2023

Action Classification in Human Robot Interaction Cells in Manufacturing

Shakra S.M. Mehak
Technological University Dublin, Ireland, d21127063@mytudublin.ie

Maria Chiara Leva
Technological University Dublin, Ireland, mariachiara.leva@tudublin.ie

John Kelleher
Technological University Dublin, Ireland, john.kelleher@tudublin.ie

See next page for additional authors

Funder: CISC project, funded from the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie grant agreement no. 955901.

Action Classification in Human Robot Interaction Cells in Manufacturing

Moving Towards Mutual Performance Monitoring Capacity

Shakra, S.M, Mehak
Technological University Dublin & Pilz Ireland, Cork, Ireland
s.mehak@pilz.ie

Maria Chiara Leva
Technological University Dublin, Dublin, Ireland
mariachiara.leva@tudublin.ie

John D. Kelleher
Technological University Dublin, Dublin, Ireland,
john.d.kelleher @tudublin.ie

Michael Guilfoyle
Pilz Ireland, Cork, Ireland
M.Guilfoyle@pilz.ie

ABSTRACT

Action recognition has become a prerequisite approach to fluent Human-Robot Interaction (HRI) due to a high degree of movement flexibility. With the improvements in machine learning algorithms, robots are gradually transitioning into more human-populated areas. However, HRI systems demand the need for robots to possess enough cognition. The action recognition algorithms require massive training datasets, structural information of objects in the environment, and less expensive models in terms of computational complexity. In addition, many such algorithms are trained on datasets derived from daily activities. The algorithms trained on non-industrial datasets may have an unfavorable impact on implementing models and validating actions in an industrial context. This study proposed a lightweight deep learning model for classifying low-level actions in an assembly setting. The model is based on optical flow feature elicitation and mobilenetV2-SSD action classification and is trained and assessed on an actual industrial activities' dataset. The experimental outcomes show that the presented method is futuristic and does not require extensive preprocessing; therefore, it can be promising in terms of the feasibility of action recognition for mutual performance monitoring in real-world HRI applications. The test result shows 80% accuracy for low-level RGB action classes. The study's primary objective is to generate experimental results that may be used as a reference for future HRI algorithms based on the InHard dataset.

CCS CONCEPTS

• **Action Classification**; • **Human Robot Interaction**; • **Deep Learning**;

KEYWORDS

Fluent HRI, Machine Learning Algorithms

ICMLT 2023, March 10–12, 2023, Stockholm, Sweden

ACM ISBN 978-1-4503-9832-9/23/03…$15.00
https://doi.org/10.1145/3589883.3589916

ACM Reference Format:
Shakra, S.M, Mehak, Maria Chiara Leva, John D. Kelleher, and Michael Guilfoyle. 2023. Action Classification in Human Robot Interaction Cells in Manufacturing: Moving Towards Mutual Performance Monitoring Capacity. In *2023 8th International Conference on Machine Learning Technologies (ICMLT 2023), March 10–12, 2023, Stockholm, Sweden.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3589883.3589916

1 INTRODUCTION

Collaborative robots require a set of functionalities that enable them to behave in close proximity to operators; among the future required is mutual performance monitoring between the human and the intelligent agent to realize some teamwork conditions. It is one of the prerequisites of dynamic robot decision-making and implies the feasibility of Human-Action Recognition [1]. Considering the ultimate objective of a natural, human-like, and flawless interaction, we acknowledge that recognizing human actions may enable robots to complement humans more effectively in a teaming structure. Furthermore, human actions in a Human Robotic Collaborative (HRC) context can be classified as higher level with generic behaviors or lower level with detailed descriptions. Low-level operations include: Grabbing item X with the left hand, seizing a screwdriver of type Y, or constricting the screw [2]. Anticipation of human intentions and sequence of actions, regardless of general or specific, is nevertheless a fundamental research topic that encourages more exploration. During manufacturing, the industrial operator's actions indicate operational intent, for which the analysis approach incorporates offline recognition and online anticipation in actual implementations. However, this form of HRI is challenging because it involves a range of unforeseen events and behaviors that are difficult for robots to interpret and respond to appropriately. Therefore, precise action prediction is imperative for proactive HRI, enabling robots to extrapolate human intentions and provide adaptive or proactive assistance [3].

