# InHARD - Industrial Human Action Recognition Dataset in the Context of Industrial Collaborative Robotics


Mejdi DALLEL
LINEACT Laboratory, CESI
University of Rouen Normandy
Rouen, France
mejdi.dallel@univ-rouen.fr
mdallel@cesi.fr

Vincent HAVARD
LINEACT Laboratory, CESI
Rouen, France
vhavard@cesi.fr

David BAUDRY
LINEACT Laboratory, CESI
Rouen, France
dbaudry@cesi.fr

Xavier SAVATIER
IRSEEM Laboratory,
ESIGELEC
Rouen, France
xavier.savatier@esigelec.fr



***Abstract*—Nowadays, humans and robots are working more closely together. This increases business productivity and product quality, leading to efficiency and growth. However, human and robot collaboration is rather static; robots move to a specific position then humans perform their tasks while being assisted by the robots. In order to get a dynamic collaboration, robots need to understand the human's intention and learn to recognize the performed actions complementing therefore his capabilities and relieving him of arduous tasks. Consequently, there is a need for a human action recognition dataset for Machine Learning algorithms. Currently available depth-based and RGB+D+S based human action recognition datasets have a number of limitations, counting the lack of training samples along with distinct class labels, camera views, diversity of subjects and more importantly the absence of actual industrial human actions in an industrial environment. Actual action recognition datasets include simple daily, mutual, or health-related actions. Therefore, in this paper we introduce an RGB+S dataset named "Industrial Human Action Recognition Dataset" (InHARD) from a real-world setting for industrial human action recognition with over 2 million frames, collected from 16 distinct subjects. This dataset contains 13 different industrial action classes and over 4800 action samples. The introduction of this dataset should allow us the study and development of various learning techniques for the task of human actions analysis inside industrial environments involving human robot collaborations.**



***Keywords— Human Action Recognition, Dataset, Deep Learning, Human-Robot Collaboration (HRC), Industry 4.0, RGB+D, Skeleton, LSTM, RNN.***


## I. INTRODUCTION

In the Industry 4.0 factories, humans and robots can fully collaborate without separation or safety barriers. With current advances in industrial environments, robots are steadily shifting into a more human populated environment, which leads to the need of possessing more cognitive abilities. Other than the fact that they need to operate efficiently and safely in populated environments, robots need to also achieve higher levels of interaction and communication with humans [1].

For the past few years, manufacturing industries has been going through a major shift, the so-called Industry 4.0. In fact, robots are now moving into more crowded environments, the thing that demands more sophisticated and complex cognitive abilities for robots [2].

Advances in Human-Robot Collaboration (HRC) have led to significant progress and development in all areas. The profitability of HRC has increased productivity allowing expanded production, thus growth, repeatability, redeployment of labor into better jobs, higher profits and reducing the exposure to repetitive or arduous tasks with less safety risks [1] [3].

The notion of human-robot collaboration can be divided into 5 levels according to the degree of interaction [4] :

1. Isolated (Cell): this type of interaction is the most basic. In this type of collaboration, the robot is completely isolated from the operator and everyone works separately in its environment.
2. Coexistence: here the robot and the operator work in the same environment but on different parts or components.
3. Synchronization: the robot and the operator share the same space and work on the same part one after the other.
4. Cooperation: the robot and the operator share the same space and work on different parts or components.
5. Collaboration: which is the highest level of interaction where the robot and the operator share the same space and work simultaneously on the same part.

Eventually, HRC is likely to become more productive, as it will gradually be more efficient and faster with the evolution of sensors. HRC is heading slowly to full automation, but it is indispensable and incontrovertible. HRC will change future companies, therefore marking the beginning of a new phase of manufacturing [5].

Refining human-robot collaboration and human action recognition is going to be the next responsibility of the industrial community aiming to achieve greater shifts in productivity, thus freeing people from difficult and dangerous tasks. However, existing human action recognition datasets mostly contain common life activities (cooking, running, drinking etc.) with a clear lack of industrial human action recognition datasets. This paper proposes to fill this gap by proposing a dataset of a manufacturing process that can help the research community to go steps forward in 3D human action recognition in industrial environments, hence facilitating human robot collaboration.


978-1-7281-5871-6/20/$31.00 ©2020 IEEE.

The rest of the paper is organized as follows: Section II examines current human action recognition methods and benchmark datasets. Section III will be dedicated to introduce the proposed dataset. In Section IV we will explain segmented/online action recognition methods. Section V presents the dataset usage metrics to follow. Finally, the last section concludes the paper with future research directions.

