Available online at www.sciencedirect.com

ScienceDirect

Procedia CIRP 127 (2024) 135–140

www.elsevier.com/locate/procedia

10th CIRP Conference on Assembly Technology and Systems (CIRP CATS 2024)

Make some Noise: Acoustic Classification of Manual Work Steps Towards Adaptive Assistance Systems

Lorenz Fink[a,*], David Kostolani[b], Thomas Trautner[a], Sebastian Schlund[b]

[a]Institute of Production Engineering and Photonic Technologies, TU Wien, Getreidemarkt 9, 1060 Vienna, Austria
[b]Institute of Management Science, Dept. of Human-Machine Interaction, TU Wien, Theresianumgasse 27, 1040 Vienna, Austria

* Corresponding author. Tel.: +43-1-58801-31125. *E-mail address:* lorenz.fink@ift.at

Abstract

With 32 million people working in the European manufacturing sector, human work still plays a crucial role in industry. However, due to lot size one manufacturing and increased quality requirements, the complexity of products and processes is growing. Therefore, numerous approaches introduce adaptive assistance systems in assembly to support workers during complex work tasks and adapt their level of assistance to the current situation. To enable adaptivity, human action recognition must be incorporated into the consideration of context. Until now, research has focused on providing context to the machine through wearable sensors or cameras. However, wearable sensors hinder worker's movements and cameras have difficulties distinguishing between work steps of high visual similarity. To mitigate these challenges, we present a new method to classify manual work steps only by their typical acoustic characteristics. The proposed method uses log-Mel spectrograms of work sounds fed into a convolutional neural network (CNN), thus learning their characteristic structure. Moreover, we present a new public dataset for the acoustic classification of manual work steps. The dataset includes typical sources of sounds in manufacturing, such as working with a bench grinder, cordless screwdriver, filing, or grabbing screws. Before feeding the data to the CNN, we apply various pre-processing and data augmentation techniques to increase generalisation capabilities. Our method can detect work steps with reliable accuracy while requiring less parameters than other techniques, proving that detecting work context through acoustics is possible and feasible.

Keywords: Deep Learning; Log-Mel Spectrogram; Adaptive Assistance; Human Action Recognition; CNN

1. Introduction

The field of assembly is undergoing a profound transformation. Driven by lot size one manufacturing and increased quality requirements, the complexity of products and processes is constantly rising. In order for workers to master this challenge, they need to be supported accordingly. Recently, the emergence of Industry 4.0 technologies and intelligent human-centered systems has presented a potential solution to tackle the increasing complexity of work, improve productivity, and help workers meet the required quality standards [14].

Within this line of research, assistance systems aim to provide cognitive and physical support to workers. Examples of assistance systems include augmented reality, virtual assistants, or robotic assistants that provide workers with the tools and information needed to perform their tasks efficiently [3]. Recently, research has focused on enhancing these systems by the ability to detect context in the work environment [7]. In manufacturing, context-awareness refers to the recognition of the work tasks currently performed, the work progress, and possible errors. By detecting this information, feedback from the system can be adapted to users and their current needs. Ultimately, the goal of such systems, also referred to as adaptive assistance systems, is to provide assistance at the right time, and hence improve user acceptance and usability [21].

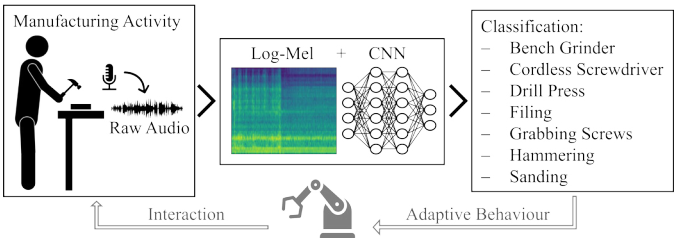

Fig. 1: High level overview of our method, towards adaptive behaviour. We created and labeled a new dataset of typical manual manufacturing sounds. This dataset was then used to train a new algorithm for the acoustic classification of assembly tasks.

