2022 26th International Conference on Pattern Recognition (ICPR)
August 21-25, 2022, Montréal, Québec, Canada

HRI30: An Action Recognition Dataset for Industrial Human-Robot Interaction

Francesco Iodice* †
Email: francesco.iodice@iit.it

Elena De Momi†
Email: elena.demomi@polimi.it

Arash Ajoudani*
Email: arash.ajoudani@iit.it

*Human-Robot Interfaces and physical Interaction HRI^2 Lab of Istituto Italiano di Tecnologia (IIT), Genoa, Italy.
†Department of Electronics, Information, and Bioengineering, Politecnico di Milano, Milan, Italy.

2022 26th International Conference on Pattern Recognition (ICPR) | 978-1-6654-9062-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICPR56361.2022.9956300

***Abstract*—Over the past years, action recognition techniques have gained significant attention in computer vision and robotics research. Nevertheless, their performances in realistic applications, despite dedicated efforts to collect and annotate medium/large datasets, remain far from satisfactory, especially when it comes to applications in the field of human-robot collaboration. In response to this shortfall, we create a dataset not dispersive in its classes but sectoral, i.e., dedicated exclusively to the industrial environment and human-robot collaboration. Specifically, we describe our ongoing collection of the 'HRI30' database for industrial action recognition from videos, containing 30 categories of industrial-like actions and 2940 manually annotated clips. We test our dataset on multiple action detection approaches and compare it with the HMDB51 and UCF101 public datasets using the best-performing approach. We define a baseline of 86.55% Top-1 accuracy and 99.76% Top-5 accuracy, hoping that this dataset will encourage research towards understanding actions in collaborative industrial scenarios. The dataset can be downloaded at the following link: 10.5281/zenodo.5833411**

I. Introduction

How to interpret human behaviour and infer an action from a video is a research topic that the computer vision communities have studied extensively over the past decade. Much of this progress is due to the creation of datasets acquired with conventional RGB cameras that have allowed the development of models that can reveal both object properties and detect human actions. Although there has been a significant amount of research on the representation and understanding of the human actions, up to now, this progress falls short on addressing the day-to-day activities of the professional workplaces such as manufacturing and service industries.

Furthermore, with the increasing flexibility and responsiveness demands of the new industrial revolution, humans and robots are expected to be co-workers and co-inhabitants in workplaces [1]. In this context, the use of collaborative robots (cobots), has demonstrated a high potential in redistribution of labor into better jobs, reducing physical efforts and promoting operator health. Hence, the recognition and understanding of human actions in such environments, and in particular, in human-cobot interaction scenarios become a crucial matter.

In this regard, we aim to provide the robotics and computer vision communities with a dataset capable of covering a large number of human actions performed in an industrial-like environment, allowing the development of models that can optimize human-robot collaboration and possibly address interaction and safety issues in industry 4.0.

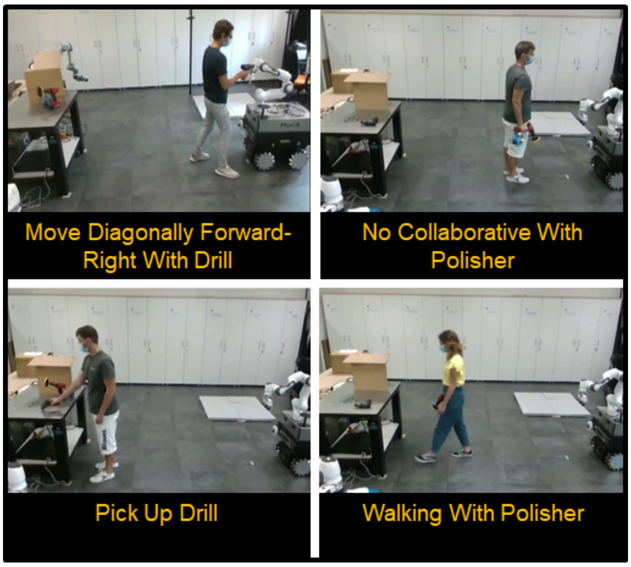
Move Diagonally Forward-Right With Drill
No Collaborative With Polisher
Pick Up Drill
Walking With Polisher

Fig. 1. Sample frames for four classes of actions in HRI30. They highlight the three semantic groups: human motion, human-object interaction, and human-robot interaction.

