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; However, this presence of workers in the shared workspace with robots decreases the productivity as the robot is not aware about the human position and intention which leads to concerns about the human safety. This issue is addressed in this work by designing a reliable safety monitoring system for collaborative robots (Cobots). The main idea here is, to significantly enhance safety by a combination of recognition of human actions using visual perception and at the same time interpreting physical human-robot contact by tactile perception. 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 cobot takes place. Two different deep learning networks are used for human action recognition and contact detection which in combination, are expected to lead to the enhancement of human safety and an increase of the level of cobot perception 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

As the manufacturing industry evolves from rigid conventional procedures of production to a much more flexible and intelligent way of automation within the frame of the Industry 4.0 paradigm, human-robot collaboration (HRC) has gained rising attention [1,2]. To increase manufacturing flexibility, the present industrial need is to develop a new generation of robots that are able to interact with human and support operators by leveraging tasks in terms of 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. It is expected that cobots actively assist operators in performing complex tasks, with highest priority on human safety in cases humans and cobots need to physically cooperate and/or share their workspace [3]. This is problematic because the current settings of cobots do not provide an adequate perception of human presence in the shared workspace. Although there are some safety monitoring systems [4-7], they can only provide a real or virtual fence for the cobot to stop or slow down when an object, including a human being, enters the defined safety zone. However, this reduces productivity in two ways as follows:

1. It is not possible to differentiate between people and other objects that enter the workspace of the cobot. Therefore, the speed is always reduced regardless of the object.
2. It is also not possible to differentiate whether an interaction with the robot should really take place or not; this also always forces a maximum reduction in speed.

This issue can only be tackled by implementing a cascaded, multi-objective safety system which primarily recognizes human actions and detects human-robot contact [4] to percept human intention in order to avoid collisions. Therefore, the primary goal of this work is to conduct a step-change in safety for HRC in enhancing the perception of cobots by providing visual and tactile feedback to the robot from which it is able to interpret the human intention. The task is divided into two parts, Human Action Recognition (HAR) using visual perception and contact type detection using tactile perception which will be subsequently investigated. Finally, by combining these subsystems, it is considered to attain a more reliable and intelligent safety system which takes advantage of considerably enhanced robot perceptional abilities.

1.1 Human Action Recognition (HAR)

Based on the existing safety regulation related to HRC applications and by inspiring from human perception and cognition ability in different situations, adding the visual perception to the cobot can enhance HRC performance. Nevertheless, the main challenge is how cobots are able to adapt to human behavior. HAR as part of visual perception plays a crucial role in overcoming this challenge and increasing productivity and safety. HAR can be used to allow the cobot keeping a safe distance with 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 that it will 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. In general, vision based HAR 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) [19, 20], Neural Networks [21, 22] and Gaussian Mixture Models (GMM) [23, 24], 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 [26, 27], RGB-D images [28 – 31] or 3D skeleton tracking and joints extraction [32 – 35] for classification of arbitrary actions. Among different types of Deep networks, Convolutional Neural Networks (CNN) stand for a popular approach which can be represented as 2D or 3D network in action recognition but still needs a large set of labeled data for training and contains many layers. The first 3D CNN for HAR has been introduced by [36 – 38] providing an average accuracy of 91 percent. Recent researches based on 3D CNN techniques [39 – 42] have obtained a high performance on the KTH dataset [39] in comparison to 2D CNN networks [43 – 46]. Yet, the maximum accuracy of these researches is reported to be at 98.5 percent but being not capable of classifying in real-time. Also, most of these articles mainly focus on action classification based on domestic scenarios [19, 47], only few have an approach for industrial scenarios [48 – 50] and a restricted number works on unsupervised human activities in presence of mobile robots [49]. Thus, there is a need to introduce a fast and more precise network for HAR in industrial applications which can be presented as a new 3D network architecture by considering an outperforming result in action classification.

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

In more and more HRC applications, there is a need of having direct, physical collaboration between human and cobot, physical Human-Robot Collaboration (pHRC) due to an unmatched degree of flexibility in the execution of various tasks. Indeed, when a cobot is performing its task, it should be aware of its contact with human. Also, from a cobots point of view, the type of this contact is not immediate obvious, due to the fact that the cobot cannot distinguish whether a human gets in contact with the robot incidentally or intentionally, when a collaborative task is executed; Therefore, it is important that the cobot needs to percept human contact with deeper understanding. Towards this goal, it is imperative firstly, to detect the human-robot contact and secondly, 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 [52, 53], but through momentum observers [52, 54 – 57], using extended state observers [56], vibration analysis models [57], finite-time disturbance observers [54], energy observers [55], or joint velocity observers [58]. Among these methods, the momentum observer is the most common method of collision detection because it provides better performance compared to the other methods, although the disadvantage is, that it requires precise dynamic parameters of the robot [59]. For this reason, machine learning approaches like artificial neural networks [62 – 64] and deep learning [65, 66] have recently been applied for collision detection based on robot sensors´ stream data due to their performance in modeling the uncertain systems with lower analytics effort.

Deep neural networks are extremely effective in feature extraction and learning complex patterns [65]. Recurrent neural networks (RNN) like long short-term memory network approaches (LSTM) are frequently used in researches for processing time series and sequential data [66 – 69]. However, the main drawback of this network is the difficulty and time consumption for training in comparison to convolutional neural networks (CNN) [70]. Additionally, current researches showed that CNN has a great performance for image processing in real time situations [22, 71 – 73] where the input data is much more complicated than 1D time series signals. As proposed in [70], a 1D-CNN, named CollisionNet, has a proper potential in detecting collision, although only incidental contacts have been considered. Moreover, depending on whether the contact is intentional or incidental, the cobot should provide an adequate response which in every case ensures the safety of the human operator. At this step, identifying the cobot link where the collision occurred, is an important information for anticipating proper robot reaction which needs to be considered in contact perception [59].