2212-8271
10.1016/j.procir.2024.07.024

In previous research, various sensors for enabling adaptivity in assistance systems have been proposed. Popular approaches include wearables such as inertial measurement units (IMUs), or cameras. While IMUs are widely used, a major limitation of this method is that they affect workers' mobility and work comfort. In addition, IMUs are prone to drift due to thermo-mechanical and flicker noise, temperature effects, or calibration errors [18]. Although cameras have been proven to provide accurate detection, recognition of tasks with visual similarity remains challenging. Moreover, only a small part of video recordings contain relevant temporal information, making most of the data redundant [8].

To mitigate these issues, we propose a novel approach for classifying assembly tasks based on their acoustic characteristics. As shown in Figure 1, our method is based on raw audio recordings of the work process, which are transformed into spectrogram images, and fed to a convolutional neural network (CNN) to perform the classification. In order to train the network, a new dataset containing sound samples of seven relevant manufacturing activities was created and open-sourced for future research. Within this work, we demonstrate that acoustic detection of assembly tasks can be performed with reliable accuracy, highlighting the potential of leveraging acoustics for context-awareness in manufacturing.

Our method aims to address current limitations of wearables and camera-based techniques and can be used to enhance these approaches by incorporating acoustics into multi-modal recognition of manual work by leveraging sensor fusion techniques. The contribution of our work is twofold. First, we introduce the first public dataset[1] for sound-based classification of work tasks. This dataset aims to address the lack of available open-source data in the manufacturing domain and enable future research on the use of acoustics recognition in assembly. Second, we present a new method for acoustic classification of manual work tasks, which serves as a baseline for future research and demonstrates the feasibility of our approach. Moreover, by open-sourcing our method[2], we aim to contribute to a faster transfer of our findings into related research, thus benefiting the community and advancing research on adaptive assistance systems in assembly.

2. Background and Related Work

In manufacturing and assembly, the trend of product customisation is increasing the complexity of labour. Under these conditions, the workforce will face increased cognitive demands to ensure high productivity and error-free production [26]. To mitigate negative effects on workers, such as increased mental or physical strain, adaptive assistance systems have been introduced as a potential solution [7]. Adaptive assistance systems refer to the implementation of context-awareness to detect the current stage of work and support workers depending on their current needs [15]. Examples of adapting assistance depending on the work context involve alerting workers upon missing work steps or preparing and handing tools required for the next step. Other applications involve adapting the content of work information displayed on an interface, or error handling [11].

To enable adaptivity, the system has to acquire data related to the user and the work progress. This can be achieved by human action recognition (HAR), which focuses on the recognition of the performed tasks [4]. Recently, HAR has been used in adaptive assistance systems to trigger action-specific assembly instructions [22] or improve human-robot collaboration during assembly tasks [28].

In previous works, different approaches for recognising assembly tasks have been proposed. One way to derive this information is through wearable sensors, such as inertial measurement units (IMUs), which are placed on the hands of the worker. The method presented in [1] utilises one sensor on each hand and convolutional neural networks (CNN) to process the data. In [25], IMUs were combined with electromyography to infer muscular activity of the forehand. However, the invasive nature of body-worn sensors reduces work comfort. Another way to perform HAR is via visual recognition with a camera. Popular methods include extracting the pose of the worker, which can be further processed with hidden Markov models [2] or LSTM networks [20], or applying CNN to detect the task from the video stream [23]. However, recognition by visual means is prone to occlusions, visual similarity of tasks, and temporal redundancy of the data.

Recently, the combination of audio-visual features has become increasingly popular in video processing. Leveraging acoustics for action recognition was first presented in [8]. By processing image-audio pairs, the work displayed similar levels of accuracy as complex visual models, while reducing the computational cost and improving the processing speed. Other works that utilise acoustics for action recognition followed, with applications in crowd activities [27] or recognition of activities performed at home [12]. Due to the discreet nature of data acquisition and the acoustic characteristics of manufacturing and assembly tasks, leveraging acoustics for action recognition holds promising potential. Despite this, to the best of our knowledge, no work has attempted to close this gap by incorporating sounds into the recognition of assembly tasks.

3. Method

In this section, we introduce "Make some Noise", a novel method for detecting assembly operations by their acoustic characteristics. This approach can be used as an alternative to current action recognition techniques in assembly, mitigating the challenges imposed by the use of wearable sensors or cameras. The overview of our approach is shown in Figure 2. In Section 3.1, we introduce the dataset including technical specifications and further describe the processing performed on the data (Section 3.2) and present the architecture of our network (Section 3.3).

