Behdad [b,**], Boyi Hu [a,*]

[a] Department of Industrial and Systems Engineering, University of Florida, FL 32611, USA
[b] Department of Environmental Engineering Sciences, University of Florida, FL 32611, USA

ARTICLE INFO

Keywords:
e-waste disassembly
Task recognition
Deep learning methods

ABSTRACT

The production of electronic waste, also known as e-waste, has risen with the growing reliance on electronic products. To reduce negative environmental impact and achieve sustainable industrial processes, recovering and reusing products is crucial. Advances in AI and robotics can help in this effort by reducing workload for human workers and allowing them to stay away from hazardous materials. However, autonomous human motion/intention perception is a primary barrier in e-waste remanufacturing. To address the research gap, this study combined experimental data collection with deep learning models for accurate disassembly task recognition. Over 570,000 frames of motion data were collected from inertial measurement units (IMU) worn by 22 participants. A novel sequence-based correction (SBC) algorithm was also proposed to further improve the accuracy of the overall pipeline. Results showed that models (CNN, LSTM, and GoogLeNet) had an overall accuracy of 88–92%. The proposed SBC algorithm improved accuracy to 95%.

1. Introduction

In comparison to the ongoing fourth industrial revolution, which focuses on the cyber-physical system by leveraging advances in AI and the internet of things, the upcoming fifth industrial revolution will place a greater emphasis on the collaboration between humans and the intelligent machines to achieve green and sustainable industrial processes (Breque et al., 2021). As our dependence on electronic products increases, the amount of electronic waste (e-waste) being produced has been increasing each year. According to the Global E-waste Monitor 2020 (Forti et al., 2020), 53.6 million tonnes of e-waste were produced worldwide in 2019, with a growth rate of 21%. As e-waste management is becoming a growing concern, there has been an increased interest in end-of-use product recovery to reduce e-waste (Zuidwijk and Krikke, 2008). To extract valuable components and materials from end-of-use electronic products, the process of disassembly is a necessary step.

While disassembly could increase the number of product components available for recovery and reuse, the process is often labor-intensive with significant occupational hazards, both physically and psychologically, for workers (Acquah et al., 2019). Currently, disassembly is still dominated by human workers. Automation and robotics-based methods are not widely seen due to the inherent complexity of the task such as demanding precision and high flexibility. In compliance with Industry 5.0 paradigm, research efforts have been put into designing and implementing robot assistants to complement human workers and mitigate work-related safety issues (Xu et al., 2021; Sajedi et al., 2022; Chen et al., 2022a; Chen et al., 2022b). In order to enable the robot assistance in the disassembly process, human task recognition is necessary due to the fact that robots need to be able to understand human behavior and activity to collaborate with them effectively. By recognizing human activity, the robot can interpret their intentions and adjust its behavior accordingly. This is critical for developing a safe and seamless collaboration between the robot and human. However, real-time human task recognition, as an essential component in the system workflow, still remains a significant challenge, limiting the use of robotic-assisted disassembly safely and effectively.

Task recognition is the process of identifying and understanding the specific task that needs to be completed, and it can be an important and necessary step in many different fields, such as AI and HRC. To name a few studies, Kiruba et al. (2019) developed a hexagonal volume local

* Corresponding author.
** Corresponding author.
E-mail addresses: yuhaochen@ufl.edu (Y. Chen), haoyuliao@ufl.edu (H.-Y. Liao), sara.behdad@essie.ufl.edu (S. Behdad), boyihu@ise.ufl.edu (B. Hu).
[1] The first two authors contributed equally to this work.

https://doi.org/10.1016/j.apergo.2023.104090
Received 19 January 2023; Received in revised form 21 April 2023; Accepted 3 July 2023
Available online 17 July 2023

binary pattern descriptor by considering the motion and temporal information, with a single RGB camera. The approach achieved body-pose recognition accuracy rates of over 84% across multiple benchmark datasets. In Sajedi et al. (2022), deep learning models for hand recognition were created to explicitly quantify the prediction uncertainty based on two-dimensional images. Zhang et al. (2021) proposed a hybrid approach to recognize human assembling actions during HRC leveraging an image-based CNN model and variable-length Markov modeling. Despite solo usage of RGB camera for achieving satisfactory results, several studies, such as (Mazhar et al., 2019; Arivazhagan et al., 2019), used both RGB camera and depth sensitivity functions of the device (RGB-D system, Microsoft Kinect) for improved hand gesture recognition.

The aforementioned approaches have been proposed to track and recognize human actions using image data. However, their real-time capabilities were not guaranteed due to the heavy computing power and advanced hardware required for image processing. Furthermore, typical computer vision issues, such as being sensitive to camera occlusions, light variations, and shiny surfaces, may compromise the performance of vision-based approaches in real-world applications (Pfister et al., 2014; Roda-Sanchez et al., 2021). To overcome these challenges, Inertial Measurement Unit (IMU), a widely used wearable technology to provide motion data for human activity recognition (Lara and Labrador, 2012; Hu et al., 2021; Luo et al., 2020a), has shown promise in industrial operation processes (Koskimaki et al., 2009; Koskimäki et al., 2013). In Roda-Sanchez et al. (2021), an experiment was undertaken to compare an RGB-D based approach against an IMU-based gesture recognition algorithm in remanufacturing context. The results indicated that the proposed IMU-based approach had recognition accuracy rates up to 8.5 times higher. Moreover, it was shown that accuracy of the RGB-D based approach differs significantly depending on the plane where movements are performed as well as other factors such as ambient luminosity and focal length, making it unsuitable for complex movements like hand flips and screw/unscrew that are commonly carried out in disassembly processes.

