IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 12 May 2024, accepted 31 May 2024, date of publication 5 June 2024, date of current version 12 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3409645

RESEARCH ARTICLE

Action Recognition and Subsequent In-Depth Analysis for Improving the Time Efficiency of Bimanual Industrial Work

RYOTA TAKAMIDO AND JUN OTA, (Member, IEEE)
Research into Artifacts, Center for Engineering (RACE), School of Engineering, The University of Tokyo, Tokyo 113-8654, Japan

Corresponding author: Ryota Takamido (takamido@race.t.u-tokyo.ac.jp)

This work was supported in part by the New Energy and Industrial Technology Development Organization (NEDO) under Project JPNP20016 and in part by the Japan Society for the Promotion of Science (JSPS) KAKENHI under Grant 20H04261.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee of the School of Engineering, University of Tokyo.

ABSTRACT Although there are numerous studies that have developed human action recognition (HAR) algorithms, they have mostly focused on accurate recognition of actions; there is a lack of knowledge on analysis and interpretation of the recognition results for identifying the critical factor causing work delay. Further, from a technical standpoint, existing algorithms have difficulty dealing with missing objects during work processes. To overcome these two limitations, this study developed a new HAR algorithm for the recognition of bimanual actions of industrial workers, termed coordinate-BiLSTM with missing object information (C-BiLSTM+MO), and proposed a multi-regression model for conducting in-depth analysis of the recognition results. The proposed HAR algorithm was verified with experimental data from two typical industrial scenarios (pick-and-place, assembly-and-disassembly). The proposed multi-regression model was applied to the recognition results of these tasks and the data from existing bimanual action recognition datasets. The results revealed that the proposed HAR model could recognize the actions of both hands over 85% of the time, for tasks including when an object is missing or appearing, and each key component included in the proposed HAR model could significantly improve the recognition performance. Further, the proposed multi-regression model can explain over 50% of the variance of work time for all seven tasks. Notably, we clarified that the parameter of asymmetricity in the action of the two hands had a significant effect on the work delay for all tasks ($p<.01$). These results suggest the benefits of in-depth analysis of recognition results to improve time efficiency.

INDEX TERMS Human action recognition, human–object interaction, machine learning, performance analysis, time and motion study.

I. INTRODUCTION

Improving the efficiency of manual workers is beneficial in many industries. Previous studies have revealed that human workers spend up to approximately 50% of their work time on non-productive activities such as waiting [1]. Conventionally, to detect such wasteful actions during work processes, analysts perform a time and motion study (TMS) [2], [3], which decomposes workers' actions into several primitive actions, such as picking or transporting a product, and visualizes the entire work process as a time series of primitive actions to identify wasteful parts. In this context, the importance and effectiveness of visualizing work processes using TMS analysis have been demonstrated in various industries [4], [5], [6], [7], [8], [9]. However, in the traditional approach a considerable amount of time is required for

The associate editor coordinating the review of this manuscript and approving it for publication was Vicente Alarcon-Aquino.

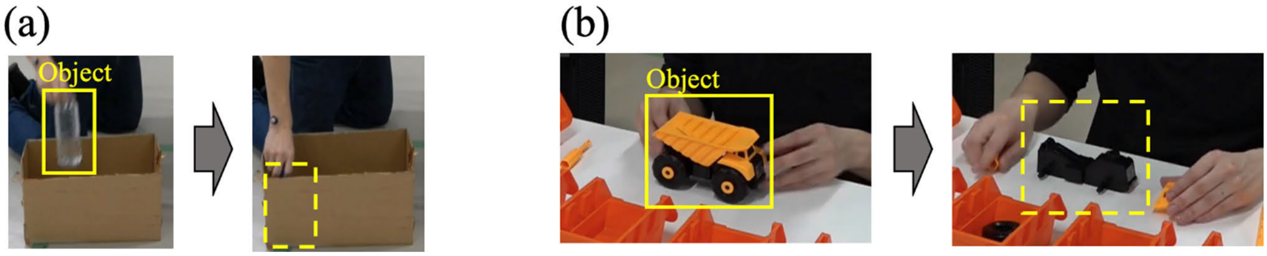

FIGURE 1. Missing (appearing) object problem for HOI detection in industrial scenario, (a) an object moves to behind the obstacle, (b) shape of an object changes during the working process.

analysts to manually assign action labels to all activities during the work process. Hence, the reduction in human resource requirement and TMS efforts is a critical issue.

One potential approach that is used to address this issue is TMS automation, which utilizes human action recognition (HAR) algorithms. Currently, studies on HAR, which aim to develop algorithms for classifying human motion data into several action labels, are the focus of considerable attention in the machine learning field [10]. Of the different HAR algorithms, sequence-to-sequence recognition algorithms, which receive the sequence input of human motion data such as joint position information and return sequence output of action labels, that best represent human motions at each time step (e.g., [11], [12]), can be used for TMS. Specifically, HAR algorithms for bimanual actions [13], [14], [15], [16], which can recognize the actions of both hands, separately, have high applicability because they can be used to capture the coordination between both hands during the working process; this is a key factor in TMS [2].

However, there are several problems associated with the aforementioned approaches that need to be addressed. The first and most critical problem is that although many HAR algorithms have been developed, specifically for industrial scenarios (e.g., [17], [18], [19], [20], [21], [22]), there is a lack of knowledge related to the analysis of recognition results to identify the cause of work delay and to evaluate time efficiency. Most previous studies aimed to achieve higher recognition accuracy, that is, "how to accurately recognize the worker's action." However, in the practical workflow analysis, in-depth analysis for identifying wasteful actions, which follows the accurate visualization of the entire process, is essential to improve the work efficiency [23]. Therefore, this lack of knowledge on "how to analyze the recognition results" decreases the effectiveness of the analysis of HAR algorithms. To identify wasteful actions and determine the relationship between different activities of the worker from the recognition results obtained using an HAR algorithm, in-depth analysis of the worker's activities at each time step and total work time should be conducted.

Additionally, existing methodologies have difficulty managing interactions with "missing objects" during the working process. That is, most industrial work includes the process of changing the states of the objects in the environment (shape or position), the work inevitably includes the object disappearing and appearing as shown in Fig. 1. This makes it difficult to apply existing algorithms, particularly the existing bimanual HAR algorithms [13], [14], [15], [16] that explicitly use object information for considering complex human–object interactions (HOI). Even if the above in-depth analysis problems could be solved, the limitation on the range of target tasks owing to the missing object problem reduces the effectiveness of HAR-algorithm analysis in practical applications. Therefore, in addition to developing an effective in-depth analysis methodology, the missing object problem should be addressed for wide application of the HAR algorithm.

II. RELATED WORK

A. TIME AND MOTION STUDY

Methods for measurement and evaluation of workers' activities has been a significant research topic. The idea of a "time and motion study," which decomposes a worker's activity into several sub-activities and identifies waste, originated from the work of Taylor, and has been further developed by others such as Gilbreth and Bedaux [24], [25]. As reported in several case studies, this concept is presently being utilized for improving work efficiency [4], [5], [6], [7], [8], [9], [26], [27], [28], [29], [30]. For example, Moktadir et al. [7] improved productivity of the manufacturing process of leather products by decomposing it into sixty sub-processes and identifying points for improvement. Duran et al. [27] achieved a 53% improvement of time efficiency by applying TMS to the process of making tea glass models by identifying redundancies from the calculation of the standardized time for each process. The traditional TMS consists of the following procedures: (1) recording all information related to the job, (2) breaking down or decomposing the job into elements, (3) examining these elements and determining the sample size, (4) recording the time to perform each elemental task using a stop-watch, (5) assessing the speed of working, (6) converting the observed time to basic time, (7) determining the allowances, and (8) determining the standard time [30]. Evidently, performing all these procedures manually requires a considerable amount of time, which can be a potential barrier for the introduction of TMS.

B. HAR ALGORITHM FOR INDUSTRIAL APPLICATIONS

Although HAR algorithms have usually been applied to analyze daily human activity using common datasets [31], many studies have attempted to develop an HAR algorithm for manufacturing lines [21], [22], hand crafting [17], [18], [19], [20], [32], [33], [34], [35], [36], [37], [38], [39], warehouse picking [40], construction sites [41], [42], [43], [44], agriculture [45], and human–robot (machine) interaction [46], [47], [48], [49], [50], [51], and other industrial tasks such as spraying [52]. The introduction and usage of HAR algorithms is a key concept for the modern manufacturing industry [53].

From a technical perspective, sequential recognition is usually used in industry scenarios to capture the details of a workflow [32]. While statistical modeling methods such as the hidden Markov model [54] or pattern matching [55] were mainly used in the period from the early 2000s to the mid-2010s, deep learning methods such as convolutional neural network (CNN)- [19], recurrent neural network (RNN)- [36], and graph neural network (GNN)-based methods [37] have become dominant in recent years. The difficulty related to the use of HAR in industrial scenarios is that there are numerous activities, the durations of these activities vary, and the performance rate of an activity does not remain constant [39]. Therefore, previous studies developed specialized HAR algorithms with custom features for each industrial scenario. For example, Hernandez et al. [40] proposed a work-monitoring system for a warehouse picking task with a specialized neural network having two RNN-based processing streams of the human skeleton and image information around both hands. Zhang et al. [34] also developed two stream networks to process both human and object information for considering the complex interaction between them in the assembly scenario. Additionally, Yan and Wan [21] recently developed an automatic monitoring system for car assembly tasks involving several workers by combining the YOLO V3 detector and VGG16, a kind of CNN-based architecture for image data processing [56]. Finally, Gammulle et al. [51] captured the hand motion of assembly operators using Kinect and classified their intention using a Gaussian mixture model to achieve efficient human–machine interaction.

Further, as a more specific HAR algorithm for capturing detailed work actions, some recent studies developed a HAR algorithm and datasets for human bimanual action recognition [13], [14], [15], [16]. While most conventional HAR algorithms assigned one single action label to each time point, those studies attempted to assign separate labels for each hand of the human worker. In particular, Dreher et al. [13] first addressed this problem by developing the ‘‘bimanual action dataset,’’ which consists of 540 human bimanual actions. They also proposed a HAR algorithm for use with the dataset. To the best of our knowledge, this is the only dataset that contains separate label information for each hand at each time step. After Dreher's work, several studies proposed various algorithms with that dataset and updated the highest recognition score. Morais et al. [14] proposed ASSIGN (asynchronous-sparse interaction graph networks) that contains two layers of spatio-temporal graph networks for detecting the HOI with multiple time spans. Xing et al. proposed PGCN (pyramid graph convolutional network), which employs a pyramidal encoder–decoder architecture consisting of an attention-based graph convolution network. This system represents the 2D or 3D spatial relations of humans and objects from the detection results in video data as a graph [16]. Finally, Qiao et al. [15] developed 2G-GCN (two-level geometric feature-informed graph convolutional network) that considers both visual and geometric features of humans and objects using two graph networks. By assigning separate labels to each hand at each time frame using these methods, key factors for executing a bimanual tasks such as hand roles and symmetry [57] can be extracted.