## II. Related Work

In this section, we will further explore Human Robot Collaboration, which is an emerging trend in the field of industrial robotics under the Industry 4.0's strategy. Then we will examine the different type of sensors used for action recognition to finally review recent human action recognition methods and some available benchmark datasets in this domain.

### A. Human Robot collaboration (HRC)

Action recognition has come to the attention of researchers in recent years. Actual methods are designed for segmented action recognition, i.e. the action type is recognized once the whole action sequence has been observed. It is also called offline action recognition. Nevertheless, as things currently stand, it would be desirable if the action could be recognized during processing i.e. instantly, which will be more convenient for real time action recognition [2].

In order to assist the human in achieving a set of goals, it is the task of the robot to estimate his intention and to act accordingly. A person can communicate his or her intention either deliberately by explicit communication or implicitly by actions. The ways of communicating intention as elaborated in [2] are shown in TABLE I.

TABLE I. Main ways of communicating intentions [2]

| | | |
|---|---|---|
| **Communication of Intention** | Speech | Explicit Information |
| | | Emotion |
| | Gesture | Head/Eyes |
| | | Communicative Gesture |
| | Action | Manipulative Gesture |
| | | Proactive Task Execution |
| | Haptic Signal | Force/Torque |
| | | Angles/Orientation |
| | Physiological Signal | Approval/Arousal |

### B. Sensors for action recognition

In order to detect actions or activities, sensors have to be used. Real-time visual sensors have been increasingly attracting researchers due to numerous advantages that come with this kind of devices. Sensors used for action recognition are divided into two branches: Image based and Non-Image based sensors [6] as shown in Fig. 1.

- For Image based sensors we find markers, depth sensors, stereo cameras and single cameras.
- For Non-image Based sensors, we find gloves, band sensors and wireless sensors.

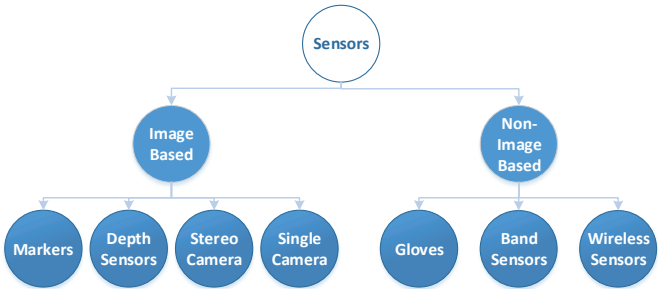


Fig. 1. Types of sensors used for action recognition

To collect our data, we will use both camera and non-camera sensors including Single cameras and Wireless sensors.

### C. Action recognition approaches

Action recognition usually consists of two steps; a feature extraction step and a classification step. Feature extraction consists in identifying distinctive features from a video sequence while being robust to noise. In the classification stage, we are interested in the possibility of identifying actions using Machine-Learning methods, while taking into account the variability to which an action class can be exposed, especially if they are executed by different subjects of different kinds and sizes, and with different speed and manner. After introducing some few benchmarks for action recognition datasets, a fewer number of online methods were proposed and evaluated on online datasets. In this section, we expose the main primitive extraction approaches used in the literature [7] as described in TABLE II.

TABLE II. Methods for Action Recognition [7]

| | **Modality** | **Recognition technique** | **Approach**[1] |
|---|---|---|---|
| **Action recognition based on RGB-D cameras** | RGB | Segmented | CNN, RNN, HTM, RBM, DBN, ISA, LSTM |
| | | Online | CNN, RNN |
| | Depth | Segmented | CNN |
| | | Online | CNN |
| | Skeleton | Segmented | CNN, RNN, HBM, HDP, DBM, DBN, SDA |
| | | Online | RNN |
| | Multi-Modal | Segmented | CNN, RNN, DBN, HCF |
| | | Online | RNN |

### D. Action recognition datasets

In TABLE III. we used a comparison between some of the current datasets made by [8] with our large-scale RGB+D action recognition dataset adding on top of that the action types involved in each dataset as well as the interactions between subjects if it exists.

