Journal of Intelligent & Robotic Systems (2022) 106:35
https://doi.org/10.1007/s10846-022-01607-2

SHORT PAPER

Dynamic Speed and Separation Monitoring Based on Scene Semantic Information

Botao Yang[1] · ShuXin Xie[1] · Guodong Chen[1] · Zihao Ding[1] · Zhenhua Wang[1]

Received: 15 September 2021 / Accepted: 24 February 2022 / Published online: 21 September 2022

Abstract
Human-robot collaboration (HRC) based on speed and separation monitoring should consider the difference of risk factors in the scene; otherwise, the sudden invasion of non-operators or routine operation of the operator may stop the robot system. In this paper, we propose a sensing network based on the fusion of multi-information to obtain scene semantic information and employ it to realize risk assessment. However, due to the influence of light on the image information sensed by RGB cameras, it is not easy to obtain accurate scene semantic information. We apply a depth camera and a thermal imager to obtain depth and infrared information to enhance the RGB images. We build a risk information database and use it to quantify the obtained scene semantic information into risk factors. The dynamic change of risk factors judges whether the distance between humans and robots is safe. The experimental results verify that the algorithm of intelligent human-robot monitoring can realize the analysis of dangerous situations and control the robot system, thereby reducing the number of false shutdowns and improving safety.

Keywords Human-robot collaboration (HRC) · Speed and separation monitoring (SSM) · Neural network · Semantics · Multi-information fusion

1 Introduction

With the growth of the demand for customized products, human-robot collaboration (HRC) has become an essential mode of production and has been widely studied [1]. Meanwhile, due to the close interaction between humans and robots in HRCs, safety has become a problem that must be taken into account in the process. Because traditional security measures lack flexibility and adaptability, they cannot solve the problem in the clustered workspace. Vision-based security measures have the advantages of strong adaptability and high intelligence, and have become a hotspot in the field of robot security in recent years [2].After international cooperative robot-related specifications are released, much research is concentrated in the field of collaborative robot safety protection. The standard of human-robot collaboration is defined in ISO/TS 15066 [3]. Collaborative operations may include one or more of the following methods: safety-rated monitored stop, hand guiding, speed and separation monitoring, and power and force limiting. Among them, the speed and separation monitoring (SSM) method is flexible and less restrictive, and commits to the requirements of human-robot interaction, so it has been widely considered.

Botao Yang and ShuXin Xie contributed equally to this work.

✉ Guodong Chen
guodongxyz@163.com

Botao Yang
20195229047@stu.suda.edu.cn

ShuXin Xie
xsx_suda@126.com

Zihao Ding
zhding@stu.suda.edu.cn

Zhenhua Wang
wangzhenhua@suda.edu.cn

[1] Jiangsu Provincial Key Laboratory of Advanced Robotics, School of Mechanical and Electrical Engineering, Soochow University, Suzhou 215123, China

As early as 2013, Marvel [4] proposed metrics that evaluate speed and separation monitoring efficacy in industrial robot environments in terms of the quantification of safety and the effects on productivity. SSM is that the maximum safe speed is relevant to the distance between the human and robot. At any time under the current speed, the robot system should stop the robot before it collides with a human. Then Andrea Maria [5] expanded on this basis and introduced the hierarchical adjustment of speed. To realize the continuous adjustment of robot speed, Shin [6] proposed a method to keep the robot running at a safe and maximum speed, which can guarantee safety and improve efficiency to the greatest extent. Byner [7] proposed the method of dynamic calculation of the safe distance between humans and robots. In addition further optimized the algorithm of SSM. In addition to adjusting the speed, some people [8, 9] have proposed avoiding danger by changing the trajectory. For example, Chen [10] used a series of virtual constraint balls as the expression of safe distance and used the obstacle avoidance algorithm based on the artificial potential field method to make the robot avoid operation and prevent dangerous collisions between humans and robots. Some approaches for collaborative robot path planning are designed to achieve adaptive obstacle avoidance in dynamic manufacturing [11]. However, this kind of method has superior spatial uncertainty and is rarely applied in practice.

