Acceleration-based Human Activity Recognition of Packaging Tasks Using Motif-guided Attention Networks

Jaime Morales
Graduate School of Information Science and Technology
Osaka University
Osaka, Japan
jaime.morales@ist.osaka-u.ac.jp

Naoya Yoshimura
Graduate School of Information Science and Technology
Osaka University
Osaka, Japan
yoshimura.naoya@ist.osaka-u.ac.jp

Qingxin Xia
Graduate School of Information Science and Technology
Osaka University
Osaka, Japan
xia.qingxin@ist.osaka-u.ac.jp

Atsushi Wada
Corporate Manufacturing Engineering Center
Toshiba Corporation
Kanagawa, Japan
atsushi3.wada@toshiba.co.jp

Yasuo Namioka
Corporate Manufacturing Engineering Center
Toshiba Corporation
Kanagawa, Japan
yasuo.namioka@toshiba.co.jp

Takuya Maekawa
Graduate School of Information Science and Technology
Osaka University
Osaka, Japan
maekawa@ist.osaka-u.ac.jp

***Abstract*—This study presents a new method for recognizing complex human activities in a logistical domain, such as packaging, using acceleration data from a body-worn sensor. Recognition of packaging tasks using standard supervised machine learning is difficult because the observed data vary considerably depending on the number of items to pack, the size of the items, and other parameters. In this study, we focus on characteristic and necessary actions (motions) that occur in a specific operation such as an action of stretching packing tape when assembling shipping boxes. We propose the use of an attention-based neural network to focus on these characteristic actions when recognizing the data. However, training of a such deep network model is a data-intensive process, and obtaining a huge amount of labeled training data in actual industrial settings is difficult. To address this problem, we employ motif-detection algorithms to detect sensor data motifs (segments corresponding to characteristic actions) that can be useful for recognizing operations in advance. Moreover, we propose that the training of the attention-based network should be guided such that it pays attention to the detected motifs, i.e., motif-guided training.**

***Keywords*—Activity recognition, Machine learning, Logistics, Packaging task**

I. Introduction

Background: Due to the growth of the e-commerce industry, logistics currently plays a central role in global and regional supply chains [1], [2]. Therefore, streamlining processes at logistics centers can significantly improve entire supply chains. For example, because Amazon ships 2.5 billion packages annually in the U.S., the costs related to the shipping processes are significant. The processes in logistics centers rely largely on manual activities performed by human employees, and this is expected to continue into the future [3]–[5] to ensure flexible responses to the fast-changing demands of customers and suppliers. Therefore, quantifying manual activities performed by workers is crucial for streamlining the processes in logistics centers, e.g., assessing and re-designing the processes and re-allocating resources. In the pervasive computing community, human activity recognition (HAR) techniques have been employed for recognition tasks in industrial domains, such as factories and logistics centers, [6]–[8] to quantify manual tasks. When an order from a customer is processed at a logistics center, a worker first picks the items specified in the order, i.e., order picking, and then another worker (or the same worker) packs the picked items. In the packaging process, a worker repetitively performs sequences of operations, such as assembling a shipping box, filling the box with the items, and closing the box.

Problems: Figure 1 shows an example of packaging-related acceleration data collected from a worker's smartwatch at an actual logistics center. In this example, a typical series of operations is iterated twice, with each iteration (i.e., work period) comprising a sequence of 6-9 operations. The goal of this study is to predict an operation class label for each time slice. The main challenges of this task are the following: (i) **Difference in sensor data**: As shown in the example, the waveforms of an operation in different periods are dissimilar because of the difference between the items to pack. Each period has a number of items and each of these items can vary in shape. (ii) **Difference in duration**: As can be seen in the example, the duration of an operation in different periods is sometimes different. For example, in the "read label" operation during the i-th period, the worker required less time to obtain the label information using the laser scanner than in the $i+1$-th period. This is because the closeness of the scanner to the

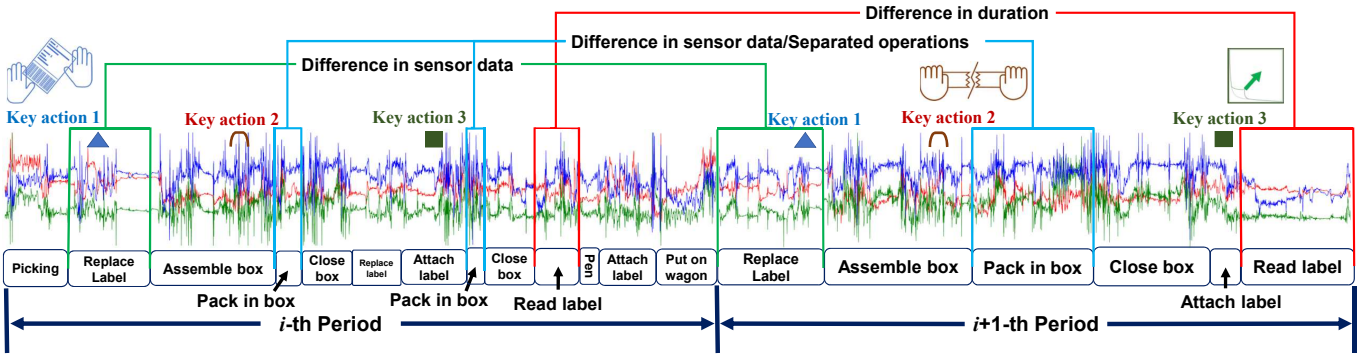

Fig. 1. Example of sensor data regarding work periods (i-th and $i+1$-th periods). The red, green, and blue lines show the x-, y-, and z-axis data, respectively. The rectangles below show the ground truth labels of operations. The "difference in sensor data" and "difference in duration" between the same operation in different periods make the recognition task difficult. The triangles, brackets, and squares show example key actions, which are characteristic actions found in specific operations (attaching a label, stretching tape, and removing glue protection, respectively).

worker's hand after misplacing it on the previous period. (iii) **Complex data**: Several existing HAR studies assume simple waveforms such as periodic waveforms, which are found in walking and running activities. In contrast, as shown in the example sensor data, the sensor data in packaging activities are highly complex because each operation involves multiple complex hand movements. (iv) **Amount of training data**: In addition to the aforementioned issues, we should address a problem related to supervised learning in an actual industrial domain. In the actual industrial domain, it is difficult to collect a huge amount of training data like ImageNet and YouTube-8M Dataset. It is for this reasons that training deep models for industrial applications is challenging.

Approach: Recognition of packaging activities at actual logistics centers is a challenging task. However, our investigation of sensor data from actual logistics centers suggests that a few characteristic actions are included in some operations. For example, in the "assemble shipping box" operation, the worker stretches the packing tape several times and this action is not observed in the other operations. Also, in the "attach address label" operation, the worker always removes the glue protection film from a label before attaching the label. Fig. 1 shows examples of characteristic actions (key actions). We believe that these key actions can be important clues for recognizing complex packaging tasks. In this study, we leverage attention mechanisms [9], [10] to focus on these key actions. The attention mechanisms are used to explicitly identify important parts within a data instance (e.g., region in an image) and to assign higher weights to the identified parts in the recognition process. However, training such complex models requires substantial training data. To address this problem, we employ motif-detection algorithms to detect sensor data motifs (characteristic sensor data segments corresponding to key actions) that can be useful for recognizing operations in advance. We propose guiding the training of the attention-based network such that it focuses on the detected motifs, i.e., motif-guided training. We propose three types of motifs that are useful for recognizing packing activities; the unique, supportive, and boundary motifs.

A Unique motif is a motif that almost exclusively occurs in an operation of interest. A supportive motif is a motif that occurs before or after a unique motif of interest, with the time difference from the corresponding unique motif being consistent. The supportive motif helps understand the temporal/sequential structures of key actions. When we assume that removing the glue protection of a label corresponds to a unique motif, a supportive motif then corresponds to the action of pasting the label on a box after removing the glue protection. A boundary motif corresponds to an action that is performed at around the start or end of an operation of interest. Identifying the starting and ending times of each operation is important for accurately recognizing sequential operations. We identify these motifs in advance for each operation using training data, and we train the neural network based on the identified motifs. The neural network contains an attention head for each motif of each operation, and the head is trained to detect the corresponding motif.