[1] CNN: Convolutional Neural Networks
RNN: Recurrent Neural Networks
HTM: Hierarchical Temporal Memory
RBM: Restricted Boltzmann Machine
DBN: Deep Belief Networks
ISA: Independent Subspace Analysis
LSTM: Long Short-Term Memory
HBM: Hierarchical Bayesian Model
HDP: Hierarchical Dirichlet Process
DBM: Deep Boltzmann Machine
SDA: Stacked Denoising Autoencoder
HCF: Hierarchical Compound Features

TABLE III. ACTION RECOGNITION DATASETS

| Dataset | Samples | Classes | Subjects | Views | Sensors | Actions type | Modalities |
|---|---|---|---|---|---|---|---|
| MSR Action3D (2010) [9] | 567 | 20 | 10 | 1 | N/A | Daily activities *(e.g. high arm wave)* | D+ Skeleton 3DJoints |
| CAD-60 (2011) [10] | 60 | 12 | 4 | - | Kinect v1 | Daily activities *(e.g. brushing teeth)* | RGB+D +Skeleton 3DJoints |
| RGBD-HuDaAct (2011) [11] | 1189 | 13 | 30 | 1 | Kinect v1 | Daily activities *(e.g. eat a meal)* | RGB+D |
| MSRDailyActivity3D (2012) [12] | 320 | 16 | 10 | 1 | Kinect v1 | Daily activities *(e.g. read book)* | RGB+D +Skeleton 3DJoints |
| Act42 (2012) [13] | 6844 | 14 | 24 | 4 | Kinect v1 | Daily living *(e.g. Drink)* | RGB+D |
| 3D Action Pairs (2013) [14] | 360 | 12 | 10 | 1 | Kinect v1 | Daily action pairs *(e.g. Wear/take off)* | RGB+D +Skeleton 3DJoints |
| Multiview 3D Event (2013) [15] | 3815 | 8 | 8 | 3 | Kinect v1 | Daily events *(e.g. read book)* | RGB+D +Skeleton 3DJoints |
| Online RGB+D Action (2014) [16] | 336 | 7 | 24 | 1 | Kinect v1 | Daily actions *(e.g. Drinking)* | RGB+D +Skeleton 3DJoints |
| Northwestern-UCLA (2014) [17] | 1475 | 10 | 10 | 3 | Kinect v1 | Daily activities *(e.g. stand up)* | RGB+D +Skeleton 3DJoints |
| UWA3D Multiview (2014) [18] | 900 | 30 | 10 | 1 | Kinect v1 | Daily activities *(e.g. jumping)* | RGB+D +Skeleton 3DJoints |
| Office Activity (2014) [19] | 1180 | 20 | 10 | 3 | Kinect v1 | Office activities *(e.g. answering-phones)* | RGB+D |
| UTD-MHAD (2015) [20] | 861 | 27 | 8 | 1 | Kinect v1+WIS | Daily actions *(e.g. swipe left)* | RGB+D +Skeleton 3DJoints +ID |
| UWA3D Multiview II (2016) [21] | 1075 | 30 | 10 | 5 | Kinect v1 | Daily activities *(e.g. jumping)* | RGB+D +Skeleton 3DJoints |
| NTU-RGB+D (2016) [8] | 56880 | 60 | 40 | 80 | Kinect v2 | Daily, mutual, and health-related actions *(e.g. drinking, sneezing, punching)* | RGB+D+ IR+ Skeleton 3DJoints |
| OAD (2016) [22] | 59 | 10 | - | 1 | Kinect v2 | Daily activities *(e.g. drinking)* | RGB+D +Skeleton 3DJoints |
| **InHARD (2019)** | **4804** | **14** | **16** | **3** | **C920 cams+ Perception Neuron 32** | **Industrial actions/ activities** | **RGB +Skeleton +3DJoints** |

As shown in TABLE III. the presented datasets simply comprise daily, mutual, and health-related actions. We notice the lack of actual industrial human actions in industrial environments, which limits the usage of HAR in industry. Therefore, this paper proposes to fill this gap by proposing an industrial human actions dataset in industrial environments.

## III. THE INHARD DATASET

In order to facilitate Human-Robot collaboration in industrial environments, the InHARD dataset is created based on a real use-case in an industrial environment. This use-case involves an assembly of various parts and components and carried out on different stages with the help of the robotic arm UR10. This latter is a Collaborative Industrial Robotic Arm that is designed for large-scale tasks by automating processes and tasks such as packaging, palletizing, assembly and pick and place. The initial configuration of the setup is exposed in Fig. 2.

