Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 September 2020 doi:10.20944/preprints202009.0119.v1

Article

A Mixed Perception Approach for Safe Human-Robot Collaboration in Industrial Automation

Fatemeh Mohammadi Amin [1], Maryam Rezayati [2], Hans Wernher van de Venn [3,*] and Hossein Karimpour [4]

[1] Institute of mechatronics system, Zurich University of Applied Science, Switzerland; mohm@zhaw.ch
[2] Institute of mechatronics system, Zurich University of Applied Science, Switzerland; rzma@zhaw.ch
[3] Institute of mechatronics system, Zurich University of Applied Science, Switzerland; vhns@zhaw.ch
[4] Mechanical engineering department, University of Isfahan, Iran; h.karimpour@eng.ui.ac.ir

* Correspondence: vhns@zhaw.ch; Tel.: +41-58-934-77-89

Abstract: Digital enabled manufacturing systems require high level of automation for fast and low-cost production but should also present flexibility and adaptiveness to varying and dynamic conditions in their environment, including the presence of human beings. This issue is addressed in this work by implementing a reliable system for real-time safe human-robot collaboration based upon the combination of human action and contact type detection systems. Two datasets containing contact and vision data are collected by using different volunteers. The action recognition system classifies human actions using the skeleton representation of the latter when entering the shared workspace and the contact detection system distinguishes between intentional and incidental interactions if a physical contact between human and robot takes place. Two different deep learning networks are used for human action recognition and contact detection which in combination, lead to the enhancement of human safety and an increase of the level of robot awareness about human intentions. The results show a promising path for future AI-driven solutions in safe and productive human–robot collaboration (HRC) in industrial automation.

Keywords: Safe physical Human-Robot Collaboration, collision detection, human action recognition, artificial intelligence, industrial automation

1. Introduction

Recently, human-robot collaboration (HRC) has gained increasing attention, evolving the manufacturing industry from rigid conventional procedures of production to a much more flexible and intelligent way of manufacturing within the frame of the Industry 4.0 paradigm [1,2]. The present industrial need is to develop a new generation of robots that support operators by leveraging tasks in terms of flexibility and cognitive skills requirements [1]. Consequently, the robot becomes a companion or so-called collaborative robot (Cobot) for flexible task accomplishment rather than a preprogrammed slave for repetitive, rigid automation. These robots are expected to actively assist operators in performing complex tasks, with highest priority on human safety in cases humans and robots need to physically cooperate and/or share their workspace [3].

This issue can only be tackled by implementing a cascaded, multi-objective safety system which primarily avoids collisions and in all other cases limits the force impact if a collision-free movement is inevitable. Ensuring safety of humans during collaboration with cobots in physical Human-Robot Interaction (pHRI) is crucial, and one of the main preconditions to answer this challenge is human intention detection [4]. Therefore, the primary goal of this work is to make a step-change in assuring safety in pHRI. The task is divided in two parts, Human Action Recognition (HAR) and contact type detection which will be subsequently investigated. At the end by combining these subsystems, it is considered to attain a reliable safety system which takes advantages of both methodologies.

© 2020 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 September 2020 doi:10.20944/preprints202009.0119.v1

1.1. Human Action Recognition (HAR)

HAR can be used to allow the robot keeping a safe distance to its human counterpart or the environment, ensuring an essential requirement for fulfilling safety in shared workspaces. Recent studies have been concentrated on visual and non-visual perception systems to recognize human actions [5]. One method amongst non-visual approaches consists of using wearable devices [6–11]; Nevertheless, applying this technology as a possible solution for an industrial situation seems at present neither feasible nor comfortable in industrial environments because of restrictions imposed to the operator's movements. On the other hand, active vision-based systems are widely used in such applications for recognizing human gestures and actions but can be significantly affected in their performance in poorly lit scenes or scenarios with large changes in lighting conditions. In general, vision-based approaches consist of two main steps: proper human detection and action classification.

As alluded by recent researches, machine learning methods are essential in recognizing human actions and interpreting them. Some traditional machine learning methods such as Support Vector Machine (SVM) [12–14], Hidden Markov Model (HMM) [15,16], neural networks [17,18] and Gaussian mixture models (GMM) [19,20], have been used for human action detection with a reported accuracy of about 70 to 90 percent. On the other hand, Deep Learning (DL) algorithms prevail as a new generation of machine learning algorithms with significant capabilities in discovering and learning complex underlying patterns from a large amount of data [21]. This algorithm provides a new approach to improve the recognition accuracy of human actions by using depth data provided by time-of-flight, depth or stereo cameras, extracting human location and skeleton pose. DL researchers either use video stream data [22,23], RGB-D images [24–27] or 3D skeleton tracking and joints extraction [28–31] for classification of arbitrary actions. Most of these articles mainly focus on action classification based on domestic scenarios [15,32], only few have an approach for industrial scenarios [33–35] and a restricted number worked on unsupervised human activities in presence of mobile robots [36].

In this work, we use a deep learning approach for real-time human action recognition in an industrial automation scenario. A convolutional analysis is applied on RGB images of the scene in order to model the human motion over the frames by skeleton-based action recognition. The artificial intelligence based human action recognition algorithm provides the core part for distinguishing between collision and intentional contact.

1.2. Contact Type Detection

Toward this goal, at the first step, it is imperative to detect robot contact with human and then distinguish between intentional and incidental contacts, a process called collision detection. Some researchers propose sensor-less procedures for detecting a collision based on the robot dynamics model [37,38], but also through momentum observers [37,39–42], using extended state observers [43], vibration analysis models [44], finite-time disturbance observers [41], energy observers [42], or joint velocity observers [45]. Among these methods, the momentum observer is the most common method of collision detection because it has better performance compared to the other methods, although the disadvantage is that it requires for precise dynamic parameters of the robot [46]. For this reason, machine learning approaches like artificial neural networks [47–49] and deep learning [50,51] have recently been applied for collision detection based on robot sensors stream data due to their fast response and low computational cost.

Deep neural networks are extremely effective in feature extraction and learning complex patterns[52]. Among these deep networks, recurrent neural networks (RNN) like long short-term memory network approaches (LSTM) are frequently used in research for processing time series and sequential data [53–56]. However, the main drawback of this network is the difficulty and time consumption for training in comparison to convolutional neural networks (CNN) [50]. Additionally, current researches showed that CNN has a great performance for image processing in real time situations [22,57–59] where the input data is much more complicated than 1D time series signals.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 September 2020 doi:10.20944/preprints202009.0119.v1

Therefore, in this part, we aim to detect and distinguish between intentional and incidental (collision) human contact by using the convolutional neural network approach to achieve a model free safety system. In the second step, depending on whether the contact is intentional or incidental, the robot should provide an adequate response which in every case ensures the safety of the human operator. At this step, identifying at which link the collision occurred, is an important information for anticipating proper robot reaction [46] which is also considered in the current work.

2. Material and Methods

2.1. Robotic Platform

The accessible platform used throughout this project is a Franka Emika robot (Panda), recognized as a suitable collaborative robot in terms of agility and contact sensitivity. The key features of this robot will be summarized hereafter; It consists of two main parts, arm, and hand. The arm has 7 revolute joints and precise torque sensors (13 bits resolution) at every joint, is driven by high efficiency brushless dc motors, and has the possibility to be controlled by external or internal torque controllers at a 1 kHz frequency. The hand is equipped with a gripper which can securely grasp objects due to a force controller. Generally, the robot has a total weight of approximately 18 kg and can handle payloads up to 3 kg.

