KoARob: Towards AI-based Safety in Human-Robot-Collaborations

Daniel Bermuth, Alexander Poeppel and Wolfgang Reif

***Abstract*— As industries face changes in population and the need for better production efficiency arises, combining human workers with robots is becoming more common. In such collaborations, besides the user experience, the safety of the human worker is a critical aspect. This paper introduces an AI-based safety system designed to maintain a safe distance from human workers to prevent injuries. To circumvent the limitations of similar safety approaches, the system uses redundant methods to detect humans and their various joints. An evaluation in a real-world scenario shows that such an AI-based system can reliably detect humans and stop robots before a collision occurs. This work proves that using AI-based systems for human detection in safety-related contexts is not impossible and creates a basis for a new generation of safety systems that can enhance future human-robot collaborations.**

I. Introduction

Following demographic changes and the demand for efficient production, many processes are being automated. Since not every task can be fully automated, because the task is too complex for a robot or needs high flexibility, a collaboration between humans and robots can be a good solution. In such an application, both can bring in their strengths to achieve a common goal. The most important aspect of such a collaboration is the safety of the human worker, since the robot can cause severe injuries if errors occur.

While there already exist several safety systems for robots, most of them are not suitable for efficient human-robot collaborations, because they are too restrictive. Commonly used systems, like light barriers [1] or light-based distance sensors [2], already stop the robot before the human is even close to the robot, while others, like *CoBots* [3] for example, only stop after a collision has occurred. Both system types are not ideal for human-robot collaborations. While contact with the hands is usually acceptable or even desired for some tasks, contact with the rest of the body should be avoided, especially the head should always be kept at a safe distance. Thus, a system is desirable that can differentiate between different body parts.

Therefore, in this work, a safety system is presented that uses redundant AI-based methods to detect humans and their various joints. The system is evaluated in a real-world scenario, in which it can be shown that the system reliably detects humans and stops the robot before a collision occurs. This makes it the first AI-based system for which, over the investigated experiment duration, a reliability was reached that is required for real production environments. This work thus lays the ground for a new generation of safety systems that can improve future human-robot collaborations.

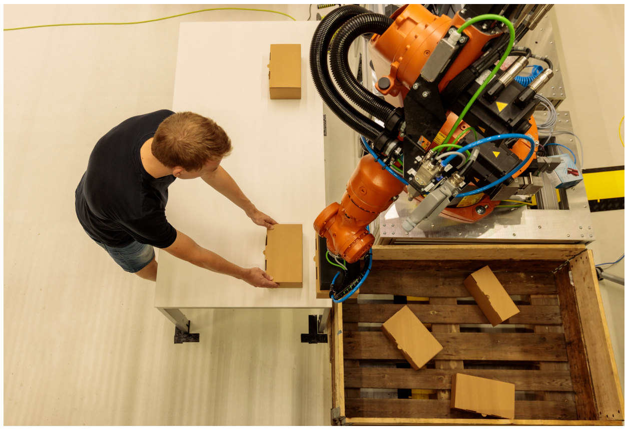


Fig. 1: Human robot collaboration without safety barriers.

II. Related Work I

To allow collaboration with human workers, the robot cell must be openly accessible, and not fenced off like in conventional industrial applications. There already exist several commercial options for this:

A simple approach for open access is to use light barriers or floor plates that detect the human worker. This is a very reliable system, but it has the disadvantage that the robot, if it is not combined with another safety system, has to stop if the worker enters the work cell, even if the worker is not in danger [4]. This can lead to a significant reduction in productivity, and true collaboration is not possible at all.

Another option is to use robots with built-in collision detection, that automatically stop if they hit an obstacle. There already exist several robots that have this feature, like the *KUKA LBR iiwa* or some of the *Universal Robots URs*. The disadvantage of these systems is that the robot only stops after a collision, which is not the best user experience for workers, and the maximum speed and payload of the robot are also restricted. There also exists some research into classifying whether contacts were intended or not [5], but from a safety perspective, those systems are not reliable enough yet to allow the robot to continue its movement.

One approach to stopping the robot before contact is the use of capacitive sensors. They can detect the human worker before a collision occurs, but they have the disadvantage that they only have a range of a few centimeters, and are sensitive to environmental influences in the workspace. One commercial example is the *Bosch APAS* system, though there exist more examples in research [6, 7].

This work is part of the project *KoARob: Combination of Assistance and Sensor Systems for Safe Human-Robot Cooperation* funded by the *Bavarian Ministry of Economic Affairs, Regional Development and Energy*.

All authors are with the Institute for Software & Systems Engineering, University of Augsburg, 86159 Augsburg, Germany. Contact: daniel.bermuth@informatik.uni-augsburg.de

A different option is the use of additional external sensor systems, like cameras or laser sensors, that detect the human from some distance. In [8] a laser scanner is used to stop the robot if a human enters the workspace. With the *Pilz-SafetyEye* there was also a commercial camera-based system that allowed the monitoring of a 3D-space using multiple safety zones around the robot. It is not available anymore. Splitting the workspace into more zones allows reducing the number of stops, but true collaboration is still not possible.

Ideally the robot system should be aware of the worker's position and only create the minimal required safety zone around the human body to prevent unnecessary stops. The following chapters will investigate how this can be achieved. After a description of the robot scenario, an overview of common safety standards is given, followed by an explanation of the proposed safety concept. The system is then evaluated, and compared to further related work.