Therefore, this paper presents a fully labeled balanced dataset of industrial-like work activities. Eleven subjects, including one left-handed, three ambidextrous, and seven right-handed, participated in the experiment performed in a laboratory environment. The actions are divided into three sets. The first set aims at human-object interaction actions and includes picking up and putting down objects and tools. The second aims at actions without interaction and collaboration, and includes human movements with and without objects and tools. The third aims at collaborative and end-collaborative actions, including work application actions and paired movement actions, to reach any point in the work area.

To the best of our knowledge, HRI30, with its 30 classes, is the largest dataset of actions in an industrial-like environment available in the literature. As there are no publicly released datasets for the industrial sector, we perform comparative benchmarks on the actions HMDB51 [2], UCF101[3], and HRI30. The approach used for action recognition is based on [4], which offers higher performance with respect the other

TABLE I
SUMMARY OF ACTION RECOGNITION DATASETS IN HUMAN-ROBOT INTERACTION APPLICATIONS.

Dataset	Year	Data Modalities	Capture	Activities	Clips
MSR Action3D RGB-D Dataset	2010	Depth sequences	A depth camera acquires the depth through structure infrared light	20	320
JPL First-Person Interaction Dataset	2013	RGB videos	Captured by Kinect	7	-
HDM05 Dataset	2007	3D Motion	Captured by 3D motion sensors (MoCap)	70	1457
CMU Motion Capture	2011	3D Motion	Captured by 3D motion sensors (MoCap)	109	2605
KIT Whole-Body Human Motion	2015	3D Motion and videos	Captured by 3D motion sensors (MoCap) and a monocular camera	-	9727
TUM Kitchen Dataset	2009	RGB video	Captured by 3D motion sensors (MoCap), four monocular cameras, environmental RFID tags and magnetic sensors	20	-
Cornell Activity Dataset	2012	RGB + Depth + Skeleton	Captured by Kinect	12	-
MoCA	2020	3D Motion and videos	Captured by 3D motion sensors (MoCap) and three cameras	20	141

approaches tested.

The rest of the paper is organized as follows: Section II presents related work; Section III details the dataset with statistical data (III-A), action list (III-B), and data acquisition method (III-C); Section IV presents the Experiments and analysis on the collected data; and finally, Section V concludes this work.

II. RELATED WORK

This section will introduce and discuss the existing datasets for action recognition in both generic and Human-Robot Interaction (HRI) contexts, and then will explain the contribution of our work.

A. Generic Action Recognition Datasets

Many approaches to predicting action labels from videos have been studied in the literature in recent years. Action datasets containing annotated videos were made available to the community to evaluate these approaches. The first most influential datasets for classifying activities are KTH [5] and Weizmann [6]. Both contain simple actions, such as walking, running, and greeting, performed against a homogeneous background with a static camera. Subsequently, the datasets became more realistic with the extraction of videos from movies, TV shows, and web platforms for sharing and viewing multimedia content (e.g. Youtube). This realism is given by the naturalness of the scenes obtained through moving cameras and messy backgrounds, making them more demanding in recognition. The Hollywood [7] dataset is an example of this approach; it collects scenes from Hollywood films grouped into twelve categories. At the same time, the Thumos14 [8], UCF50 [9], UCF-Sports [10], and Olympic Sports [11] datasets are traceable to collections of videos extracted from YouTube, which point to the complexity of the action-focused exclusively on sports activities very complex. The subsequent datasets in the order of size, HMDB51 [2] and UCF101 [3], are presented as extensions of UCF50 [9]. They collect short videos of simple actions with a semantic organization divided into five groups: general-facial actions, facial actions with object manipulation, general body movements, body movements with object interaction, and body movements for human interaction, for HMDB51 [2]. While human-object interaction, body-motion only, human-human interaction, playing musical instruments, sports, are categorized in UCF101 [3].

B. Action Recognition Dataset in the HRI Context

These datasets are ideal for developing supervised classifiers for action recognition, thanks to their manual annotation method, which allows for the crop and capture of a single action in a short video. However, the number of classes to cover the most relevant actions, which involve humans in daily life, is insufficient due to the annotation method, which is effort-demanding and time-consuming. In this regard, recent large-scale datasets such as ActivityNet [12], Sports-1M [13], Something-Something [14], YouTube-8M [15], and Kinetics [16] rely on the automatic annotation method. While favouring a more significant number of classes than previous datasets, this method is potentially noisy, i.e., it introduces an unknown amount of label noise.