2.2. Camera Systems

The vision system is based on a multi-sensor approach using two Kinect V2 cameras for monitoring the environment and tackle the risk of occlusion. The Kinect V2 has a depth camera with resolution of 512 x 424 pixels with a field of view (FoV) of 70.6° x 60° and the color camera has a resolution of 1920 x 1080px with a FoV of 84.1° x 53.8°. So, this sensor as one of the RGB-D Cameras can be used for human body and skeleton detection.

2.3. Standard robot collision detection:

A common collision detection approach is defined as [46]

$$cd(\mu(t)) = \begin{cases} True & \text{if } |\mu(t)| > \epsilon_\mu \\ False & \text{if } |\mu(t)| \leq \epsilon_\mu \end{cases} \quad (1)$$

where *cd* is the collision detection output function which maps the selected monitoring signal $\mu(t)$ into a collision state, TRUE or FALSE. ϵ_μ indicates a threshold parameter, which determines the sensitivity for detecting a collision.

2.4. Deep learning approach

A Convolutional Neural Network (CNN) model performs classification in an end-to-end manner and learns data patterns automatically which is different to the traditional approaches where the classification is done after feature extraction and selection [60]. In this paper, a combination of 3D-CNN for HAR and 1D-CNN for contact type detection has been utilized. The following subsections describe each network separately.

2.4.1. Human Action Recognition Network

Since human actions can be interpreted by analyzing the sequence of human body movements such as arms and legs and placing them in a situational context, the consecutive skeleton images are used as inputs for our 3D-CNN network which was successfully applied for real-time action recognition. In this section, the 3D-CNN which is shown in Figure 1, classifies HAR to five states, namely: Passing, Observation, Dangerous Observation, Interaction, and Fail.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 September 2020 doi:10.20944/preprints202009.0119.v1

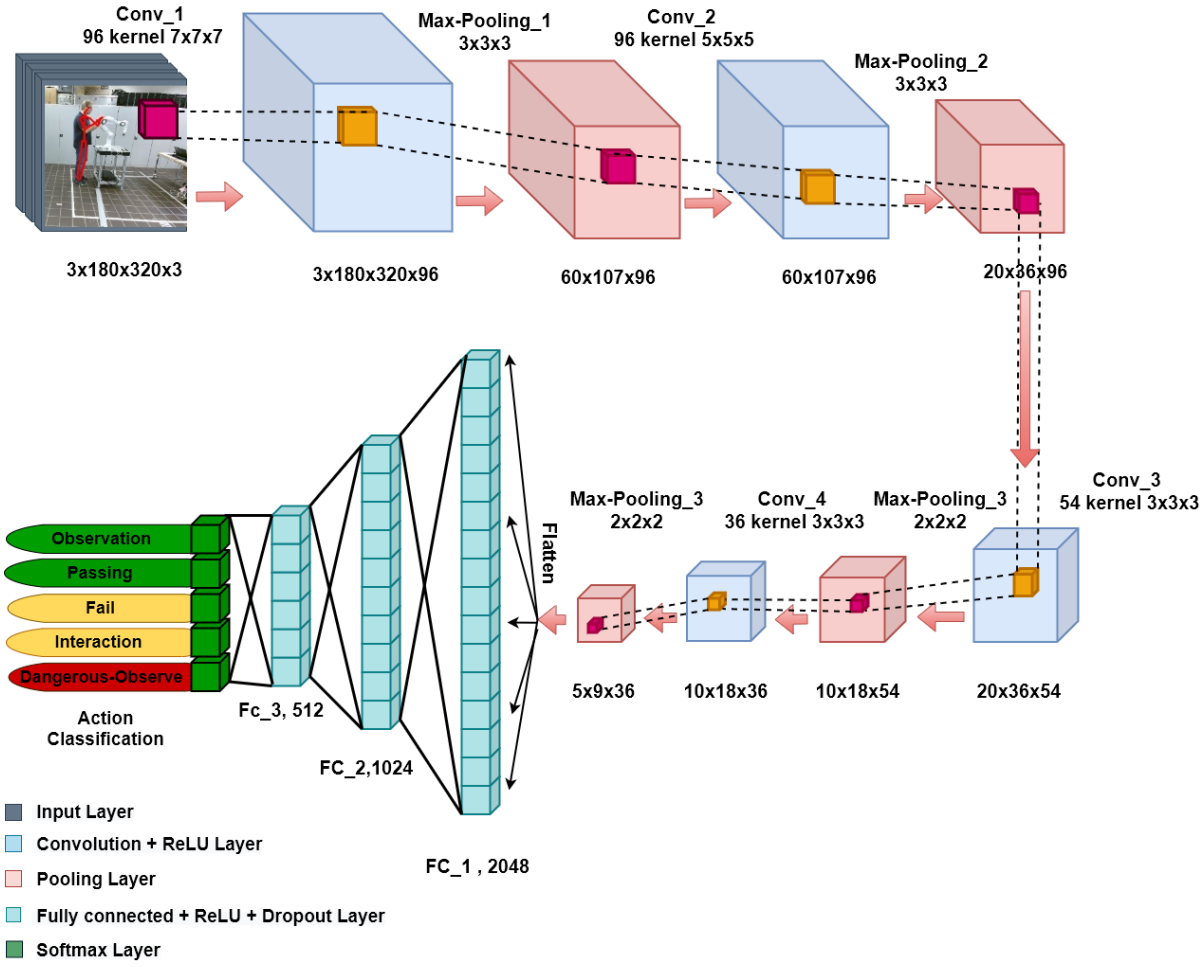

Figure 1: 3D CNN for Human Action Recognition

- Input layer

The input layer has 4 dimensions, $N_{image\text{-}width} \times N_{image\text{-}height} \times N_{channel} \times N_{frame}$. The RGB image of Kinect V2 has a resolution of 1980×1080 pixels which is decreased to 320×180 for reducing the trainable parameters and network complexity. So $N_{image\text{-}width}$, $N_{image\text{-}height}$, and $N_{channel}$ are 320, 180, and 3 respectively. N_{frame} indicates the total number of frames in the image sequence which is 3 in this research.

- Layers

As shown in Figure 3, the proposed CNN is composed of fifteen layers, consisting of 4 convolutional layers, 4 pooling layers, 3 fully connected layers followed by 3 dropout layers and a Softmax layer for predicting actions. Over 10 million parameters must be trained for establishing a map to action recognition.

