

OpenPack: A Large-scale Dataset for Recognizing Packaging Works in IoT-enabled Logistic Environments

Naoya Yoshimura, Jaime Morales, Takuya Maekawa, Takahiro Hara
Graduate School of Information Science and Technology, Osaka University, Japan
{yoshimura.naoya, jaime.morales, maekawa, hara}@ist.osaka-u.ac.jp

Abstract—Unlike human daily activities, existing publicly available sensor datasets for work activity recognition in industrial domains are limited by difficulties in collecting realistic data as close collaboration with industrial sites is required. This also limits research on and development of methods for industrial applications. To address these challenges and contribute to research on machine recognition of work activities in industrial domains, in this study, we introduce a new large-scale dataset for packaging work recognition called OpenPack. OpenPack contains 53.8 hours of multimodal sensor data, including acceleration data, keypoints, depth images, and readings from IoT-enabled devices (e.g., handheld barcode scanners), collected from 16 distinct subjects with different levels of packaging work experience. We apply state-of-the-art human activity recognition techniques to the dataset and provide future directions of complex work activity recognition studies in the pervasive computing community based on the results. We believe that OpenPack will contribute to the sensor-based action/activity recognition community by providing challenging tasks. The OpenPack dataset is available at <https://open-pack.github.io>.

Index Terms—datasets, work activity, activity recognition

I. INTRODUCTION

In factories and logistics centers, human workers continue to perform important roles in adapting to the fast-changing demands of customers and suppliers [1], [2]. In 2021, Amazon shipped over 5 billion packages in the U.S.¹ with 1.6 million employees², emphasizing the need for efficient shipping in supply chains. Digitization is applied in industry to streamline human work and assist in decision-making as part of Industry 4.0. In Industry 4.0, data from sensors and IoT devices (e.g., connected handheld terminals) is used to recognize and streamline human work activities. Therefore, activity recognition for human workers in industry has become a notable research topic in pervasive computing [3]–[8].

However, the following challenges should be addressed to enhance work activity recognition studies.

- **Lack of datasets for industrial domains:** Fig. 1 shows a typical series of packaging tasks iterated several times, with each iteration of the process (i.e., period) comprising a sequence of operations in which the acceleration data indicate the complexity of operations. However, the amount of publicly

available datasets for industrial domains containing complex activities remains limited, and many of the available activity recognition datasets focus only on simple daily activities (e.g., walking and running) [9]–[11].

- **Limited modality:** Many public datasets for manual tasks provide only vision-related modalities [12]–[14]. However, because many types of manufacturing equipment and storage systems are installed in industrial environments, occlusion tends to pose challenges in applying vision-only approaches.

- **Unavailability of readings from IoT-enabled devices:** Although digitization is progressing in actual industrial domains as part of the development of Industry 4.0, to our knowledge, no datasets on activity recognition that include both sensory data on human motions and readings from IoT-enabled devices operated by the human workers are publicly available.

- **Lack of rich metadata:** Many of the available datasets do not provide a rich set of metadata related to manual works such as a set of the items to be packed, which limits the understanding of recognition results and the design of new, enhanced research tasks.

To address the challenges regarding complex work recognition, in this study, we propose a new multimodal dataset for packaging work recognition in logistics, called OpenPack (Fig. 1). OpenPack consists of 20,161 instances of activities (operations) and 53,286 instances of actions with 9 types of modalities captured from 16 distinct subjects with various levels of experience performing packaging tasks. The main features of OpenPack are summarized as follows.

- (1) **OpenPack is the largest multimodal work activity dataset in the industrial domain**, including sensory data from body-worn inertial measurement units (IMUs), depth images, and LiDAR point clouds for use in research on multi-/crossmodal, IMU-only, and vision-only work activity recognition according to conditions in an expected target environment.

- (2) **OpenPack provides a rich set of metadata** such as subjects' levels of experience in packaging work as well as their physical characteristics, enabling the design of various research tasks (e.g., assessment of workers' performance) in addition to basic work activity recognition.

- (3) **OpenPack is the first large-scale dataset for complex packaging work recognition that contains readings from IoT-enabled devices.** Leveraging high-confidence readings

¹<https://www.seattletimes.com/business/amazon-is-upss-biggest-customer-and-biggest-competitive-threat>

