

# Visual quality and safety monitoring system for human-robot cooperation

---

Kozamernik, Nejc; Zaletelj, Janez; Košir, Andrej; Šuligoj, Filip; Bračun, Drago

*Source / Izvornik:* **The International Journal of Advanced Manufacturing Technology, 2023, 128, 685 - 701**

**Journal article, Accepted version**

**Rad u časopisu, Završna verzija rukopisa prihvaćena za objavljivanje (postprint)**

<https://doi.org/10.1007/s00170-023-11698-2>

*Permanent link / Trajna poveznica:* <https://urn.nsk.hr/urn:nbn:hr:235:176788>

*Rights / Prava:* [In copyright](#)/[Zaštićeno autorskim pravom.](#)

*Download date / Datum preuzimanja:* **2025-02-17**

*Repository / Repozitorij:*

[Repository of Faculty of Mechanical Engineering  
and Naval Architecture University of Zagreb](#)





# Visual quality and safety monitoring system for human-robot cooperation

Nejc Kozamernik<sup>1</sup> · Janez Zaletelj<sup>2</sup> · Andrej Košir<sup>2</sup> · Filip Šuligoj<sup>3</sup> · Drago Bračun<sup>1</sup> 

Received: 26 December 2022 / Accepted: 31 May 2023  
© The Author(s) 2023

## Abstract

Efficient workspace awareness is critical for improved interaction in cooperative and collaborative robotic applications. In addition to safety and control aspects, quality-related tasks such as the monitoring of manual activities and the final quality assessment of the results are also required. In this context, a visual quality and safety monitoring system is developed and evaluated. The system integrates close-up observation of manual activities and posture monitoring. A compact single-camera stereo vision system and a time-of-flight depth camera are used to minimize the interference of the sensors with the operator and the workplace. Data processing is based on a deep learning to detect classes related to quality and safety aspects. The operation of the system is evaluated while monitoring a human-robot manual assembly task. The results show that the system ensures a high level of safety, provides reliable visual feedback to the operator on errors in the assembly process, and inspects the finished assembly with a low critical error rate.

**Keywords** Human-robot cooperation · Vision systems · Safety · Quality · Assembly supervision

## 1 Introduction

Manual activities in combination with robots are useful for complex tasks that require human skill, precision, understanding of seen, and a high degree of self-control [1, 2]. In human-robot cooperation (HRC), it is necessary to support the worker by appropriate safety measures and, if possible, by monitoring task performance. Any system that supports

HRC should first and foremost provide safety by ensuring that no collisions are possible between the robot, the human, and the equipment. This is usually done by using additional sensors, preferably tactile or visual sensors, and supporting control systems [3]. It is desirable that safety systems not burden the worker physically, by equipping him with sensors or beacons, or psychologically, by making him feel that he is under constant surveillance. A crucial component of HRC is to allow the worker to freely set up the work environment according to their preferences, work habits, and the task at hand [4]. This means that the positions of parts and tools can move freely within the workplace, which presents an additional challenge to safety and monitoring systems.

### *Research objectives and proposed solution*

An important aspect of manual work is the quality of task performance, since manual work in practice has a high variability. To this end, it would be desirable for the perception system to monitor the manual tasks, provide visual feedback to the operator on the correctness of the work operations, and perform a final quality check of the work result at the end.

Machine vision systems are the most sensible choice for monitoring manual tasks and ensuring safety without burdening the worker with additional sensors [5]. Previous studies have demonstrated that research on collaborative robotics or perception technologies has been predominantly concen-

✉ Drago Bračun  
drago.bracun@fs.uni-lj.si

Nejc Kozamernik  
nejc.kozamernik@fs.uni-lj.si

Janez Zaletelj  
janez.zaletelj@fe.uni-lj.si

Andrej Košir  
andrej.kosir@fe.uni-lj.si

Filip Šuligoj  
filip.suligoj@fsb.hr

<sup>1</sup> Faculty of Mechanical Engineering, University of Ljubljana, Aškerčeva ulica 6, Ljubljana 1000, Slovenia

<sup>2</sup> Faculty of Electrical Engineering, University of Ljubljana, Tržaška cesta 25, Ljubljana 1000, Slovenia

<sup>3</sup> Faculty of Mechanical Engineering, University of Zagreb, Ivana Lučića 5, Zagreb 10000, Croatia

trated on individual tasks such as safety, ergonomics, or part inspection. In this study, we introduce a multi-functional system that is capable of performing various tasks related to safety and quality simultaneously.

The Visual Quality and Safety Monitoring (VQSM) system proposed in this study combines two monitoring systems: one for close-range manual work monitoring and another for far-range human posture monitoring. The close-range system is capable of identifying the hands of the operator and quality-related aspects to monitor assembly and assess the final product quality. Meanwhile, the posture monitoring system provides an additional layer of safety implementation. The integration of both systems results in a comprehensive set of functions that can be summarized as follows:

- Monitoring of manual work by detecting assembly errors and forwarding them to the operator screen,
- Final quality control to determine if the assembly is correct,
- Increased safety through continuous monitoring of the operator's hands at close range and human posture at far range, and the additional conditioning that controls the robot's movements based on the results of the final quality inspection.

The goal of this study is to demonstrate that the integration of close-view monitoring (Sect. 2.1) and posture monitoring (Sect. 2.2) can offer a cost-effective multifunctional solution, which is capable of performing different tasks simultaneously. This research stands out due to its minimalist hardware design, which utilizes cutting-edge building blocks. Specifically, a SCSV (Single Camera Stereo Vision) system, which is capable of capturing views from two distinct angles within a single image. This eliminates the need for synchronizing and capturing images from multiple cameras, resulting in a streamlined system that simplifies the hardware and image processing aspects, consolidating them into a single, efficient pipeline.

The continuation of Sect. 1 begins with a review of related works. Then, in Sect. 2, a description of the proposed methods and technological components is given. In Sect. 3, the proposed methods are tested in a series of experiments involving human operators. Finally, a discussion follows in Sect. 4, and a conclusion of the research is given in Sect. 5.

#### *Related work*

A review of the main safety systems that have been proposed and applied in industrial robotic environments and proven to contribute to the achievement of safe HRC work is given in [6] and in [7]. Literature review regarding the use of vision systems in HRC safety [8] shows that they can be classified into four categories: (a) systems that compute

the distance between several points and moving obstacles, for example, between robotic joints and a human utilizing a depth camera; (b) collision avoidance systems; (c) human intent detection systems; and (d) systems that use visualization and monitoring of safety zones created by projecting virtual optical barriers with lasers or projectors. For many computer vision tasks, such as classification, segmentation, or object detection, the deep learning (DL) models have become a standard. Powered by massive databases, modern GPU's, and fast developing algorithms, there are tasks where DL can achieve state-of-the-art performance near human level [9]. The use of DL in human action recognition is the subject of intensive research because it can provide the basis for robot action planning. It is also a promising tool for supervision of manual assembly tasks. Object detection is aimed at determining the presence and localization of objects within an image. These objects can be categorized into several predefined classes; alternatively, only one type of object can be searched.

Zamora-Hernández et al. [10] proposed an object recognition-based architecture for monitoring an operator during manual assembly in a manufacturing cell to reduce potential errors during the manufacturing process. Their approach focuses on identifying tools, components, and actions in the assembly process using deep learning techniques and a general-use language for describing actions. An investigation of DL as a data driven technique for continuous human motion monitoring and future human-robot collaboration needs prediction is provided by Wang et al. [11]. The method achieved recognition accuracy of over 96% in an engine assembly case study. In [12], Zhang et al. proposed an industrial part recognition algorithm based on the DL real-time object recognition model YOLOv3 in intelligent assembly, and Park et al. [13] presented an empirical study on a process management system that recognizes specific engine parts in a ship assembly line using YOLO to estimate the process rate for each workshop in real time. H. Rajnathsing et al. [14] proposed a monitoring system for the shared HRC workspace that complements the robot to locate the human operator and always ensure that a minimum safe distance is maintained from the robot to its human partner. The monitoring system consists of four neural networks, namely, an object detector, two neural networks responsible for evaluating detections, and a simple custom speech recognizer. Although effective, their approach is solely focused on ensuring safety, while relying only on three independent 2D cameras and detections provided by a deep object detector, without taking into account the operator's three-dimensional pose. A smart operator advice model for a human-robot coexisting assembly line is proposed by Wang et al. in [15], which similarly uses three independent cameras and a deep learning algorithm to detect relevant classes. While the study

presents an innovative operator advice and guidance system, it also only considers 2D image processing technology and does not address safety concerns in detail. Collaboration between human operators and industrial robots in assembly operations with a focus on safety and simplified interaction involving wearable devices used by the operator is presented by Papanastasiou et al. in [16]. They employed a manual guidance module, a contact sensor named “safety skin,” and a vision system for recognition and tracking of objects. They also used advanced user interfaces, including audio and haptic commands accompanied by augmented reality technology, to support the operator and provide awareness by visualizing information related to production and safety aspects. The system heavily relies on additional equipment that has to be worn by the operator, which increases complexity compared to solely vision-based methods and could affect the operator’s comfort and performance. Q. Xiong et al. [17] presented an integrated method that uses optical flow images to encode temporal information of human motion as input to a two-flow CNN structure for simultaneous analysis of spatial and temporal information of human motion. Transfer learning is investigated to transfer the feature extraction capability of a pre-trained CNN to a production scenarios. While Xiong et al.’s approach improves human action recognition accuracy, their method only addresses this issue and does not assess safety or provide quality monitoring. H. Liu et al. [18] present a context awareness-based collision-free human-robot collaboration system that can provide human safety and assembly efficiency at the same time. The system can plan robotic paths that avoid colliding with human operators while still reach target positions in time. J. Zhang et al. [19] present a method to analyze visual observations of human actions in an assembly environment and predict the future motion trajectory of the human operator for online robot action planning and execution. Although H. Liu et al.’s and J. Zhang et al.’s methods are highly advanced in terms of safety and control aspects, their works solely focus on those tasks and do not consider any assembly contextual information. Flacco et al. [20] presented a fast method for evaluating distances between the robot and moving obstacles (including humans) based on the concept of depth space. Their approach emphasizes on collision avoidance with arbitrary obstacles, yielding promising outcomes. Nevertheless, it lacks the capability to recognize human activities in the context of human-robot collaboration. Liu et al. [21] presented a DL-based multimodal fusion architecture that includes three modalities: voice command, hand motion, and body motion. Three unimodal models are first trained to extract features, which are then fused to share the representation. The authors demonstrate the accuracy of the fused model in comparison to the unimodal models through experiments. However, the inves-

tigation is restricted to command recognition, and there is no further exploration of the human-robot collaboration context.