The input layer is followed by a convolution layer with 96 feature maps of size 7^3. Subsequently, the output is fed to the Rectified Linear Unit (ReLU) activation function. ReLU is the most suitable activation function for this work, as it is specially designed for image processing and it can keep the most important features of the input. In addition, it is easier to train and usually achieves better performance, which is significant for real-time applications. Next layer is a max pooling layer with size and stride of 3. The filter size of the next convolutional layers decreases to 5^3 and 3^3 respectively with stride 1 and zero padding. Then, Max-pooling windows decline to 2^3 with stride of 2. The output of the last pooling layer is flattened out for the fully connected layer input. The fully connected layers consist of 2024, 1024, 512 neurons, respectively. The last step is to use a Softmax level for activity recognition.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 September 2020 doi:10.20944/preprints202009.0119.v1

2.4.2. Contact Detection Network

For contact detection, a deep network which is shown in Figure 2 is proposed. In this scheme, a 1D-CNN which is a multi-layered architecture with each layer consisting of few one-dimensional convolution filters, is used. It includes one network for classification of 5 states, which were defined as:

- No-Contact: no contact is detected within the specified sensitivity
- Intentional_Link5: an intentional contact at robot link 5 is detected
- Incidental_Link5: a collision at robot link 5 is detected
- Intentional_Link6: an intentional contact at robot link 6 is detected
- Incidental_Link6: a collision at robot link 6 is detected

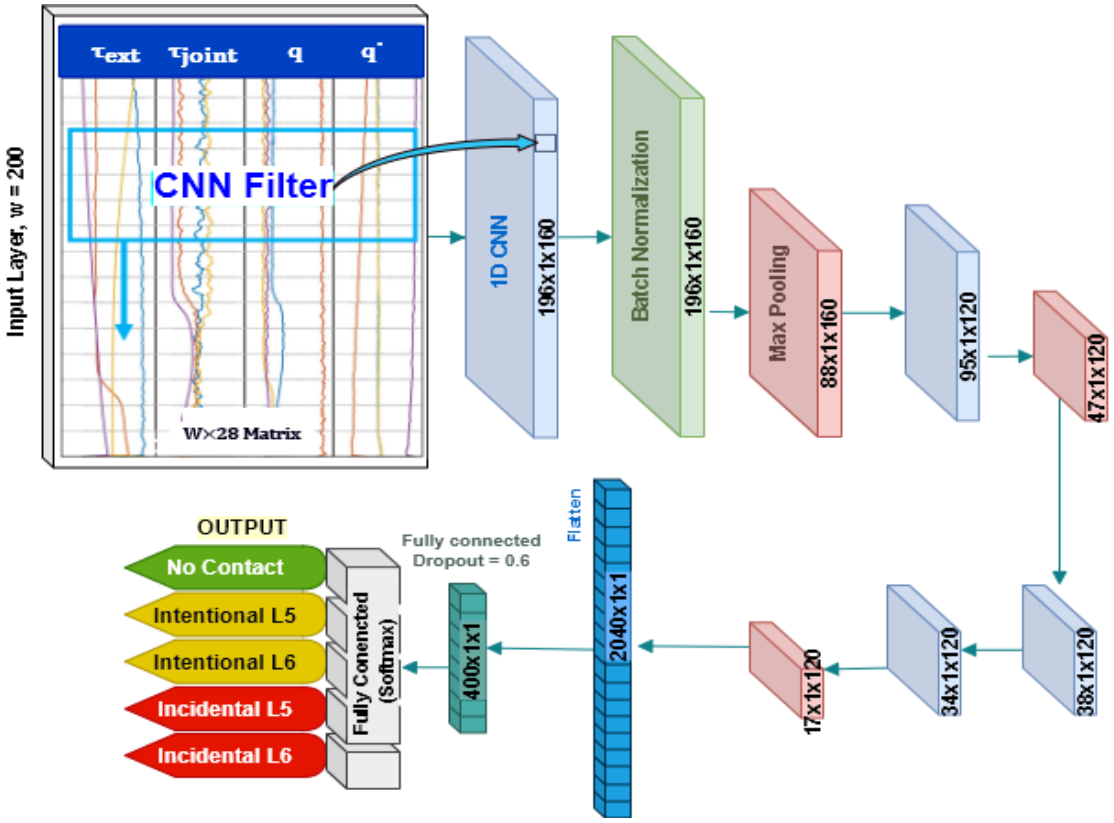

Figure 2: Contact Detection Network Diagram

- Input vector

In this paper, the input vector represents a time series of robot data as

$$x = \begin{bmatrix} \tau_J^0 & \tau_{ext}^0 & q^0 & \dot{q}^0 \\ \tau_J^1 & \tau_{ext}^1 & q^1 & \dot{q}^1 \\ \vdots & \vdots & \vdots & \vdots \\ \tau_J^W & \tau_{ext}^W & q^W & \dot{q}^W \end{bmatrix} \quad (2)$$

And

$$\tau_J^i = \left[\tau_{J1}^i \; \tau_{J2}^i \; \tau_{J3}^i \; \tau_{J4}^i \; \tau_{J5}^i \; \tau_{J6}^i \; \tau_{J7}^i \right] \quad (3)$$

$$\tau_{ext}^i = \left[\tau_{ext1}^i \; \tau_{ext2}^i \; \tau_{ext3}^i \; \tau_{ext4}^i \; \tau_{ext5}^i \; \tau_{ext6}^i \; \tau_{ext7}^i \right] \quad (4)$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 September 2020 doi:10.20944/preprints202009.0119.v1

$$q^i = \left[q_1^i \; q_2^i \; q_3^i \; q_4^i \; q_5^i \; q_6^i \; q_7^i \right] \quad (5)$$

$$\dot{q}^i = \left[\dot{q}_1^i \; \dot{q}_2^i \; \dot{q}_3^i \; \dot{q}_4^i \; \dot{q}_5^i \; \dot{q}_6^i \; \dot{q}_7^i \right] \quad (6)$$

where τ_J, τ_{ext}, q, and $\dot{q}$ indicate joint torque, external torque, joint position, and joint velocity, respectively. W is the size of a window over the collected signals which stores time-domain samples as an independent instance for training the proposed models. Hence, the input vector is W×28, and in this research, by selecting 100, 200, and 300 samples for W, three different networks were trained to compare the influence of this parameter.

- Layers (generalization)

As shown in Figure 2, the designed CNN is composed of eleven layers. In the first layer of this model, the convolution process maps the data with 160 filters. The kernel size of this layer is optimally considered 5 to obtain a faster and sensitive enough human contact status; a parameter higher than 5 led to an insufficient network's response as it is more influenced by past data rather than near to present data. To normalize the data and avoid overfitting, especially due to the different maximum force patterns of every human, a Batch Normalization is used in the second layer. Furthermore, the size of all max pooling layers is chosen as 2, and ReLU function is considered as the activation function, due to reasons already mentioned before.

2.5. Central Decision Maker (CDM)

To determine the level of safety for the human cooperator, a Central Decision Maker system (CDM) is designed by combining the results of the two parts, ARN and CDN, using rules which is shown in Figure 3. Different human actions and contact types are categorized in three level of safety, namely safe, caution, and danger, with color-code as green, yellow, and red.