²<https://www.macrotrends.net/stocks/charts/AMZN/amazon/number-of-employees>

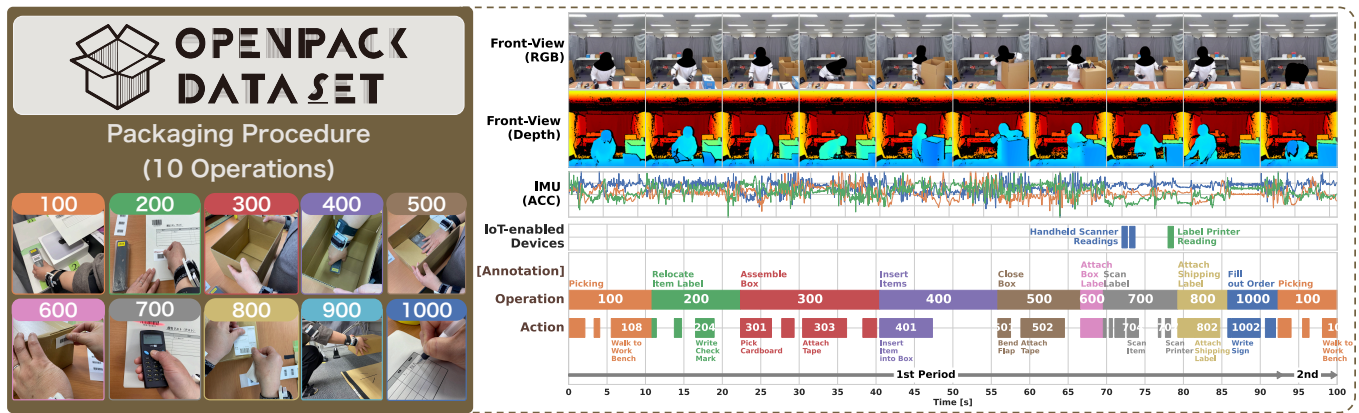


Fig. 1: Illustration and example sensor data of the OpenPack dataset. A subject iterated a typical series of packaging works, with each iteration of the process (i.e., period) comprising a sequence of complex operations.

from IoT-enabled devices, which strongly relate to the activities performed (e.g., handheld scanner readings relate to “scan label” operations), is expected to be a key enabler to precise recognition in real-world applications.

The key contributions of this research are:

- To the best of our knowledge, OpenPack is the largest dataset on industrial work activity recognition. We expect this data to contribute to the pervasive computing community by providing a challenging task, that is, to perform complex activity recognition via sensor and sparse IoT data.
- We apply state-of-the-art recognition methods to OpenPack and provide future research directions for complex work activity recognition by analyzing the recognition results.

II. RELATED WORK

Although many multimodal, vision-based, and IMU-based datasets for daily activity recognition have been made publicly available [15]–[18], the number of publicly available datasets for work activity recognition in industrial domains remains limited. Table I summarizes the attributes of datasets on human activities and manual labor³. To the best of our knowledge, the LARa dataset [5], [19] is the only dataset on work activity recognition in logistics. However, this dataset does not have class labels of types of operations such as packing items and assembling a box, and the subjects has no or limited experience in working at real logistics centers. Moreover, the above datasets do not provide readings from IoT-enabled devices or a rich set of metadata.

The InHARD dataset [13] and the ABC Bento packaging dataset [20] are designed to accelerate human-robot collaboration in industrial settings. The InHARD dataset consists of RGB and 3D keypoint data from 16 subjects collected while they were assembling various parts and components. The ABC Bento Packaging dataset is a dataset that captures activities

related to bento packaging. This dataset focuses on common mistakes made by bento manufacturers, such as “forgetting to put in ingredients,” and provides labels only for outlier types. The ABC Bento Packaging dataset is quite small to be applied to data-driven algorithms, such as deep learning. These vision-based datasets also lack sensor data modalities. In contrast, OpenPack is a large-scale dataset containing 20,161 work operation instances with multimodal sensor data.

Datasets of various complex procedural activities are also available. Specifically, many multimodal/vision-based cooking activity datasets are available such as CMU-MMAC [15], 50 Salads dataset [21], Breakfast Actions Dataset [22], EPIC-KITCHENS [23], and the Cooking Activity Dataset [24]. Vision-based datasets focused on procedural activities other than cooking include IKEA-ASM [14] and Assembly101 [25] for assembling furniture and toys, and COIN [12], a collection of instructional videos collected from YouTube. As noted above, many multimodal or vision-based datasets on manual tasks in daily life have been made publicly available. In contrast, the availability of public multimodal datasets for industrial domains remains limited, and OpenPack is the first large-scale dataset for activity recognition in industrial domains. This may be attributed to the difficulties in collecting datasets for industrial domains compared to everyday tasks. Collecting data for industrial applications requires close collaboration with industrial engineers working in an actual target environment to coordinate an experimental environment with various equipment for the target task, to define a set of activity labels by obtaining an actual work instruction document used in the target industrial environment, and to employ workers with experience in the target task as research subjects.

III. DEVELOPING OPENPACK DATASET

OpenPack (<https://open-pack.github.io>) is the first multimodal large-scale dataset for activity recognition in industrial domains. 16 distinct participants packed 3,956 items in 2,048 shipping boxes in total, and the total duration of our dataset is 53.8 hours, consisting of 104 data collection sessions. Open-

³Activity Class: When action classes are defined in a hierarchical manner, the numbers of classes in different levels are shown with separator “+”. In the case of OpenPack, they correspond to # of operations and actions.

TABLE I: Overview of public datasets of human activities/works. D: Depth, Acc: Acceleration data, Gyro: Gyroscopic data, Ori: Orientation sensor data, EDA: Electrodermal activity, BVP: Blood volume pulse, Temp: Temperature.