Various types of cameras, stereo vision, and 3D measuring systems or their combination are used for a digital description of the HRC scene. J. Arents et al. [3] published a review of HRC articles published between 2010 and 2021, which show that 3D cameras are the most often utilized sensors for human tracking or gesture recognition in HRC environments. Such an example is an advanced RGB-D vision system presented by Olesen et al. [22]. The authors presented a collaborative robot cell assembled with off-the-shelf components designed for random bin-picking and robotic assembly applications. They used an optimized version of YOLO to detect the arbitrarily placed components of the mobile phone on the working space. The success rate of the final assembly was not ideal due to multiple sources of potential errors; additionally, the system featured only basic human awareness. A stereo vision system [23] is developed for safety monitoring in human-robot collaboration cell production. A multi camera system is used to capture images for tracking of color areas on the human operator and to produce three pairs of stereo vision to improve the robustness towards lost tracking and occlusion tolerance. Their investigation centered on joint position tracking and utilized traditional computer vision methods rather than deep learning. M. Melchiorre et al. [24] introduced a control strategy for human-robot hand-over tasks, where human pose estimation was done using a duplex Kinect v2 sensors system to reduce problems related to occlusions of the sensors. While the cited paper presents an effective control strategy for human-robot hand-over tasks, their prediction scheme is limited to a small portion of the human arm.

In summary, most research focuses on human perception, object perception, various quality aspects of objects, human-robot cooperation, and human assistance in performing various isolated tasks. However, there is little research that covers various complete solutions for quality-related estimation of a process in HRC involving manual work. Moreover, integration of such systems is usually demanding in terms of both hardware and software, as multiple sensors and equipment need to be assembled and synchronized. In order to overcome the aforementioned limitations, in this paper, we propose a ready to use system that prioritizes quality-related functions by maximizing the capabilities provided by DL, in addition to safety and operator assistance. Furthermore, the challenge of system complexity is addressed from a hardware perspective by proposing a vision system that uses a minimalist hardware design. At the same time, the challenge of designing a user-friendly and intuitive-to-use system is addressed.

## 2 VQSM system implementation

To illustrate the operation of the VQSM system, we begin by considering the HRC workstation depicted in Fig. 1. In this setup, the operator and the robot share a common *workspace* with no physical barriers separating them. The operator can freely arrange the tools and assembly parts on the worktable (8) as per their requirement and perform the assembly task within a smaller portion of the worktable, referred to as the *assembly space* (7). Once the assembly is completed and verified as correct, an industrial robot (4) picks up the finished assembly and transfers it to the next operation. The robot is allowed to move only when the *workspace* is safe, which means that both hands are in a safe location (6).

The VQSM system is designed to provide safety and quality inspection by continuously monitoring the workplace and holding the robot's movements if the operator's hands enter the *workspace* during the robot's operation. The VQSM system supervises the assembly process by using the Single Camera Stereo Vision (SCSV) system (2), which captures stereo images (1) that are analyzed by a deep object detector to identify quality-related classes such as various defects and the correctness of the final assembly (discussed in Sect. 2.1). The analysis results are displayed on the screen (5) to provide visual feedback to the operator. The system incorporates an enhanced safety function by combining the hand detection signal from the SCSV deep object detector with a posture monitoring system implemented by the Kinect (3) (discussed in Sect. 2.3).

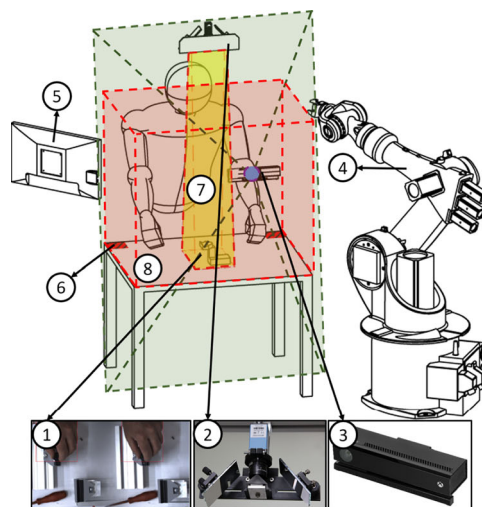
Figure 2 displays the main functions and data flow of the VQSM system where the SCSV image is fed into a deep object detector YOLOv3 [25]. The object detector returns predicted bounding boxes (image coordinates of the ori-

gin, height, and width) along with their objectiveness score and class probabilities of detected objects. Prediction is additionally analyzed to achieve the following functionality (discussed in Sect. 2.1):

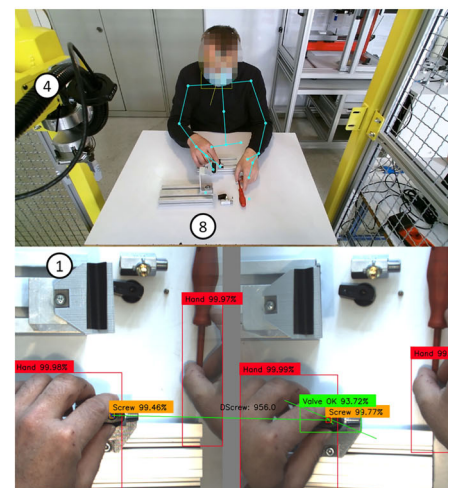
- *Assembly supervision*: the system continuously checks for different classes of errors, which can occur during assembly. By relaying them to the computer display, it warns the operator about errors in assembly procedure.
- *Final assembly inspection*: check whether the final assembly is correct and returns *assembly OK* signal. Complexity of this section depends on the inspection task and is specific to particular object of assembly. For the demonstrated valve assembly, the handle orientation is checked as well as the screw whether it is tight or just placed into the corresponding hole. For this purpose, the depth information from stereo vision assists in verifying relations between assembly components.
- *Safety*: the object detector continuously checks whether the operator's hands are present in SCSV field of view and generates *SCSV hands* signal.
- *Localization*: since the operator is able to configure the workplace according to preferences, the robot pick-up holder can be positioned anywhere within the SCSV field of view. The exact position of the assembled object in 3D space is determined utilizing disparity and triangulation principle.

In continuation of this chapter, we provide details of the proposed system, including hardware setup, training of the object detection model for assembly supervision, localization, computation of safety score from the human skeleton

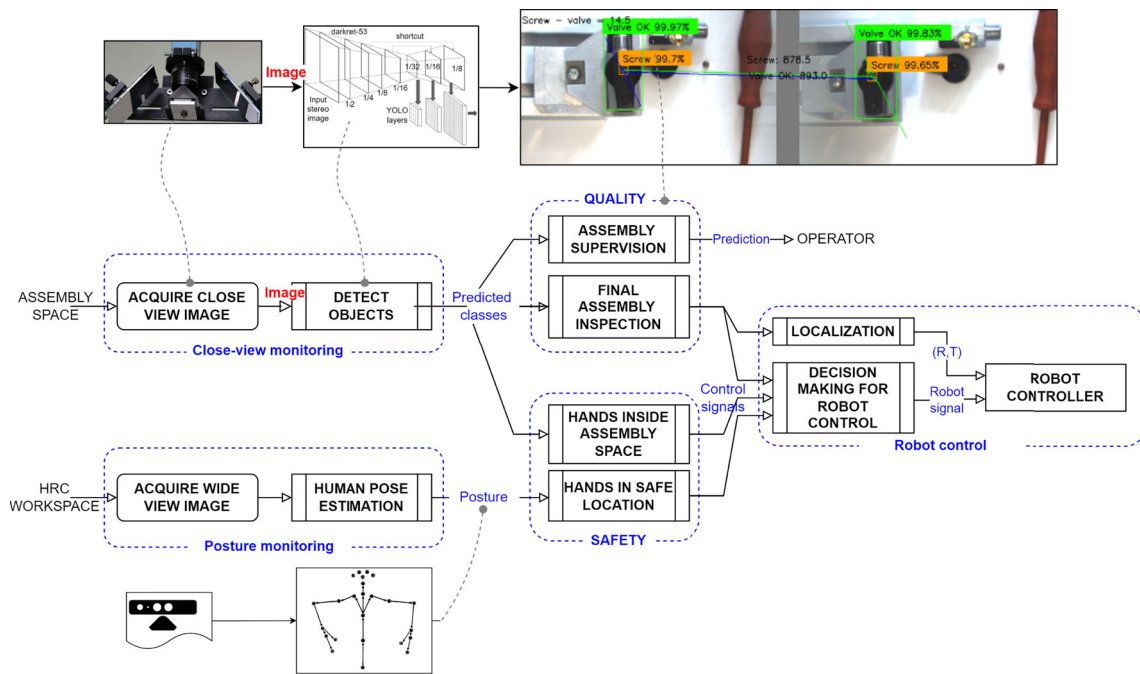
**Fig. 1** **a** The VQSM system setup. **b** The views of the Kinect (3) and SCSV system (2). SCSV system captures a stereo image (1) of the *assembly space* (7), which is part of the worktable (8), and analyzes it by a deep object detector. The analysis results are displayed on the screen (5) to provide visual feedback to the operator. Safe location for hands (6) and the robot (4)