With the development of sensor technology, people realize that perception is a significant factor that limits the algorithm's performance. Therefore, researchers attempt to integrate more information into robot safety monitoring. Marvel [12] proposed a task-based off-line security risk assessment platform for cooperative robot systems. The risk assessment factors include end tools, personnel identity, task type, and duration. Lucci [13] put forward a new idea that combines force information and speed with distance monitoring and proposed a human-robot safety monitoring mode that combines speed with distance monitoring and force control. Kim [14] introduced the radar sensor into the safety monitoring of the robot. Kumar [15] used multiple TOF sensors to improve the performance of speed and separation monitoring. Mazhar [16] uses the knectv2 camera to obtain three-dimensional information and uses gesture recognition to improve the safety and interactivity of the compliant robot. Aliev [17] evaluated the safety risk factors for robot systems offline and used the machine learning method to judge the possible danger.

In the above method, a variety of sensors are in place to obtain the human-robot distance as accurately as possible. However, in the process of collaboration, not only the distance between human and robot but also the semantic information in the scene such as task type, dangerous type, and end tool will affect the safety. The danger degree of robots in high-speed cutting tasks is obviously higher than that in a low-speed grasping task. At the same time, robot operators and non-operators, due to the different cognitions of robot working paths and operation modes, they should have different degrees of danger. If it is not distinguished, it is easy to cause accidental shutdown due to excessive deceleration during normal operation by operators or collisions when non-operators invade. The above method does not analyze the semantic information of the dangerous target in detail or can only carry out offline task-based hazard analysis, which cannot be adjusted in real-time. Therefore, we proposed a dynamic speed and separation monitoring algorithm based on semantic information of the scene, which could obtain information such as the number and types of hazardous targets in the scene, and dynamically adjust the safe distance between humans and robots. The overall flow of the method is shown in Fig. 1.

Neural networks have long been used in industry for fault prognosis [18] . In recent years, it has been used for detection. Y Sun [19] used multimodal information and neural networks to detect dangerous targets in automatic driving scenes, which reduced the impact of light on detection. Some people use lidar to obtain point cloud information [20], segment the point cloud information and obtain spatial semantic information. However, due to the large scale of point cloud data in large scenes, the extraction efficiency is relatively low, and the working environment of the robot is often complex, other objects will affect the detection of risk targets in the area close to the space position of the robot, there is a lot of noise in the point cloud information obtained by lidar. It is difficult to eliminate the interference and obtain accurate semantic information only using the depth information obtained by lidar or depth camera. Therefore, we also add temperature information. The heating of robot body and human body can be highlighted in the temperature image to enhance the risk target. We use various sensors to obtain the RGB, depth, and temperature images of the whole scene and propose a neural network for data fusion and semantic information extraction. We transformed the obtained risk information into dynamic risk factors through the risk information database, regulating the safe distance dynamically. The algorithm proposed in this paper can make the robot dynamically perceive the danger degree of the scene according to the semantic information of the scene, and adopt appropriate collision avoidance strategies for different dangerous scenes, to reduce the probability of the robot's false stop and collision.

Our primary contributions are:

- proposing a dynamic speed and distance monitoring algorithm based on scene semantic information;
- integrating the multimodal information into the neural network for detection and segmentation, and designing the semantic perception network of multi-information fusion (MSNet);
- establishing the database of risk information to realize the dynamic update of risk factors.