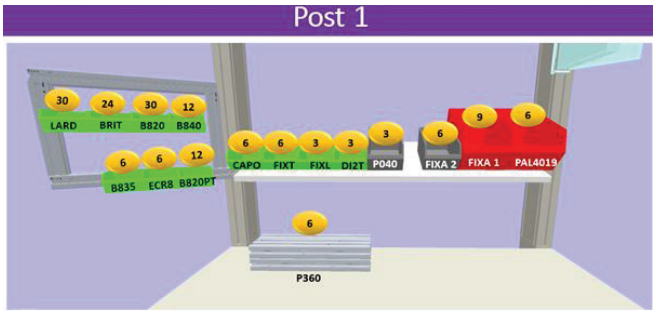


Fig. 2. Initial configuration of the setup

Participants have to consequently follow and consult instruction sheets in order to pick and assemble the right components together to finally get a final sub-system as shown in Fig. 3.

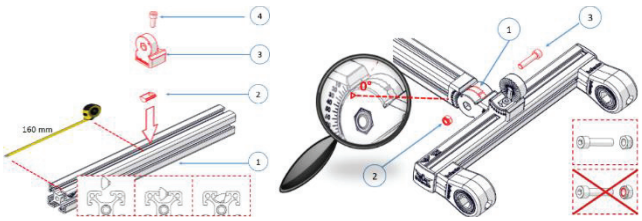


Fig. 3. First and final operations of the manipulation

To help the community to go a step further in Human robot collaboration in industrial environments, we are making this dataset publicly available[2]. The latter comprises all data's modalities and thereafter we will explain the processing to be done on these data.

We also provide a dataframe with all info including Filename, Subject, Operation, Action low/high level label, Action start/end, Duration etc. in order to facilitate the dataset handling and use. An extract of the dataframe is described in Fig. 4.

| File_name | Subject | Operation | Action_label | Meta_action_number | Meta_action_label | Action_start_bvh_frame | Action_end_bvh_frame |
|---|---|---|---|---|---|---|---|
| P01_R01... | P01 | OP010 | [OP010] Consult sheets | 2 | Consult sheets | 323 | 1071 |

Fig. 4. Dataframe of the InHARD dataset info

In this section, we present the details of the proposed InHARD dataset, including data modalities and the different sensors used through the entire data collection. Thereafter we elaborate action classes, subjects and the different views captured.

### A. Subjects

We invited 16 distinct subjects for our data collection to perform an industrial system assembly task with the help of an UR10 Cobotic arm. A consistent ID number is assigned to each subject. Each subject performs the task two up to four times. Thus, we ensure that short-term actions are captured.

[2] https://github.com/vhavard/InHARD
https://recherche.cesi.fr/inhard-industrial-human-action-recognition-dataset/

### B. *Modalities and views*

To collect our dataset, we gathered data using two modalities through different sensors.

**Skeleton modality**

We used a "Combination Perception Neuron 32 Edition v2" motion sensor to capture the skeletal data delivered at a frequency of 120 Hz. Skeleton data comprises the 3D locations (Tx, Ty and Tz) of 17 major body joints detected and tracked throughout the entire scene and 3 rotations around each axis (Rx, Ry and Rz) as shown in Fig. 5. Skeleton data are saved into BVH format files and stored in the Skeleton/ folder of the InHARD dataset.

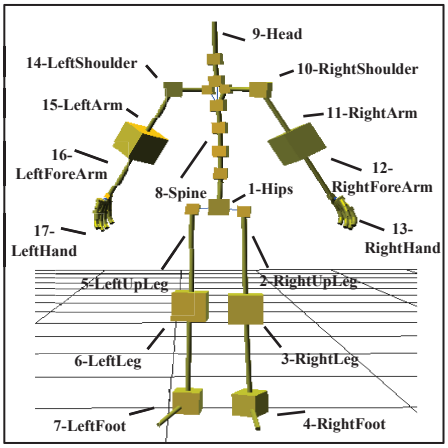


Fig. 5. Configuration of the 17 body joints in the InHARD dataset Views

**Video modality and point of views**

We also used three "Logitech Webcams C920" situated in three different posts covering three different views (top, left side and right side) to capture RGB data. The three video camera streams are gathered into one RGB video file captured with a 1280x720 resolution and a 30 f/s framerate (FPS).

