A Supporting Technique for Comparative Analysis of Factory Work by Skilled and Unskilled Workers using Neural Network with Attention Mechanism

Qingxin Xia[1] Atsushi Wada[2] Takanori Yoshii[2] Yasuo Namioka[2] Takuya Maekawa[1]

Abstract: This study presents a method for identifying significant activity differences between skilled and unskilled factory workers by a neural network with an attention mechanism using wrist-worn accelerometer sensor data collected in real manufacturing. To discover skill knowledge from skilled workers, industrial engineers manually identify activity differences between skilled and unskilled workers, which is likely to obtain skill knowledge, by watching video recordings or sensor data. However, a factory has many workers, and manual comparison between pairs of workers is time-consuming. We propose an attention-based neural network to visualize the importance of input segments that contribute to the classification output, which is useful to identify activity differences between workers. Our proposed method consists of three phases: (1) network training that classifies the skilled/unskilled worker classes and the attention layers can be trained to emphasize the input segments with significant activity differences, (2) detecting activity differences, which uses attentions to map the input segments to select candidates of input segments containing activity differences, and (3) identifying corresponding activities of the candidates of inputs on the other worker class. For instance, when an action of screwing by a skilled worker is identified by the attention mechanism, the corresponding sensor data segment of the screwing action by an unskilled is identified in his sensor data. To qualitatively analyze the result of candidates of input segments detected by the attention mechanism, we ask industrial engineers to assess skill knowledge within each candidate, 7 out of 11 segments imply useful skill knowledge.

1. Introduction

In a factory, the skill level of workers will influence the quality of products and deteriorate work productivity [11]. Therefore, it is an urgent need for industrial engineers to help unskilled workers to improve their work skills. In the current situation, engineers manually compare the unskilled worker with the skilled worker by watching videos or sensor data collected from every worker to find activity differences between them. By learning from the activity differences, the unskilled workers can modify their movements based on the skilled workers to get to the higher skill level fast [4]. Unfortunately, there are many workers in a factory, which is time-consuming for industrial engineers to manually locate the differences in minutes between skilled and unskilled groups.

Figure 1 shows two segments of raw sensor data collected from the right wrists of skilled and unskilled workers, respectively. As shown in the figure, a complete work process (i.e., work period) is comprised of a sequence of operations, such as "$Scan_1$", "Set", "$Label_1$", etc. However, some activities performed in an operation for the skilled and unskilled workers are different. For example, the pictures in Figure 1 presents different postures for the skilled/unskilled workers of attaching a label in the "$Label_3$" operation, where the skilled worker bends his spine for a smaller degree to perform the work satisfactorily. In contrast, the unskilled worker is more likely to suffer back pain caused by the large curvature of the spine. After discovering the differences between the skilled and unskilled workers, the industrial engineer can guide the unskilled workers for better performance. This study aims to identify activity differences between the skilled/unskilled workers by using acceleration data collected from the workers' wrists. However, it is challenging to compare activity differences by simply calculating sensor data differences. Since even the same activities in different periods performed by the same worker can cause sensor data differences. Therefore, we focus on identifying candidates of segments with significant activity differences

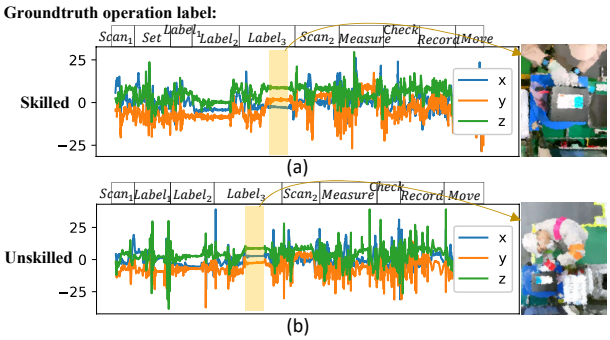

Fig. 1: An example of a series of operations collected from two workers' right wrists by using a 3-axes accelerometer. The rectangles indicate an activity difference between workers, which implies a skill knowledge about working posture to reduce body burden

[1] Osaka University, Graduate School of Information Science and Technology, Suita, Osaka, 5650871, Japan
[2] Toshiba Corporation, Corporate Manufacturing Engineering Center, Yokohama, Kanagawa, 2350017, Japan

between the skilled/unskilled workers.