Existing deep learning methods may predict human behaviors and human–object dynamics based on RGB, depth, and skeletal data, or they can fuse multi-data models [4] [5]. Throughout most proposed solutions, researchers relied on conventional CNN [6] ,and RNN [7] models, which required a substantial quantity of training data. Although various datasets, such as MSR-Action3D [8], UCF50 [9], NTU RGB+D [10] , HMDB51 [11] , provide a variety of data classes to enable HRI algorithms, the majority of them are

specific to everyday life or health-care activities [2]. Nearly all of the offered algorithms are trained using datasets from [8-10, 11]. There is a lack of datasets and methodologies that offer genuine industrial actions from the perspective of industrial activities. This work aims to present an action classification framework trained on low-level action frames for cognitive HRC applications.

The major contribution of our study is as follows:

- We investigated the industrial directions of fluent HRI applications as human action or intention prediction, which is complex and preliminary in literature.
- We proposed an action recognition deep learning model for fluent HRI applications for dynamic HRC scenarios. The proposed method addresses two challenges for the HRI action recognition algorithm, the structural feature information and computational less expensive classification models i.e., optical flow (OpF) combined with MobilnetV2-SSD.
- The experimental outcomes demonstrate the significance of the proposed system using industrial activity dataset. However, the dataset is new for the scientific community and can provide a direction for designing HRI algorithms.

The remaining sections are structured as follows. Section II covers the existing research relevant to our subject area. Section III discusses the proposed methodology. In section IV, the implementation parameters and experimental outcomes are described. Finally, section V concludes the study and directs to future work.

2 RELATED WORKS

Human and industrial robots have strengths and behavior within the industrial manufacturing process. They complement each other as a team member, leading to the development of HRI applications. As a co-worker, the robot should quickly determine human participants' intentions and future positions in the HRC environment [12]. Many studies have been proposed in past on human motion [13] estimation and action recognition [14]. Within the machine-assisted human interaction approaches, the types of collaborative intelligence tasks more frequently addressed by the retrieved collection of research works are human intention prediction and human motion recognition, specifically for human-robot collaboration applications. These two tasks have been addressed separately in some works but are connected and commonly tackled together. Deep learning approaches have become popular to ensure efficient recognition of human activities in HRC applications. To our knowledge, four studies were found addressing human motion recognition for close human-robot collaboration, with two using CNN [15] [16]. Whereas [17] have used ANNs for the classification of human gestures for robot instructions, moreover [18] used discrete Hidden Markov Model (HMM) for human gesture recognition. For human motion intention prediction [19] [20], applied HMM. The author in [21] proposed RNN-based prediction of motor intention projection of upper-limb action in an HRC environment. For identifying human behaviors, inconsistent SURF matches are eliminated by adding a human detector. A supervised learning method based on a three-dimensional convolutional neural network (CNN) has been developed to infer action representations from the movements of operators during the assembly of an optical controller [22]. Meanwhile, [23] introduced a deep wide network for 3D-feature extractor that connects robotic correctness control and safety requirements for collision avoidance. The methods, as mentioned earlier, are good at collecting fine-grained details but inflexible when associating visual patterns of the same activity from multiple perspectives, not to mention the high computing cost for long-duration video clips. In contrast, skeleton-based action recognition is computationally effective but lacks low-level details. Another problem with traditionally used deep learning models is that they require substantial training datasets. A study [24], addressed the aforementioned problem and proposed a deep convolutional Generative adversarial network (GAN) for the situation where the amount of training data available is insufficient. However, GAN model does not have an intrinsic evaluation metric for better training.

In order to achieve high kinematics, statistical models are also renowned for their data-intensive requirements [25]. For instance, Gaussian Mixture Models (GMM) and space partitioning are standard techniques that incorporate statistical models. These techniques optimize probability by fitting gaussian model ensembles to data in high-dimensional domains. In this sense, [26] introduced a multi-model communication framework for HRI. They used Bayesian inference for intention prediction through motions. Similarly, [27] presented the probabilistic dynamic movement primitive (PDMP) for determining human intent and estimating human hand movements during an online operation. The challenge with PDMP is that the framework is built on one-shot learning; therefore, it requires a unique set of features for the training dataset, and feature engineering should be performed accurately. Although several machine learning and statistical models have shown significant improvement, these models must be trained on industrial action datasets for HRI applications to be practical. Visual sensors have been successfully used to record diverse human actions in manufacturing setup [25]. The authors in [25] captured a range of small industrial activities, including entering and exiting a work cell, indicating at a target, exerting pressure, and grabbing for an object to control robot actions. Different industrial operations, such as assembly tasks, tool handling, and more dynamical and less prescriptive activities, such as maintenance or inspection interventions, can be monitored with high precision using a video-based tracking system. Regarding industrial human actions, there is an absence of industrial activity datasets. This issue has been partially addressed by [28], which offers the industrial human activity recognition dataset. The dataset comprises a range of industrial assembly activities in multiple formats, i.e., RGB and skeleton. However, since it is relatively recent, its merit must be thoroughly discussed within the scientific community.