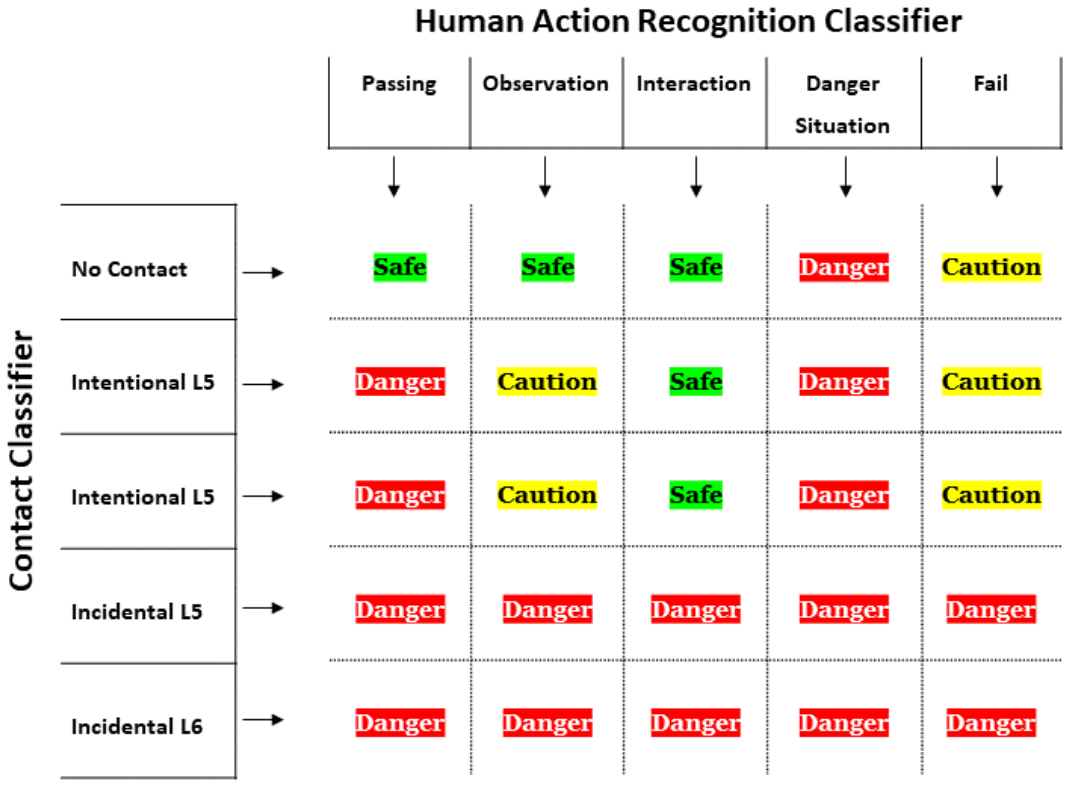

Figure 3: Central Decision Maker rules

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 September 2020 doi:10.20944/preprints202009.0119.v1

2.6. Data Collection

2.6.1. Human Action Recognition

The HAR data is collected simultaneously from different views by two Kinect V2 cameras recording the scene of an operator moving next to a robot performing repetitive motions. As Kinect V2 library in Linux is not precise and does not project human skeleton in RGB images, depth coordinates are converted to RGB coordinates as follows

$$x_{rgb} = x_d \times \frac{PD_{xrgb}}{PD_{xd}} + \frac{C_{xrgb} \times PD_{xd} - C_{xd} \times PD_{xrgb}}{PD_{xrgb} \times PD_{xd}} \quad (7)$$

$$y_{rgb} = y_d \times \frac{PD_{yrgb}}{PD_{yd}} + \frac{C_{yrgb} \times PD_{yd} - C_{yd} \times PD_{yrgb}}{PD_{yrgb} \times PD_{yd}} \quad (8)$$

where (C_{xrgb}, C_{yrgb}) and (C_{xd}, C_{yd}) are RGB and depth image centers, respectively. PD shows the number of pixels per degree for depth and RGB images, respectively equal to 7×7 and 22 ×20 [61][62].Then, the RGB images, which are supplemented with the skeleton representation in each frame, are collected as dataset. The collection rate by considering the required time for saving the images was 22 frames/second. Both cameras start collecting data once the human operator enters the environment. The collected images are then sorted into 5 different categories and labeled accordingly.

2.6.2. Contact Detection

The data acquired at the robot joints during a predefined motion were collected as shown in Figure 2, in a collision-free state and during interaction with the operator, at a sampling rate of 200Hz (one sample per 5ms). Then, a frame of W-window with 200ms latency passed through the entire data gathered, preparing it to be used as training data for the input layer of the designed network. Thanks to the default cartesian contact detection ability of the Panda robot, those contact data is used as a trigger to stop recording data after contact occurrence. Consequently, the last W-samples of each collision trial data is considered as input for training the network. For assuring comprehensiveness of the gathered data, each trial is repeated 10 times with different scenes, including touched links, direction of motion, line of collision with the human operator, and contact type (intentional or incidental). Additionally, each sample is labeled according to the mentioned sequence.

2.7. Training hardware and API setup

In the training of a network by using Graphic Processor Units (GPU), memory plays an important role in reducing the training time. In this research, a powerful computer with NVIDIA Quadro P5000 GPU, Intel Xeon W-2155 CPUs, and 64 GB of RAM is employed for modeling and training the CNN networks using the Keras library of TensorFlow. To enable CUDA and GPU-acceleration computing, a GPU version of TensorFlow is used, and in consequence, the training process is speeded up. The total runtime of the vision network trained with 30,000 image sequences was about 12 hours for 150 epochs, while it was less than 5 minutes for training contact networks.

2.8. Real time interface

The real time interface for collecting dataset and implementing the trained network on the system was provided by Robotics Operating System (ROS) on Ubuntu 18.04 LTS. Figure 4 shows the hardware and software structure used in this work. Two computers execute the vision networks for each camera separately and publish the action states at the rate of 200 Hz on ROS. Furthermore,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 September 2020 doi:10.20944/preprints202009.0119.v1

CDN and CDM are executed on another pc at the same rate, connected to the robot controller for receiving the robot torque, velocity and position data of joints 5 and 6.

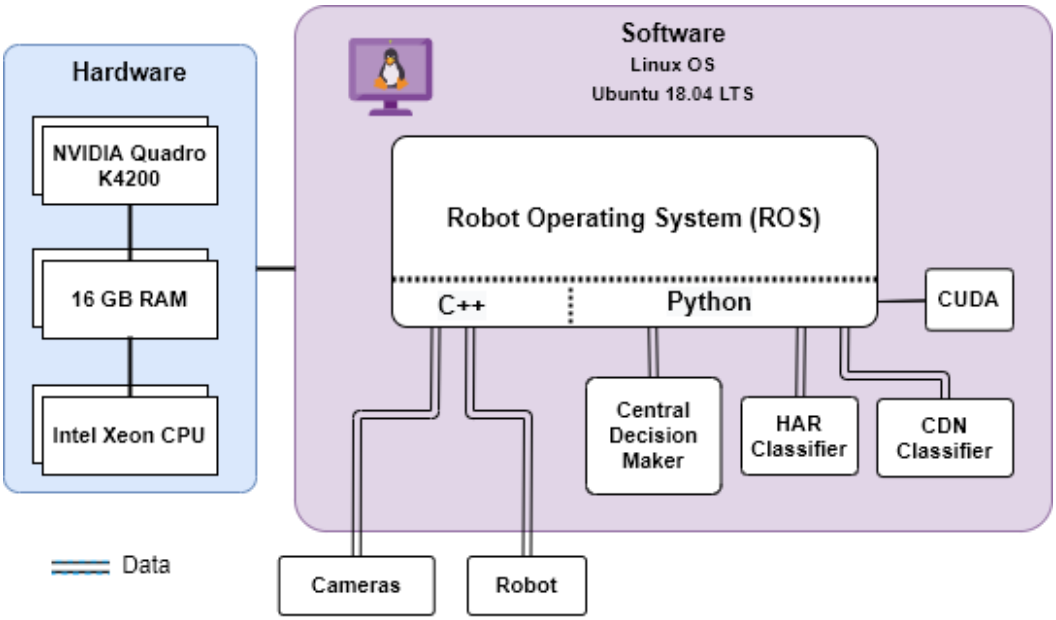

Figure 4: Real-time interface of complex system

3. Results

In order to evaluate the performance of the proposed system, the following metrics is used. A first evaluation consists of an offline testing, for which the results are compared based on the key figures Precision and Recall, defined as follows:

$$Precision = \frac{tp}{tp + fp} \tag{9}$$

$$Recall = \frac{tp}{tp + fn} \tag{10}$$

Where t_p is the amount of the predicted true positive samples, f_p represents the amount of the predicted false positive samples and f_n is the count of predicted false positive classes. Accuracy calculation follows later.