In this study, we propose a neural network model based on the attention mechanism to help the engineer quickly find candidates of activity differences between the skilled and unskilled workers by using acceleration data collected from the workers' wrists. A neural network that is trained to classify a time-series for a period data into skilled/unskilled workers captures differences between the two classes, and the attention mechanism weights the input segments based on the data differences between the two classes. Therefore, we can extract segments of the input series with high attention values from attention layers as the candidate segments, which are likely to contain activity differences. However, attention only represents the weight of its corresponding input instance, after getting a candidate segment from an input in a worker class, it is necessary to identify the corresponding segment on the other worker class (e.g., as shown in Figure 1, if a candidate is an attaching label action in the "$Label_3$" for the skilled worker, the engineer should locate that action of the unskilled worker to compare the differences). However, it is difficult to directly find corresponding segments using raw sensor data, as the activities done by different workers are different. In this study, we leverage a clustering algorithm for the latent representations in the neural network to roughly eliminate individual differences. So that latent representations with similar activities can be clustered into the same cluster. Finally, the system will output every candidate of activity differences and their corresponding segments at the other worker class to the industrial engineer for skill assessment. We assume that significant differences between skilled and unskilled workers are more likely to contain skill knowledge that can help the unskilled workers promote their skills. So, we ask an industrial engineer to assess the skill knowledge inside the candidate segments detected by our proposed method to analyze our method quantitatively.

The contributions of this study are summarized as follows.

- Our method provides an attention-based neural network to automatically detect candidates of activity differences between skilled/unskilled workers from their acceleration data, which reduces the time costs for engineers to spend on navigating activity differences.
- We propose a novel way of detecting corresponding activities between skilled/unskilled workers by using a clustering approach for the latent representations of the neural network by eliminating individual differences.

In the rest of this paper, we first review studies for evaluating work performance. We then present the design of the proposed method and evaluate the method using sensor data collected in actual factories.

2. Related Work

There is a growing interest in studying individual's work performance to help amateurs improve their skill level fast. Many of the earliest work uses archival records, rating scales, and job knowledge tests [1] for work performance assessment. However these metrics have drawbacks, the results reported by workers are not accurate, and some unconscious skills of workers can not be detected [1]. Therefore, recent studies try to introduce activity data collected by electronic devices to complement traditional assessments. For example, Mirjafarl et al. [8] use mobile phones, wearables, and beacons to study behavior differences between higher and lower work performers in companies. Das Swain et al. [2] leverages sensors in commodity devices to quantify the daily activities of workers.

While the classifiers the above studies used are mainly applied for recognizing long duration of activities, which do not consider the small differences of same activities done by different persons, many recent studies implement an attention component in neural networks, trying to discover more detailed information. For instance, Zeng et al. [13] developed two attention models: temporal attention and sensor attention for detecting important signals and sensor modalities, respectively, which can be applied to identify the most important activities and sensor modalities for detecting Parkinson disease. Murahari et al. [9] also proposed an attention model as a data-driven approach for exploring the relevant temporal context in time-series data. Maekawa et al. [6] also present an attention-based neural network over animals' trajectories to detect segments in trajectories that are characteristic of one group, enabling biologists to focus on these specific segments and formulating new hypotheses.

Besides, there are many studies that use autoencoder to extract features for an individual dataset automatically. According to the special structure, the autoencoder transforms input data to a lower space and then retrieves it, which is very useful to compress data and extract features [10]. For guiding the autoencoder to generate better features, Xie et al. [12] and Guo et al. [3] focus on learning feature representations and a clustering assignment simultaneously for helping the network extract better features for what they expected.

In this study, we combine the advantages of the attention component and autoencoder. Our neural network tries to identify activity differences between skilled/unskilled workers with high attention values, while the corresponding data segment on the other worker by clustering the latent representations of the autoencoder.

3. Methodology

3.1 Preliminaries

In this study, at least two workers performing one type of work, one skilled worker and one unskilled worker. The data collected from each type of work are triaxial accelerometer data of the workers' both wrists. Each worker has multiple time-series data, with each time-series corresponding to a period.

Our study aims to identify candidates of segments with significant activity differences between the skilled/unskilled workers (e.g., the attaching label activity performed by the skilled worker) and detect the corresponding segments of each candidate on the rest of the worker (e.g., the cor-

responding attaching label activity found in the unskilled worker).

There are two assumptions in our study:

- Skill knowledge is more likely to exist in significant activity differences between skilled and unskilled workers.
- Data segments with significant activity differences contribute more to skilled/unskilled worker classification.

According to the first assumption, we can evaluate the effectiveness of our method in identifying activity differences by validating if skill knowledge exists in the candidates of segments or not. Based on the second assumption, we design our method that classifies a time series into a skill or unskilled class while focusing on important data segments.

