3rd Conference on Production Systems and Logistics

# Towards A Flexible Approach To Transfer Machine Operation Know-How From Experts To Beginners With AI

Timo Leitritz[1], Martina Köhler[1], Christian Jauch[1]
[1]*Fraunhofer Institute for Manufacturing Engineering and Automation IPA, Stuttgart, Germany*

**Abstract**

Training new users at a production machine is a time intensive and expensive task. To reduce the effort in this task we examine the possibilities of enhancing a production machine with a system that is able to learn from its users and teach inexperienced users this knowledge: Self-Learning and Self-Explanatory Machine SLEM. The learning process of SLEM relies on watching an experienced user working on a machine using camera-based human activity recognition which predicts the acitivities based on the estimated human skeleton in the video stream. SLEM must be able to work with little data to reduce the learning time as much as possible. Thus, this paper shows that training an activity recognition model solely on one experienced individual's actions can lead to comparatively high activity recognition accuracy despite the low data variety. The results show that training on a single-person dataset can reach relatively high accuracy levels and is a suitable way of training the model in the industrial setting. For the teaching process, in which the system has to compare the actual activities with the target acitivities to give feedback, the activity recognition has to run in real-time. Different amounts of input data for the activity recognition model are examined and lead to a configuration with little accuracy loss and sufficient latency performance.

**Keywords**

Machine Learning; Human Activity Recognition; Pose Estimation; Skeleton; Industrial; Online Activity Recognition

## 1. Introduction

Highly complex machines in manufacturing companies need to be operated and maintained by human experts. Furthermore, a lot of older machines are expensive to upgrade even though a lot of the older knowledge about the machine is missing, so they require specialized knowledge from experts. This poses challenges whenever a machine expert leaves the company and is not able to transfer their knowledge to more inexperienced workers. Growing employee turnover causes further important knowledge to get lost. Additionally, teaching inexperienced workers is not only a time intensive task, but also causes considerable machine downtime.

The developed system from the publicly funded project "Self-Learning and Self-Explanatory Machine" (SLEM) tried to solve this issue using a variety of AI-based methods. A camera-based system was able to learn the manual interaction with the machine from an experienced user (subsequently referred to as "experts"). Furthermore, the system was able to generate a guideline for optimal machine operation from this gained knowledge to help inexperienced users ("beginners").

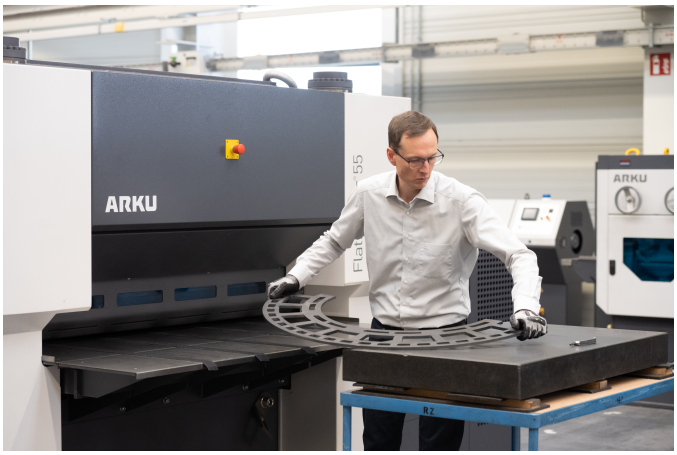


Figure 1: ARKU leveler FlatMaster® 55

This paper presents the results of the SLEM project on skeleton based activity recognition using PoseC3D [1], as well as comparing two types of scenario driven datasets: “Expert” and “Mixed”. The former contains data only from one expert user, whereas the latter combines data from a variety of differently skilled users. A skeleton based Human Activity Recognition (SHAR) was chosen due to the highly generalized options for pose recognition that are available as well as due to the ability to abstract from the image based representation of a human, that can differ greatly regarding the appearance, e.g. in cloth type or hair color. The following topics are addressed in this paper:

- Recording, annotating and preprocessing a scenario driven dataset from a real world use case at a company, see Figure 1
- Integrate PoseC3D in an application oriented real time activity recognition pipeline for an actual industrial use case
- Examination of accuracy on “Expert” and “Mixed” dataset
- Examination of different input clip sizes and pose estimation intervals to optimize system latency