(a) The VQSM system setup



(b) Kinect (top) and SCSV (bottom) views



**Fig. 2** Functions and data flow in the VQSM system. The SCSV system observes the assembly space and uses deep learning object detection to locate and detect classes of interest, such as the operator’s hands and assembly defects. At the end of the assembly process, classes related to the quality of the finished part are also detected to determine if the part is acceptable. Posture monitoring provides redundant safety by monitoring the upper body skeleton and verifying that the hands are in a safe location. The close range and posture monitoring signals are used

in decision making for robot control to determine what state the system is in, i.e., empty workspace, manual assembly, assembly finished workspace not safe, and robot action. The robot action is allowed when both hands are in a safe position and the correctly assembled object is in the pickup holder. The system does not require any specific gestures from the operator to indicate the completion of the assembly process to the robot

data, overall data processing, and definition of quality and safety signals.

### 2.1 Close-view monitoring of manual work

The close-view monitoring of *assembly space* is carried out using a Single Camera Stereo Vision (SCSV) system to perform object detection and quality control. The choice of this system was based on the need for a small and compact vision system where one image contains a view from two different angles, eliminating the need for synchronizing and simultaneously capturing images from multiple cameras [26]. The complete vision system from the hardware to image processing is simplified to just one pipeline. Figure 3 shows an optical setup of the SCSV system. Such an optical arrangement provides a close view of the task, like that of a worker assembling a product. The top view area, i.e., assembly space covered by the SCSV measures approximately 200 x 300 mm.

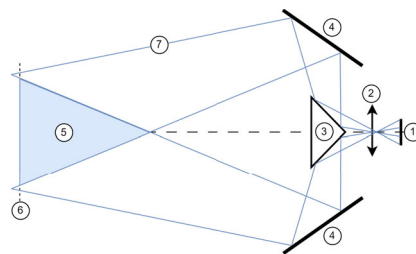
The SCSV system is calibrated using a checkerboard and a camera calibration software [27]. Object coordinates in 3D space are calculated using intrinsic camera parameters and

disparity as described in [28]. To facilitate robot localization tasks, it is necessary to establish a common coordinate frame between the work table and the robot. This is achieved through a calibration process with the checkerboard. The mirrors of the SCSV were fixed during all experiments.

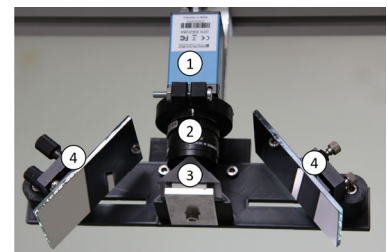
### Deep-learning based object detection

State-of-the-art deep learning-based object detection methods are grouped typically into single-stage or two-stage methods. Two-stage methods, such as Faster R-CNN [29] or Mask R-CNN [30], prioritize accuracy by first finding regions of interests and then sending the region proposals down the pipeline for object classification and bounding-box regression. Single-stage methods, such as YOLO [25] and SSD [31], prioritize speed and treat object detection as a simple regression problem by taking an input image and learning the class probabilities and bounding box coordinates. Such models are faster and are appropriate for real-time object detection but suffer from lower accuracy rates compared to two-stage models due to lower number of region proposals for detection.

**Fig. 3** The camera's image detector (1) captures an image produced by a lens (2). A 90° prism (3), featuring mirrored surfaces, splits the view field into two directions. The views, illustrated by light rays (7), are directed to a common observation point (5) by two additional mirrors (4) rigidly mounted in specific positions and orientations. The mirrors (4) are positioned and rotated to ensure that the virtual positions of the stereo cameras match an approximate distance between human eyes. The intersection of the two views (5) and the depth of field between the camera and the worktable (6) ranges from 0.3 to 0.9 m



(a) Schematic of the SCSV setup



(b) Photograph of the SCSV setup

**Fig. 4** Object class examples and data split

| Class                       | Example 1 | Example 2 | Training dataset<br>(no. instances) | Test dataset<br>(A1)<br>(no. instances) |
|-----------------------------|-----------|-----------|-------------------------------------|---|
| 1) Missing handle           |           |           | 158                                 | 40                                      |
| 2) Missing screw            |           |           | 476                                 | 254                                     |
| 3) Wrong handle orientation |           |           | 508                                 | 200                                     |
| 4) Valve OK                 |           |           | 883                                 | 488                                     |
| 5) Screw                    |           |           | 1327                                | 514                                     |
| 6) Hand                     |           |           | 4631                                | 2441                                    |
| Total number of frames      |           |           | 3414                                | 882                                     |

To reduce the labeling effort of the training dataset, other methods designed to train with limited data could also be considered, such as few-shot [32] or weakly supervised object detection, which has already been used for robotics applications [33]. Training with reduced data was not the key issue of this research, and the VQSM system is envisioned to be implemented in mass-production scenarios, where the same product is manufactured in large quantities for a long time and the data-labeling process pays off.

Object detection in VQSM system implements YOLOv3 [34] model as it reaches a good compromise between speed and accuracy and has good real time capabilities. It assigns multiple anchor boxes to feature maps in three different resolutions. Small objects are accurately detected from the anchors in low-level feature maps with small receptive fields and large objects from the anchors in feature maps with large receptive fields. YOLOv3 uses a powerful Darknet 53 as its backbone with several sets of residual blocks. This single-stage model can achieve similar accuracy as two-stage Faster R-CNN while maintaining real-time efficiency.

Most HRC support systems usually focus on the aspect of safe human-robot interaction and recognize classes related to the human body or the robot. Quality-related classes should be selected based on the manual task and quality requirements. To do this, the assembly process must be carefully considered in advance to create a list of scenarios in which it is correct and in which the operator has performed an incorrect action. Thus, the final list of object classes consists of the correct states and, if possible, all incorrect states.

Sometimes, it is impossible to predict all errors in advance because some errors are rare. In such cases, anomaly detection methods can help identify unusual cases. The assembly process is an exception in this regard, as there are only a limited number of variants where components might not fit together properly or might be missing altogether. Furthermore, we have the ability to either simulate these errors in reality or create synthetically rendered images from CAD models, which means that large supervised learning datasets can be created to support the desired capabilities.

For the demonstrated valve assembly task, the first group of classes relates to assembly supervision and possible errors in the assembly procedure: (1) *missing handle*, (2) *missing screw*, and (3) *wrong handle orientation*. The second group of classes refers to the final quality inspection. Class (4) is a correctly assembled valve, i.e., *valve OK*, class (5) *screw* is used for disparity search to determine if the screw is tightened. Safety-related class (6) denotes the operator's hand, i.e., *hand* (see Fig. 4). As mentioned above, the choice of classes depends on the assembly task. If only a safety function is required, the classes related to the operator's body parts and some potentially hazardous equipment are needed.

## Training of the object detector

As an input for training of the object detector, several videos of valve assembly are captured with the help of four operators (three men and one woman). The operators were allowed to set up the workspace according to their preferences. The only stipulation was that the activities should take place within the SCSV field of view. Each video contained two attempts of the valve assembly. In the first, operators were instructed to intentionally assemble the valve incorrectly to collect data for all object classes, and in the second, the task was performed in the correct manner. The videos were captured in RGB color space with a resolution of 1920 x 1080 pixels and a frame rate of 10 frames/s.

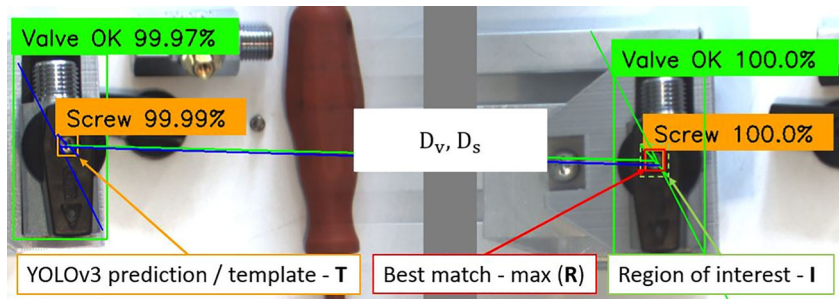
A total of 3414 frames containing objects of interest were extracted from the videos, manually labeled using Dark-Label software, and used as the training dataset. Since the images captured by the SCSV system contain a view from two different angles, they were divided into two sub-images corresponding to the left and right view, each with a half resolution of 920 x 1080 pixels. The 80-pixel wide gap between them is excluded due to optical overlap. For each image, only the left or right view image is labeled, with about half of the labeled images belonging to each. Image augmentation techniques are used to enhance the training dataset. Transformations include vertical and horizontal flipping, scaling of exposure and saturation, hue rotation, and random cropping at various scales. The scaling factors and hue rotation value are drawn randomly from uniform distributions. The factors for scaling exposure and saturation are between 2/3 and 3/2, and the values for random hue rotation are between  $-36^\circ$  and  $36^\circ$ . The random area of the crop with respect to the area of the original image is between 1/4 and 1. By applying multiple transformations, seven additional images are created from each original training image in each epoch, increasing the number of training images from 3414 to 27,312.