3.2 Method Overview

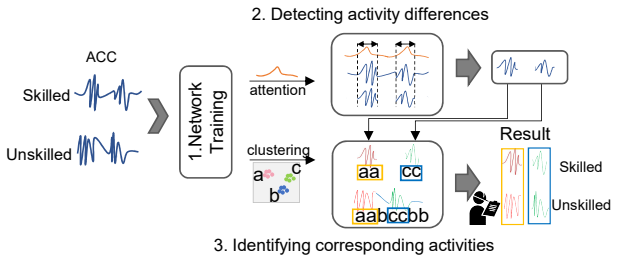

Fig. 2: Overview of the proposed method

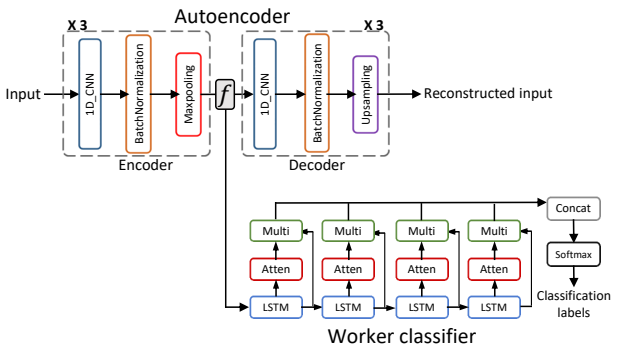

Fig. 3: Overview of the neural network

Figure 2 shows an overview of the proposed method, which mainly consists of three phases: (1) network training, (2) detecting activity differences, and (3) identifying corresponding activities. Initially, we collect acceleration data from both skilled and unskilled workers' wrists for training the network. After training, we calculate and keep attention values of each hidden layer as well as the latent representations of the network for the next step. In the *detecting activity differences* phase, we select two representative inputs for both classes, respectively. For each representative, we extract segments with high attention values as candidate segments of activity differences. The *identifying corresponding activities* phase first symbolizes input series according to the clustering result, then, for each candidate of segment in phase (2), finds the corresponding segment on the other worker class's representative input. Finally, the system will output a pair of segments for each candidate of segment. The pair of segments is composed of a candidate of segment and the corresponding segment on the other worker class's representative input. With the help of the paired segments, the industrial engineer can easily compare the activity differences between skilled and unskilled workers and identify useful skill knowledge from them.

3.3 Network Training

The neural network shown in Figure 3 is composed of two parts: (1) an autoencoder that extracts latent representations of input series and (2) a skilled/unskilled worker classifier to fine-tune the latent representations to detect activity differences between the two classes.

We assume that X denotes an input sequence of the neural network, which corresponds to a period of data for a worker. The class label $y \in [0, 1]$ is associated with every X, where 0 and 1 correspond to the skilled and unskilled workers, respectively.

3.3.1 Network Architecture

In order to detect candidate segments of activity differences, we employ an attention-based neural network to classify time-series from the skilled and unskilled workers, as shown in Figure 3. The autoencoder architecture consists of three encoding blocks as well as three decoding blocks. A single encoder block consists of a 1D-CNN layer plus a BatchNormalization layer and a MaxPooling layer. A decoder block consists of a 1D-CNN layer plus a BatchNormalization layer and a UpSampling1D layer. The 1D-CNN layer captures a salient waveform from the input, and the MaxPooling layer compresses the size of the 1D-CNN outputs by choosing the maximum value within a scale. By combining with the two layers, we can get the latent representation f, where the feature vector at each timestep corresponds to a short-term data within input series. Besides, in the worker classifier architecture, four stacks of LSTM layers are connected to the encoder's output for extracting long-term dependencies in the data used for classifying skilled and unskilled workers. Blocks labeled "LSTM" include LSTM and BatchNormalization layers. Blocks named "Atten" are calculated from the output of every "LSTM" using Eq. 1, which calculates the attention weight of the "LSTM" layer output. For every time-series inputs, a corresponding attention series is computed in each "Atten" layer, with the input series with a higher attention weight being more important over the whole input series for the skilled/unskilled worker classification. Blocks labeled "Multi" multiply the attention and the outputs of the "LSTM" to emphasize important timings for classification. Blocks "Concatenate" and "Softmax" refer to the concatenate and softmax layers, respectively.

We assume that T denotes the length of the latent representations f, z_t is a D dimensional vector representing f at time t ($t \in \{1, ..., T\}$), h_t is a real-valued hidden-state vector at time t output by an "LSTM" layer. The equation of calculating attention at time t is denoted as follows:

$$\alpha_t = \exp(z_t) / \sum_{s=1}^{T} \exp(z_t) \quad (1)$$

$$z_t = \tanh(W h_t + b) \quad (2)$$

where W and b are the weight matrix and bias, respectively.