| Domain | Datasets | Type of Task | Recording Length | Activity Class ³ | # of Work Period | Annotated Instances | Subjects | Modality | IoT Devices | View | Year |
|---------------------|-------------------|--------------------|------------------|-----------------------------|------------------|---------------------|----------|---|-------------|-------|---------------|
| Multi-modal | MHAD | Daily Actions | 82m | 11 | N/A | 660 | 12 | RGB+D+Keypoints+Acc+Mic | No | 12 | 2013 |
| | UTD-MHAD | Daily Actions | 9m | 27 | N/A | 180 | 8 | RGB+D+Keypoints+Acc+Gyro | No | 1 | 2015 |
| | MMAAct | Daily Actions | 17h35m | 37 | N/A | 36,764 | 20 | RGB+D+Keypoints+Acc+Gyro+Ori+WiFi+Pressure | No | 4+Ego | 2019 |
| Cooking | CMU-MMAC | Cooking | - | 5 | 186 | 186 | 39 | RGB+D+Keypoints+Acc+Mic | Yes (RFID) | 5 | 2010 |
| | 50 Salads | Cooking | 4h | 52 | 50 | 2,967 | 25 | RGB+D+Acc | Yes (Acc) | 1 | 2013 |
| Procedural Activity | COIN | Instruction Video | 476h38m | 180 | N/A | 46,354 | N/A | RGB (Youtube) | No | N/A | 2019 |
| | IKEA-ASM | Furniture Assembly | 35h16m | 33 | 371 | 16,764 | 48 | RGB+D+Keypoints | No | 3 | 2021 |
| | Assembly101 | Toy Assembly | 42h+ | 202 | 362 | 1M+ | 53 | RGB | No | 12 | 2022 |
| Industrial | InHARD | Industrial Actions | 18h30m | 14 + 72 | 38 | 4,800 | 16 | RGB+Keypoints (3D) | No | 3 | 2020 |
| | ABC Bento | Bento Packaging | 3h22m | 10 | 199 | 151 | 4 | MOCAP | No | 1 | 2021 |
| | LARa | Picking Packaging | 14h50m | 8 + 19 | 324 125 | 8,878 2,103 | 16 10 | RGB+Keypoints (3D)+Acc | No | 1 | 2020 +2022 |
| | OpenPack (v1.0.0) | Packaging | 53h50m | 10 + 32 | 2,048 | 20,161 | 16 | D+Keypoints+LiDAR+Acc+Gyro+Ori+EDA+BVP+Temp | Yes | 2 | 2023 |

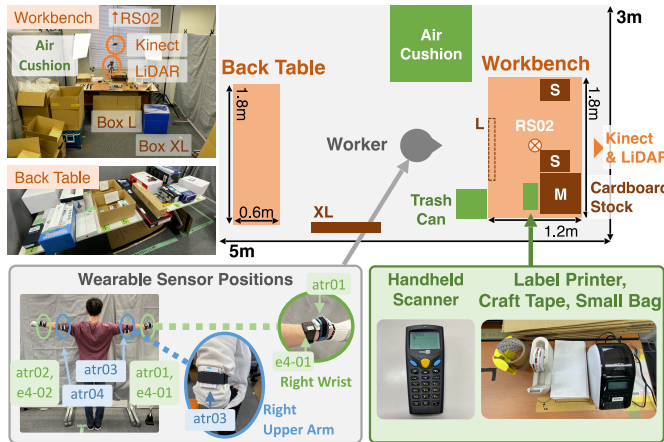


Fig. 2: Environmental setup and wearable sensor positions

Pack is the largest industrial dataset that includes both vision and wearable sensor data with precise labels by annotators.

In the following, we present an overview of the target activity of the dataset, the process by which it was collected, and how the data were annotated.

A. Packaging Work

As shown in Fig. 1, a typical series of complex operations is iterated, with each iteration (i.e., period) comprising a sequence of operations such as assembling a shipping box and filling the box with items. In a given work period, a worker processes a single shipping order consisting of 10 pieces of work. That is, the worker picks items in the shipping order, double-checks the items, assembles a shipping box, fills the

box with the items, and so forth to complete the order. When performing specific operations, the worker uses IoT-enabled devices such as a handheld barcode scanner, and the operation is recorded and transmitted by the device.

Because the size of items to be packed, the number of items, and the size of shipping items depend on shipping orders, sensor data collected in different work periods and the duration of the same operation in different work periods vary. The task of recognizing specific operations is challenging owing to these characteristics of packaging work.

B. Data Collection

1) *Collection Environment*: We collected data in a dedicated environment shown in Fig. 2. With the help of industrial engineers, we constructed a 3m × 5m environment designed to simulate an actual workspace in a warehouse. The environment mainly comprised a workbench, a back table on which items were placed after being picked from shelves by another worker, boxes containing air cushions, and a trash can. A handheld barcode scanner, a printer, and craft tapes used for packaging were located on the right side of the table. Four types of cardboard boxes were available for packing, including small, medium, large, and extra-large sizes.

2) *Subjects*: We invited 16 subjects (5 males and 11 females) to participate in our data collection process. The ages of the subjects were ranged from 20s and 50s. 12 subjects had experience in packaging work ranging from 1 month to 4 years. In addition, 4 subjects did not have work experience. Data from these four subjects can be valuable for recognizing the operations of workers newly involved in the work environment and analyzing the learning curve.

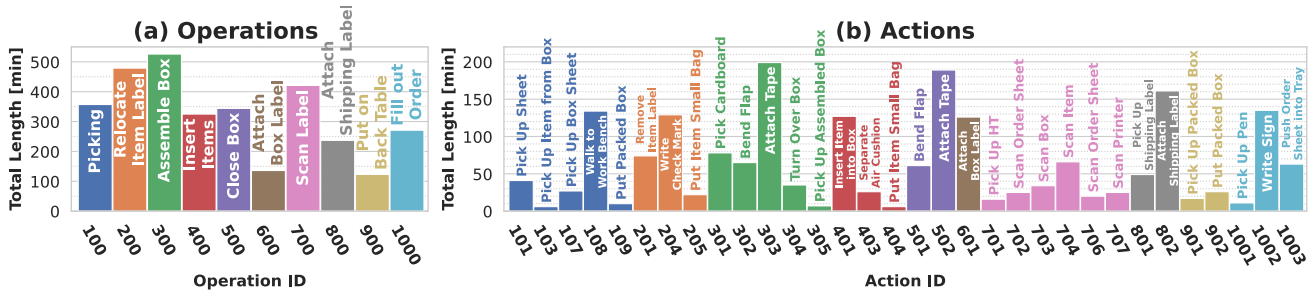


Fig. 3: Distribution of the total lengths of the annotated activities. The horizontal axis shows activity IDs.