Even though prior research has demonstrated promising results in task recognizing using IMUs, a major challenge is the lack of publicly available datasets. Unlike video datasets, which are well-established in both RGB (Schuldt et al., 2004; Liu et al., 2009; Zhou et al., 2018) and RGB-D (Wang et al., 2012; Koppula et al., 2013) modalities, there are currently far fewer IMU datasets. The authors are aware of IMU datasets for task recognition in common daily activities (Zhang and Sawchuk, 2012), air-writing (Tripathi et al., 2021), and gait analysis (Luo et al., 2020b), etc., however, they are even rarer for industrial operations (Dallel et al., 2020). To the best of our knowledge, no dataset for human task recognition in the e-waste disassembly process is currently available, which significantly restricts the broad attention from the machine learning and artificial intelligence communities.

Given the fact that the e-waste disassembly process could present many challenges that are not conducive to the application of computer vision methods, such as camera occlusions, light variations, and shiny surfaces, the primary objective of this study is to develop deep learning models that use IMU data to accurately detect disassembly tasks in complex disassembly settings of consumer electronics (Fig. 1). To fill the gap of lacking IMU datasets in the e-waste disassembly process, a set of experiments have been conducted to obtain human motion data from wearable IMUs. Then, Convolution Neural Networks (CNN), Long Short-Term Memory (LSTM), and GoogLeNet models were trained with the collected motion data and their performance were compared. Finally, a novel sequence-based correction approach was developed to increase the accuracy of the task recognition model. The main contributions of the study are:

- *The development of IMU-based deep learning models for task recognition allows intelligent systems, such as robot assistants, to detect human intention during complex disassembly processes.*

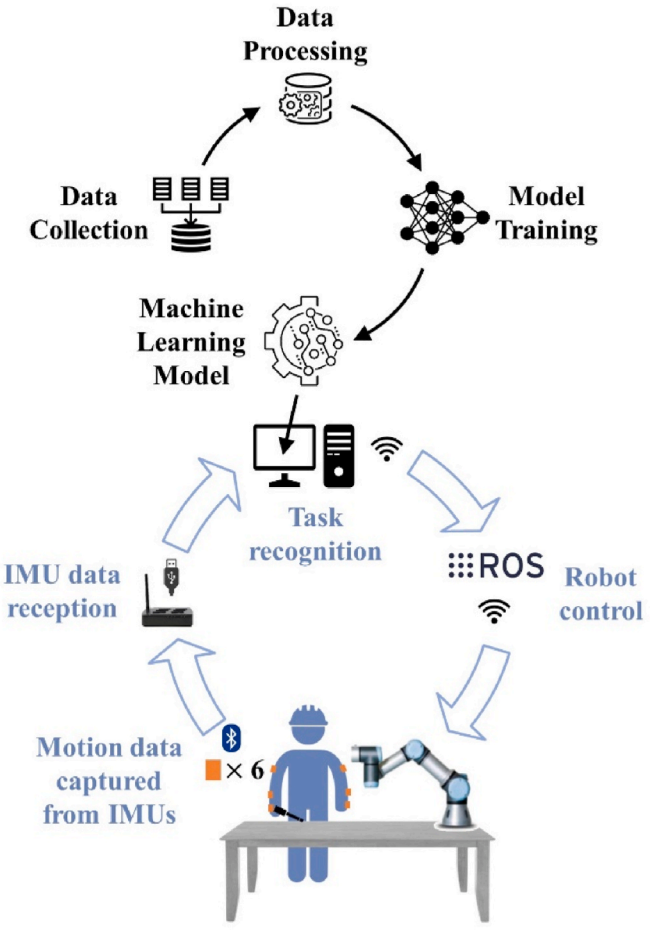


Fig. 1. The conceptual overview of the proposed task recognition method and an example application of robotic-assisted disassembly.

- *A sequence-based correction approach is proposed for enhancing the accuracy of the task recognition model in real-time based on a known disassembly sequence.*

2. Methods

2.1. Participants

Twenty-two participants (fourteen males and eight females) were recruited from the university students to participate in this study. Their mean (SD) age, height, and body weight were 25.4 (4.3) years, 174.6 (10.3) cm, and 65.4 (20.0) kg. All participants reported being healthy, having normal or corrected to normal vision with contact lenses, and without any musculoskeletal injuries that required medical treatment in the past 12 months. The majority of the participants, 20 out of 22, self-reported as being right-handed, while one participant claimed to be left-handed and the other claimed to be ambidextrous. Participants completed informed consent before any data collection, and the experimental protocol was approved by the University of Florida Institutional Review Board (IRB202200211).