However, as mentioned in Section I, the common problem of these studies is a lack of knowledge of how to analyze and evaluate the recognition results to identify the cause of a work delay or the increase in time efficiency. Although many studies have focused on the development of accurate recognition algorithms, from the viewpoint of industrial applications, improvements in the work process ‘‘after recognition’’ is also essential. Therefore, a gap exists between the focus of current studies (recognition part) and requirements of practical applications. Further, the existing bimanual HAR algorithms [13], [14], [15], [16] assume the possibility of explicit detection of the objects; hence, recognition of the HOI with missing or appearing objects during the working process is problematic. This complicates the application of HAR algorithms to industrial tasks that frequently contain missing or appearing objects such as assembly.

III. METHODS

A. AIM AND APPROACH OF THIS STUDY

Based on the literature reviewed, this study aims to address two critical issues that hinder the HAR-algorithm application for automating the TMS analysis. The first issue is the lack of knowledge on ‘‘how to analyze the recognition results,’’ which complicates making specific improvements and re-designing the workflow to increase the time efficiency. To address this issue, we build a new statistical model that can explain the variance of work time using the recognized bimanual label sequence information from an HAR algorithm. The model contains some key variables from the aspects of bimanual hand coordination such as the asymmetry of the two hands or waiting time for one hand; therefore, we can identify the key variable responsible for the improvement of the time efficiency by referencing the weights of each variable.

Second, to address the missing or appearing object problem, which limits the range of target tasks of the analysis using an HAR algorithm, we propose a new HAR algorithm called coordinate-BiLSTM with missing object information (C-BiLSTM+MO) that can recognize an HOI with missing objects during the work process. The objective of this algorithm is to detect the object's missing or appearance point in the image as the clue for identifying certain specific actions

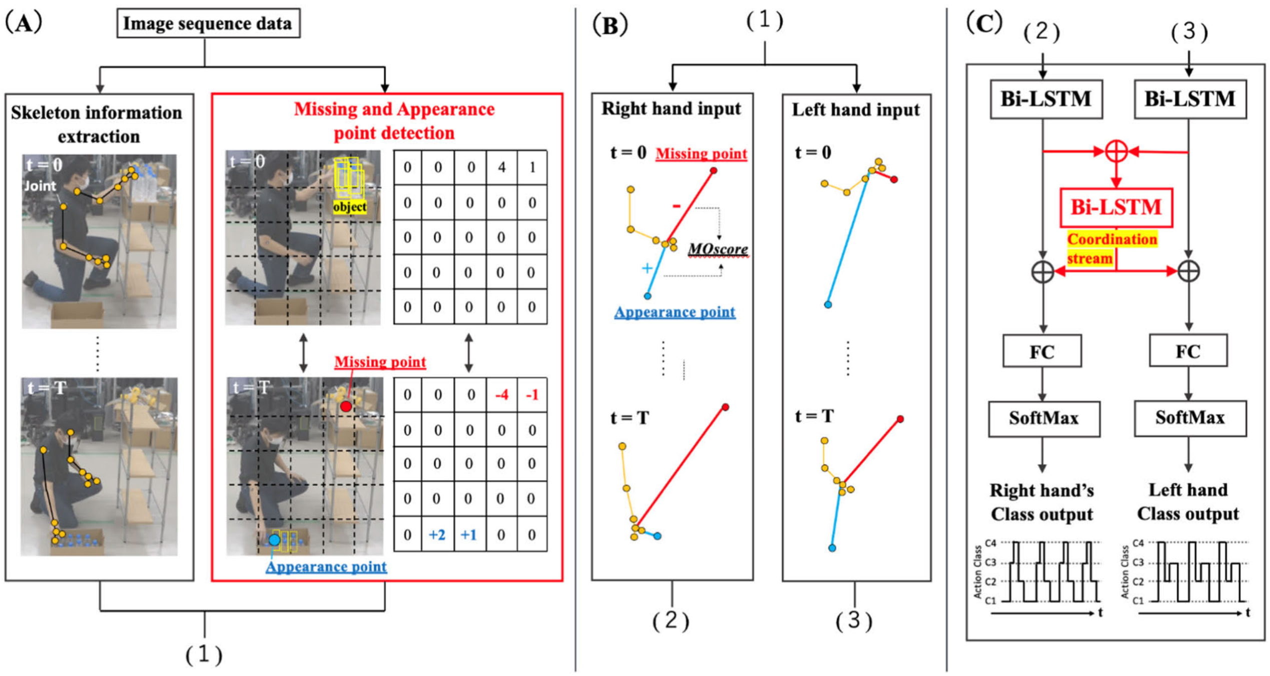

FIGURE 2. Overview of proposed method: C-BiLSTM+MO. (A) Input image sequence data are processed, and the human skeleton and object's missing (appearance) points are extracted; (B) Generation of the input data of each hand including MOscore, and (C) performing HAR for each hand using the proposed network architecture.

(for example, removing a part). This reduces the required information needed to identify human-object interactions, and enable us to extend the range of target tasks.

B. PROPOSED HAR ALGORITHM FOR ADDRESSING THE MISSING OBJECT PROBLEM

Fig. 2 presents an overview of the proposed HAR, namely the C-BiLSTM+MO algorithm. This algorithm converts the image sequence of a worker's hand motion into two sequences of corresponding action labels for each hand, in a frame-by-frame manner. The basic architecture is a stacked BiLSTM that can deeply process the time-series data along both forward and backward directions and accurately represent the complex temporal dependencies of the time-series data [58], [59]. Lack of information on the coordination between two hands negatively impacts performance of bimanual HAR [15]; hence, our algorithm has a specific BiLSTM layer called the "coordinate stream" for considering the complex coordination between the hands. To reduce the complexity and the cost for the measurement system, we adopt skeleton data extracted from a single RGB camera image as human motion information.

Similar to existing bimanual HAR algorithms [13], [14], [15], [16], the proposed model uses both human motion and object information as input data. However, while existing methods continuously detect and use object information such as a name or bounding box, our approach is to only use the initial and end states of the object for calculating where the object disappeared or appeared. Specifically, we considered that the fact of "object missing or appearing" itself increases the probability of some specific worker action. If the worker's hand is close to the object missing point, it suggests a high probability of an action that changes the position or shape of an object (Fig. 1). Although identifying "interaction points" such as the midpoint between human and target objects is the common approach for human–object interaction recognition (e.g., [60]), we extend it to deal with the missing and appearance objects. Information on the interaction with a missing object is defined as the Missing Objects' Interaction Score (MOscore) and used as a feature variable for HAR (right side of Fig. 2 (A) and (B)). This reduction of the requirements for implementation will facilitate industrial applications of the algorithm.

C. PROCESS OF C-BILSTM+MO

In the process of C-BiLSTM+MO, the image sequence of the motion of the manual worker was first recorded using an RGB camera. The recorded images were then processed using skeleton recognition software, and two-dimensional joint position information of the two shoulders, elbows, wrists, hands, thumbs, and the hand tips were extracted (Fig. 2 (A)).

In addition, to evaluate and quantify the HOI with missing and appearing objects, the missing objects and appearance points were determined using the YOLOv4 object recognition algorithm [61]. Fig. 3 shows the details of the calculation process. First, the YOLOv4 detector defines the bounding boxes of all the objects included in the image sequence in $T \times S$ seconds at the beginning and end of the video

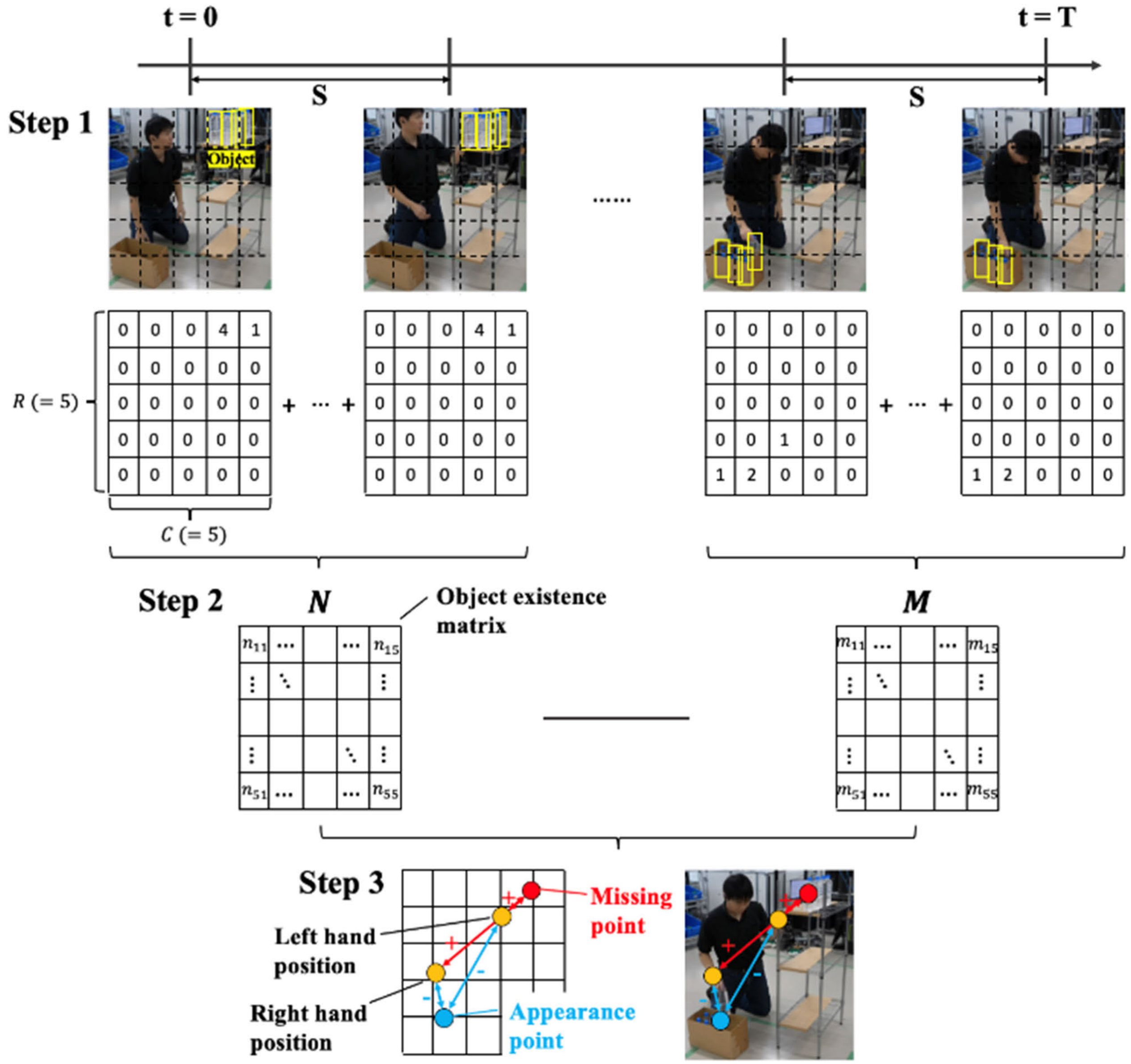

FIGURE 3. Process of calculating missing and appearance points. First, the object existence matrices that represent the position and number of target objects included in each image frame are calculated (Step 1) and summarized in the initial and end phases (Step 2). Missing objects and appearance points were detected by subtracting both matrices (Step 3).