## 2. Related work

Human Activity Recognition (HAR) in industrial settings promises many benefits for assistance systems in manual work. An AI-based assistance system with implemented HAR is able to automatically detect and predict what the human is currently doing or even going to do. Based on this information the assistance system is able to react, e.g. by giving support or security warnings. The required activities must first be learned by the HAR method on annotated data and the granularity of the defined activities can vary greatly, from detecting “walking” and “standing still” to precise gestures. In the recent years, many new methods for HAR have been developed, using a variety of different deep learning architectures such as 2D-CNN, 3D-CNN, RNN, GCN and Transformers. In general, these approaches can be further differentiated by the type of input data they use:

- Integrated sensors, e.g. accelerometer data [2] [3]
- Image and video data [4] [5] [6]
- Skeleton data [1] [7] [8]

Data from integrated sensors, such as accelerometer data, is a common data source for HAR since it is available in every smartphone and is a rather inexpensive technology. The detail of activities detected based on such data is limited and mainly focuses on detecting specific movements of the person wearing the sensor. Using image and video data offers a similar data source as a human uses to recognize activities and scenes with the sense of sight. Thus a higher detail of activities can be recognized with the cost of significantly

more computation in training and execution of the HAR approach. In addition, a great amount of diverse annotated data is required to train a generalized model and being able to recognize activities on people with varying appearances. In contrast to that, using skeleton data for HAR solves this issue by abstracting from the visual appearance of the person and focusing on keypoints of the human skeleton.

Furthermore, different types of input data can be combined, which was shown in [9], where a unified framework using visual and skeleton features was able to recognize human activities.

Several publications also focused on industrial settings [10] [11] [12]. An end-to-end approach using only skeleton data as input was developed by [10]. A video of an industrial scenario was given and the output was an activity class. First, skeleton data was predicted for each frame in the video through pose estimation, specifically using the stacked hourglass model. Afterwards, leveraging a spatial transformer (STN), the skeleton was further augmented towards translation, rotation, and scaling in order to give the data more human body pose diversity. A graph convolutional network (GCN) [7] was used in the end in order to preserve the relationship between the limbs from which activity scores were calculated. Due to the nature of the augmentation by the STN, the authors found that some activities – such as walking back and forth – are not rotation invariant and disadvantaged the HAR unnecessarily.

A slightly different industry application scenario is the interaction with robots. Since many industrial robots work as human assistances, it becomes more and more important for them to know what the human is doing. [11] used 3D skeleton data and a random forests binary classifier in order to classify primitive movement actions in real time, since such actions could already assist in human-robot interaction.

Another publication focused on using HAR as a way to make interactions between users and industrial machines more natural [12]. The authors used images as raw input for the networks and tested various convolutional networks especially in regards to their execution time. This is particularly important in order to fit into their Natural Machine Operation (NaMO) framework that was developed in order to create a fluid workflow between human and machine and to fit HAR into this workflow. The authors show that it is possible to train HAR models on small industrial datasets.

In general, HAR poses many challenges in industrial settings. Overfitting on small datasets, as well as the effort of having to manually annotate these datasets. The latter poses special challenges when there are no commonly agreed upon classes in such individual cases. However, by using a skeleton-based approach the amount of data necessary to train a generalized model can be reduced, since the skeleton data provides an abstraction of appearance. Thus, this paper focuses on training HAR with limited datasets in terms of variety by only using data from a single person and evaluating the result on two separate individuals, which was not examined in detail by the above publications [10] [11] [12].

## 3. Methods

Section 3.1 explains the core deep learning pipeline in terms of the model architecture as well as the training and inference pipeline to provide a detailed understanding of the used HAR approach. The data acquisition and preparation techniques are explained in Section 3.2, focusing on all relevant steps to prepare the dataset for training. This is especially important since training and inference were conducted on a custom industrial dataset with custom activity classes. Section 3.3 describes the used metrics in the following experiments to prepare for the evaluation in Chapter 4.