The YOLOv3 model is trained for 60 epochs using the Adam optimizing algorithm with the initial learning rate of  $1e-4$ . The batch size is set to four, and for each batch, the resolution of the input images randomly varied from 384 x 384 to 448 x 448 pixels with a step of 32 pixels. The training process took about 11 h on a Nvidia GeForce RTX 2060 GPU.

The performance of the object detector is evaluated on a new dataset created from previously unused video (Test dataset A1 described in Sect. 3). In this dataset, both left and right view stereo images are labeled in each frame, resulting in a higher number of object instances per frame (see Fig. 4, which shows the number of instances for each class and the size of dataset). Using the standard average precision (AP) evaluation metric for object detection, the trained object detector achieved values of 0.982 for the *missing handle* class, 0.898 for *missing screw*, 0.735 for *wrong handle orientation*, 0.945 for *valve OK*, 0.986 for *screw*, and 0.858



**Fig. 5** Disparity search for the class screw. The object bounding box predicted by the object detector in the left stereo image is used as a template **T**. The bounding box of the same object in the right stereo image is enlarged by 20% and used as the region of interest **I**



for the *hand* object class. The calculated mean average precision (mAP) for all classes was 0.901.

### Inference and post-processing

The detector uses as input full-frame images containing both the left and right stereo views. This means that only one inference is required to detect the objects in both stereo views, which is another advantage of SCSV. Before the image is fed into the object detector, the left and right views are separated by an 80 pixel wide vertical gray strip to remove overlap effects. Then, the image is rescaled to 896 x 448 pixels and as such used as input to the detector. Predictions with an objectivity score above 0.7 are used in further decision making, and intersection over union (IoU) threshold of 0.5 is used for non-maximum suppression (NMS). Since object detection for monitoring the assembly process is performed simultaneously for both left and right stereo images, it is possible that different (conflicting) classifications will occur on them. For that reason, the object with the lower objectivity score is suppressed so as not to confuse the operator.

### Disparity search

Inspection and localization for the robot pick-up tasks are based on the disparity calculation. In the search for disparities, we rely on the bounding boxes of the object detector predicted for the same object in the left and right stereo images. The disparity search between the bounding boxes only lacks accuracy. In order to improve disparity search accuracy, a template matching method based on the normalized correlation coefficient from OpenCV [35] is used. Example of best matching result for the class *screw* is depicted in Fig. 5.

Disparity of *screw*  $D_s$  is calculated between the center  $x$  coordinates of the best match bounding box in the right image  $x_{s,R}$  and the left bounding box  $x_{s,T}$  predicted by the object detector:

$$D_s = x_{s,R} - x_{s,T} \quad (1)$$

To determine whether the *screw* is tightened or not, the disparities between the classes *valve OK* and *screw* are sub-

tracted:

$$D_t = D_v - D_s \quad (2)$$

Disparity of the class *valve OK*, i.e.,  $D_v$ , is determined in a similar manner as is described for the class *screw*.

Experiments showed that the proposed system could correctly predict if the screw is tightened or not, with the difference in  $D_t$  of an untightened screw being  $18 \pm 0.5$  pixels and of a tightened screw  $14 \pm 0.5$  pixels depending on the rotation of the valve and the precision of the predicted bounding boxes. In practice, this equals to approximately 4 mm difference in depth. In subsequent system evaluation experiments (Sect. 3), the threshold value of  $D_t$  is set at 16 pixels, i.e., if the calculated  $D_t$  is below the threshold, the screw is considered tightened and *assembly OK* signal is set to True.

## 2.2 Posture monitoring

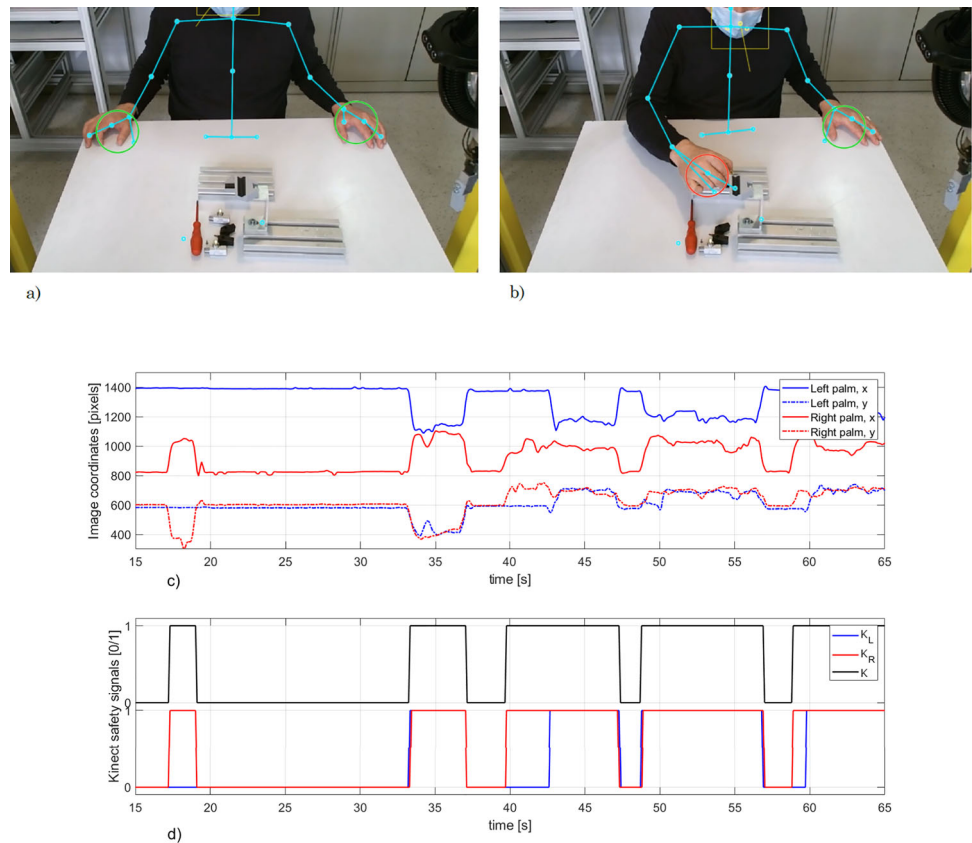
The vision system for posture monitoring is positioned toward the operator as depicted in Fig. 1 (3) to capture the upper part of the body and to avoid robotic arm in obstructing the line of sight to the operator. MS Kinect One (version 2) was chosen as it provides a compact solution for industrial environment and is used for various research purposes [36, 37]. The sensor provides an integrated full-HD RGB image and 512x424 pixel depth image within 4.5m range, at 30 frames per second. Furthermore, it provides real-time monitoring of the human operator (human skeleton detection) by tracking the movements of their limbs, i.e., hands and palms.

Figure 6 illustrates monitoring of the body skeleton and testing if the hands are in a safe location, what is indicated by green circles in a safe location (a) and red elsewhere (b). The  $x$  and  $y$  image coordinates of both hands as function of time are plotted in (c). The *Kinect hands* safety signal  $K$  (d) is determined as

$$K(t) = K_R(t) \text{ OR } K_L(t) \quad (3)$$

The right and left-hand safety signals  $K_R(t)$  and  $K_L(t)$  take up 1 when the hand is inside the workspace and 0 in a

**Fig. 6** Kinect hands safety signals: **a** hands in safe position; **b** right hand inside assembly space; **c** the system estimates the position of each hand; **d** the system generates hands safety signals



safe location. For example, at 42 s, the right hand is inside the assembly area as shown in (b),  $K_R = 1$ ,  $K_L = 0$ , and the Kinect hands signal is  $K = 1$ .

### 2.3 Decision making for robot control

The Kinect system estimates the operator’s posture and provides the *Kinect hands* signal, which is 1 if the hands are within the workspace and 0 if they are in a safe location. The SCSV system monitors the assembly space for object detection and quality control, returning the *SCSV hands* signal, which is 1 if the hands are within the assembly space and 0 if they are not. The SCSV system also provides the *Assembly OK* signal, which is 1 if the assembly is correct and 0 if not.

The system’s operation can be described using final state machine methodology (see Fig. 7), with four different states denoted as  $S_0$ ,  $S_1$ ,  $S_2$ , and  $S_3$ .