The 3 cameras placed captures three different views of the same action. For each setup, two cameras were placed at the same height but at two different horizontal angles: -45° and +45° to capture both left and right sides. The third camera is placed on top of the subjects to capture the top view. Camera 1 always observes top views and is displayed on the top left quarter of the RGB video. Camera 2 observes left side views and is shown on the top right quarter of the RGB video. Lastly, Camera 3 observes right side views and is displayed on the bottom right quarter of the RGB video as shown in Fig. 6.

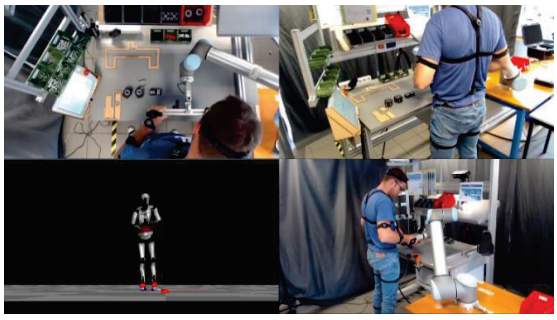

Fig. 6. Skeleton data in BVH format (bottom left).
1280x720 RGB videos containing 3 different views (Top, Left and Right)

### C. *Action classes*

In order to label the actions, we used a video annotation tool called Anvil [23]. It offers multi-layered annotation based on a user-defined coding scheme. During coding the user can see color-coded elements on multiple tracks in time-alignment. Some special features are inter-level links, non-time objects, time tracks, analysis of coding agreements, synchronization management of flows with a different FPS, 3D visualization of Motion Capture data (.BVH) and a project tool to manage entire corpuses of annotation files.

We have identified 14 low-level action classes and 72 high-level action classes where actions are a lot more accurate. For the time being, we will use only low-level action classes described in TABLE IV.

TABLE IV. INHARD LOW/HIGH LEVEL ACTION CLASSES

| ID | Meta action label | Action label |
|---|---|---|
| **0** | No action | No action |
| **1** | Consult sheets | Consult sheets |
| **2** | Turn sheets | Turn sheets |
| **3** | Take screwdriver | Take screwdriver |
| **4** | Put down screwdriver | Put down screwdriver |
| **5** | Picking in front | Catch Profile P040<br>Catch Fixation FIXA1<br>Catch Fixation FIXA2<br>Catch Pillow block bearing PAL4019 |
| **6** | Picking left | Catch Fixation FIXT<br>Catch Fixation FIXL<br>Catch Cover CAPO<br>Catch Nut ECR8<br>Catch Bolt B835<br>Catch Bolt B840<br>Catch Bolt B820<br>Catch Fixture key LARD<br>Catch Bolt B820PT<br>Catch Strap clamps BRIT |
| **7** | Take measuring rod | Take measuring rod |
| **8** | Put down measuring rod | Put down measuring rod |
| **9** | Take component | Catch Fixation FIXT<br>Catch Fixation FIXL<br>Catch Cover CAPO<br>Catch Nut ECR8<br>Catch Bolt B835<br>Catch Bolt B840<br>Catch Bolt B820<br>Catch Fixture key LARD<br>Catch Bolt B820PT<br>Catch Strap clamps BRIT |
| **10** | Put down component | Put down Fixation FIXT<br>Put down Fixation FIXL<br>Put down Cover CAPO<br>Put down Nut ECR8<br>Put down Bolt B835<br>Put down Bolt B840<br>Put down Bolt B820<br>Put down Fixture key LARD<br>Put down Bolt B820PT<br>Put down Strap clamps BRIT<br>Put down Fixation FIXT<br>Put down Fixation FIXL<br>Put down Cover CAPO<br>Put down Nut ECR8<br>Put down Bolt B835<br>Put down Bolt B840<br>Put down Bolt B820<br>Put down Fixture key LARD<br>Put down Bolt B820PT<br>Put down Strap clamps BRIT |
| **11** | Assemble system | Place LARD on Profile P360-1<br>Place FIXA1 on LARD at 160mm<br>Screw FIXA1 with B820<br>Place LARD on P360-1<br>Place CAPO on P360-1<br>Place BRIT1 and BRIT2 on P360-2<br>Place FIXL on P360-2<br>Place FIXA2 on P360-2<br>Screw FIXL with B820PT<br>Screw FIXA2 with B820<br>Place BRIT1 and BRIT2 on P040 |

| ID | Meta action label | Action label |
|---|---|---|
| | | Place P040 on P360-2<br>Screw P040 with B820PT<br>Place LARD on P040 (P360-2)<br>Place FIXA1 on P360-2<br>Place FIXA2 on P360-2<br>Place DI2T on P360-2<br>Place ECR8 on DI2T<br>Screw P360-2 with B835 |
| **12** | Take subsystem | Put down Profile P360-1<br>Put down Lower Part |
| **13** | Put down subsystem | Catch P360-1<br>Catch P360-2 |

The assembly task involves seven operations. Each operation implies around 15 actions. During the entire manipulation, a subject carries out between 100 and 180 actions. Fig. 7 shows the distribution of the number of actions carried out by each subject as well as total actions duration.