### 3.1 Deep learning pipeline

The pipeline from video stream to classified activity regarding SHAR consists of two main parts: Pose extraction and activity recognition. For activity recognition PoseC3D [1] was selected as promising SHAR architecture as it achieved the highest accuracy on the NTU60 cross-view test dataset as of January 2022

[13]. It was implemented in the widely used MMAction2 framework by OpenMMLab [14]. PoseC3D uses a 3D-CNN to recognize activities from a 3D heatmap volume, generated from the skeleton data of the video. The key is to convert 2D pose information to a heatmap that encodes the skeleton information of a single frame and can be processed efficiently by a CNN architecture. Each pose keypoint (e.g. elbow, eye, knee) is represented with a Gaussian distribution at the corresponding location as one channel in the resulting heatmap. These heatmaps are generated per frame of the input video clip and then stacked together to a 3D-Heatmap that represents the pose information of the video clip. For pose extraction the same two stage approach was used as in [1]. This approach consists of a Faster R-CNN [15] object detector to detect all people in the image and a HRNet [16] pose estimator to extract the skeleton information per person. The object detector is implemented in the MMDetection framework [17] whereas the pose estimation model is implemented in the MMPose framework [18]. The pose estimator is trained on the COCO [19] body keypoints dataset.

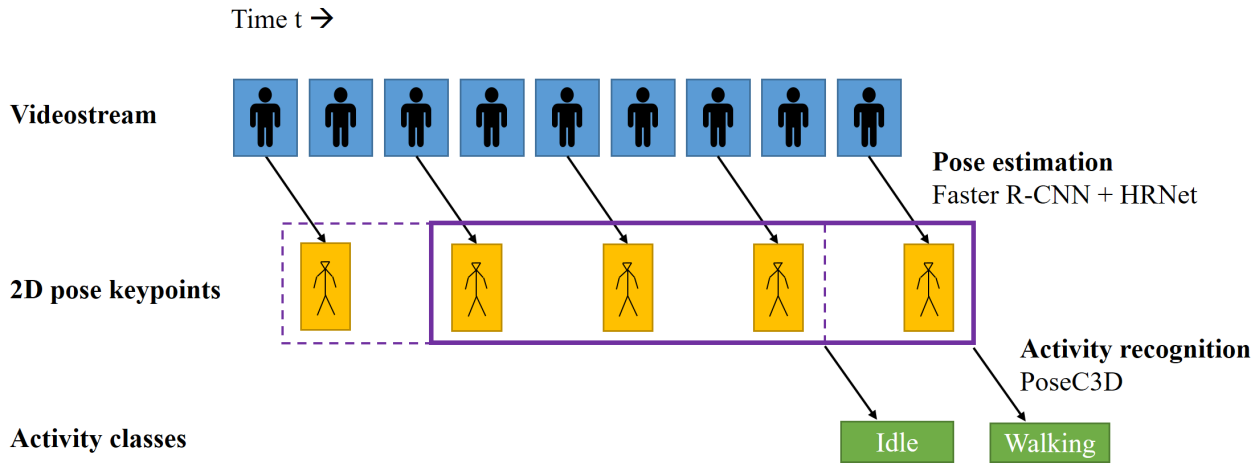


Figure 2: Activity recognition pipeline illustration. The pose estimation is only applied on every second frame, which is due to the pose estimation interval $I_p = 2$. The violet rectangle represents the sliding window of pose data that acts as input for the activity recognizer PoseC3D.

Figure 2 illustrates the SHAR pipeline in an online scenario. In contrast to the NTU60 offline scenario, PoseC3D has to classify video clips of constant length instead of varying length. This is why "Uniform Sampling" as proposed in [1] was not applied in this work. In the online pipeline shown in this paper, PoseC3D takes a fixed amount of $N_{PC3D}$ poses including the current and $N_{PC3D} - 1$ prior poses as inputs. The pose estimation can be executed on each frame to extract the poses, but with a given framerate of 30 fps, this is very computationally expensive. Alternatively, poses can be estimated once every $p$ frames in a pose estimation interval $I_p$. Estimated poses are saved in a ring buffer of length $N_{PC3D}$ to be partially reused for the next activity recognition execution.

