3rd Conference on Production Systems and Logistics

Industrial Human Activity Prediction and Detection Using Sequential Memory Networks

Tadele Belay Tuli[1], Valay Mukesh Patel[1], Martin Manns[1]
[1]*PROTECH - Institut für Produktionstechnik, Universität Siegen, Germany*

Abstract

Prediction of human activity and detection of subsequent actions is crucial for improving the interaction between humans and robots during collaborative operations. Deep-learning techniques are being applied to recognize human activities, including industrial applications. However, the lack of sufficient dataset in the industrial domain and complexities of some industrial activities such as screw driving, assembling small parts, and others affect the model development and testing of human activities. The InHard dataset (Industrial Human Activity Recognition Dataset) was recently published to facilitate industrial human activity recognition for better human-robot collaboration, which still lacks extended evaluation. We propose an activity recognition method using a combined convolutional neural network (CNN) and long short-term memory (LSTM) techniques to evaluate the *InHard* dataset and compare it with a new dataset captured in a lab environment. This method improves the success rate of activity recognition by processing temporal and spatial information. Accordingly, the accuracy of the dataset is tested using labeled lists of activities from IMU and video data. A model is trained and tested for nine low-level activity classes with approximately 400 samples per class. The test result shows 88% accuracy for IMU-based skeleton data, 77% for RGB spatial video, and 63% for RGB video-based skeleton. The result has been verified using a previously published region-based activity recognition. The proposed approach can be extended to push the cognition capability of robots in human-centric workplaces.

Keywords

Human-robot collaboration; human activity recognition; deep-learning; InHard dataset

1. Introduction

Today human-robot collaboration (HRC) is becoming an essential part of the industry for achieving better quality products in less time. In this regard, robots' cognition capabilities are expected to be enriched with the prediction and detection of human activities [1]. Such an approach that helps recognize human actions and activities may enable robots to complement human motions and activities in shared smart workplaces [2,3]. Lists of human activities in HRC may require semantical descriptions and definitions, which can be described either at a higher level with generalized actions or at a lower level with detailed descriptions. Generalized actions refer to, e.g., reach, pick, put, turn, and assemble. In contrast, low-level actions describe details such as reaching with the left hand to object A, taking a type B screwdriver, and tightening the screw. Prediction and detection of human action and sequence of activities, either higher or lower, is still basic research requiring further investigation.

The classical approach in HRC task planning most often consists of pre-programmed logic that remains fixed for the given cycle operations. In a typical assembly, the robot executes the desired task in collaboration with the human operator in a programmed sequence. It hardly responds to changes in human performance which is one of the main barriers to HRC. Most recently, industries have been focusing on improving human and robot interaction in a shared workplace to complete a task efficiently and safely with flexibility in production processes. However, such a level of HRC is challenging as it involves many unpredictable events and actions, which are difficult for robots to understand and act accordingly. In this aspect, the robot must possess a cognitive capability and ability to understand various actions performed by the human operator to predict the subsequent possible action in order to carry out the task.

This work aims to present the activity recognition method using combined convolutional neural network (CNN) and long short-term memory (LSTM) techniques to evaluate the *InHard* dataset compared to a newly captured dataset in a lab environment.

2. Related Works

The current research that focuses on human activity recognition for HRC can be discussed from the aspects of types of human activities in industrial environments, methods employed for recognizing human actions, and generated datasets for human activity recognition (HAR).

2.1 Human activity types for HRC.

In today's industry 4.0 era, many researchers have brought humans and robots closer to the industrial environment. In [4], a detailed overview of HRC in industry 4.0 regarding various levels of humans and robots working together is given by solving safety issues using a particular collaborative robot (cobots). Furthermore, how the HRC can improve the efficiency of the industrial process by eliminating the uncomfortable, repetitive work of human operators is discussed. Similarly, [5] has conducted a survey to test the HRC process by measuring trust between humans and robots in an open workspace executing pick and place tasks. Work has to detail discussion about various safety factors and trust factors of HRC that can affect the productivity and efficiency of the process. Further, [6] has presented an HRC from a technical point of view. Various methods for human intention estimation through machine learning algorithms, robot action planning, and human-robot joint action planning are discussed and compared. A more detailed scenario of the industry is presented by [7], and it gives detailed information about the various industrial activities such as assembly activities, tool handling activities, and non-deterministic activities which are non-reparative, such as repairing activity or inspection, and also demonstrate that fusion of inertial measuring unit (IMU) sensors and video-based tracking system can be used to capture these activities with high precision. Similarly, [8] has also presented a work which includes modeling of industrial activities using a fusion of various motion capture sensors. It provides the detailed information of small industrial activities such as handling of nuts and bolts for assembly of the product and also to model various hand gestures movements to control the robot action. A visual sensor-based approach e.g., red-green-blue-depth (RGB+D) cameras have been also employed to capture various human activities in the industrial environments. Some of these activities include entering, leaving a work cell (movement), pointing to an object, waving (gesture), picking, and moving parts (object handling), applying pressure, reach to an object [8–10].