In addition to the datasets introduced previously, there are datasets in the literature collected from RGB, depth sequences and skeleton sequences, which are frequently used to evaluate HRI methods on robotic platforms. These datasets, shown in Table I, can be grouped into whole-body actions and kitchen actions.

The MSR Action3D [17], JPL First-Person interaction [18], HDM05 [19] and CMU motion capture [20] datasets are associated with the first group containing whole-body movements. Their data collection methods include depth sensors for the first dataset, a GoPro2 camera mounted on the head of a humanoid model for the second one, and motion capture sensors for the last two ones. Compared to the latter, the remaining datasets, recording both body movement and interactive objects in the actions, were collected from multiple devices. For example, the KIT Whole-Body Human Motion dataset [21] collects data from video recordings, auxiliary data (e.g. force data), anthropometric measurements and motion capture sensors. The TUM kitchen dataset [22] uses motion capture sensors, three fixed cameras, one mounted camera (above the head), environmental RFID tags, and magnetic sensors to detect when a door or drawer is opened. The Cornell activity dataset [23] adopts the Kinect sensor, with RGB video, depth images and skeleton sequences. Finally, the MoCA bi-

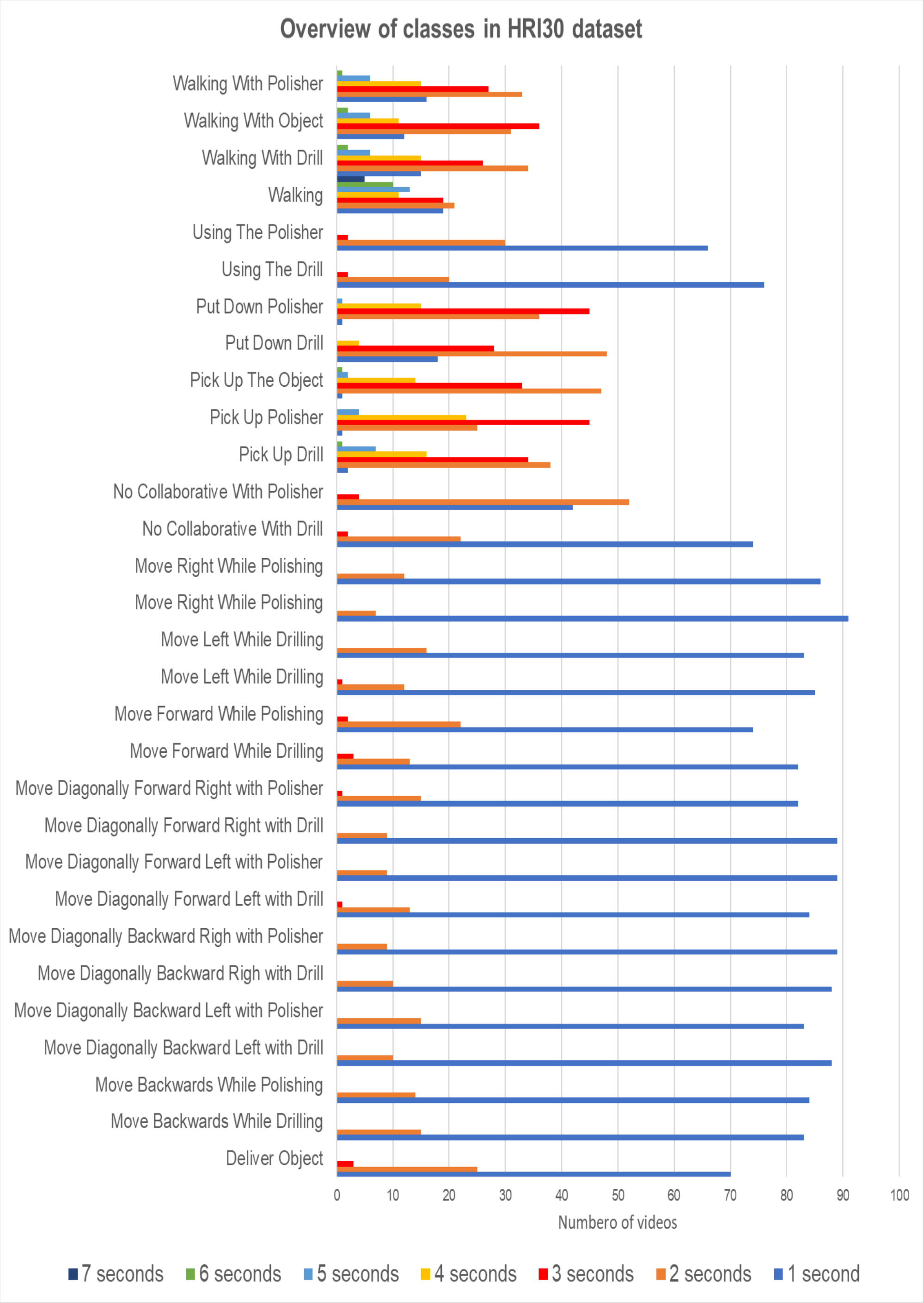

Fig. 2. Overview of classes and distribution of videos by duration for each class.

modal dataset [24] collects the data by motion capture sensors and video sequences captured from three views.

C. Contribution

A thorough analysis of the existing human action recognition datasets demonstrates that only few HRI datasets are available that target real-world applications, all of which are adapted to home settings. Therefore, given the shortage of datasets in industrial tasks, we aim to provide the community with a dataset created in a laboratory setting that includes actions commonly performed within manufacturing and service industries. In addition, the proposed dataset, with its dimensions of about 3000 videos, meets the requirements of deep learning algorithms for the development of intelligent learning models for action recognition and imitation in HRI applications.

The proposed dataset includes thirty classes and has a semantic representation that can be grouped into three categories: human-object interaction, body movement only and human-robot collaboration. To evaluate its efficiency, we tested with popular action recognition methods TSN [25], ir-CSN [26], TIN [28], and SlowOnly [27]; using the latter, we compared its accuracy with the HMDB51 [2] and UCF101 [3] datasets.