The motivation of this work is two-fold: first, to introduce a future perspective for installing cognitive abilities in fluent HRI applications for an industrial setting, and second, to explore datasets [28], that can serve as training data for HRI algorithms. The presented work is the preliminary step to our project, which intends to introduce a mutual performance monitoring capacity of human-robot collaborative applications.

3 MATERIAL AND METHODS

Complex deep neural networks for action/object recognition can attain high accuracy, but they require significant calculation and

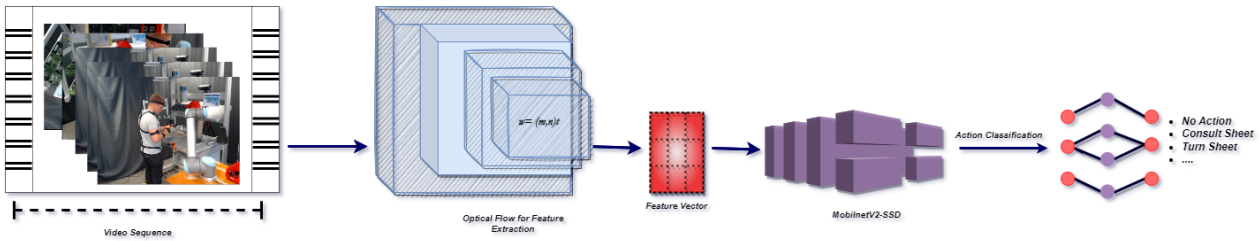


Figure 1: The proposed architectural design for the action classification approach

configuration parameters within the model. These models can be unsuitable for embedded devices and responsive applications. Based on [29, 30], we propose MobilnetV2-SSD combined with optical flow (*OpF*)feature elicitation, as illustrated in Figure-01.

We use *OpF* method for feature elicitation from three separate frames i.e., left, right and top which are fused into feature vector. We annotated the key action frames using machine learning libraries. Further, we transform the RGB frames into feature vectors, including spatial and temporal information for three distinct orientations (left, right, and top views). The feature vector is then passed to MobilnetV2 with SSD (Single Shot MultiBox Detector) networks for action classification. We implemented the depth-wise separable convolution method of calculus convolution to minimize model size and computation. Each input channel is separately mapped to its corresponding output channel by the depth-wise separable convolution. The proposed model is applicable for embedded architectures and controller applications such as robots since mobilnetV2 is lightweight and requires less time to execute [31]. To make our study self-contained, we first briefly overview optical flow for feature elicitation and mobilnetV2-SSD. Subsequently, model implementation and conducted experiments have been discussed in next section.

3.1 Spatio-Temporal Optical Flow based Feature Extraction

Human activities are represented as a series of frames in which spatial and temporal information changes over time while performing an operation. The frame sequence can be in any format, such as RGB, depth, skeleton, etc. The proposed method considered RGB videos comprised of a series of moving frames where human participant is performing a specific activity. We adopted OpF from [32] for feature elicitation in the proposed method. OpF is effective for feature recognition due to its robustness to appearance, even when temporal consistency is insufficient.

The relevance of flow precision at edges and for minor displacements is relatively significant for action recognition. In OpF, observing the instantaneous velocity of the pixel motion on frames can reveal an object's structure and motion relationship in a video. Assuming, however, that time is continuous, and the spectrum of motion is not violent, then dt time yields the following equations [33].

$$I(x, y, t) = I(x + dx, y + dy, t + dt) \tag{1}$$

$$I(x, y, t) = I(x, y, t) + \frac{\partial I}{\partial x}dx + \frac{\partial I}{\partial y}dy + \frac{\partial I}{\partial t}dt \tag{2}$$