2.2. Experimental data collection

Desktop computers have been selected for the study due to their wide application and contribution to the total e-waste generation. Six components were targeted for removal in a fixed sequence: 1) the thumbscrew, 2) the cover, 3) the hard disk drive, 4) the fan, 5) the heat sink, and 6) the memory module (RAM). The specific product used in the

study was a Dell OptiPlex 7050 Micro desktop computer (Dell Inc., Round Rock, TX). The components were selected based on their potential for reuse, recycling, or recovery. For example, as a general practice, hard drives are expected to be manually pulled out before starting the recycling process due to potential privacy concerns. On the other hand, during end-of-life electronic recycling and disassembly operations, RAM modules are usually easier to reuse and yield more favorable economic returns. The disassembly sequence was determined mainly due to the physical and design constraint. For example, the heat sink and RAM can only be removed after the fan to gain access. While it is true that different strategies or sequences for product component extraction may influence the results, or there may be other confounding factors, they are not the focus of the study, which aims to develop an IMU-based human task recognition method during e-waste disassembly operations.

Before the formal data collection, adequate training was provided to the participants based on the manufacturer's teardown removal guide to ensure that they were comfortable and confident in completing the disassembly tasks. Subsequently, participants were tasked to remove the six components in the predetermined sequence. Each component had to be positioned in the exact spot, as shown in Fig. 2. Each participant repeated the disassembly process five times, with each trial lasting 90 s, on average. To capture human motion, participants wore six IMU sensors (MVN Awinda, Xsens Technologies BV, Enschede, Netherlands) while performing disassembly tasks. The six sensors were placed on the left & right hands, left & right forearms, and left & right upper arms. The sampling frequency was set at 60 Hz. Fig. 3 shows a participant using the hand tool to remove the heat sink while wearing the six IMU sensors set.

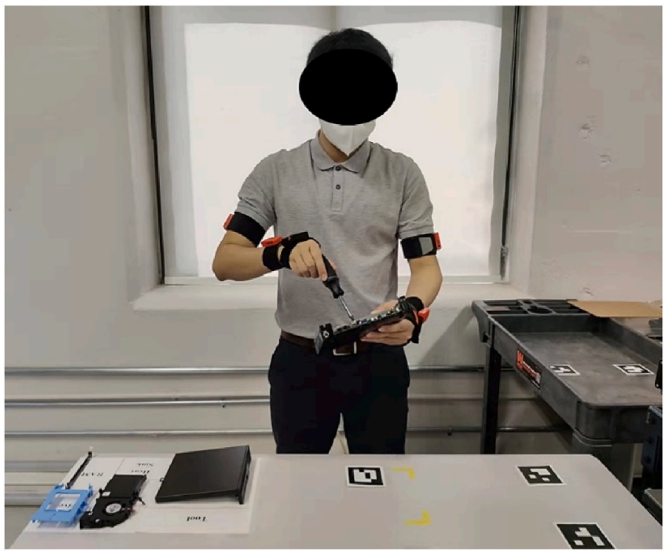


Fig. 3. A participant is performing the removal of the heat sink with a hand tool.

2.3. Initial data processing

Over 570,000 frames of motion data were collected from 22 participants. Joint angle and segment position were exported using MVN Analyze (Xsens Technologies BV, Enschede, Netherlands) and used as inputs for task recognition. For the joint angle, bilateral side of elbow ulnar deviation, protonation, and flexion angles were utilized. For the segment position, both the right hand and left hand's tri-axial coordinates (x, y, z) were used. In addition, a sliding window of 4 frames (66.7 ms) was applied to segment and augment the time-series data obtained from IMUs (please refer to Results and Discussion section for the determination of the window length). Thus, the input of the models at each time was a vector with a size of 1 × 1 × 48 (2 sides × 2 variables × 3 channels × 4 frames).

The output of the models was the task label of the current frame, i.e., 0: the thumbscrew disassembly, 1: the cover disassembly, 2: the hard disk drive disassembly, 3: the fan disassembly, 4: the heat sink disassembly, or 5: the RAM disassembly. The ground truth labels were marked manually by researchers. The implementation of the sliding window is shown in Fig. 4. The input and output of the task recognition models are summarized in Table 1. The input was the time-series motion data of the current frame plus the previous 3 frames, i.e., T. The output is the task label of the current frame, i.e., Y.

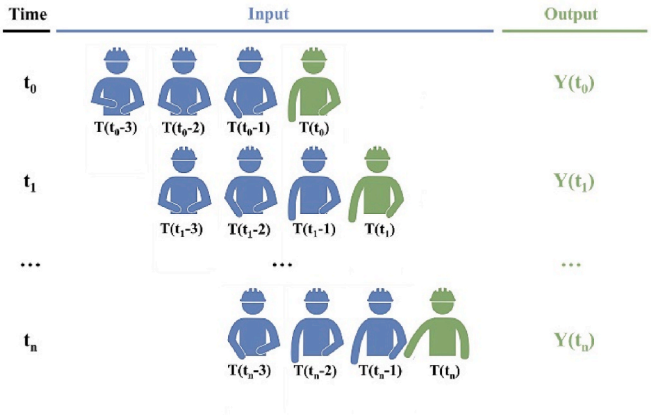


Fig. 4. The sliding window concept that was implemented in the study.

Since the two variables (joint angle and segment position) had different units, and the range of each channel was substantially different, the min-max normalization was applied to restrict the range of values from 0 to 1.