One subject was left-handed. Each subject was assigned a consistent identifier throughout the entire dataset.

3) *Data Modalities*: OpenPack provides nine data modalities, including acceleration, gyroscope, quaternion, blood volume pulse (BVP), electrodermal activity (EDA) data, body temperature, keypoints, a LiDAR point cloud, and depth images. Fig. 2 illustrates the positions of wearable sensors. Four IMU units were attached to the subject’s left and right wrists and upper arms to collect acceleration data on three axes, as well as gyroscope and quaternion data at 30 Hz. In addition, two Empatica E4 sensors were attached to the subject’s left and right wrists to collect BVP and EDA signals at 64 Hz and 4 Hz, respectively, in addition to acceleration data at 32 Hz. Kinect and LiDAR sensors were installed as front-view cameras and RealSense (RS02) as a top-view camera as shown in Fig. 2. The LiDAR sensor was considered effective in accurately tracking the subject’s position when they were away from the workbench. We included the BVP and EDA sensors because they are expected to be used for analyzing the internal status of subjects in specific situations. OpenPack also provides operational logs of IoT-enabled devices, i.e., the handheld scanner and label printer, in the environment. The operational logs of the handheld scanner, for example, contain a time-stamp of a scan and an identifier of a scanned item. These are highly reliable sources of information for recognizing the scanning operation.

4) *Data Collection Procedure*: Before data collection, each subject received instructions related to the outline of the experiment and the operations to be performed based on the instruction document. Subsequently, we obtained the informed consent of the subjects, who then practiced the packaging work by performing work periods up to five times. Subsequently, we activated and calibrated the sensors and attached the wearable sensors to the subjects. The subjects iterated up to five data collection sessions within 6 h (including a 60-min lunch break). At the beginning of each session, the subject received 20 order sheets and then sequentially processed the sheets. That is, the subject completed 20 shipping boxes in the session. A 15-min break was included between two consecutive sessions.

5) *Scenarios*: The difficulties in packaging work recognition depend on various factors found in logistics centers. Four scenarios are prepared to simulate the difficulties. **Scenario 1**: This is the most simple scenario in which workers follow the

work instructions as accurately as possible. **Scenario 2**: Some logistics centers do not require workers to strictly follow work instruction documents. Here, we encouraged the subjects to alter the procedure of operations for more efficient operation. **Scenario 3**: Depending on the items to be packed and/or surrounding situations, a worker performs irregular actions or activities. Here, three irregular situations/actions were added: (1) use an already assembled box, (2) put small items into an additional bag, and (3) pick up items for several order sheets at the same time. **Scenario 4**: Due to seasonal events or flash sales, task volumes for each worker may temporarily increase and workers have to move more quickly. Here, we rushed the subjects by introducing an auditory alarm to simulate a busy working time.

6) *Metadata*: OpenPack provides a rich set of metadata, which is mainly composed of subject- and order-related metadata. The subject-related metadata contains information regarding each of the participants’ experience in packaging tasks, as well as their dominant hand, gender, and age.

Order-related metadata contains information regarding items and an order sheet processed by a subject in a work period. OpenPack assumes an online order management system and provides information regarding an identifier of an order and a set of items to pack in the order. The management system also stores information regarding an identifier, product code, and the size of each item. These identifiers are used to manage items in an order management system. In contrast, product codes are unique numbers for each item, which are widely used in retail sales and enable information to be retrieved regarding an item, such as product name, product type, and price. OpenPack also provides this information.

C. Annotations

After data collection, the data were labeled by three expert annotators by reference to the RGB images with the help of industrial engineers. It took about 1 year to complete the annotation task. Two types of labels are available: (1) work operation and (2) action.

1) *Work Operations*: Operation classes used in OpenPack were defined based on an instruction document used in an actual logistics center. The document specifies a sequence of operations performed by a worker, and each worker in the center performs operations according to the document. Therefore, the basic activities performed by all workers, i.e.,

operations, were used to label the dataset. Our dataset contains ten classes of operations shown in Fig. 3 (a). Note that many other logistics centers also utilize patterns of operations very similar to those used in this study.

2) *Actions*: An instructional document also contains a description of each operation that explains how to perform the operation. For example, a description of the “relocate item label” operation is given as “*Remove the label from the items and place it on the bottom margin of the packaging list. Check the product name and quantity on the list and label with a ballpoint pen.*” Based on the description, we also defined 32 action classes included in operations as shown in Fig. 3 (b). For example, Fig. 1 shows that the “assemble box” operation is composed of four actions, including “pick up cardboard,” “bend flap,” “attach tape,” and “turn over box.” The action classes are useful for a manager in a logistics center to assess the status of a job in progress in detail.

Note that action labels were not assigned to every time step, as shown in Fig. 1. We did not annotate all the atomic actions included in the operations because there are many meaningless and inconsistent body movements, for example, the transition from one action to the next. Therefore, we created action labels that specified meaningful and consistent atomic actions, as described in the work instruction document.

For the details of the data collection and annotation, see the dataset website (<https://open-pack.github.io>).