2.2 Methods for human activity recognition

Methods that have been employed for HAR can be considered into two big categories. The first is a statistical model, and the second is a deep learning-based model for activity prediction and detection.

Statistical models are known for their data-intensive requirement in order to generate high-quality motions [8]. Common approaches utilizing statistical models include Gaussian Mixture Models (GMM) or space partitioning (e.g., using k-means or principal component analysis (PCA) based linear mapping). GMM is a probabilistic model that maximizes expectation by fitting mixtures of Gaussian models to samples in high dimensional spaces. Deep learning-based models have been implemented for modeling human activities [11–13] based on video or skeleton. Open pose for two-dimensional pose estimation presented in [14] has employed a multi-stage CNN for extracting spatial features for human action recognition. Deep convolutional generative adversarial networks (GAN) have also been used to classify human activities even for fewer training datasets [15]. Further investigations for human motion generation or synthesis for enabling human interaction with smart machine systems that may involve higher-level human intention prediction and detection and lower-level details of actions have been shown in [16,17].

2.3 Human activity recognition datasets.

Dataset for HAR that publicly available includes HMDB51 [18], UCF50 [19], NTU RGB+D dataset [20], MSR-Action3D [21], and InHARD [10]. HMDB51 was introduced by [18], consisting of approximately seven thousand realistic video clips from sources such as movies and web series. The dataset consists of 51 classes of general day-to-day life activities such as jumping, laughing, kissing, and others, with 100 samples in each category. Another dataset in a similar category is UCF50 [19], which has offered 50 activity classes collected from online platforms like YouTube. The activities offered in the dataset include horse riding, pull-ups, diving, running, skipping, etc. Later, the activities are extended into 101 classes with the same human activity category, which is called UCF101. Both datasets offer only RGB data at a resolution of 320 x 240 with a fixed frame rate of 25 frames per second (fps).

The kinetics dataset introduced by [22] consists of a significant dataset for HAR with 700 activity classes with more than 700 video par classes. Each video is captured from YouTube videos lasting for ten seconds. Types of activities included in the data set are human-to-human or human-to-object interactions such as shaking hands, hugging, steering the car, and brushing the floor. NTU RGB+D dataset presented by [20] offers a diverse range of activity classes. Types of action are divided into three categories: eleven mutual activities like pushing, kicking, etc., nine health-related activities such as sneezing, staggering, etc., and 40 daily activities like drinking, reading, etc. It consists of approximately fifty-seven thousand samples in RGB + Depth and in skeleton format. In addition, MSR-Action3D presented by [21] has been a choice for skeleton-based activity recognition. Dataset offers 567 depth map sequences with 20 different hand gesture activity class-like horizontal arm waves, drawing a circle with an arm. Depth maps are captured using a depth camera sensor and are available in 640 x 240 resolution of recorded sequences

Though many datasets offer a diverse range of activity classes to facilitate HAR processes, most of them are related to daily life or health-related activities. From an industrial HAR point of view, there is a lack of a dataset that offers industrial activities, which is further addressed by [10] and presented *InHard* dataset (Industrial Human Activity Recognition Dataset). *InHard* demonstrates the actual industrial activity in an industrial environment, and the dataset is publicly released to facilitate the research progress in the field. The dataset provides various industrial assembly activities in the skeleton and RGB video format to facilitate HAR in the industrial environment. Moreover, it has not been well evaluated by the scientific community.

3. Methodology

A Panasonic 4K camcorder was employed in our experiment to obtain the video from the right side at 45 degrees. Similarly, the Xsens Awinda IMU system is used to capture the joint motions for comparison to the *InHard* (c. [10]) dataset. However, both datasets comprise different settings such as frame rate, skeleton

joint numbers, and motion capturing systems. Therefore, we must resample the time to make a consistent frame rate, re-arrange the data structure and retarget the skeleton before comparing.

The *InHard* dataset has offered an actual use case of industrial activities such as assembly of a part, picking components, measuring components, and representing actual industrial setup. The dataset comprises RGB video and IMU data for different participants (person) and in an adequate quantity. The participant's task in the *InHard* dataset is to assemble a component following instruction sets with the help of the UR10 robotic arm, using screws and hooks and a tool such as a screwdriver (c. [10,23]). The same activity with a different job (e.g., assembly of gear components) is proposed to reproduce *InHard* activities. The RGB video from the top view is used to analyze the spatial activities, while the RGB camera from the side helps to acquire (c. Figure 1)

(a) (b) (c) (d) (e)

Figure 1 Validation and comparison of the InHard and newly captured datasets in the lab environment for similar activities but different sensing systems; InHard dataset (a) IMU skeleton, (b) Open pose overlay, (c. [7]), New dataset (c) IMU skeleton, (d) Open pose overlay, (e) RGB spatial.