III. Scenario description

The checking of returned packages is a common task for many online shops. The packages are returned by customers and need to be checked for completeness and damage so that they can be put back into stock if they are still in good condition. This task is normally done by a human worker, because the quality inspection is hard to automate. The worker needs to open the packages, check the contents and sort the packages into different categories. In the examined example scenario, the parcels are delivered on a Euro-pallet. The worker takes the parcels from the pallet and puts them on a table in front of them. The parcels are then opened, and the contents are checked. After the check is completed, the worker sorts the parcels into two different boxes, depending on the condition of the contents. This task is very repetitive and can lead to physical strain for the worker.

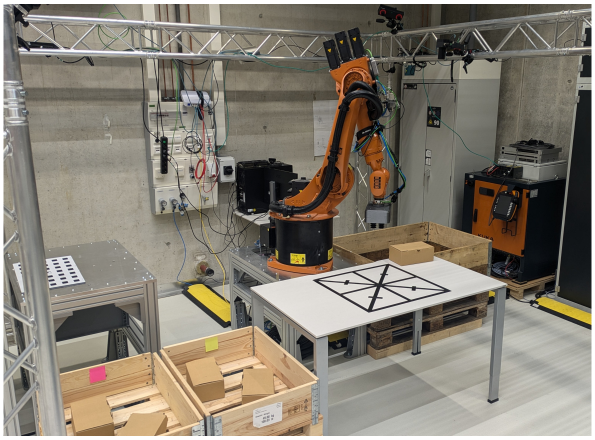


Fig. 2: The robot setup.

The task of the robot in the *KoARob* example is thus to assist the worker by fetching the parcels from the pallet and placing them on the table. This reduces the physical strain on the workers and allows them to focus on the quality inspection. Since robot and worker work close together, the robot system needs to be able to detect the worker and ensure their safety. Also, in case a package is too damaged for the robot to lift it with the suction gripper, the worker should be able to fetch it manually.

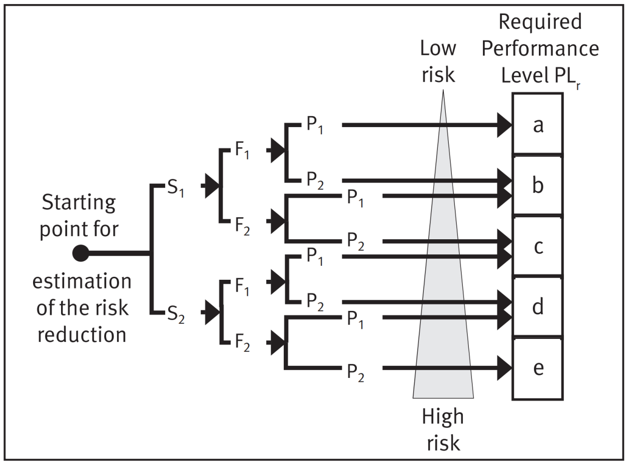
Figure 2 shows the setup for the application. The robot is a *KUKA KR16* with a custom suction gripper. It lifts the packages from the pallet on the right side and places them onto the table, alternating between two drop positions. After the worker has inspected the contents of each parcel, he sorts them into two boxes on the left side (yellow-green for reusable ones, pink-red for waste). The worker is watched by five *Roboception rc-visard* stereo-cameras mounted on the trusses, while another one detects the locations of the packages on the pallet. If the robot is unable to lift a package, the worker can walk around the table and fetch it manually.

IV. Safety analysis

The safety of the system is a critical aspect, especially in a scenario where the robot is working close to a human worker. To ensure the safety of a human worker each system has to follow safety regulations. This chapter will present a short overview of the generally common procedure. The safety of the system needs to be analyzed in different steps according to the ISO 13849-1 standard. First, the risks of the robot are identified. Then the safety measures are defined to reduce the risks. Finally, the safety measures are evaluated to ensure that they are sufficient to reduce the risks to an acceptable level. In the case of using a robot, additional standards like ISO 10218-1 and ISO 10218-2, and for a collaboration robot additionally ISO 15066, need to be considered as well.

A. Risk assessment

In general, after the risks of the system are identified, the risks need to be analyzed to determine the required performance level (PL_r) of the safety system. Depending on their severity they can be classified into one of five levels a-e, for example by following the simple procedure in Figure 3.



S – Severity of injury

S1 = Slight (normally reversible injury)
S2 = Serious (normally irreversible injury or death)

F – Frequency and/or exposure to hazard

F1 = Seldom to less often and/or exposure time is short
F2 = Frequent to continuous and/or exposure time is long

P – Possibility of avoiding hazard or limiting harm

P1 = Possible under specific conditions
P2 = Scarcely possible

The possibility of avoiding the hazard is further specified by:

- Speed with which the hazard arises (e.g. quickly or slowly)
- Possibilities for hazard avoidance (e.g. by escaping)
- Practical safety experiences relating to the process
- Operation by experts or non-professionals
- Operation with or without supervision

Fig. 3: Risk assessment according to ISO 13849 [4].

Due to space constraints, we concentrate on a small selection of the most important risks:

- The robot places a parcel on the table while the human has its hand on the put-down position:
  - $S1$ (suction gripper is somewhat flexible)
  - $F2$ (could occur more than once per hour)
  - $P1$ (robot moves slow enough that human could pull back hand)
  - $PL_r = b$ (according to Figure 3)
- Robot drops the parcel, it falls on human foot: $S1$ (parcel weight $< 5kg$), $F2$, $P1$, $PL_r = b$
- Robot moves on teached path and hits human: $S1$ (robot moves slow, no major squeeze points), $F2$, $P1$ (lot of space around robot), $PL_r = b$
- Robot has an error and leaves the workspace: $S2$, $F1$, $P2$, $PL_r = d$
- Robot has an error and moves too fast: $S2$, $F1$, $P2$, $PL_r = d$

Each risk needs to be handled by a safety measure that has at least the required performance level. One safety system might secure multiple risks together. So in this case, if only one safety system shall be used, a PL_d is required, as ISO 10218-1 requires for robots in general, if no detailed risk assessment was done.