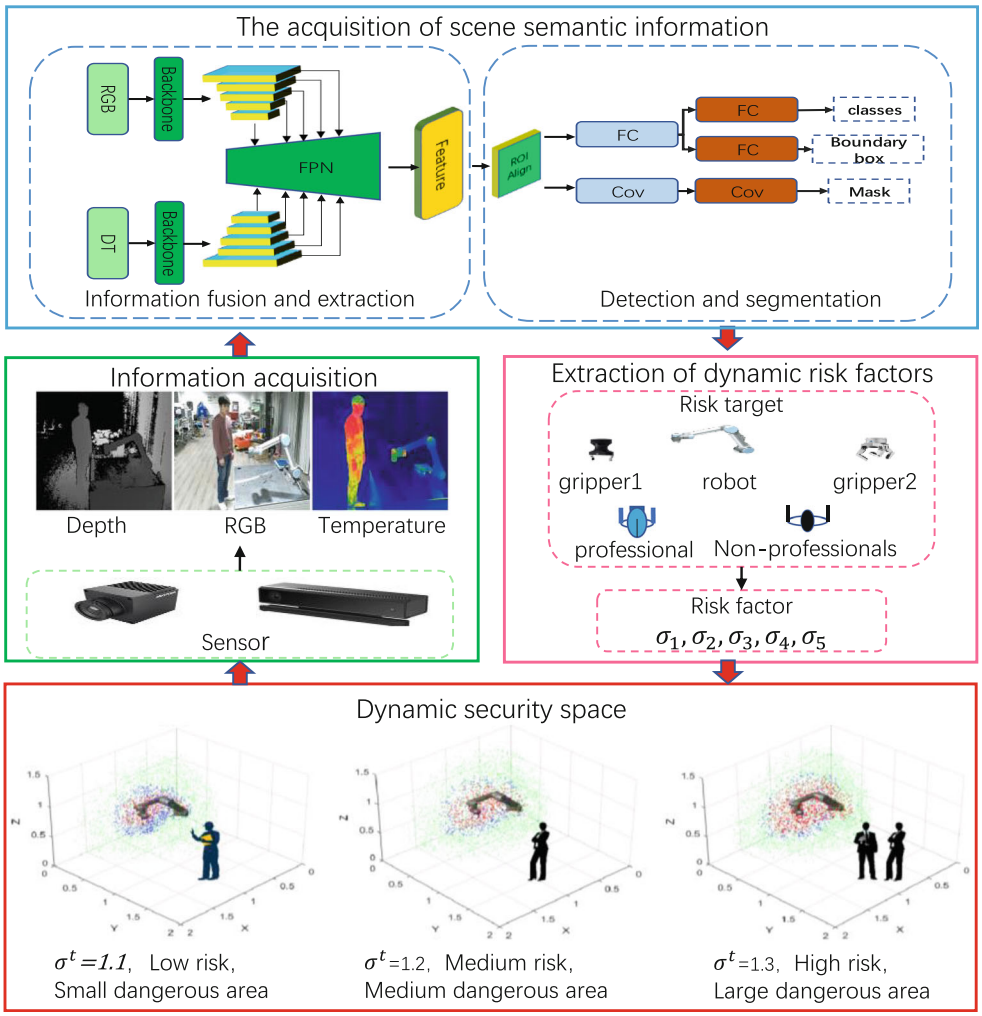

Fig. 1 Dynamic speed and separation monitoring based on scene semantic information

2 Scene Semantic Extraction Based on Multi-Information

2.1 Data Analysis

According to the regulation of risk assessment in ISO/TS 15066, in the workspace of a robot, many factors are related to safety, including the type of robot body, end tools, personnel type, position, and movement speed in the scene. According to the above factors, we selected five targets as risk samples. They are robot body, gripper1, gripper2, operator and non-operator. We obtained the RGB image of the dangerous target in the workspace and found that the RGB image quality is poor in the backlight environment, which is difficult to detect normally. So we try to use a variety of sensors to get images. After testing, it is found that the depth image and infrared thermal imaging are less affected by illumination and can obtain much better image information than ordinary RGB images.

As shown in Fig. 2, we get three kinds of images after registration and gray processing, and draw the gray distribution histogram. The information types of the temperature image and depth image are similar, so we directly fuse the two images on the channel, and after fusion, we call them DT fusion images. The fusion method of the DT fusion image is as follows: first, the depth image and temperature image are transformed into a gray image, and then (1) is used to add pixels to obtain the value of each pixel of the fused image.

$$A_i = \gamma D_i + \delta T_i. \tag{1}$$

where A_i is the pixel value of the fused image; D_i is the pixel value of the depth image; T_i is the pixel value of the temperature image; $\gamma$and δ are the weight factors of the depth image and the temperature image respectively.