The methods for HAR based on video and skeleton dataset is described in three categories in subsection 3.1, and 3.2.

3.1 Human activity definition and dataset curation

Before the pre-processing task, it is necessary to explore the dataset to remove any unwanted data. Only an adequate and equal number of samples are provided for the deep learning method for training. Following is the list of nine low-level activity classes of both skeleton and RGB data with several samples in each class.

Table 1 Number of samples in each activities class.

Activity Classes	Assemble System	No Action	Picking Front	Picking Left	Put Down Component	Put Down Screwdriver	Take Component	Take Screwdriver	Turn Sheets
No. of samples	1378	500	456	641	385	461	485	420	224

It is essential to provide equal training data for each activity class to ensure proper learning of the network and the overall accuracy of the network. As deep learning algorithms require a large amount of data for good performance, we have only considered more than 200 samples classes. To facilitate HRC in the industry, the *InHard* dataset community has generalized assembly activities and presented activities that are used in many industrial assembly processes. Activities are divided into low-level activities, consisting of nine action classes (see Table 1), and high-level activities, which comprise 72 detailed action classes for more accurate activity detection.

3.2 HAR modeling based on sequential and temporal memory networks

A fusion of CNN and LSTM (see Figure 2) is proposed to take advantage of both networks as CNN handles spatial information and LSTM takes care of temporal data information. Video frames are given as input to CNN using a pre-trained network inception V3. The inception V3 model is one of the commonly used computer vision techniques for tasks such as object detection. It is pre-trained on the ImageNet dataset consisting of one thousand categories. Inception V3 architecture is built with symmetric and asymmetric blocks. The block includes a series of smaller convolutions, average pooling, and max pooling for faster training and processing of image data [24]. The last output layer of the CNN network is removed to obtain the feature vector. Then this feature vector becomes an input to the LSTM to learn temporal dependency and give the final classification.

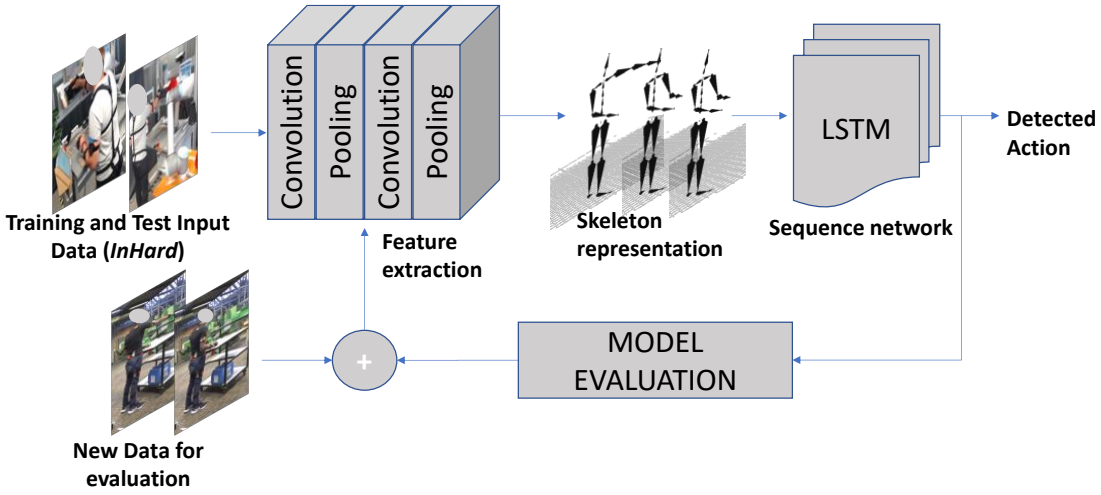


Figure 2 Fusion of CNN and LSTM architecture for action recognition and model evaluation using InHard and new dataset.