Then, we multiply attention and the output of "LSTM" in "Multi" as follows:

$$H = \sum_{t=1}^{T} \alpha_t h_t \quad (3)$$

The output of the classifier is calculated by the "Softmax" block, whose target is the class label y for classifying the skilled and unskilled workers.

3.3.2 Network Training

The loss of the network is composed of two components: reconstruction loss L_a and binary cross-entropy loss L_c. The reconstruction loss aims to learn latent representations in an unsupervised manner while preserving intrinsic local structure in data. The binary cross-entropy loss is responsible for learning features that differentiate skilled and unskilled workers. The overall loss function of the network is defined as follows:

$$L = L_a + \lambda L_c \quad (4)$$

where the parameter λ controls the trade-off between L_a and L_c.

3.4 Detecting Activity Differences

In this phase, we select a representative input for each worker and then extract candidate segments with significant activity differences according to the attention values for each representative. We argue that the skill of workers can be discovered from their repeated work periods, and their performance of the same task will be consistently similar [5].

To detect the consistent skill of each worker, we first look for a representative input (period) for each worker that is most similar to all the remaining input instances of the worker, in other words, the centroid of all instances. We use DTW algorithm to compare the similarity between two time-series data. For an input series X_i in the skilled or unskilled worker class, we calculate the DTW distance between X_i and each of the remaining inputs within the same class. The best X_i is supposed to have the minimal overall DTW distances in the class, selected as the centroid instance of the class.

After getting the centroid instances for both skilled/unskilled classes, the corresponding attention values of the centroid instances can be applied to extract candidate segments of activity differences. According to the structure of the worker classifier showing in Figure 3, there are four layer-wise attention time-series generated from each LSTM layer of the classifier, which focus on different time scales. Since a value of attention reveals the weight of the corresponding temporal position of input series in terms of the contributions to the final classification result, segments with high attention values will be more dissimilar in different worker classes. Therefore, for each attention layer, we extract candidates of input segments with the highest attention values in the top-k%.

In the next section, we will introduce how to detect the corresponding segment of each candidate on the other centroid instance.

3.5 Identifying Corresponding Activities

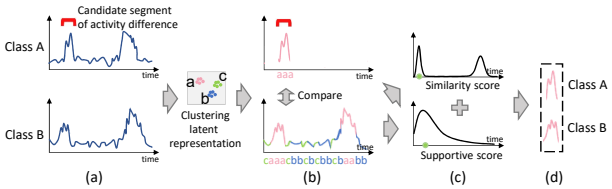

Fig. 4: Procedure of identifying corresponding segments

To extract skill knowledge from candidates of segments, it is important to find the corresponding segment on the other worker class to investigate the actual activity differences between the workers. In this section, we propose a clustering approach to find the corresponding segment for each candidate with latent representations of the neural network. Because activities performed by different workers result in sensor data differences, it is impractical to find the corresponding segments by simply comparing sensor data similarity. Instead, we look for similar latent representations of the neural network.

Figure 4 introduces the main idea of detecting the corresponding segment for a candidate segment in one worker class. Since the latent representations of similar activities are supposed to be similar, we leverage the clustering method to eliminate the differences between workers so that similar activities in different workers can be clustered into the same cluster. Four steps are implemented to identify the corresponding activities. Firstly, we employ a clustering approach to cluster all data points in every timestep of the latent representations f. Then, for every centroid instance, we symbolize the raw sensor data according to the clustering result. Next, for each candidate, which is detected in Section 3.3, we identify the corresponding segment on the other centroid instance by using a sliding window across the whole data. Finally, each candidate and its corresponding segment will be offered to the industrial engineer for skill assessment.

3.5.1 Clustering Latent Representations

As Figure 1 shows, the raw sensor data between workers are different due to the different movements in workers. However, compared with different activities (e.g., the first scan activity in the skilled worker comparing with the

attaching label activity in the unskilled worker), data belonging to similar activities (e.g., the attaching label activity for both skilled and unskilled worker) has more similar latent representations. Therefore, we leverage a clustering algorithm to roughly cluster the latent representations to eliminate the difference between individuals, with similar activities being clustered to the same clusters. We assume that $f^D_{X_i,t}$ indicates the feature representation of input series X_i at time t with D channels. We leverage a k-Means algorithm to cluster every data points in $k(\leq X)$ clusters $S = \{S_1, S_2, ..., S_k\}$. This process is formulated to find S that yields the minimum overall inner-cluster distance as follows:

$$\underset{S}{\operatorname{argmin}} \sum_{m=1}^{k} \sum_{f^D_{X_i,t} \in S_m} ||f^D_{X_i,t} - \mu_m||^2 \tag{5}$$

where μ_m is the center of S_m. Since the CNN layers of the autoencoder will not change the temporal relationship of the data points in the input series, the temporal relationship of data points in the latent space will still correspond to the points in the input series. Therefore, we can roughly label data points of every input series based on the clustering result of the latent representations at the corresponding timesteps.