B. System assessment

After defining a safety function for each risk, the performance level of each safety measure has to be analyzed. For this, the following three factors are most important: the function's structural category, the reliability, and the runtime parallel testing of possible system errors. The categories describe the structure of the system, whether it is basic (B), uses better components (1), has additional runtime tests (2), and also includes redundancy (3/4). The reliability is given by the mean time to dangerous failure ($MTTF_d$), which is measured in years and converted to four different levels ($not\,suitable\,(<$ 3a), $low\,(\geq$ 3a), $medium\,(\geq$ 10a), $high\,(\geq$ 30a)). The runtime parallel testing is measured through the diagnostic coverage (DC), which is the percentage of dangerous failures that are detected by the safety system, which is also converted to four levels ($none\,(< 60\%)$, $low\,(\geq\ 60\%)$, $medium\,(\geq\ 90\%)$, $high\,(\geq\ 99\%)$). The overall performance level of the evaluated safety function can then be taken from the diagram in Figure 4 (there also exists another option to directly calculate the PFH_D value if required).

Note that the standard has additional requirements, for example regarding testing of common causes of failures or the specification and verification of software, but the aforementioned factors are the most important ones for the following evaluation.

V. Safety Concept

To simplify development, the safety functions are divided into three main parts: safe robot positioning, safe human detection, and distance monitoring. The safe robot positioning ensures that the robot is not moving too fast or at the wrong positions, while the safe human detection and the distance monitoring ensure that the robot stops if a human gets too close.

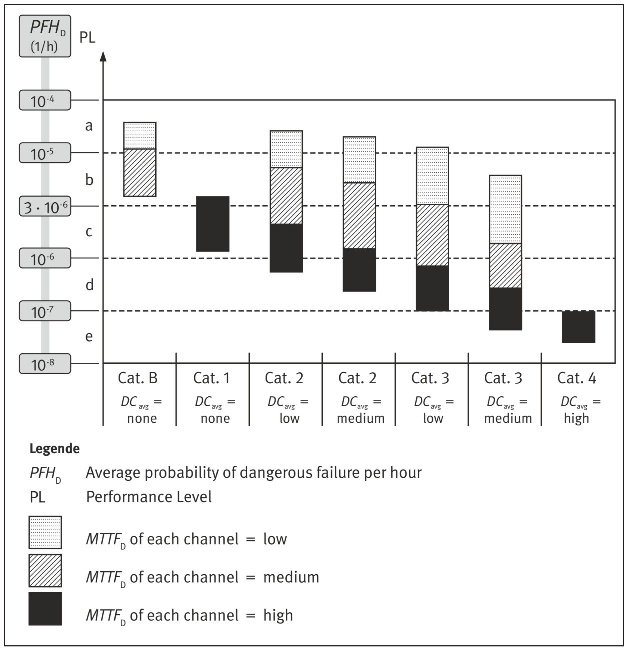


Fig. 4: System assessment according to ISO 13849 [4].

A. Safe Robot Position

As found in the risk analysis, leaving the workspace or moving too fast are two very critical risks that can lead to severe injuries. Especially too rapid acceleration needs to be detected quickly. To reach the required performance level a redundant system with at least two channels is needed. Often the robot's software already provides safety functions to monitor the robot's position and speed. If this is not the case or the robot does not output its position redundantly, a second channel is needed to verify the robot's position. This could for example be achieved using an *Inertial Measurement Unit (IMU)* attached to the robot's gripper, a camera system detecting *AprilTags* [9] attached to the robot's joints, or *Wire Draw Encoders* [10] to measure the robot's position directly. Since the focus of this work is on reliable AI-based human detection, it will be assumed that the robot's position output is safe and reliable.

B. Safe Human Detection

To introduce redundancy in the human pose estimation, three different approaches were developed in this project:

The first, published as *VoxelKeypointFusion* [11], uses a top-down approach with *RTMPose* [12] to first detect humans in each 2D image by their bounding boxes and then their joints (keypoints) in the box cutouts. The heatmap representations of the keypoints are then triangulated in voxelized space to estimate 3D joint proposals at overlapping positions. The resulting proposals are then merged and assigned in a bottom-up approach to complete persons. The second, called *PlaneSweepPosePlus* is a combination of *RTMO* [13] with *PlaneSweepPose* [14] to make it end-to-end runnable, including some additional improvements. It uses a bounding-box approach to detect locations of 2D persons with their joints in a single step, and a mixture of depth estimation and per-view consistency to estimate the 3D poses. The third, published as *SimpleDepthPose* [15], predicts 2D keypoints using a bottom-up approach with a modified *HigherHrNet* [16] model that only predicts directly visible joints. Then the corresponding distances of each joint are extracted from the cameras' depth images. The resulting 3D poses per view are merged, and, after dropping outliers, averaged to get the final 3D poses.

One important aspect here is that all three models use different architectures to detect the human joints, to increase the probability that at least one of them detects the human if the others have problems. For example, a top-down approach normally achieves better localization results, because they can cut out their boxes with higher resolutions, while in contrast to bottom-up approaches, they are more likely to miss the complete person if only very few joints are visible and no box is predicted. Using depth information allows the accurate detection of persons even if they are visible in only one camera, but if the depth information is noisy, the resulting poses can be very inaccurate. A triangulation-based approach needs at least two cameras to detect a person, but can also estimate positions of occluded joints if the 2D models correctly predict/estimate those joints.