### 3.2 Data acquisition and preparation

To acquire data that can be used to train and test these models, videos of workers of different skill levels (beginner, intermediate, expert) during usual tasks on a metal leveler machine were recorded (see Figure 1). This was done as part of a Data Driven User Needs Assessment [20], which aims to analyze user experiences of the workspace based on machine data, visual data, eye tracking as well as interviews with the machine operator. Roughly 2 h 40 min (6 expert, 2 intermediate and 3 beginner videos) of 2D-RGB video material in total was recorded with a video resolution of 2560 by 1440 pixels using a Microsoft Kinect Azure DK camera sensor. In an offline post-processing step the 2D pose data of each frame is extracted from all videos by the pose estimation method described in Section 3.1.

The videos were annotated frame-by-frame, where each frame received exactly one class label. The analysis of the video data resulted in the decision on nine activity classes for working on the leveler machine: *"0 - Idle", "1 - Walk", "2 - Measure with tool", "3 - Rearrange part", "4 - Test part with hands", "5 - Pick object", "6 - Place object", "7 - Interact with machine interface", "8 - Feed part to machine".*

The annotated videos recordings were split into a training and validation dataset with two variants (see Table 1) and one test dataset according to Table 1. In addition, the training and validation dataset was split into training (90%) and validation dataset (10%) after sampling and balancing the dataset. The goal of the sampling was to convert the dataset of large videos with mixed activity annotations into short video clips of fixed size and exactly one activity class. From each video recording, video clips of a fixed clip size were sampled by moving a sliding window with a step size of one over the recording, resembling the online pipeline scenario from Section 3.1. Resulting clips with annotations from more than one activity class were discarded. Furthermore, the balancing was applied due to large differences in the class occurrences in the dataset. A new data set with balanced class occurrences was derived from the sample of clips. The resulting dataset only contained pose data and no image data since PoseC3D only requires pose data.

Table 1: Overview of the amount of video scenarios in training, validation and test datasets

| Dataset name | Expert | Intermediate | Beginner | Total |
|---|---|---|---|---|
| Training/validation "Mixed" | 5 | 1 | 2 | 8 |
| Training/validation "Expert" | 5 | 0 | 0 | 5 |
| Test | 1 | 1 | 1 | 3 |

### 3.3 Evaluation metrics

For evaluation and comparison, the accuracy of the resulting predictions was calculated. More specifically, top-K accuracy was used with $K = 1$ and $K = 5$. This means that a prediction was classified as correct, if the actual class occurs in the $K$ highest confidence predictions.

Furthermore, a distinction between "total" top-1 accuracy as well as scenario specific accuracies "Beginner", "Intermediate" and "Expert" was added. The "total" top-1 accuracy is calculated on the whole test dataset, whereas the scenario accuracies only took into account predictions from the corresponding scenario.

In order to evaluate the runtime performance of the activity recognition pipeline, "PoseC3D" latency and "total" latency were measured. The former measured the inference time for a single prediction of the PoseC3D model. The latter took into account the amount of time necessary to run the pose estimation as well as the activity recognition for the specified pose interval and clip input size over a time period of 1.6 s of a video stream. This equals 48 frames at 30 fps.

## 4. Experiments

This section describes the conducted experiments that were used to examine various aspects of SHAR. The first experiment focused on the performance difference of PoseC3D on the two industrial dataset variants "Expert" and "Mixed" to investigate the impact of dataset with limited variety. Afterwards, an experiment that compared several model input configurations was conducted with different pose estimation intervals and input clip sizes to analyze the trade-off between accuracy and latency for a real-time HAR scenario.

### 4.1 PoseC3D performance on "Expert" vs. "Mixed" dataset variants

PoseC3D achieves state-of-the-art accuracy on the common and publicly available dataset NTU60 [21]. This experiment aims to show the performance of PoseC3D on the custom industrial dataset. Furthermore, it was

simultaneously examined how well the model can learn from the specialized "Expert" dataset variant in contrast to training on the "Mixed" dataset. It was assumed that models trained on purely the "Expert" dataset would have issues generalizing towards the "Beginner" and "Intermediate" videos.

Training hyper parameters were similar to the default MMAction2 configuration for PoseC3D: Learning rate was set to 0.01 with SGD and batch size adjusted to 16. All clips in the dataset had a length of 48 pose frames. PoseC3D's clip input size was set to 48, so no "Uniform Sampling" was used as the frame number matches the dataset video clip size. A starting checkpoint of PoseC3D that was pre-trained on the NTU60 dataset was used. The training ran on a single NVIDIA Tesla V100 GPU with no augmentation applied. The random crop and flip operations were explicitly removed in the PoseC3D data pipeline, in order to create a baseline without any augmentation that can be built upon in future experiments.