3.5.2 Symbolizing Series Data

As shown in Figure 4, we symbolize the centroid instances of workers to characters based on the clustering results, where different characters represent different clusters (e.g., in the figure, as the latent representations are clustered to three clusters, which are labeled as a, b and c, the corresponding data points of the centroid instances are therefore colored to "red", "blue" and "green", respectively). We symbolize the centroid instances according to the clustering results since the clustering algorithm eliminates the small differences between workers, indicating that similar activities have the same characters.

Next, we will introduce the idea of finding the corresponding segment by using the symbols.

3.5.3 Finding Corresponding Segment

Now that we use the symbolized centroid instances to detect their corresponding segments. In Figure 4, we find a candidate segment of activity difference in class A. Our aim is to find the corresponding segment for class B. After symbolizing the centroid instances of class A and B, we compare the symbolized candidate segment for class A with the other worker class B by sliding along the complete sequence for class B. We calculate two scores: (1) similarity score and (2) supportive score, to find the corresponding segments for class B.

The similarity score represents the cluster similarity between two segments. Since latent representations of similar activities will be clustered into the same clusters, two segments with similar symbols are supposed to show similar activities. For instance, when the symbolized candidate segment for class A is "aaa", the most similar segment in class B "caaacbbcbcbbcbaabb" will be "aaa" with the similarity score of 3. We calculate similarity score $Score_{simi}$ by counting the number of paired points with the same clusters based on the following formula:

$$C_t = \begin{cases} 1, & if S(i_t) = S(j_t) \\ 0, & otherwise \end{cases} \tag{6}$$

where $S(i_t)$ and $S(j_t)$ represent the $t-th$ character of symbolized segment i and j for class A and B, respectively.

$$Score_{simi} = \sum_{t=j_1}^{j_{|j|}} C_t \tag{7}$$

where $j_{|j|}$ represents the last data point in segment j.

In addition, we calculate the supportive score. Since the order of the operations done by workers is predefined by industrial engineers, the occurrence timing of the corresponding activities (in class B) should be similar to that of the candidate segment (in class A). Therefore, we employ the supportive score $Score_{supp}$ to find segments in similar timings, which are calculated as follows:

$$Score_{supp} = \frac{1}{||Seg_i - Seg_j||^2} \tag{8}$$

where Seg_i and Seg_j are the elapsed times of the starting points of the segments from the beginning in series i and j, respectively.

We sum up the above two scores as the overall score. The segment with the highest overall score on the other series (i.e., class B) is supposed to be the most similar segment corresponding to the candidate segment (i.e., a candidate in class A).

Finally, the pair of a segment for each candidate and its corresponding segment will be provided to the industrial engineer for skill assessment.

Table 1: Overview of recorded datasets

Worker	Number of instances	skill level	Dataset
1	38	Skilled	Screwing
2	41	Unskilled	
3	42	Skilled	Final check
4	44	Unskilled	

4. Evaluation

4.1 Dataset

We evaluated the proposed method using two datasets from four individuals working in a real factory. Every worker wears two smartwatches (Sony SmartWatch3 SWR50) on each wrist, collecting acceleration data with an approximate sampling rate of 60Hz. Table 1 shows an overview of the recorded datasets. In the dataset of "Screwing", workers were employed to install screws on circuit boards, consisting of many predefined operations, such as setting, screwing, recording data, etc. The workers in the dataset of "Final check" are required to check final products and record

checking results, consisting of attaching labels, scanning labels, measuring box, etc. The skill level of each worker is labeled by an industrial engineer based on the workers' job performances in real manufacturing.

4.2 Evaluation Methodology

In order to measure if the proposed method can find the corresponding data segment on the other worker class, we calculate Mean Squared Error (MSE) of the starting time of the segment on the other worker to the groundtruth. We provide the results for the proposed method against other comparative methods, leaving one score out to evaluate the effectiveness of our scoring metrics. The methods to be tested are listing as follows:

- **Proposed**: This is the proposed method.
- **W/o Simi**: The proposed method without using the similarity score when comparing the similarity between a candidate segment and segments on the other worker class.
- **W/o Supp**: The proposed method without using the supportive score when comparing the similarity between a candidate segment and segments on the other worker class.