The second evaluation is based on real-time testing; The tests have shown promising results in early trials, the following YouTube video gives an impression of the performance.
https://www.youtube.com/watch?v=ED_wH9BFJck

3.1. Dataset

Regarding the vision category, the dataset consisting of 33050 images is divided into 5 classes, including Interaction, Observation, Passing, Fail, and Dangerous Observation, Figure 5 representing the different possible actions of a human operator during robot operation. Contact detection dataset [63] with 1114 samples is subdivided into 5 classes, namely No-contact, Intentional_Link5, Intentional_Link6, Incidental_Link5, Incidental_Link6, which determine the contact state on the last two links including their respective type, incidental or intentional.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 September 2020 doi:10.20944/preprints202009.0119.v1

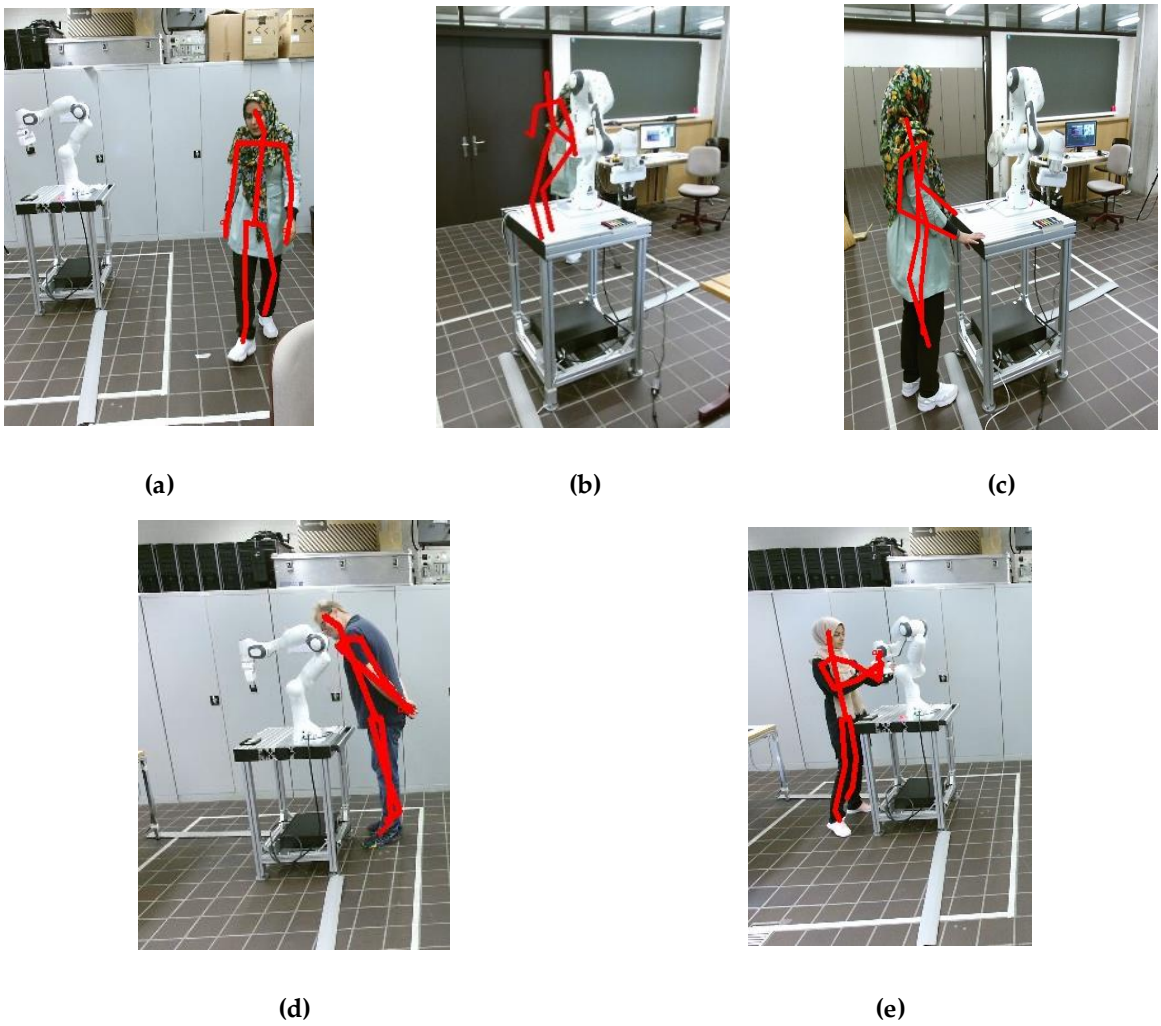

Figure 5: Type of human actions: (a) Passing: Operator is just passing by, without paying attention to the robot; (b) Fail: Blind spots or occlusion of the visual field may happen for a camera, in this situation the second camera takes over detection; (c) Observation: Operator enters the working zone, without any interaction intention and stands next to the robot; (d) Dangerous Observation: Operator proximity is too close, especially his head is at danger of collision with the robot; (e) Interaction: Operator enters the working zone and prepares to work with the robot.

3.2. Comparison between Networks

3.2.1. Action recognition

For optimizing efficiency in HAR, two different networks, 2D and 3D, were tested, the latter indicating a significant outcome in both real-time and off-line testing cases. These two networks are compared with respect to the results of 150 training epochs, in Table 1 and Table 2**.** As it is clear, the 3D network shows superiority in terms of Accuracy, Precision and Recall.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 September 2020 doi:10.20944/preprints202009.0119.v1

Table 1: Precision and Recall of two trained networks for Human Action Recognition

Network	2D		3D	
	Precision	**Recall**	**Precision**	**Recall**
Observation	0.99	0.99	**1.00**	**1.00**
Interaction	1.00	1.00	**1.00**	**1.00**
Passing	1.00	1.00	**1.00**	**1.00**
Fail	1.00	1.00	**1.00**	**1.00**
Dangerous Observation	0.98	0.96	**0.98**	**0.99**
Accuracy	0.9956		**0.9972**	

Table 2: Confusion Matrix for different classes in HRC

Network		2D					3D				
		Observation	Interaction	Passing	Fail	Dangerous Observation	Observation	Interaction	Passing	Fail	Dangerous Observation
True Labels	Observation	3696	7	2	0	5	3751	6	2	1	7
	Interaction	13	4130	0	0	1	8	4030	0	0	0
	Passing	2	0	1145	0	0	1	0	1160	0	0
	Fail	0	0	0	593	0	0	0	0	588	0
	Dangerous Observation	12	1	0	0	313	2	0	0	0	359

3.2.2. Contact detection

To evaluate the influence of the size of the sampling window (w) on the precision of the trained networks, three different size dimensions of 100, 200, and 300 unity are selected, corresponding to 0.5, 1, 1.5 seconds of sampling period duration. 70% of the dataset are selected for training and 30% for testing. Each network is trained with 300 epochs and the results are shown in **Table 3** and **Error! Reference source not found.**. Window size of 200 and 300 unities provide a good precision for identifying the contact status, in contrast to w=100 which is not satisfactory. Furthermore, by comparing the result of 200-window and 300-window networks, 200-window network has a better precision and recall.