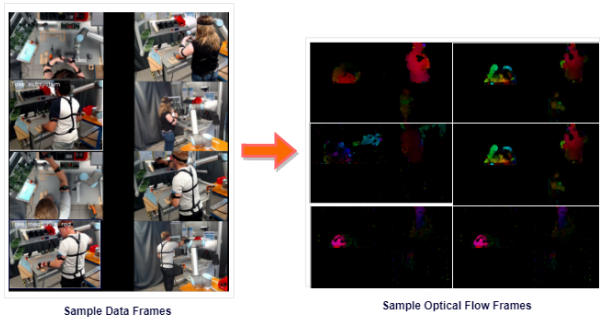


Figure 2: The Illustration of optical flow feature frames

$$I(x, y, t) + \frac{\partial I}{\partial x}dx + \frac{\partial I}{\partial y}dy + \frac{\partial I}{\partial t}dt = 0 \tag{3}$$

(2) and (3) is the form (1) obtains when Taylor Expansion is applied. We obtain following by dividing both sides of the equation by dt:

$$I_x m + I_y n + I_t = 0 \tag{4}$$

$$I_x = \frac{I}{x}\ ;\ I_y = \frac{I}{y} \tag{5}$$

$$m = \frac{dx}{dt}\ ;\ \ n = \frac{dy}{dt} \tag{6}$$

Where m and n are the horizontal and vertical optical flow field components, I_x and I_y indicate the shift in light intensity between adjacent pixels in the m and n orientations of the same frame. It reflects the temporal shift of the same pixel at an adjacent interval. Consequently, $w = (m, n)t$ forms the optical flow at that interval. In our method, we compute the feature vector for each pixel in each frame using dense optical flow. We obtain a two-channel array containing optical flow vectors(m, n) that determine magnitude and direction, (m) and (n)correspond to the saturation and plane value respectively. Figure 2 shows the color-coded version of optical frames for enhanced visibility.

3.2 Overview of MobilnetV2-SSD

MobileNetV2 is an efficient model that uses depth-separable convolutions to build lightweight deep convolutional neural networks and facilitates a framework for mobile and controller-based vision applications [34]. The depth-wise separable convolutions are more resource-efficient alternative to other neural network models, where required resources increase proportionally to their size. It

divides the traditional convolution layer into depth-based convolution and point-based pooling layers. The depth-based convolutional component performs the filtering, whereas the point-based layer corresponds to channel-specific feature mappings. Prior to combining the outcome with 1×1 convolution, depth-wise convolution splits the input channels and filters into multiple streams. The network input is $DF \times DF \times \mathrm{M}$, where the kernel size is DK, and the network comprises n output channels. The cost of depth convolution is computed as [29]:

$$C = DF \times DF \times \mathrm{DK} \times \mathrm{M} + \mathrm{M} \times \mathrm{N} \times \mathrm{DF} \times \mathrm{DF} \tag{7}$$

The average weight vector on an input filter is computed by performing computations on each filter. Depth convolution maps exclusively on each channel. Therefore, there is an equal distribution of input and output channels. Two main mobilnetV2 architecture elements are linear bottlenecks between layers and fast interconnections in bottlenecks [35]. After depth-wise separable convolution, the ReLU function applied to low-dimensional features may lead to data loss. In order to prevent data loss, the ReLU is replaced by a linear function within the linear bottleneck of the convolutional block. Like the residual blocks, the bottleneck blocks consist of an input followed by many bottlenecks and expansions. The bottlenecks include all pertinent information, but the expansion layer provides an execution layer that enables a nonlinear tensor transition. Subsequently, SSD adds extra bits to predict the action and creates an action-occurrence probability vector [31]. SSD is a fast model for classification using a single deep neural network and can execute multi-target detection by concurrently predicting target class and label. The SSD model employs a forward-feeding convolution network where VGG16 [34], serves as the network's backbone, followed by six additional feature layers. Using six separate feature layers, the feature map size is modified to distinguish activities of varying scales. Then multi-scale discretized edges are developed on respective data layers to predict the offset of the default frame at various scales and aspect ratios, along with the confidence associated with it. The actual SSD's convolution layer can be replaced with the MobileNet's depth-wise separable layer to solve the problem of excessive parameter size and model execution efficiency.

4 EXPERIMENTS

This section details the experiments conducted and the performance of the proposed method for recognizing the operator's actions in the assembly environment.