HRI30 is the first dataset focusing on industrial-like tasks. The dataset aims to provide human actions performed exclusively within a collaborative industrial environment between humans and robots.

III. Dataset Details

In this section, we present the statistical data of the videos, the actions contained, and the method of acquisition and annotation of human actions.

A. Statistics

In HRI30, we selected 30 classes (Fig.3), each associated with an action category, with a total of 2940 clips, 98 clips for each action class. The classes are of three types: Human-Object Perception, Body-Motion Only, and Human-Robot Collaboration. There are 11 subjects involved, including one left-handed, three ambidextrous, and seven right-handed.

TABLE II
Split lists Train/Test.The table shows the number of videos for each list.

Split	Train	Test
1	2100	840
2	2310	630
3	1890	1050

B. Action list

The Train/Test split is random and follows the rules shown in Table II.
Most of the videos have a duration between 1 and 2 seconds (Fig.2), and all have a fixed frame rate of 30 FPS and a resolution of 720×480, respectively.

The list of actions included in HRI30 includes: Deliver Object, Move Backwards While Drilling, Move Backwards While Polishing, Move Diagonally Backwards Left with Drill, Move Diagonally Backwards Left With Polisher, Move Diagonally Backwards Right With Drill, Move Diagonally Backwards Right With Polisher, Move Diagonally Forwards Left with Drill, Move Diagonally Forwards Left With Polisher, Move Diagonally Forwards Right With Drill, Move Diagonally Forwards Right With Polisher, Move Forwards While Drilling, Move Forwards While Polishing, Move Left While Drilling, Move Left While Polishing, Move Right While Drilling, Move Right While Polishing, No Collaborative With Drill, No Collaborative With Polisher, Pick Up Drill, Pick Up Polisher, Pick Up The Object, Put Down Drill, Put Down Polisher, Using The Drill, Using The Polisher, Walking, Walking With Drill, Walking With Object, Walking With Polisher.

C. Data Acquisition

Video is obtained by positioning the Realsense D435i camera at the height of approximately 2 meters and 30 cm to provide visual coverage of the entire working area. The data distribution is uniform among the classes, which helps avoid data imbalance during the training of classifier and includes videos cut to contain only the intended action. The annotation of time boundaries, where a task is performed in a video, is manual. To deal with this manual process, we rely on editing software [1] to speed up this time-consuming step and obtain a curated set of action instances. The nomenclature of video files takes its cue from [3] and has the following form:

$$v\_L\_gG\_cC.avi$$

Where L, G, and C represent the action class label, the group, and the clip number, respectively. Audio is not stored in the action collection.