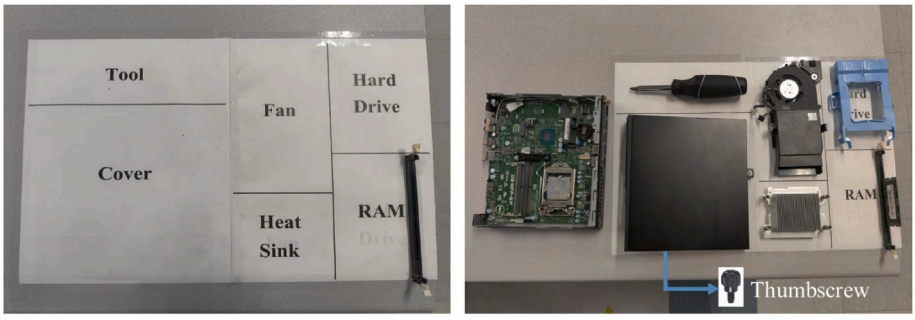


Fig. 2. Target positions for the six components (left). An example picture of a completed disassembly trial (right). Note that the target position of the thumbscrew is identical to that of the cover.

Table 1
The input and output of task recognition models.

Variable	Channel	Input lead time T	Output label Y
Joint angle	1 *Right/left elbow ulnar deviation* 2. *Right/left elbow pronation* 3. *Right/left elbow flexion*	T(t), T (t-1), T (t-2), T (t-3)	0: screw 1: cover 2: hard disk drive 3: fan
Segment position	1 *Right/left x* 2 *Right/left y* 3 *Right/left z*		4: heat sink 5: RAM

2.4. Model selection and architecture

After data collection, the next step is to predict the disassembly tasks using IMU motion data. Three architectures, i.e., CNN, LSTM, and GoogLeNet, were trained and compared based on their performance.

2.4.1. CNN

The CNN architecture used in this study involves the application of 2D transposed convolution in the first layer since the input was time-series data with a size of $1 \times 1 \times 48$. The 2D transposed convolution converted the input size from $1 \times 1 \times 48$ to $224 \times 224 \times 3$, which is a commonly used size in the transfer learning models such as GoogLeNet and ResNet-50 (He et al., 2016). The batch normalization was applied to convolution layers. The optimizer was stochastic gradient descent (SGD) with a 0.001 learning rate and 0.9 momentum. The loss function was the cross-entropy loss.

2.4.2. LSTM

The LSTM network is proposed by Hochreiter and Schmidhuber to implement feedback connections in traditional feedforward neural networks (Hochreiter and Schmidhuber, 1997). In this study, five different hidden sizes of LSTM, i.e., 16, 32, 64, 128, 256, and 512, were selected as general settings suggested in the previous literature (Hu et al., 2018). The learning rate was 0.01. The optimizer was the adaptive moment estimation (Adam). The decay learning rate was applied with 0.1 for every 7 epochs. The cross-entropy loss function was applied because of the classification problem. The output layer was the linear activation function. After training and testing the models with five different hidden sizes, the hidden size with the best performance was selected for the task recognition.

2.4.3. GoogLeNet

GoogLeNet is a widely accepted model developed by Google in 2014 (Szegedy et al., 2015). There are 22 layers in GoogLeNet, which was trained by one million images with one thousand types of objects. In this study, we applied transfer learning by using GoogLeNet's pre-parameters. The learning rate was 0.001 with a momentum of 0.9 and the SGD optimizer. The decay learning rate was 0.1 for every 7 epochs. To satisfy the input size requirement of GoogLeNet, we modified the structure of GoogLeNet by adding one additional layer of a 2D transposed convolution at the beginning with the batch normalization to convert the size to $224 \times 224 \times 3$. The shape of 2D transposed convolution was the same as the CNN model.

2.5. Model implementation and training

The experiment data was divided into training, validation, and testing. Three random trials of each participant were selected for training and the remaining two trials were used for validation and testing. The total number of trials for training, validation, and testing was 66 (22 participants $\times$ 3 trials), 22 (22 participants $\times$ 1 trial), and 22, respectively. Before training, the data was sufficiently shuffled. The number of training, validation, and testing samples were 375668, 93917, and 108710, respectively. All training and testing were performed on a desktop computer with a i9-10900K CPU @ 3.70 GHz, an NVIDIA Quadro RTX 4000 GPU, and 64 GB RAM.

2.6. Sequence-based correction algorithm

In addition to the training of the aforementioned prediction models, we introduced a sequence-based correction (SBC) algorithm that can further improve the task recognition accuracy. The proposed algorithm includes the following rules:

- ***Rule 1*:** Given the input Y(t), which is the frame t task label obtained from the prediction model, subtract 1 from it, i.e., Y(t)-1, if the result doesn't equal the task label of the previous frame, i.e., Y (t-1), then replace Y(t) with Y (t-1).
- ***Rule 2*:** If Y(t)-1 equals the task label of the previous frame, i.e., Y (t-1), then check the task labels of the next 180 frames and replace Y(t) with the task label that appears most often in the next 180 frames. The replaced task label can only be Y(t) or Y (t-1) due to the sequence-based constraint.

Since all participants were tasked to perform the disassembly task in a fixed sequence, the output labels should adhere to the sequence-based requirement. For example, if the current task is removing the thumbscrew (label 0), the following task should be either removing the thumbscrew (label 0) or removing the cover (label 1). Other task labels are not reasonable. Based on the idea, rule 1 was applied to add the sequence-based constraint.