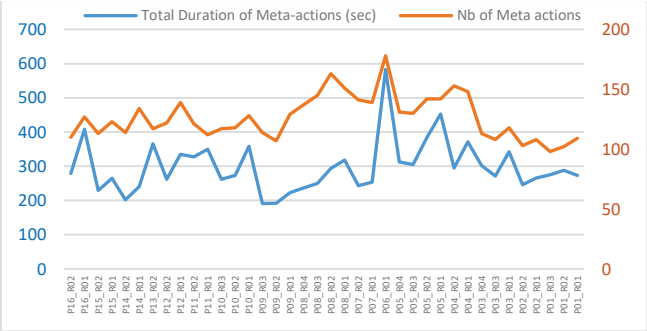


Fig. 7. PivotChart of meta-actions number and duration per subject

In Fig. 8, we represent actions total duration distribution per all subjects. As can be seen, all meta-actions take between 0.5 to 26.9 seconds. As can been seen, the meta-action "Assemble system" has a high variance since it occurs when the operator assembles several parts. Therefore, it can take a few seconds on some operation steps and several tens of seconds on other ones. However, InHARD also provides the action label, which is more precise. In addition, we ensured that there is a large dispersion of difficult actions all along the entire scene, creating variability between actions.

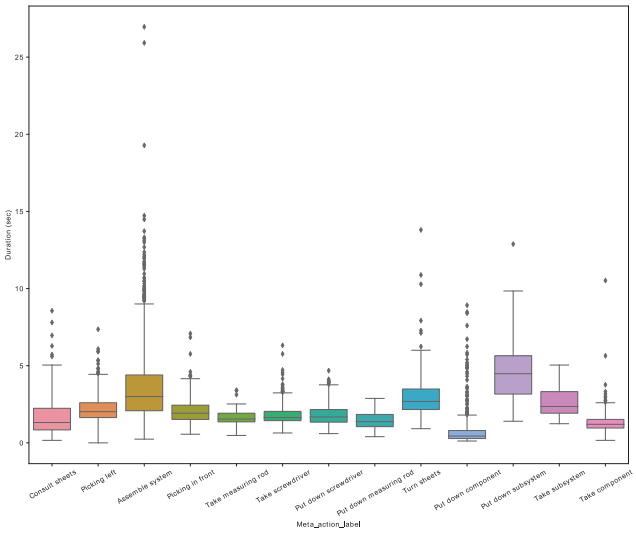


Fig. 8. Meta-actions duration distribution

## IV. Online/Segmented Action Recognition

There are two types of action recognition: segmented (offline) and unsegmented (online) recognition.

- Segmented recognition: given a video sequence $V = \{v_0, \dots, v_{n-1}\}$, segmented recognition aims to determine if a frame $v_t$ at time $t$ matches an action among the pre-set M classes i.e. action recognition is performed after observing the full video sequence.

- Online recognition: it is defined as the detection of the action on the fly within a long video sequence, as early as possible without using any further information, unlike segmented action recognition, which uses the entire video sequence to determine the action.

Online recognition is complicated since, in addition to determining the label of the action, an online action recognition system should also predict the start and end times of each action. To have a better grasp of these techniques, Fig. 9 demonstrate the process of segmented and online action recognition.

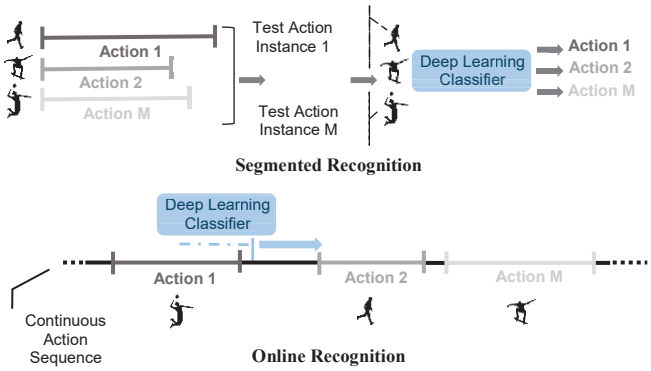


Fig. 9. Segmented/Online Recognition

Despite the fact that online action recognition is vital, there is hardly few works that have been proposed for this matter. Furthermore, Recurrent Neural Network (RNN) has not been well exploited and studied for efficient temporal localization of actions. The majority of the methods aim for a segmented recognition that carries out the recognition after observing the full video sequence. Previous works mostly use a sliding window to localize actions that splits the video sequence into overlapped clips before action recognition/classification is performed on each clip. Such a design has low computational efficiency.