(Fig. 3. Step 1). Here, T represents the time length of the image sequence, and S represents the ratio of the initial and end phases to the entire time sequence, which was set to 0.05 in this study. Subsequently, the number of objects included in each $R \times C$ divided region (set as $R = C = 5$, in this study) of each image was counted and summarized for the duration, and the existence matrices N and M were calculated (Fig. 3, Step 2). If the detector can find the target object, the element in the matrix becomes positive, and the value increases in response to the frequency and number of objects found during the $T \times S$ timespan. The difference of the two matrices (L) represents changes in the status (missing/appearance) of the target objects because of the entire work process:

$$L = N - M = \begin{bmatrix} n_{11} - m_{11} & \cdots & n_{1C} - m_{1C} \\ \vdots & \ddots & \vdots \\ n_{R1} - m_{R1} & \cdots & n_{RC} - m_{RC} \end{bmatrix} = \begin{bmatrix} l_{11} & \cdots & l_{1C} \\ \vdots & \ddots & \vdots \\ l_{R1} & \cdots & l_{RC} \end{bmatrix}. \quad (1)$$

The plus and minus signs of the indices l_{rc} indicate the missing appearance of the object during the work process. That is, if target objects that were not seen in the initial span appear in the end phase, the value of index l_{rc} becomes negative. If the position of the object does not change during the entire work process, the value becomes zero; hence, the existence of fixed objects is unaffected. In addition, because matrix L can be calculated even when there are no target objects at either the initiation or end of the work, it can deal with actions that include missing or appearing objects, such as placing the objects into an opaque container. The missing object and appearance points $\boldsymbol{P_m}(x_m, y_m), \boldsymbol{P_a}(x_a, y_a)$ are defined by calculating the weighted mean of each index in matrix L:

$$\begin{cases} \boldsymbol{P_m}\left(x_m, y_m\right) = \dfrac{\sum l_{rc} \times \boldsymbol{p_{rc}}}{\sum l_{rc}} l_{rc} > 0 \\ \boldsymbol{P_a}\left(x_a, y_a\right) = \dfrac{\sum \left|l_{rc}\right| \times \boldsymbol{p_{rc}}}{\sum \left|l_{rc}\right|} l_{rc} < 0, \end{cases} \quad (2)$$

where $\boldsymbol{p_{rc}} = (x_{rc}, y_{rc})$ is the position vector of the center of $R \times C$ divided region.

After calculating the missing and appearing points, we calculated a *MOscore* that represents how close the hand is to an object's missing or appearance points. Given $\boldsymbol{h_r}$ and $\boldsymbol{h_l}$ are the two-dimensional positions of the right and left hands, respectively, they are calculated using the following equation (Fig. 3, Step 3):

$$\begin{cases} MOscore_r = \dfrac{1}{1 + \sqrt{\left|\boldsymbol{P_m} - \boldsymbol{h_r}\right|}} - \dfrac{1}{1 + \sqrt{\left|\boldsymbol{P_a} - \boldsymbol{h_r}\right|}} \\ MOscore_l = \dfrac{1}{1 + \sqrt{\left|\boldsymbol{P_m} - \boldsymbol{h_l}\right|}} - \dfrac{1}{1 + \sqrt{\left|\boldsymbol{P_a} - \boldsymbol{h_l}\right|}} \end{cases} \quad (3)$$

If the worker's hand is close to the missing point and away from the appearance point, the value of the *MOscore* approaches +1 and its differentiation becomes positive. Here, even if there is no difference in the positions of the hands, the *MOscore* can differ depending on the position of the missing and appearance points. This enables us to discriminate similar actions, such as attaching a screw versus detaching a screw with the same pose, because the former occurs near the missing point ($MOscore > 0$), and the latter occurs near the appearance point ($MOscore < 0$). If one of the missing or appearance points cannot be defined, the *MOscore* is calculated using only one term in (3), which can be defined.

Finally, the network input variable was generated and action recognition was performed for each hand (Fig. 2 (C)). The input variables were composed of 2-dimensional position information of the aforementioned joints (six joints for each hand), the *MOscore*, and its differentiation. These variables were normalized and combined, and two sets of input data were generated for the recognition of the left- and right-hand action. These processing flows produced two predicted action-label sequences for each hand at each time step using a frame-by-frame analysis.

D. PROPOSED MODEL FOR ANALYZING THE OBTAINED BIMANUAL LABEL SEQUENCE

This study also conducted subsequent in-depth analysis of the obtained label sequence data to identify the cause of work delay and evaluate time efficiency. To quantitatively evaluate the relationship between work delay and its potential factors, we adopted a multiple regression model for the analysis. Specifically, we proposed the multiple regression model to explain the variance of work time among several repeated operations with the following four feature variables, which are considered as potential factors affecting the working efficiency in TMS [2], [3], [23], [24].

(1) Asymmetry in the action of the two hands (A): This variable represents the asymmetry in the movement left and right hands. The importance of effective hand coordination for working efficiency was emphasized in TMS as "both hands should move and stop at the same time" [2]. To quantify the A, all labels are roughly divided into either "dynamic label" or "static label." Subsequently, the summation of time durations having asymmetric movement intensity (dynamic vs. static) was counted and defined as this feature. If the timing of initiation and end of both hand actions differs, the value of this variable increases.

(2) Waiting time (W): This represents the summation of the duration that both hands have the "idle" label indicating that the hand does not do anything regarding the target process. The waiting time also depends on the timing of completion of the previous process [62]; hence, if this parameter makes a large contribution to the explanation of the working time, the engineers should verify the connection with the previous process.

(3) Total number of right-hand actions (N_r): This represents the total number of right-hand actions included in the obtained label sequence. A previous study revealed that productivity is improved by eliminating unnecessary movements and simplifying individual tasks [29]. If the work time increases as a result of a worker performing extra actions, this parameter makes a large contribution to the work time variance. Specifically, this was defined as the number of time frames that have right-hand action labels that are from the ones in the previous frame.

(4) Total number of left-hand actions (N_l): This represents the total number of left-hand actions included in the obtained label sequence. By evaluating the effect of both N_r and N_l on the working time, we can obtain more detailed insights about the current process and identify points for improvement.

Using these features, we constructed the following multiple regression model to predict the deviation from the average working time $\bar{WT}$ among repeated operations:

$$\begin{aligned} WT - \bar{WT} = w_1(A - \bar{A}) + w_2(W - \bar{W}) + w_3\left(N_r - \bar{N_r}\right) \\ + w_4\left(N_l - \bar{N_l}\right), \end{aligned} \quad (4)$$

where $w_1 - w_4$ indicates the multiple regression coefficient of each independent variable. When there were several data generated from different task domains, this multiple regression model was fitted to each domain separately, and identified the cause of work time variance in each domain. If a large asymmetricity in the actions of the two hands action causes a delay in a specific task, the contribution and significance of A for

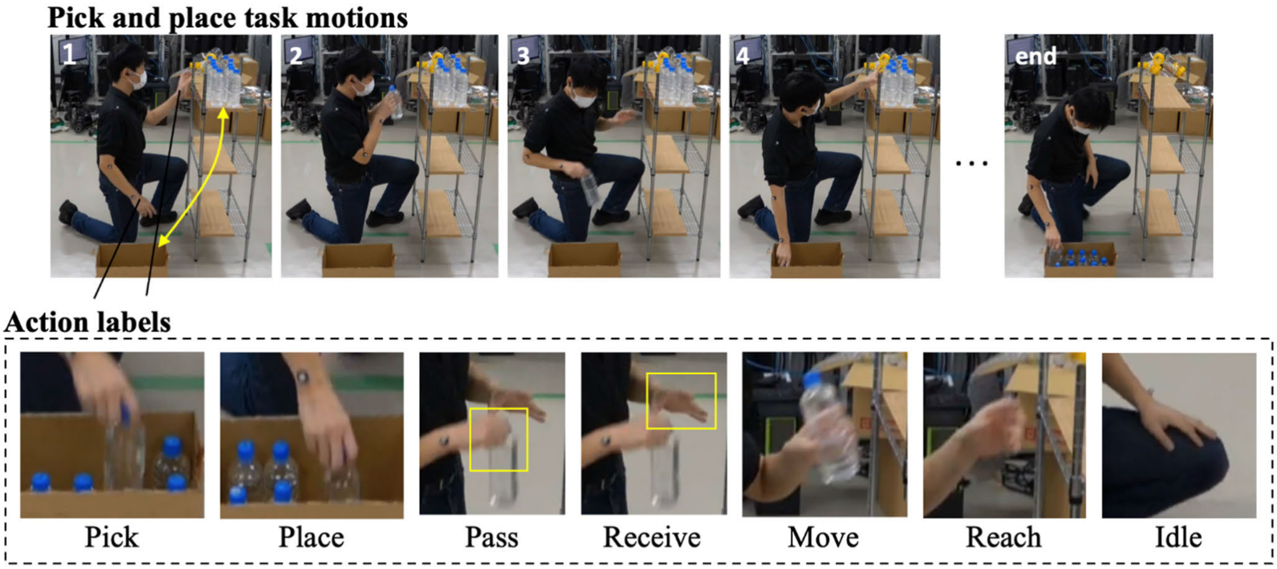

FIGURE 4. Details of the pick-and-place task. The worker tried to move all target objects from a shelf to a container, and vice versa. The motion of the worker is continuously classified into seven labels at each video frame.

work time prediction becomes larger, and if the waiting time (W) is the main factor, it becomes the significant factor in the task. Therefore, through this multiple regression analysis with the Equation (4), we can obtain more in-depth insights into the task than through visualization of the recognition action sequences alone.

IV. VERIFICATION OF PROPOSED METHODS

A. VALIDATION OF THE PROPOSED HAR ALGORITHM

Although the dataset collected by Dreher et al. [13] for human bimanual motion recognition is available, it does not include missing object information because the data were collected under ideal situations where the object was continuously captured in the image. Moreover, the task used to generate the dataset was relatively simpler than that in a practical industrial situation that involves actions such as unscrewing and removing only one screw. Therefore, we conducted experiments to collect more industry-oriented data for the verification of the proposed approach.

