Comparative Analysis of High- and Low-Performing Factory Workers with Attention-Based Neural Networks

Qingxin Xia[1], Atsushi Wada[2], Takanori Yoshii[2], Yasuo Namioka[2], and Takuya Maekawa[1(✉)]

[1] Graduate School of Information Science and Technology, Osaka University, Osaka 5650871, Japan
{xia.qingxin,maekawa}@ist.osaka-u.ac.jp
[2] Corporate Manufacturing Engineering Center, Toshiba Corporation, Yokohama, Kanagawa 2350017, Japan
{atsushi3.wada,takanori.yoshii,yasuo.namioka}@toshiba.co.jp

Abstract. This study presents a new method that supports the comparative analysis of works performed by high- and low-performing factory workers. Our method, based on explainable deep learning, automatically detects a sensor data segment that potentially contains knowledge about the skill of works by analyzing acceleration sensor data from high- and low-performing workers. Our evaluation with industrial engineers using sensor data from actual factory workers revealed that 78% of sensor data segments detected by our method included knowledge about skill.

Keywords: Attention networks · Work performance · Wearable sensor · Factory work

1 Introduction

1.1 Background

The skill level of a factory worker of assembly work significantly influences the productivity of a production system in which the worker is involved. Assembly work is a common part of line production systems and typically involves factory workers performing a repetitive work process consisting of a sequence of operations, such as setting a board on a workbench and screwing parts onto the board. Mistakes and delays in each work period, that is, one iteration of the entire sequence of operations, are accumulated, significantly deteriorating the overall performance of the production system. Therefore, developing skills for low-performing workers is crucial to improve the efficiency of assembly work. For this purpose, industrial engineers have manually compared video recordings of works by a high-performing worker with those of a low-performing worker

T. Hara and H. Yamaguchi (Eds.): MobiQuitous 2021, LNICST 419, pp. 469–480, 2022.
https://doi.org/10.1007/978-3-030-94822-1_26

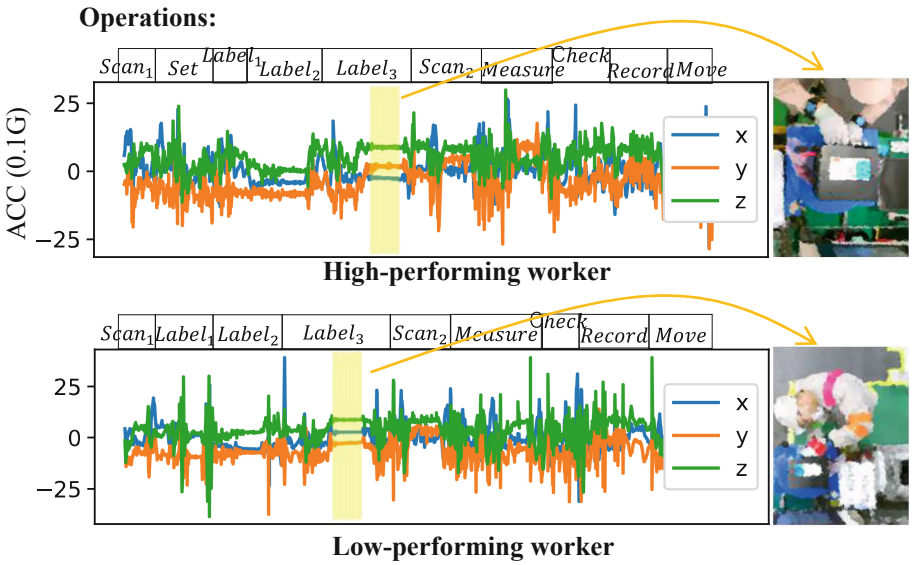

Fig. 1. Example time-series of acceleration data from two workers' right wrists. The yellow rectangles indicate segments that potentially contain knowledge about skill extracted by our method.

to extract hidden knowledge about skills [6], which is useful for training low-performing workers. However, because there are many workers in the factory, it incurs huge costs for industrial engineers to analyze the work manually.