Deep learning methods, such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) have proved a very high performance on temporal dynamics modeling and feature extraction on segmented clips. However, the design of an efficient online action recognition system, which utilizes neural networks for untrimmed data, is not quite well studied.

## V. Experiments and Usage Metrics

In this section, we propose a set of the InHARD dataset usage metrics for future utilization by the research community in order to ensure fair evaluation of the dataset between different approaches. First, we will define training and validation sets as follows:

$S_{train}$ = {P01_R01, P01_R03, P03_R01, P03_R03, P03_R04, P04_R02, P05_R03, P05_R04, P06_R01, P07_R01, P07_R02, P08_R02, P08_R04, P09_R01, P09_R03, P10_R01, P10_R02, P10_R03, P11_R02, P12_R01, P12_R02, P13_R02, P14_R01, P15_R01, P15_R02, P16_R02}.

$S_{val}$ = {P01_R02, P02_R01, P02_R02, P04_R01, P05_R01, P05_R02, P08_R01, P08_R03, P09_R02, P11_R01, P14_R02, P16_R01}.

**Online metrics**

For this metrics, we will use the Meta-actions interval $I_{ma_i}^{(v_j)}$ of the action $i$ in the file $j$ and the estimated interval $\hat{I}_{MA_i}^{(v_j)}$

$$I_{ma_i}^{(v_j)} = [t_{s,i}^{(v_j)}, t_{e,i}^{(v_j)}] \text{ and } \hat{I}_{ma_i}^{(v_j)} = [\hat{t}_{s,i}^{(v_j)}, \hat{t}_{e,i}^{(v_j)}] \quad (1)$$

with $t_{s,i}^{(v_j)}$ and $t_{e,i}^{(v_j)}$ the start and end time ground truth of the meta action $i$ in the file $j$ and $\hat{t}_{s,i}^{(v_j)}$ and $\hat{t}_{e,i}^{(v_j)}$ the estimated ones. Let us also define the Intersection over Union and the associated accuracy as:

$$IoU_{ma_i}^{(v_j)} = \frac{I_{ma_i}^{(v_j)} \cap \hat{I}_{ma_i}^{(v_j)}}{I_{ma_i}^{(v_j)} \cup \hat{I}_{ma_i}^{(v_j)}} \quad (2)$$

$$a_{ma_i}^{(v_j)} = \begin{cases} 1 \text{ if } IoU_{ma_i}^{(v_j)} > \theta = 0.6 \\ 0 \text{ else} \end{cases} \quad (3)$$

Therefore, online accuracy will be computed as:

$$acc = \frac{\sum_j^{\#files} \sum_{i=0}^{\#ma^{(v_j)}} a_{ma_i}^{(v_j)}}{\#ma^{(all)}} \quad (4)$$

with #files represents the number of files in the validation set, $\#ma^{(v_j)}$ the number of meta actions in the file $j$ and $\#ma^{(all)}$ the total number of meta actions in the validation set files.

**Segmented metrics**

For these metrics, a single clip file is generated for each meta action of the training and validation sets. Evaluation metrics are: **accuracy**, **precision**, **recall** and **F1 score**.

## VI. Conclusion and Future Work

In this paper, we introduced a large-scale RGB+Skeleton action recognition dataset named "Industrial Human Action Recognition Dataset (InHARD)". Our dataset includes 4804 different action samples spread over 38 videos collected from 14 industrial action classes. In comparison with existing action recognition datasets, which comprises only daily, mutual, and health-related actions, ours proposes actual industrial actions from real use-case scenarios in an industrial environment hoping that it would help the community to step forward in Human Robot Collaboration in industrial environments. For the best of our knowledge, it is the first industrial action recognition dataset to be proposed. We also proposed a set of the dataset usage metrics for future utilization to allow fair evaluation between the different approaches.