The target task in the experiment consists of two representative industrial scenarios. One is the pick-and-place task, in which the target object from a container is manually picked up and moved to a shelf or vice versa (Fig. 4). The pick-and-place task is the basic object of interactive work such as arranging products on a shelf or warehouse picking that can be observed in many industries. As this task involves a large change in the object position, it usually includes objects that appear and disappear. In addition, we tested the proposed method on a more complex task: assembly and disassembly tasks with miniature cars (Fig. 5). Product assembly and disassembly are representative manual tasks in industries that require dexterous coordination between the hands and various interactions with objects. In this task, the human worker tried to assemble the miniature car by attaching 13 parts to the body frame or disassemble it by removing the parts from the frame. As this task involves a major change in object shape, it also includes missing and appearing objects. This study was approved by the Ethics Committee of the School of Engineering, University of Tokyo.

1) DATA COLLECTION

In the pick-and-place task, one male worker carried a total of ten target objects (500 ml plastic bottles) for one trial, and the position of the lined-up objects on the shelf (DOSHISHA, WSD6012-4) was varied for three patterns (upper, middle, and lower layers of the shelf). The heights of the three layers were 100 cm, 75 cm, and 15 cm for the upper, middle, and lower layers, respectively. During the target task, the worker was allowed to use both hands and asked to minimize work time as much as possible. The worker could change and adjust the initial position of the shelf and container for each trial to increase the work efficiency.

The motion was recorded using an RGB video camera (SONY, FDR-AX45A) at 30 fps from nine different angles, and 18 trials (three object positions × 2 task patterns × 3 iterations) were performed for each angle. As a result, we recorded 162 pick-and-place motions (9 angles × 18 trials) with 1620 individual object-transportation actions. The data from three trials were excluded owing to recording errors. The mean required time for completing the task is 18.3 s (549 frames). After completing the recording, seven action labels (Fig. 4) were manually assigned to each hand in all the video frames (87,840 frames).

For the assembly and disassembly task, the same worker assembled or disassembled a miniature car (THEXIN, MY5032A Cement Mixer) 60 times. In the assembly task, the worker attached parts such as screws or tires to the body frame until all of them were attached. Parts such as screws or tires

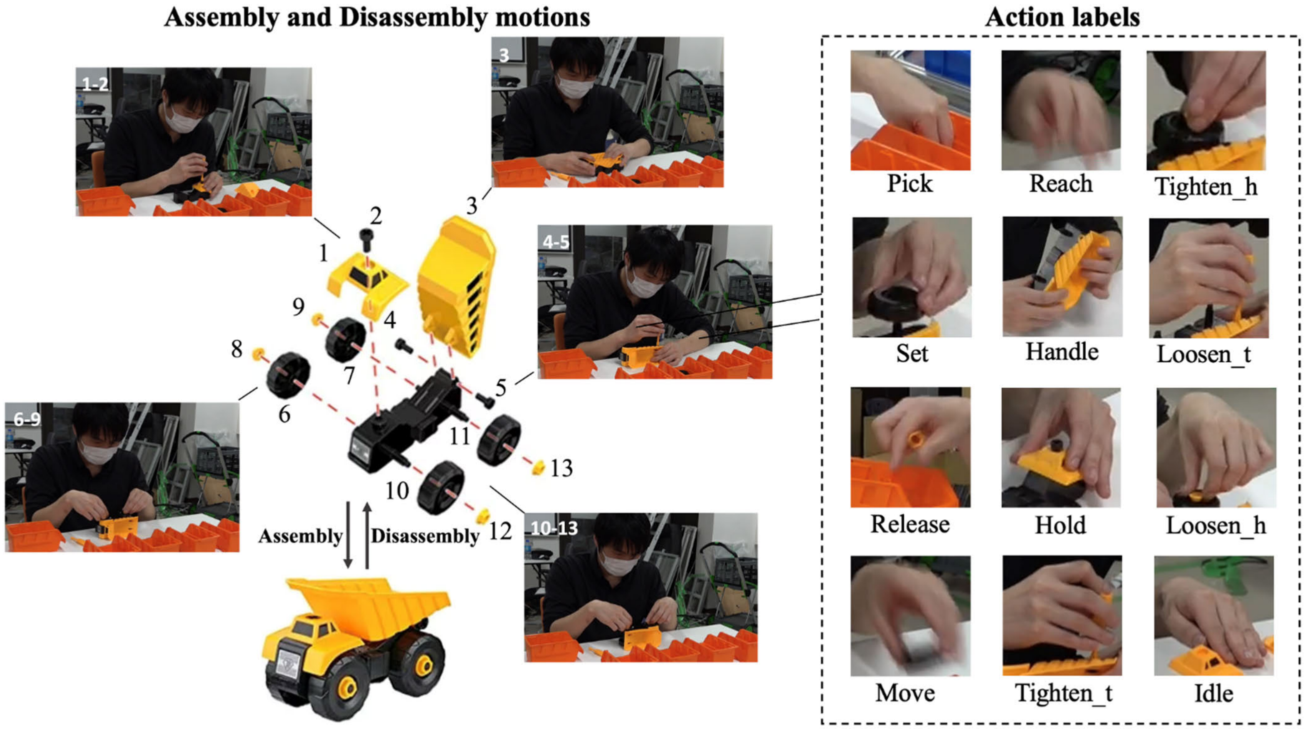

FIGURE 5. Details of the assembly and disassembly tasks. The worker tries to attach (or detach) a total of thirteen parts on the frame. The worker's action was classified into twelve action labels.

were initially stored in containers, the body frame was set on the center of the desk, and the screwdriver was also placed on the desk. In the disassembly task, the worker removed all parts from the car and placed the parts into the container. The worker could change and adjust the arrangement of the initial position of the body frame, containers, and screwdrivers for each trial. To increase the validity of the camera angle differences, three synchronized cameras with different angles were used for recording and changing the position every ten trials. Synchronized cameras were used to reduce the number of iterations and worker fatigue. A total of 360 videos were recorded (six angles × 3 cameras × 10 trials × 2 task patterns (assembling and disassembling)), and among them, the data of one trial were excluded owing to unsuccessful recording. The mean recording time and number of frames for one trial were 35.7 s and 1,073 frames, respectively. Twelve action labels were manually assigned to both hands in all the frames (385,207 frames).

2) DATA PROCESSING

To generate the input data of the proposed model based on the data collected from the experiment, all recorded images were processed using skeleton tracking software (VisionPose, NEXT-SYSTEM), and the time series of two-dimensional joint position data and the differentiation of shoulders, elbows, wrists, hands, thumbs, and index fingers were collected. Subsequently, a low-pass filter with a cutoff frequency of 6 Hz was used to smooth the data. To reduce the effect of differences in the recording angles, the skeleton data were normalized. Twelve input variables containing human skeletal information were generated for both hands.

Additionally, the MOscore and its differentiation were calculated across all time spans and normalized to generate the input data for the network by following the process mentioned in the previous section (Fig. 3). The recognition of the target object was performed using a pretrained YOLOv4 detector (CSP-DarkNet-53 framework with the COCO dataset) on the MATLAB deep learning toolbox (MATLAB 2022b, The Math Works, Inc.). The detector attempted to detect the position of the "bottle" object in the image and return its bounding box in the pick-and-place task. In addition, because there was no existing object detection model for the target object (miniature car) in Experiment 2, we manually trained the YOLO-v4 detector to detect the target object using the image sequences of six videos with different camera angles (approximately 6,000 images in total). We assigned "object" label and bounding box information only when all parts were attached to the frame, therefore, if some parts were removed from the body frame, the detector would not find the object, and the result would be "object missing". The detection threshold was set to 0.2, and the durations of the initial and end sequences were set to be $S = 0.05$ (5%) for each experiment. The calculated MOscore and its differentiation were also normalized and combined with the joint position information to generate input data with 28 dimensions (14 variables × 2 hands) for the proposed model for all trials.

3) TRAINING AND PERFORMANCE EVALUATION

We trained the proposed HAR model and evaluated its recognition performance based on a generated dataset of the input variables and action labels. During the training process, a leave-one-angle-out cross-validation was performed for the pick-and-place task, and a leave-one-synchronized-pair-out (three different angles) cross-validation was performed for the assembly task. The number of hidden layers for each Bi-LSTM cell was set as 128. The batch size was set to 64, the number of maximum epochs was set to 100, and the learning rate was set to 1e-4. To avoid overfitting of the training data, the training process was monitored and manually interrupted as required. The proposed model was trained to minimize the summation of the two cross-entropy losses. The training time for the proposed model was approximately 15 min and 45 min for each task obtained by running three GPUs (NVIDIA RTX A5000) in parallel on a MATLAB platform with a deep learning toolbox and parallel computing toolbox. Recognition performance was evaluated by calculating the average value of the accuracy, macro-averaged precision (maP), recall (maR), and F1-score (maF1) by considering the differences in the frequency of occurrence of each action.

Additionally, to verify the effect and role of each component included in the proposed model (C-BiLSTM+MO) such as usage of missing object information, we compared its recognition performance with that of the following three simplified models:

(1) Single LSTM with no object information (single LSTM): This model processes human motion information using a single LSTM layer and does not contain coordination stream or object information.

(2) Coordinate-LSTM with no object information (C-LSTM): This model processes human motion information using two LSTM layers with a coordinate stream. By comparing the above results, we can confirm the effectiveness of the coordination stream.

(3) Coordinate-LSTM with object information (C-LSTM+MO): this model processes both human motion and object information. By comparing this model with the above results, we can confirm the effectiveness of the missing object information.

We verified the effectiveness of the proposed model by comparison with the simplified models rather than with existing algorithms because there is no existing method that meets the following criteria: (1) assigns action labels for each hand separately, (2) does not require continuous object detection, (3) recognizes in a sequence-to-sequence (frame-by-frame) manner. However, insights regarding the critical component, namely, the missing object information would increase the performance of the HAR algorithm in practical industrial scenarios through comparison with the above models.

B. VALIDATION OF THE MULTIPLE REGRESSION MODEL FOR WORK TIME VARIANCE

To verify the proposed multiple regression model, we fitted the model to two different datasets. One was the bimanual label sequence data of Dreher's datasets (bimanual action dataset [13]) and the other was the recognized label sequence data from the above recognition tests. The former was used to test whether the model could explain the variance of the work time in the simplified task when the correct label sequence information was given, namely, to verify the model itself. The latter was used to verify the applicability of the model in a more practical context by applying it to noisy, non-perfect recognition results with more industry-oriented scenarios.

Specifically, for the Dreher datasets [13], we selected five tasks (cooking, cooking with bowls, cereals, hard drive, free hard drive) with an average work time exceeding 10 s for the nine different tasks included in the datasets. The data of each task included the label sequence information of both hands and total work time for each of the 10 trials made by the six workers. For the pick-and-place and assembly and disassembly tasks, the output recognition results from the proposed HAR model were directly used for model fitting. The data from one of the three synchronized cameras for the assembly and disassembly task were used to avoid duplication. Table 1 presents the mean work time and its standard deviation, and minimum and maximum work time for each task. Table 2 shows the definition of static and dynamic labels for calculation of A. The model of (4) was fitted by a least-squares method for each task to verify whether the proposed statistical model could explain the variance of work time based on the bimanual label sequence information. Variance inflation factors (VIF) were used to detect the multicollinearity of the independent variables with the value of 10 as the criteria. The adjusted R-squared value, RMSE (root mean squared error) and significance of the fitting result (p-value) were calculated to evaluate the fitness of the multi-regression model for each task; the multiple regression coefficient, t-value and p-value of each parameter were also calculated to identify the dominant factor causing work time variance in each task.