Contributions: i) We propose a new activity-recognition method that involves guiding the training of an attention-based network such that it pays attention to important sensor data motifs. ii) We introduce three types of motifs that are useful for recognizing packaging tasks—unique, supportive, and boundary motifs. iii) We investigate the effectiveness of the proposed method using sensor data collected at an actual logistics center.

II. Related Work

Methods for identifying and improving factory work using wearable sensor technologies have been extensively studied [6]–[8], [11], [12] due to the increasing interest in smart manufacturing as well as Industry 4.0 [13]–[15]. For instance, Koskimäki et al. [16] collected data from wrist-worn accelerometers to ensure that all required actions are performed. They identified operations such as screwing and hammering using a k-nearest neighbor search. Ward et al. [17] recognized individual operations using hidden Markov models (HMMs)

as well as a linear discriminative classifier from inertial acceleration and audio data collected during woodworking activities. Stiefmeier et al. [18], [19] used acceleration data acquired during manual works such as maintenance work and bicycle repair, and employed an HMM or template matching to recognize the data.

Rueda et al. [7] employed a parallel approach based on CNN for multiple inertial measurement units (IMUs). Their method involves preparing a block composed of convolutional and max-pooling layers for each IMU and merging all the parallel paths for the IMUs on a single fully connected layer. This method was evaluated on sensor data collected in a logistics scenario. Reining et al. [8] used CNNs to identify feature representations of activities performed during the order-picking process. They employed CNNs to process motion capture data to recognize specific motions that are considered as characteristic features of multiple activities in the order-picking system. Tao et al. [6] employed CNNs to recognize six different standard assembly operations performed consecutively; they converted multi-channel IMU data and electromyographic data into images and trained the CNNs on the images. In contrast, we enhance training of attention-based neural networks through the motif-guided training. To address the scarcity of training data, Xia et al. [12], [20] developed a motif-based particle filter to recognize factory tasks in an unsupervised manner under a real work environment. However, the aforementioned studies assume that workers sequentially perform pre-defined operations in a pre-defined order within the expected duration of each operation specified in an instruction document.

Several studies have explored motif-detection algorithms for activity and gesture recognition [21]–[23]. Minnen et al. [21] discovered motif seeds using a minimum description length (MDL) criterion and then refined the motif seeds by splitting, merging, and extending the motifs. Maekawa et al. [23] measured the duration of each work period on a production line in an unsupervised manner by discovering a motif (corresponding to an atomic action) that appears only once in each work period. Dissanayake et al. [24] predicted location classes of a room such as a kitchen, and restroom, where a user is located by discovering location-specific sensor data motifs by calculating a score that represents the "location specificity" of each motif. These studies suggest that atomic actions can be successfully extracted from acceleration data using symbol-based motif detection algorithms.

III. Activity Recognition with Motif-guided Attention Network

A. Preliminaries

This study assumes that workers wear a body-mounted accelerometer. Each worker sequentially processed orders from customers. A work period for processing an order included one iteration of overall operations, as shown in Fig. 1. We assumed that training data consisting of labeled sensor data segments were available, with each labeled segment corresponding to the sensor data of one training period. Each label attached to a segment specified the starting and ending times of an operation included in the period and a class label of the operation as shown in Fig. 1.

B. Overview

The proposed method involves training and test phases. In the training phase, which is depicted in Fig. 2, we first preprocess sensor data of training periods (standardization) and then find unique, supportive, and boundary motifs from the training data. For each motif, we calculate a motif-occurrence sequence from sensor data of each training period. Note that a motif-occurrence sequence is a time-series that specifies when a motif of interest occurs. We train our motif-guided attention network (MGA-Net) on the labeled sensor data and motif-occurrence sequences. In the test phase, we preprocess a sensor data segment of a test period and then feed the preprocessed segment into the trained MGA-Net. MGA-Net outputs a class label for each data point within the input segment.

C. Motif Identification

Here, we first symbolize the input time-series. We then find useful motifs from the sequence of symbols. We start by extracting candidate motifs with a length of l_m from a sequence of symbols corresponding to the initial working period in the training data. We use sliding time windows with a stride length of 1 (symbol) and window size of l_m to extract the candidate motifs. Note that candidate motifs for each operation are extracted from a segment corresponding to the operation in the first period. We then compute an occurrence sequence for each motif candidate from a sequence of symbols for each training period. Thereafter, we calculate the scores of each candidate to identify the unique, supportive, and boundary motifs from among the candidates. Finally, we select the unique, supportive, and boundary motifs for each operation. Fig. 3 illustrates the procedures the symbolization, extracting candidate motifs, and calculation of motif occurrence sequences.

1) Preprocessing: We first employ principal component analysis (PCA) [25] to reduce the dimensionality of the sensor data. Because three-axis time-series data collected from a worker were used in our experiment (acceleration data from a sensor attached to the dominant hand), we obtained one-axis time-series data by taking the first principal component since they will be converted into a sequence of symbols in the next process.

To find motifs efficiently, we preprocess the one-dimensional time-series. We first use piecewise aggregate approximation (PAA) [26] to reduce the number of data points within the time-series. In PAA, data points in a data segment within a time window with the length of δ are averaged and then it is used as a representative value of the window. We then symbolize the down-sampled data as shown in Fig. 3 (upper). In brief, we convert each downsampled numerical value into a symbol based on the value range associated with each symbol [27], e.g., a time-series is symbolized as *aabcddbaa*, where the same character belongs to the same value range.

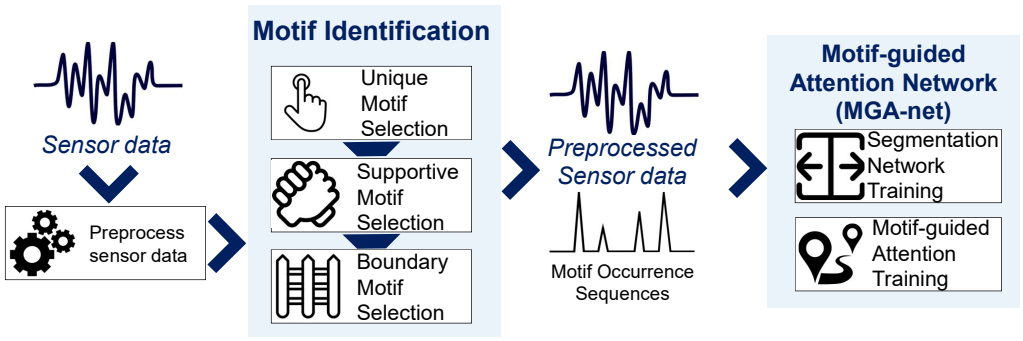

Fig. 2. Overview of the training phase of the proposed method

Then, we extract motif candidates with a length of l_m for each operation. We first obtain a symbol sequence corresponding to the operation of interest in the first training period. Then, we use a sliding time window with a length of l_m and a stride of one symbol to extract motif candidates with a length of l_m in the symbol sequence as shown in Fig. 3 (upper right).

2) Calculating Occurrence Sequence of Motif: From the calculated sequence of symbols for each training period, we compute the occurrence sequence of each extracted candidate motif by using a sliding time window as shown in Fig. 3 (lower). The length of the window is l_m, and the window moves along the sequence of symbols with a stride of 1. We calculate the similarity between the extracted candidate motif and a series of symbols within each time window, obtaining a series of similarity values. When calculating the similarity, we leverage the Levenshtein distance metric [28], which is widely used to compare two symbol sequences.

We then process the similarity sequence to transform it into an occurrence sequence. Because we require the positions where a key action (motif) is performed, we convert all similarity values below the threshold th_s to zero. As a result, we obtain an occurrence sequence, which is a numeric time-series exhibiting high similarity values at the occurrences of the motif of interest.