Table 2: PoseC3D accuracy results after training on "Mixed" and "Expert" dataset for 10 epochs with $I_p = 1$ and $N_{PC3D} = 48$

| | Top-1 accuracy | | | | Top-5 accuracy |
|---|---|---|---|---|---|
| Model name | Total | Beginner | Intermediate | Expert | Total |
| PC3D-48-1 „Mixed" | 0.456 | **0.554** | 0.418 | 0.449 | 0.910 |
| PC3D-48-1 „Expert" | **0.508** | 0.433 | **0.487** | **0.584** | **0.952** |

The results of the evaluation in Table 2 show that PoseC3D did not reach the same accuracy level as on NTU60, where it achieved a cross-subject accuracy of 0.941. The best total accuracy result occurred when training on the "Expert" dataset with a total top-1 accuracy of 0.508 whereas the training on the "Mixed" dataset resulted in a drop of 10.2 %. Top-5 accuracy is close between both variants. Since the test dataset includes different amounts of samples per scenario (312 "Beginner", 713 "Intermediate" and 512 "Expert" samples), total top-1 accuracy favors the model that performs better on the "Intermediate" or "Expert" scenario. After training on the "Expert" dataset the top-1 accuracy is significantly better on the "Expert" and the "Intermediate" test data than on the "Beginner" data. This is assumed to be due to the "Expert" training dataset being more similar to the "Expert" and "Intermediate" subsets of the test data than to the "Beginner" test data. However, the model trained on the "Expert" dataset is still able to perform reasonably well on "Beginner" data with a top-1 accuracy of 0.433. This also proves the assumption that the skeleton based approach provided a solid abstraction from the visual appearance of the workers in the original video recordings. Additional augmentation techniques may improve these accuracies further and compensate the lack of variance in the expert dataset variant.

The generally worse performance of PoseC3D on the industrial dataset could be explained by the following data annotation quality problems and the features used for the actual activity recognition. For one, some sections of the video contain ambiguous activities in respect to the defined activity classes, e.g. picking up an object with one hand and measuring a part with a tool in the other hand. This aspect can also be seen in the confusion matrix in Figure 3. It shows the misclassification behavior of the "expert" model on the industrial test dataset. Activity classes *"0 - Idle", "1 - Walking", "2 - Measure with tool", "5 - Pick object", "7 - Interact with machine interface"* and *"8 - Feed part to machine"* were mostly classified correctly. Class *"4 - Test part with hands"* was often misclassified as class *"2 - Measure with tool"*, since it was also a very stationary activity and looked similar in the footage. Class *"5 - Pick object"* and *"6 - Place object"* was often confused with each other, presumably due to it being the same activity but only in reverse. Class *"3 - Rearrange part"* was misclassified most often, which is probably due to its similarity to other classes like *5, 6* and *8*.

These results show how important choosing the correct activity classes is. Some classes were not distinguishable from each other using only skeleton data and may require image features to be distinguished. A way to incorporate image features into the pipeline was shown in [9]. In addition, a very precise definition

of activities and a significant amount of samples for each of the activities is essential for good recognition results.