D. Ethics Corner

The data collection was carried out in accordance with the Declaration of Helsinki and the protocol was approved by the ethics committee Azienda Sanitaria Locale (ASL) Genovese N.3 (Protocol IIT_HRII_ERGOLEAN 156/2020).

IV. Experiments

This section contains evaluations that show the challenge in compressing HRI30 tasks for current computer vision algorithms. We use these evaluations to choose the best approach in the literature to benchmark our video dataset against standard video datasets.

A. Implementation details

We use an open-source machine learning library, Pytorch, widely used for its flexibility and computational power. We use an Nvidia RTX 2080 Super video card for training, tuning, and testing. The input frames of each phase are first sampled one every four frames per clip and then resized to 256 x 256. For the training and tuning phase, the images are flipped randomly. The learning rate is 0.001, and the stochastic gradient descent (SGD) optimizer [31] is used.

B. Evaluation metrics

For evaluation, we use accuracy as a classification metric. It is suitable for both binary and multiclass classification problems. Moreover, it is a valid evaluation choice for classification problems on well-balanced and unbiased datasets or without class imbalances.

To quantify the accuracy of the methods, we use:

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} \tag{1}$$

It is the proportion of correct predictions (both true positives and true negatives) over the total number of cases examined, with TP = True Positive, FP = False Positive, TN = True Negative, FN = False Negative.

[1] Adobe Premiere Pro

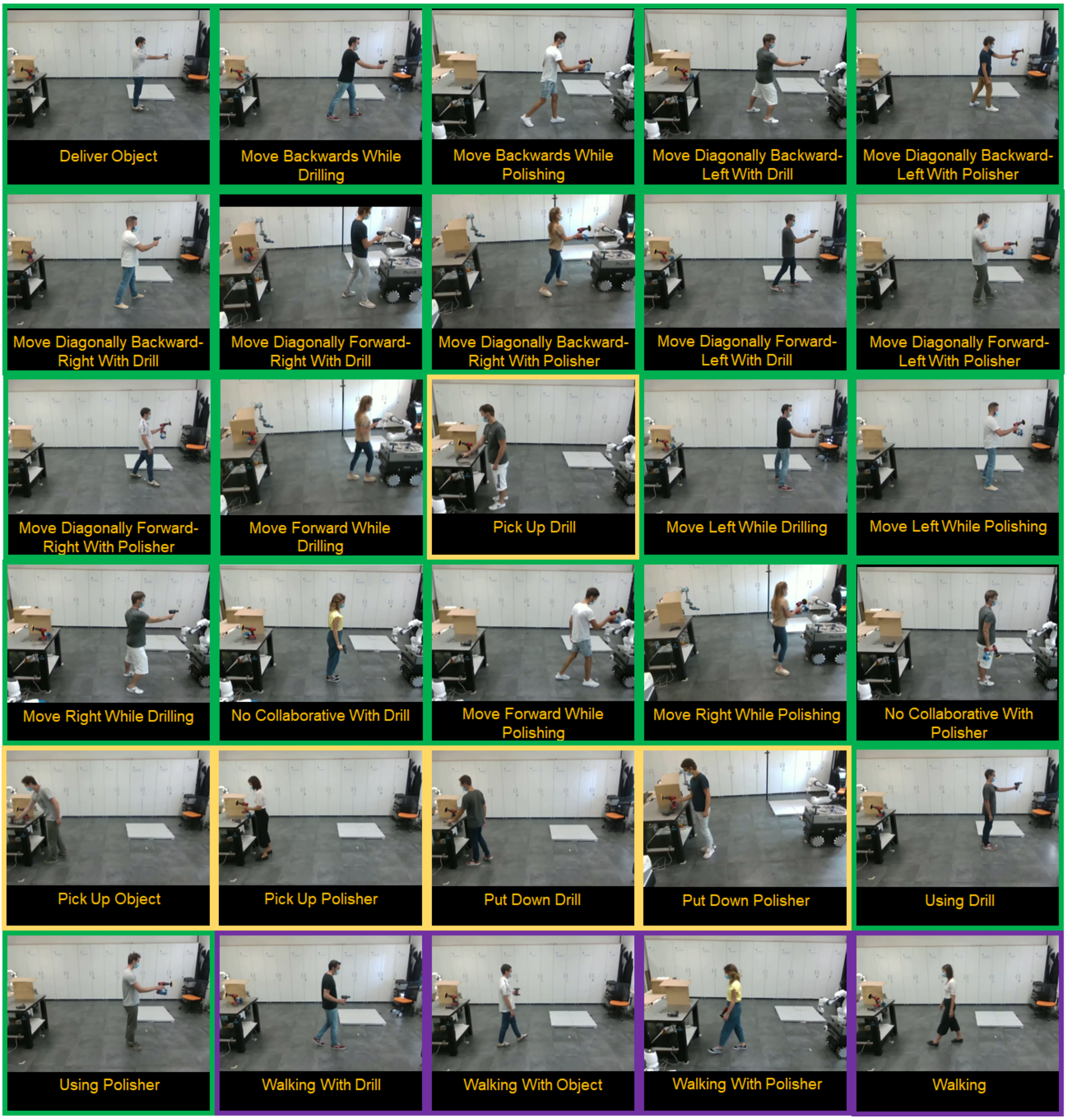
Deliver Object
Move Backwards While Drilling
Move Backwards While Polishing
Move Diagonally Backward-Left With Drill
Move Diagonally Backward-Left With Polisher
Move Diagonally Backward-Right With Drill
Move Diagonally Forward-Right With Drill
Move Diagonally Backward-Right With Polisher
Move Diagonally Forward-Left With Drill
Move Diagonally Forward-Left With Polisher
Move Diagonally Forward-Right With Polisher
Move Forward While Drilling
Pick Up Drill
Move Left While Drilling
Move Left While Polishing
Move Right While Drilling
No Collaborative With Drill
Move Forward While Polishing
Move Right While Polishing
No Collaborative With Polisher
Pick Up Object
Pick Up Polisher
Put Down Drill
Put Down Polisher
Using Drill
Using Polisher
Walking With Drill
Walking With Object
Walking With Polisher
Walking