After calculating an occurrence sequence for each motif candidate of each operation for each training period, we discard motifs that do not frequently occur in the operation of interest. Finally, for each target operation, we select the 10 most frequent motif candidates in each occurrence of the operation of interest in all subsequent training periods. We use this shortlist to select the three defining motifs (unique, supportive, and boundary) for each operation.

3) Identifying Unique Motif: A unique motif that usually appears in a specific operation is useful for identifying that operation. Moreover, the unique motif should not be observed in any other operation. Therefore, we calculate a uniqueness score for each candidate motif belonging to operation O_j using the motif-occurrence sequences of the candidate motif extracted from training data:

$$\mathbf{U}_{score} = \frac{1}{N_t} \sum_{i=1}^{N_t} \frac{\mathbf{o}_i \mathbf{B}_i^{\mathrm{T}}}{T^i} \tag{1}$$

where N_t is the number of training periods, T^i is the length of the i-th period, $\mathbf{o}_i$ is an occurrence time-series of the i-th period for the candidate motif, and $\mathbf{B}_i$ is a time-series whose element value at time t is 1 when the operation label matches that of the motif candidate and is -1 otherwise.

$$\mathbf{B}_i^t = \begin{cases} 1, & \text{Operation at time } t \text{ is } O_j \\ -1, & \text{otherwise} \end{cases} \tag{2}$$

When the motif of interest occurs during the operation of interest, $\mathbf{o}_i \mathbf{B}_i^{\mathrm{T}}$ takes a large positive value. In contrast, when the motif of interest frequently occurs at operations different from the operation of interest, $\mathbf{o}_i \mathbf{B}_i^{\mathrm{T}}$ takes a large negative value because the value of $\mathbf{B}_i^t$ is negative when the operation of interest does not occur. We consider the candidate motif with the best score as a unique motif of the operation. Note that when identifying a unique motif, we vary the value of l_m to generate motif candidates (and their occurrence sequences) and select the candidate with the best score.

4) Identifying Supportive Motif: A unique motif can only locate a single type of action, although an operation consists of a sequence of actions. Therefore, it is still difficult to model the structure of an operation using only a unique motif. To recognize the operation efficiently, capturing the sequential structure of actions is important. We employ a supportive motif for the unique motif for the operation O_j that occurs before or after the unique motif, where the time difference between the unique motif and supportive motif is consistent in each period. To find the supportive motif from among the candidate motifs of operation O_j, we calculate a score for each candidate motif as follows. Note that this calculation is performed only for candidates whose time difference from the unique motif is longer than t_{su} in the first training period.

$$S_{score} = \frac{|\mathbf{T}_{\mathbf{\Delta}}|}{N_t} \sqrt{\frac{\sum_i (\mathbf{T}_{\mathbf{\Delta}i} - \overline{\mathbf{T}_{\mathbf{\Delta}}})^2}{|\mathbf{T}_{\mathbf{\Delta}}|}} \tag{3}$$

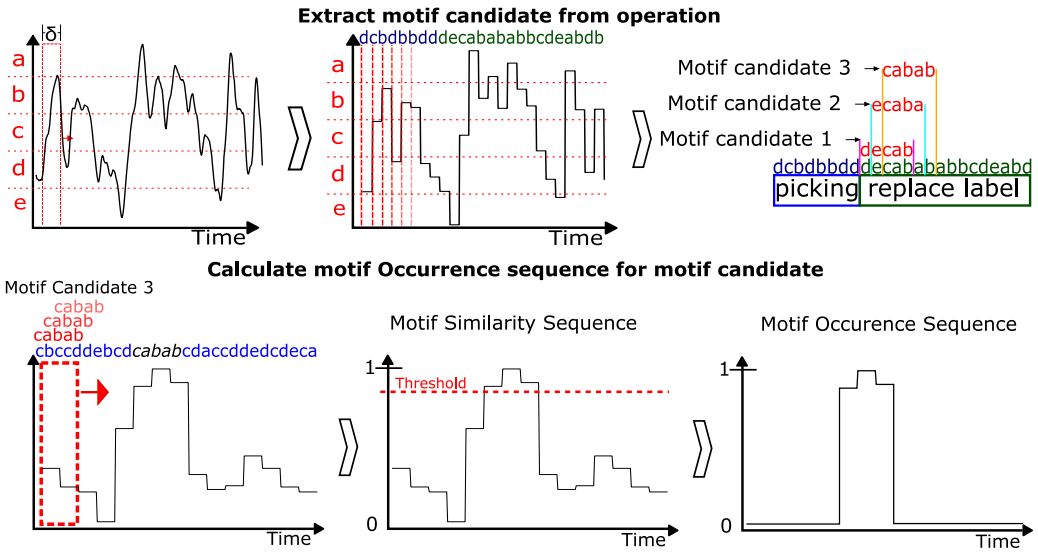

Fig. 3. Overview of motif extraction and motif occurrence sequence calculation.

where $\mathbf{T_\Delta}$ is a vector representing the time difference between the occurrence of the unique motif and the nearest occurrence of the candidate motif in the training periods. $\mathbf{T_\Delta}_i$ is the i-th element of $\mathbf{T_\Delta}$. The differences are calculated by subtracting the timestamp of the candidate-motif occurrence from that of nearest the unique-motif occurrence within a period. By calculating $\sum_i(\mathbf{T_\Delta}_i - \overline{\mathbf{T_\Delta}})^2$, we can measure the stability of the temporal difference between the occurrences of the unique and candidate motifs. We select the candidate motif with the best score as a supportive motif for the operation.

5) Identifying Boundary Motif: Unlike unique and supportive motifs, a boundary motif appears at the boundary between two consecutive operations. Because we employ a segmentation neural network to find the starting and ending times of each operation, it is important to know where the boundary between operations lies. A boundary motif for operation O_j is selected from among the candidate motifs of operation O_j that were not selected as a unique or supportive motif. We calculate a score for each candidate motif based on the temporal difference between the occurrence of the candidate motif and the end of the designated operation, as follows:

$$B_{score} = \frac{|\mathbf{B_\Delta}|}{\overline{\mathbf{B_\Delta}} N_t} \sqrt{\frac{\sum_i(\mathbf{B_\Delta}_i - \overline{\mathbf{B_\Delta}})^2}{|\mathbf{B_\Delta}|}} \tag{4}$$

where $\mathbf{B_\Delta}_i$ is the temporal distance between the ending time of operation O_j in the i-th period and the nearest occurrence of the candidate motif. By calculating $\sum_i(\mathbf{B_\Delta}_i - \overline{\mathbf{B_\Delta}})^2$, we can measure the stability of the occurrences of the candidate motif. In addition, $\overline{\mathbf{B_\Delta}}$ in the denominator increases the score of a candidate that appears close to the boundaries. We select the candidate motif with the best score as a boundary motif for the operation.

D. Motif-guided Attention Network (MGA-Net)

The input of our motif-guided attention network (MGA-Net) is a three-dimensional time-series of preprocessed sensor data. When we train MGA-Net, the occurrence sequences of unique, supportive, and boundary motifs are also used to train the attention heads in MGA-Net. The output of MGA-Net is a series of an estimate (class label) for each data point.

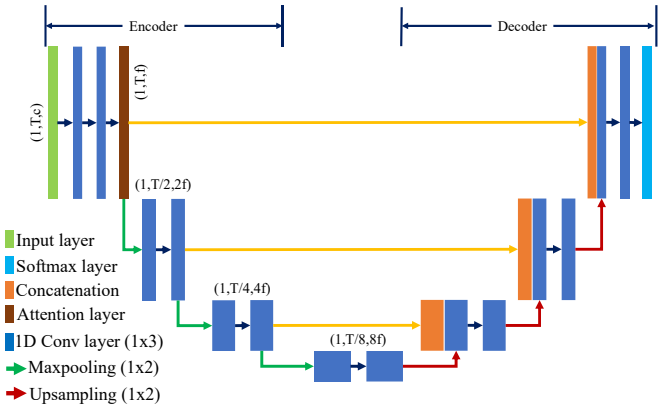

Fig. 4. MGA-Net structure. f is the number of nodes (kernels) in the convolution layer in the first encoding block. T is the input length. c is the number of input channels ($c = 3$).