[1] https://doi.org/10.48436/hv20e-zzb35
[2] https://github.com/finklorenz/MakeSomeNoise

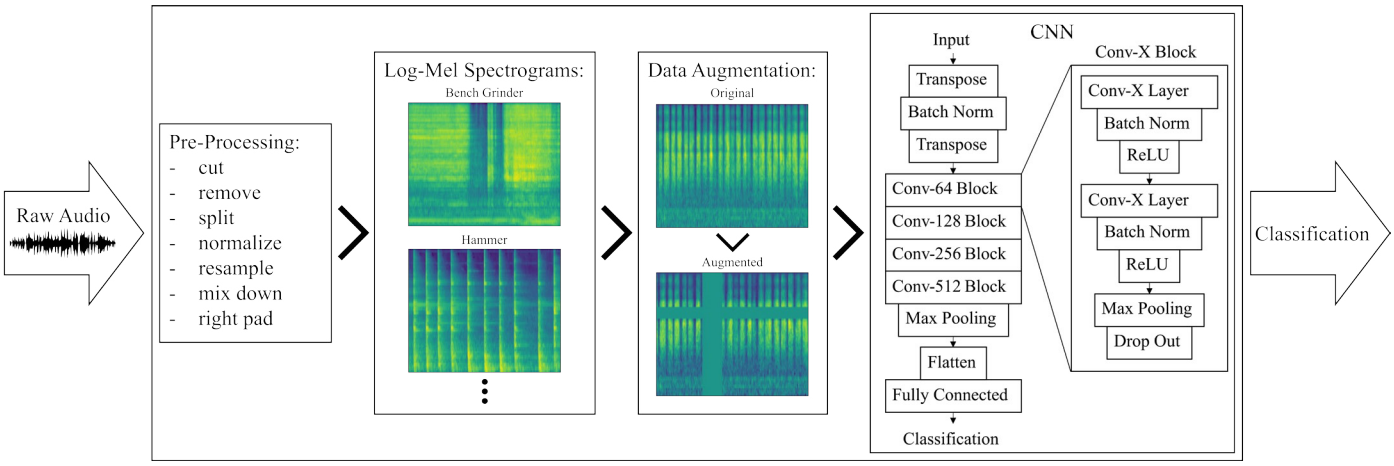

Fig. 2: Overview of the pre-processing, augmentations, and the proposed architecture.

3.1. Dataset

The lack of accessible datasets serves as a great challenge for research on adaptivity in assembly assistance. Currently, a dataset that contains sounds specific to the manufacturing and assembly domain is not available. Therefore, we developed a new dataset for action classification via acoustics and made it publicly available [6].

The dataset contains a selection of typical production activities, which can be classified based on their typical sounds. The different classes and their distribution are depicted in Figure 3. In total, almost 7 hours of data was recorded and, after pre-processing, the recorded dataset consists of 6,117 audio clips, each 4 seconds long.

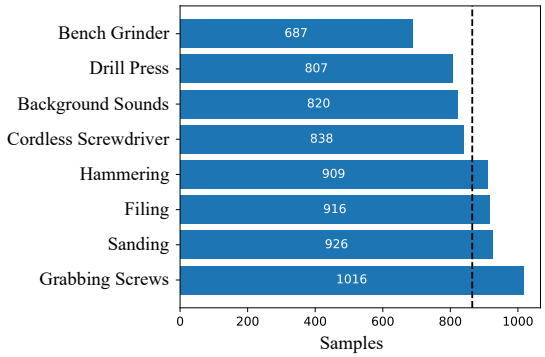

Fig. 3: Distribution of the different labels. Dashed line represents the mean sample size per class.

The dataset was constructed in a laboratory facility under realistic settings, including background noise from machinery. The samples were recorded using the built-in microphones of an iPhone and of a notebook with different apps. The position of the microphone was varied, and executions of noise production (e.g., pace, intensity, movement pattern) were regularly changed to increase data variance. The variance is further increased by samples with a varying number of audio channels, sampling rate, and bit depth.

In order to include “unknown” sounds not represented in our dataset, such as human conversations, sneezing, or ambient noise, we have introduced an additional class, the background class, for training purposes. The background class was sampled from *ESC-50: Dataset for Environmental Sound Classification* [17], which contains various different environmental sounds. This was done to avoid possible misclassifications and increase the robustness of the classification. In practice, this class aims to label all outputs which are unknown to the model as “unknown background noise”.