Table 2 System configuration for deep learning model

Feature	Description
Hardware configuration	All the deep learning model pre-processing and training are done using a high-performance computing (HPC) cluster, which has a specification of AMD 3.35 GHz CPU, NVIDIA Tesla V100 GPU, and 128 GB of RAM.
Data preparation	A model is trained for 14 low-level activity classes with approx. 400 samples per class. 77% of data is used for training, 3% for training validation and the remaining 20% of data is used to evaluate the trained model.
Optimizer	Adamax

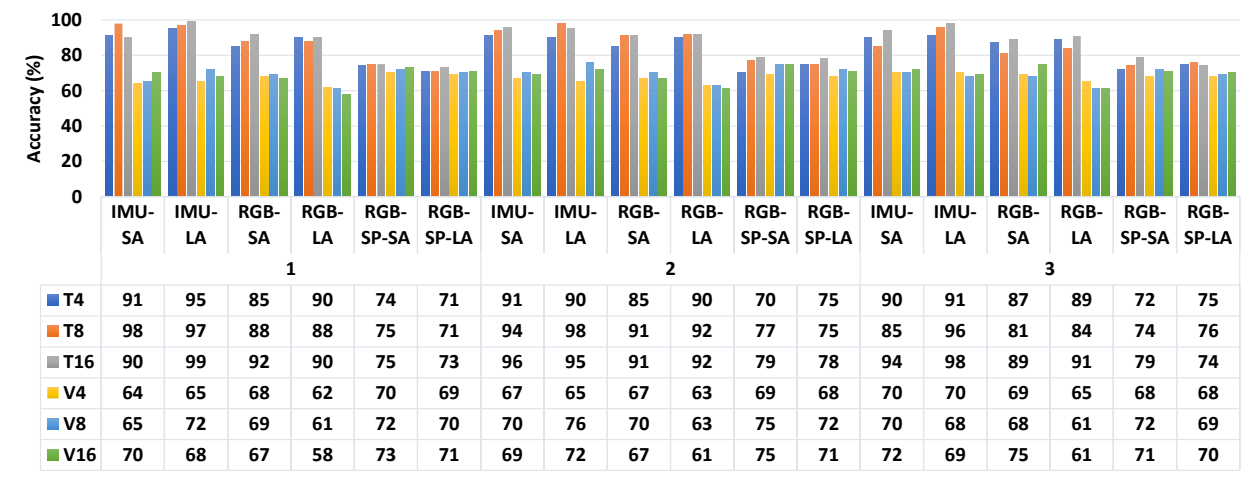
Performance of the *InHard* dataset is tested using a deep learning model: Long Short-Term Memory (LSTM) and Convolution Neural Network (CNN). The accuracy of the dataset is tested using segmented activity recognition using both Skeleton and RGB data. The system configuration details are shown in Table 2. The implementation comprises IMU-based joint data and RGB-based skeleton pose data to further predict and detect human activity in the same time domain. Pose detection and LSTM technique are applied to RGB video data to extract the skeleton pose from video using an open pose library [14]. LSTM processes the extracted skeleton pose data and classifies the activity. LSTM is used to process the temporal dependency of extracted features and classify the activity at the end. We have used a single LSTM model to train skeleton data (BVH files); for RGB data (.mp4 files) training, CNN and LSTM are fused in which CNN is used to extract the spatial feature (resolution) of video frames.

As we are dealing with activity recognition tasks that may include different lengths, we have employed an LSTM model as a final output model for all cases. It is necessary to note that LSTM takes temporal features as an input in which the time-stamp indicates the length of the activities. Some of the *InHard* dataset activities are of different lengths, and this required us to perform training of the model using two different lengths of activity for each method. Accordingly, the activity length is categorized as short and long length activity in which for short length activity, the first 30 seconds of data is considered, and for prolonged length activity, the first 60 seconds of data is considered for each case.

Before starting the training of deep learning models, it is necessary to clean and pre-process the training and test data. The cleaning and pre-processing of the *InHard* dataset include dividing the data frame into smaller batches, extracting skeleton data, and labeling its class. The training process is similar for both Skeleton (BVH) and video-skeleton cases. In both, hierarchical skeleton data is converted into vectors along with its class label. Besides, the data frame is divided into smaller batches depending on the length of the activity (short and long). Then these batches are used to train the LSTM model. The human activity recognition network code is publicly available [25]. The Tensorflow and Keras framework has been implemented for the training deep-learning model. Tensorflow is an open-source platform that provides various machine learning libraries, and Keras is a user-friendly high-level API that runs on top of Tensorflow [26].

4. Result

Based on systematically selected Tensorflow and Keras framework parameters, the early patience is to 4, and the learning rate is set to 0.02. The loss function is set to *categorical cross-entropy* as it is the default choice for classification.