1.2 Research Goal

Owing to the recent progress in sensing technologies, studies on work management and analysis using wearable sensors have been actively carried out [3,8,15,16]. Wearable sensors are a promising technology for achieving smart manufacturing because they enable the capture of fine-grained activities by workers that are difficult to capture by cameras in a factory where many obstacles exist. This study focuses on a line production system and attempts to help extract knowledge about skill from sensor data collected from high- and low-performing workers performing the same work process in order to support manual comparative analysis by industrial engineers. Specifically, we attempt to automatically detect a sensor data segment from a high-performing worker that potentially contains knowledge about skill, and then provide it to the industrial engineers with a video recording capturing that moment. Note that, because it is difficult for industrial engineers to understand the meaning of skill hidden in the data by watching only the data (or video), we also find a corresponding sensor data segment from the low-performing worker and present it to the industrial engineers with the segment from the high-performing worker. For example, when a segment corresponding to a screwing operation from a high-performing

worker's data is detected, we also find a segment corresponding to the screwing operation from a low-performing worker's data. Comparing these segments (video recordings) permits the industrial engineer to understand the meaning of the skills hidden in the data.

Figure 1 shows example segments extracted by our proposed method (highlighted in yellow) described below. The segments (and video from a top-down view) reveal different postures between the high- and low-performing workers when attaching a label, where the high-performing worker kneels down to the level of the workbench by bending the knees to perform the work. In contrast, the low-performing worker largely bends the spine to perform the operation, and is more likely to suffer back pain. The industrial engineer can guide the low-performing worker based on the information.

1.3 Challenges and Approaches

This study has two technical challenges. The first challenge is to extract a candidate sensor data segment containing meaningful knowledge about skill by analyzing only the sensor data. In this study, we hypothesize that it is possible to extract meaningful knowledge about skills from candidate segments with the following two characteristics. (i) Sensor data from different periods by a worker are somewhat different. Segments containing a sensor data pattern (e.g., characteristic hand movement) that are found in all these periods are expected to be important and essential in the work process of interest [5]. (ii) When a sensor data pattern with the above characteristics is only available in data from the high-performing worker, the probability of the sensor data pattern relating to skill is high.

To find candidate segments with the above characteristics, we leverage explainable deep learning. Because deep learning does not require manual feature design, which is difficult for an industrial engineer for each work process, the engineer can obtain desired segments by simply feeding raw sensor data into a deep model. This is the advantage of the deep learning approach, and our experiment revealed that a simple signal matching-based approach did not work in our task. In this study, we first build a recurrent neural network that classifies time-series sensor data corresponding to a period into a high- or low-performing worker class. Because the network is trained to discriminate the sensor data of a high-performing worker from those of a low-performing worker with high accuracy, the network is expected to identify sensor data patterns containing the above two characteristics. This is because a sensor data pattern that is found in all the data from the high-performing worker but not in the low-performing worker data (and vice versa) is an important clue in the classification task. Therefore, we can leverage the trained network to automatically identify segments that potentially contain knowledge of skills. Our idea of revealing important sensor data patterns identified by the neural network, which is regarded as a black box, leverages attention mechanisms [14]. The attention mechanism provides information about importance (attention) for each data point in time series, enabling

us to find a candidate segment with high attention that potentially contains knowledge about skill.

The above procedure extracts a candidate segment from the high-performing worker. We then extract a segment from the low-performing worker corresponding to the segment from the high-performing worker, which is the second challenge. Assume that a candidate segment corresponding to the screwing is detected from the high-performing worker data. This means that the screwing operation is different between the high- and low-performing workers; it is difficult to find a segment of screwing from the low-performing worker by using sensor data similarity. Our idea to address this issue is to introduce an autoencoder [10] into an attention-based network. The autoencoder extracts latent representations (compressed representations) of input data points while preserving the main components of the original data points, making it easier to find segments that are semantically similar to each other, that is, sensor data segments corresponding to the same operation.