In addition, to quantitatively analyze the attention mechanism for detecting activity differences between skilled and unskilled workers, we discussed with industrial engineers whether the detected activity differences include skill knowledge or not. Detailed discussion is showing in Section 4.5.

4.3 Results of Identifying Corresponding Segments

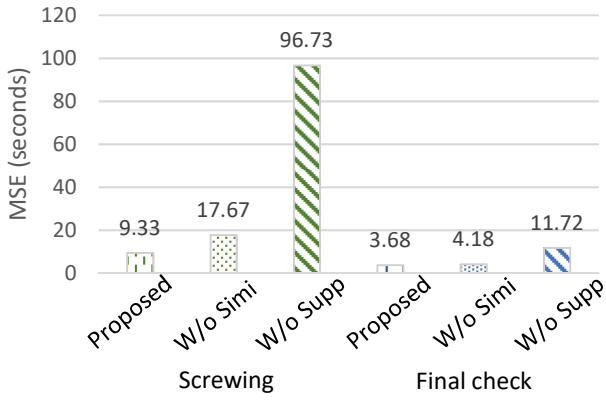

Fig. 5: MSE of the three methods on two datasets

As shown in Figure 5, we calculate the sum of MSE for all candidates of segments in each dataset and compare the performance of three methods on both datasets. The proposed method achieved the lowest MSE in both datasets, which is the most robust method among the three methods. We observe that W/o Supp in the "Screwing" dataset has the highest MSE, which is because that the candidate segment corresponding to a screwing action occurs several times in each instance, many segments corresponding to screwing actions on the other class will have a high similarity score to the candidate segment, but they do not belong to the same operation. W/o Simi also shows a high MSE on both datasets since the starting times of an activity performed in different periods are different. The similarity score detects the most similarity data segment on the other class to find the corresponding activity.

4.4 Performance of Network Structure

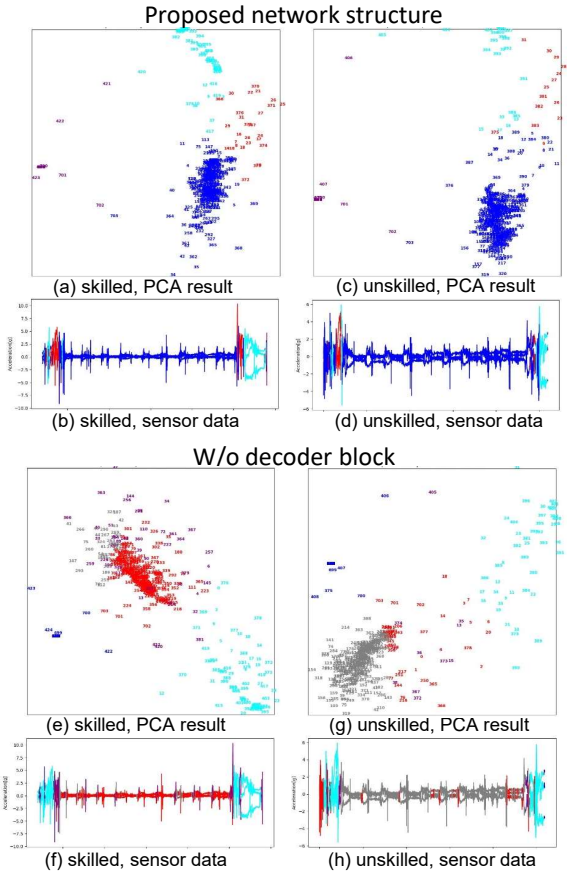

Fig. 6: Clustering results of latent representations for "Screwing" dataset, different color shows different clusters. (a), (c), (e), and (g) show latent representations visualized by PCA. (b), (d), (f), and (h) show clustering results visualized on raw acceleration data of right wrists

In this section, we discuss if the combination of the autoencoder and worker classifier can improve the performance of our method. As shown in Figure 6, we compare the Proposed network structure with the W/o decoder block, in order to evaluate if the autoencoder can protect the structure of latent representations similar to the input series. As can be seen, the Proposed network structure has similar distributions of latent representations between the skilled and unskilled workers, with similar activities clustered into the same cluster (e.g., the sensor data segments with blue color are corresponding to screwing activities). Whereas the W/o decoder block shows different distributions of latent representations between the skilled and unskilled workers, resulting in similar activities in different workers clustered into different clusters.