IV. DATASET ANALYSIS

In this section, we provide some statistics on the dataset to show the diversity of packaging work in terms of activity length and period length.

1) *Annotation Summary*: Figure 3 summarizes the total length for each class.

The total lengths of operation and action labels are 53.8 hours and 37.8 hours, respectively. The numbers of operation and action labels are 20,161 and 53,286, respectively. As shown in Fig. 3, there is considerable variation in the total lengths of labeled segments in the activities. This reflects a large difference in the time required to perform different work operations/actions and their occurrence frequency. To recognize these activities in OpenPack, this class imbalance problem must be considered. This could be challenging, especially for action recognition.

2) *Variation in Operations and Period Length*: Even for the same work activities, the data exhibit a wide variation depending on different situations. Fig. 4 (a) shows the distribution of the length of two operations; “relocate item label,” and “attach box label.” Simple operations such as “attach box label,” which are not affected by the number of items in an order sheet, only exhibit small variations in length. In contrast, operations such as “relocate item label,” which depend on the numbers and sizes of items, exhibit long-tailed distributions.

Figure 4 (b1) shows the distribution of period lengths calculated for each user and session. The working speeds of the subjects vary significantly. For example, the work speed of U0204 is particularly fast; this worker completed a single

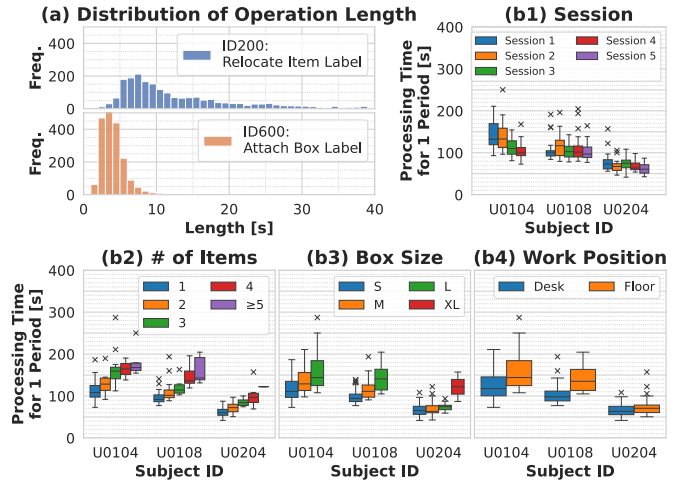


Fig. 4: (a) Distribution of the length of the two operation classes. (b) Distribution of the period lengths with different sessions, item numbers, box sizes, and work positions.

period in 70.1 seconds on average, while the average for all workers is 96.4 seconds. The effects of the number of items to be packed, box size, and work location, on the period length are summarized in Fig. 4 (b2–b4). As the number of items and box size increases, the period length tends to increase. In addition, work performed with boxes placed on the floor tended to take longer time than work performed on a workbench, owing to the time required to move items and tools. Although these three factors are not independent, they have a significant impact on work activities and makes recognition difficult.

V. EVALUATION & BENCHMARK

A. Evaluation Methodology

We benchmarked state-of-the-art activity recognition methods on the OpenPack dataset in the four typical settings.

Acceleration data from the workers’ left wrists, i.e., non-dominant hands, were used as inputs. Models were trained to recognize 10 work operations at 1 Hz resolution. Macro average of F1-measure calculated for each scenario was used as evaluation metrics. We prepared the following 6 models as baselines: CNN [26], U-Net [27], DeepConvLSTM (DCL) [26], DCL with Self-attention (DCL-SA) [28], Conformer-HAR [29], and LOS-Net(-R) [8]. Models are trained with the Adam optimizer with ExponentialLR learning rate scheduler for up to 500 epochs with early stopping. We repeated the training five times with different random seeds, and the average score is shown.

B. Results

1) *Cross-user Activity Recognition with a Sufficient Amount of Training Data (Data-rich Setting)*: This setting is designed to confirm the upper bounds of recognition performance of complex work activities for deep models that require large amounts of training data. This data-rich setting assumes that

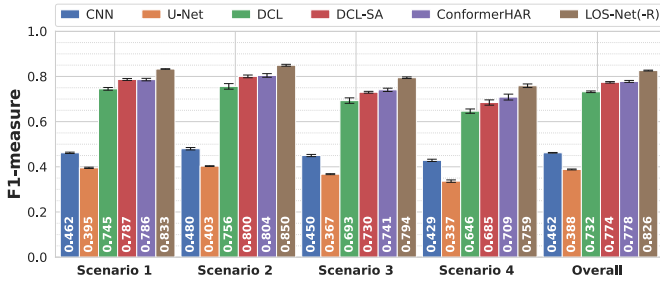


Fig. 5: Results of cross-user work operation recognition with a sufficient amount of training data

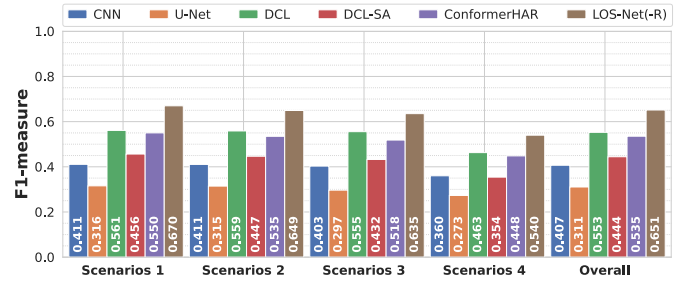


Fig. 7: Results of work operation recognition with a limited amount of training data from a known worker.