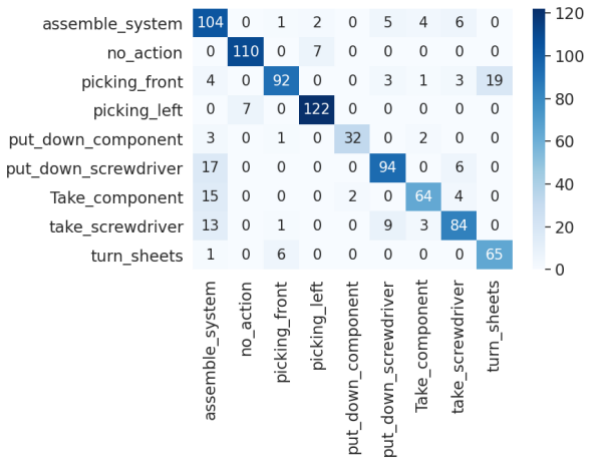
	1						2						3					
	IMU-SA	IMU-LA	RGB-SA	RGB-LA	RGB-SP-SA	RGB-SP-LA	IMU-SA	IMU-LA	RGB-SA	RGB-LA	RGB-SP-SA	RGB-SP-LA	IMU-SA	IMU-LA	RGB-SA	RGB-LA	RGB-SP-SA	RGB-SP-LA
T4	91	95	85	90	74	71	91	90	85	90	70	75	90	91	87	89	72	75
T8	98	97	88	88	75	71	94	98	91	92	77	75	85	96	81	84	74	76
T16	90	99	92	90	75	73	96	95	91	92	79	78	94	98	89	91	79	74
V4	64	65	68	62	70	69	67	65	67	63	69	68	70	70	69	65	68	68
V8	65	72	69	61	72	70	70	76	70	63	75	72	70	68	68	61	72	69
V16	70	68	67	58	73	71	69	72	67	61	75	71	72	69	75	61	71	70

Figure 3 Comparison of model parameters for short activity (SA) or long activity (LA) for training (T) and validation (V) phase.

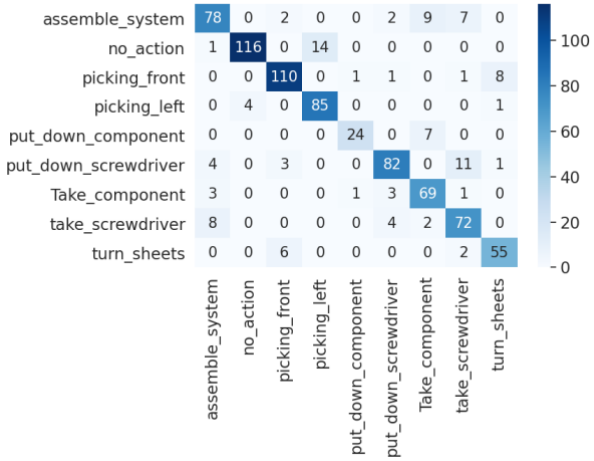
For tuning the models, we have altered different parameters such as optimizer, size of input batches, and the number of hidden layers to obtain optimal parameters. The tuning parameters are epoch: 100, optimizer: 'Adam' and 'Adamax', number of hidden layers: 1, 2, and 3, number of batches: 4, 8, and 16, and the same training process is repeated for two activities length data, i.e., short and long (see Figure 3).

Training result is presented for skeleton (BVH) and video-skeleton data. For skeleton (BVH) and video-skeleton data, the LSTM model is used and trained using a different model structure and parameters for short and long activity length types. The same training procedure is applied for RGB video data on CNN and LSTM network fusion. Training and validation accuracy for each model structure, parameters, and activity length is compared in Figure 3. The result shows the percentage of accuracy for RGB-based skeleton short activity and prolonged activity (RGB-SA and RGB-LA), RGB-based spatial video for short and long activity

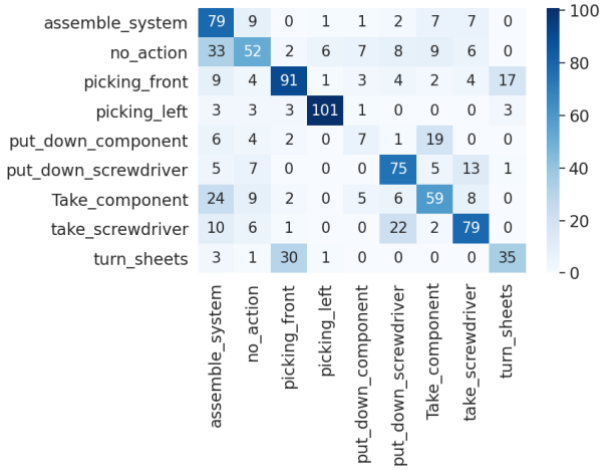
(RGB-SP-SA and RGB-SP-LA), and IMU-based skeleton for short and long activity (IMU-SA and IMU-LA) data for different numbers of hidden layers and batch sizes (4, 8, and 16).