4.1 Dataset and Experimental Setup

We assessed the proposed algorithm on the InHard dataset [28], comprising RGB multi-view videos (Left, right & top views). The RGB videos incorporate annotations with multiple layers depending on a user-defined coding scheme. The dataset contains spatio-temporal information, including inter-level connections, time tracks, and non-temporal objects. With almost 2 million frames, the dataset contains fourteen low-level action classes and seventy-two high-level action classes for classification problems. The low-level action classes of RGB data vary across multiple samples. Providing equal training data for each action while training a model is important to ensure proper training and overall network accuracy. We have considered 220 samples for each class at this point. Since the duration of every action is variable, we observed 60 seconds of information. The optical flow features are extracted in our experiments using the OpenCV tools. The Keras framework is employed in the deep learning step to develop the neural network structure with a model size of 14MB and 3.4 million parameters. The experiment uses the NVIDIA CUDA framework 10.1 and the cuDNN 8.0.3 library.

We selected categorical cross-entropy as the loss function, the standard classification parameter, and set the learning rate to 0.02 to facilitate our experiments. The hardware setup employed for the experiments contains an Intel Core i9-9900X processor, 64 GB of RAM, and an NVIDIA Geforce RTX 2080TI * 2 graphic card.

4.2 Performance Metrices

To estimate the performance of the proposed system, we compute the overall accuracy, which is defined as:

$$accuracy = \frac{accurately\ identified\ action\ instances}{Total\ number\ of\ action\ instances}$$

Furthermore, we use F1-score for a comprehensive evaluation of the classifier's performance which is defined:

$$F1 - Score = 2\frac{Precision \times Recall}{Precision + Recal}$$

F1 score is generally used to determine the weighted average of precision and recall. The F1 score demonstrates the model's performance is satisfactory even when competing in imbalanced classes [35].

4.3 Results Analysis

This section provides a concise summary of the findings from our experiments. For model evaluation, we conducted two experiments. At the beginning of our research, we employed all of the low-level RGB action classes that the Inhard dataset provided. We found that a few classes negatively impact the overall model performance and have a low total number of key feature points. The challenge we confront with these data classes is that the spatial location of a point varies abruptly. With increasing spatial distance, OpF algorithm often underperform for this problem, which is highly prevalent when identical feature points exist. The dataset for these classes must be reassessed to enhance the features or eliminate the noisy data points, which is part of another study. The results of our experiments are presented in Tables 01 and 02 accordingly. In our second experiment, we identified feature-rich classes for which we do not need to perform excessive feature engineering. *{Assembly system, picking left, Put down screwdriver, Put down subsystem, Take screwdriver, Take subsystem}* are used in our second experiments. The proposed action classification algorithm attained an accuracy of 80% during validation and 89% during training, as shown in figure 03, the. However, the model confuses identical actions, such as Take the screwdriver and Take subsystem.

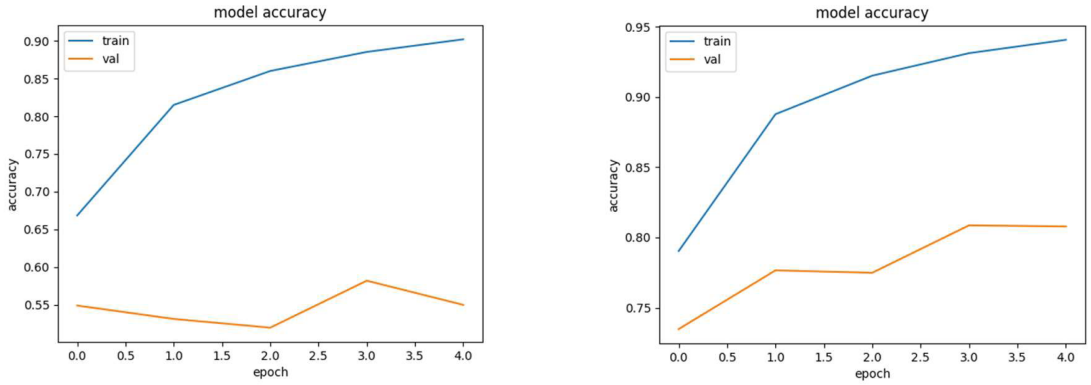
The validation results of the proposed method demonstrate advancements in the operator's action recognition accuracy reported in [4], which implements a conventional convolutional neural network (CNN) combined with long short-term memory (LSTM)

Table 1: Experimental results with all action classes on the Inhard Dataset

Meta-Action Class labels	Precision	Recall	F1- Score
No Action	0.60	0.76	0.67
Consult sheets	0.14	0.63	0.23
Turn sheets	0.39	0.39	0.39
Take screwdriver	0.55	0.73	0.63
Put down screwdriver	0.84	0.78	0.81
Picking in front	0.61	0.12	0.20
Picking left	0.71	0.46	0.56
Take measuring rod	0.69	0.52	0.59
Put down measuring rod	0.95	0.75	0.84
Take component	0.61	0.16	0.26
Put down component	0.50	0.46	0.48
Assemble System	0.58	0.53	0.55
Take sub system	0.79	0.49	0.61
Put down subsystem	0.85	0.23	0.37