The  $S_0$  state corresponds to an empty workspace where no hands or assembly parts are present. When hands are detected by either vision system, the system transitions to state  $S_1$ , in which manual assembly is carried out until the assembly is finished and recognized as OK. This triggers a transition to state  $S_2$ , where the assembly is OK and ready for robot task, but the workspace is not safe due to the presence of hands. The robot action is carried out in state  $S_3$ , where the workspace is

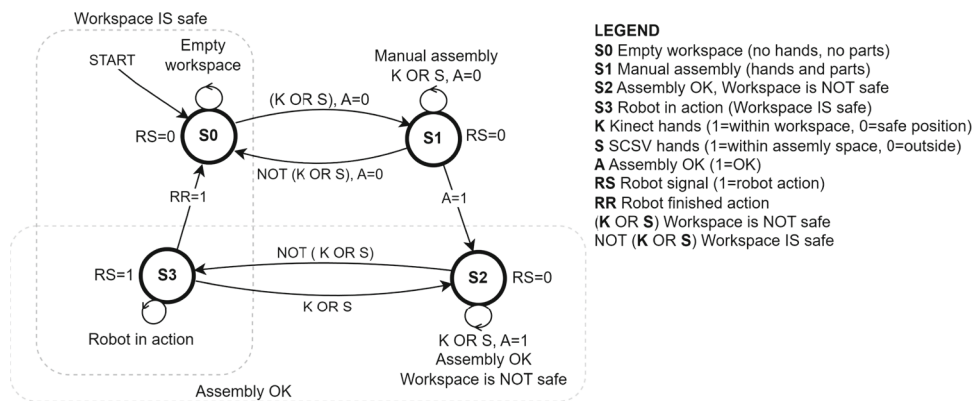
safe and the assembly is OK. The system transitions to this state when hands are out of *workspace* in a safe position and the condition NOT (*Kinect hands* OR *SCSV hands*) is true. If this condition is violated during the robot action, the system transitions back to state  $S_2$  and robot motion is held. Signal for robot motion expressed in Boolean’s logic is thus

$$\text{Robot signal} = \text{NOT} (\text{Kinect hands OR SCSV hands}) \text{ AND } (\text{AssemblyOK}). \quad (4)$$

Once the state  $S_3$  is reached, the robot approaches the assembled object from the side at a safe distance above it, based on the provided location by the SCSV system. The robot then positions its gripper on top of the object and moves down to grasp it. We used a pickup holder to limit the orientation of the object. This entire pickup procedure is designed to minimize the view obstruction of the operator for the Kinect and SCSV system.

### 3 Evaluation of the VQSM system

Evaluation was performed by nine independent experiments with three different human operators. The goal was to estimate the system performance in terms of quality and safety



**Fig. 7** The system can be in four different states. The initial state  $S_0$  corresponds to an empty workspace where no hands or assembly parts are present. To describe an additional states and the transitions between them in the diagram, we introduce short abbreviations for the signals as follows:  $K$  for *Kinect hands*, monitored by the wide-view system,  $S$  for *SCSV hands*, and  $A$  for *Assembly OK*, which are both monitored by the close-view system. When hands are detected by one of the vision systems ( $K$  OR  $S$ ), the system enters the  $S_1$  state, where manual assembly is performed. It remains in this state until assembly is complete and the object detector detects it as OK  $A=1$ . This triggers the transition

to the  $S_2$  state, where the part is OK and ready for the robot task, but the workspace is not safe due to presence of hands. The robot action is performed in the  $S_3$  state, where the workspace is safe and the assembly is OK. The transition to this state occurs when the hands are in a safe position and the condition NOT ( $K$  OR  $S$ ) is satisfied. If this condition is violated at any time during the robot motion, e.g., if the operator's hands are no longer in a safe position, the system returns to the  $S_2$  state and the robot is hold. The action is resumed after the workspace is safe again. When the robot action is finished, which is signaled by the robot controller  $RR=1$ , the system transits back to initial  $S_0$  state

monitoring as defined in Sect. 3.1. The human operators are referred to as persons A, B, and C. None of them participated in the acquisition of the training data set of the object detector. Each operator performed three experiments, e.g., A1-A3, with 2 valve assembly attempts in each. Persons A and B were experienced operators who were familiar with the operation of the VQSM system and with the process of assembly. To simulate unpredictable behavior more realistically, operator C was completely inexperienced and did not know the operation of the system or the assembly procedure.

The operator was provided with the required assembly parts, e.g., a valve body, a handle, a fixing screw and tools like a screwdriver, and two holders. The assembly holder was provided to support the valve body during the assembly, while the pick-up holder was raised by approximately 100 mm above the table and is used for holding the valve during the final inspection, localization, and robot pick-up task. The operator could freely move assembly parts and tools within the assembly space. The correct manual assembly procedure is shown in Fig. 8.

Experimental scenarios were carefully designed for each of the experiment. Person A intentionally made planned mistakes during assembly to create a data set to evaluate the *assembly supervision* and final assembly inspection aspect of the VQSM system. He was consistent in putting his hands in a safe location when the assembly was complete.

Person B worked according to the assembly instructions for all experiments but intentionally kept his hands in the

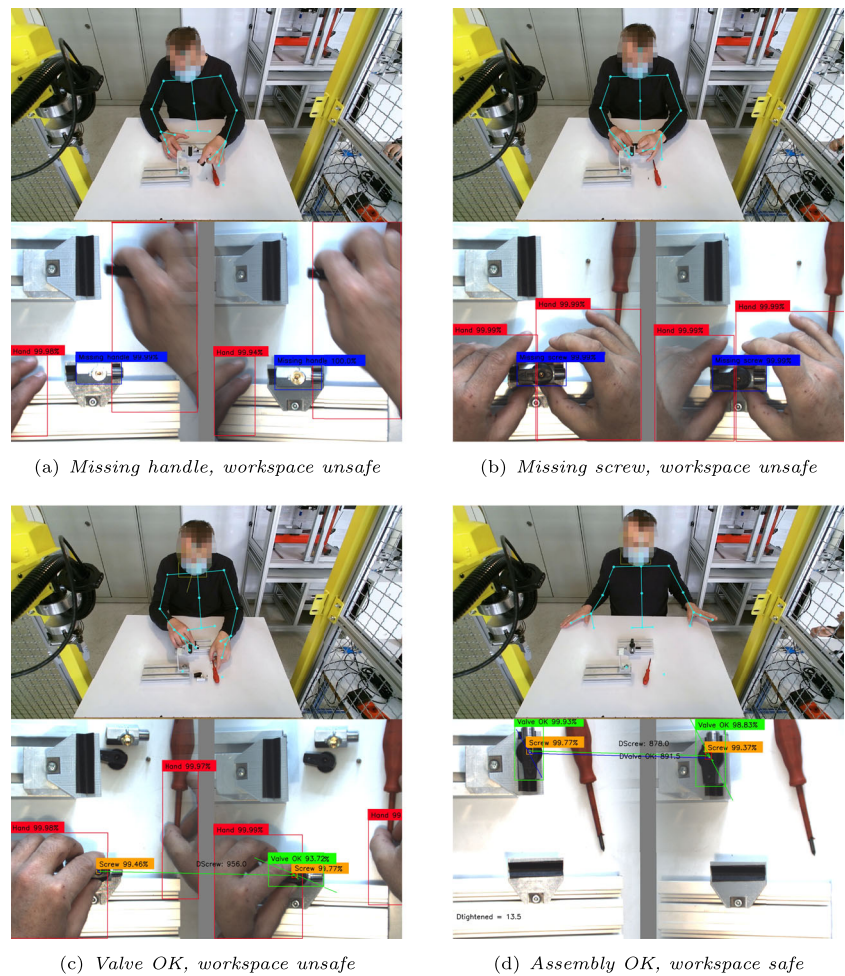
*workspace* for an additional 1–3 s each time after removing them from the *assembly space*.

Person C received incomplete instructions. Before the first experiment, C1, the person saw the correctly assembled valve and where it must be placed at the end. There was no additional information or help from the assembly supervision. Consequently, he was unable to successfully assemble the valve, and experiment C1 was terminated prematurely. In experiment C2, the operator was allowed to observe the SCSV predictions during assembly and cautioned to keep his hands in a safe location at the end. In experiment C3, detailed oral instructions were provided on the assembly procedure.

The evaluation was performed on the same experimental setup used to record the training dataset, since the environment in an industrial setting is usually also highly controlled and conditions such as worktable (background in our case) and other external conditions that could degrade the performance of the object detector do not vary substantially.

Figure 9 shows signal graph for all experiments A1-C3 stacked in the horizontal axis. Vertical black dotted lines mark the separation between different experiments. Vertical axis shows different signals, blue lines represent predicted positive values, and red their positive ground truths. Successful prediction occurs when prediction overlaps with ground truth, i.e., the red and blue line of the same class occur at the same time, and it is unsuccessful when it does not, i.e., only red or only blue.

**Fig. 8** Processed images captured by the Kinect (top) and SCSV (bottom) systems during experiment A1, illustrating the valve assembly process: **a** the valve body is placed into the assembly holder, **b** the handle is attached in the correct orientation, and **c** it is secured with a screw by placing it into the corresponding hole and manually tightening it with a screwdriver. Finally, **d** the assembled valve is moved into the pick-up holder



Below are some more details about signal capture. The SCSV videos are recorded at a frame rate of 10 fps, and Kinect signals are downsampled from 30 fps to 10 fps by averaging to enable data fusion. The safety-related signals *Kinect hands* and *SCSV hands* are labeled and processed at 10 fps for fast response, because robot motion must be terminated immediately in unsafe situations.

Since manual assembly is a relatively slow process and the image scene does not change much between frames at a frame rate of 10 fps, the videos are additionally downsampled to 1-second sections. Processing the input videos into 1-second sections reduces the noise of the signals and makes the labeling process more convenient. Ground truth in videos is labeled according to the following rule: if some class of interest occurred in any frame within 1-second, the ground truth label for a complete 1-second section is assigned to that class. If two or more classes occurred within 1-second section, the prevailing class is labeled as the ground truth. *Assembly supervision* and final *assembly inspection* signals are generated by averaging class scores within 1-second section, and the class with the highest average score is assigned to it.