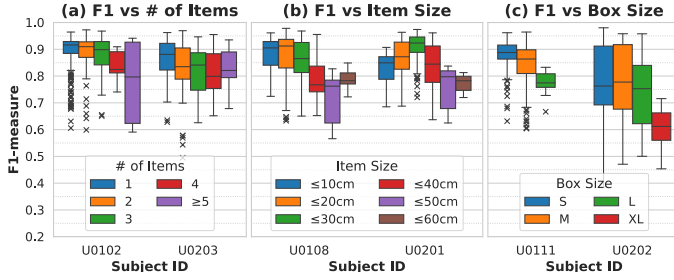


Fig. 6: Distributions of F1-measure in different conditions (LOS-Net(-R); Cross-user setting)

a sufficient amount of training data from other users are provided, i.e., the cross-user setting. Therefore, we conducted the leave-one-subject-out cross validation.

Figure 5 shows the results of work operation recognition. Long-term context must be extracted to estimate the class to which each action belongs based on the previous and next actions because most operations are similar to each other in their movements. Thus, LOS-Net(-R), which has a module for long-term context extraction, achieved the highest F1-measures of 0.83 in Scenario 1. In contrast, CNN and U-Net models perform relatively poorly at extracting long-term contexts, as shown in the experimental results. DCL-SA slightly outperformed DCL, which demonstrates the effectiveness of the self-attention mechanism on the work activity recognition task. The scores for Scenario 4 were lower than those of the other scenarios because the subjects were rushed. In reality, it is difficult to prepare such a large amount of training data like this benchmark setting. It is necessary to develop a model that can recognize with the same high accuracy even if the amount of training data is limited.

Figure 6 shows the distributions of F1-measures on the period basis in different conditions: (a) the number of items to be packaged, (b) the size of items, and (c) the size of used boxes. The recognition performance decreases with an increase in the item count or size because there are limited samples corresponding to them in the training dataset. Specifically, the difference in the item size significantly affects the worker's movements. As shown in Fig. 6 (b), the F1-measure of U0201 for " ≤ 30 cm" is higher than that of the other conditions. This might be owing to the larger hand movements in " ≤ 30 cm", which makes recognition easier. In reality, the item distribution

is also biased, which may degrade the recognition performance. Developing techniques to mitigate this degradation by utilizing item metadata would be an interesting research topic.

2) *Work Activity Recognition with a Limited Amount of Training Data from a Known Worker (Data-scarce Setting):* When work instruction documents are not very strict, workers perform tasks in different ways, which makes cross-user operation recognition more difficult. However, preparing a sufficient amount of labeled training data from a target worker is expensive. Therefore, the objective of this setting was to investigate the performance of recognizing workers' activities with a limited amount of training data collected from a target worker. In this data-scarce setting, a limited amount of training data (1 session of data) from each target subject was used. Each model was trained on data from the 3rd session and calculated the F1-measure for each scenario using the remaining data from the same subject. There are only 20 periods in one session, but the annotation took roughly 5 hours.

Figure 7 shows the results of the average of F1-measure for all users in each scenario. In Scenario 1, LOS-Net(-R) achieved the F1-measure of 0.67, but they were 0.16 pts lower than the results of the first setting. The recognition performances largely deteriorated in this data-scarce setting even when data from the target subject was included in the training set. Interestingly, F1-measures of U-Net and DCL-SA were lower than CNN and DCL, respectively, in the data-scarce setting. A model with large trainable parameters or a self-attention module is likely to require more training data. Therefore, methods should be developed to facilitate training of such state-of-the-art modules and architectures, such as self-attention, with limited training data [7].

3) *Cross-user Activity Recognition with Different Amount of Training Data:* We investigate the relationship between the amount of training data used and recognition accuracy. We used a fixed test set and the model was trained by varying the number of remaining workers. Data from four workers, i.e., U0104, U0108/U0203, U0110/U0110, and U0207, was used as a test set.

Fig. 8 shows the results. In Scenario 1 and Scenario 2, recognition performance improved significantly until the number of training subjects reached a total of 4, after which performance improvements became moderate. In contrast, in Scenarios 3 and Scenario 4, the score gradually improved

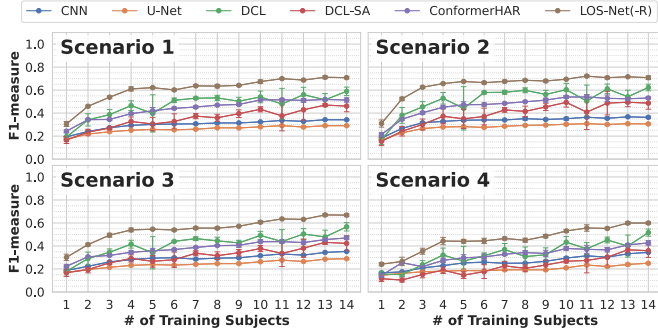


Fig. 8: Results of work activity recognition with a limited amount of training data from source workers

with more training subjects, indicating the negative effect of variations in sensor data on recognition performance in these scenarios. Therefore, the development of methods that can deal with variations in the data is crucial. For example, for Scenario 4, in which the subjects were rushed, a model that is robust against differences in working speed must be developed, such as using data augmentation or a bottom-up approach that detects actions that are less sensitive to the differences in speed first and then estimates work operations.