Other common challenges for vision-based detectors, like unusual clothing, severe occlusions or challenging lighting, which also would affect the above approaches, can be mitigated by making use of the controlled setup. The company owner can influence the workers' clothing and the system developer can optimize the number and positions of the cameras to reduce occlusions, as well as the lighting conditions.

The three models were evaluated across multiple publicly available datasets (*Panoptic Studio* [17], *Human3.6m* [18], *Campus & Shelf* [19], *MVOR* [20]), and compared to many other open-sourced state-of-the-art approaches (*VoxelPose* [21], *Faster-VoxelPose* [22], *MvP* [23], *PRGnet* [24], *TEMPO* [25], *SelfPose3d* [26], *mvpose* [27], *mv3dpose* [28], *PartAwarePose* [29], *OpenPTrack* [30]). It was found that regarding the performance of the models, on datasets yet unseen to them to test their generalization capabilities, the three models described before were in summary the most reliable. Further details about the evaluations and their results can be found in the papers of *VoxelKeypointFusion* [11] and *SimpleDepthPose* [15], which also include links to their source codes and any pre-trained models.

C. Distance Monitoring

To simplify the redundant execution of the safety function, to reduce hardware requirements, and to ensure explainability, a simple concept for distance monitoring is chosen. The robot and the human are represented by two voxelmaps, which define their space occupancy (see Figure 5). If a voxel is occupied by the robot and the safety-hull of the human occupies it, the robot is stopped. If the voxel is freed again by one of them, the robot can continue to move. Besides that, the safety system also ensures that both voxelmaps are up to date, so if one voxelization module stops sending new messages, the robot is automatically stopped.

Since the person detectors can differentiate between different body parts, the human pose voxelization process can define different safety distances for different body parts. For the scenario three different distances for the head, torso and hands were chosen. So the system can ensure a larger distance for the worker's head and a lower distance for his hands.

An additional buffer was introduced to prevent the robot from constantly jerking in borderline cases when the human is located directly at the robot's safety boundary zone and the voxels at the boundary are sometimes occupied and sometimes released due to slight movement. Therefore, this second safety module aggregates the output of the voxelmap occupancy over a certain period of time (currently 100ms) and only releases the robot if all states in it are ”free”. This module has the additional advantage that other modules can also trigger a stop outside the voxelmap architecture. A camera monitoring module for example could generate a stop if it detects that the camera has moved or is no longer providing current images.

The occupancy evaluation runs on a single Raspberry-Pi-4, and takes about 1-3ms (at $5cm$ voxel size). It should therefore also be possible to run the calculation on somewhat weaker but safety-certified hardware.

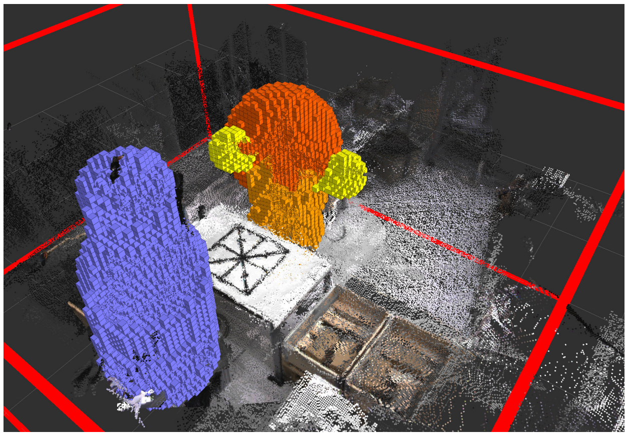


Fig. 5: Voxel-based distance monitoring. The human is represented by the orange/yellow voxels (depending on the limb type), and the robot by the purple/blue ones.

D. System architecture

Figure 6 shows the overall software pipeline, and how the three safety parts are combined. As interface to the KUKA robot *RSI* is used to receive the robot's position and to update the *override* speed to stop it. For communication between the other modules *ROS2* is used. The modular software architecture allows an easy exchange of the safety modules.

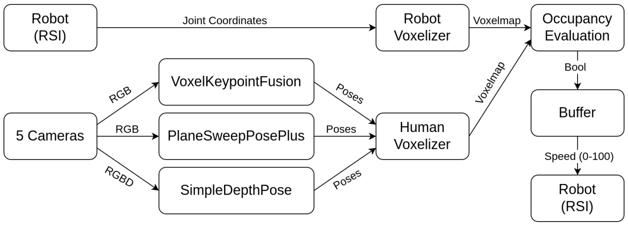


Fig. 6: Software pipeline.

VI. Evaluation

The safety as well as the usability of the system were finally evaluated using the setup described in Chapter III.

A. Pose Estimation

To assess the reliability of the person recognition in the given scenario, a dedicated dataset was created and labeled. To create the dataset, various people were asked to process the ”returns handling” scenario with the robot. In the scenario, the robot placed several packages with different contents on top of the table in front of the worker, which they then inspected and, depending on whether an object was present or not, sorted them into two boxes for reusable returns or waste. Meanwhile, the robot was monitored by the safety system described before and stopped automatically if the distance between the person and the robot fell below a certain threshold.

The images from the cameras were recorded during this process, and afterward, parts of the sequences were labeled. A semi-automatic procedure was implemented for this, in which one of the algorithms (regularly switched to prevent biases) created an initial suggestion for the joint positions, which was then corrected manually to repair any misidentifications. A total of 1100 images from 10 people were labeled. The sequences were labeled at a frame rate of $10Hz$ with a few seconds of consecutive frames and then a larger gap before the next sequence. See Figure 7 for an example of the labeled dataset.