### 3.1 Target criteria for evaluation of the proposed system

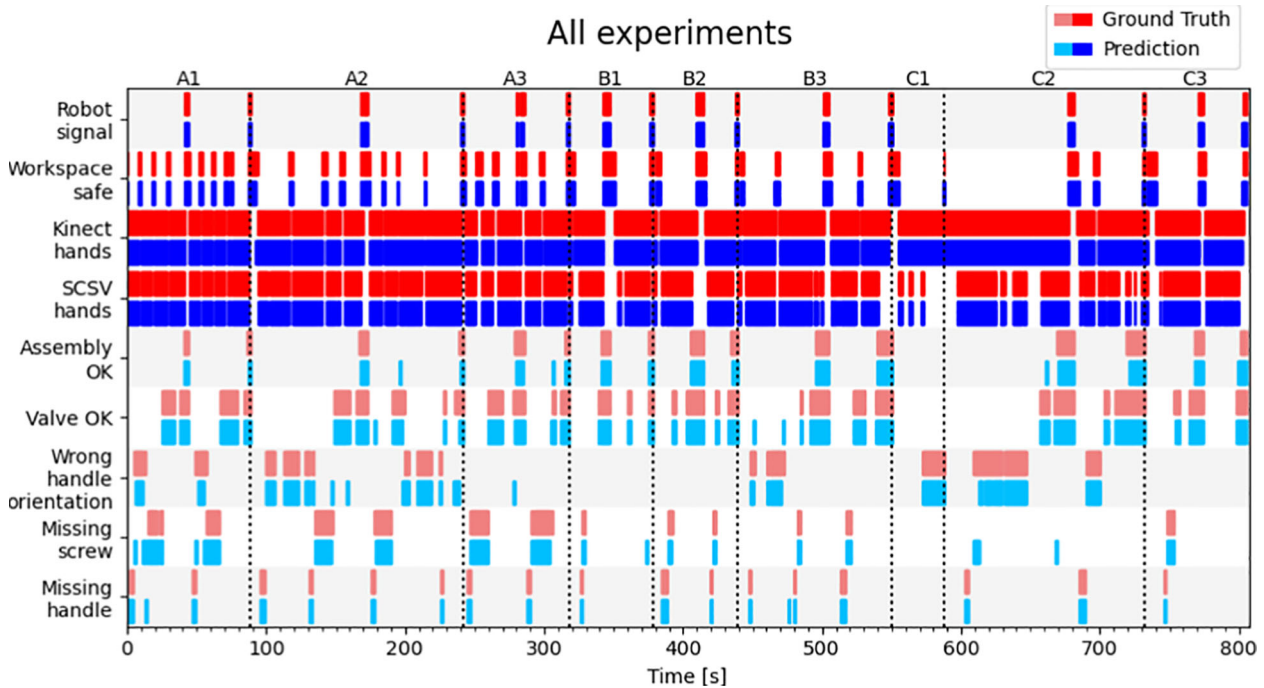
By comparing predicted and ground truth sections, the number of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) instances is counted for each signal. Typical classification evaluation metrics are precision, recall, and F1 score. Precision represents the ratio between the true positives and the total number of positive predictions:

$$precision = TP / (TP + FP) \tag{5}$$

Precision is the more important metric when false detections are costlier than overlooked cases. Recall is the ratio between the TP and the total number of labeled positive instances:

$$recall = TP / (TP + FN) \tag{6}$$

Recall is more important when the overlooked detections are costlier than false predictions. *F1-score* combines recall and precision into a single measure and is calculated as the



**Fig. 9** Signal graph for experiments A1-C3 stacked in the horizontal timeline. The vertical axis shows different signals, blue lines represent predicted positive values, and red their positive ground truths. Successful prediction occurs when prediction overlaps with ground truth, i.e., red and blue line of the same class occur at the same time, and unsuccessful when it does not, i.e., only red or only blue line occurs. Signals

related to final assembly inspection (*assembly OK*) and assembly supervision (*missing handle*, *missing screw*, *wrong handle orientation*, *valve OK*) are colored in lighter shades of red and blue, and signals related to safety are in darker shades (*SCSV hands*, *Kinect hands*, *workspace safe*, *robot signal*)

harmonic mean of both metrics:

$$F1\text{-score} = 2TP / (2TP + FP + FN) \tag{7}$$

The common classification measures for critical and non-critical error rate evaluation are also *false positive rate* (FPR) and *false negative rate* (FNR):

$$FPR = FP / N \tag{8}$$

$$FNR = FN / P \tag{9}$$

*Assembly supervision* is treated as a multiclass classification problem since the object being assembled can only

be in one particular state at any moment during the manual assembly. For assembly supervision, precision is identified as the relevant metric. Incorrect predictions are costlier than overlooked cases since they have the potential to confuse the operators and make them take incorrect actions.

*Final assembly* inspection is treated as a binary classification problem, since the only output signal of interest is whether the object is correctly assembled or not. In the valve assembly scenario, this combines the *valve OK* class and the appropriate  $D_i$  value. The assessment of quality was based on the critical error rate, which is determined by the false positive rate. This rate corresponds to the ratio between falsely classified instances as *assembly OK* and the number of all instances of incorrectly assembled objects.

**Table 1** A confusion matrix for object detection classes relevant to the assembly supervision task

|                 |                          | True class     |               |                          |          |
|-----------------|--------------------------|----------------|---------------|--------------------------|----------|
|                 |                          | Missing handle | Missing screw | Wrong handle orientation | Valve OK |
| Predicted class | Missing handle           | 34             | 2             | 0                        | 0        |
|                 | Missing screw            | 0              | 80            | 9                        | 1        |
|                 | Wrong handle orientation | 0              | 1             | 106                      | 7        |
|                 | Valve OK                 | 0              | 2             | 2                        | 217      |

*Safety assurance* is also treated as a binary classification problem, since the only information of interest is whether or not there is a risk of collision with the operator. The hand detection signals *SCSV hands* and *Kinect hands* are first analyzed independently. Here, the critical error rates are represented by the calculated false negative rates, that is, how many unsafe instances were not successfully detected relative to all unsafe instances. The safety signal *workspace safe* combines hand detection signals from both systems using logic NOR function. Therefore, in this case, the critical error rate is represented by the calculated false positive rate, that is, how many unsafe instances is classified as safe in relation to all unsafe instances.

*Robot signal* is the final signal received by the robot, and it combines *workspace safe* with the *assembly OK* signal. The assessment of the VQSM system safety was based on the critical error rate of the *robot signal* and is represented by the calculated false positive rate.

### 3.2 Evaluation of assembly supervision

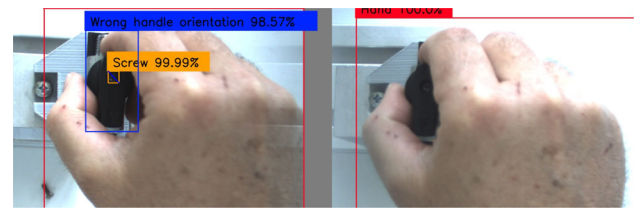
In this section, we aim to evaluate the performance of the SCSV system in terms of *assembly supervision*. Table 1 shows the confusion matrix for all classes relevant to *assembly supervision* task. The relevant performance metrics recall, precision, and F1 score are shown in Table 2. Assembly precision is calculated as a weighted average across all classes.

Tests have shown that the precision is affected by several phenomena, e.g., missing or incorrect predictions due to partial occlusion of objects by the hands during the assembly process, as shown in right image in Fig. 10a. In addition, there are also misclassifications that occurred in some ambiguous cases even though there was no obvious occlusion of the object. Such an example can be seen in the right image in Fig. 10b, where the valve with the handle mounted in the wrong orientation and a missing screw was classified as a valve with *missing screw*, although priority should be given to the class *wrong handle orientation* as indicated in the training data set.

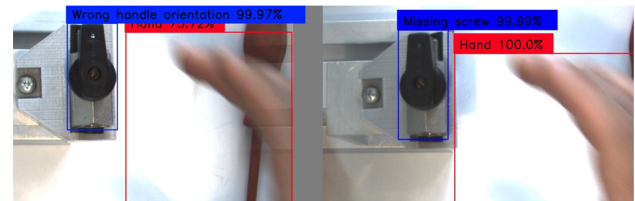
Sometimes, when the hand is holding a screwdriver, a lack of hand prediction occurs or the screwdriver is confused

**Table 2** Recall, precision, and F1-score for classes relevant to *assembly supervision* task and their weighted averages

|                          | Precision | Recall | F1 - score |
|--------------------------|-----------|--------|------------|
| Missing handle           | 0.919     | 1      | 0.958      |
| Missing screw            | 0.851     | 0.952  | 0.899      |
| Wrong handle orientation | 0.922     | 0.883  | 0.902      |
| Valve OK                 | 0.982     | 0.96   | 0.971      |
| Assembly precision       | 0.94      | 0.943  | 0.938      |



(a)



(b)

**Fig. 10** An example of prediction failure: **a** objects are not detected on the right image due to partial occlusion by hand; **b** class *missing screw* assigned on the right image even though priority should be given to the class *wrong handle orientation*

with a hand due to the similar color and size when the palm is upright.

### 3.3 Evaluation of final assembly inspection

Table 3 shows the confusion matrix for the signal *assembly OK*. The notation *positive* means that the valve was correctly assembled and the screw tightened, and *negative* that it was not. The calculated metrics for the signal *assembly OK* are shown in Table 4.