4) *Recognition with Multi-/limited-modalities*: Multi-modal activity recognition is a hot research topic and this approach has the potential to achieve precise recognition compared to single-modal recognition. However, because it is not practical to ask workers to wear multiple sensors on a daily basis, recognition technologies using limited modalities are an important research focus area. Here, we evaluated the recognition performance of the various sensor combinations. We used the setting as Section V-B1, i.e., leave-one-subject-out CV. LOS-Net(-R) was used for wearable sensor modalities with early fusion techniques [30] and ST-GCN [31] was used for keypoints.

Figure 9 shows the results. Since most subjects are right-handed, the recognition accuracy of the combination including the right wrist sensor is higher than that of the other combinations. However, the addition of gyros and quaternions to capture more detailed hand movements resulted in limited improvement in accuracy with the simple early fusion techniques. ST-GCN exhibited relatively lower performance, probably due to the occlusions of boxes. Based on these results, we believe there is considerable room for performance improvement by sensor fusion.

VI. RESEARCH DIRECTIONS AND POTENTIAL TASKS

In this section, based on the analysis and benchmark results, we highlight the possible research directions that can be explored with OpenPack.

- **Metadata-aided activity recognition**: The benchmark results showed that recognition performances for some models was low for activities affected by conditions such as the combination of items to pack. Because information about an order sheet that a worker is currently processing is commonly managed by an online order management system, the perfor-

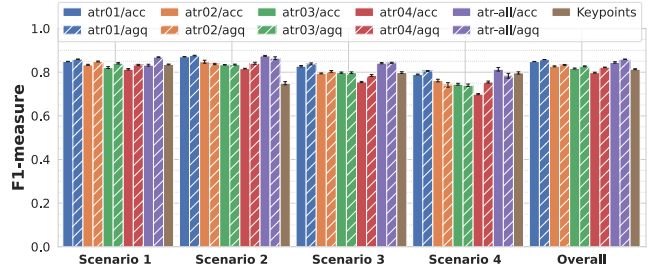


Fig. 9: Results of work operation recognition with various sensor combinations. “atr01” to “atr04” are IMU sensors. The input modality is acceleration only for “/acc” and a combination of acceleration, gyro, and quaternion for “/agq”.

mance of activity recognition methods can be enhanced with information about the order such as the number and size of items. For example, we can switch activity recognition models depending on the characteristics of orders, and information about an order, such as the number of items, can be used as prior knowledge because the duration of related operations is proportional to the number of items.

- **Speed-invariant activity recognition**: The working speed significantly affected recognition performance in the benchmark experiment. The working speed can vary between workers and depending on situations. Therefore, development of speed-invariant activity recognition is important, such as data augmentation techniques enabling recognition performance that is robust to variations in working speed and speed-agnostic feature extraction methods.

- **Fusion with high-confidence modalities**: The class imbalance problem and sensor data variations are inevitable in complex work activity recognition, and result in performance deterioration for specific difficult activity classes. Therefore, fusion with high-confidence modalities such as readings from IoT-enabled devices, such as a bar-code scanner, is crucial.

In addition to the operation and action recognition tasks using sensor data, OpenPack with a set of rich metadata and annotations enables various designs of research tasks, which includes the following: (1) transfer learning across sensor positions, across subjects, and across modalities [18], (2) skill assessment using sensor data and metadata related to work experience as ground truth [32], (3) counting the number of necessary actions or the number of packed items using sensor data [33], (4) estimating workers’ levels of fatigue using sensor and physiological data, and (5) detecting mistakes and accidents in the work process [25].

VII. CONCLUSION

This study presented a new large-scale dataset for packaging work recognition called OpenPack dataset. Based on the analysis and benchmark results on OpenPack, we provided future research directions for complex work activity recognition. We believe that human activity recognition methods developed based on OpenPack are applicable to many complex work activities in industrial domains, and OpenPack can be used as a baseline dataset for complex work activity recognition.

Ethical Statement: This study is approved by the ethical committee of the institute of the authors.

Acknowledgement: This work is partially supported by JSPS KAKENHI JP21H03428, JP21H05299, JP21J10059, JST ACT-X JPMJAX200T, and JST Mirai JP21473170. We greatly appreciate Chikako Kawabe, Kana Yasuda, and Makiko Otsuka for their efforts in developing the OpenPack dataset. We would also like to express our appreciation to Dr. Namioka, Toshiba Corporation for the support in data collection.