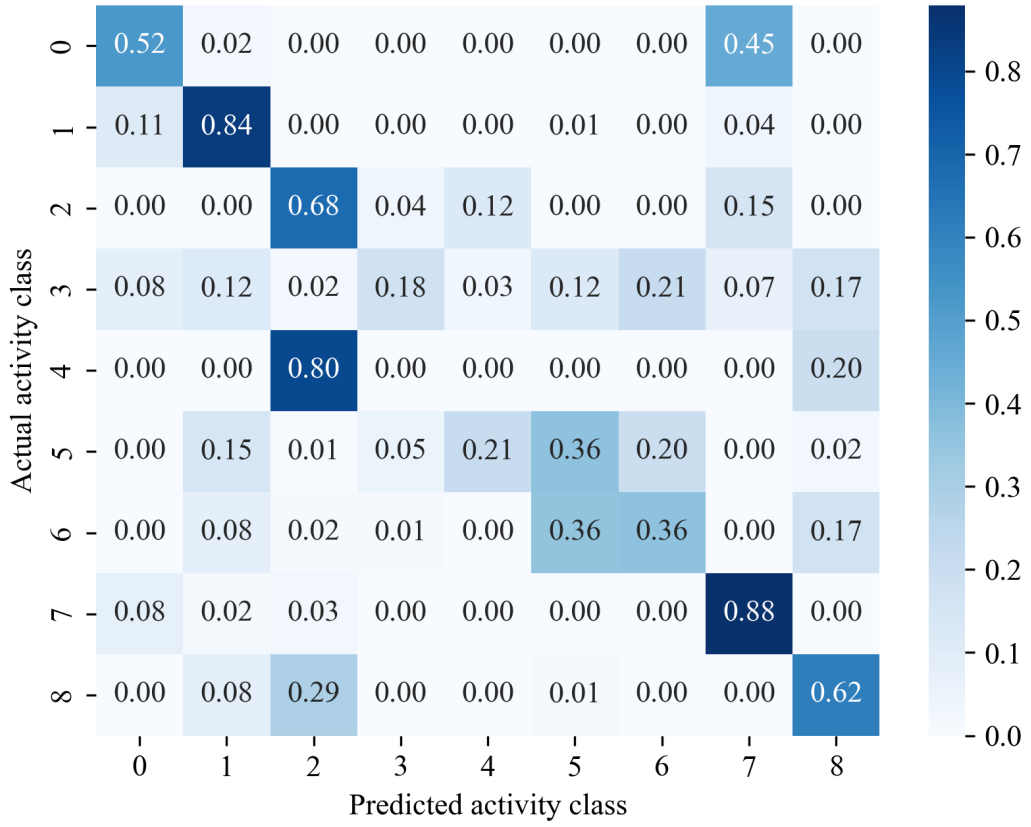


Figure 3: Confusion matrix of PoseC3D "Expert" with $I_p = 1$ and $N_{PC3D} = 48$ after 10 epochs. Cell numbers specify the ratio of occurrence in respect to the total occurrences in the specific row. Activity labels correspond to the following classes: "*0 - Idle", "1 - Walking", "2 - Measure with tool", "3 - Rearrange part", "4 - Test part with hands", "5 - Pick object", "6 - Place object", "7 - Interact with machine interface", "8 - Feed part to machine".*

### 4.2 PoseC3D input data variation

As PoseC3D in its baseline configuration with an input clip size of 48 is rather computationally expensive, the accuracy results and latency under multiple variations of input clip size $N_{PC3D} \in \{48, 24, 16, 8, 4\}$ and the respective pose estimation interval $I_p \in \{1, 2, 3, 6, 12\}$ were examined. These combinations were selected to retain the same time window of 1.6 s for the input clip. Depending on these parameters, the 48-frame clips were subsampled. For example, in the case of *PC3D-24-2* only every other pose frame was collected due to $I_p = 2$, resulting in 24-frame clips that correspond to the input clip size of 24. The training parameters were identical to the experiment of Section 4.1., except for the change in input clip size of the model and changing pose estimation interval in data pre-processing. The experiment was again performed on a NVIDIA Tesla V100 GPU. These model variants were trained on the "Expert" dataset, because it is more relevant to the SLEM use case. It was assumed that there is a trade-off between accuracy and latency, especially since lower latency was achieved by using fewer frames that could potentially cause the model to not receive enough data in order to make an accurate prediction.

Top-1 accuracy in the results from Table 3 look very similar across all model configurations, although *PoseC3D-16-3* achieved the top result with a top-1 total accuracy of 0.513. In contrast to that, top-5 accuracy varied more and showed *PC3D-48-1* on the high end with 0.952 and *PC3D-4-12* on the low end with 0.850. These results directly mirrored the reduction in input clip size and pose data per time interval. At last, latency results showed that the smallest model input size with $N_{PC3D} = 4$ took the least time to recognize the

activities. The total latency result revealed, that only the *PC3D-4-12* variant was able to be executed in real time in the given time window of 1.6 s and on the given hardware, because the total latency did not surpass this time threshold. To lower the latency even further there are several major ways:

- Optimizing model execution using optimized inference libraries like NVIDIA's TensorRT
- Scale up hardware, e.g. with more or faster GPU (or comparable specialized hardware)
- Replace pose estimation method with a faster model combination: Faster R-CNN takes 124 ms for the detection and HRNet 93 ms for pose extraction, in contrast to 19 ms of *PC3D-4-12* for activity recognition. Replacing the R-CNN based object detector with a faster single shot detector like YOLOX [22] could improve latency significantly with a small loss in accuracy.