V. RESULTS AND DISCUSSIONS

A. VERIFICATION OF THE PROPOSED HAR MODEL

Tables 3 and 4 present the comparisons of the recognition results of the seven models, and Fig. 6 and 7 show the confusion matrices for each hand and examples of the predicted label sequences obtained using the proposed method. The total accuracies for the right- and left-hand action recognition are 0.89 and 0.87 for the pick-and-place task and 0.91 and 0.91 for the assembly task, respectively. Therefore, the proposed algorithm can well recognize the bimanual actions in both the industrial scenarios even though it does not use the explicit and continuous object information.

By comparing the proposed method (C-BiLSTM+MO) with its simplified models (Single LSTM, C-LSTM and C-LSTM+MO), we can confirm the effect of setting each component. First, from the comparison of the results of single LSTM and C-LSTM, the increase in total accuracy is in the range of 10–14% and 18–22% for each task, respectively. This demonstrates the importance of considering the

TABLE 1. **Work tme variance of the target tasks.**

Dataset	Task	Sample size	Mean work time (s)	Work time range (s)
Bimanual actions dataset [12]	Cooking	60	16.3±6.7	8.4-34.5
	Cooking with bowls	60	14.5±3.4	8.4-20.5
	Cereals	60	16.5±3.2	12.0-26.0
	Hard drive	60	17.0±4.3	10.4-29.9
	Free hard drive	60	17.4±4.5	7.9-26.7
Our experiments	Pick-and-place	159	18.3±2.7	11.5-24.3
	Assembly and disassembly	120	35.7±6.6	25.2-50.5

TABLE 2. **Definition of the static and dynamic labels in each dataset.**

Dataset's name	Static labels (0)	Dynamic labels (1)
Bimanual action datasets [12]	idle, retreat, hold, pour, saw, screw, drink	approach, lift, place, cut, hammer, stir, wipe
Pick-and-place	idle, receive	move, reach, pick, place, pass
Assembly and disassembly	idle, hold, loosen_h, loosen_t, release, set, tighten_h, tighten_t	handling, move, reach, pick

coordination information between both hands for bimanual HAR. Previous studies also revealed a large decrease in accuracy when there is no information regarding the interaction between them [15]. Additionally, on comparing C-LSTM and C-LSTM+MO, the total accuracy increased by 6% and 1–2% for each task, respectively, by adding the MOscore and its differentiation. Therefore, the information about the missing and appearing object increases the recognition performance in each industrial scenario. This is the novel finding revealed by the results of this study. The effect of missing object information in the assembly task may be lower because of the range of motion of the hand was smaller in the task; there was no clear difference in the MOscore for each action. Finally, the introduction of bidirectional architecture improved the recognition accuracy by 4–5% and 7–12% for each task. Hence, considering the long-term interaction of bimanual action can increase recognition performance, particularly for a complex task such as manual assembly. Although this study adopts two typical tasks in the practical industrial scenario, the insights with regard to the effects of each component can be transferred to other tasks.

As a case where the proposed method did not work, the "set" action to attach parts to the body frame by contacting them exhibited the worst recognition accuracy (52.4%). As shown in Figure 6, the action was frequently confused with the "hold" action to fix the body frame so that it did not move when attaching parts with the other hand. Because both actions were performed near the appearance point with slow speed movements, the potential reason is the high similarity of those actions from both aspects of MOscore and movement of body joints. In addition, errors tended to increase for the action labels with fewer observations (e.g., "idle" label of the right hand in the pick and place task). Finally, errors also increased in timing of transitions between actions

TABLE 3. Recognition results of each model in the pick-and-place task.

	Right hand				Left hand			
HAR model	Total Accuracy	maP	maR	maF1	Total Accuracy	maP	maR	maF1
Single LSTM	0.69	0.74	0.58	0.65	0.62	0.56	0.52	0.54
C-LSTM	0.79	0.76	0.75	0.76	0.76	0.76	0.70	0.73
C-LSTM+MO	0.85	0.84	0.82	0.83	0.82	0.82	0.76	0.79
C-BiLSTM+MO (Proposed)	**0.89**	**0.88**	**0.86**	**0.87**	**0.87**	**0.87**	**0.83**	**0.85**

TABLE 4. Recogntion results of each model in the assembly and disassemhly task.

	Right hand				Left hand			
HAR model	Total Accuracy	maP	maR	maF1	Total Accuracy	maP	maR	maF1
Single LSTM	0.55	0.57	0.48	0.52	0.65	0.56	0.39	0.46
C-LSTM	0.77	0.79	0.70	0.74	0.83	0.73	0.73	0.73
C-LSTM+MO	0.79	0.81	0.72	0.76	0.84	0.75	0.75	0.75
C-BiLSTM+MO (Proposed)	**0.91**	**0.91**	**0.89**	**0.90**	**0.91**	**0.87**	**0.90**	**0.88**

(e.g., "reach" to "pick"). This may be because of the difficulty in manually assigning consistent labels for the ambiguous "mid-term action".

B. VERIFICATION OF THE PROPOSED MULTI-REGRESSION MODEL

Table 5 presents the goodness of fit of the proposed model to each target task. The proposed multi-regression model could explain more than 50% of the variance of work time (R-squared value > 0.5 for all tasks). Fig. 8 shows the relationships between the predicted work delay from mean value (WT - $\bar{WT}$) and the true value. As shown in the figure and table, the proposed model can well explain the variance of the work time for each repeated operation using the bimanual label sequence information, even if noise causes imperfect recognition results (87–89% accuracy

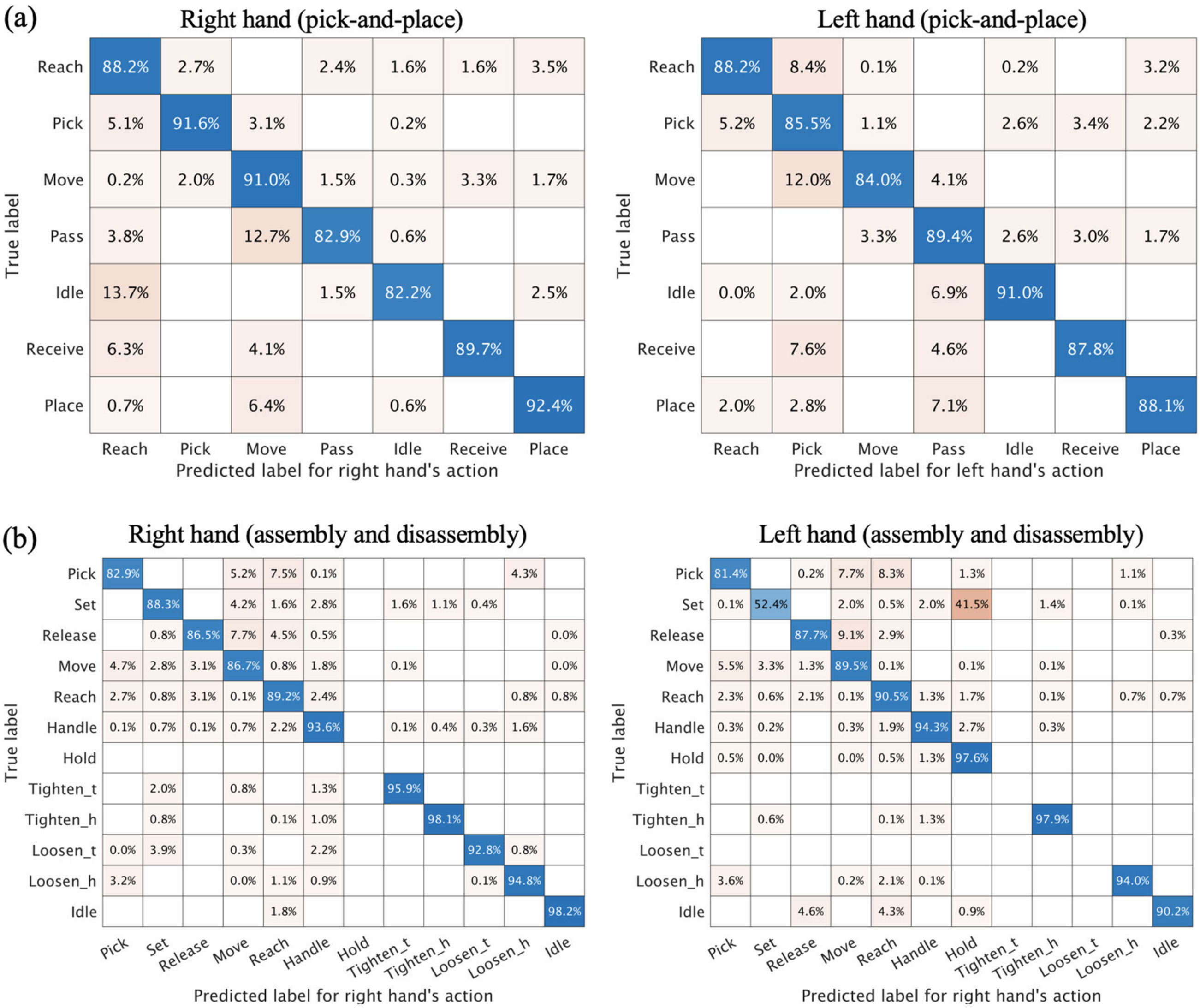

FIGURE 6. Confusion matrixes for each hand recognition task. (a) pick-and-place task (b) assembly-and-disassembly task.

for pick-and-place, 91% for assembly and disassembly task).

To get more detailed insights, Table 6 shows the coefficient, t-value, and p-value of each feature variable of TMS in each task. As there is no variance in the number of actions of each hand in the cooking tasks, $\boldsymbol{N_r}$ and $\boldsymbol{N_l}$ were excluded from the analysis. The results of the multicollinearity test with VIF showed that no variables exceeded the criterion value. In Table 6, the asymmetricity of both hands (A) have significant effects on all even tasks (p <0.01). This suggests that as mentioned in the conventional TMS [2], the coordination of both hands is a key factor for work efficiency. Remarkably, as the coefficients of A are above 1.0 for most tasks, the duration of the asymmetric phase of both hands results in a further work delay in the entire process. This may be due to the additional time that the ''preceding'' hand waits until the other hand to recovers from the misalignment between them.

Further, waiting time (W) is also a significant factor for all tasks. Although there are no other processes before and after the target task, the difference in time interval between the signal to start and the initiation of movement was represented as the W and reflected in the overall work time. Finally, the number of actions of both hands ($\boldsymbol{N_r}$ and $\boldsymbol{N_l}$) have a significant effect on the task of cooking with bowls and cereals, but not in other tasks. The reason for relatively minimal effects may be that increased primitive actions sometimes have a positive effect on work efficiency. For example, when a portion of ''idle'' labels are split into other action labels, the overall work time will decrease. Therefore, it may be important to consider what type of movements increased or decreased. It can be seen from the result that these parameters have a negative coefficient in the task of cooking with bowls, and negative coefficient in cereals.