With this background and considering the fact that contact properties´ patterns of incidental and intentional states are different according to the contacted link, we aim to use supervised learning, convolutional neural network, to have a model-free contact detection. Indeed, not only does the proposed system detect the contact, which in other papers [59, 70, 74, 75] is named collision detection, it can also recognize the types of contact, incidental or intentional, provide information about contacted link, and consequently increase the robot awareness and perception about human intention during physical contact.

2. Material and Methods

2.1. Mixed perception Terminology and design

We hypothesize that combining two types of perception, visual and tactile, in a Mixed Perception approach can enhance the safety of human during collaborating with a robot by additional information to the robot's perception spectrum. It is easy to imagine that a robot then would be able to see and feel a human in its immediate vicinity at the same time. Using visual perception, a robot can notice:

- a human entering the shared workspace (Passing),
- a human observing its tasks when he/she is near to the robot and wants to supervise the robot task (Observation),
- a dangerous situation when the human is not in a proper situation to do collaboration or observation which can threaten human safety (Dangerous Observation),
- a human interaction when the human is close to robot and doing the collaboration (Interaction).

However, it is difficult using a pure vision-based approach to distinguish between dangerous observation and interaction and to differentiate between incidental and intended contact types not only for a machine but also for a human. Therefore, at this stage, considering both types of perception, vision and haptics, is of significance. As indicated above, this approach is able to increase the safety and can be like a supervisory unit to the vision part as the latter can fail due to occlusion effects.

To support our hypothesis, we first choose the approach of developing two separate networks for human action and contact recognition, which meet the requirements for human-robot collaboration and real-time capability. These networks will be examined and discussed with regard to their appropriateness and their results. As a first step, we want to determine in this paper whether a logical correlation of the outputs of the two networks is theoretically able to provide a reasonable expansion of the perception spectrum of a robot for human-robot collaboration. We want to find out what the additional information content is and how it can specifically be used to further increase the safety and with that possibly also additional performance parameters of HRC solutions like short cycle time, low downtime, high efficiency and high productivity. The concrete merging of the two networks in a common application represents an additional stage of our investigations, which is not subject of this work. The results of the present investigation however shall provide evidence that the use of AI in robotics is able to open up significant new possibilities and enables robots to achieve their operational objectives in close cooperation with humans. Enhanced perceptional abilities of robots are future key features to shift the existing technological limits and open up new fields of application in industry and beyond.

2.2. 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.

2.3. Camera Systems

The vision system is based on a multi-sensor approach using two Kinect V2 cameras for monitoring the environment to 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.

2.4. Standard robot collision detection:

A common collision detection approach is defined as [59]

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

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

2.5. 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 [76]. 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.

2.5.1 Human Action Recognition Network

Since human actions can be interpreted by analysing the sequence of human body movements involving 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. These categories are based on the most feasible situations which may occur during human-robot collaboration:

- Passing: a human operator occasionally needs to enter the robot's workspace which is speciefied due to the fix position of the robot but without any intention to actively intervene the task execution of the robot.
- Interaction: a human operator wants to actively intervene the robot's task execution, which can be the case due to correct a TCP path or to help the robot if it gets stuck.
- Observation: the robot is working on its own and a human operator is about to observe and check the working process from within the robot's workspace.
- Dangerous Observation: the robot is working on its own and a human operator is about to observe the working process. Due to the proximity of exposed body parts (head and upper extremities) to the robot in the shared workspace, there is a high potential of life-threatening injury in case of a collision.
- Fail: one or all system cameras are not able to detect the human operator in the workspace due to occlusion by the robot itself or other artefacts in the working area.

Input layer

The input layer has 4 dimensions, $N_{channel} \times N_{image\text{-}height} \times N_{image\text{-}width} \times N_{frame}$. 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 $N_{channel}$, $N_{image\text{-}height}$, and $N_{image\text{-}width}$ are 3, 180, and 320 respectively. N_{frame} indicates the total number of frames in the image sequence which is 3 in this research.

Layers

As shown in Figure 1, the proposed CNN is composed of fifteen layers, consisting of four convolutional layers, four pooling layers, three fully connected layers followed by three dropout

layers and a Softmax layer for predicting actions. Convolutional layers utilize for feature extraction by applying filters and pooling layers specifically use to reduce the dimensionality of the input. This layer performs based on the specific function, such as max pooling or average pooling, which extracts the maximum or medium value in a particular region. Fully connected layers are like a neural network for learning non-linear features as represented by the output of convolutional layers. In addition, dropout layers as a regularization layer try to remove overfitting in the network. Over 10 million parameters have to be trained for establishing a map to action recognition.