RGB images mainly contain category information, while depth and temperature images mainly contain edge information. The types of information are different. Therefore, we use two CNN networks to extract features from RGB images and DT fusion images and find that the gray distribution is similar in some intermediate stages of extraction. The contour information of humans and robots in the middle feature layer of the

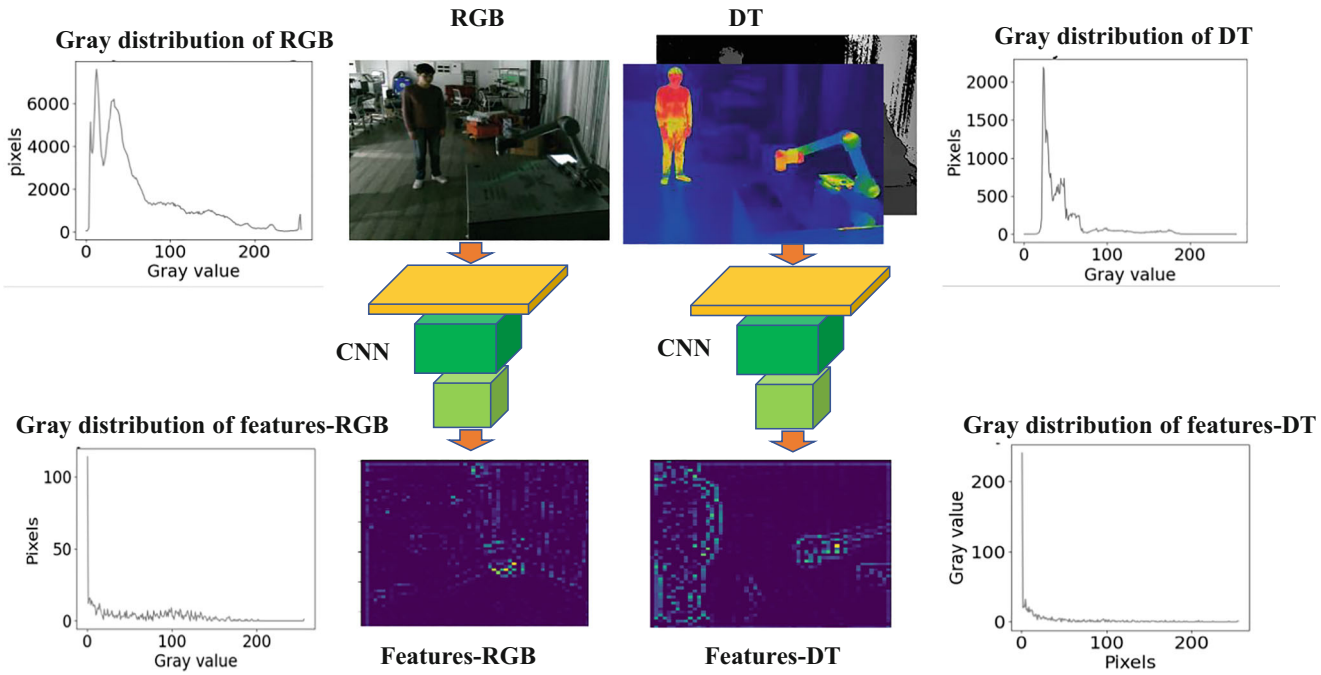

Fig. 2 Gray-level distribution of original image and characteristic image

DT image is obvious, and the two features show good fusibility, as shown in Fig. 2.

2.2 MSNet

Object detection is one of the basic tasks in the field of computer vision. In recent years, the rise of deep learning [21] has significantly improved object detection performance and brought significant progress to object detection [22]. In human-robot interaction, neural networks have also been widely used to improve the intelligence degree of robots [23–25]. Then the convolution [26] is used to extract the features from the image and add the features to achieve multimodal information fusion. Therefore, to improve the robustness and accuracy of the network under different light conditions, this paper proposes a deep neural network that integrates RGB, depth, and temperature image information to obtain semantic image information.

Mask R-CNN [27] is a flexible and universal object instance segmentation framework. It can detect the objects in the images and output a high-quality segmentation result for each object. In order to distinguish dangerous targets in dark and high light ratio environments and reduce the influence of light and background on recognition, we designed MSNET on the basis of Mask R-CNN. As shown in Fig. 3, there is a feature pyramid structure for the fusion of multi-information in MSNET. The feature pyramid structure has realized feature fusion on multi-scale. Using the characteristics of depth images and infrared thermal imaging which are insensitive to light and have obvious edge features, the RGB image was enhanced, and the robustness of the network in different lighting environments was improved.