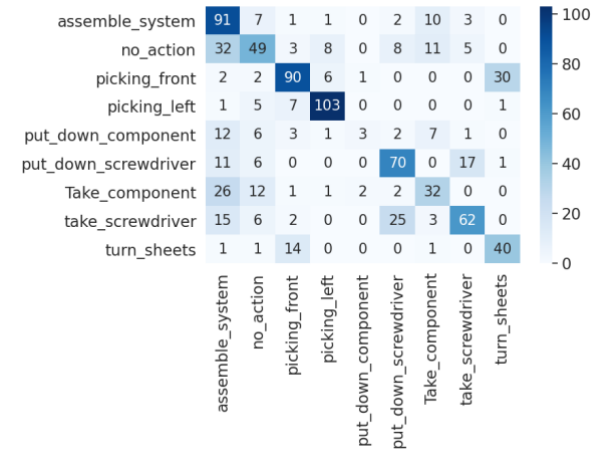
(a) Short activity from IMU skeleton (IMU-SA)



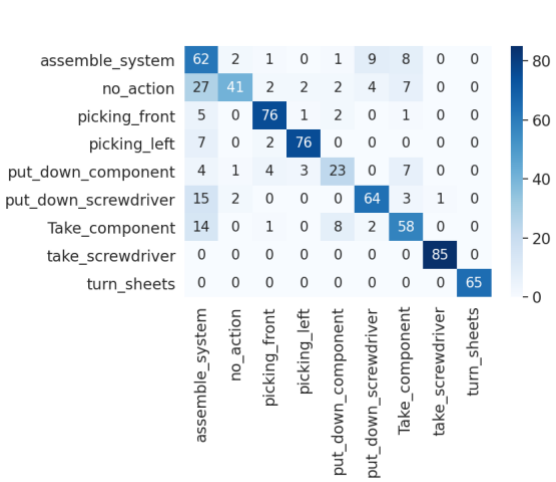
(b) Long activity from IMU skeleton (IMU-LA)



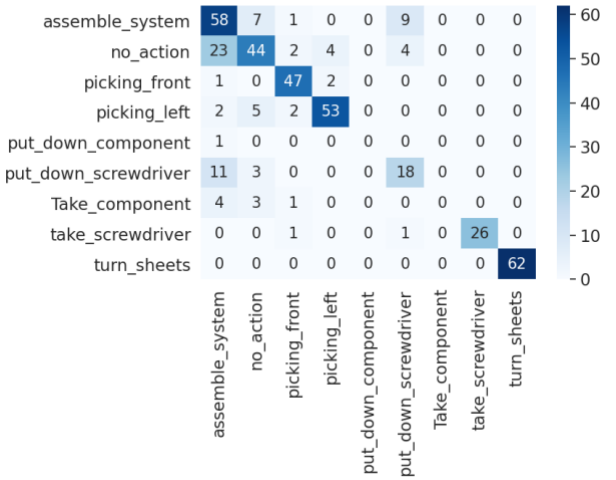
(c) Short activity from RGB skeleton (RGB-SA)



(d) Long activity from RGB skeleton (RGB-LA)



(e) Short activity from RGB spatial (RGB-SP-SA)



(f) Long activity from RGB spatial (RGB-SP-LA)

Figure 4 Confusion matrix representing the comparison of the predicted and detected actions.

The quantitative results based on Figure 3 show that short-activity recognition based on *RGB video skeleton* has achieved model training accuracy of 89% and validation accuracy of 75%. The training accuracy remains 89% for the long-activity, while the validation accuracy drops to 65%. Using the model for evaluation on a test dataset yields 62% and 63% accuracy for short and long activity, respectively. Considering the *IMU skeleton,* the highest training accuracy and validation accuracy of short-activity are 94% and 72%, respectively. The long-activity is 98% for training accuracy and 76% for validation. Using the training weights of the model with the highest training accuracy and validation accuracy, the evaluation with test data yields evaluation accuracy of 85% for short-length activity and 88% for long-activity. For RGB video, a fusion of CNN and LSTM achieved 81% and 78% training and validation accuracy for short activity length and 80% and 72% training and validation accuracy for long activity length. Fusion of model has shown a 77% and 74% evaluation accuracy on test data.

A confusion matrix, which has a size of *n x n*, where *n* is the number of activities, is used to evaluate how accurately the model can classify the activities (c.[27]). The matrix compares the actual activities with the predicted activities (see Figure 4).

5. Discussion

To facilitate HRC, we have tested the *InHARD* dataset for industrial activity recognition using deep learning techniques for two modes of data: IMU skeleton data, and video-skeleton data, among which skeleton activity recognition has shown 88% evaluation accuracy, RGB video model gave 77%, and video-skeleton data is with 63% of evaluation accuracy on *InHard* dataset.

The model's prediction is better when activity length is long because it has acquired more data frame (i.e., the batch size consists of 60 seconds of activity) than in short-length activity where the duration is 30 seconds. Skeleton data provides detailed information about each human pose, more training data, and improved accuracy. Thus, it has been possible to distinguish the classification of activities having slight differences, such as putting down the screwdriver and picking up the screwdriver accurately.