The results show that the signal *assembly OK* has a low critical error rate. False-positive signals occurred when the operator transferred the valve from one fixture to another, resulting in the inability to calculate the disparity  $D_t$  due to unconventional alignment of the assembled part in the pick-up holder or due to partial occlusions. Although some critical errors can be made during final inspection of the assembly, they never overlap with the *robot signal* because the operator's hands are still inside the assembly space. However, the *assembly OK* signal has a higher non-critical error rate due to numerous false-negative predictions. These predominate in the phase in which the operator has just placed the valve

**Table 3** Confusion matrix for the *assembly OK* signal

|                 |          | True class |          |
|-----------------|----------|------------|----------|
|                 |          | Positive   | Negative |
| Predicted class | Positive | 93         | 5        |
|                 | Negative | 10         | 692      |

**Table 4** False positive rate, false negative rate, precision, and recall for the signal *assembly OK*

|             | FPR<br>(Critical error rate) | FNR<br>(Non-critical error rate) | Precision | Recall |
|-------------|------------------------------|----------------------------------|-----------|--------|
| Assembly OK | 0.007                        | 0.097                            | 0.948     | 0.902  |

in the robot's pick-up holder, covering the valve again or moving the pick-up holder with their hands.

### 3.4 Evaluation of safety

Table 5 shows the confusion matrix for the signals *SCSV hands* and *Kinect hands* in-dependently in the first two rows. The third row shows the *workspace safe* signal, which is a logical NOR combination of the *SCSV hands* and *Kinect hands* signals, and the last row shows the *robot signal*. The notation positive for *SCSV hands* and *Kinect hands* signals indicate that the hands were present, and negative indicates that they were not. In case of the *workspace safe* signal, the notation positive marks that the workspace is safe, and the notation negative indicates that the workspace is unsafe, while *robot signal* additionally considers the signal from the *final assembly inspection*. The notation positive denotes that all conditions are met for the robot to pick up the valve, and the notation negative indicates that they are not.

The calculated metrics for safety-related signals are shown in Table 6. The critical error rates for each signal are highlighted in red. The *SCSV hands* signal has the lowest critical error rate. However, there were some cases where the operator's hand was partially inside the assembly space but was not detected because the operator had a screwdriver in his hand, as explained earlier in Sect. 3.2 or due to optical overlap of the stereo images. False positive predictions occurred frequently when an object other than a hand was detected, such as the screwdriver, which is often seen in the operator's hand and has a similar color to the palm. Compared to *SCSV* predictions, the *Kinect hands* signal has a slightly higher critical and non-critical error rate. Experiments have shown that the *Kinect* system had problems determining the

precise joint position for some operators sitting with their lower body covered.

The combined signal *workspace safe* means that there are no body parts in the *workspace* and no risk of collision between the robot and the operator. Table 5 shows the confusion matrix for the resulting *workspace safe* signal, where the notation positive indicates the workspace is safe and the notation negative shows that it is unsafe. This signal has a lower critical error rate than the standalone *Kinect* system but is still larger than the standalone *SCSV* system because the *SCSV* system only observes a smaller portion of the *workspace*.

Before the robot is allowed to move, another condition is considered. The assembly must properly performed. Table 5 shows the confusion matrix for the resulting *robot signal*, where the positive notation means that all conditions are met for the robot to pick up the valve, and the negative notation means that they are not. The additional condition of a correctly assembled valve lowers the critical error rate and prevents the robot from moving if the valve is not mounted correctly. When this condition is met, the operator also expects the robot to move and is more alert to the possibility of a collision.

## 4 Discussion

The experimental results can be summarized as follows:

- The *SCSV* system (i) provided the operator with visual aid for the assembly at 94% average precision, (ii) reliably inspected the final assembly with a low critical error rate of 0.7%, and (iii) ensured a safe *assembly space* by hand detection with a critical error rate of 1.1%.

**Table 5** Confusion matrix for safety signals

| Predicted class |          | True Class |          | Signal         |
|-----------------|----------|------------|----------|----------------|
|                 |          | Positive   | Negative |                |
| SCSV hands      | Positive | 5195       | 3        | SCSV hands     |
|                 | Negative | 58         | 2795     |                |
| Kinect hands    | Positive | 6687       | 58       | Kinect hands   |
|                 | Negative | 95         | 1211     |                |
| Workspace safe  | Positive | 1211       | 86       | Workspace safe |
|                 | Negative | 58         | 6696     |                |
| Robot signal    | Positive | 430        | 25       | Robot signal   |
|                 | Negative | 17         | 7579     |                |

**Table 6** False positive rate, false negative rate, precision, and recall for signals

| Signal         | FPR   | FNR   | Precision | Recall |
|----------------|-------|-------|-----------|--------|
| SCSV hands     | 0.001 | 0.011 | 0.999     | 0.989  |
| Kinect hands   | 0.045 | 0.014 | 0.991     | 0.986  |
| Workspace safe | 0.013 | 0.045 | 0.934     | 0.954  |
| Robot signal   | 0.003 | 0.038 | 0.945     | 0.962  |

- The wide-angle Kinect ensured a safe *workspace* by estimating human posture and hand location with a critical error rate of 1.4%. High critical errors rate is a consequence of a partial occlusion of the lower half of the human body and strong ambient light.
- The combined VQSM system improves hand detection in the *workspace* to a critical error rate of 1.3%.
- The addition of assembly inspection capability improves the safety of the VQSM system by allowing the robot to move only after a properly assembled object has been placed in its designated holder and no hands are in the work area. The critical error rate was improved to 0.3%. The errors were short lived and occurred during transitions between safe and unsafe conditions when only a small portion of the human body was present in the *workspace*. There was practically no likelihood of a collision between the operator and the robot.

Table 7 summarizes the critical error rates for signals related to safety. Since SCSV system that produces *SCSV hands* signal supervises a narrower assembly space, only *Kinect hands*, *workspace safe*, and *robot signal* safety signals are directly comparable.

We applied proportion Z-test to test significance of differences, and all *P*-values are below risk level  $\alpha = 0.05$ , that is, proportions are statistically different.

Although the SCSV system exhibits high precision and low critical error rates, its field of view limits hand detection to assembly space only and does not cover complete workspace. For that reason, the Kinect system is used to perform supervision of a human posture in a larger working area. However, experiments showed that the Kinect system sometimes has trouble estimating the precise joint position

**Table 7** Safety assessment in terms of critical error rates for independent signals and the combined VQSM system

| System         | Critical error rate |
|----------------|---------------------|
| Kinect hands   | FNR = 1.4%          |
| SCSV hands     | FNR = 1.1%          |
| Workspace safe | FPR = 1.3%          |
| Robot signal   | FPR = 0.3%          |

due to the operator's sitting position with the lower part of the body occluded from view. It is also widely accepted that Kinect system does not perform well in the presence of strong ambient light [38]. These issues could be reduced by using alternatives to the Kinect sensor, e.g., the Intel RealSense sensor, which uses either time-of-flight or IR stereo vision-based estimation of depth, or Microsoft Kinect Azure, which also uses time-of-flight technology. For human skeleton estimation, these sensors would be paired with deep learning models. Such changes would provide additional flexibility in optimizing human recognition, but at the cost of additional hardware and system complexity.

The *assembly supervision* and *final assembly inspection* in 1-second time sections do not represent real-time operation (10 frames/s) of the VQSM system, but rather operation with a 1-second delay before a decision is made. It should be noted that robot motion is still interrupted at any moment in high framerate if a single frame with a hand is detected by either system. The 1-second delay for initiating robot motion after part inspection and reaching a safe condition is acceptable. The evaluation based on 1-second video sections is also fully justified in terms of assembly supervision, since the operator is unlikely to make a wrong decision based on a single wrong prediction, but is likely to decide according to the prevailing predictions and ignore the noise. Improvements in assembly supervision and final inspection could be achieved by providing additional training data for the object detector. Manually collected and labeled data in a real environment, as demonstrated in this work, or multiple instances of correctly and incorrectly assembled objects could be artificially generated using CAD models of the product. This approach is worth exploring in the future as it could provide a large amount of training images and reduce the effort required for human labeling. Thus, the assembly supervision presented in this paper is a viable option for many variations of different products.

The experiments that included operator C show that the displayed predictions of the SCSV system can be a good help for an inexperienced worker. Without the predictions displayed during experiment C1, the operator assembled the valve incorrectly and did not figure out what the error was (*incorrect handle orientation*). When the operator was shown the predictions during experiment C2, he was able to use them to identify the errors during assembly and correctly assemble both valves (the first assembly attempt took 81 s and the second took 46 s). When the operator was given additional instructions and already had some experience, the assembly times in trial C3 improved significantly (first trial 29 s and second trial 24 s). Since valve assembly is a relatively simple task, we cannot measure any improvement in efficiency for the experienced operators, but such a system could be invaluable in more complex operations or in training new operators.



## 5 Conclusions

In this study, we present a VQSM system for human-robot cooperation. The system ensures the safety and enables quality control during the execution of work operations and the final product. The installation of the system and the impact of the sensors on the operator and the workspace are minimized by using a compact SCSV system for close observation of the manual task and a wide-angle Kinect for posture monitoring.

The operation of the system was demonstrated and evaluated using an industrial case study for manual assembly. The YOLOv3 object detector successfully predicted all classes of interest and proved to be a suitable choice for quality and safety-related tasks due to its high accuracy and low critical error rates. This confirms the premise that supervised learning is a highly efficient method when used for assembly supervision, as we want to detect quality features that have been defined in advance by experienced personnel.