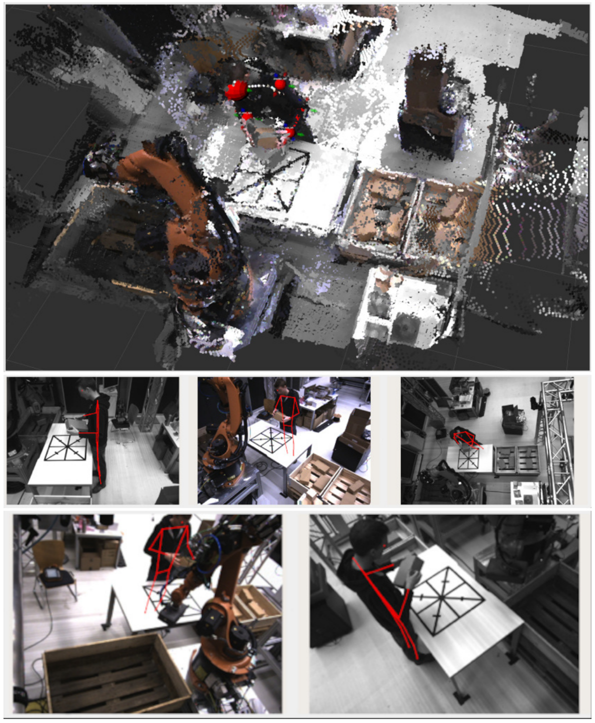


Fig. 7: Evaluation dataset example of a 3D pose and their projection into the five camera views.

The three models were evaluated using the standard metrics in computer vision. *Percentage of Correct Keypoints (PCK)* calculates the percentage of how many keypoints were detected with an error lower than the given threshold in millimeters. *Mean Per Joint Prediction Error (MPJPE)* calculates the mean error of all joints for each person and then averages over all persons, dropping persons with an error of above $500mm$ as not matched. In total 13 keypoints are evaluated (2 shoulders, 2 hips, 2 elbows, 2 wrists, 2 knees, 2 ankles, 1 nose/head). The *Recall* shows the percentage of persons with an average joint error regarding the ground-truth lower than the given threshold. *Invalid* counts the percentage of predictions that were not matched to any ground-truth label, and *F1* combines it with the *Recall@500* score into a single value. More details can be found in the paper of *VoxelKeypointFusion* [11]. The *FPS* states the end-to-end (images-to-poses) inference speed, evaluated on a single *Nvidia-3090*.

In Table I it can be seen that *VoxelKeypointFusion* and *PlaneSweepPosePlus* which both only use color images had problems in correctly detecting all persons (*Recall@500*). In contrast, *SimpleDepthPose* recognized all persons and was also the fastest. In comparison to other datasets evaluated previously, the lack of synchronization of the cameras was found to result in some performance decreases. For hardware reasons only the most recent image from each camera could be used instead of generating all images simultaneously. If the person moves quickly, the joints can be recorded at different points in the respective images. *VoxelKeypointFusion* in particular then has problems with triangulation, since the heatmap rays no longer overlap correctly. *SimpleDepthPose* also had problems with the quality of the depth data and additional extensive occlusions, this algorithm had always performed most reliable in the previous comparisons.

Since, as Table I shows, no algorithm reliably recognizes all joints within $500mm$ (PCK@500), they need to be combined to improve the results. Using their diverse architectures, this should lead to better reliability. Two different

Method	PCK@100/500		MPJPE	Recall@100/500		Invalid	F1	FPS
VoxelKeypointFusion	82.8	**99.4**	59.1	90.2	99.6	**0.0**	**99.8**	6.2
PlaneSweepPosePlus	**86.5**	99.2	**45.9**	**94.8**	99.3	0.7	99.3	8.6
SimpleDepthPose	62.5	97.9	112	51.3	**100**	11.1	94.1	**22.7**

TABLE I: Results of single approaches on the new *koarob* dataset

Method	minPCK@99%	minPCK@99.9%	minPCK@100%	non-close-percentage
close persons	0.20	0.30	0.45	21.4 (stop frames)
fused keypoints	0.15	0.25	0.30	11.7 (extra joints)

TABLE II: Results of merged approaches on the new *koarob* dataset. $minPCK$ in meters.

Method	Head	Shoulders	Elbows	Wrists	Hips	Knees	Ankles
close persons	0.30	0.25 | 0.20	0.20 | 0.30	0.20 | 0.30	0.25 | 0.25	0.25 | 0.30	0.45 | 0.30
fused keypoints	0.25	0.15 | 0.25	0.15 | 0.30	0.15 | 0.25	0.20 | 0.25	0.25 | 0.30	0.30 | 0.30

TABLE III: Results per joint (*minPCK@100%*, *left*|*right*) of merged approaches on *koarob*

approaches were investigated for this purpose. The first one checks in the first step whether all algorithms have recognized the same number of persons and then whether all 3D joint coordinates of these persons are also close to each other. If not, a stop is triggered. The second approach just merges the results of all algorithms, and thus detects the persons at all suggested positions (which then can simply overlap).

As can be seen in Table II, both combination approaches now recognize all joints reliably (minimal distance in which all joints are detected ($minPCK$@100%) is less than $500mm$). With the second, however, the minimum required safety space all human joints are recognized within is somewhat smaller, which allows closer cooperation between humans and robots. Both concepts have their own advantages and disadvantages. The first approach stops the robot in 21% of the images (although these are not expected to be randomly distributed, but position-dependent, which means if the worker enters an area with poor visibility, the robot remains stationary all the time). On the other hand, it can be assumed that the accepted images have a good pose accuracy. The second approach does not stop the robot directly, but increases the amount of occupied voxels. Approximately 12% of the occupied joint points are further than $500mm$ away from the actual ones.