3.2. Data Transformation and and Augmentation

The audio recordings were divided into samples of uniform length. Those Samples that were too short to extract relevant information from were removed, while samples that were slightly too short but still contained the bulk of the information were right-padded. The samples were also normalised between [-1.0, 1.0], resampled to 44.1 kHz, and mixed down to mono audio. For training purposses, a 80/20 training to validation split was applied.

After preprocessing, the raw audio signals were transformed to Mel spectrograms. Then, the spectrograms were transformed from the power-amplitude scale to the decibel scale (log-Mel spectrogram). The number of Fast Fourier Transformation (FFT) bins was set to 1,024, which is the same as the window length. We use a hop length of 512 and the number of Mel filters was set to 64. The resulting size of the spectrogram is [64,345], examples are shown in Figure 2.

To increase robustness during training, several data augmentation techniques were implemented and applied with a probability of 70%. Raw audio files are randomly cropped to 90% at their beginning and end. Furthermore, Gaussian noise was added with a random signal-to-noise ratio (SNR) between 0.001 and 0.05. The short-time Fourier transform (STFT) of the spectrogram was stretched in time without modifying the pitch for a rate between 0.8 and 1.2. Lastly, we use masking both in the frequency domain and in the time domain at a random position with a maximum of 20% of the length.

3.3. Model

The proposed CNN model, shown in Figure 1, was originally inspired by the VGG architecture [24], with modifications for improved detection on spectrograms. The network consists of four convolutional blocks and a classification head, each block further consisting of two convolutional layers with a ReLU as an activation function. Furthermore, we utilise batch normalisation and a dropout layer in each block for improved stability and regularisation. Overall, the model contains almost 4.7 million trainable parameters.

To normalise the acoustic frequencies of action, we introduced a batch-wise frequency normalisation layer as the first block of our network. This is achieved by two transposing function, with a batch normalisation layer in between. As different materials in manufacturing might produce different frequencies of manual operations, such as filing or drilling, we aim to normalise the data by centering the mean and standard deviation per frequency bin over the whole mini-batch. In the end, this aims to better normalise the acoustic characteristics of the dataset.

To train the network, we use cross entropy loss, and AdamW as the optimiser. Further details on the hyperparameters can be found in our GitHub repository[3].

4. Results

To demonstrate the value of our approach, we performed an evaluation in three different scenarios. First, in 4.1, we present the results computed on the dataset introduced in Section 3.1. In order to compare our method to strong baselines, we performed an evaluation on the UrbanSound8k dataset (4.2). Lastly, we provide experimental results from a realistic simulation of a manufacturing workflow, where classes are sequentially ordered (4.3).

4.1. Results on the Presented Dataset

To evaluate the presented architecture, the classifier was trained on the dataset as described in Section 3, including a background class sampled from ESC-50 [17], until a convergence was reached. We have achieved an overall validation accuracy of 97.76%, and the resulting validation metrics were computed per class and are depicted in Table 1.

The metrics presented demonstrate the feasibility of our approach, given the low computational cost of our architecture. From Table 1, accurate and well-distributed recognition metrics can be observed for each class. Considering the similarities in the movement patterns among the represented classes, such as *sanding* and *filing*, this demonstrates that the use of sounds can aid in the recognition of assembly tasks.

[3] https://github.com/finklorenz/MakeSomeNoise

Table 1: Multiclass metrics per label.

Classes	Precision	Recall	F1
Bench grinder	0.9856	1	0.9928
Cordless screwdriver	0.9824	0.9882	0.9853
Drill press	0.9877	0.9938	0.9908
Filing	0.9785	0.9838	0.9811
Grab screws	0.9952	0.9952	0.9952
Hammer	0.9891	0.9628	0.9757
Sanding	0.9892	0.9892	0.9892

4.2. Ablation Experiments: UrbanSound8k Dataset

To compare our method to existing models, we performed additional experiments on the UrbanSound8k dataset [19], which is considered as one of the most challenging datasets for sound classification tasks. To create log-Mel spectrograms, we use the same settings as described in Section 3.2, with the exception of resampling to 22.05 kHz, as proposed in [9], and setting the number of Mel filters to 128. Moreover, similar to [9], we introduced label-smoothing as an additional regularisation, and one-cycle learning rate scheduler for faster convergence. We use a batch size of 128 and set the learning rate of 0.001. We did not further optimise or change the architecture proposed in Section 3.3. We trained the model from scratch using the official 10-fold cross-validation split.