Table 2: Experimental results of high feature action classes on the Inhard Dataset

Meta-Action Class labels	Precision	Recall	F1- Score
Assemble System	0.76	0.93	0.83
Picking in front	0.86	0.90	0.88
Picking left	0.91	0.87	0.89
Put down screwdriver	0.72	0.68	0.70
Put down subsystem	0.99	0.83	0.90
Take screwdriver	0.66	0.51	0.58
Take subsystem	0.73	0.63	0.68



(a) (b)

Figure 3: Overall training and validation accuracy using MobilnetV2-SSD low-level industrial human action classification for assembly setting-(a) for all action classes, (b) high feature action classes

model on the same dataset. The problem with the conventional CNN+LSTM model is computational time and model size, which may not be appropriate for fluent HRI applications, where response time does matter. Our methodology could be suited for dynamic HRI applications where the recognition model requires environmental structure information in addition to motion and objects which OpT covers. Additionally, for the classification, the mobilnetV2-SSD is appropriate in terms of computational complexity.

Human action recognition applications are expanding rapidly in computer vision, surveillance systems, human-machine interaction, and human-object interaction. Several approaches have been introduced in the literature for determining human actions using spatio-temporal data. However, it remains a challenging field regarding fluent HRI and human-robot collaborative applications, mainly low-level operator actions in manufacturing work cells. Although the proposed algorithm performs well, it still suffers from numerous scarcities such as the inability to distinguish across relatively similar postures (Take screwdriver, Take measuring rod, where the operator position is primarily static) and dynamic activities. In most instances, the relationship between completed actions is inconsistent throughout the activities, making recognition even more difficult.

5 CONCLUSION AND FUTURE OUTLOOKS

In this paper, we introduced an operator's action recognition algorithm based on optical flow and mobilnetV2-SSD for HRC environment. We investigated that the optical flow may infer the operator's movements, objects and the structure of environment-related items. We concluded that the presented model performs well but faces a few limitations. Different classes in the dataset demand a high feature engineering procedure, and recognition of static pose and dynamic actions remains difficult. However, predicting operator's behavior and identifying the following actions are essential for enhancing HRI's collaborative operations. Integrating activity detection algorithms into the cognitive assembly line can transform the human-robot teaming structure. Appropriate datasets in the industrial domain and a more responsive algorithm can foster human-robot team collaboration. As a potential application for industrial environment, we will introduce a human-robot teaming structure driven primarily by visual recognition framework in future work.

ACKNOWLEDGMENTS

The authors gladly acknowledge the CISC project, funded from the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie grant agreement no. 955901.

REFERENCES

[1] Mohammad Anvaripour, Mahta Khoshnam, Carlo Menon, and Mehrdad Saif. 2020. FMG- and RNN-Based Estimation of Motor Intention of Upper-Limb Motion in Human-Robot Collaboration. *Front. Robot. AI* 7, December (2020), 1–13. DOI:https://doi.org/10.3389/frobt.2020.573096

[2] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition. *Cvpr* (2017), 6299–6308.

[3] Eva Coupeté, Fabien Moutarde, and Sotiris Manitsaris. 2015. Gesture Recognition Using a Depth Camera for Human Robot Collaboration on Assembly Line. *Procedia Manuf.* 3, Ahfe (2015), 518–525. DOI:https://doi.org/10.1016/j.promfg.2015.07.216

[4] Mejdi Dallel, Vincent Havard, David Baudry, and Xavier Savatier. 2020. InHARD-Industrial Human Action Recognition Dataset in the Context of Industrial Collaborative Robotics. *Proc. 2020 IEEE Int. Conf. Human-Machine Syst. ICHMS 2020* (2020), 7–12. DOI:https://doi.org/10.1109/ICHMS49158.2020.9209531

[5] Mejdi Dallel, Vincent Havard, Yohan Dupuis, and David Baudry. 2022. A Sliding Window Based Approach with Majority Voting for Online Human Action Recognition using Spatial Temporal Graph Convolutional Neural Networks. *ACM Int. Conf. Proceeding Ser.* (2022), 155–163. DOI:https://doi.org/10.1145/3529399.3529425