1) Network Structure: Figure 4 shows the structure of MGA-Net, which is based on the U-Net topology for segmentation of time-series data [29]. The network consists of three encoding and three decoding blocks. Each encoding block consists of two one-dimensional convolution layers and one max-pooling layer. Each decoding block consists of two one-dimensional convolution layers and one upsampling layer. Another part of the original U-Net topology is the concatenation layer at the beginning of each decoder block,

where the output from the second convolution layer of the corresponding encoding block is concatenated with that of the upsampling layer. This network structure enables us to use a complete period of data as a single input. It also permits us to output an estimate (class prediction) for each data point in the input time-series.

As can be seen in Fig. 4, we include a multi-head attention mechanism at the end of the first encoding block to identify unique, supportive, and boundary motifs for each operation. This attention mechanism consists of one attention head dedicated to identifying each specific motif of an operation (i.e., three heads for each operation), with each of the attention heads analyzing the input time-series independently to focus more on identifying the motif. We introduce this attention mechanism into the first encoding block to focus even on small actions that would be lost in the deeper layers. The attention layer is composed of $3 \times K$ attention heads, where K is the number of operation classes. Each attention head individually calculates the time-series of attention according to Equation 5. The length of the time-series of the attention is identical to that of the output of the first encoding block. The time-series of the attention is calculated as follows.

$$\mathbf{a} = \text{softmax}(\tanh(WZ^{\mathrm{T}})), \tag{5}$$

where W represents trainable parameters of the weight of an attention head, and $Z \in \mathbb{R}^{T \times f}$ is the output of the last convolution layer in the first encoding block. T is the length of the input time-series, and f is the number of nodes in the convolution layer in the first encoding block. An attention value in $\mathbf{a}$ at time t shows the importance (i.e., attention) of the data point at t. The attention time-series is multiplied by the output of the last convolution layer in the first encoding block to highlight important segments (corresponding to motifs) in the output of the first encoding block. It should be noted that different attention heads have different trainable weight parameters (i.e., W).

2) Motif-guided Training: To enhance the training of the network, the attention heads in MGA-Net are trained to detect motifs using motif-occurrence sequences. We prepared a total of three occurrence sequences corresponding to unique, supportive, and boundary motifs for each operation and each period. These sequences specify where the motifs occur in a period and can therefore be useful for training the attention heads prepared for the operation. We train MGA-Net via backpropagation based on Adam [30] by minimizing the loss function calculated using the occurrence sequences and estimates:

$$E(\theta, \mathbf{o}) = \frac{1}{N_T}\sum_{i=1}^{N_T}\mathcal{L}_c(\theta) + \lambda_2 \frac{1}{N_T}\sum_{i=1}^{N_T}\mathcal{L}_a(\theta_a, \mathbf{o}^i) + \lambda_3 \frac{1}{N_T}\sum_{i=1}^{N_T}\mathcal{L}_s(\theta), \tag{6}$$

where θ is a set of trainable parameters of MGA-Net, N_T is the number of training periods, θ_a represents the trainable parameters of the attention heads in θ, $\mathbf{o}^i$ is a set of motif-occurrence sequences in the i-th period, and λ_2 and λ_3 are the trade-off hyperparameters of the three loss functions. $\mathcal{L}_c(\cdot)$ denotes the cross-entropy loss associated with label prediction for the i-th training period. $\mathcal{L}_a(\cdot, \cdot)$ represents the loss involved in attention-guided training for the i-th training period, which is formulated as follows:

$$\mathcal{L}_a(\theta_a, \mathbf{o}^i) = \sum_{j=1}^{K*3}\sum_{t=1}^{T} ||\mathbf{a}_t^j - \mathbf{o}_t^{i,j}|| \tag{7}$$

where $\mathbf{a}_t^j$ is the attention value of the j-th attention head at time t, and $\mathbf{o}_t^{i,j}$ is the value of the j-th occurrence sequence at time t. Note that because the sum of elements in an attention series for a period is 1, we normalize $\mathbf{o}^{i,j}$, which is the j-th occurrence sequence, to ensure that the sum of elements in $\mathbf{o}^{i,j}$ is 1. Furthermore, MGA-Net has three attention heads for each operation (i.e., $K*3$). Because this loss function denotes the difference between the attention time-series computed by an attention head of interest and an occurrence time-series of the corresponding motif, we can guide the training of the attention head such that the attention head detects occurrences of the motif.

$\mathcal{L}_s(\cdot)$ represents the segmentation loss for the i-th training period. The segmentation loss is introduced to mitigate the skewed difference among class appearances during the training period [29] and is formulated as follows.

$$\mathcal{L}_s(\theta) = \sum_{t=1}^{T}\sum_{k=1}^{K} -\frac{\mathbf{P}_k}{\mathbf{P}} \ln(f(x)_{k_t}) \tag{8}$$

where $\mathbf{P}_k$ is the number of data points belonging to class k within the i-th period, and $\mathbf{P}$ is the total number of data points belonging to any class within the i-th period.

The combination of these three losses allows us to fine-tune the parameters required to train the segmentation network as well as the attention mechanism. Thus, we can identify the locations of the key actions that describe each operation, while satisfying the requirements for accurately determining the start and end times of each operation.

IV. Evaluation

A. Dataset

We evaluated the proposed method using two acceleration datasets. The first one (LOGI dataset) was collected from four workers at a real logistics center. The data were collected from a smartwatch (Sony SmartWatch3 SWR50) worn by the workers on their dominant hand, with a sampling rate of approximately 30 Hz. The total duration of the dataset is about five hours. The collected data were manually labeled to generate ground truths using video recordings. Fig. 5 shows an example of the images captured. This dataset contains ten operations (activity classes) to recognize: "picking," "replace label," "assemble box," "pack in box," "close box," "attach label," "read label," "put on cart," "assemble box," "use pen," and "unassigned (others)." Table I shows an overview of the dataset. Note that sensor data from some periods are used only for training because of issues related to data-collection failure or labeling.

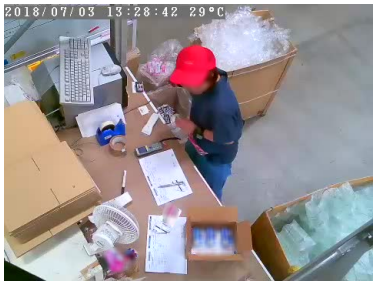

Fig. 5. Example of Worker 1 performing a packaging task

TABLE I
OVERVIEW OF LOGI DATASET

Worker	1	2	3	4
Number of work periods	71	50	48	86
Periods only for training	9	7	12	3
Total duration (seconds)	2467	3927	5136	6258
Avg duration (seconds)	34	77	105	72
Worker proficiency	skilled	beginner	mid-level	mid-level
Experience (years)	15	0.5	20	10
Dominant hand	right	right	left	right

The second is the **L**ogistic **A**ctivity **R**ecognition Ch**a**llenge (LARa) dataset [31] that contains accelerometer data from 14 subjects performing product picking and packaging on 3 different simulated scenarios. We used the accelerometer data collected from the dominant wrist with an approximate sampling rate of 100Hz. We used the data from 6 of the 8 workers who performed scenarios 2 and 3 that contain operations performed in the LOGI dataset, such as "scan code/label" or "seal/close box." This dataset contains 6 operations (activity classes) to recognize: "standing," "walking," "cart," "handling (upwards)," "handling (centred)," and "handling (downwards)." Note that 2 workers were excluded from the evaluation due to their small number of working periods derived from excessive noise and faulty sensor readings. We used 5.6 hours of data containing 168 working periods (28 periods per worker on average).