The aim of rule 2 is to check when the next task has begun. For example, if Y(t) and Y (t-1) are equal to 1 and 0, respectively, it is required to check whether task 1 has started or Y(t) is a prediction error. To do this, the task labels of the next 3 s (3 s $\times$ 60 Hz = 180 frames, the duration of 3 s is empirically defined) are used to check whether the next task starts. If the number of label 1 is more than the number of label 0 among the 180 frames, it is assumed that task 1 has started, thus Y(t) is considered to be correct and no correction is needed. Otherwise, Y(t) is regarded to be a prediction error, meaning task 1 has not started yet, in which case Y(t) is changed to Y (t-1) = 0. The pseudocode of the proposed SBC algorithm is presented below:

Algorithm 1. Sequence-based correction (SBC) algorithm

Algorithm 1 Sequence-based correction (SBC) algorithm

Input: The frame t task label obtained from the prediction model, i.e., Y(t)
Output: The corrected task label of frame t, i.e., Y'(t)

```
function SBC(Y(t))
    if Y(t)-1 doesn't equal to Y(t-1) then
        replace Y(t) with Y(t-1). Y'(t) stands for the result
        return Y'(t)
    else if Y(t)-1 is equals to Y(t-1) then
        check task labels of the next 180 frames
        if amount of Y(t) is larger than amount of Y(t-1) in Y(t) to Y(t+180) then
            return Y(t)
        else
            replace Y(t) with Y(t-1). Y'(t) stands for the result
            return Y'(t)
        end if
    end if
end function
```

Given that some of these modifications need the knowledge of the "future" 180 frames in advance, in this study, the SBC algorithm was launched 3 s after the machine learning model started running.

3. Results and Discussion

The accuracy is the primary metric utilized to compare the performance of prediction models. In addition, the performance of the proposed SBC sequence-based correction algorithm is evaluated by applying it to the model with the highest accuracy. Furthermore, a comparison of the models' training time and input-output delay are compiled.

3.1. CNN

The optimal length of the sliding window was determined by training the CNN model. As shown in Fig. 5, the highest prediction accuracy (88%) is achieved with a sliding window of 4 frames (66.7 ms), i.e., t~(t-3). Hence, the window length of 4 frames was used for the training of all models in this study. Fig. 6 shows the normalized confusion matrix of the CNN model, which summarizes its classification performance. This matrix represents the actual vs. predicted labels of the task, where the diagonal elements denote the percentage of correct predictions, and the off-diagonal elements represent the percentage of incorrect predictions. The CNN model achieved an overall accuracy of 88%.

3.2. LSTM

Six different hidden sizes of LSTM were trained in this study. The testing accuracy increases as the hidden size increases (Table 2). The testing accuracy converged to 91% when the hidden size was more than or equal to 128. The hidden size refers to the dimension of the hidden state. The results indicate that the hidden size should be over 128 to accurately depict the complexity of task recognition during disassembly. Fig. 7 shows the testing results for LSTM with the hidden size of 128.

3.3. GoogLeNet

Fig. 8 summarizes the testing results of GoogLeNet. The overall testing accuracy was 92%, higher than the 88% accuracy of CNN. More specifically, the GoogLeNet prediction accuracy range from 81 to 96 percent for each task, whereas CNN performance varies from 70 to 96 percent (Fig. 6). Significant increases in task recognition accuracy were observed in task 1: the cover disassembly, task 2: the hard disk drive disassembly, task 3: the fan disassembly, and task 6: the RAM disassembly. GoogLeNet outperformed CNN since, in contrast to the serial architecture of CNN, the inception layers of GoogLeNet allow for concurrent training of multiple convolutional and pooling layers.

GoogLeNet also outperformed LSTM even though it had a 91% overall accuracy. While there is only a 1% difference in the overall accuracy between the testing results of LSTM-HS128 (Fig. 7) and GoogLeNet (Fig. 8), their prediction accuracy for each individual task was substantially different. Although the GoogLeNet showed lower accuracies than the LSTM-HS128 in task 3: the fan disassembly and task 4:

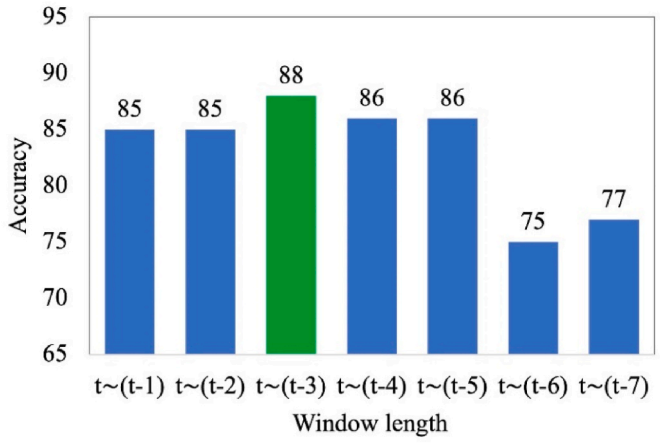


Fig. 5. Different window lengths and their corresponding accuracies.

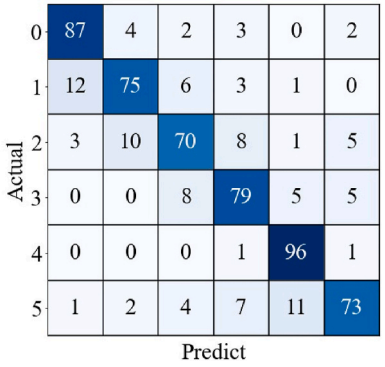


Fig. 6. The normalized confusion matrix of CNN on the testing set with 88% accuracy.