In the future work, we intend to finish developing an end-to-end regression classification LSTM network in order to evaluate our dataset using the proposed metrics.

## Acknowledgment

Parts of this work has been performed within the AGIRH project and has been funded in the framework of the Normandy Regional Council.

## References

[1] S. Robla-Gómez, V. M. Becerra, J. R. Llata, E. Gonzalez-Sarabia, C. Torre-Ferrero and J. Perez-Oria, "Working together: A review on safe human-robot collaboration in industrial environments," *IEEE Access,* vol. 5, pp. 26754-26773, 2017.

[2] A. Bauer, D. Wollherr and M. Buss, "Human-Robot Collaboration: a Survey.," *I. J. Humanoid Robotics,* vol. 5, pp. 47-66, 3 2008.

[3] V. Havard, B. Jeanne, M. Lacomblez and D. Baudry, "Digital twin and virtual reality: a co-simulation environment for design and assessment of industrial workstations," *Production & Manufacturing Research,* vol. 7, pp. 472-489, 2019.

[4] A. A. Malik and A. Bilberg, "Collaborative robots in assembly: A practical approach for tasks distribution," *Procedia Cirp,* vol. 81, pp. 665-670, 2019.

[5] A. Vysocky and P. Novak, "Human-Robot collaboration in industry," *MM Science Journal,* vol. 9, pp. 903-906, 2016.

[6] H. Liu and L. Wang, "Gesture recognition for human-robot collaboration: A review," *International Journal of Industrial Ergonomics,* vol. 68, pp. 355-367, 2018.

[7] P. Wang, W. Li, P. Ogunbona, J. Wan and S. Escalera, "RGB-D-based human motion recognition with deep learning: A survey," *Computer Vision and Image Understanding,* vol. 171, pp. 118-139, 2018.

[8] A. Shahroudy, J. Liu, T.-T. Ng and G. Wang, "NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis," 2016.

[9] W. Li, Z. Zhang and Z. Liu, "Action recognition based on a bag of 3d points," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, 2010.

[10] J. Sung, C. Ponce, B. Selman and A. Saxena, "Human activity detection from RGBD images," in *Workshops at the twenty-fifth AAAI conference on artificial intelligence*, 2011.

[11] B. Ni, G. Wang and P. Moulin, "Rgbd-hudaact: A color-depth video database for human daily activity recognition," in *2011 IEEE international conference on computer vision workshops (ICCV workshops)*, 2011.

[12] J. Wang, Z. Liu, Y. Wu and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[13] Z. Cheng, L. Qin, Y. Ye, Q. Huang and Q. Tian, "Human daily action analysis with multi-view and color-depth data," in *European Conference on Computer Vision*, 2012.

[14] O. Oreifej and Z. Liu, "Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013.

[15] P. Wei, Y. Zhao, N. Zheng and S.-C. Zhu, "Modeling 4d human-object interactions for event and object recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013.

[16] G. Yu, Z. Liu and J. Yuan, "Discriminative orderlet mining for real-time recognition of human-object interaction," in *Asian Conference on Computer Vision*, 2014.

[17] J. Wang, X. Nie, Y. Xia, Y. Wu and S.-C. Zhu, "Cross-view action modeling, learning and recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[18] H. Rahmani, A. Mahmood, D. Q. Huynh and A. Mian, "HOPC: Histogram of oriented principal components of 3D pointclouds for action recognition," in *European conference on computer vision*, 2014.

[19] K. Wang, X. Wang, L. Lin, M. Wang and W. Zuo, "3d human activity recognition with reconfigurable convolutional neural networks," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014.

[20] C. Chen, R. Jafari and N. Kehtarnavaz, "Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *2015 IEEE International conference on image processing (ICIP)*, 2015.

[21] H. Rahmani, A. Mahmood, D. Huynh and A. Mian, "Histogram of oriented principal components for cross-view action recognition," *IEEE transactions on pattern analysis and machine intelligence,* vol. 38, pp. 2430-2443, 2016.

[22] Y. Li, C. Lan, J. Xing, W. Zeng, C. Yuan and J. Liu, "Online Human Action Detection using Joint Classification-Regression Recurrent Neural Networks," 4 2016.

[23] M. Kipp, L. F. Hollen, M. C. Hrstka and F. Zamponi, "Single-Person and Multi-Party 3D Visualizations for Nonverbal Communication Analysis.," in *LREC*, 2014.