B. Evaluation Methodology

We used leave-one-period-out cross-validation (i.e, worker-dependent models) and leave-one-worker-out cross-validation (within each dataset) for this evaluation. The proposed method was evaluated by using the weighted average F1-score based on an estimate for each sample. To evaluate the effectiveness of the proposed method, we compared it with window- and sample-based methods.

1) Window-based methods: The input of a window-based method is a sensor-data segment in a time window with a size of 30 data points. We used a sliding time window with a stride length of one point. For window-based methods using deep learning, we employed the cross-entropy loss function based on the Adam optimizer. The learning rate was 0.01, and the training period was 100 epochs, with a batch size of 128.
- **LSTM**: We used a network composed of four LSTM layers with the sigmoid activation function and an output layer.
- **DeepConvLSTM**: This is a more recent baseline considered as the state-of-the-art for HAR using convolutional LSTM networks. We selected a variation of the model used in [32].
- **Conv-IMU**: This method [7] uses parallel branches for multi-channel multi-sensor data streams. Even though the advantage of using the parallel branches dedicated to multiple IMUs is not applied when we use only one IMU in our experiment, we consider that Conv-IMU achieves state-of-the-art due to its ability to reduce temporal variability of predictions with its asynchronous loss calculation (see [7] in detail). Therefore, we use it as a baseline for comparison.
- **Random forest**: This employs the random forest algorithm. The number of trees is 100. We extracted statistical features from each window; the standard deviation, zero-crossings, and energy are extracted for each axis, and the root mean square for the three axes based on prior studies [33], [34].

2) Sample-based methods: The input of a sample-based method is a sensor-data segment corresponding to a period. A sample-based method outputs a class estimate for each data point. The training period was 100 epochs, with a batch size of 4. A batch corresponds to a sensor-data segment of a period. The Adam optimizer was used, and the learning rate was 0.01.
- **MGA-Net**: This is the proposed method.
- **Stateful-LSTM**: This is the variant of the LSTM method. We use the same network structure as in the LSTM method.
- **U-net**: This is the base model for our method and is based on a neural network for time-series segmentation [29]. The network consists of three encoding and three decoding blocks. For each encoder and decoder block, we used two 1D convolutional layers with a kernel size of 3 and a max-pooling/upsampling layer with a kernel size of 2.
- **W/o MGT**: This is the variant of the proposed method without the motif-guided training. This method is also regarded as the fusion of the state-of-the-art segmentation [29] and attention [35] HAR networks.

Table II shows the experimental parameters.

TABLE II
EXPERIMENTAL PARAMETERS USED FOR EVALUATION

Parameter	Value	Parameter	Value
PAA window size	10	t_{su}	2 seconds
# symbols	10	f	30
l_m	[2,3,5,7,10,15,20]	λ_2	125
th_s	0.9	λ_3	1

C. Results

1) Leave-one-period-out CV: The left part of Fig. 6 shows the mean F1-scores over all the ten workers in the two datasets, showing that MGA-Net significantly outperformed the other

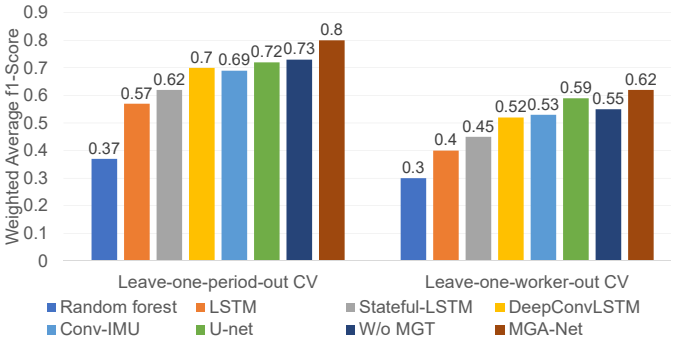

Fig. 6. Average F1-score over the ten workers for the methods

TABLE III
WEIGHTED AVERAGE F1-SCORE FOR ALL THE TESTED METHODS AND ALL WORKERS FOR THE LOGI DATASET. THESE ARE THE RESULTS OF LEAVE-ONE-PERIOD-OUT CROSS VALIDATION.

Worker	1	2	3	4	AVG
Random forest	0.33	0.31	0.32	0.36	0.33
LSTM	0.57	0.51	0.53	0.61	0.55
Stateful-LSTM	0.60	0.54	0.57	0.63	0.58
DeepConvLSTM	0.58	0.55	0.53	0.66	0.58
Conv-IMU	0.56	0.46	0.65	0.69	0.60
U-Net	0.67	0.47	0.60	0.65	0.59
W/o MGT	0.73	0.54	0.63	0.69	0.65
MGA-Net	**0.83**	**0.62**	**0.74**	**0.81**	**0.75**

methods. Table III shows the mean F1-score of MGA-Net for the LOGI dataset. Surprisingly, MGA-Net achieved the F1-scores higher than 0.80 for Workers 1 and 4 even though the observed data are very complex compared to data used in standard HAR studies. In contrast, as can be seen in Table III, the F1-scores of the state-of-the-art methods (DeepConvLSTM, Conv-IMU, U-net and W/o MGT) were lower than 0.7 in many cases. The mean F1-score for the LSTM method was 0.55, indicating the difficulty of this task. The mean F1-score for the Stateful-LSTM method was slightly better than that of LSTM, suggesting that, compared with LSTM, Stateful-LSTM could capture long-term dependency in the data because its input is longer than that of LSTM. The mean F1-score for the Random Forest method was 0.33. As expected, the traditional machine learning method lacks the ability to properly recognize this type of complex data. We employed a complex model (attention-based model) because the attention-based method is superior to simpler methods trained on limited data. As indicated by the results, the attention-based method (W/o MGT and MGA-Net) could outperform the simpler methods. Specifically, the motif-guided training compensated for the scarcity of the training data. The mean F1-score over the workers for MGA-Net was higher than the mean F1-scores of the state-of-the-art methods by about 15-20%. In addition, MGA-Net outperformed W/o MGT by 10%, indicating the significant contribution of the proposed motif-guided training.

Figure 7 shows ground truth labels and predictions by the methods for a work period of Worker 3. As shown in the figure, MGA-Net has few false detections with short duration. This can be because MGA-Net could precisely detect operation boundaries. Even when we used the attention-based neural network for segmentation (i.e., W/o MGT), we could not prevent short-duration false detections.

As can be seen in Fig. 7, the results of U-net contain many false detections with long duration due to the overall similarity in sensor data between different operations. The results of LSTM and DeepConvLSTM also contain many false detections with long duration. In contrast, MGA-Net can recognize operations by identifying short characteristic segments (motifs). Conv-IMU could capture the overall trend of the operation sequence. However, the results of Conv-IMU also contain many false detections with short duration.

TABLE IV
WEIGHTED AVERAGE F1-SCORE FOR SELECTED METHODS AND WORKERS FROM LARA DATASET. THESE ARE THE RESULTS OF LEAVE-ONE-PERIOD-OUT CROSS VALIDATION

Worker	7	8	9	10	13	14	AVG
Random forest	0.42	0.38	0.40	0.42	0.40	0.41	0.41
LSTM	0.50	0.62	0.57	0.63	0.59	0.56	0.58
Stateful-LSTM	0.56	0.67	0.62	0.69	0.65	0.62	0.64
DeepConv LSTM	0.79	0.77	0.74	0.76	0.79	0.81	0.78
Conv-IMU	0.77	0.72	0.76	0.75	0.79	0.74	0.75
U-Net	0.80	0.81	0.80	0.81	0.81	0.80	0.80
W/o MGT	0.78	0.77	0.80	0.78	0.79	0.79	0.78
MGA-Net	**0.84**	**0.82**	**0.83**	**0.85**	**0.85**	**0.82**	**0.83**

Table IV shows the results of the LARa dataset. MGA-Net achieved the best performance for all the workers. Because the LARa dataset is easier to recognize than the LOGI dataset (6 vs. 10 classes), we believe that the performance of MGA-Net is close to the upper bound of the recognition performance of this dataset. Therefore, the difference of performance between MGA-Net and U-net for the LARa dataset is smaller than that of the LOGI dataset.