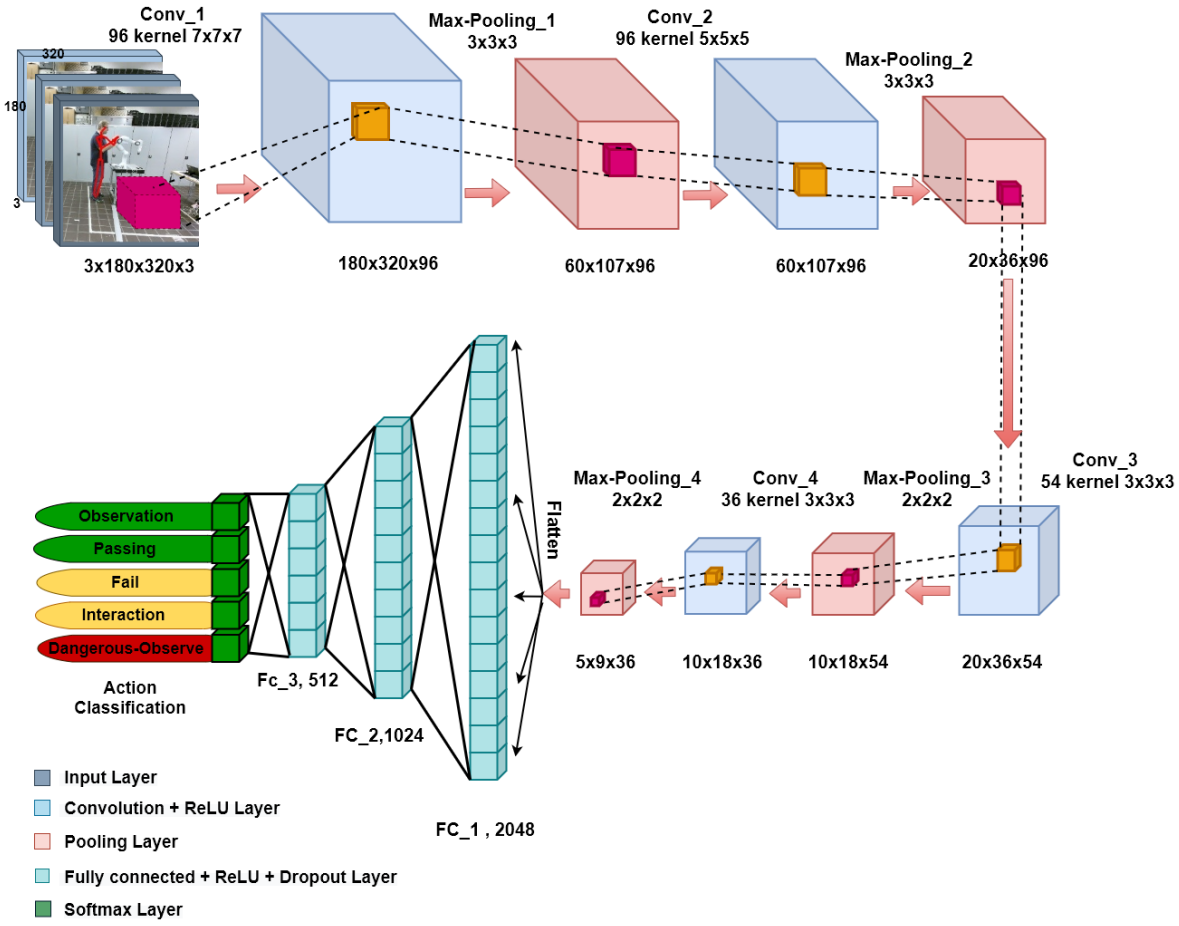


Figure 1: 3D CNN for Human Action Recognition

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 specifically 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.

2.5.2 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. In this research, just two links of the robot which are more likely to be used as contact points during physical human-robot collaboration, considered which indeed, does

not influence the general approach used in this paper; Therefore, 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

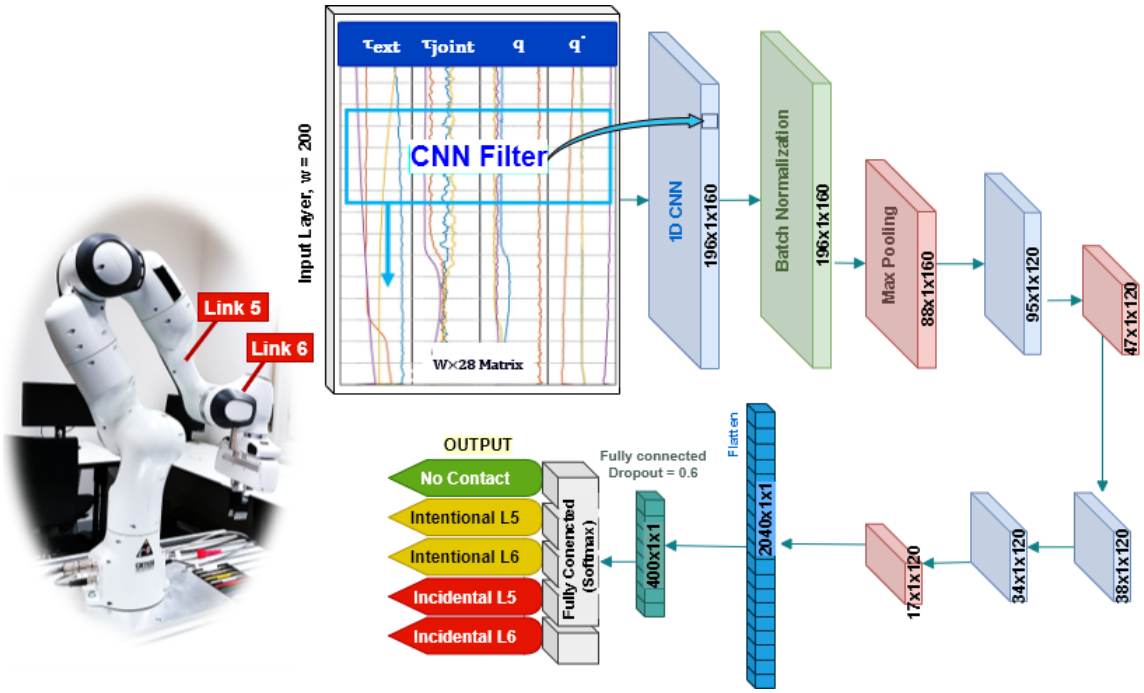


Figure 2. Contact Detection Network Diagram.