Table 3: Precision and recall of trained networks for contact detection with different window size

w	100	200	300	100	200	300
		Precision			Recall	
No-Contact	0.94	0.99	0.98	0.94	1.00	1.00
Intentional_Link5	0.74	0.91	0.89	0.84	0.91	0.84
Intentional_Link6	0.68	0.97	0.91	0.64	0.90	0.91
Incidental_Link5	0.61	0.89	0.83	0.61	0.93	0.89
Incidental_Link6	0.69	0.96	0.96	0.57	0.96	0.93
Accuracy	0.78	0.96	0.93			

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 September 2020 doi:10.20944/preprints202009.0119.v1

Table 4: Confusion matrix of trained networks for contact detection with different window size

Window Size		100					200					300				
		No-Contact	Intentional_Link5	Intentional_Link6	Incidental_Link5	Incidental_Link6on	No-Contact	Intentional_Link5	Intentional_Link6	Incidental_Link5	Incidental_Link6	No-Contact	Intentional_Link5	Intentional_Link6	Incidental_Link5	Incidental_Link6
True Labels	No-Contact	166	0	9	0	1	342	0	3	0	1	167	0	3	0	0
	Intentional_Link5	0	86	12	19	0	0	93	4	4	1	0	86	5	5	1
	Intentional_Link6	8	1	59	2	17	0	3	83	0	0	0	5	84	0	3
	Incidental_Link5	0	15	1	33	5	0	6	0	50	0	0	10	0	48	0
	Incidental_Link6	3	0	11	0	31	0	0	2	0	52	0	1	0	1	50

4. Discussion

Human Robot Collaboration has recently gained a lot of interest and received many contributions on both theoretical and practical aspects, including sensor development [64], design of robust and adaptive controllers [65,66], learning robots force-sensitive manipulation skills [67], human interfaces [68,69] and so on. In addition, some companies attempted to introduce collaborative robots so that HRC become more suited to enter in manufacturing applications and production lines. However, Cobots available on the market have limited payload/speed capacities because of safety concerns which limits HRC application to some light tasks with low productivity.

On the other hand, according to the norms for HRC operations [70], it is not essential to observe a strict design or limit the power of operations if the human safety factor can be ensured in all its aspects; In this regard, an intelligent safety system has been developed in this research to detect hazardous situations, also assessing Human Intention Awareness (HIA) whether in physical contact with the robot or not. As a result, our studies show that the different forms of collaboration such as coexistence, cooperation, etc. with their different safety requirements can be reduced to a single scenario. In this safety scenario, the robot reacts by being able to detect human intention and thus ensuring safety in all work situations. Thus, a smart robot will take care of the safety of humans from entering the shared workspace to physical interaction in order to jointly accomplish a task.

Another advantage of this system is that the robot would be smart enough to take care about safety norms depending on the conditions and consequently, could operate at an optimum speed during HRC applications. In other words, current safety requirements in most cases stop or drastically slow down the robot when human enters a shared workspace. However, with the proposed safety system, based on the robot's awareness, it is possible to implement a reasonable trade-off between security and productivity, which will be discussed in more detail in our future research.

5. Conclusion

The efficiency of safety and productivity of Cobots in HRC can be improved if they can easily recognize complex human actions and differentiate between multitude types of contact. In this paper, a safety system composed of visual and physical interaction detection systems is proposed to improve the productivity in HRC applications by making the robot aware of human intentions with the ability to distinguish between intentional and incidental contact. In a first step, the system is purposed to detect human intention once he enters the workspace and just in case of hazardous situations, the robot would adapt or stop accordingly which can lead to higher productivity. On the other hand, if there is any contact between robot and human, the system would decide about the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 September 2020 doi:10.20944/preprints202009.0119.v1

final situation (collision or intentional contact) based on the defined rules, and by considering both HAR and contact detection system outputs.

Acknowledgments: This research was a part of an international project between Zurich University of Applied Science ZHAW and University of Isfahan. Authors acknowledge the Leading House of South Asia and Iran at Zurich University of Applied Science for funding this research.

The persons recognizable in the pictures have consented to private and / or commercial use - publication, distribution, use, processing and transfer - in digital and print form by the photographer or by third parties.

7. References

1. Robla-Gomez, S.; Becerra, V.M.; Llata, J.R.; Gonzalez-Sarabia, E.; Torre-Ferrero, C.; Perez-Oria, J. Working Together: A Review on Safe Human-Robot Collaboration in Industrial Environments. *IEEE Access* **2017**, *5*, 26754–26773, doi:10.1109/ACCESS.2017.2773127.
2. Nikolakis, N.; Maratos, V.; Makris, S. A cyber physical system (CPS) approach for safe human-robot collaboration in a shared workplace. *Robot. Comput. Integr. Manuf.* **2019**, *56*, 233–243, doi:10.1016/j.rcim.2018.10.003.
3. Villani, V.; Pini, F.; Leali, F.; Secchi, C. Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications. *Mechatronics* **2018**, *55*, 248–266, doi:10.1016/j.mechatronics.2018.02.009.
4. Losey, D.P.; McDonald, C.G.; Battaglia, E.; O'Malley, M.K. A Review of Intent Detection, Arbitration, and Communication Aspects of Shared Control for Physical Human–Robot Interaction. *Appl. Mech. Rev.* **2018**, *70*, doi:10.1115/1.4039145.
5. Zhang, H.B.; Zhang, Y.X.; Zhong, B.; Lei, Q.; Yang, L.; Du, J.X.; Chen, D.S. A comprehensive survey of vision-based human action recognition methods. *Sensors (Switzerland)* **2019**, *19*, 1–20, doi:10.3390/s19051005.
6. Otim, T.; Díez, L.E.; Bahillo, A.; Lopez-Iturri, P.; Falcone, F. Effects of the Body Wearable Sensor Position on the UWB Localization Accuracy. *Electronics* **2019**, *8*, 1351, doi:10.3390/electronics8111351.
7. Moschetti, A.; Cavallo, F.; Esposito, D.; Penders, J.; Di Nuovo, A. Wearable sensors for human-robot walking together. *Robotics* **2019**, *8*, 38, doi:10.3390/ROBOTICS8020038.
8. Otim, T.; Díez, L.E.; Bahillo, A.; Lopez-Iturri, P.; Falcone, F. A Comparison of Human Body Wearable Sensor Positions for UWB-based Indoor Localization. In Proceedings of the 10th International Conference Indoor Positioning Indoor Navigat; Piza, Italy, 2019; pp. 165–171.
9. Rosati, S.; Balestra, G.; Knaflitz, M. Comparison of different sets of features for human activity recognition by wearable sensors. *Sensors (Switzerland)* **2018**, *18*, doi:10.3390/s18124189.
10. Xu, Z.; Zhao, J.; Yu, Y.; Zeng, H. Improved 1D-CNNs for behavior recognition using wearable sensor network. *Comput. Commun.* **2020**, *151*, 165–171, doi:10.1016/j.comcom.2020.01.012.
11. Xia, C.; Sugiura, Y. Wearable Accelerometer Optimal Positions for Human Motion Recognition. In Proceedings of the LifeTech 2020 - 2020 IEEE 2nd Global Conference on Life Sciences and Technologies; Institute of Electrical and Electronics Engineers Inc., 2020; pp. 19–20.
12. Qian, H.; Mao, Y.; Xiang, W.; Wang, Z. Recognition of human activities using SVM multi-class classifier. *Pattern Recognit. Lett.* **2010**, *31*, 100–111, doi:10.1016/j.patrec.2009.09.019.
13. Reddy, K.K.; Shah, M. Recognizing 50 human action categories of web videos. *Mach. Vis. Appl.* **2013**, *24*, 971–981, doi:10.1007/s00138-012-0450-4.
14. Manosha Chathuramali, K.G.; Rodrigo, R. Faster human activity recognition with SVM. In Proceedings of the International Conference on Advances in ICT for Emerging Regions, ICTer 2012 - Conference

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 September 2020 doi:10.20944/preprints202009.0119.v1

Proceedings; 2012; pp. 197–203.