The results per joint in Table III show that the localization error notably differs between the joint types. Especially the ankles, which were highly occluded in the evaluated scenario, have a larger error. By adding more cameras to reduce occlusions it is expected that this position error can be reduced. Instead of the partition in head, torso and limbs, used while running the evaluation (as shown in Figure 5), one could also use a custom voxel-occupation radius per joint type. In most applications only a close distance to the hands and arms is of importance, the rest of the body can stay further away.

B. Safety System

Regarding the assessment of the safety system according to Figure 4, the category of the system is B, since no redundancy is used for the distance monitoring and the three human detection methods were merged into a single channel. While there are a few runtime tests, like for recent timestamps, the overall DC is below 60%. In combination this would match to the first column of Figure 4. To reach a certifiable PL with the safety system, one would now need to show a longer $MTTF_d$ duration. While the evaluation showed the reliability of the presented safety system for a short time, a much longer testing phase would be needed to show the reliability for at least 3 years to reach a suitable $MTTF_d$ level.

One option could be to test the system in real industrial applications in which there are either no serious risks of injuries to persons or in which protection is provided by other measures and sensors. Another option if the reliability already was shown for some longer time, would be to reduce the possible risks that they require a lower PL_r. For example, in the current scenario, this could be possible if the risks are split into two parts, one caused by the robot which is secured by a PL_d system, and a second one that contains only the human risks which only require PL_b. One of course, and according to ISO 10218-1 as well, would need to do a detailed risk analysis then, which covers all normally possible risks. This would have the benefit that the $MTTF_d$ which needs to be verified could be lower in the first step. In a running application, the $MTTF_d$ could then be further evaluated to prove that it also meets the requirements for higher PL.

VII. Related Work II

There is already some existing research into tighter safety zones around the human workers to compare against:

In [31] multiple *Time of Flight (ToF)*-cameras are used to detect obstacles that trigger a stop if they reach a dynamically adjusted safety zone. In [32] two *ToF*-cameras are used to detect humans in the path of the robot. Another early research project is [33], in which multiple depth-cameras are used to partition the workspace into three different types, background, robot and human/unknown. Robots and humans received an additional safety-hull and if they overlapped the robot is stopped. A similar concept was presented in [34], including additional collision avoidance. All those systems lack safety certification and thus can not be used in production environments. One commercially available and safety-certified system following the same approach is *Veo Freemove* [35], which also separates between background, robot and person/unknown using multiple cameras. To reach redundancy, besides redundant calculations, each of the *ToF*-cameras has two sensors that can be cross-checked. The main disadvantage of purely depth-based systems is that they only distinguish between large unknown objects and background, a finer distinction between human body parts, like head or hands, is not possible. Therefore such a system can not allow acceptable contacts with a human hand and prevent dangerous contacts with the head at the same time.

To distinguish between the body parts one could use a marker-based approach in which the workers are required to wear special clothing that can easily be detected by sensor systems. Examples of such systems are the commercially available *Vicon Motion Capture Suit* or other research works like [36–38]. This might be a very reliable system, even though none of the authors have added a detailed safety analysis. Such a system, however, has the disadvantage that the workers have to wear special suits, which can be uncomfortable and restrict their movement, or in some use-cases is not possible at all.

Regarding AI-based safety systems, which would solve the aforementioned problems, and which would be able to differentiate between body parts without using tracking suits, there are only a few existing research projects so far: In [39], a voxel-based concept was extended with information from a human pose estimation system to annotate voxel-types. They mentioned a latency of around $0.8s$ for voxels of $4cm$ size and only used a single camera system which is susceptible to occlusions. The detection reliability was not evaluated. The authors of [40] used multiple modalities including thermal images for background removal and segmentations before detecting human bounding-boxes. Those bounding-boxes were then used to separate the point-cloud into robot, background and human points. They reported a miss-rate of at least 4%, depending on the modalities they used. The approach of [41] used an object detection system to detect body parts in the workcell and another neural network to predict whether the object is inside the safety distance or not. While they were able to correctly detect all intrusions in their test dataset, the approach is not transferable to other scenarios, since every time the second network needs to be retrained. In [42] a camera with a top-down view was used to detect the heads of humans in the workspace. A cylindrical safety zone was then created around the person and depending on the distance the robot's speed was reduced. The authors reported a recall of 91.5% for the detection in the danger zone. The system is also not able to distinguish between body parts. Unlike those works, the presented system was able to detect all persons as well as all their joints in the workspace with 100% reliability. The evaluation also closely followed the requirements of current safety standards.

In different domains, the project *KI Absicherung* [43, 44] explored the usage of AI-based safety systems for autonomous vehicles and developed both models and architectural concepts for such a system. In the currently ongoing project *safe.trAIn* [45] a similar approach is followed for the railway domain, using a driverless regional train as example.

VIII. Conclusion

This work presented a safety concept for human-robot collaboration using AI-based sensor processing, which was evaluated in a real-world scenario. To improve the performance, three different human-pose detectors were combined. The system was able to reliably detect humans and stop the robot if the distance between the robot and any of the human's joints dropped below a certain threshold.