Input vector

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

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

and

$$\tau_J^i = \left[\tau_{j1}^i \; \tau_{j2}^i \; \tau_{j3}^i \; \tau_{j4}^i \; \tau_{j5}^i \; \tau_{j6}^i \; \tau_{j7}^i \right] \tag{3}$$

$$\tau_{ext}^i = \left[\tau_{ext\,1}^i \; \tau_{ext\,2}^i \; \tau_{ext\,3}^i \; \tau_{ext\,4}^i \; \tau_{ext\,5}^i \; \tau_{ext\,6}^i \; \tau_{ext\,7}^i \right] \tag{4}$$

$$q^i = \left[q_1^i \; q_2^i \; q_3^i \; q_4^i \; q_5^i \; q_6^i \; q_7^i \right] \tag{5}$$

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

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

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.

2.6 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. The human skeleton is detected using the Kinect library based on the random Forest decision method [77]. As the Kinect V2 library in Linux is not precise and does not project human skeleton in RGB images, the 3D joint position in depth coordination was extracted and 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 [78, 79]. Then, the RGB images, which are supplemented with the skeleton representation in each frame, are collected as dataset. The sample 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 while it is assumed that the robot is stationary in a structured environment. The collected images are then sorted into 5 different categories and labeled accordingly based on the skeleton position and configuration and with respect to the fixed base position of the robot.

2.6.2 Contact Detection

The data acquired at the robot joints during a predefined motion with a speed of 0.5 m/s were collected in three states, contact-free, during interaction with the operator, and collision-like contacts, at a sampling rate of 200Hz (one sample per 5ms). In this part, collecting collision-like contact data is challenging as the dedicated operator induces the collision intentionally [70]; However, the collision can be considered to happen like in a normal situation where the human is standing with no motion and the robot is performing its task, while the impact happens. Indeed, a data analysis shows, that it can be clearly distinguished from object and intentional contacts and therefore can be used at least as similar samples of real collision data. 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 the dataset and implementing the trained network on the system was provided by Robotics Operating System (ROS) on Ubuntu 18.04 LTS. Figure 3 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, 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.

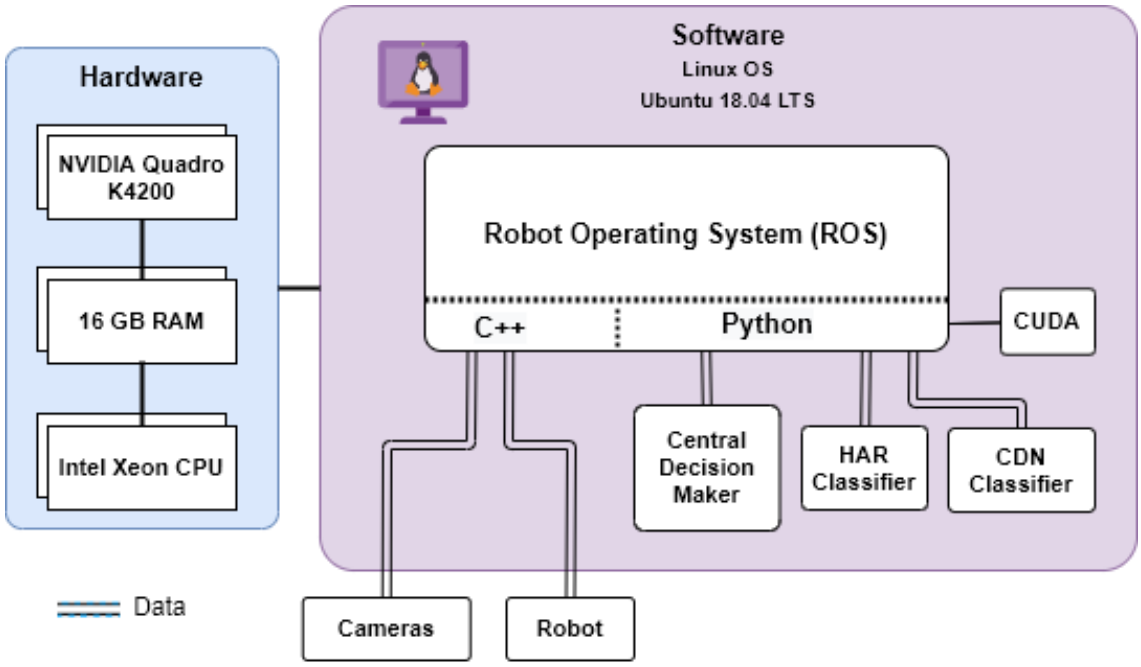


Figure 3: Real-time interface of complex system

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, recall, and accuracy, defined as follows:

$$Precision = \frac{tp}{tp + fp} \quad (9)$$

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

$$Accuracy = \frac{tp + tn}{tp + fn + tn + fp} \tag{11}$$

where t_p is the amount of the predicted true positive samples, t_n is the number of data points labeled as negative correctly, f_p represents the amount of the predicted false positive samples, and f_n is the count of predicted false positive classes.

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 (Due to safety reasons the velocity of the robot is reduced to an amount which is considered to be safe according to ISO 10218).
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 4 representing the different possible actions of a human operator during robot operation. The contact detection dataset [80] 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.