Table 2
Testing results of LSTM with different hidden sizes.

Hidden size	Testing accuracy
16	89%
32	89%
64	90%
128	91%
256	91%
512	91%

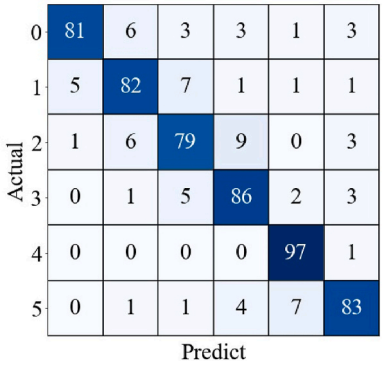


Fig. 7. The normalized confusion matrix of LSTM-HS128 on the testing set with 91% accuracy.

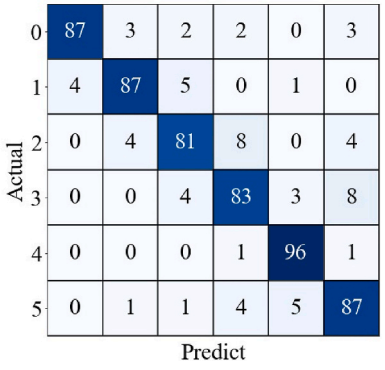


Fig. 8. The normalized confusion matrix of the GoogLeNet on the testing set with 92% accuracy.

the heat sink disassembly, significantly higher accuracies could be seen in task 0: the thumbscrew disassembly, task 1: the cover disassembly, task 2: the hard disk drive disassembly, and task 5: the RAM disassembly.

A possible reason for the superior performance of GoogLeNet over both the CNN and LSTM models could be attributed to its deeper and wider architecture, which allows it to capture complex features and their interrelationships more effectively than the CNN and LSTM models. The observed improvements in task recognition accuracy suggest that GoogLeNet is better able to distinguish between different disassembly tasks, which can ultimately lead to a more efficient and effective disassembly process.

3.4. Sequence-based correction algorithm

Since GoogLeNet demonstrated the highest performance, the proposed sequence-based correction algorithm was applied to GoogLeNet. The result is shown in Fig. 9. By utilizing the information of the known fixed task sequence and a majority voting mechanism to correct the task classification results given by GoogLeNet, the proposed SBC algorithm enhances the overall task recognition accuracy of GoogLeNet from 92% to 95%. More specifically, a significant increase in prediction accuracy was observed for every individual task. Except for task 2, all disassembly tasks achieved recognition accuracies above 90%. Furthermore, after applying the algorithm, there were only three possible prediction outcomes for a certain task: 1) the task is labeled correctly; 2) the task is wrongly identified as the one immediately preceding it; and 3) the task is wrongly identified as the one immediately following it.

3.5. Model training time and input-output delay

As mentioned in Methods section, the aforementioned models were trained and tested using data collected during the simulated desktop disassembly tasks. There are 375668, 93917, and 108710 samples in the training, validation, and testing data sets. All training and testing were performed on a desktop computer with a i9-10900K CPU @ 3.70 GHz, an NVIDIA Quadro RTX 4000 GPU, and 64 GB RAM. In addition to the testing accuracy, the training time and the input-output delay of each model were recorded (Table 3).

While GoogLeNet showed a higher testing accuracy (92%), it requires more training time and has a longer input-out time delay than CNN and LSTM-HS128 as it has a more complicated architecture. However, we argue that the input-output delay of GoogLeNet (5.27 ± 5.48 ms), which is less than the frame rate of IMU system (16.67ms), is quick enough to fulfill the need of the real-time application during collaborative human-robot e-waste disassembly processes.

Due to the lack of previous research in disassembly task recognition during HRC and the scarcity of publicly available IMU datasets, no benchmark exists for us to use as a comparison. Most comparable to our research is the work in Wen et al. (2019), which proposed an image-based 3D CNN model for human assembly task recognition during HRC. The recognition accuracy was 82% and the real-time implementation was not guaranteed. While the comparison may not be fair, the performance of our proposed model, especially GoogLeNet with 92% accuracy, still supports the feasibility of IMU-based deep learning models for task recognition during collaborative human-robot e-waste disassembly.

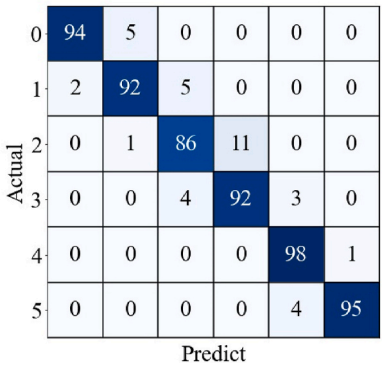


Fig. 9. The normalized confusion matrix of the GoogLeNet with SBC algorithm on the testing set. The overall accuracy is 95%.

Table 3
The testing accuracy, training time, and input-out delay of each model.

Model	Testing accuracy	Training time (s)	Input-output delay (ms)
CNN	88%	6491	0.66 ± 1.76
LSTM-HS128	91%	1265	0.26 ± 1.30
GoogLeNet	92%	12509	5.29 ± 4.23
GoogLeNet + SBC	95%	12509	(3000 + 5.53) ± 4.33

In addition to the model selection, we also introduced a sequence-based correction algorithm to further improve the task recognition performance. As shown in Table 3, after applying the SBC algorithm to GoogLeNet, its testing accuracy increased from 92% to 95%. The training time remained constant since the SBC algorithm did not affect the model training process.