The video-skeleton method using open pose techniques shows poor results compared to the IMU methods. It classifies the activities; however, it gets confused between similar activities such as *Take the screwdriver* and *Put Down Screwdriver*. As the open pose technique highly depends on the person's view for detecting key points on the body for mapping to the skeleton, considering a camera position in a proper orientation with minimum occlusion possibility is crucial for obtaining better results. With the implemented open pose technique, only the required region of interest is considered from the RGB video to avoid noise that may affect the model accuracy. The comparison results with different datasets captured in the lab environment shows less than fifty percent success ratio from the captured twenty operations, while the IMU skeleton accuracy is 67%. On the other hand, the RGB video model (CNN+LSTM) has shown 60% accurate detection for some activities captured in the lab with Panasonic 4K camera. The accuracy evaluation of the proposed methods is still below the accuracy of the human activity recognition that has been published in [17], which employs region-based joint configuration. Reproducible workplace setup does not necessarily yield the same output for reasons such as motion capturing systems, body size variation, and implementation complexities.

Overall results show that human activity recognition for industrial setup is still challenging to detect activities when robots are considered in the loop accurately. Due to the skeleton, body size, capturing system, and model parameters, repeated activities performed in different workplace settings are not straightforward to reproduce. Open source datasets such as *InHard* are helpful to investigate optimal settings for motion capturing and modeling, allowing to exploit the opportunities and identify the inherent challenges regarding activity prediction and detection techniques. However, more datasets must be employed before generalizing

human activities and actions detections. This may facilitate the path toward sustainable human and robot collaboration.

6. Conclusion and future outlooks

Human activity recognition in the cognitive production system may change the way humans and machines interact and cooperate for completing tasks. Gathering sufficient data that helps to extensively evaluate the performance and limitations of existing methods is still challenging. The main reasons discussed are model accuracy, data validity, and activity duration. Employing multi-systems for human motion data acquisition such as IMU and Optical cameras, methods such as CNN+LSTM approaches are evaluated for their accuracy. The overall result shows open research questions regarding motion capturing methods, feature mapping, and labeling. Nevertheless, the proposed approach has the potential to improve the way robots learn human motion behavior as co-partners. Future works will address real-time activity recognition with an extended cognitive capability in human-centric workplaces.

Acknowledgments

The authors acknowledge the European Regional Development Fund (EFRE) within the project SMAPS (grant number: 0200545) for the financial support. Similarly, the authors would like to acknowledge the ZIMT department for *Scientific Computing* services at the Universität Siegen.

References

[1] Li S, Wang R, Zheng P, Wang L. Towards proactive human–robot collaboration: A foreseeable cognitive manufacturing paradigm. J Manuf Syst 2021;60:547–52. https://doi.org/10.1016/j.jmsy.2021.07.017.

[2] Pichler A, Akkaladevi SC, Ikeda M, Hofmann M, Plasch M, Wögerer C, et al. Towards Shared Autonomy for Robotic Tasks in Manufacturing. Procedia Manufacturing 2017;11:72–82. https://doi.org/10.1016/j.promfg.2017.07.139.

[3] Ji Z, Liu Q, Xu W, Liu Z, Yao B, Xiong B, et al. Towards Shared Autonomy Framework for Human-Aware Motion Planning in Industrial Human-Robot Collaboration. 2020 IEEE 16th International Conference on Automation Science and Engineering (CASE), 2020, p. 411–7. https://doi.org/10.1109/CASE48305.2020.9217003.

[4] Vysocky A, Novak P. Human - Robot collaboration in industry. MM Science Journal 2016;2016:903–6. https://doi.org/10.17973/MMSJ.2016_06_201611.

[5] Kumar S, Savur C, Sahin F. Survey of Human–Robot Collaboration in Industrial Settings: Awareness, Intelligence, and Compliance. IEEE Transactions on Systems, Man, and Cybernetics: Systems 2021;51:280–97. https://doi.org/10.1109/TSMC.2020.3041231.

[6] Bauer A, Wollherr D, Buss M. Human–robot collaboration: a survey. Int J Human Robot 2008;05:47–66. https://doi.org/10.1142/S0219843608001303.

[7] Hartmann B. Human worker activity recognition in industrial environments. KIT Scientific Publishing; 2011. https://doi.org/10.5445/KSP/1000022235.

[8] Roitberg A, Somani N, Perzylo A, Rickert M, Knoll A. Multimodal Human Activity Recognition for Industrial Manufacturing Processes in Robotic Workcells. Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle Washington USA: ACM; 2015, p. 259–66. https://doi.org/10.1145/2818346.2820738.

[9] Voronin V, Zhdanova M, Semenishchev E, Zelenskii A, Cen Y, Agaian S. Action recognition for the robotics and manufacturing automation using 3-D binary micro-block difference. Int J Adv Manuf Technol 2021. https://doi.org/10.1007/s00170-021-07613-2.