3.2 Comparison between Networks

3.2.1 Human action recognition

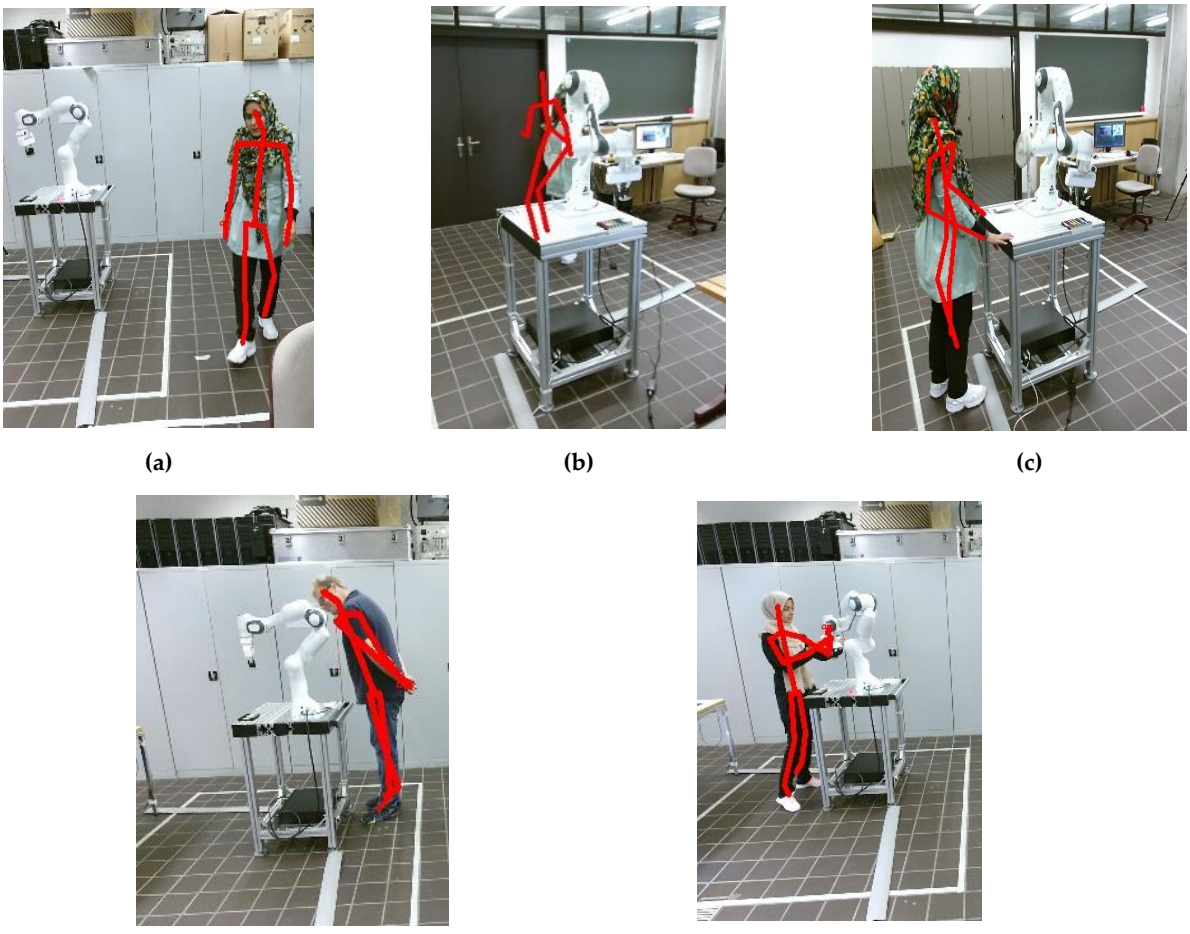
For optimizing efficiency in HAR, two different networks, 2D and 3D, were tested, the latter indicating a significant outcome in both real-time and off-line testing cases. These two networks are compared with respect to the results of 150 training epochs, in Table 1**.** The confusion matrix can be considered as a good measurement to deliver the overall performance in the multi-category classification systems. As it is shown in Table 2, each row of table represents the actual label, and each column indicate the predicted labels, which can also show the number of fail prediction in every class. As shown in Table 1, both networks have promising result in classifying “Interaction”, “Passing”, and “Fail” states. However, these networks have lower, but sufficient, accuracy in classifying the “Dangerous Observation” category due to the lack of 3rd dimensional (depth) information in the network input. By considering the confusion matrix shown in Table 2, it is obvious that the networks did not precisely distinguish between “Interaction”, “Observation” and “Dangerous Observation” caused by the similarity of these three classes. With regard to the condition of the experimental setup where the location of cameras and robot base are fixed, the current approach has enough accuracy but for a real industry case, we need to add a true 3D representation of the human skeleton and the robot arm in our network input.

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



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

Figure 4: 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.

3.2.2 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 is selected for training and 30% for testing. Each network is trained with 300 epochs and the results are shown in Table 3 and Table 4. 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 the 200-window and 300-window networks, the 200-window network provides 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			

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	242	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

3.2.3. Mixed Perception Safety monitoring

Every perception system designed separately to detect human intention according to Figure 5 (a, b), is regarded as a preliminary condition for the mixed perception system shown in Figure 5 (c). As shown in Figure 5, for proper safety monitoring, the robot is programmed to categorize human safety into three levels, Safe, Caution, and Danger, with its respective color-codes green/yellow, orange, and red. Safe level consist of two states, indicating whether the cobot has physical contact with human (yellow) or not (green). Considering only visual perception or only tactile perception in

determining the safety level does not provide sufficient information compared to the mixed perception system. For instance, in Green Safe state of mixed perception, the robot can have a higher speed and in consequence, increased productivity, while in the other perception systems, Green Safe does not give this confidence to the robot to be faster, consequently it should be more conservative about possible collisions. Thus, this higher information content can increase human safety and the robot's productivity of pHRC systems. Already a simple logical composition of the results (Figure 5 (c)) shows a significantly higher information content and thus a possible increase in safety and productivity in human-robot collaboration. However, it might be, that the mixed perception approach will have multiple effects on the safety of HRC. Therefore, we will examine in detail the influence of the two subsystems on the overall performance and quality of the entire system at a later stage.