As shown in Fig. 3, the MSNET proposed in this paper uses the traditional two-stage network structure as the backbone [28]. First, it carries out feature extraction and region generation, which are used to extract the feature layer and generate the candidate target bounding box. Then, it uses the candidate target bounding box and ROI Align layer to extract features from each feature layer. The network has three branches corresponding to the category, candidate box, and mask. The difference is that in the initial stage of feature image generation, two backbones extract the features of the RGB image and DT fusion image respectively. The two features extracted are fused by the feature pyramid structure proposed in this paper, which reduces the interference of external light on the image and improves the detection ability of the network for different targets.

As shown in Fig. 4, the feature pyramid structure is a multiscale feature fusion structure. The backbone network of the RGB branch is used to extract features rich in category information. The backbone network of the DT branch is used to extract edge and position information. The two kinds of information are superimposed on four different scale levels through the feature pyramid structure to enhance the effective feature information. After superimposition, the fusion feature layer of each scale is output to enhance the feature extraction ability of the feature pyramid. Because the sensitivity of different feature layers to different scale targets is different, the output of intermediate feature layers with different scales after fusion will have a better effect on different scale targets. The

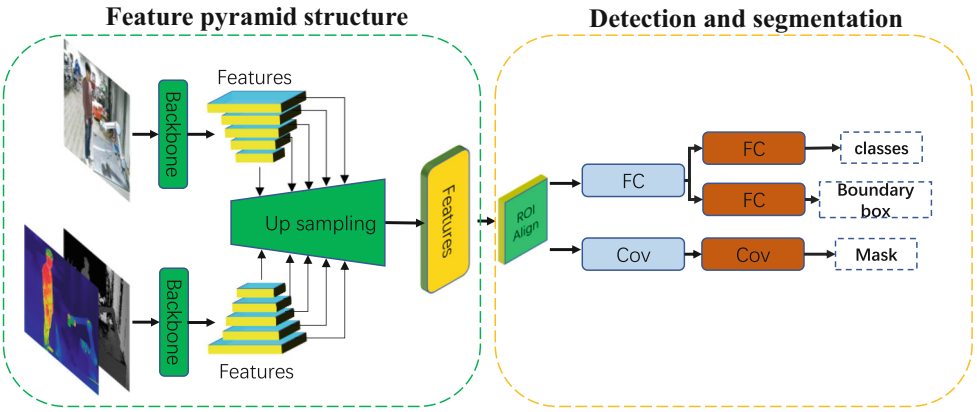

Fig. 3 MSNET

specific structure is shown in Fig. 4. The two CNN networks extract features from the input 512*512 three-channel images. The feature extraction network is divided into five stages. Each stage consists of two 3 × 3 convolutions with one stride and a 2 × 2 maximum pooling layer with two strides. Each stage reduces the size of the feature map by one time, starting from stage 2. The output feature maps of each stage are not the only input to the next stage but also input to the feature fusion pyramid after convolution. The feature maps in the pyramid are up-sampled by linear interpolation and added with feature maps of different sizes from the feature extraction network. Each layer of the feature map in the feature pyramid will be output for later detection and segmentation. This structure can fuse the features of three kinds of images at different scales to improve the performance of the network under different illumination conditions.

3 Speed and Separation Monitoring Based on MSNET

3.1 Standard Speed and Separation Monitoring Algorithm

According to the definition of SSM given in ISO/TS 15066, when the dangerous target moves beyond the safe distance, the robot will stop to ensure human safety. The calculation method of the safety distance is shown in (2).

$$S_p(t_0) = S_h + S_r + S_s + C + Z_d + Z_r \tag{2}$$

Where S_p is the total safe distance. S_h is the moving distance of a human, $S_r$is the reaction distance of the robot;and Ss is the stopping distance of the robot. Their definitions are given by Eqs. (3), (4), and (5). C is the pre-intrusion distance, which