Since the system, and especially the human joint detection algorithms, are not specifically designed or trained for this package-returning scenario, but allow a general detection of humans in some workspace, the system is not limited to this specific application. Instead, it can be used in a wide range of other applications where a robot needs to work close to a human worker. A more complex scenario could be in bike manufacturing, for example, when the human integrates the different cables into the bike frame. The robot can hold and move the frame, to ergonomically improve the positioning, while the worker needs to contact the frame with his hands at the same time to hold and insert the cables.

The main contribution of this work is showing that using AI-based systems for human detection in safety-related contexts is not impossible. Unlike previous works that used AI-based concepts, it was shown, even though only for a short time so far, that the system can detect persons and their joints in the workspace with 100% reliability. This is the most important requirement for a new type of safety system, because if it is not completely reliable, the system will not be usable in any production environment.

In the future, the investigated safety concept will allow the use of specific benefits of AI-based approaches, like the ability to differentiate between different body parts. This enables an optimization of the safety distances depending on the joint type, for example to allow closer contacts with the hands than with the head, which can then lead to higher collaboration productivity. Besides stopping the system before a collision occurs, one could further argue, in line with ISO 15066, that transient contacts with limited speeds are acceptable. Because the speed (and force) limits for limbs, and especially the hands, are higher (the contact impulse is lower because they can be pushed away more easily), the now available ability to differentiate between joint types would allow faster movements of the robot in certain spaces and thus again higher productivity.

One of the main remaining challenges is to prove the reliability of the system over a longer time, which is needed to reach a certifiable performance level. However, this is expected to be solvable with further research. Although a general safety certification is left for future work, the presented system is already usable in some applications. For example, in an application using an already safe *CoBot*, the system could be used to avoid contact with certain body parts, to improve the user experience of the workers.

IX. Acknowledgment

The authors would like to thank their project partners *Michael Suppa* and *Felix Ruess* from *Roboception GmbH* as well as *Sebastian Pieniack* and *Moritz Hottelet* from *MRK-Systeme GmbH* for their support with the robot setup.

References

[1] "Sick light barrier." [Online]. Available: https://www.sick.com/de/en/catalog/products/safety/safety-light-beam-sensors/detem/c/g422854?tab=overview

[2] "Sick laser scanner." [Online]. Available: https://www.sick.com/de/en/catalog/products/safety/safety-laser-scanners/microscan3/c/g295657?tab=overview

[3] "KUKA LBR iiwa." [Online]. Available: https://www.kuka.com/en-de/products/robot-systems/industrial-robots/lbr-iiwa

[4] "Functional safety of machine controls - Application of EN ISO 13849," *Deutsche Gesetzliche Unfallversicherung e. V. (DGUV)*, 2019. [Online]. Available: https://www.dguv.de/medien/ifa/en/pub/rep/pdf/reports-2019/report0217e/rep0217e.pdf

[5] C. Y. Wong, L. Vergez, and W. Suleiman, "Vision-and tactile-based continuous multimodal intention and attention recognition for safer physical human–robot interaction," *IEEE Transactions on Automation Science and Engineering*, 2023.

[6] A. Hoffmann, A. Poeppel, A. Schierl, and W. Reif, "Environment-aware proximity detection with capacitive sensors for human-robot-interaction," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 145–150.

[7] S. E. Navarro, B. Hein, and H. Wörn, "Capacitive tactile proximity sensing: from signal processing to applications in manipulation and safe human-robot interaction," in *Soft Robotics: Transferring Theory to Application*. Springer, 2015, pp. 54–65.

[8] E. Helms, R. D. Schraft, and M. Hagele, "rob@work: Robot assistant in industrial environments," in *Proceedings. 11th IEEE Int. Workshop on Robot and Human Interactive Communication*. IEEE, 2002, pp. 399–404.

[9] E. Olson, "AprilTag: A robust and flexible visual fiducial system," in *2011 IEEE international conference on robotics and automation*. IEEE, 2011, pp. 3400–3407.

[10] "Sick wire draw encoders." [Online]. Available: https://www.sick.com/de/en/catalog/products/motion-control-sensors/wire-draw-encoders/c/g286652

[11] D. Bermuth, A. Poeppel, and W. Reif, "VoxelKeypointFusion: Generalizable Multi-View Multi-Person Pose Estimation," *arXiv preprint arXiv:2410.18723*, 2024.

[12] T. Jiang, P. Lu, L. Zhang, N. Ma, R. Han, C. Lyu, Y. Li, and K. Chen, "RTMPose: Real-Time Multi-Person Pose Estimation based on MMPose," *arXiv preprint arXiv:2303.07399*, 2023.

[13] P. Lu, T. Jiang, Y. Li, X. Li, K. Chen, and W. Yang, "RTMO: Towards High-Performance One-Stage Real-Time Multi-Person Pose Estimation," *arXiv preprint arXiv:2312.07526*, 2023.

[14] J. Lin and G. H. Lee, "Multi-view multi-person 3D pose estimation with plane sweep stereo," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 886–11 895.

[15] D. Bermuth, A. Poeppel, and W. Reif, "SimpleDepthPose: Fast and Reliable Human Pose Estimation with RGBD-Images," *arXiv preprint arXiv:2410.18723*, 2025.

[16] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang, "Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2020, pp. 5386–5395.

[17] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh, "Panoptic studio: A massively multiview system for social motion capture," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3334–3342.

[18] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1325–1339, jul 2014.

[19] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic, "3D pictorial structures for multiple human pose estimation," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1669–1676.

[20] V. Srivastav, T. Issenhuth, A. Kadkhodamohammadi, M. de Mathelin, A. Gangi, and N. Padoy, "MVOR: A multi-view RGB-D operating room dataset for 2D and 3D human pose estimation," *arXiv preprint arXiv:1808.08180*, 2018.