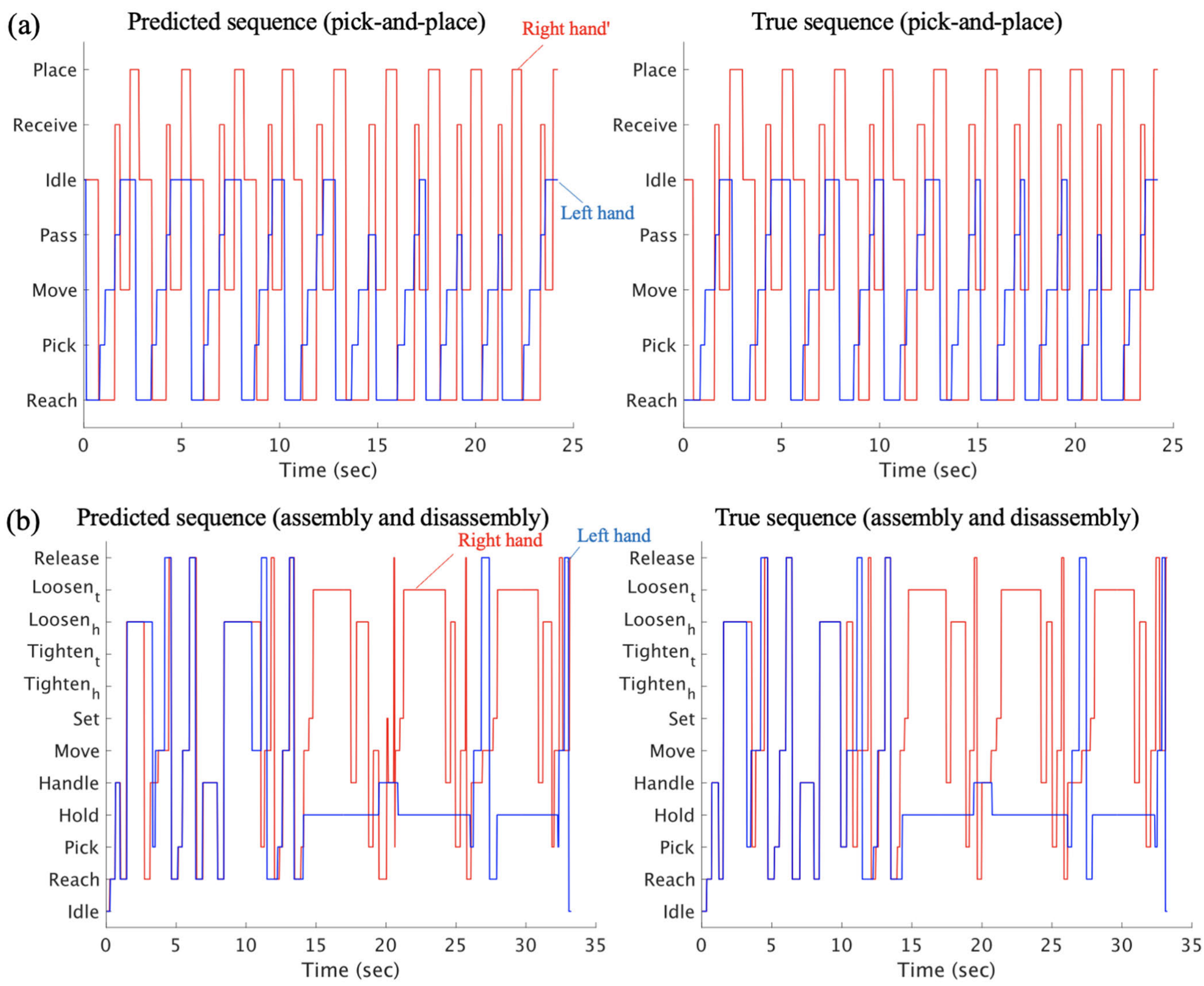

FIGURE 7. Predicted action label sequences for each hand by the proposed model (left figure) and true sequence (right figure). (a) pick-and-place task, (b) assembly-and-disassembly task. Results of one test trial (disassembly) are shown as an example of the action sequence.

TABLE 5. Results oe the multi-regression analysis for the bimanual label sequencl of each task. **: $\beta < 01$.

Dataset	Task	adjusted R-squared value	RMSE (s)	*p*-value
Bimanual actions dataset [12]	Cooking	0.98	0.92	**
	Cooking with bowls	0.98	0.50	**
	Cereals	0.90	0.99	**
	Hard drive	0.60	2.83	**
	Free hard drive	0.55	3.08	**
Our experiments	Pick-and-place	0.76	1.41	**
	Assembly and disassembly	0.54	4.41	**

Finally, the proposed model showed relatively low R-squared value for the free hard drive task, and assembly and disassembly task (Table 5). The common feature of these two tasks is the large variation in the execution time within the same action label, that is, these two tasks include the action to turn the screw using the tool to install it in its

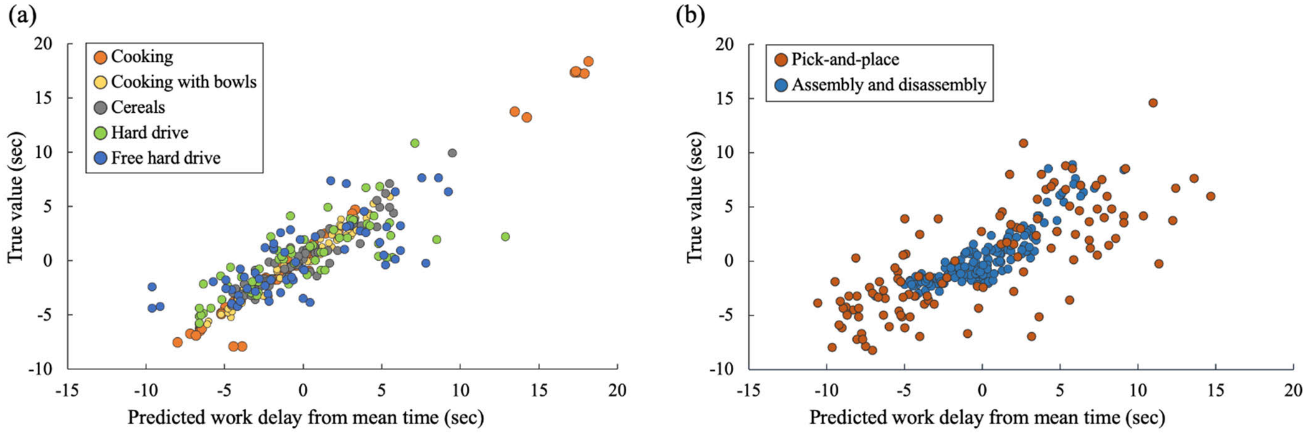

FIGURE 8. Relationship between the predicted work time delay from mean time by the proposed model (a) the task in the Bimanual actions dataset [12], (b) the task performed in the experiment.

TABLE 6. Contribution of each feature variable por multi-regression analysis.

	Task	Feature value	coefficient	t-value	p-value
Bimanual actions dataset [12]	Cooking	Asymmetricity (*A*)	1.09	8.04	**
		Waiting time (*W*)	1.09	51.3	**
	Cooking with bowls	Asymmetricity (*A*)	1.08	12.45	**
		Waiting time (*W*)	1.03	49.64	**
		Number of right- hand actions (N_r)	-1.06	-4.41	**
		Number of left -hand actions (N_l)	0.08	0.33	n.s.
	Cereals	Asymmetricity (*A*)	1.62	9.58	**
		Waiting time (*W*)	1.08	14.88	**
		Number of right -hand actions (N_r)	0.16	2.58	*
		Number of left-hand actions (N_l)	0.28	4.95	**
	Hard drive	Asymmetricity (*A*))	1.93	4.82	**
		Waiting time (*W*)	1.81	4.27	**
		Number of right-hand actions (N_r)	0.03	0.22	n.s.
		Number of left-hand actions (N_l)	0.55	0.97	n.s.
	Free hard drive	Asymmetricity (*A*))	2.14	5.45	**
		Waiting time (*W*)	0.95	2.29	*
		Number of right-hand actions (N_r)	0.54	1.47	n.s.
		Number of left-hand actions (N_l)	0.18	0.86	n.s.
Our experiments	Pick-and-place	Asymmetricity (*A*))	1.93	4.82	**
		Waiting time (*W*)	1.81	4.27	**
		Number of right-hand actions (N_r)	0.03	0.22	n.s.
		Number of left-hand actions (N_l)	0.55	0.97	n.s.
	Assembly and disassembly	Asymmetricity (*A*))	2.14	5.45	**
		Waiting time (*W*)	0.95	2.29	*
		Number of right-hand actions (N_r)	0.54	1.47	n.s.
		Number of left-hand actions (N_l)	0.18	0.86	n.s.

**: p<.01, n.s.: non-significance (p>.05)

proper place. Because this action requires precise alignment of screws and tools, the worker sometimes took longer time to adjust the position of them, and sometimes did not. Therefore, this increases the variance of motion time within same label and may decrease the prediction accuracy of the model.

C. GENERAL DISCUSSION

In summary, insights about the dominant factor causing a work delay can be first obtained by the in-depth analysis developed in this study. Although some information can be obtained from visualizing the work activities as a series of label sequences, by conducting further analysis we can increase the interpretability and explainability of the recognition results for the users. Moreover, the results of this study suggest the importance of bimanual recognition for understanding the detailed process of industrial manual workers because the asymmetricity of both hands' action was a dominant factor for all target tasks in this study. Although there are few studies that address the problem of separate recognition of both hands' action as discussed in a previous study [13], it is necessary to identify the cause of a work delay and improve efficiency. From this aspect, the method of dealing with missing objects proposed in this study can contribute to expand applicable tasks from one in an ideal experimental room to more practical environments.

From general aspects, discussing the applicability of the proposed method to other task domains is worthwhile. Within many industrial tasks, the task that satisfies the following conditions has a high possibility: (a) the environment allows to capture the visual (image) information of workers' upper body, (b) one or both the locations of missing and appearance points of the target objects can be identified, (c) the missing and appearance points of the target object is caused by a worker's action. From these aspects, the proposed method is highly applicable to some industrial tasks such as packaging or food processing at the same location (e.g., a workbench). Conversely, to increase the generalizability of the proposed method, further improvements such as integrating the force or sound information to address the case where vision sensor is not available is needed.