4.5 Discussion

Since we do not have groundtruth labels to evaluate the activity differences, we asked industrial engineers to assess if skill knowledge exists in the candidates of segments detected by the attention network, as our ultimate goal is to improve the efficiency to help the engineers to find skill knowledge. Tables 2 and 3 present the skill information of each activity candidate detected in the "Screwing" and "Final check" datasets, respectively. The "No." column represents the indices of candidate segments detected in each dataset. The "Activity description" column briefly describes the main activities performed in the corresponding candidates. The skill knowledge types are shown in the "Skill knowledge" column, which was decided by the industrial engineers. In the "Other" column, we offer some extra information related to the activities for assessment, such as the average duration of the activities for the skilled/unskilled workers.

When determining if an activity candidate has skill knowledge, the engineer follows the "Principles of motion economy" strategy proposed in [7] to identify the skill type of the candidates. The types of skill knowledge in our study are listed as follows:

- Time conservation (1): Even temporary delay of work by a man or machine should not be encouraged.
- Time conservation (2): Two or more jobs should be performed upon at the same time, or two or more operations should be carried out on a job simultaneously if possible.
- Arrangement of the work place (1): Arrange the height of the workplace and chair for comfortable sitting and standing.
- Arrangement of the work place (2): Tools, materials, and controls should be located close to and in front of the operator.
- Arrangement of the work place (3): There should be a definite and fixed place for all tools and materials.

In the dataset of "Screwing", the No.2 candidate shows that both the skilled and unskilled workers have a high overall waiting time, in which the waiting time for the skilled worker is two times longer than the unskilled. The information indicates that workers upstream of the production line of the current workers have a lower work efficiency, which will deteriorate the productivity of the whole production line. According to the strategy of "Time conservation (1)", it is advised to shorten the delay by upstream workers by assigning more workers upstream the production line. No.3 and No.4 correspond to the *6-th* and the last screwing actions, respectively. The average duration of the screwing action for the skilled worker is slightly shorter than the unskilled worker since the skilled worker can use both hands for different activities simultaneously. For example, while controlling the screwing tool with the right hand, the skilled worker uses his left hand to set a new screw into the next hole. However, the unskilled worker cannot perform these activities at the same time. Based on the idea of "Time conservation (2)", the unskilled worker should learn how to cooperate with both hands to promote work efficiency. As for the skill information in the No.1 candidate, the major difference between the workers is that they use a different hand to set circuit boards, which is not recognized as a skill knowledge from the engineer's point of view.

In the dataset of "Final check", the skill knowledge of the No.1 candidate also belongs to "Time conservation (2)", where the skilled worker places his left hand on the box to control the height of the label and decides the timing to stick the label by the right hand. In contrast, the unskilled worker holds the sides of the label with both hands, spending a more extended time locating the appropriate place to stick the label. The No.2 candidate shows a different skill knowledge, in which the height of the workplace for the unskilled work is not suitable for him, so that he usually bends over to attach the label. As a result, the unskilled worker suffers a heavier body burden and feels tired soon. The "Arrangement of the work place (1)" strategy suggests that the engineer coordinates a suitable workplace for workers to reduce body burden. In the No.3 candidate, workers need to remove a label from a box, then attaching the label back to the box. After removing the label, the skilled worker places the label in front of the box, while the unskilled worker sticks the label far from the box, which takes more time to finish the activity. The corresponding strategy advises the label be placed close to the box. The main difference between workers in the No.6 candidate is that the skilled worker sometimes scans the box before he moves the box to the work table, while the industrial engineer requires workers to process the box just over the work table to ensure every operation can be performed appropriately. Therefore, the skill knowledge we observed from the No.6 candidate was not recommended by the engineer. Unfortunately, it is challenging to evaluate skill knowledge for the No.5 and No.7 candidates, as these activity candidates happen at the end of the work data without available video information.

However, there are some limitations to the attention-based network. Our method cannot detect small activity differences, which is also likely to contain useful skill information. For example, the way of holding the measuring tool is an important skill knowledge to get accurate measuring results, but the activity difference between the skilled and unskilled workers mainly different in the direction of the wrist, which does not have high attention values.

As a result, we extracted 11.7% and 19.3% candidate segments over the whole centroid inputs in "Screwing" and "Final check" datasets, respectively, in which 7 out of 11 segments include skill information identified by the industrial engineer. The attention mechanism is a practical supporting technique to be adopted to detect skill knowledge within significant activity differences.