Table 3: Accuracy and latency results of PoseC3D with variation of input clip size $N_{PC3D}$ and pose estimation interval $I_p$ after 10 epochs of training on "Expert" dataset. Latency results are measured on a NVIDIA Tesla V100 GPU.

| Model configuration | | | Top-1 accuracy | Top-5 accuracy | Latency (ms) | |
|---|---|---|---|---|---|---|
| Name | $N_{PC3D}$ | $I_p$ | Total | Total | PoseC3D | Total |
| PC3D-48-1 | 48 | 1 | 0.508 | **0.952** | 161 | 18144 |
| PC3D-24-2 | 24 | 2 | 0.504 | 0.918 | 94 | 7464 |
| PC3D-16-3 | 16 | 3 | **0.513** | 0.904 | 56 | 4368 |
| PC3D-8-6 | 8 | 6 | 0.491 | 0.880 | 31 | 1984 |
| PC3D-4-12 | 4 | 12 | 0.498 | 0.850 | **19** | **940** |

## 5. Conclusion

This work shows how the state-of-the-art skeleton based HAR model PoseC3D can be applied to an industrial setting to enable real time activity recognition for advanced assistance systems like SLEM. With SLEM, first steps were taken towards a system to assist new workers in machine operation and reduce teaching effort. Specifically, it was of special interest how expert knowledge can be preserved and used to guide less experienced workers. Leveraging a machine learning based approach to HAR enables the possibility to adapt to a wide variety of machine types without any reprogramming necessary for the underlying code base.

Skeleton based HAR proved useful in reducing complexity of the input data as well as abstracting away from the few and select people that are going to be observed in the industrial setting. Resulting accuracies of the conducted experiments show that even with data from only one person, skeleton based HAR is able to recognize activities of other people.

Experiments with different variations in PoseC3D's input clip sizes and the interval used for pose estimation during online activity recognition with PoseC3D make clear that inference latency can be reduced by a significant amount without noticeably affecting accuracy on the industrial dataset. It was shown which pipeline configuration is able to provide real time activity recognition.

In the future, we continue to improve on this approach by examining promising aspects. Class activities that are hard to detect with skeleton data will need further consideration, e.g. by combining high-level image features with skeleton data [9]. Additionally, we want to directly compare skeleton based HAR approaches with image based methods in the future. Optimization of the total system latency is crucial to use a HAR system in a real world application and will be further assessed. Augmentation of pose data promises to expand the available dataset and enable the model to generalize better on people with different sizes and different habits in executing activities. Finally, the amount of data required to train a robust HAR model for industrial applications must be thoroughly studied with respect to application complexity and activity classes.

**Acknowledgements**

The SLEM project is funded by the Baden-Württemberg Ministry of Economics, Labor and Housing as part of the Baden-Württemberg AI Innovation Competition.


## References

[1] Duan, H., Zhao, Y., Chen, K., Shao, D., Lin, D., Dai, B., 2021. Revisiting Skeleton-based Action Recognition. CoRR abs/2104.13586.

[2] Ignatov, A., 2018. Real-time human activity recognition from accelerometer data using Convolutional Neural Networks. Applied Soft Computing 62, 915–922.

[3] Antar, A.D., Ahmed, M., Ahad, M.A.R., 2019. Challenges in sensor-based human activity recognition and a comparative analysis of benchmark datasets: a review, in: 2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR), pp. 134–139.

[4] Hutchinson, M., Samsi, S., Arcand, W., Bestor, D., Bergeron, B., Byun, C., Houle, M., Hubbell, M., Jones, M., Kepner, J., Kirby, A., Michaleas, P., Milechin, L., Mullen, J., Prout, A., Rosa, A., Reuther, A., Yee, C., Gadepally, V., 2020 - 2020. Accuracy and Performance Comparison of Video Action Recognition Approaches, in: 2020 IEEE High Performance Extreme Computing Conference (HPEC). 2020 IEEE High Performance Extreme Computing Conference (HPEC), Waltham, MA, USA. 9/22/2020 - 9/24/2020. IEEE, pp. 1–8.