Finally, limitations exist in applying the proposed method to the real-world applications although this study addressed some gaps to use the HAR algorithm for TMS analyses in industrial scenarios. First, because real scenarios inevitably include undesirable cases for action recognition, such as including unknown (anormal) actions or class imbalance, the proposed method should be extended to address those cases by adding some processes of unknown action rejection [63] or data augmentation [64]. Moreover, from a practical perspective, it is desirable to introduce unsupervised or semi-supervised learning methods [65], [66] because manual label assignment for preparing the training dataset is time-consuming, particularly for the bimanual recognition task, which can be a barrier for application. Finally, as mentioned previously, the statistical model should also be extended to cover more variated works. For example, by introducing the explainable-AI technique that emphasizes the interpretability and explainability of the model may be useful as an alternative for dealing with more complicated causes in the real industrial scenarios [67]. If sufficient amount of data can be collected from the work environment, more detailed findings could be obtained by replacing the simple regression model with such a complex deep-learning model.

VI. CONCLUSION

This study addressed two critical issues in the existing HAR algorithm for practical industrial applications. To overcome the difficulty of dealing with missing and appearing objects, we proposed a new HAR model, C-BiLSTM+MO, and verified its recognition performance in two experiments with typical industrial scenarios (pick-and-place, assembly-and-disassembly). Further, to address the lack of knowledge of how to analyze recognition results, we proposed the multi-regression model with four key features in TMS to identify the dominant factor causing a time delay. The results revealed that our proposed multi-regression model can explain over 50% of variance of the work time in the seven different tasks. Specifically, the asymmetricity in the actions of the two hands has a significant effect on work delay in all tasks, which suggests the importance and effectiveness of bimanual recognition and subsequent in-depth analysis in industrial scenarios. The limitations of the methodologies used in this study should be addressed to make further improvements for enhancing it suitability for practical applications.

ACKNOWLEDGMENT

This article is one of the achievements of joint research, and forms jointly owned copyrighted material of the ROBOT Industrial Basic Technology Collaborative Innovation Partnership.

REFERENCES

[1] M. C. Gouett, C. T. Haas, P. M. Goodrum, and C. H. Caldas, "Activity analysis for direct-work rate improvement in construction," *J. Construction Eng. Manage.*, vol. 137, no. 12, pp. 1117–1124, Dec. 2011, doi: 10.1061/(asce)co.1943-7862.0000375.

[2] R. M. Barnes, *Motion and Time Study: Design and Measurement of Work*. Hoboken, NJ, USA: Wiley, 1991.

[3] F. E. Meyers and J. R. Stewart, *Motion and Time Study for Lean Manufacturing*. Upper Saddle River, NJ, USA: Prentice-Hall, 2002.

[4] K. S. Al-Saleh, "Productivity improvement of a motor vehicle inspection station using motion and time study techniques," *J. King Saud Univ., Eng. Sci.*, vol. 23, no. 1, pp. 33–41, Jan. 2011, doi: 10.1016/j.jksues.2010.01.001.

[5] N. Saibani, A. A. Muhamed, M. F. Maliami, and R. Ahmad, "Time and motion studies of manual harvesting methods for oil palm fruit bunches: A Malaysian case study," *Jurnal Teknologi*, vol. 74, no. 3, pp. 77–83, May 2015, doi: 10.11113/jt.v74.4555.

[6] A. Kunz, M. Zank, T. Nescher, and K. Wegener, "Virtual reality based time and motion study with support for real walking," *Proc. CIRP*, vol. 57, pp. 303–308, Jan. 2016, doi: 10.1016/j.procir.2016.11.053.

[7] M. A. Moktadir, S. Ahmed, F. T. Zohra, and R. Sultana, "Productivity improvement by work study technique: A case on leather products industry of Bangladesh," *Ind. Eng. Manage.*, vol. 6, no. 1, 2017, Art. no. 1000207, doi: 10.4172/2169-0316.1000207.

[8] C. Prakash, B. P. Rao, D. V. Shetty, and S. Vaibhava, "Application of time and motion study to increase the productivity and efficiency," *J. Phys., Conf. Ser.*, vol. 1706, no. 1, Dec. 2020, Art. no. 012126, doi: 10.1088/1742-6596/1706/1/012126.

[9] O. Bongomin, J. I. Mwasiagi, E. O. Nganyi, and I. Nibikora, "Improvement of garment assembly line efficiency using line balancing technique," *Eng. Rep.*, vol. 2, no. 4, Apr. 2020, Art. no. e12157, doi: 10.1002/eng2.12157.

[10] N. Gupta, S. K. Gupta, R. K. Pathak, V. Jain, P. Rashidi, and J. S. Suri, "Human activity recognition in artificial intelligence framework: A narrative review," *Artif. Intell. Rev.*, vol. 55, no. 6, pp. 4755–4808, Aug. 2022, doi: 10.1007/s10462-021-10116-x.

[11] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 677–691, Apr. 2017, doi: 10.1109/TPAMI.2016.2599174.

[12] Y. Hu, X.-Q. Zhang, L. Xu, F. Xian He, Z. Tian, W. She, and W. Liu, "Harmonic loss function for sensor-based human activity recognition based on LSTM recurrent neural networks," *IEEE Access*, vol. 8, pp. 135617–135627, 2020, doi: 10.1109/ACCESS.2020.3003162.

[13] C. R. G. Dreher, M. Wächter, and T. Asfour, "Learning object-action relations from bimanual human demonstration using graph networks," *IEEE Robot. Autom. Lett.*, vol. 5, no. 1, pp. 187–194, Jan. 2020, doi: 10.1109/LRA.2019.2949221.

[14] R. Morais, V. Le, S. Venkatesh, and T. Tran, "Learning asynchronous and sparse human-object interaction in videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16036–16045, doi: 10.1109/CVPR46437.2021.01578.

[15] T. Qiao, Q. Men, F. W. B. Li, Y. Kubotani, S. Morishima, and H. P. H. Shum, "Geometric features informed multi-person human-object interaction recognition in videos," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Oct. 2022, pp. 474–491.

[16] H. Xing and D. Burschka, "Understanding spatio-temporal relations in human-object interaction using pyramid graph convolutional network," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2022, pp. 5195–5201, doi: 10.1109/IROS47612.2022.9981771.

[17] Q. Jiang, M. Liu, X. Wang, M. Ge, and L. Lin, "Human motion segmentation and recognition using machine vision for mechanical assembly operation," *SpringerPlus*, vol. 5, no. 1, p. 1629, Dec. 2016, doi: 10.1186/s40064-016-3279-x.

[18] W. Tao, Z.-H. Lai, M. C. Leu, and Z. Yin, "Worker activity recognition in smart manufacturing using IMU and sEMG signals with convolutional neural networks," *Proc. Manuf.*, vol. 26, pp. 1159–1166, Jan. 2018, doi: 10.1016/j.promfg.2018.07.152.

[19] Z. Wang, R. Qin, J. Yan, and C. Guo, "Vision sensor based action recognition for improving efficiency and quality under the environment of industry 4.0," *Proc. CIRP*, vol. 80, pp. 711–716, Jan. 2019, doi: 10.1016/j.procir.2019.01.106.

[20] M. Al-Amin, W. Tao, D. Doell, R. Lingard, Z. Yin, M. C. Leu, and R. Qin, "Action recognition in manufacturing assembly using multimodal sensor fusion," *Proc. Manuf.*, vol. 39, pp. 158–167, Jan. 2019, doi: 10.1016/j.promfg.2020.01.288.

[21] Z. Wang and J. Yan, "Multi-sensor fusion based industrial action recognition method under the environment of intelligent manufacturing," *J. Manuf. Syst.*, vol. 74, pp. 575–586, Jun. 2024, doi: 10.1016/j.jmsy.2024.04.019.

[22] J. Yan and Z. Wang, "YOLO v3 + VGG16-based automatic operations monitoring and analysis in a manufacturing workshop under industry 4.0," *J. Manuf. Syst.*, vol. 63, pp. 134–142, Apr. 2022, doi: 10.1016/j.jmsy.2022.02.009.

[23] N. Kumar, S. S. Hasan, K. Srivastava, R. Akhtar, R. K. Yadav, and V. K. Choubey, "Lean manufacturing techniques and its implementation: A review," *Mater. Today, Proc.*, vol. 64, pp. 1188–1192, Jan. 2022, doi: 10.1016/j.matpr.2022.03.481.

[24] D. R. Towill, "Industrial engineering the Toyota production system," *J. Manage. Hist.*, vol. 16, no. 3, pp. 327–345, Jun. 2010, doi: 10.1108/17511341011051234.

[25] M. Hazarika, U. S. Dixit, and J. P. Davim, "History of production and industrial engineering through contributions of stalwarts," in *Manufacturing Engineering Education*. Hull, U.K.: Chandos Publishing, 2019, pp. 1–29, doi: 10.1016/B978-0-08-101247-5.00001-0.

[26] M. N. B. C. Ani and S. A. B. A. Hamid, "Analysis and reduction of the waste in the work process using time study analysis: A case study," *Appl. Mech. Mater.*, vol. 660, pp. 971–975, Oct. 2014, doi: 10.4028/www.scientific.net/amm.660.971.

[27] C. Duran, A. Cetindere, and Y. E. Aksu, "Productivity improvement by work and time study technique for Earth energy-glass manufacturing company," *Proc. Econ. Finance*, vol. 26, pp. 109–113, Jan. 2015, doi: 10.1016/S2212-5671(15)00887-4.

[28] P. H. A. Cury and J. Saraiva, "Produção de lentes orgânicas no Pólo industrial de manaus," *Gestão Produção*, vol. 25, no. 4, pp. 901–915, Jul. 2018, doi: 10.1590/0104-530x2881-18.

[29] M. Waseem, U. Ghani, T. Habib, S. Noor, and T. Khan, "Productivity enhancement at molding compound manufacturing plant by applying time and motion analysis," *Mehran Univ. Res. J. Eng. Technol.*, vol. 40, no. 4, pp. 761–774, Oct. 2021, doi: 10.22581/muet1982.2104.07.

[30] R. Pisuchpen and W. Chansangar, "Modifying production line for productivity improvement: A case study of vision lens factory," *Songklanakarin J. Sci. Technol.*, vol. 36, no. 3, pp. 345–357, 2014.

[31] C. Chen, R. Jafari, and N. Kehtarnavaz, "UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 168–172, doi: 10.1109/ICIP.2015.7350781.

[32] C. Chen, T. Wang, D. Li, and J. Hong, "Repetitive assembly action recognition based on object detection and pose estimation," *J. Manuf. Syst.*, vol. 55, pp. 325–333, Apr. 2020, doi: 10.1016/j.jmsy.2020.04.018.

[33] M.-A. Zamora-Hernández, J. A. Castro-Vargas, J. Azorin-Lopez, and J. Garcia-Rodriguez, "Deep learning-based visual control assistant for assembly in industry 4.0," *Comput. Ind.*, vol. 131, Oct. 2021, Art. no. 103485, doi: 10.1016/j.compind.2021.103485.

[34] J. Zhang, P. Wang, and R. X. Gao, "Hybrid machine learning for human action recognition and prediction in assembly," *Robot. Comput.-Integr. Manuf.*, vol. 72, Dec. 2021, Art. no. 102184, doi: 10.1016/j.rcim.2021.102184.