Experimental results show that in terms of safety, neither the SCSV nor the Kinect performs ideally as a standalone system, but our approach demonstrated improved performance when the two work in parallel. The safety signals from both systems are logically fused, and the resulting signal improves the critical error rate of hand detection throughout the workspace. The risk of the robot colliding with the operator in the assembly space is practically non-existent, especially when the condition of a correctly assembled part is considered. Since the operator does not have to make any special gestures when assembling the object but only places his hands in a safe place, the ease of use and intuitiveness is considered high.

Future research will take advantage of existing CAD models of the objects being assembled to create images for training the object detector, hence reducing the need for manual labeling, since synthetically rendered images can increase the performance of deep learning models [39]. Such an approach could also prove useful in other areas, e.g., in production planning when taking into account assembly times or when analyzing videos for possible bottlenecks in various processes.

**Author Contributions** The authors contributed equally to the work. All authors have read and agreed to the published version of the manuscript.

**Funding** This work was supported by the Ministry of Higher Education, Science and Technology of the Republic of Slovenia, research programs P2-0270 (Manufacturing Technologies and Systems) and P2-0246 ICT4QoL (Information and Communications Technologies for Quality of Life).

## Declarations

**Conflict of interest** The authors declare no competing interests.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

1. Yang C, Zhu Y, Chen Y (2022) A review of human-machine cooperation in the robotics domain. *IEEE Transactions on Human-Machine Systems* 52(1):12–25. <https://doi.org/10.1109/THMS.2021.3131684>
2. Ajoudani A et al (2018) Progress and prospects of the human-robot collaboration. *Auton Robot* 42. <https://doi.org/10.1007/s10514-017-9677-2>
3. Arents J et al (2021) Human-robot collaboration trends and safety aspects: a systematic review. *J Sensor Actuator Netw* 10. <https://doi.org/10.3390/jsan10030048>
4. Kim W et al (2019) Adaptable workstations for human-robot collaboration: a reconfigurable framework for improving worker ergonomics and productivity. *IEEE Robot Autom Mag* 26. <https://doi.org/10.1109/MRA.2018.2890460>
5. Bonci A, Cheng PDC, Indri M, Nabissi G, Sibona F (2021) Human-robot perception in industrial environments: a survey 21(5):1–29. <https://doi.org/10.3390/s21051571>
6. Robla-Gomez S et al (2017) Working together: a review on safe human-robot collaboration in industrial environments. *IEEE Access* 5:26754–26773. <https://doi.org/10.1109/ACCESS.2017.2773127>
7. Bai Y et al (2019) An investigation of security approaches in industrial robots. 2019 5th International Conference on Control, Automation and Robotics, ICCAR 2019 103–110. <https://doi.org/10.1109/ICCAR.2019.8813393>
8. Halme R-J et al (2018) Review of vision-based safety systems for humanrobot collaboration. *Procedia CIRP* 72:111–116. <https://www.sciencedirect.com/science/article/pii/S2212827118301434>. <https://doi.org/10.1016/j.procir.2018.03.043>, 51st CIRP Conference on Manufacturing Systems
9. Feng X, Jiang Y, Yang X, Du M, Li X (2019) Computer vision algorithms and hardware implementations: a survey. *Integration* 69:309–320. <https://doi.org/10.1016/j.vlsi.2019.07.005>
10. Zamora-Hernández MA, Castro-Vargas JA, Azorin-Lopez J, Garcia-Rodriguez J (2021) Deep learning-based visual control assistant for assembly in industry 4.0. *Comput Ind* 131:103485. <https://doi.org/10.1016/j.COMPIND.2021.103485>
11. Wang P, Liu H, Wang L, Gao RX (2018) Deep learning-based human motion recognition for predictive context-aware human-robot collaboration. *CIRP Ann* 67:17–20. <https://doi.org/10.1016/j.cirp.2018.04.066>
12. Zhang J, Liu FL, Wang RW (2020) Research on industrial parts recognition algorithm based on yolo v3 in intelligent assembly. *Guangdianzi Jiguang/Journal of Optoelectronics Laser* 31:1054–1061. <https://doi.org/10.16136/j.joel.2020.10.0264>

13. Park J, Jang W, Yoo N (2021) An empirical study on process management system using yolo-based parts recognition. *J Phys Conf Ser* 1888. <https://doi.org/10.1088/1742-6596/1888/1/012024>
14. Rajnathsing H, Li C (2018) A neural network based monitoring system for safety in shared work-space human-robot collaboration. *Ind Robot* 45:481–491. <https://doi.org/10.1108/IR-04-2018-0079>
15. Wang KJ, Santoso D (2022) A smart operator advice model by deep learning for motion recognition in human-robot coexisting assembly line. *Int J Adv Manuf Technol* 119. <https://doi.org/10.1007/s00170-021-08319-1>
16. Papanastasiou S et al (2019) Towards seamless human robot collaboration: integrating multimodal interaction. *Int J Adv Manuf Technol* 105. <https://doi.org/10.1007/s00170-019-03790-3>
17. Xiong Q, Zhang J, Wang P, Liu D, Gao RX (2020) Transferable two-stream convolutional neural network for human action recognition. *J Manuf Syst* 56:605–614. <https://doi.org/10.1016/j.jmsy.2020.04.8007>
18. Liu H, Wang L (2021) Collision-free human-robot collaboration based on context awareness. *Robot Comput Integr Manuf* 67:101997. <https://doi.org/10.1016/j.rcim.2020.101997>
19. Zhang J, Liu H, Chang Q, Wang L, Gao RX (2020) Recurrent neural network for motion trajectory prediction in human-robot collaborative assembly. *CIRP Ann* 69:9–12. <https://doi.org/10.1016/j.cirp.2020.04.077>
20. Flacco F, Kröger T, Luca AD, Khatib O (2012) A depth space approach to human-robot collision avoidance. *Proceedings - IEEE International Conference on Robotics and Automation* 338–345. <https://doi.org/10.1109/ICRA.2012.6225245>
21. Liu H, Fang T, Zhou T, Wang L (2018) Towards robust human-robot collaborative manufacturing: multimodal fusion. *IEEE Access* 6:74762–74771. <https://doi.org/10.1109/ACCESS.2018.2884793>
22. Olesen AS, Gergaly BB, Ryberg EA, Thomsen MR, Chrysostomou D (2020) A collaborative robot cell for random bin-picking based on deep learning policies and a multi-gripper switching strategy. *Procedia Manuf* 51:3–10. <https://doi.org/10.1016/j.promfg.2020.10.002>
23. Tan JTC, Arai T (2011) Triple stereo vision system for safety monitoring of human-robot collaboration in cellular manufacturing. *Proceedings - 2011 IEEE International Symposium on Assembly and Manufacturing. ISAM 2011:1–6*. <https://doi.org/10.1109/ISAM.2011.5942335>
24. Melchiorre M, Scimmi LS, Mauro S, Pastorelli SP (2021) Vision based control architecture for human-robot hand-over applications. *Asian J Control* 23. <https://doi.org/10.1002/asjc.2480>
25. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2016-December*. <https://doi.org/10.1109/CVPR.2016.91>
26. Duvioubourg L, Ambellouis S, Lefebvre S, Cabestaing F (2007) Obstacle detection using a single camera stereo sensor 979–986. <https://doi.org/10.1109/SITIS.2007.26>
27. Bouguet J-Y (2013) Camera calibration toolbox for matlab (computer software). <http://www.vision.caltech.edu/bouguetj/calibdoc/index.html>
28. Bračun D, Sluga A (2015) Stereo vision based measuring system for online welding path inspection. *J Mater Process Technol* 223:328–336. <https://doi.org/10.1016/j.jmatprotec.2015.04.023>
29. Ren S, He K, Girshick R, Sun J (2017) Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39. <https://doi.org/10.1109/TPAMI.2016.2577031>
30. He K, Gkioxari G, Dollár P, Girshick R (2020) Mask r-cnn. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42. <https://doi.org/10.1109/TPAMI.2018.2844175>
31. Liu W et al (2016) Ssd: single shot multibox detector. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 9905 LNCS. <https://doi.org/10.1007/978-3-319-46448-02>
32. Antonelli S et al (2022) Few-shot object detection: a survey. *ACM Comput Surv*. <https://doi.org/10.1145/3519022>
33. Maiettini E, Tikhonoff V, Natale L (2021) Weakly-supervised object detection learning through human-robot interaction. *IEEE-RAS International Conference on Humanoid Robots 2021-July*. <https://doi.org/10.1109/HUMANOIDS47582.2021.9555781>
34. Redmon J, Farhadi A (2018) Yolov3: an incremental improvement. *arXiv:1804.02767*
35. OpenCV object detection (2022). <http://opencv.org>
36. Springer S, Seligmann GY (2016) Validity of the kinect for gait assessment: a focused review. *Sensors (Switzerland)* 16. <https://doi.org/10.3390/s16020194>
37. Han J, Shao L, Xu D, Shotton J (2013) Enhanced computer vision with microsoft kinect sensor: a review. *IEEE Transactions on Cybernetics* 43. <https://doi.org/10.1109/TCYB.2013.2265378>
38. Obdrzalek S et al (2012) Accuracy and robustness of kinect pose estimation in the context of coaching of elderly population. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. <https://doi.org/10.1109/EMBC.2012.6346149>
39. Sampaio I, Machaca L, Viterbo J, Guérin J (2021) A novel method for object detection using deep learning and cad models. <https://doi.org/10.5220/0010451100750082>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.