15. Berg, J.; Reckordt, T.; Richter, C.; Reinhart, G. Action recognition in assembly for human-robot-cooperation using hidden Markov Models. *Procedia CIRP* **2018**, *76*, 205–210, doi:10.1016/j.procir.2018.02.029.
16. Hoang Le Uyen Thuc; Shian-Ru Ke; Jenq-Neng Hwang; Pham Van Tuan; Truong Ngoc Chau Quasi-periodic action recognition from monocular videos via 3D human models and cydic HMMs. In Proceedings of the The 2012 International Conference on Advanced Technologies for Communications; IEEE, 2012; pp. 110–113.
17. Hasan, H.; Abdul-Kareem, S. Static hand gesture recognition using neural networks. *Artif. Intell. Rev.* 2014, *41*, 147–181.
18. Sharma, S.; Modi, S.; Rana, P.S.; Bhattacharya, J. Hand gesture recognition using Gaussian threshold and different SVM kernels. In Proceedings of the Communications in Computer and Information Science; Springer Verlag, 2018; Vol. 906, pp. 138–147.
19. Cho, W.H.; Kim, S.K.; Park, S.Y. Human action recognition using hybrid method of hidden Markov model and Dirichlet process Gaussian mixture model. *Adv. Sci. Lett.* **2017**, *23*, 1652–1655, doi:10.1166/asl.2017.8599.
20. Piyathilaka, L.; Kodagoda, S. Gaussian mixture based HMM for human daily activity recognition using 3D skeleton features. In Proceedings of the Proceedings of the 2013 IEEE 8th Conference on Industrial Electronics and Applications, ICIEA 2013; 2013; pp. 567–572.
21. Wang, P.; Liu, H.; Wang, L.; Gao, R.X. Deep learning-based human motion recognition for predictive context-aware human-robot collaboration. *CIRP Ann. - Manuf. Technol.* **2018**, *67*, 17–20, doi:10.1016/j.cirp.2018.04.066.
22. Ullah, A.; Ahmad, J.; Muhammad, K.; Sajjad, M.; Baik, S.W. Action Recognition in Video Sequences using Deep Bi-Directional LSTM with CNN Features. *IEEE Access* **2017**, *6*, 1155–1166, doi:10.1109/ACCESS.2017.2778011.
23. Zhao, Y.; Man, K.L.; Smith, J.; Siddique, K.; Guan, S.U. Improved two-stream model for human action recognition. *Eurasip J. Image Video Process.* **2020**, *2020*, 1–9, doi:10.1186/s13640-020-00501-x.
24. Wang, P.; Li, W.; Ogunbona, P.; Wan, J.; Escalera, S. RGB-D-based human motion recognition with deep learning: A survey. *Comput. Vis. Image Underst.* **2018**, *171*, 118–139, doi:10.1016/j.cviu.2018.04.007.
25. Gao, J.; Yang, Z.; Sun, C.; Chen, K.; Nevatia, R. TURN TAP: Temporal Unit Regression Network for Temporal Action Proposals. In Proceedings of the International Conference on Computer Vision; 2017.
26. Gu, Y.; Ye, X.; Sheng, W.; Ou, Y.; Li, Y. Multiple stream deep learning model for human action recognition. *Image Vis. Comput.* **2020**, *93*, 103818, doi:10.1016/j.imavis.2019.10.004.
27. Srihari, D.; Kishore, · P V V; Kiran Kumar, · E; Anil Kumar, · D; Teja, · M; Kumar, K.; Prasad, · M V D; Raghava Prasad, C. A four-stream ConvNet based on spatial and depth flow for human action classification using RGB-D data. *Multimed. Tools Appl.* **2020**, *79*, 11723–11746, doi:10.1007/s11042-019-08588-9.
28. Yan, S.; Xiong, Y.; Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the 32nd AAAI Conference on Artificial Intelligence, AAAI 2018; 2018; pp. 7444–7452.
29. Si, C.; Jing, Y.; Wang, W.; Wang, L.; Tan, T. Skeleton-Based Action Recognition with Spatial Reasoning and Temporal Stack Learning. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **2018**, *11205 LNCS*, 106–121, doi:10.1007/978-3-030-01246-5_7.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 September 2020 doi:10.20944/preprints202009.0119.v1

30. Cheng, J. Skeleton-Based Action Recognition with Directed Graph Neural Networks. *Comput. Vis. Pattern Recognit.* **2019**, 7912–7921.
31. Trăscău, M.; Nan, M.; Florea, A.M. Spatio-temporal features in action recognition using 3D skeletal joints. *Sensors (Switzerland)* **2019**, *19*, doi:10.3390/s19020423.
32. Olatunji, I.E. Human Activity Recognition for Mobile Robot. *J. Phys. Conf. Ser.* **2018**, *1069*, 4–7, doi:10.1088/1742-6596/1069/1/012148.
33. Akkaladevi, S.C.; Heindl, C. Action recognition for human robot interaction in industrial applications. In Proceedings of the 2015 IEEE International Conference on Computer Graphics, Vision and Information Security, CGVIS 2015; 2016; pp. 94–99.
34. Ding, Y.; Xu, W.; Liu, Z.; Zhou, Z.; Pham, D.T. Robotic Task Oriented Knowledge Graph for Human-Robot Collaboration in Disassembly. *Procedia CIRP* **2019**, *83*, 105–110, doi:10.1016/j.procir.2019.03.121.
35. Roitberg, A.; Perzylo, A.; Somani, N.; Giuliani, M.; Rickert, M.; Knoll, A. Human activity recognition in the context of industrial human-robot interaction. *2014 Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. APSIPA 2014* **2014**, doi:10.1109/APSIPA.2014.7041588.
36. Duckworth, P.; Hogg, D.C.; Cohn, A.G. Unsupervised human activity analysis for intelligent mobile robots. *Artif. Intell.* **2019**, *270*, 67–92, doi:10.1016/j.artint.2018.12.005.
37. Cao, P.; Gan, Y.; Dai, X. Model-based sensorless robot collision detection under model uncertainties with a fast dynamics identification. *Int. J. Adv. Robot. Syst.* **2019**, *16*, 172988141985371, doi:10.1177/1729881419853713.
38. Haddadin, S.; Albu-Schaffer, A.; De Luca, A.; Hirzinger, G. Collision Detection and Reaction: A Contribution to Safe Physical Human-Robot Interaction. In Proceedings of the 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems; IEEE, 2008; pp. 3356–3363.
39. de Luca, A.; Mattone, R. Sensorless Robot Collision Detection and Hybrid Force/Motion Control. In Proceedings of the Proceedings of the 2005 IEEE International Conference on Robotics and Automation; IEEE; pp. 999–1004.
40. Xiao, J.; Zhang, Q.; Hong, Y.; Wang, G.; Zeng, F. Collision detection algorithm for collaborative robots considering joint friction. *Int. J. Adv. Robot. Syst.* **2018**, *15*, 172988141878899, doi:10.1177/1729881418788992.
41. Cao, P.; Gan, Y.; Dai, X. Finite-Time Disturbance Observer for Robotic Manipulators. *Sensors* **2019**, *19*, 1943, doi:10.3390/s19081943.
42. Luca, A.; Albu-Schaffer, A.; Haddadin, S.; Hirzinger, G. Collision Detection and Safe Reaction with the DLR-III Lightweight Manipulator Arm. In Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems; IEEE, 2006; pp. 1623–1630.
43. Ren, T.; Dong, Y.; Wu, D.; Chen, K. Collision detection and identification for robot manipulators based on extended state observer. *Control Eng. Pract.* **2018**, *79*, 144–153, doi:10.1016/J.CONENGPRAC.2018.07.004.
44. Min, F.; Wang, G.; Liu, N. Collision Detection and Identification on Robot Manipulators Based on Vibration Analysis. *Sensors* **2019**, *19*, 1080, doi:10.3390/s19051080.
45. Haddadin, S. *Towards safe robots : approaching Asimov's 1st law*; ISBN 9783642403088.
46. Haddadin, S.; De Luca, A.; Albu-Schaffer, A. Robot Collisions: A Survey on Detection, Isolation, and Identification. *IEEE Trans. Robot.* **2017**, *33*, 1292–1312, doi:10.1109/TRO.2017.2723903.
47. Sharkawy, A.-N.; Koustoumpardis, P.N.; Aspragathos, N. Neural Network Design for Manipulator