[21] H. Tu, C. Wang, and W. Zeng, "VoxelPose: Towards Multi-Camera 3D Human Pose Estimation in Wild Environment," in *European Conference on Computer Vision (ECCV)*, 2020.

[22] H. Ye, W. Zhu, C. Wang, R. Wu, and Y. Wang, "Faster VoxelPose: Real-time 3D Human Pose Estimation by Orthographic Projection," in *European Conference on Computer Vision (ECCV)*, 2022.

[23] T. Wang, J. Zhang, Y. Cai, S. Yan, and J. Feng, "Direct Multi-view Multi-person 3D Human Pose Estimation," *Advances in Neural Information Processing Systems*, 2021.

[24] S. Wu, S. Jin, W. Liu, L. Bai, C. Qian, D. Liu, and W. Ouyang, "Graph-based 3D multi-person pose estimation using multi-view images," in *ICCV*, 2021.

[25] R. Choudhury, K. M. Kitani, and L. A. Jeni, "TEMPO: Efficient multi-view pose estimation, tracking, and forecasting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 14 750–14 760.

[26] V. Srivastav, K. Chen, and N. Padoy, "SelfPose3d: Self-Supervised Multi-Person Multi-View 3D Pose Estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 2502–2512.

[27] J. Dong, W. Jiang, Q. Huang, H. Bao, and X. Zhou, "Fast and Robust Multi-Person 3D Pose Estimation from Multiple Views," 2019.

[28] J. Tanke and J. Gall, "Iterative Greedy Matching for 3D Human Pose Tracking from Multiple Views," in *German Conference on Pattern Recognition*, 2019.

[29] H. Chu, J.-H. Lee, Y.-C. Lee, C.-H. Hsu, J.-D. Li, and C.-S. Chen, "Part-aware measurement for robust multi-view multi-human 3D pose estimation and tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1472–1481.

[30] M. Munaro, A. Horn, R. Illum, J. Burke, and R. B. Rusu, "OpenPTrack: People tracking for heterogeneous networks of color-depth cameras," in *IAS-13 Workshop Proceedings: 1st Intl. Workshop on 3D Robot Perception with Point Cloud Library*. Citeseer, 2014, pp. 235–247.

[31] F. Vicentini, N. Pedrocchi, M. Giussani, and L. M. Tosatti, "Dynamic safety in collaborative robot workspaces through a network of devices fulfilling functional safety requirements," in *ISR/Robotik 2014; 41st International Symposium on Robotics*. VDE, 2014, pp. 1–7.

[32] L. Wang, B. Schmidt, and A. Y. Nee, "Vision-guided active collision avoidance for human-robot collaborations," *Manufacturing Letters*, vol. 1, no. 1, pp. 5–8, 2013.

[33] "Sensor fusion for human safety in industrial workcells, author=Rybski, Paul and Anderson-Sprecher, Peter and Huber, Daniel and Niessl, Chris and Simmons, Reid," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012, pp. 3612–3619.

[34] H. Liu and L. Wang, "Collision-free human-robot collaboration based on context awareness," *Robotics and Computer-Integrated Manufacturing*, vol. 67, p. 101997, 2021.

[35] "Freemove," *Veo Robotics Inc*, 2024. [Online]. Available: https://www.veobot.com/freemove,https://www.veobot.com/blog/2020/10/16/how-collaborative-is-your-robot-a-practical-approach

[36] J. T. C. Tan and T. Arai, "Triple stereo vision system for safety monitoring of human-robot collaboration in cellular manufacturing," in *2011 IEEE International Symposium on Assembly and Manufacturing (ISAM)*. IEEE, 2011, pp. 1–6.

[37] P. A. Lasota, G. F. Rossano, and J. A. Shah, "Toward safe close-proximity human-robot interaction with standard industrial robots," in *2014 IEEE international Conference on Automation Science and Engineering (CASE)*. IEEE, 2014, pp. 339–344.

[38] A. Cherubini, R. Passama, A. Crosnier, A. Lasnier, and P. Fraisse, "Collaborative manufacturing with physical human–robot interaction," *Robotics and Computer-Integrated Manufacturing*, vol. 40, pp. 1–13, 2016.

[39] L. Antao, J. Reis, and G. Gonçalves, "Voxel-based space monitoring in human-robot collaboration environments," in *2019 24th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*. IEEE, 2019, pp. 552–559.

[40] M. Costanzo, G. De Maria, G. Lettera, C. Natale, and D. Perrone, "A multimodal perception system for detection of human operators in robotic work cells," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. IEEE, 2019, pp. 692–699.

[41] H. Rajnathsing and C. Li, "A neural network based monitoring system for safety in shared work-space human-robot collaboration," *Industrial Robot: An International Journal*, vol. 45, no. 4, pp. 481–491, 2018.

[42] L. M. Amaya-Mejía, N. Duque-Suárez, D. Jaramillo-Ramírez, and C. Martinez, "Vision-based safety system for barrierless human-robot collaboration," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 7331–7336.

[43] "Abschlussbericht," *KI Absicherung: Safe AI for Automated Driving*, 2022. [Online]. Available: https://www.ki-absicherung-projekt.de/en/

[44] T. Fingscheidt, H. Gottschalk, and S. Houben, *Deep neural networks and data for automated driving: Robustness, uncertainty quantification, and insights towards safety*. Springer Nature, 2022.

[45] *safe.trAIn: Safe AI using a driverless regional train as an example*, 2024. [Online]. Available: https://safetrain-projekt.de/en/