While our proposed algorithm enhanced the testing accuracy, a significantly higher input-out delay was observed (~3s). The SBC algorithm was launched 3 s after the machine learning model, which resulted in a delay of about 3 s. However, it is worth noting that recognizing a disassembly task needs longer timespans than recognizing an action as a task often consists of several actions. In our experiment, each of the six tasks took more than 3 s to complete. That being said, there would not be any missed tasks during the ~3s delay. Furthermore, from the operational safety perspective, the delay in task recognition gives the robot assistant a buffer of time before executing any motions, preventing any concurrent movement of the human worker and the robot that might lead to safety issues such as surprise to the worker or even collisions. In summary, even though there is an input-output delay, but the improved accuracy of the proposed algorithm may demonstrate a worthwhile trade-off. It should be noted that the setting of 180 frames (3s) in the SBC algorithm was defined empirically for the proof-of-concept purpose. If the task changes, this setting might be drastically altered. However, our preliminary results reveal that the proposed sequence-based correction algorithm can be used to improve the accuracy for any process with a fixed sequence. The future research will focus on determining the optimal time setting and how it relates to the task's characteristics.

4. Conclusions

This study proposed a framework to recognize human tasks during robotic-assisted disassembly processes. The framework consists of two main elements: (1) a prediction architecture for the task recognition, and (2) a sequence-based correction algorithm. CNN, LSTM, and GoogLeNet architectures have been used for task recognition, and a two-rule tuning algorithm was proposed to improve the task recognition accuracy further.

An experimental study has been conducted on the disassembly of a desktop computer to collect data needed for accurate prediction of the disassembly tasks. Over 570,000 frames of motion data were collected from 22 participants. The highest task recognition accuracy was achieved by GoogLeNet on the IMU motion data. Furthermore, the proposed correction algorithm was able to improve the accuracy from 92% to 95%.

The study was limited by the lack of validation in a real-world job setting. For example, the performance of the model on individuals who

were not included in the study remains unknown and requires further testing. Furthermore, the effective integration of the proposed task recognition technique into real-world applications, such as HRC, needs further investigation. Several ways in which the study can be extended are as follows: First, a human-in-the-loop experiment can be carried out to evaluate the model performance in real-time. Moreover, when implementing HRC, concerns on human physical and mental safety need to be further discussed (Lu et al., 2022; Chen et al., 2022c). Second, instead of merely relying on IMU or image data, a hybrid model using data fusion may improve both the accuracy and robustness of task recognition during the disassembly process (Amorim et al., 2021). Finally, when there are more than one feasible disassembly sequences, the determination of the optimal disassembly sequence is crucial as it is a prerequisite for the proposed sequence-based correction algorithm.

Declaration of competing interest

The authors declare that no potential conflict of interest that could have appeared to influence the work reported in this paper.

Acknowledgment

This investigation was made possible by Award No. 2026276 from the National Science Foundation (NSF). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of NSF. The authors would like to thank Mustafa Ozkan Yerebakan and Shuyan Xia for their assistance with the data collection. We would also like to thank Yue Luo and Shihui Ruan for assistance with the preparation of Figs. 1 and 4.

References

Acquah, A.A., et al., 2019. Processes and challenges associated with informal electronic waste recycling at Agbogbloshie, a suburb of Accra, Ghana. Proc. Hum. Factors Ergon. Soc. Annu. Meet. 63 (1), 938–942.

Amorim, A., Guimares, D., Mendona, T., Neto, P., Costa, P., Moreira, A.P., 2021. Robust human position estimation in cooperative robotic cells. Robot. Comput. Integrated Manuf. 67, 102035.

Arivazhagan, S., Shebiah, R.N., Harini, R., Swetha, S., 2019. Human action recognition from RGB-D data using complete local binary pattern. Cognit. Syst. Res. 58, 94–104.

Breque, M., De Nul, L., Petridis, A., 2021. Industry 5.0: towards a Sustainable, Human-Centric and Resilient European Industry. European Commission, Directorate-General for Research and Innovation, Luxembourg, LU.

Y. Chen, Y. Luo, and B. Hu, "Towards next generation cleaning tools: factors affecting cleaning robot usage and proxemic behaviors design," Front. Electron., p. 14. doi: 10.3389/felec.2022.895001..

Chen, Y., Yang, C., Gu, Y., Hu, B., 2022b. Influence of mobile robots on human safety perception and system productivity in wholesale and retail trade environments: a pilot study. IEEE Trans. Human-Mach. Syst. 52 (4), 624–635.

Chen, Y., Luo, Y., Yerebakan, M.O., Xia, S., Behdad, S., Hu, B., 2022c. Human workload and ergonomics during human-robot collaborative electronic waste disassembly. In: 2022 IEEE International Conference on Human-Machine Systems (ICHMS), pp. 1–6.

Dallel, M., Havard, V., Baudry, D., Savatier, X., 2020. InHARD-industrial human action recognition dataset in the context of industrial collaborative robotics. In: 2020 IEEE International Conference on Human-Machine Systems (ICHMS), pp. 1–6.