[10] DALLEL M, HAVARD V, BAUDRY D, SAVATIER X. InHARD - Industrial Human Action Recognition Dataset in the Context of Industrial Collaborative Robotics. 2020 IEEE International Conference on Human-Machine Systems (ICHMS), 2020, p. 1–6. https://doi.org/10.1109/ICHMS49158.2020.9209531.

[11] Cho NJ, Lee SH, Suh IH. Modeling and evaluating Gaussian mixture model based on motion granularity. Intel Serv Robotics 2016;9:123–39. https://doi.org/10.1007/s11370-015-0190-1.

[12] Ji S, Xu W, Yang M, Yu K. 3D Convolutional Neural Networks for Human Action Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 2013;35:221–31. https://doi.org/10.1109/TPAMI.2012.59.

[13] Wang P, Liu H, Wang L, Gao RX. Deep learning-based human motion recognition for predictive context-aware human-robot collaboration. CIRP Annals 2018;67:17–20. https://doi.org/10.1016/j.cirp.2018.04.066.

[14] Cao Z, Hidalgo G, Simon T, Wei S-E, Sheikh Y. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. IEEE Transactions on Pattern Analysis and Machine Intelligence 2021;43:172–86. https://doi.org/10.1109/TPAMI.2019.2929257.

[15] Shi X, Li Y, Zhou F, Liu L. Human Activity Recognition Based on Deep Learning Method. 2018 International Conference on Radar (RADAR), 2018, p. 1–5. https://doi.org/10.1109/RADAR.2018.8557335.

[16] Liu H, Qu D, Xu F, Zou F, Song J, Jia K. A Human-Robot Collaboration Framework Based on Human Motion Prediction and Task Model in Virtual Environment. 2019 IEEE 9th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER), 2019, p. 1044–9. https://doi.org/10.1109/CYBER46603.2019.9066603.

[17] Manns M, Tuli TB, Schreiber F. Identifying human intention during assembly operations using wearable motion capturing systems including eye focus. Procedia CIRP 2021;104:924–9. https://doi.org/10.1016/j.procir.2021.11.155.

[18] Kuehne H, Jhuang H, Garrote E, Poggio T, Serre T. HMDB: A large video database for human motion recognition. 2011 International Conference on Computer Vision, 2011, p. 2556–63. https://doi.org/10.1109/ICCV.2011.6126543.

[19] Reddy KK, Shah M. Recognizing 50 human action categories of web videos. Machine Vision and Applications 2013;24:971–81. https://doi.org/10.1007/s00138-012-0450-4.

[20] Shahroudy A, Liu J, Ng T-T, Wang G. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA: IEEE; 2016, p. 1010–9. https://doi.org/10.1109/CVPR.2016.115.

[21] Li W, Zhang Z, Liu Z. Action recognition based on a bag of 3D points. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, San Francisco, CA, USA: IEEE; 2010, p. 9–14. https://doi.org/10.1109/CVPRW.2010.5543273.

[22] Smaira L, Carreira J, Noland E, Clancy E, Wu A, Zisserman A. A Short Note on the Kinetics-700-2020 Human Action Dataset. ArXiv:201010864 [Cs] 2020.

[23] DALLEL M, HAVARD V, BAUDRY D, SAVATIER X. InHARD - Industrial Human Action Recognition Dataset in the Context of Industrial Collaborative Robotics 2020. https://doi.org/10.5281/zenodo.4003541.

[24] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, p. 2818–26. https://doi.org/10.1109/CVPR.2016.308.

[25] Tuli TB, Patel VM, Manns M. HARNets: Human Activity Recognition Networks Based on Python Programming Language. 2022. https://doi.org/10.5281/zenodo.6366665.

[26] Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow, Large-scale machine learning on heterogeneous systems. 2015. https://doi.org/10.5281/zenodo.4724125.

[27] Powers DMW. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation, 2008.

Biography

Tadele Belay Tuli is a researcher and Ph.D. candidate at the University of Siegen, Germany. He received his M.Sc. in Mechatronics Engineering, robotics curriculum from the University of Trento, in Italy. His research interest covers human motion behavior modeling for human-robot collaboration, human activity recognition, and path planning for 3D printing.

Valay Mukesh Patel is an M.Sc. student at the University of Siegen, Germany, in Mechatronics Engineering. His research interest is automation and human activity recognition.

Martin Manns has been head of the Chair of Production Automation and Assembly (FAMS) at the University of Siegen since 2016. Before his appointment, Univ.-Prof. Dr.-Ing. Martin Manns has worked in production research at Daimler AG (2009-2016) and Henkel KGaA (2007-2009) and as a post doctorate fellow at the University of Winsor, Ontario, Canada (2006-2007).