[6] Yonghao Dang, Fuxing Yang, and Jianqin Yin. 2020. DWnet: Deep-wide network for 3D action recognition. *Rob. Auton. Syst.* 126, (2020), 103441. DOI:https://doi.org/10.1016/j.robot.2020.103441

[7] Mickael Delamare, Cyril Laville, Adnane Cabani, and Houcine Chafouk. 2021. Graph convolutional networks skeleton-based action recognition for continuous data stream: A sliding window approach. *VISIGRAPP 2021 - Proc. 16th Int. Jt. Conf. Comput. Vision, Imaging Comput. Graph. Theory Appl.* 5, Visigrapp (2021), 427–435. DOI:https://doi.org/10.5220/0010234904270435

[8] Yong Du, Wei Wang, and Liang Wang. 2015. Hierarchical recurrent neural network for skeleton based action recognition. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 07-12-June, (2015), 1110–1118. DOI:https://doi.org/10.1109/CVPR.2015.7298714

[9] Muhammad Hassan, Tasweer Ahmad, Ali Farooq, Syed Asghar Ali, Syed Rizwan hassan, and Nudrat Liaqat. 2014. A Review on Human Actions Recognition Using Vision Based Techniques. *J. Image Graph.* 2, 1 (2014), 28–32. DOI:https://doi.org/10.12720/joig.2.1.28-32

[10] Andreas Hofmann. 2019. A Model-Based Human Activity Recognition for Human – Robot Collaboration. (2019), 2–9.

[11] Phat Nguyen Huu, Huong Nguyen Thi Thu, and Quang Tran Minh. 2021. Proposing a Recognition System of Gestures Using MobilenetV2 Combining Single Shot Detector Network for Smart-Home Applications. *J. Electr. Comput. Eng.* 2021, (2021). DOI:https://doi.org/10.1155/2021/6610461

[12] Hema S. Koppula and Ashutosh Saxena. 2016. Anticipating Human Activities Using Object Affordances for Reactive Robotic Response. *IEEE Trans. Pattern Anal. Mach. Intell.* 38, 1 (2016), 14–29. DOI:https://doi.org/10.1109/TPAMI.2015.2430335

[13] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. 2011. HMDB: A large video database for human motion recognition. *Proc. IEEE Int. Conf. Comput. Vis.* (2011), 2556–2563. DOI:https://doi.org/10.1109/ICCV.2011.6126543

[14] Ivan Laptev, Inria Willow, Computer Science, and Ecole Normale Suprieure. 2010. Efficient feature extraction , encoding and classification. (2010). Retrieved from http://youtube.com/t/press_statistics

[15] Qinghua Li, Zhao Zhang, Yue You, Yaqi Mu, and Chao Feng. 2020. Data Driven Models for Human Motion Prediction in Human-Robot Collaboration. *IEEE Access* (2020). DOI:https://doi.org/10.1109/ACCESS.2020.3045994

[16] Shufei Li, Pai Zheng, Junming Fan, and Lihui Wang. 2022. Toward Proactive Human-Robot Collaborative Assembly: A Multimodal Transfer-Learning-Enabled Action Prediction Approach. *IEEE Trans. Ind. Electron.* 69, 8 (2022), 8579–8588. DOI:https://doi.org/10.1109/TIE.2021.3105977

[17] Wanqing Li, Zhengyou Zhang, and Zicheng Liu. 2010. Action recognition based on a bag of 3D points. *2010 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. - Work. CVPRW 2010* 2010, (2010), 9–14. DOI:https://doi.org/10.1109/CVPRW.2010.5543273

[18] Hongyi Liu and Lihui Wang. 2017. Human motion prediction for human-robot collaboration. *J. Manuf. Syst.* 44, (2017), 287–294. DOI:https://doi.org/10.1016/j.jmsy.2017.04.009

[19] Zhen Tao Liu, Fang Fang Pan, Min Wu, Wei Hua Cao, Lue Feng Chen, Jian Ping Xu, Ri Zhang, and Meng Tian Zhou. 2016. A multimodal emotional communication based humans-robots interaction system. *Chinese Control Conf. CCC* 2016-Augus, (2016), 6363–6368. DOI:https://doi.org/10.1109/ChiCC.2016.7554357