We present the results in comparison to other relevant baselines in Table 2. As discussed in [10], we do not include works that use pre-training or validation methods other than the official cross-validation split in order to ensure transferability and reproducibility.

Table 2: Comparison of state-of-the-art methods on UrbanSound8K dataset.

Method	Feature	Acc (%)	Param
Baseline [19]	MFCC	68	SVM
Piczak-CNN [16]	Log-Mel	73.7	26 M
Pyramid-Combined CNN [5]	Spectrogram	78.14	20 M
ESResNet (mono) [10]	STFT	79.91	23.53 M
AemNet WM1.0 [13]	Log-Mel	81.5	5 M
EAT-S [9]	Raw data	**85.5**	5.3 M
Our Method	Log-Mel	77.98	**4.69 M**

We demonstrate that our method achieves a high accuracy on the UrbanSound8k dataset, while requiring less parameters than other models. In practice, this results in a lower computational demand during training, as well as lower memory requirements during the deployment.

4.3. Tests on Continuous Workflow

So far, we have presented an evaluation on datasets that are sampled from a well-distributed sample set. However, we believe that such metrics are not directly transferable to real-life assembly, due to the disturbances inherent in the work process

and the environment. Such disturbances may include overlapping actions, or samples that do not contain any actions at all, both possibly degrading the results achievable during deployment.

In order to evaluate our approach during a continuous and realistic workflow, and not only on individual data samples, complete sequential workflows were recorded and processed with the network. In the process, the seven classifiable assembly tasks were performed directly one after the other. To process the audio stream, we implemented a sliding-window approach with a window size of 4 seconds. We set the window to move forward by 0.5 seconds after each sample, thus overlapping with the previous windows by 3.5 seconds. During the testing, we deliberately included difficult cases, such as activities executed for a shorter time than the size of the sliding window or merging of two activities to deceive the network and create ambiguous inputs.

In total, 11 tests were carried out, each between 1-1.5 minutes long. Of the 11 tests, 62 of the 77 recorded assembly activities were recognised correctly (80.5%). An analysis has shown that, on average, our method requires approximately 1.5-2 seconds to recognise the sound of the assembly step. We note that this delay is inherent to the sliding window approach, as the window must contain an appropriate number of data items related to the action in order to classify the action correctly.

5. Discussion

5.1. Potential of Acoustics in Classification of Assembly Steps

Despite the success of leveraging acoustics in the area of data science, especially for human action recognition, no previous works were able to evaluate the potential of acoustic classification of assembly tasks given the lack of available data. Hence, our work presents the first experimental foundation and aims to motivate and guide future research in this area. The results highlight the ability to detect manual work steps by accurately classifying their characteristic sounds.

Because of the materials used in manufacturing and the sounds they produce, as well as the repetitive pattern of certain work tasks (e.g., filing, hammering), we hypothesise that many assembly tasks and their respective sounds have clear decision boundaries. By utilising acoustics, we show that less complex models, such as the one presented in our work, are sufficient to exploit these clear boundaries.

5.2. Implications on the Design of Adaptive Assistance

CNNs in combination with spectrogram images are the most widely used methodology to classify sounds, mainly because they are computationally cheap to implement. Compared to other works, especially in the area of camera-based HAR, we demonstrate that sophisticated models may not be necessary to recognise assembly tasks. Thus, models that utilise acoustics could provide a similar recognition accuracy with lower computational demand. This makes deployment directly at the edge possible, without the need for powerful computational resources. Therefore, this approach could be highly suitable for context-awareness in assembly and thus adaptive assistance systems.

We also demonstrate that acoustic classification is a viable alternative to wearable sensors, which reduce work comfort. Moreover, wearables suffer from calibration issues or sensor noise. In practice, we believe that a multi-modal approach, that is, combining acoustics with wearables or cameras, holds promising potential for the design of assistance systems in assembly.