Fig. 3. Thirty diverse action categories included in the HRI30 dataset are shown in a sample frame. The color of frame borders specifies to which action type they belong: Human-Object Perception, Body-Motion Only and Human-Robot Collaboration.

TABLE III
COMPARATIVE ANALYSIS OF WELL-KNOWN ACTION RECOGNITION METHODS ON THE 3 SPLIT LISTS OF THE HRI30 DATASET.

Network			**Train/Test Split List 1**		**Train/Test Split List 2**		**Train/Test Split List 3**	
Method	Backbone	Pretrained	Top-1 Accuracy	Top-5 Accuracy	Top-1 Accuracy	Top-5 Accuracy	Top-1 Accuracy	Top-5 Accuracy
TSN	Resnet-50	-	74.05	97.2	73.98	99.05	73.71	98.86
IRCSN	Resnet-50	IG65M	79.17	99.88	74.64	99.84	77.67	99.62
TIN	Resnet-50	-	62.10	93.78	77.14	98.89	65.14	96.76
SlowOnly	Resnet-50	Kinetics-400	86.55	99.76	83.49	99.84	82.43	99.90
SlowOnly	Resnet-50	Imagenet	64.29	99.79	65.11	99.68	60.48	99.14

Our experiments consider two accuracy types: the conventional top-1 accuracy that extracts the maximum value from the final outputs of the softmax and the top-5 accuracy calculated to measure the frequency with which the predicted class falls within the first 5 values of the softmax distribution.

C. Comparison of different networks for action recognition

In order to demonstrate the strengths and weaknesses of the proposed dataset, we performed a comparative analysis on a set of well-known action recognition methods.

- **The time segment network** (TSN) [25] is based on modeling the long-term temporal structure.
- **The temporal interlacing network** (TIN) [28] weaves spatial representations from past to future and vice versa instead of learning temporal features.
- **The interaction-reduced channel-separated network** (ir-CSN) is a variation of the CSN network [26] that factorizes 3D convolutions by separating channel interactions and Spatio-temporal interactions and achieves higher accuracy at a low computational cost. Ir-CSN differs from CSN in the reduced number of channels used in its architecture.
- **SlowOnly** is a variation of the SlowFast [27] network. The latter is one of the latest networks for action recognition. It is based on encoding motion in a 'fast' path operating at high frame rates and simultaneously capturing semantics through a 'slow' path with low frame rates. SlowOnly considers only the slow path.