REFERENCES

- [1] R. Michel, "2016 warehouse/dc operations survey: Ready to confront complexity," Nov 2016. [Online]. Available: https://www.logisticsmgmt.com/article/2016_warehouse_dc_operations_survey_ready_to_confront_complexity
- [2] 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, p. 101864, 2020.
- [3] S. Inoue, P. Lago, T. Hossain, T. Mairitha, and N. Mairitha, "Integrating activity recognition and nursing care records: The system, deployment, and a verification study," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 3, 2019. [Online]. Available: <https://doi.org/10.1145/3351244>
- [4] 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, pp. 1–23, 2019.
- [5] F. Niemann, C. Reining, F. Moya Rueda, N. R. Nair, J. A. Steffens, G. A. Fink, and M. Ten Hompe, "Lara: Creating a dataset for human activity recognition in logistics using semantic attributes," *Sensors*, vol. 20, no. 15, p. 4083, 2020.
- [6] 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, 2020.
- [7] J. Morales, N. Yoshimura, Q. Xia, A. Wada, Y. Namioka, and T. Maekawa, "Acceleration-based human activity recognition of packaging tasks using motif-guided attention networks," in *Proceedings of the IEEE International Conference on Pervasive Computing and Communications*, 2022, pp. 1–12.
- [8] N. Yoshimura, T. Maekawa, T. Hara, A. Wada, and Y. Namioka, "Acceleration-based activity recognition of repetitive works with lightweight ordered-work segmentation network," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 2, 2022.
- [9] D. Roggen, A. Calatroni, M. Rossi, T. Holleczeck, K. Förster, G. Tröster, P. Lukowicz, D. Bannach, G. Pirkel, and A. Ferscha, "Collecting complex activity datasets in highly rich networked sensor environments," in *Proceedings of the International Conference on Networked Sensing Systems*, 2010, pp. 233–240.
- [10] A. Reiss and D. Stricker, "Introducing a new benchmarked dataset for activity monitoring," in *Proceedings of the International Symposium on Wearable Computers*, 2012, pp. 108–109.
- [11] B. Barshan and M. C. Yüsek, "Recognizing daily and sports activities in two open source machine learning environments using body-worn sensor units," *Computer Journal*, vol. 57, no. 11, pp. 1649–1667, 2014.
- [12] Y. Tang, D. Ding, Y. Rao, Y. Zheng, D. Zhang, L. Zhao, J. Lu, and J. Zhou, "COIN: A large-scale dataset for comprehensive instructional video analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1207–1216.
- [13] M. Dallel, V. Havard, D. Baudry, and X. Savatier, "Inhard-industrial human action recognition dataset in the context of industrial collaborative robotics," in *Proceedings of the IEEE International Conference on Human-Machine Systems*, 2020, pp. 1–6.
- [14] Y. Ben-Shabat, X. Yu, F. Saleh, D. Campbell, C. Rodriguez-Opazo, H. Li, and S. Gould, "The IKEA ASM dataset: Understanding people assembling furniture through actions, objects and pose," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 847–859.
- [15] E. H. Spriggs, F. De La Torre, and M. Hebert, "Temporal segmentation and activity classification from first-person sensing," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2009, pp. 17–24.
- [16] F. Offi, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley mhad: A comprehensive multimodal human action database," in *Proceedings of the IEEE Workshop on Applications of Computer Vision*, 2013, pp. 53–60.
- [17] 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 *Proceedings of the IEEE International Conference on Image Processing*, 2015, pp. 168–172.
- [18] Q. Kong, Z. Wu, Z. Deng, M. Klinkigt, B. Tong, and T. Murakami, "MMAct: A large-scale dataset for cross modal human action understanding," in *Proceedings of the IEEE International Conference on Computer Vision*, October 2019, pp. 8657–8666.
- [19] F. Niemann, S. Lüdtkke, C. Bartelt, and M. Ten Hompe, "Context-aware human activity recognition in industrial processes," *Sensors*, vol. 22, no. 1, p. 134, 2022.
- [20] S. S. Alia, K. Adachi, N. Nahid, H. Kaneko, P. Lago, and S. Inoue, "Bento packaging activity recognition challenge," 2021. [Online]. Available: <https://dx.doi.org/10.21227/cwhs-t440>
- [21] S. Stein and S. J. McKenna, "Combining embedded accelerometers with computer vision for recognizing food preparation activities," in *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2013, pp. 729–738.
- [22] H. Kuehne, A. Arslan, and T. Serre, "The language of actions: Recovering the syntax and semantics of goal-directed human activities," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 780–787.
- [23] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray, "Scaling egocentric vision: The EPIC-KITCHENS dataset," in *Proceedings of the European Conference on Computer Vision*, 2018.
- [24] P. Lago, S. Takeda, K. Adachi, S. S. Alia, M. Matsuki, B. Benai, S. Inoue, and F. Charpillet, "Cooking activity dataset with macro and micro activities," 2020. [Online]. Available: <https://dx.doi.org/10.21227/hyzz-9m49>
- [25] F. Sener, D. Chatterjee, D. Shelepov, K. He, D. Singhania, R. Wang, and A. Yao, "Assembly101: A large-scale multi-view video dataset for understanding procedural activities," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21 096–21 106.
- [26] F. Ordóñez and D. Roggen, "Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, 2016.
- [27] Y. Zhang, Z. Zhang, Y. Zhang, J. Bao, Y. Zhang, and H. Deng, "Human activity recognition based on motion sensor using U-Net," *IEEE Access*, vol. 7, pp. 75 213–75 226, 2019.
- [28] S. P. Singh, M. K. Sharma, A. Lay-Ekuakille, D. Gangwar, and S. Gupta, "Deep convlstm with self-attention for human activity decoding using wearable sensors," *IEEE Sensors*, vol. 21, no. 6, pp. 8575–8582, 2020.
- [29] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.
- [30] S. Münzner, P. Schmidt, A. Reiss, M. Hanselmann, R. Stiefelhagen, and R. Dürichen, "CNN-based sensor fusion techniques for multimodal human activity recognition," in *Proceedings of the ACM International Symposium on Wearable Computers*, 2017, pp. 158–165.
- [31] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018, pp. 7444–7452.
- [32] Q. Xia, A. Wada, T. Yoshii, Y. Namioka, and T. Maekawa, "Comparative analysis of high-and low-performing factory workers with attention-based neural networks," in *Proceedings of International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, 2022, pp. 469–480.
- [33] Y. Nishino, T. Maekawa, and T. Hara, "WeakCounter: Acceleration-based repetition counting of actions with weakly supervised learning," in *Proceedings of the International Symposium on Wearable Computers*, 2021, pp. 144–146.