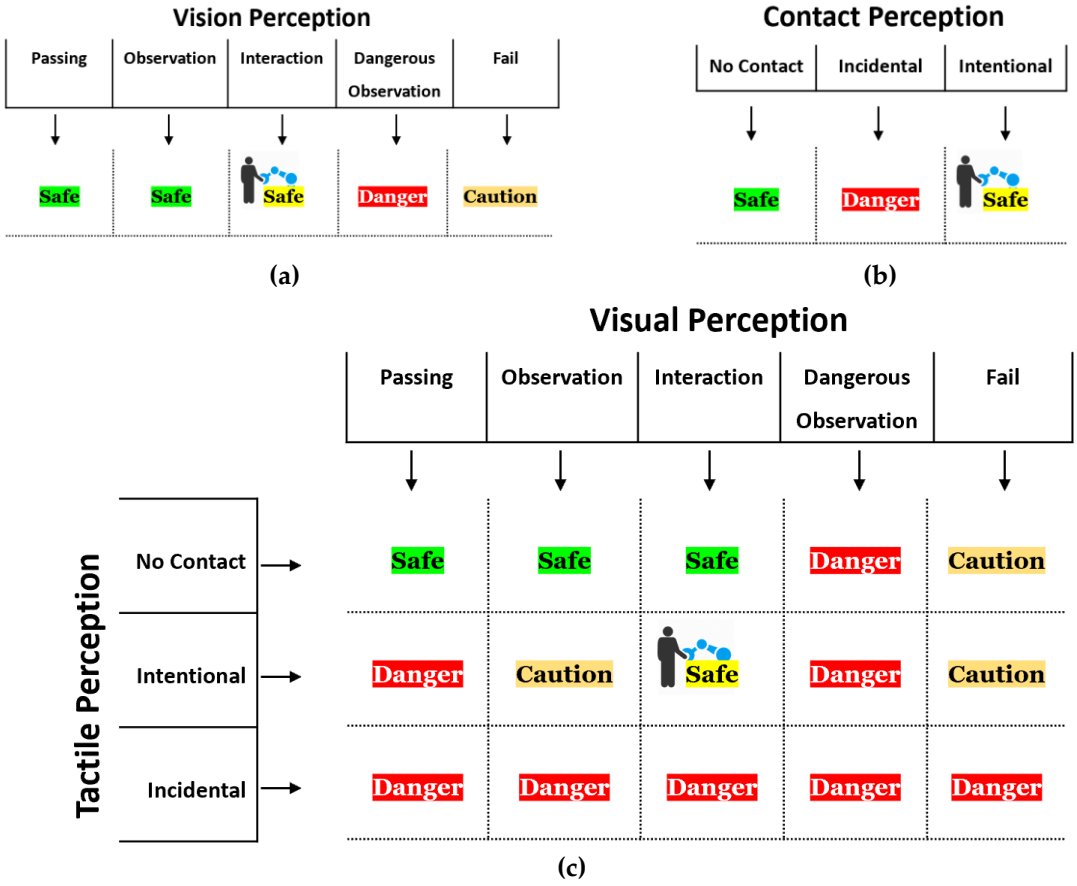


Figure 5: Safety perception spectrum in a) Visual Perception, b) Contact Perception, c) Mixed Perception safety systems

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 [81], design of robust and adaptive controllers [82, 83], learning robots force-sensitive manipulation skills [84], human interfaces [85, 86] and the like. Besides, some companies attempt to introduce collaborative robots in order HRC to become more suited to enter 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 very limited productivity. On the other hand, according to the norms for HRC operations [87], it is not essential to observe a strict design or to limit the power of operations if human safety can be ensured in all its aspects. In this regard, an intelligent safety system as the mixed perception approach has been proposed in this research to detect hazardous situations to take care of the human safety from entering the shared workspace to physical interaction in order to jointly accomplish a task by taking advantage of visual

and tactile perceptions. Visual perception detects human actions in the shared workspace; Meanwhile, tactile perception identifies human-robot contacts. One of the relevant researches in human action recognition focusing on industrial assembly application is mentioned in [88]. By taking advantage of RGB image and 3D CNN network, the authors of the mentioned paper classified human action during assembly and achieved 82% accuracy [88], while our visual perception system shows higher accuracy of 99.7% by adding human skeleton to the RGB series as the network input. Although our skeleton detection using Kinect library can be slightly affected by lighting conditions, it detects the human skeleton in near 30 FPS, which is essential for fast human detection in real-time HRC applications [89]. Indeed, using deep learning approaches like OpenPose [90] and AlphaPose[91] can omit lightening problems [92]; However, their respective detection rate are 22 [90] and 23 FPS [91], which needs more researches to be faster and applicable in safety monitoring systems. Besides, among contact detection approaches in the state of the art, there are two similar works investigating collision detection using 1D-CNN. The authors of [93] compared both approaches, CollisionNet [70] and FMA [93], where the accuracy was 88% and 90%, respectively, featuring a detection delay of 200ms [93]; While our procedure in tactile perception (what is called collision detection in the state of the art literature [59, 70, 74, 75]) reached 99% with 80ms detection delay. For detecting contact type and robot joint, the accuracy was higher than 89% up to 96% which in turn needs more research to achieve a higher accuracy.