All the networks have the Resnet-50 [29] backbone; for SlowOnly, we use two pre-trained models, one on Imagenet [30], the other on kinetics-400 [16].
Table III shows the results of the action recognition networks on the three random Train/Test split lists in the dataset, with all networks learning best on the first split list except the TIN and SlowOnly pre-trained Imagenet networks, which learn best on the second split list. For all split lists, the additional data from ImageNet [30] was not as helpful for the SlowOnly network as the pre-training on Kinetics [16], and most baselines had almost comparable performance in terms of Top-1 accuracy, except SlowOnly pre-trained on kinetics400 [16], which performed better. Furthermore, the relatively low Top-1 accuracy and high Top-5 accuracy highlight the difficulty of these action recognition methods in predicting the correct action, despite the fact that it is always recognized among the top five actions. This issue stems from the dataset's complexity, which includes complicated backdrop clutter and action category similarity. Such similarity has a significant impact within dynamic collaborative industrial environments because the robot performs different tasks depending on the perceived human intention, such as activating when walking with an object toward it and then grabbing the object or reaching an ergonomic position with the object to the human to initiate collaboration when walking with a tool toward it.

D. Comparison with traditional action recognition datasets

In addition to HRI30, we also analyzed standard video datasets to compare the difficulty in action recognition on the same model.

- **HMDB51** [2]. It has 6,766 videos distributed in 51 action categories. The dataset has background clutter and variations in the movement pattern.
- **UCF101** [3]. This dataset is among the largest benchmarks available for action recognition. UCF101 has about 13320 videos for 101 share classes.

All three datasets are balanced and include trimmed videos. In addition, they have three splits of train/test data, and performance is measured by the average classification accuracy across the splits.

TABLE IV
COMPARISON ANALYSIS ON THE SLOWONLY MODEL BETWEEN THE HRI30 DATASET AND THE STANDARD VIDEO DATASETS FOR ACTION RECOGNITION.

Dataset	Top-1 Accuracy	Top-5 Accuracy
HMDB-51 Dataset	65.95	91.05
UCF-101 Dataset	92.78	99.43
HRI30 Dataset	86.55	99.76

The analysis of the three datasets uses the best-performing approach obtained in the previous comparison. Indeed, Table IV shows the accuracy of the three datasets on the SlowOnly network pre-trained on Kinetics-400 [16] and with the Resnet-50 [29] backbone. These results show that our dataset achieves higher accuracy results than HMDB51 but lower than the UCF-101 dataset and highlight its versatility both for the vision community, as HRI30 poses a significant challenge for state-of-the-art recognition methods and robotics, as it enables the creation of viable action recognition application frameworks for human-robot interaction in industrial environments.

V. CONCLUSION

This paper proposed HRI30, the largest action dataset suitable for Industry 4.0 with 30 action categories and slightly less than 3.000 video clips. Its data collection was performed on eleven subjects performing various repetitions of each activity to provide a good amount of annotated training data samples. HRI30 differs from the existing action recognition datasets in being the first dataset made available to the community to focus its videos on human-robot interaction for industrial-like tasks. Given the performance level of the state-of-the-art machine vision representative algorithms, our dataset proves to be a good starting point. Future work will focus on extending the number of classes and videos for each by considering more viewpoints with more cameras for the occlusion problem. In the meantime, we hope this dataset will be useful for researchers who are or will be working in human action recognition for Industry 4.0.

ACKNOWLEDGMENT

This work was supported by the ERC-StG Ergo-Lean (Grant Agreement No. 850932).

References

[1] Ajoudani, Arash, et al. ”Progress and prospects of the human–robot collaboration.” Autonomous Robots 42.5 (2018): 957-975.

[2] Jhuang, H., et al. ”A large video database for human motion recognition.” Proc. of IEEE International Conference on Computer Vision. Vol. 4. No. 5. 2011.

[3] Soomro, Khurram, Amir Roshan Zamir, and Mubarak Shah. ”UCF101: A dataset of 101 human actions classes from videos in the wild.” arXiv:1212.0402 (2012).