2) Leave-one-worker-out CV: Table V shows the results of the LOGI dataset. In addition, Table VI shows the results of the LARa dataset. MGA-Net achieved the best performance for many workers. The F1-scores for leave-one-worker-out cross validation were poorer than those for leave-one-period-out cross validation. This is because of the high degree of freedom for packaging work, resulting in large differences in sensor data between workers. The performance of W/o MGT was rather poorer than that of U-net. In contrast, the performance of MGA-Net was rather higher than that of U-net. This indicates that capturing key actions commonly found in different workers using a simple attention-based method (i.e., W/o MGT) was difficult. However, the motif-guided training appears to facilitate the detection of key actions commonly found in training workers.

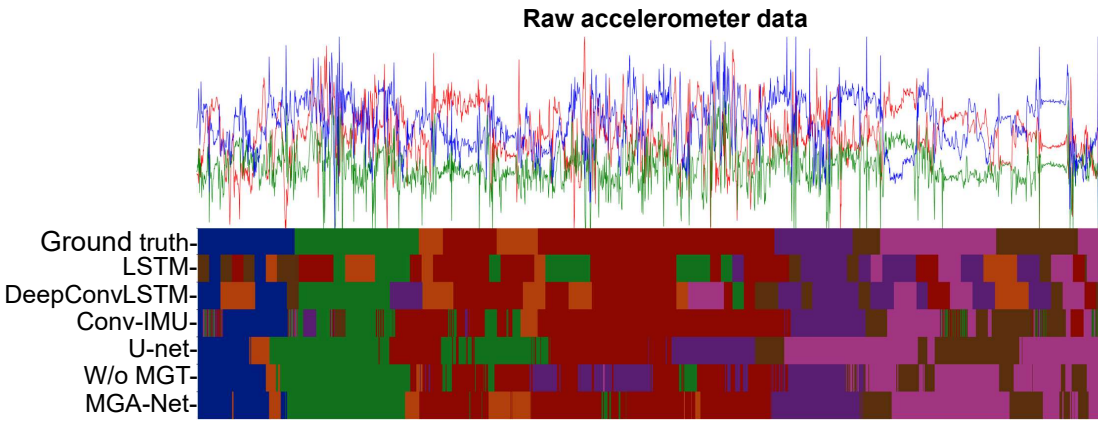

Fig. 7. Ground truth and predictions by the methods for an example working period of Worker 3. The horizontal axis shows time.

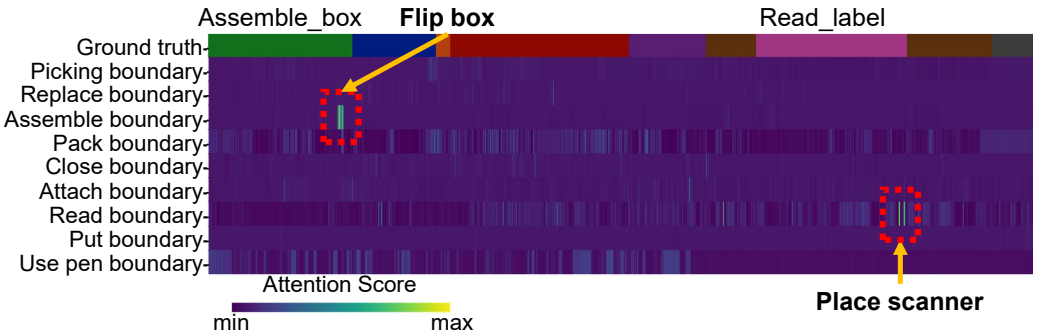

Fig. 8. Time-series of attention values for attention heads of boundary motifs on several operations in an example period from Worker 1. The attention heads focus on actions (flip box and place scanner actions) performed at the end of operations.

TABLE V
WEIGHTED AVERAGE F1-SCORE FOR ALL METHODS AND WORKERS FROM LOGI DATASET. THESE ARE THE RESULTS OF LEAVE-ONE-WORKER-OUT CROSS VALIDATION

Worker	1	2	3	4	AVG
Random forest	0.25	0.22	0.25	0.28	0.25
LSTM	0.25	0.12	0.32	0.36	0.26
Stateful-LSTM	0.23	0.17	0.3	0.37	0.27
DeepConvLSTM	0.27	0.16	0.32	0.39	0.29
Conv-IMU	0.33	0.34	0.39	0.41	0.37
U-Net	**0.51**	0.36	0.40	**0.42**	0.42
W/o MGT	0.45	0.31	0.43	0.35	0.39
MGA-Net	**0.51**	**0.39**	**0.46**	0.40	**0.44**

TABLE VI
WEIGHTED AVERAGE F1-SCORE FOR SELECTED METHODS AND WORKERS FROM LARA DATASET. THESE ARE THE RESULTS OF LEAVE-ONE-WORKER-OUT CROSS VALIDATION

Worker	7	8	9	10	13	14	AVG
Random forest	0.35	0.33	0.36	0.30	0.33	0.33	0.34
LSTM	0.43	0.56	0.48	0.44	0.51	0.49	0.49
Stateful-LSTM	0.55	0.63	0.48	0.61	0.47	0.64	0.56
DeepConv LSTM	0.67	0.68	0.67	0.68	0.71	**0.73**	0.69
Conv-IMU	0.60	0.63	0.67	0.61	0.68	0.62	0.63
U-Net	0.72	0.72	0.68	**0.70**	0.69	0.68	0.70
W/o MGT	0.69	0.68	0.65	0.60	0.66	0.65	0.65
MGA-Net	**0.75**	**0.75**	**0.73**	0.67	**0.77**	**0.73**	**0.73**

D. Discussion

1) Contributions of Motifs: Here we investigate the contributions of the three types of motifs. We have prepared the following methods.

- **MGA-Net (U)**: The variant of the proposed method that uses only unique motifs.
- **MGA-Net (U+S)**: The variant of the proposed method that does not use boundary motifs.
- **MGA-Net (U+B)**: The variant of the proposed method that does not use supportive motifs.
- **MGA-Net (B)**: The variant of the proposed method that uses only boundary motifs.

Figure 9 shows the mean F1-scores of the methods and MGA-Net (U+S+B) for the LOGI dataset. The performance of MGA-Net (B) was almost identical to that of MGA-Net (U). Since a unique motif almost exclusively occurs in an operation of interest, it is useful to identify a segment corresponding to the operation. Because identification of the starting and ending times, i.e., boundary, plays a key role in time-series

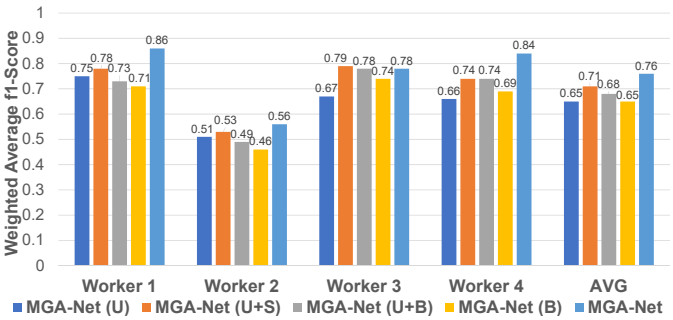

Fig. 9. Contributions of motifs. We used ten randomly selected testing periods for this evaluation. The remaining periods from the same worker have been used as training.