5. Conclusion

This paper proposes an attention-based neural network to identify significant activity differences between workers, which is applied to support industrial engineers to find skill

Table 2: Skill information for the dataset of Screwing

No.	Activity description	Skill knowledge	Other
1	Set box and push button	-	Avg. duration of skilled/unskilled (6.7s/7.2s)
2	Wait for the next box	Time conservation (1)	Total duration of skilled/unskilled (296.6s/102.4s)
3	Screwing	Time conservation (2)	Avg. duration of skilled/unskilled (4.4s/4.6s)
4	Screwing	Time conservation (2)	Avg. duration of skilled/unskilled (4.6s/5.8s)

Table 3: Skill information for the dataset of Final check

No.	Activity description	Skill knowledge	Other
1	Attach a small label on the box	Time conservation (2)	Avg. duration of skilled/unskilled (6.6s/8.9s)
2	Attach a large label on the box	Arrangement of the work place (1)	Avg. duration of skilled/unskilled (8.9s/10.8s)
3	Stick the large label on table	Arrangement of the work place (2)	-
4	Rotate the box to check information	-	Avg. duration of skilled/unskilled (1.7s/2.3s)
5	Bring the box to other place	-	No camera information
6	Set the box on the table	Arrangement of the work place (3)	Change activity order, not recommended
7	Bring the box to other place	-	No camera information

knowledge from workers. We employ the attention mechanism to emphasize and visualize important input segments and design a clustering method to eliminate individual differences. According to the candidates of activity differences detected by attentions, the industrial engineer can find useful skill knowledge from workers. In the future, we desire to increase the neural network's ability to identify small activity differences with skill knowledge.

6. Acknowledgments

This work is partially supported by JST CREST JPMJCR15E2, JSPS KAKENHI Grant Number JP16H06539 and JP17H04679.

References

[1] John P Campbell, Jeffrey J McHenry, and Lauress L Wise. Modeling job performance in a population of jobs. *Personnel Psychology*, 43(2):313–575, 1990.

[2] Vedant Das Swain, Koustuv Saha, Hemang Rajvanshy, Anusha Sirigiri, Julie M Gregg, Suwen Lin, Gonzalo J Martinez, Stephen M Mattingly, Shayan Mirjafari, Raghu Mulukutla, et al. A multisensor person-centered approach to understand the role of daily activities in job performance with organizational personas. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(4):1–27, 2019.

[3] Xifeng Guo, Long Gao, Xinwang Liu, and Jianping Yin. Improved deep embedded clustering with local structure preservation. In *IJCAI*, pages 1753–1759, 2017.

[4] Yohei Kawase and Manabu Hashimoto. Analysis of skill improvement process based on movement of gaze and hand in assembly task. In *International Conference on Computer Analysis of Images and Patterns*, pages 15–26. Springer, 2019.

[5] Aftab Khan, Sebastian Mellor, Rachel King, Balazs Janko, William Harwin, R Simon Sherratt, Ian Craddock, and Thomas Plötz. Generalized and efficient skill assessment from imu data with applications in gymnastics and medical training. *ACM Transactions on Computing for Healthcare*, 2(1):1–21, 2020.

[6] Takuya Maekawa, Kazuya Ohara, Yizhe Zhang, Matasaburo Fukutomi, Sakiko Matsumoto, Kentarou Matsumura, Hisashi Shidara, Shuhei J Yamazaki, Ryusuke Fujisawa, Kaoru Ide, et al. Deep learning-assisted comparative analysis of animal trajectories with deephl. *Nature Communications*, 11(1):1–15, 2020.

[7] Fred E Meyers and James Robert Stewart. *Motion and time study for lean manufacturing*. Pearson College Division, 2002.

[8] Shayan Mirjafari, Kizito Masaba, Ted Grover, Weichen Wang, Pino Audia, Andrew T Campbell, Nitesh V Chawla, Vedant Das Swain, Munmun De Choudhury, Anind K Dey, et al. Differentiating higher and lower job performers in the workplace using mobile sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(2):1–24, 2019.

[9] Vishvak S Murahari and Thomas Plötz. On attention models for human activity recognition. In *The 2018 ACM International Symposium on Wearable Computers*, pages 100–103, 2018.

[10] Andrew Ng et al. Sparse autoencoder. *CS294A Lecture notes*, 72(2011):1–19, 2011.

[11] Nancy L Stokey. Human capital, product quality, and growth. *The Quarterly Journal of Economics*, 106(2):587–616, 1991.

[12] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International Conference on Machine Learning*, pages 478–487. PMLR, 2016.

[13] Ming Zeng, Haoxiang Gao, Tong Yu, Ole J Mengshoel, Helge Langseth, Ian Lane, and Xiaobing Liu. Understanding and improving recurrent networks for human activity recognition by continuous attention. In *Proceedings of the 2018 ACM International Symposium on Wearable Computers*, pages 56–63, 2018.