[4] Feichtenhofer, Christoph, et al. ”Slowfast networks for video recognition.” Proceedings of the IEEE/CVF international conference on computer vision. 2019.

[5] Schuldt, Christian, Ivan Laptev, and Barbara Caputo. ”Recognizing human actions: a local SVM approach.” Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.. Vol. 3. IEEE, 2004.

[6] Blank, Moshe, et al. ”Actions as space-time shapes.” Tenth IEEE International Conference on Computer Vision (ICCV’05) Volume 1. Vol. 2. IEEE, 2005.

[7] Laptev, Ivan, et al. ”Learning realistic human actions from movies.” 2008 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2008.

[8] Jiang, Yu-Gang, et al. ”THUMOS challenge: Action recognition with a large number of classes.” (2014).

[9] Reddy, Kishore K., and Mubarak Shah. ”Recognizing 50 human action categories of web videos.” Machine vision and applications 24.5 (2013): 971-981.

[10] Rodriguez, Mikel D., Javed Ahmed, and Mubarak Shah. ”Action mach a spatio-temporal maximum average correlation height filter for action recognition.” 2008 IEEE conference on computer vision and pattern recognition. IEEE, 2008.

[11] Niebles, Juan Carlos, Chih-Wei Chen, and Li Fei-Fei. ”Modeling temporal structure of decomposable motion segments for activity classification.” European conference on computer vision. Springer, Berlin, Heidelberg, 2010.

[12] Caba Heilbron, Fabian, et al. ”Activitynet: A large-scale video benchmark for human activity understanding.” Proceedings of the ieee conference on computer vision and pattern recognition. 2015.

[13] RKarpathy, Andrej, et al. ”Large-scale video classification with convolutional neural networks.” Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2014.

[14] Goyal, Raghav, et al. ”The” something something” video database for learning and evaluating visual common sense.” Proceedings of the IEEE international conference on computer vision. 2017.

[15] Abu-El-Haija, Sami, et al. ”Youtube-8m: A large-scale video classification benchmark.” arXiv preprint arXiv:1609.08675 (2016).

[16] Kay, Will, et al. ”The kinetics human action video dataset.” arXiv preprint arXiv:1705.06950 (2017).

[17] Li, Wanqing, Zhengyou Zhang, and Zicheng Liu. ”Action recognition based on a bag of 3d points.” 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops. IEEE, 2010.

[18] Ryoo, Michael S., and Larry Matthies. ”First-person activity recognition: What are they doing to me?” Proceedings of the IEEE conference on computer vision and pattern recognition. 2013.

[19] Müller, Meinard, et al. ”Mocap database hdm05.” Institut für Informatik II, Universität Bonn 2.7 (2007).

[20] [online] Available: http://mocap.cs.cmu.edu/. April 26, 2011.

[21] Mandery, Christian, et al. ”The KIT whole-body human motion database.” 2015 International Conference on Advanced Robotics (ICAR). IEEE, 2015.

[22] Tenorth, M.; Bandouch, J.; Beetz, M. The TUM kitchen data set of everyday manipulation activities for motion tracking and action recognition. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops, Kyoto, Japan, 27 September–4 October 2009; pp. 1089–1096.

[23] Sung, Jaeyong, et al. ”Unstructured human activity detection from rgbd images.” 2012 IEEE international conference on robotics and automation. IEEE, 2012.

[24] Nicora, Elena, et al. ”The MoCA dataset, kinematic and multi-view visual streams of fine-grained cooking actions.” Scientific Data 7.1 (2020): 1-15.

[25] Wang, Limin, et al. ”Temporal segment networks: Towards good practices for deep action recognition.” European conference on computer vision. Springer, Cham, 2016.

[26] Tran, Du, et al. ”Video classification with channel-separated convolutional networks.” Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.

[27] Feichtenhofer, Christoph, et al. ”Slowfast networks for video recognition.” Proceedings of the IEEE/CVF international conference on computer vision. 2019.

[28] Shao, Hao, Shengju Qian, and Yu Liu. ”Temporal interlacing network.” Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. No. 07. 2020.

[29] He, Kaiming, et al. ”Deep residual learning for image recognition.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[30] Deng, Jia, et al. ”Imagenet: A large-scale hierarchical image database.” 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009.

[31] Ruder, Sebastian. ”An overview of gradient descent optimization algorithms.” arXiv preprint arXiv:1609.04747 (2016).