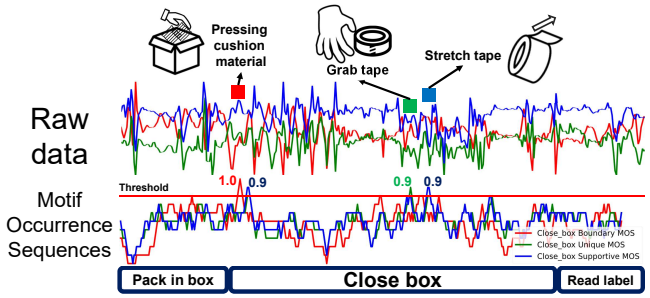

Fig. 10. Example of motif occurrence sequences for Close box operation in worker 1. Relevant atomic actions found by each motif type within an operation are highlighted.

TABLE VII
ANALYSIS OF ATTENTION VALUES IN MGA-NET FOR LOGI DATASET

	Unique motif	Supportive motif	Boundary motif
Avg. in Op. of interest	0.439×10^{-3}	0.448×10^{-3}	$\mathbf{0.601}\times10^{-3}$
Max. in Op. of interest	$\mathbf{6.12}\times10^{-3}$	1.89×10^{-3}	$\mathbf{24.7}\times10^{-3}$
Avg. outside Op. of interest	0.0614×10^{-3}	0.113×10^{-3}	0.206×10^{-3}
Max. outside Op. of interest	0.906×10^{-3}	0.209×10^{-3}	0.128×10^{-3}

segmentation, boundary motifs also contributed to the accurate segmentation. In the results of Worker 2, the mean F1-score of MGA-Net (B) was poorer than that of MGA-Net (U). This can be because an operation by Worker 2 was sometimes separated into several segments in a work period because of mistakes, making it difficult to find useful boundary motifs.

MGA-Net (U+S) slightly outperformed MGA-Net (U+B), indicating that the contribution of supportive motifs is higher than that of boundary motifs. Supportive motifs permit us to capture temporal information of actions, which is also effective to predict the duration of an operation, when used in combination with unique motifs.

2) Analysis of Motif Identification: Here we analyze the results of the motif identification method. As mentioned in Section III-C, because the process for selecting motifs varies by operations and by motif types, various motions with different motion lengths can be identified as relevant motifs. Fig. 10 shows an example of the output (motif occurrence sequences before thresholding) from the motif identification method for the operation "Close box." We can observe that both the unique and boundary motifs were able to identify specific atomic actions that can be attributed to the "Close box" operation. For example, the pressing of cushion materials before closing the sides of the box can be the start of the operation in every iteration and is a unique action that can only be present in the "Close box" operation. In the case of actions found by the unique and supportive motifs, we can understand its relative importance to the operation given that only two operations in the sequence use tape but the motion simplicity makes it hard to find perfect matches and opens the possibility for incorrect appearances on the same or different operations.

In Fig. 10, a motif occurrence sequence of the boundary motif shows a similarity score of "1.0" in its region that indicates a perfect match to its original seed, while both the unique and supportive motifs have "0.9" matches within the operation. All the three motifs have no matches in the adjacent operations, but we note that for the selected motif occurrence sequence of the supportive motif, there are matches with "0.9" and "1.0" similarity values in two other operations (Assemble box and Replace label). The appearances of high similarity values on other operations/actions for the supportive motif can be attributed to its selection method, while the unique and boundary motifs are selected giving a high priority to its relative number of appearances outside the target operation. In contrast, the supportive motif mostly considers its time and appearance dependence with the unique motif.

3) Contributions of Attention: Here, we qualitatively analyze attention values output by an attention head of MGA-Net. First, we obtained a set of attention values in each label of an operation for which the attention head is responsible. Subsequently, we computed the average and maximum attention value over the set of attention values. Finally, we computed their average (i.e., average of average attention values and average of maximum attention values) over all labels of the operation of interest. In addition, we computed the average and maximum values outside the labels of the operation of interest. Table VII shows the results. It can be seen that the average attention value in the labels of the operation of interest was approximately 10 times as high as that outside the labels, indicating that the trained attention heads could focus on operations for which they are responsible. The maximum attention value for boundary motifs in the labels of the operation of interest was considerably higher than that outside the labels. This may be because attention heads for boundary motifs focus only on short-duration segments around operation boundaries as shown in Fig. 8.

4) Amount of Training Data: Figure 11 shows the transitions in the mean F1-scores of the methods when we randomly reduced the number of training periods. As can be seen in this figure, MGA-Net outperformed the other methods even when

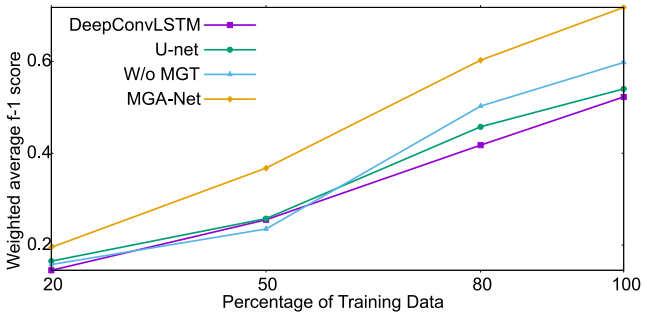

Fig. 11. Transitions in mean F1-scores of the LOGI dataset for each method when we changed the amount of training data. These results are the averages for the four workers. We used ten randomly selected testing periods for this evaluation. The remaining periods from the same worker were used as a pool of training periods. We sampled periods from the pool and used them as training data. We changed the percentage of periods sampled from the pool.

we reduced the amount of training data. However, because these deep-learning methods are data-intensive, the F1-scores significantly decreased when limited training data were used. Accurate prediction with limited training data is our important future work.

V. Conclusion

This study presented a new method for recognizing complex packaging works using a body-worn acceleration sensor. We focused on characteristic actions (motions) that occur in packaging works, and propose the use of an attention-based neural network to focus on these characteristic actions when recognizing the data. The proposed method was evaluated on sensor data collected at an actual logistics center, and it significantly outperformed the state-of-the-art methods used in the HAR domain. As a part of our future study, we plan to evaluate our method on activity data in other applications such as factory work and animal activities [36], [37] by detecting activity-specific motifs. In addition, we plan to apply our method to daily life applications such as predicting a location class (location semantics) of a room by leveraging location-specific motifs [24], [38] as described in the related work section.

Dataset

A dataset of logistics works collected by our research group will be available soon at https://getty708.github.io/open-pack-dataset [39].

Acknowledgement

This work is partially supported by JSPS KAKENHI Grant Number JP16H06539, JP17H04679, JP21H03428, 21H05299, and JP21K19769.

References

[1] X. Zou and B. Smith, "An empirical study on relationship between regional logistics industry development and economic growth based on logistic model," in *International Conference on Education, Management and Information Technology*. Atlantis Press Jinan, China, 2015, pp. 859–866.

[2] D. Schlögl and H. Zsifkovits, "Manuelle Kommissioniersysteme und die Rolle des MenschenManual Picking Systems and Human Factors," *BHM Berg- und Hüttenmännische Monatshefte*, vol. 161, no. 5, pp. 225–228, may 2016. [Online]. Available: https://link.springer.com/article/10.1007/s00501-016-0481-7

[3] M. Klumpp, M. Hesenius, O. Meyer, C. Ruiner, and V. Gruhn, "Production logistics and human-computer interaction—state-of-the-art, challenges and requirements for the future," *The International Journal of Advanced Manufacturing Technology*, vol. 105, no. 9, pp. 3691–3709, 2019.

[4] C. Reining, F. Niemann, F. Moya Rueda, G. A. Fink, and M. ten Hompel, "Human activity recognition for production and logistics—a systematic literature review," *Information*, vol. 10, no. 8, p. 245, 2019.

[5] V. Yavas and Y. D. Ozkan-Ozen, "Logistics centers in the new industrial era: A proposed framework for logistics center 4.0," *Transportation Research Part E: Logistics and Transportation Review*, vol. 135, mar 2020.

[6] 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," *Procedia Manufacturing*, vol. 26, pp. 1159–1166, jan 2018.