In this study, combining the result of both abovementioned intelligent systems is presented using a safety perception spectrum to examine the potential of the mixed perception approach in safety monitoring of collaborative workspaces. The result shows that even with a simple combination of both systems, the performance of safety monitoring can be improved as each system separately doesn't have enough perception of the collaborative workspace. Furthermore, this research suggests that the different forms of collaboration such as coexistence, cooperation, etc. with their different safety requirements can be reduced to a single scenario using mixed perception as the robot would be able to "see" humans and "percept" external contacts.

As a result of this safety scenario, the robot reacts by being able to detect human intention, determining human safety level, and thus ensuring safety in all work situations. Another advantage of the proposed 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 robots' awareness using the presented mixed perception approach, it is possible to implement a reasonable trade-off between safety and productivity, which will be discussed in more detail in our future research.

In this research, there are two limitations concerning data collection: the collision occurred intentionally, and we did not gather data when the human and/or the robot move at high speed, which can be extremely dangerous for the human operator. As can be prooved, the speed of the robot has an insignificant influence on the result, since the model has learned the dynamics of the robot in the presence or absence of human contact with normalized input data. On the other hand, if the human operator wants to grab the robot at high speed with the intention of working with it, this could be classified as a collision by the model due to its clear difference between contactless and intentional data patterns. However, this only increases the false positive error of the collision class (i.e. this would then be mistakenly perceived as a collision by the robot), which does not represent a safety problem in this case.

In addition, the current work focuses on a structured environment with fixed cameras and a stationary robot base position, which has yet to be generalized for an unstructured environment. In principle, however, this does not restrict the generality of this approach, since for cobots only the corresponding position of the robot base has to be determined for the proximity detection to a human operator. In our ongoing work, we are trying to use some methods to tackle these problems. Moreover, with the current software and hardware, a sampling rate of HAR and contact detection networks are 30 Hz and 200Hz, respectively, while for the mixed perception system, there is a need of synchronization of the result of both systems.

5. Conclusion

The efficiency of safety and productivity of cobots in HRC can be improved if cobots are able to easily recognize complex human actions and can differentiate between multitude contact types. In this paper, a safety system using a mixed perception is proposed to improve the productivity and safety in HRC applications by making the cobot aware of human actions (visual perception), with the ability to distinguish between intentional and incidental contact (tactile perception). The Vision perception system is based on a 3D CNN algorithm for human action recognition, which unlike the latest HAR methods, was able to achieve 99.7% accuracy in an HRC scenario. The HAR system is intended to detect human action once the latter enters the workspace and only in case of hazardous situations, the robot would adapt its speed or stop accordingly which can lead to higher productivity. On the other hand, the tactile perception by focusing on the contact between robot and human, can decide about the final situation during pHRC. The contact detection system by taking advantage of the contact signal patterns and 1D-CNN network, was able to distinguish between the incidental and intentional contact and recognize the impacted cobot's link. According to the experimental result, with respect to traditional and new methods, our proposed model is obtained the highest accuracy of 96% in tactile perception.

Yet, based on our experimental results, visual and tactile perceptions are not sufficient enough separately for intrinsically safe robotic applications, since each system exhibits some lack of information which may cause less productivity and safety. By considering this fact, the mixed perception, by taking advantage of both visual and tactile perception, can enhance productivity and safety. Although a simple safety perception spectrum of the mixed perception is proposed which needs more research to enhance its intelligence, it shows higher resolution in compared to each single perception system.

As a future work for our system, we will extend our research regarding to multi-contact and multi-person detection which is highly beneficial for the latest Industry 4.0 safety considerations.

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 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 cyclic 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.
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. Baccouche, M.; Mamalet, F.; Wolf, C.; Garcia, C.; Baskurt, A. Sequential deep learning for human action recognition. In Proceedings of the Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer, Berlin, Heidelberg, **2011**; Vol. 7065 LNCS, pp. 29–39.
33. Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D Convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 221–231, doi:10.1109/TPAMI.2012.59.
34. Latah, M. Human action recognition using support vector machines and 3D convolutional neural networks. *Int. J. Adv. Intell. Informatics* **2017**, *3*, 47–55, doi:10.26555/ijain.v3i1.89.
35. Almaadeed, N.; Elharrouss, O.; Al-Maadeed, S.; Bouridane, A.; Beghdadi, A. A Novel Approach for Robust Multi Human Action Detection and Recognition based on 3-Dimentional Convolutional Neural Networks. *Pattern Recognit. Lett.* **2019**.
36. Arunnehru, J.; Chamundeeswari, G.; Bharathi, S.P. Human Action Recognition using 3D Convolutional Neural Networks with 3D Motion Cuboids in Surveillance Videos. In Proceedings of the Procedia Computer Science; Elsevier B.V., **2018**; Vol. 133, pp. 471–477.
37. Asghari-Esfeden, S.; Io, S.; Sznaier, M.; Camps, O. Dynamic Motion Representation for Human Action Recognition *. In Proceedings of the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); **2020**; pp. 557–566.
38. Fan, H.; Luo, C.; Zeng, C.; Ferianc, M.; Que, Z.; Liu, S.; Niu, X.; Luk, W. F-E3D: FPGA-based acceleration of an efficient 3D convolutional neural network for human action recognition. In Proceedings of the Proceedings of the International Conference on Application-Specific Systems, Architectures and Processors; Institute of Electrical and Electronics Engineers Inc., **2019**; Vol. 2019-July, pp. 1–8.
39. Recognition of human actions Available online: https://www.csc.kth.se/cvap/actions/ (accessed on Oct 8, **2020**).