5.3. Limitations and Future Work

Due to the interactive nature of assembly assistance systems, recognition of assembly steps should be performed with a low level of latency. Only in this scenario, assistance can be adapted according to the context of the environment. As a result of the windowing-method chosen for live classification, there is a time delay of 1.5-2 seconds before the system catches up with the classification. This could be a constraint for real-time applications. At the same time, we note that a high-level comparison in terms of *reactiveness* cannot be performed due to the current lack of metrics provided by other works. Therefore, we would like to encourage future works to include such metrics, to enable practical evaluation and support real-life deployment.

When compared to other established datasets, such as UrbanSound8k, our model displays a lower precision than the results obtained on our presented dataset. This is primarily due to the fact that our dataset was sampled in one single environment, resulting in a higher degree of similarity. On the contrary, the sounds within established datasets exhibit great variance. Moreover, the range of the activities represented in our dataset is limited. While aiming to present a proof-of-concept for acoustics classification, we suggest and encourage future research to investigate the impact of a larger number of classes and a greater variety of recordings on classification performance. Moreover, subsequent research efforts should focus on refining more complex networks to reduce classification time.

Furthermore, we note that acoustic classification may not be suitable for all actions and environments. Certain assembly actions may not be distinguished by sounds alone, and sometimes the sounds of activities may be too quiet or impacted by ambient noise to be distinguished. We believe that a combination with other modalities, such as cameras or wearables holds the potential to address this issue.

6. Conclusion

In this work, we propose a novel approach for the classification of assembly tasks by leveraging acoustics. Our method aims to resolve the challenges current techniques for the classification of manual work tasks face and open up new areas of research on the use of acoustics in action recognition in manufacturing. To address the lack of open data, we introduced the first public dataset containing sounds of relevant manual work

tasks and demonstrated that they can be detected with a high accuracy. In the future, we aim to explore this area of research further, providing additional baselines and deploying the concept in novel applications.

Acknowledgements

The Institute of Production Engineering and Photonic Technologies would like to acknowledge the funding of this work by the European Commission through the Horizon Europe FLEX4RES project with Grant Agreement Number 101091903.

The Department of Human-Machine Interaction would like to acknowledge the funding of this work by the Austrian Research Promotion Agency as a part of the project A2P (FFG-891196).

References

[1] Al-Amin, M., Qin, R., Tao, W., Doell, D., Lingard, R., Yin, Z., Leu, M.C., 2022. Fusing and refining convolutional neural network models for assembly action recognition in smart manufacturing. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science 236, 2046–2059.

[2] Berger, J., Lu, S., 2022. A multi-camera system for human detection and activity recognition. Procedia CIRP 112, 191–196. doi:https://doi.org/10.1016/j.procir.2022.09.071. 15th CIRP Conference on Intelligent Computation in ManufacturingEngineering, 14-16 July 2021.

[3] Buchholz, V., Kopp, S., 2023. Adaptive assistance systems: Approaches, benefits, and risks, in: The Digital Twin of Humans: An Interdisciplinary Concept of Digital Working Environments in Industry 4.0. Springer.

[4] Bulling, A., Blanke, U., Schiele, B., 2014. A Tutorial on Human Activity Recognition Using Body-Worn Inertial Sensors. ACM Comput. Surv. 46. doi:10.1145/2499621.

[5] Demir, F., Turkoglu, M., Aslan, M., Sengur, A., 2020. A new pyramidal concatenated CNN approach for environmental sound classification. Applied Acoustics 170. doi:10.1016/j.apacoust.2020.107520.

[6] Fink, L., 2022. MSDv1: Manufacturing Sound Dataset for Classification of Work-related Actions and their Sound, TU Wien. doi:10.48436/hv20e-zzb35.

[7] Funk, M., Dingler, T., Cooper, J., Schmidt, A., 2015. Stop Helping Me - I'm Bored! Why Assembly Assistance Needs to Be Adaptive, in: Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers, Association for Computing Machinery, New York, NY, USA. pp. 1269–1273. doi:10.1145/2800835.2807942.

[8] Gao, R., Oh, T.H., Grauman, K., Torresani, L., 2020. Listen to look: Action recognition by previewing audio, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10457–10467.

[9] Gazneli, A., Zimerman, G., Ridnik, T., Sharir, G., Noy, A., 2022. End-to-end audio strikes back: Boosting augmentations towards an efficient audio classification network. arXiv:2204.11479.