[7] F. Moya Rueda, R. Grzeszick, G. A. Fink, S. Feldhorst, and M. Ten Hompel, "Convolutional neural networks for human activity recognition using body-worn sensors," in *Informatics*, vol. 5, no. 2. Multidisciplinary Digital Publishing Institute, 2018, p. 26.

[8] C. Reining, M. Schlangen, L. Hissmann, M. T. Hompel, F. Moya, and G. A. Fink, "Attribute representation for human activity recognition of manual order picking activities," in *the 5th international Workshop on Sensor-based Activity Recognition and Interaction*, 2018, pp. 15–17.

[9] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations (ICLR 2015)*, sep 2015. [Online]. Available: https://arxiv.org/abs/1409.0473v7

[10] S. Mahmud, M. Tanjid Hasan Tonmoy, K. Kumar Bhaumik, A. K. Mahbubur Rahman, M. Ashraful Amin, M. Shoyaib, M. Asif Hossain Khan, and A. Ahsan Ali, "Human activity recognition from wearable sensor data using self-attention," in *Frontiers in Artificial Intelligence and Applications*, vol. 325. IOS Press BV, mar 2020, pp. 1332–1339. [Online]. Available: http://arxiv.org/abs/2003.09018

[11] J.-B. Cordonnier, A. Loukas, and M. Jaggi, "On the relationship between self-attention and convolutional layers," *arXiv preprint arXiv:1911.03584*, 2019.

[12] Q. Xia, A. Wada, J. Korpela, T. Maekawa, and Y. Namioka, "Unsupervised factory activity recognition with wearable sensors using process instruction information," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 2, p. 60, 2019.

[13] M. Bauer, L. Jendoubi, and O. Siemoneit, "Smart factory–mobile computing in production environments," in *the MobiSys 2004 Workshop on Applications of Mobile Embedded Systems (WAMES 2004)*, 2004.

[14] D. Lucke, C. Constantinescu, and E. Westkämper, "Smart factory-a step towards the next generation of manufacturing," in *Manufacturing Systems and Technologies for the New Frontier*. Springer, 2008, pp. 115–118.

[15] A. Radziwon, A. Bilberg, M. Bogers, and E. S. Madsen, "The smart factory: Exploring adaptive and flexible manufacturing solutions," *Procedia Engineering*, vol. 69, pp. 1184–1190, 2014.

[16] H. Koskimäki, V. Huikari, P. Siirtola, P. Laurinen, and J. Röning, "Activity recognition using a wrist-worn inertial measurement unit: A case study for industrial assembly lines," in *17th Mediterranean Conference on Control and Automation (MED 2009)*, 2009, pp. 401–405.

[17] J. A. Ward, P. Lukowicz, and G. Tröster, "Gesture spotting using wrist worn microphone and 3-axis accelerometer," in *the 2005 Joint Conference on Smart Objects and Ambient Intelligence: Innovative context-aware services: usages and technologies*, 2005, pp. 99–104.

[18] T. Stiefmeier, D. Roggen, and G. Tröster, "Fusion of string-matched templates for continuous activity recognition," in *11th IEEE International Symposium on Wearable Computers (ISWC 2007)*, 2007, pp. 41–44.

[19] T. Stiefmeier, G. Ogris, H. Junker, P. Lukowicz, and G. Tröster, "Combining motion sensors and ultrasonic hands tracking for continuous activity recognition in a maintenance scenario," in *10th IEEE International Symposium on Wearable Computers (ISWC 2006)*, 2006, pp. 97–104.

[20] Q. Xia, J. Korpela, Y. Namioka, and T. Maekawa, "Robust Unsupervised Factory Activity Recognition with Body-worn Accelerometer Using Temporal Structure of Multiple Sensor Data Motifs," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 3, pp. 1–30, sep 2020. [Online]. Available: https://dl.acm.org/doi/10.1145/3411836

[21] D. Minnen, T. Starner, M. Essa, and C. Isbell, "Discovering characteristic actions from on-body sensor data," in *International Symposium on Wearable Computers, ISWC 2016*. IEEE Computer Society, 2006, pp. 11–20.

[22] E. Berlin and K. V. Laerhoven, "Detecting Leisure Activities with Dense Motif Discovery," in *the 2012 ACM Conference on Ubiquitous Computing (UbiComp '12)*. New York, New York, USA: ACM Press, 2012.

[23] T. Maekawa, D. Nakai, K. Ohara, and Y. Namioka, "Toward practical factory activity recognition: Unsupervised understanding of repetitive assembly work in a factory," in *the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp 2016)*, ser. UbiComp '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 1088–1099. [Online]. Available: https://doi.org/10.1145/2971648.2971721

[24] T. Dissanayake, T. Maekawa, T. Hara, T. Miyanishi, and M. Kawanabe, "Indolabel: Predicting indoor location class by discovering location-specific sensor data motifs," *IEEE Sensors Journal*, 2021.

[25] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin, "The elements of statistical learning: Data mining, inference, and prediction," *Math. Intell.*, vol. 27, pp. 83–85, 11 2004.

[26] E. Keogh, K. Chakrabarti, M. Pazzani, and S. Mehrotra, "Locally adaptive dimensionality reduction for indexing large time series databases," in *the 2001 ACM SIGMOD International Conference on Management of Data*. New York, NY, USA: Association for Computing Machinery, 2001, p. 151–162. [Online]. Available: https://doi.org/10.1145/375663.375680

[27] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A symbolic representation of time series, with implications for streaming algorithms," in *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, 2003, pp. 2–11.

[28] L. Yujian and L. Bo, "A normalized levenshtein distance metric," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 1091–1095, 2007.

[29] Y. Zhang, Y. Zhang, Z. Zhang, J. Bao, and Y. Song, "Human activity recognition based on time series analysis using u-net," *arXiv preprint arXiv:1809.08113*, 2018.

[30] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[31] F. Niemann, C. Reining, F. Moya Rueda, N. R. Nair, J. A. Steffens, G. A. Fink, and M. Ten Hompel, "Lara: Creating a dataset for human activity recognition in logistics using semantic attributes," *Sensors*, vol. 20, no. 15, p. 4083, 2020.

[32] G. Montúfar, R. Pascanu, K. Cho, and Y. Bengio, "On the number of linear regions of deep neural networks," in *the 27th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'14. Cambridge, MA, USA: MIT Press, 2014, p. 2924–2932.

[33] J. Korpela, K. Takase, T. Hirashima, T. Maekawa, J. Eberle, D. Chakraborty, and K. Aberer, "An energy-aware method for the joint recognition of activities and gestures using wearable sensors," in *International Symposium on Wearable Computers (ISWC 2015)*, 2015, pp. 101–108.

[34] T. Maekawa, Y. Kishino, Y. Yanagisawa, and Y. Sakurai, "Wristsense: wrist-worn sensor device with camera for daily activity recognition," in *2012 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, 2012, pp. 510–512.

[35] H. Ma, W. Li, X. Zhang, S. Gao, and S. Lu, "Attnsense: Multi-level attention mechanism for multimodal human activity recognition." in *IJCAI*, 2019, pp. 3109–3115.

[36] J. Korpela, H. Suzuki, S. Matsumoto, Y. Mizutani, M. Samejima, T. Maekawa, J. Nakai, and K. Yoda, "Machine learning enables improved runtime and precision for bio-loggers on seabirds," *Communications biology*, vol. 3, no. 1, pp. 1–9, 2020.

[37] T. Maekawa, D. Higashide, T. Hara, K. Matsumura, K. Ide, T. Miyatake, K. D. Kimura, and S. Takahashi, "Cross-species behavior analysis with attention-based domain-adversarial deep neural networks," *Nature communications*, vol. 12, no. 1, pp. 1–11, 2021.

[38] M. Tachikawa, T. Maekawa, and Y. Matsushita, "Predicting location semantics combining active and passive sensing with environment-independent classifier," in *the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, 2016, pp. 220–231.

[39] N. Yoshimura, J. Morales, and T. Maekawa, " OpenPack: Public multimodal dataset for packaging work recognition in logistics domain," Jan. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.5909087