[5] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., van Gool, L., 2019. Temporal Segment Networks for Action Recognition in Videos. IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (11), 2740–2755.

[6] Feichtenhofer, C., Fan, H., Malik, J., He, K., 2019. SlowFast Networks for Video Recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

[7] Yan, S., Xiong, Y., Lin, D., 2018. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. AAAI 32 (1).

[8] Lee, J., Ahn, B., 2020. Real-Time Human Action Recognition with a Low-Cost RGB Camera and Mobile Robot Platform. Sensors 20 (10).

[9] Luvizon, D.C., Picard, D., Tabia, H., 2018. 2d/3d pose estimation and action recognition using multitask deep learning, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5137–5146.

[10] Jiao, Z., Jia, G., Cai, Y., 2020. Ensuring Computers Understand Manual Operations in Production: Deep-Learning-Based Action Recognition in Industrial Workflows. Applied Sciences 10 (3), 966.

[11] Akkaladevi, S.C., Heindl, C., 2015. Action recognition for human robot interaction in industrial applications, in: 2015 IEEE International Conference on Computer Graphics, Vision and Information Security (CGVIS), pp. 94–99.

[12] Bexten, S., Schmidt, J., Walter, C., Elkmann, N., 2021. Human Action Recognition as part of a Natural Machine Operation Framework, in: 2021 26th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA), pp. 1–8.

[13] Meta AI, 2022. Papers with Code - NTU RGB+D Benchmark (Skeleton Based Action Recognition). https://paperswithcode.com/sota/skeleton-based-action-recognition-on-ntu-rgbd. Accessed 4 February 2022.

[14] MMAction2 Contributors, 2020. OpenMMLab's Next Generation Video Understanding Toolbox and Benchmark. https://github.com/open-mmlab/mmaction2. Accessed 4 February 2022.

[15] Ren, S., He, K., Girshick, R., Sun, J., 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (6), 1137–1149.

[16] Sun, K., Xiao, B., Liu, D., Wang, J., 2019. Deep High-Resolution Representation Learning for Human Pose Estimation, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE.

[17] Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C.C., Lin, D., 2019. MMDetection: Open MMLab Detection Toolbox and Benchmark. CoRR abs/1906.07155.

[18]MMPose Contributors, 2020. OpenMMLab Pose Estimation Toolbox and Benchmark. https://github.com/open-mmlab/mmpose. Accessed 4 February 2022.
[19]Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context, in: European conference on computer vision, pp. 740–755.
[20]Wiedenroth, S.J., Denecke, J., Effenberger, I., 2020. Approach to Understand Learner Needs and Usage of the Outcome for Creativity Techniques Adressing the Learner Experience in Workplaces, in: ICERI2020 Proceedings. 13th annual International Conference of Education, Research and Innovation, Online Conference. 09.11.2020 - 11.11.2020. IATED, pp. 4365–4372.
[21]Shahroudy, A., Liu, J., Ng, T.-T., Wang, G., 2016. Ntu rgb+ d: A large scale dataset for 3d human activity analysis, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1010–1019.
[22]Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J., 2021. YOLOX: Exceeding YOLO Series in 2021. CoRR abs/2107.08430.

## Biography

**Timo Leitritz** (*1995) has been research associate at Fraunhofer Institute for Manufacturing Engineering and Automation IPA since November 2020. Timo Leitritz received a Master's degree in Mechatronics and Information technology at the Karlsruhe Institute of Technology. His main research topics are human activity recognition and AI acceleration.

**Martina Köhler** (*1993) has been research associate at Fraunhofer Institute for Manufacturing Engineering and Automation IPA since April 2021. Martina Köhler achieved her Masters of Science degree at the University of Ulm in Cognitive Systems. Her main research topics are human pose estimation and user experience.

**Christian Jauch** (*1990) has been a research associate at the Fraunhofer Institute for Manufacturing Engineering and Automation IPA in Stuttgart, Germany since 2015 and studied technical cybernetics at the University of Stuttgart. He works in the Image and Signal Processing department and leads the Scene Analysis group. His work focuses on hand pose estimation and activity recognition in industrial scenarios.