1.4 Contributions

- This is the first study to analyze factory work using attention-based explainable deep learning.
- We extract a segment that potentially contains knowledge about skill using our network composed of an attention mechanism and an autoencoder.
- We performed qualitative and quantitative evaluations of our method using sensor data from actual factories with industrial engineers.

2 Related Work

Early studies on knowledge extraction related to factory work have relied on self-reporting [1,2], making it difficult to extract implicit knowledge about skills that can be included in an operation performed subconsciously. Recent studies have introduced activity data collected using electronic devices [4]. For example, Mirjafarl et al. [12] used mobile phones, wearables, and beacons to study differences in daily behavioral patterns (e.g., sleep) between higher and lower work performers in companies. Das Swain et al. [3] also leveraged sensors in commodity devices to investigate the relationship between work performance and the daily activities of workers. Many prior studies on analyzing work performance using sensor data collected during factory work build machine learning models that predict workers' work scores [7,13]. In contrast, our method tries to detect a data segment that potentially contains knowledge about skills using explainable deep learning.

Few recent studies leverage attention mechanisms to analyze time-series behavioral data. Zeng et al. [17] developed two attention models: temporal attention and sensor attention for detecting important sensor data segments and sensor modalities, respectively, which can be applied to identify the most important sensor data patterns and sensor modalities for detecting Parkinson's disease.

Maekawa et al. [9] also applied an attention-based neural network to animals' trajectories in order to detect segments in trajectories that are characteristic of one group, enabling biologists to focus on these specific segments and formulating new hypotheses. In contrast, this study proposes a network composed of an attention mechanism and autoencoder to enable a comparative analysis of factory works by industrial engineers.

3 Factory Work Analysis with Attention-Based Network

3.1 Preliminaries and Overview

In this study, we assume that high- and low-performing workers perform the same work process, with smartwatches worn on each worker's wrists recording three-axis accelerometer data. Multiple time-series data, with each time series corresponding to a period, from each worker are given.

Our method is composed of two steps. First, we train an attention-based neural network to automatically identify candidate segments that potentially contain knowledge about skills. Then, for each candidate, we detected the corresponding segments of the other worker.

3.2 Network for Comparative Analysis of Factory Work

Network Architecture. We designed an attention-based neural network to classify time series from high- and low-performing workers, as shown in Fig. 2. The autoencoder architecture, which is responsible for extracting feature representation f that preserves main components of an input, consists of three encoding blocks, including an 1-D convolutional layer (1D CNN), a batch-normalization layer (BatchNormalization) and a maxpooling layer (Maxpooling), and three decoding blocks, including an "1D CNN," a "BatchNormalization," and an upsampling layer (Upsampling). In addition, in the attention-based worker classifier architecture, four stacks of long short-term memory (LSTM) layers are connected to the encoder's output to extract long-term dependencies in the data used for high- and low-performer classification, enabling the extraction of candidate segments at various time scales. A block labeled "LSTM" includes LSTM and BatchNormalization layers. A block named "Atten" processes the output of "LSTM" using Eq. 1, which calculates the attention weight of the "LSTM" layer output. For each time-series input, a corresponding attention series is computed in each "Atten" layer, with the input series with a higher attention weight being more important over the entire input series for classification. A Block labeled "Mul" multiplies the attention and the outputs of "LSTM" to emphasize important timings for classification. Blocks "Concatenate" and "Softmax" refer to the concatenate and softmax layers, respectively. The equation of calculating attention at time t is denoted as follows:

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

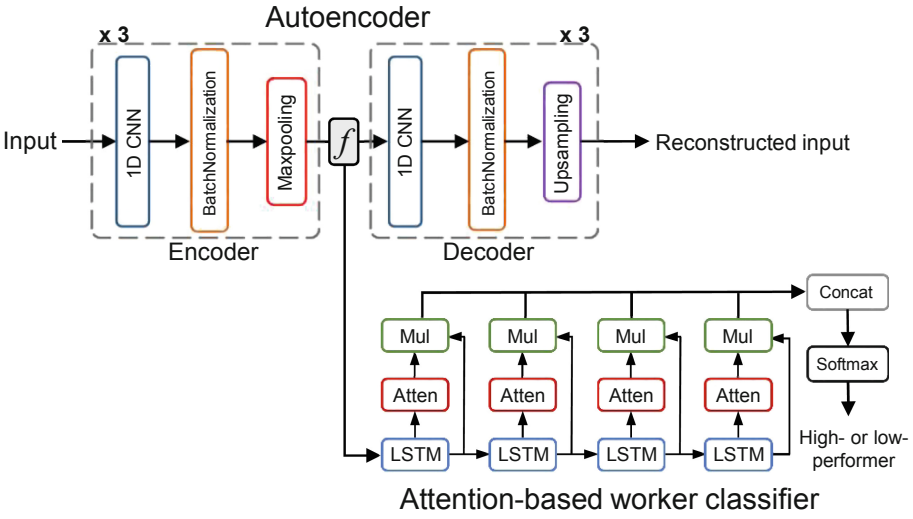

Fig. 2. Overview of proposed network.

$$z_t = \tanh(Wh_t + b) \tag{2}$$

where T is the length of the latent representation f, z_t is a D-dimensional vector calculated by "Atten" at time t ($t \in \{1, ..., T\}$), and h_t is a hidden-state vector at time t output by "LSTM." W and b are the weight matrix and bias in "Atten," respectively.

Network Training. The loss of the network is composed of two components: reconstruction loss L_a (mean squared error between input and reconstructed input) and binary cross-entropy loss L_c. The reconstruction loss aims to learn latent representations using the autoencoder in an unsupervised manner while preserving the main components in the data. The binary cross-entropy loss is responsible for the binary classification (high- vs. low-performers). The overall loss function of the network is defined as $L = L_a + \lambda L_c$, where λ controls the trade-off between L_a and L_c.

Detecting Candidate Segments with Attention. Because the time-series data of different periods performed by a worker are different, we first find a representative input (period) for each worker that is most similar to all the remaining periods of the worker (i.e., the centroid of all instances). We use the dynamic time warping (DTW) algorithm to calculate the distance between each pair of time-series data, with the centroid instance giving the smallest overall distances.

After obtaining a centroid period for each worker, the corresponding attention values of the centroid from the trained network are used to extract the candidate segments. Because the attention values reveal the importance of each data point

in the time series for predicting the skill level of performers, we extract segments with the top-k attention values as candidates for each layer (except the 1st layer), and then merge overlapping candidate segments.

3.3 Detecting Corresponding Segments

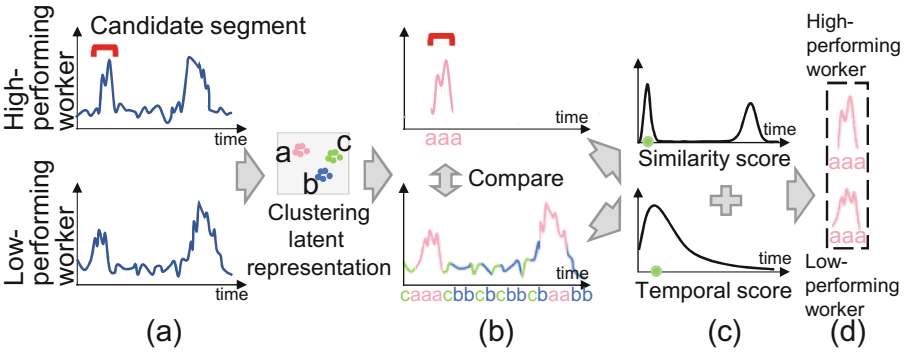

Fig. 3. Overview of the procedures to detect corresponding segments. (a) We start by clustering all data points using f to cluster similar activities. (b) We then symbolize the data points of high- and low-performers. (c) For each candidate of the high performer, we detect a corresponding segment of the low performer using a combination of similarity and temporal scores. (d) We present a pair of the candidate and corresponding segment to industrial engineers.