[20] Zitong Liu, Quan Liu, Wenjun Xu, Zhihao Liu, Zude Zhou, and Jie Chen. 2019. Deep learning-based human motion prediction considering context awareness for human-robot collaboration in manufacturing. *Procedia CIRP* 83, (2019), 272–278. DOI:https://doi.org/10.1016/j.procir.2019.04.080

[21] Ren C. Luo and Licong Mai. 2019. Human Intention Inference and On-Line Human Hand Motion Prediction for Human-Robot Collaboration. *IEEE Int. Conf. Intell. Robot. Syst.* 1 (2019), 5958–5964. DOI:https://doi.org/10.1109/IROS40897.2019.8968192

[22] Khan Muhammad, Salman Khan, Mohamed Elhoseny, Syed Hassan Ahmed, and Sung Wook Baik. 2019. Efficient Fire Detection for Uncertain Surveillance Environment. *IEEE Trans. Ind. Informatics* 15, 5 (2019), 3113–3122. DOI:https://doi.org/10.1109/TII.2019.2897594

[23] Pedro Neto, Miguel Simão, Nuno Mendes, and Mohammad Safeea. 2019. Gesture-based human-robot interaction for human assistance in manufacturing. *Int. J. Adv. Manuf. Technol.* 101, 1–4 (2019), 119–135. DOI:https://doi.org/10.1007/s00170-018-2788-x

[24] Kishore K. Reddy and Mubarak Shah. 2013. Recognizing 50 human action categories of web videos. *Mach. Vis. Appl.* 24, 5 (2013), 971–981. DOI:https://doi.org/10.1007/s00138-012-0450-4

[25] Alina Roitberg and Markus Rickert. P259-Roitberg. 259–266.

[26] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang Chieh Chen. 2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* (2018), 4510–4520. DOI:https://doi.org/10.1109/CVPR.2018.00474

[27] Amir Shahroudy, Jun Liu, Tian Tsong Ng, and Gang Wang. 2016. NTU RGB+D: A large scale dataset for 3D human activity analysis. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 2016-Decem, (2016), 1010–1019. DOI:https://doi.org/10.1109/CVPR.2016.115

[28] Xiaoran Shi, Yaxin Li, Feng Zhou, and Lei Liu. 2018. Human Activity Recognition Based on Deep Learning Method. *2018 Int. Conf. Radar, RADAR 2018* (2018). DOI:https://doi.org/10.1109/RADAR.2018.8557335

[29] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.* (2015), 1–14.

[30] Shuyang Sun, Zhanghui Kuang, Lu Sheng, Wanli Ouyang, and Wei Zhang. 2018. Optical Flow Guided Feature: A Fast and Robust Motion Representation for Video Action Recognition. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* (2018), 1390–1399. DOI:https://doi.org/10.1109/CVPR.2018.00151

[31] Aaquib Tabrez and Bradley Hayes. 2019. Improving Human-Robot Interaction Through Explainable Reinforcement Learning. *ACM/IEEE Int. Conf. Human-Robot Interact.* 2019-March, (2019), 751–753. DOI:https://doi.org/10.1109/HRI.2019.8673198

[32] Tadele Belay Tuli, Valay Mukesh Patel, and Martin Manns. 2022. CONFERENCE ON PRODUCTION SYSTEMS AND LOGISTICS CPSL 2022 3 rd Conference on Production Systems and Logistics Industrial Human Activity Prediction and Detection Using Sequential Memory Networks. (2022), 62–72. Retrieved from https://doi.org/10.15488/12144

[33] Xianhe Wen, Heping Chen, and Qi Hong. 2019. Human Assembly Task Recognition in Human-Robot Collaboration based on 3D CNN. *9th IEEE Int. Conf. Cyber Technol. Autom. Control Intell. Syst. CYBER 2019* (2019), 1230–1234. DOI:https://doi.org/10.1109/CYBER46603.2019.9066597

[34] Hong Bo Zhang, Yi Xiang Zhang, Bineng Zhong, Qing Lei, Lijie Yang, Ji Xiang Du, and Duan Sheng Chen. 2019. A comprehensive survey of vision-based human action recognition methods. *Sensors (Switzerland)* 19, 5 (2019), 1–20. DOI:https://doi.org/10.3390/s19051005

[35] Jianjing Zhang, Hongyi Liu, Qing Chang, Lihui Wang, and Robert X. Gao. 2020. Recurrent neural network for motion trajectory prediction in human-robot collaborative assembly. *CIRP Ann.* 69, 1 (2020), 9–12. DOI:https://doi.org/10.1016/j.cirp.2020.04.077