[10] Guzhov, A., Raue, F., Hees, J., Dengel, A., 2021. Esresnet: Environmental sound classification based on visual domain models, in: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE. pp. 4933–4940.

[11] Heinz-Jakobs, M., Oestreich, H., Wrede, S., Röcker, C., 2022. User expectations regarding design dimensions of adapative assistance systems, in: 2022 15th International Conference on Human System Interaction (HSI), IEEE. pp. 1–7.

[12] Jung, M., Chi, S., 2020. Human activity classification based on sound recognition and residual convolutional neural network. Automation in Construction 114, 103177.

[13] Lopez-Meyer, P., del Hoyo Ontiveros, J.A., Lu, H., Stemmer, G., 2021. Efficient end-to-end audio embeddings generation for audio classification on target applications, in: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 601–605. doi:10.1109/ICASSP39728.2021.9414229.

[14] Mark, B.G., Rauch, E., Matt, D.T., 2021. Industrial Assistance Systems to Enhance Human–Machine Interaction and Operator's Capabilities in Assembly, in: Matt, D.T., Modrak, V.Z.H. (Eds.), Implementing Industry 4.0 in SMEs: Concepts, Examples and Applications. Springer International Publishing, Cham, pp. 129–161.

[15] Oestreich, H., Heinz-Jakobs, M., Sehr, P., Wrede, S., 2022. Human-centered adaptive assistance systems for the shop floor, in: Human-Technology Interaction: Shaping the Future of Industrial User Interfaces. Springer, pp. 83–125.

[16] Piczak, K.J., 2015a. Environmental sound classification with convolutional neural networks, in: 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), pp. 1–6. doi:10.1109/MLSP.2015.7324337.

[17] Piczak, K.J., 2015b. ESC: Dataset for Environmental Sound Classification, in: Proceedings of the 23rd ACM International Conference on Multimedia, Association for Computing Machinery, New York, NY, USA. pp. 1015–1018. doi:10.1145/2733373.2806390.

[18] Roth, E., Möncks, M., Bohné, T., Pumplun, L., 2020. Context-Aware Cyber-Physical Assistance Systems in Industrial Systems: A Human Activity Recognition Approach.

[19] Salamon, J., Jacoby, C., Bello, J.P., 2014. A dataset and taxonomy for urban sound research, in: Proceedings of the 22nd ACM International Conference on Multimedia, Association for Computing Machinery, New York, NY, USA. p. 1041–1044. doi:10.1145/2647868.2655045.

[20] Schirmer, F., Kranz, P., Schmitt, J., Kaupp, T., 2023. Anomaly detection for dynamic human-robot assembly: Application of an lstm-based autoencoder to interpret uncertain human behavior in hrc, in: Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction, pp. 881–883.

[21] Schlund, S., Kostolani, D., 2022. Towards designing adaptive and personalized work systems in manufacturing. Digitization of the work environment for sustainable production , 81.

[22] Schröder, M., Ritter, H., 2017. Deep learning for action recognition in augmented reality assistance systems, in: ACM SIGGRAPH 2017 Posters.

[23] Sener, F., Chatterjee, D., Shelepov, D., He, K., Singhania, D., Wang, R., Yao, A., 2022. 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, pp. 21096–21106.

[24] Simonyan, K., Zisserman, A., 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition, in: Bengio, Y., LeCun, Y. (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.

[25] Tao, W., Lai, Z.H., Leu, M.C., Yin, Z., 2018. Worker activity recognition in smart manufacturing using imu and semg signals with convolutional neural networks. Procedia Manufacturing 26, 1159–1166. doi:https://doi.org/10.1016/j.promfg.2018.07.152. 46th SME North American Manufacturing Research Conference, NAMRC 46, Texas, USA.

[26] Torres, Y., Nadeau, S., Landau, K., 2021. Evaluation of fatigue and workload among workers conducting complex manual assembly in manufacturing. IISE transactions on occupational ergonomics and human factors 9, 49–63.

[27] Wang, W., Seraj, F., Havinga, P.J., 2020. A sound-based crowd activity recognition with neural network based regression models, in: Proceedings of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments, pp. 1–8.

[28] Zhang, J., Wang, P., Gao, R.X., 2021. Hybrid machine learning for human action recognition and prediction in assembly. Robotics and Computer-Integrated Manufacturing 72, 102184.