To support comparative analysis by industrial engineers, we then find the corresponding sensor data segment of each candidate segment. Figure 3 introduces the main idea of detecting a corresponding segment for a candidate, which is composed of four steps. First, we employed the k-means algorithm to cluster all data points using their latent representations f to group similar activities of the high- and low-performers into the same cluster, as latent representations of similar activities are supposed to be similar. Then, we symbolized each data point according to the clustering result. For each candidate, which was detected in the previous procedure, we identified a corresponding segment of a centroid period of the other worker by using a sliding time window across the entire data of the centroid period. For each time window, we calculated a combination of similarity and temporal scores, where the similarity score calculated the similarity between the symbolized candidate segment and symbol segments in the time window (inverse of Hamming distance), and the temporal score evaluates the temporal distance between the occurrence timings of the two segments in the periods because the occurrence times of the same operation in a period are expected to be similar between the high- and low-performing workers. Finally, we selected the segment with the highest score as the corresponding segment of the candidate segment.

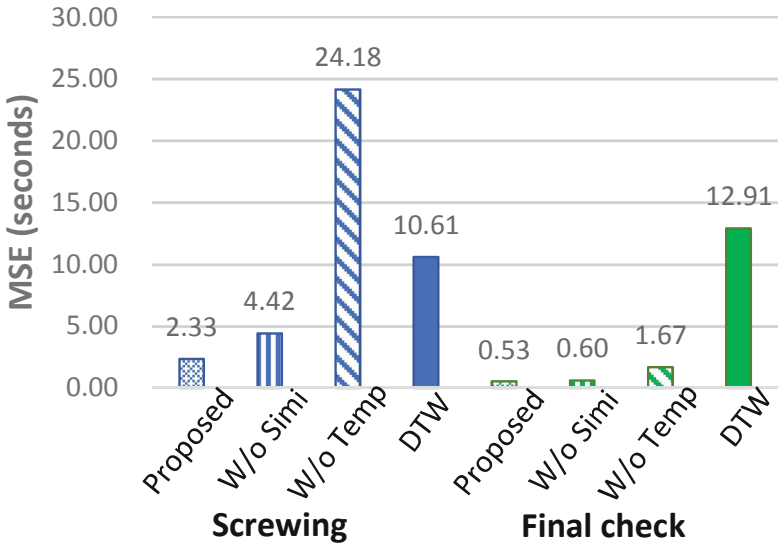

Fig. 4. Errors of the methods for detecting corresponding segments in two data sets.

4 Evaluation

4.1 Data Set

We used two acceleration data sets from four workers in a real factory using Sony SmartWatch3 SWR50 with a sampling rate 60 Hz. In the data set named "Screwing," workers were employed to install screws on the circuit boards. The number of periods for the high- and low-performers was 38 and 41, respectively, and the total duration of the data was 2337 and 2713 (s), respectively. In the data set named "Final check," workers were required to check the final products and record results; the number of periods for the high- and low-performers was 42 and 44, respectively, and the total duration was 2380 and 2797 (s), respectively.

4.2 Evaluation Methodology

We performed quantitative analysis to evaluate the performance of detecting corresponding segments. We prepared the following methods: (i) Proposed: The proposed method. (ii) W/o Simi: The proposed method without using the similarity score, (iii) W/o Temp: The proposed method without using the temporal score, and (iv) DTW: A method using raw sensor data and the DTW algorithm to find a corresponding segment. We calculated the mean squared error (MSE) between the estimated starting time of each corresponding segment by a method and the ground truth, which was manually identified.

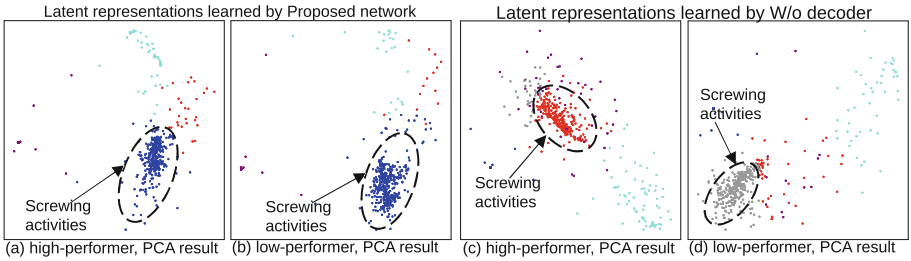

Fig. 5. Clustering result of latent representations for "Screwing" data set visualized by PCA. Different colors indicate different clusters.