40. Hoang, V.D.; Hoang, D.H.; Hieu, C. Le Action recognition based on sequential 2D-CNN for surveillance systems. In Proceedings of the Proceedings: IECON 2018 - 44th Annual Conference of the IEEE Industrial Electronics Society; Institute of Electrical and Electronics Engineers Inc., **2018**; pp. 3225–3230.
41. Kim, J.H.; Won, C.S. Action Recognition in Videos Using Pre-Trained 2D Convolutional Neural Networks. *IEEE Access* **2020**, *8*, 60179–60188, doi:10.1109/ACCESS.2020.2983427.
42. Singh, I.; Zhu, X.; Greenspan, M. Multi-Modal Fusion With Observation Points For Skeleton Action Recognition. In Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP); IEEE, **2020**; pp. 1781–1785.
43. Weng, J.; Luo, D.; Wang, Y.; Tai, Y.; Wang, C.; Li, J.; Huang, F.; Jiang, X.; Yuan, J. *Temporal Distinct Representation Learning for Action Recognition*; **2020**.
44. 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.
45. 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.
46. 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.
47. 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.
48. 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.
49. 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.
50. 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.
51. 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.
52. 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.
53. 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.
54. Cao, P.; Gan, Y.; Dai, X. Finite-Time Disturbance Observer for Robotic Manipulators. *Sensors* **2019**, *19*, 1943, doi:10.3390/s19081943.
55. 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.
56. 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.

57. 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.
58. Haddadin, S. *Towards safe robots : approaching Asimov's 1st law*; ISBN 9783642403088.
59. 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.
60. Sharkawy, A.-N.; Koustoumpardis, P.N.; Aspragathos, N. Neural Network Design for Manipulator Collision Detection Based Only on the Joint Position Sensors. *Robotica* **2019**, 1–19, doi:10.1017/S0263574719000985.
61. 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.
62. 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**).
63. 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.
64. Dine, K.M. El; Sanchez, J.; Ramón, J.A.C.; Mezouar, Y.; Fauroux, J.-C. Force-Torque Sensor Disturbance Observer using Deep Learning. **2018**.
65. 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.
66. 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.
67. Yang, K.; Wang, X.; Quddus, M.; Yu, R. Predicting Real-Time Crash Risk on Urban Expressways Using Recurrent Neural Network. **2019**.
68. 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.
69. Li, S.; Li, W.; Cook, C.; Zhu, C.; Gao, Y. Independently Recurrent Neural Network (IndRNN): Building a Longer and Deeper RNN **2018**, 5457–5466.
70. 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.
71. Zhang, B.; Wang, L.; Wang, Z.; Qiao, Y.; Wang, H. Real-Time Action Recognition With Enhanced Motion Vector CNNs **2016**, 2718–2726.
72. 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.
73. 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.
74. Birjandi, S.A.B.; Kuhn, J.; Haddadin, S. Observer-Extended Direct Method for Collision Monitoring in Robot Manipulators Using Proprioception and IMU Sensing. *IEEE Robot. Autom. Lett.* **2020**, *5*, 954–961, doi:10.1109/LRA.2020.2967287.

75. BaradaranBirjandi, S.A.; Haddadin, S. Model-Adaptive High-Speed Collision Detection for Serial-Chain Robot Manipulators. *IEEE Robot. Autom. Lett.* **2020**, doi:10.1109/LRA.2020.3015187.
76. 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.
77. T. K. Ho Random Decision Forests Tin Kam Ho Perceptron training. In Proceedings of the Proceedings of 3rd International Conference on Document Analysis and Recognition; **1995**; pp. 278–282.
78. Kinect for Windows v2 Available online: https://docs.depthkit.tv/docs/kinect-for-windows-v2.
79. 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.
80. Rezayati, M.; van de Venn, H.W. Dataset: Collision Detection in Physical Human Robot Interaction. **2020**, *1*, doi:10.17632/CTW2256PHB.1.
81. 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.
82. 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.
83. 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.
84. Johannsmeier, L.; Gerchow, M.; Haddadin, S. *A framework for robot manipulation: Skill formalism, meta learning and adaptive control*; **2019**; Vol. 2019-May; ISBN 9781538660263.
85. 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.
86. Weistroffer, V.; Paljic, A.; Callebert, L.; Fuchs, P. *A Methodology to Assess the Acceptability of Human-Robot Collaboration Using Virtual Reality*; **2013**.
87. Normung, E.K. für TECHNICAL SPECIFICATION ISO / TS Robots and robotic devices —. **2016**, *2016*.
88. Wen, X.; Chen, H.; Hong, Q. Human Assembly Task Recognition in Human-Robot Collaboration based on 3D CNN. In Proceedings of the 9th IEEE International Conference on Cyber Technology in Automation, Control and Intelligent Systems, CYBER 2019; Institute of Electrical and Electronics Engineers Inc., **2019**; pp. 1230–1234.
89. Liu, H.; Wang, L. Gesture recognition for human-robot collaboration: A review. Int. J. Ind. Ergon. **2018**, 68, 355–367, doi:10.1016/j.ergon.2017.02.004.
90. Cao, Z.; Simon, T.; Wei, S.-E.; Sheikh, Y. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *XXX*.
91. Fang, H.-S.; Xie, S.; Tai, Y.-W.; Lu, C.; Jiao Tong University, S.; YouTu, T. RMPE: Regional Multi-Person Pose Estimation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV); **2017**.
92. Robertini, N.; Bernard, F.; Xu, W.; Theobalt, C. Illumination-invariant robust multiview 3d human motion capture. In Proceedings of the Proceedings - 2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018; Institute of Electrical and Electronics Engineers Inc., **2018**; Vol. 2018-January, pp. 1661–1670.
93. Anvaripour, M.; Saif, M. Collision detection for human-robot interaction in an industrial setting using force myography and a deep learning approach. In Proceedings of the Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics; Institute of Electrical and Electronics

Engineers Inc., **2019**; Vol. 2019-Octob, pp. 2149–2154.