Forti, V., Baldé, C.P., Kuehr, R., Bel, G., 2020. The Global E-Waste Monitor 2020. vol. 120. In: United Nations University (UNU), International Telecommunication Union (ITU) & International Solid Waste Association (ISWA), Bonn/Geneva/Rotterdam.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.

Hu, J., et al., 2018. pRNN: a recurrent neural network based approach for customer churn prediction in telecommunication sector. In: 2018 IEEE International Conference on Big Data (Big Data), pp. 4081–4085.

Hu, B., Li, S., Chen, Y., Kavi, R., Coppola, S., 2021. Applying deep neural networks and inertial measurement unit in recognizing irregular walking differences in the real world. Appl. Ergon. 96, 103414.

Kiruba, K., Shiloah, E.D., Sunil, R.R.C., 2019. Hexagonal volume local binary pattern (H-VLBP) with deep stacked autoencoder for human action recognition. Cognit. Syst. Res. 58, 71–93.

Koppula, H.S., Gupta, R., Saxena, A., 2013. Learning human activities and object affordances from rgb-d videos. Int. J. Robot Res. 32 (8), 951–970.

Koskimaki, H., Huikari, V., Siirtola, P., Laurinen, P., Roning, J., 2009. Activity recognition using a wrist-worn inertial measurement unit: a case study for industrial assembly lines. In: 2009 17th Mediterranean Conference on Control and Automation, pp. 401–405.

Koskimäki, H., Huikari, V., Siirtola, P., Röning, J., 2013. Behavior modeling in industrial assembly lines using a wrist-worn inertial measurement unit. J. Ambient Intell. Hum. Comput. 4 (2), 187–194.

Lara, O.D., Labrador, M.A., 2012. A survey on human activity recognition using wearable sensors. IEEE Commun. Surveys Tutorials 15 (3), 1192–1209.

Liu, J., Luo, J., Shah, M., 2009. Recognizing realistic actions from videos 'in the wild,. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1996–2003.

Lu, L., Xie, Z., Wang, H., Li, L., Xu, X., 2022. Mental stress and safety awareness during human-robot collaboration - Review. Appl. Ergon. 105, 103832.

Luo, Y., Zheng, H., Chen, Y., Giang, W.C.W., Hu, B., 2020a. Influences of smartphone operation on gait and posture during outdoor walking task. Proc. Hum. Factors Ergon. Soc. Annu. Meet. 64 (1), 1723–1727.

Luo, Y., Coppola, S.M., Dixon, P.C., Li, S., Dennerlein, J.T., Hu, B., 2020b. A database of human gait performance on irregular and uneven surfaces collected by wearable sensors. Sci. Data 7 (1), 1–9.

Mazhar, O., Navarro, B., Ramdani, S., Passama, R., Cherubini, A., 2019. A real-time human-robot interaction framework with robust background invariant hand gesture detection. Robot. Comput. Integrated Manuf. 60, 34–48.

Pfister, A., West, A.M., Bronner, S., Noah, J.A., 2014. Comparative abilities of Microsoft Kinect and Vicon 3D motion capture for gait analysis. J. Med. Eng. Technol. 38 (5), 274–280.

Roda-Sanchez, L., Garrido-Hidalgo, C., García, A.S., Olivares, T., Fernández-Caballero, A., 2021. Comparison of RGB-D and IMU-based gesture recognition for human-robot interaction in remanufacturing. Int. J. Adv. Des. Manuf. Technol. 1–13.

Sajedi, S., Liu, W., Eltouny, K., Behdad, S., Zheng, M., Liang, X., 2022. Uncertainty-assisted image-processing for human-robot close collaboration. IEEE Rob. Autom. Lett. 7 (2), 4236–4243.

Schuldt, C., Laptev, I., Caputo, B., 2004. Recognizing human actions: a local SVM approach. vol. 3. In: Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004, pp. 32–36.

Szegedy, C., et al., 2015. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9.

Tripathi, A., Mondal, A.K., Kumar, L., Prathosh, A.P., 2021. SCLAiR: supervised contrastive learning for user and device independent airwriting recognition. IEEE Sens. Lett. 6 (2), 1–4.

Wang, J., Liu, Z., Wu, Y., Yuan, J., 2012. Mining actionlet ensemble for action recognition with depth cameras. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1290–1297.

Wen, X., Chen, H., Hong, Q., 2019. Human assembly task recognition in human-robot collaboration based on 3D CNN. In: 2019 IEEE 9th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER), pp. 1230–1234.

Xu, W., Cui, J., Liu, B., Liu, J., Yao, B., Zhou, Z., 2021. Human-robot collaborative disassembly line balancing considering the safe strategy in remanufacturing. J. Clean. Prod. 324, 129158.

Zhang, M., Sawchuk, A.A., 2012. USC-HAD: a daily activity dataset for ubiquitous activity recognition using wearable sensors. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing, pp. 1036–1043.

Zhang, J., Wang, P., Gao, R.X., 2021. Hybrid machine learning for human action recognition and prediction in assembly. Robot. Comput. Integrated Manuf. 72, 102184.

Zhou, B., Andonian, A., Oliva, A., Torralba, A., 2018. Temporal relational reasoning in videos. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 803–818.

Zuidwijk, R., Krikke, H., 2008. Strategic response to EEE returns:: product eco-design or new recovery processes? Eur. J. Oper. Res. 191 (3), 1206–1222.