[35] A. Malaisé, P. Maurice, F. Colas, F. Charpillet, and S. Ivaldi, "Activity recognition with multiple wearable sensors for industrial applications," in *Proc. 11th Int. Conf. Adv. Comput.-Hum. Interact. (ACHI)*, Mar. 2018, pp. 1–7.

[36] N. Shen, Z. Feng, J. Li, H. You, and C. Xia, "Action fusion recognition model based on GAT-GRU binary classification networks for human–robot collaborative assembly," *Multimedia Tools Appl.*, vol. 82, no. 12, pp. 18867–18885, May 2023, doi: 10.1007/s11042-022-14123-0.

[37] C. Chen, X. Zhao, J. Wang, D. Li, Y. Guan, and J. Hong, "Dynamic graph convolutional network for assembly behavior recognition based on attention mechanism and multi-scale feature fusion," *Sci. Rep.*, vol. 12, no. 1, p. 7394, May 2022, doi: 10.1038/s41598-022-11206-8.

[38] P. Lou, J. Li, Y. Zeng, B. Chen, and X. Zhang, "Real-time monitoring for manual operations with machine vision in smart manufacturing," *J. Manuf. Syst.*, vol. 65, pp. 709–719, Oct. 2022, doi: 10.1016/j.jmsy.2022.10.015.

[39] A. Mastakouris, G. Andriosopoulou, D. Masouros, P. Benardos, G.-C. Vosniakos, and D. Soudris, "Human worker activity recognition in a production floor environment through deep learning," *J. Manuf. Syst.*, vol. 71, pp. 115–130, Dec. 2023, doi: 10.1016/j.jmsy.2023.08.020.

[40] J. Hernandez, G. Valarezo, R. Cobos, J. W. Kim, R. Palacios, and A. G. Abad, "Hierarchical human action recognition to measure the performance of manual labor," *IEEE Access*, vol. 9, pp. 103110–103119, 2021, doi: 10.1109/ACCESS.2021.3095934.

[41] J. Ryu, J. Seo, H. Jebelli, and S. Lee, "Automated action recognition using an accelerometer-embedded wristband-type activity tracker," *J. Construct. Eng. Manage.*, vol. 145, no. 1, Jan. 2019, Art. no. 04018114.

[42] J. Zhang, L. Zi, Y. Hou, M. Wang, W. Jiang, and D. Deng, "A deep learning-based approach to enable action recognition for construction equipment," *Adv. Civil Eng.*, vol. 2020, pp. 1–14, Nov. 2020, doi: 10.1155/2020/8812928.

[43] Z. Li and D. Li, "Action recognition of construction workers under occlusion," *J. Building Eng.*, vol. 45, Jan. 2022, Art. no. 103352, doi: 10.1016/j.jobe.2021.103352.

[44] H. Guo, Z. Zhang, R. Yu, Y. Sun, and H. Li, "Action recognition based on 3D skeleton and LSTM for the monitoring of construction workers' safety harness usage," *J. Construct. Eng. Manage.*, vol. 149, no. 4, Apr. 2023, Art. no. 04023015.

[45] A. Anagnostis, L. Benos, D. Tsaopoulos, A. Tagarakis, N. Tsolakis, and D. Bochtis, "Human activity recognition through recurrent neural networks for human–robot interaction in agriculture," *Appl. Sci.*, vol. 11, no. 5, p. 2188, Mar. 2021, doi: 10.3390/app11052188.

[46] X. Li, "Human–robot interaction based on gesture and movement recognition," *Signal Process., Image Commun.*, vol. 81, Feb. 2020, Art. no. 115686, doi: 10.1016/j.image.2019.115686.

[47] Y. Zhang, K. Ding, J. Hui, J. Lv, X. Zhou, and P. Zheng, "Human-object integrated assembly intention recognition for context-aware human–robot collaborative assembly," *Adv. Eng. Informat.*, vol. 54, Oct. 2022, Art. no. 101792, doi: 10.1016/j.aei.2022.101792.

[48] N. Kozamernik, J. Zaletelj, A. Košir, F. Šuligoj, and D. Bračun, "Visual quality and safety monitoring system for human–robot cooperation," *Int. J. Adv. Manuf. Technol.*, vol. 128, nos. 1–2, pp. 685–701, Sep. 2023, doi: 10.1007/s00170-023-11698-2.

[49] D. Moutinho, L. F. Rocha, C. M. Costa, L. F. Teixeira, and G. Veiga, "Deep learning-based human action recognition to leverage context awareness in collaborative assembly," *Robot. Comput.-Integr. Manuf.*, vol. 80, Apr. 2023, Art. no. 102449, doi: 10.1016/j.rcim.2022.102449.

[50] T. B. Tuli and M. Manns, "Explainable human activity recognition based on probabilistic spatial partitions for symbiotic workplaces," *Int. J. Comput. Integr. Manuf.*, vol. 36, no. 12, pp. 1783–1800, Dec. 2023, doi: 10.1080/0951192x.2023.2177742.

[51] H. Gammulle, D. Ahmedt-Aristizabal, S. Denman, L. Tychsen-Smith, L. Petersson, and C. Fookes, "Continuous human action recognition for human-machine interaction: A review," *ACM Comput. Surv.*, vol. 55, no. 13, pp. 1–38, Dec. 2023, doi: 10.1145/3587931.

[52] J. Peng, A. Kimmig, D. Wang, Z. Niu, X. Tao, and J. Ovtcharova, ''Intention recognition-based human–machine interaction for mixed flow assembly,'' *J. Manuf. Syst.*, vol. 72, pp. 229–244, Feb. 2024, doi: 10.1016/j.jmsy.2023.11.021.

[53] Z. Jiao, G. Jia, and Y. Cai, ''Ensuring computers understand manual operations in production: Deep-learning-based action recognition in industrial workflows,'' *Appl. Sci.*, vol. 10, no. 3, p. 966, Feb. 2020, doi: 10.3390/app10030966.

[54] S. Mohsen, A. Elkaseer, and S. G. Scholz, ''Industry 4.0-oriented deep learning models for human activity recognition,'' *IEEE Access*, vol. 9, pp. 150508–150521, 2021, doi: 10.1109/ACCESS.2021.3125733.

[55] A. Roitberg, N. Somani, A. Perzylo, M. Rickert, and A. Knoll, ''Multimodal human activity recognition for industrial manufacturing processes in robotic workcells,'' in *Proc. ACM Int. Conf. Multimodal Interact.*, Nov. 2015, pp. 259–266, doi: 10.1145/2818346.2820738.

[56] T. Stiefmeier, D. Roggen, G. Ogris, P. Lukowicz, and G. Tr, ''Wearable activity tracking in car manufacturing,'' *IEEE Pervasive Comput.*, vol. 7, no. 2, pp. 42–50, Apr. 2008, doi: 10.1109/MPRV.2008.40.

[57] K. Simonyan and A. Zisserman, ''Very deep convolutional networks for large-scale image recognition,'' 2014, *arXiv:1409.1556*.

[58] A. Onan and M. A. Toçoglu, ''A term weighted neural language model and stacked bidirectional LSTM based framework for sarcasm identification,'' *IEEE Access*, vol. 9, pp. 7701–7722, 2021.

[59] W. Kong and H. Li, ''Combining adaptive time-series feature window and stacked bidirectional LSTM for predicting tool remaining useful life without failure data,'' *Int. J. Adv. Manuf. Technol.*, vol. 121, nos. 11–12, pp. 7509–7526, Aug. 2022, doi: 10.1007/s00170-022-09771-3.

[60] T. Wang, T. Yang, M. Danelljan, F. S. Khan, X. Zhang, and J. Sun, ''Learning human-object interaction detection using interaction points,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4116–4125.

[61] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, ''YOLOv4: Optimal speed and accuracy of object detection,'' 2020, *arXiv:2004.10934*.

[62] J. C. Chen, Y. Li, and B. D. Shady, ''From value stream mapping toward a lean/sigma continuous improvement process: An industrial case study,'' *Int. J. Prod. Res.*, vol. 48, no. 4, pp. 1069–1086, Feb. 2010, doi: 10.1080/00207540802484911.

[63] W. Bao, Q. Yu, and Y. Kong, ''Evidential deep learning for open set action recognition,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13329–13338.

[64] H. Rao, S. Xu, X. Hu, J. Cheng, and B. Hu, ''Augmented skeleton based contrastive action learning with momentum LSTM for unsupervised action recognition,'' *Inf. Sci.*, vol. 569, pp. 90–109, Aug. 2021, doi: 10.1016/j.ins.2021.04.023.

[65] A. K. M. Nor, S. R. Pedapati, M. Muhammad, and V. Leiva, ''Abnormality detection and failure prediction using explainable Bayesian deep learning: Methodology and case study with industrial data,'' *Mathematics*, vol. 10, no. 4, p. 554, Feb. 2022, doi: 10.3390/math10040554.

[66] H. Bi, M. Perello-Nieto, R. Santos-Rodriguez, P. Flach, and I. Craddock, ''An active semi-supervised deep learning model for human activity recognition,'' *J. Ambient Intell. Hum. Comput.*, vol. 14, no. 10, pp. 13049–13065, Oct. 2023, doi: 10.1007/s12652-022-03768-2.

[67] F. de Arriba-Pérez, S. García-Méndez, J. Otero-Mosquera, F. J. González-Castaño, and F. Gil-Castiñeira, ''Automatic generation of insights from workers' actions in industrial workflows with explainable machine learning: A proposed architecture with validation,'' *IEEE Ind. Electron. Mag.*, early access, Jun. 23, 2023, doi: 10.1109/MIE.2023.3284203.

RYOTA TAKAMIDO received the B.E. degree from the School of Engineering, Nagoya University, Japan, in 2015, and the M.E. and Ph.D. degrees from the Graduate School of Education and Human Development, Nagoya University, in 2016 and 2021, respectively.

Since 2021, he has been a Project Researcher with the Research into Artifacts, Center for Engineering (RACE), The University of Tokyo. His research interests include human motor control and learning, human–machine interaction, robot motion planning, sports science, and cognitive science.

JUN OTA (Member, IEEE) received the B.E., M.E., and Ph.D. degrees from the Faculty of Engineering, The University of Tokyo, in 1987, 1989, and 1994, respectively.

From 1989 to 1991, he worked at Nippon Steel Corporation. In 1991, he was a Research Associate at The University of Tokyo. He became a Lecturer and an Associate Professor, in 1994 and 1996, respectively. From 1996 to 1997, he was a Visiting Scholar at Stanford University. In April 2009, he became a Professor with the Graduate School of Engineering, The University of Tokyo. In June 2009, he became a Professor with Research into Artifacts, Center for Engineering (RACE), The University of Tokyo. Since 2019, he has been a Professor with RACE, School of Engineering, The University of Tokyo. His research interests include multiagent robotic systems, embodied-brain systems science, design support for large-scale production/material handling systems, and human behavior analysis and support.