We also performed a qualitative analysis to evaluate the ability of our attention-based network to detect knowledge about skills. The industrial engineers judged whether each candidate segment extracted by our method contained knowledge about the skill.

4.3 Results

Performance of Detecting Corresponding Segments. Figure 4 shows the MSE over all candidate segments in each data set. The proposed method achieved significantly small MSEs in the both data sets. In contrast, the DTW method had large MSEs on the datasets, indicating the difficulties in finding corresponding segments by only calculating raw sensor data similarity.

Figure 5 presents the clustering results of the latent representations corresponding to the two centroid instances of the "Screwing" data set visualized by principal component analysis (PCA). When we did not use the decoder block, the distributions of latent representations corresponding to the high- and low-performers are different. In contrast, our method generates similar distributions for high- and low-performers (e.g., distributions for screwing located at almost the same positions in Fig. 5). This result indicates that our idea of introducing the autoencoder enables the identification of corresponding segments (operations) of the other worker.

Analysis of Detected Segments by Attention. We qualitatively analyzed our method by asking industrial engineers to assess whether useful knowledge about skill exists in the detected segments, who followed the "principles of motion economy" strategy [11], which defines a set of rules to improve the manual work and reduce fatigue by manufacturing workers.

Table 1 shows knowledge about skill for each candidate segment detected in the two data sets, in which 7 out of 9 segments included knowledge of skill identified by the industrial engineers. The detected knowledge about skill was classified into three groups based on the principles of motion economy, which are (1) arrangement of the work place (No. 3 in "Final check"), (2) time conservation (No. 2 in "Screwing" and No. 5 in "Final check"), and (3) use of human body (No. 3, 4 in "Screwing" and No. 1, 2 in "Final check"). As mentioned above,

we demonstrated that our method detected segments with knowledge of skill with high precision. In general, it takes a much longer time than the duration of the sensor data to analyze the data by an industrial engineer. We believe that our method will significantly reduce the efforts of engineers regarding the manual screening of long-term sensor data from many factory workers. While it is difficult to evaluate the recall of our method because the ground truth is unknown, the engineers noticed that our method could not detect one minute action of grabbing a tool that contains knowledge during the experiment. Improvement of our method to detect such minute actions is our important future work.

Table 1. Skill information for screwing and final check

No.	Detected operations in screwing	Knowledge
1	Set box and push button (4.23s)	-
2	Wait for next item (4.92s)	Quickly complete previous operations
3	Screwing (3.43s)	Use both hands simultaneously
4	Screwing (3.85s)	Use both hands simultaneously

No.	Detected operation in final check	Knowledge
1	Attach small label on box (4.63s)	Use left hand to guide labeling
2	Attach large label on box (4.10s)	Bend knees to reduce load
3	Stick large label on table (2.50s)	Put labels close to worker
4	Rotate box for final-check (3.03s)	-
5	Set box on table (4.63s)	Optimize the order of operations

5 Conclusion

In this study, we proposed an attention-based explainable neural network to extract sensor data segments that potentially contain knowledge about skill, which was applied to support industrial engineers to find knowledge about skill from workers. We employed the attention mechanism to emphasize important timings as candidate segments and clustered similar activities to detect corresponding segments. The results proved that industrial engineers can efficiently find knowledge about skill from the detected segments by using our method.

As part of our future work, we plan to increase the ability of our method to detect minute actions of workers with useful knowledge about skill.

Acknowledgements. This study is partially supported by JSPS JP16H06539 and JP21K19769.