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 September 2020 doi:10.20944/preprints202009.0119.v1

Collision Detection Based Only on the Joint Position Sensors. *Robotica* **2019**, 1–19, doi:10.1017/S0263574719000985.

48. Sharkawy, A.-N.; Koustoumpardis, P.N.; Aspragathos, N. Human–robot collisions detection for safe human–robot interaction using one multi-input–output neural network. *Soft Comput.* **2019**, 1–33, doi:10.1007/s00500-019-04306-7.
49. Liu, Z.; Hao, J. Intention Recognition in Physical Human-Robot Interaction Based on Radial Basis Function Neural Network Available online: https://www.hindawi.com/journals/jr/2019/4141269/ (accessed on Jun 8, 2020).
50. Heo, Y.J.; Kim, D.; Lee, W.; Kim, H.; Park, J.; Chung, W.K. Collision Detection for Industrial Collaborative Robots: A Deep Learning Approach. *IEEE Robot. Autom. Lett.* **2019**, *4*, 740–746, doi:10.1109/LRA.2019.2893400.
51. Dine, K.M. El; Sanchez, J.; Ramón, J.A.C.; Mezouar, Y.; Fauroux, J.-C. Force-Torque Sensor Disturbance Observer using Deep Learning. **2018**.
52. Pham, C.; Nguyen-Thai, S.; Tran-Quang, H.; Tran, S.; Vu, H.; Tran, T.H.; Le, T.L. SensCapsNet: Deep Neural Network for Non-Obtrusive Sensing Based Human Activity Recognition. *IEEE Access* **2020**, *8*, 86934–86946, doi:10.1109/ACCESS.2020.2991731.
53. Gao, X.; Shi, M.; Song, X.; Zhang, C.; Zhang, H. Recurrent neural networks for real-time prediction of TBM operating parameters. *Autom. Constr.* **2019**, *98*, 225–235, doi:10.1016/J.AUTCON.2018.11.013.
54. Yang, K.; Wang, X.; Quddus, M.; Yu, R. Predicting Real-Time Crash Risk on Urban Expressways Using Recurrent Neural Network. **2019**.
55. Masood, S.; Srivastava, A.; Thuwal, H.C.; Ahmad, M. Real-Time Sign Language Gesture (Word) Recognition from Video Sequences Using CNN and RNN. In; Springer, Singapore, 2018; pp. 623–632.
56. Li, S.; Li, W.; Cook, C.; Zhu, C.; Gao, Y. Independently Recurrent Neural Network (IndRNN): Building a Longer and Deeper RNN 2018, 5457–5466.
57. Zhang, B.; Wang, L.; Wang, Z.; Qiao, Y.; Wang, H. Real-Time Action Recognition With Enhanced Motion Vector CNNs 2016, 2718–2726.
58. Jin, C.-B.; Li, S.; Do, T.D.; Kim, H. Real-Time Human Action Recognition Using CNN Over Temporal Images for Static Video Surveillance Cameras. In; Springer, Cham, 2015; pp. 330–339.
59. Pathak, D.; El-Sharkawy, M. Architecturally Compressed CNN: An Embedded Realtime Classifier (NXP Bluebox2.0 with RTMaps). In Proceedings of the 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC); IEEE, 2019; pp. 0331–0336.
60. Ullah, I.; Hussain, M.; Qazi, E.-H.; Aboalsamh, H. An automated system for epilepsy detection using EEG brain signals based on deep learning approach. *Expert Syst. Appl.* **2018**, *107*, 61–71, doi:10.1016/J.ESWA.2018.04.021.
61. Kinect for Windows v2.
62. Kim, C.; Yun, S.; Jung, S.W.; Won, C.S. Color and depth image correspondence for Kinect v2. *Adv. Multimed. Ubiquitous Eng.* **2015**, *352*, 111–116, doi:10.1007/978-3-662-47487-7_17.
63. Rezayati, M.; van de Venn, H.W. Dataset: Collision Detection in Physical Human Robot Interaction. **2020**, *1*, doi:10.17632/CTW2256PHB.1.
64. Cirillo, A.; Cirillo, P.; De Maria, G.; Natale, C.; Pirozzi, S. A Distributed Tactile Sensor for Intuitive Human-Robot Interfacing. **2017**, doi:10.1155/2017/1357061.
65. Khoramshahi, M.; Billard, A. A dynamical system approach to task-adaptation in physical human–robot interaction. *Auton. Robots* **2019**, *43*, 927–946, doi:10.1007/s10514-018-9764-z.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 September 2020 doi:10.20944/preprints202009.0119.v1

66. Xiong, G.L.; Chen, H.C.; Xiong, P.W.; Liang, F.Y. Cartesian Impedance Control for Physical Human–Robot Interaction Using Virtual Decomposition Control Approach. *Iran. J. Sci. Technol. - Trans. Mech. Eng.* **2019**, *43*, 983–994, doi:10.1007/s40997-018-0208-3.

67. Johannsmeier, L.; Gerchow, M.; Haddadin, S. *A framework for robot manipulation: Skill formalism, meta learning and adaptive control*; 2019; Vol. 2019-May; ISBN 9781538660263.

68. Yang, C.; Zeng, C.; Liang, P.; Li, Z.; Li, R.; Su, C.Y. Interface Design of a Physical Human-Robot Interaction System for Human Impedance Adaptive Skill Transfer. *IEEE Trans. Autom. Sci. Eng.* **2018**, *15*, 329–340, doi:10.1109/TASE.2017.2743000.

69. Weistroffer, V.; Paljic, A.; Callebert, L.; Fuchs, P. *A Methodology to Assess the Acceptability of Human-Robot Collaboration Using Virtual Reality*; 2013;

70. Normung, E.K. für TECHNICAL SPECIFICATION ISO / TS Robots and robotic devices —. **2016**, *2016*.