References

1. Bakker, A.B., Tims, M., Derks, D.: Proactive personality and job performance: the role of job crafting and work engagement. Hum. Relat. **65**(10), 1359–1378 (2012)
2. Campbell, J.P., Mchenry, J.J., Wise, L.L.: Modeling job performance in a population of jobs. Pers. Psychol. **43**(2), 313–575 (1990)
3. Das Swain, V., et al.: A multisensor person-centered approach to understand the role of daily activities in job performance with organizational personas. Proc. ACM Interact. Mob. Wearable Ubiquit. Technol. **3**(4) (2019). https://doi.org/10.1145/3369828.
4. Hölzemann, A., Van Laerhoven, K.: Using Wrist-Worn activity recognition for basketball game analysis. In: Proceedings of the 5th International Workshop on Sensor-Based Activity Recognition and Interaction, iWOAR 2018. Association for Computing Machinery, New York (2018). https://doi.org/10.1145/3266157.3266217
5. Khan, A., et al.: Generalized and efficient skill assessment from IMU data with applications in gymnastics and medical training. ACM Trans. Comput. Healthc. **2**(1) (2021). https://doi.org/10.1145/3422168
6. Johnson, T.L., Fletcher, S., Baker, W., Charles, R.: How and why we need to capture tacit knowledge in manufacturing: case studies of visual inspection. Appl. Ergon. **74**, 1–9 (2019). https://doi.org/10.1016/j.apergo.2018.07.016. https://www.sciencedirect.com/science/article/pii/S0003687018302278
7. Lin, S., et al.: Sensing personality to predict job performance. In: Proceedings of the Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, CHI EA 2019. Association for Computing Machinery, New York (2019)
8. Maekawa, T., Nakai, D., Ohara, K., Namioka, Y.: Toward practical factory activity recognition: unsupervised understanding of repetitive assembly work in a factory. In: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp 2016, pp. 1088–1099. Association for Computing Machinery, New York (2016). https://doi.org/10.1145/2971648.2971721
9. Maekawa, T., et al.: Deep learning-assisted comparative analysis of animal trajectories with DeepHL. Nat. Commun. **11**(1), 1–15 (2020)
10. Masci, J., Meier, U., Cireşan, D., Schmidhuber, J.: Stacked convolutional auto-encoders for hierarchical feature extraction. In: Honkela, T., Duch, W., Girolami, M., Kaski, S. (eds.) ICANN 2011. LNCS, vol. 6791, pp. 52–59. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21735-7_7
11. Meyers, F.E., Stewart, J.R.: Motion and Time Study for Lean Manufacturing. Pearson College Division (2002)
12. Mirjafari, S., et al.: Differentiating higher and lower job performers in the workplace using mobile sensing. Proc. ACM Interact. Mob. Wearable Ubiquit. Technol. **3**(2) (2019). https://doi.org/10.1145/3328908
13. Ushada, M., Okayama, T., Suyantohadi, A., Khuriyati, N., Murase, H.: Kansei engineering-based artificial neural network model to evaluate worker performance in small-medium scale food production system. Int. J. Ind. Syst. Eng. **27**(1), 28–47 (2017)
14. Vaswani, A., et al.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS 2017, pp. 6000–6010. Curran Associates Inc., Red Hook (2017)
15. Xia, Q., Korpela, J., Namioka, Y., Maekawa, T.: Robust unsupervised factory activity recognition with Body-Worn accelerometer using temporal structure of multiple sensor data motifs. Proc. ACM Interact. Mob. Wearable Ubiquit. Technol. **4**(3) (2020). https://doi.org/10.1145/3411836

16. Xia, Q., Wada, A., Korpela, J., Maekawa, T., Namioka, Y.: Unsupervised factory activity recognition with wearable sensors using process instruction information. Proc. ACM Interact. Mob. Wearable Ubiquit. Technol. **3**(2) (2019). https://doi.org/10.1145/3328931
17. Zeng, M., et al.: Understanding and improving recurrent networks for human activity recognition by continuous attention. In: Proceedings of the 2018 ACM International Symposium on Wearable Computers, ISWC 2018, pp. 56–63. Association for Computing Machinery, New York (2018). https://doi.org/10.1145/3267242.3267286