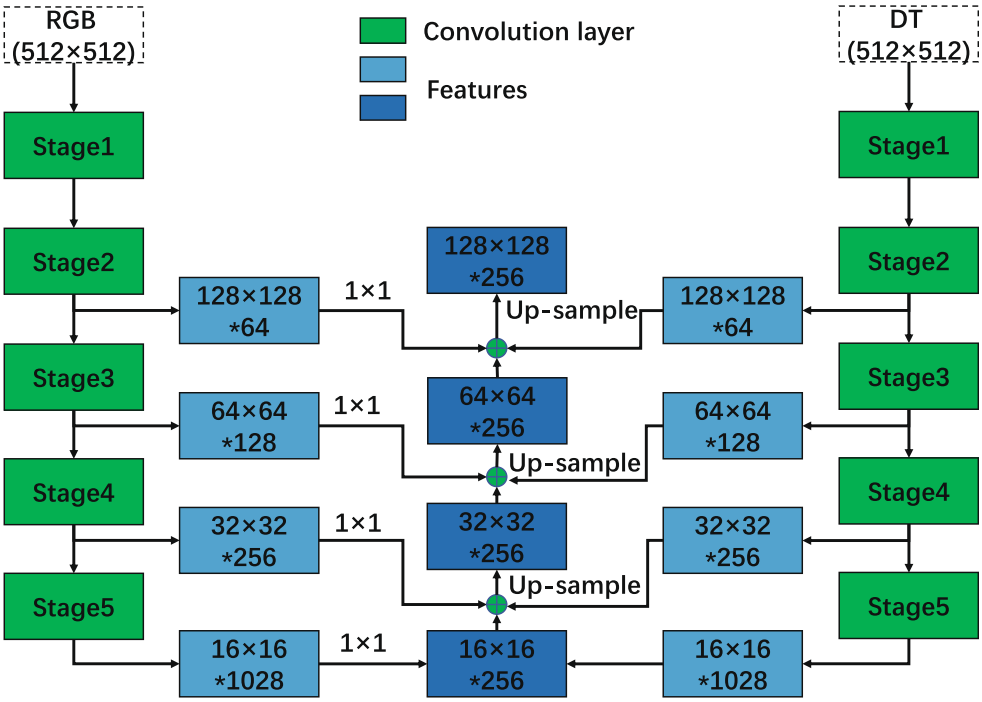

Fig. 4 Feature pyramid structure for the fusion of multi-information

indicates the distance that the dangerous target can intrude into the monitoring range before being detected. $Z_d + Z_r$ represents the position uncertainty of the robot and human. Because we are facing a huge scene, it can be ignored here.

$$S_h = v_h(T_r + T_s) \tag{3}$$

$$S_r = v_r T_r \tag{4}$$

$$S_s = \frac{v_r^2}{2a_s} \tag{5}$$

v_h is the speed of human movement, v_r is the speed of the robot, T_r is the reaction time of the robot, and a_s is the braking acceleration of the robot. T_s is the braking time of the robot.

3.2 Analysis of Human-Robot Dangerous Behavior

We collect the position of the key edge points of the human body when the operators and non-operators enter during the movement of the robot. The key edge points of the human body are obtained according to the mask output by the neural network. There are six points in total, which are the highest and lowest points of the mask, the leftmost and rightmost points, and the two intersections of the vertical line between the highest point and the lowest point and the mask. Because collisions often occur at the convex edge of the human body, this method can better get the most dangerous point of the human body, as shown in Fig. 5. In the process of robot movement, we counted the data of operators and non-operators entering and leaving the monitoring range 50 times respectively, and the sampling interval was 1 s. We find that the position of the operator is more concentrated and less intruded into the trajectory of the robot, However, the movement of non-operators is relatively irregular, and often intrudes into the trajectory of the robot because they are not familiar with it According to this characteristic, we classify the danger of different kinds of targets and propose an intelligent speed and distance monitoring algorithm. The dangerous targets detected from the neural network are introduced into the monitoring algorithm in the form of danger coefficients, to realize the dynamic adjustment of the human-robot safe distance and the dynamic adjustment of the tool radius and dangerous target radius.

3.3 M-SSM

The M-SSM proposed in this paper is based on the standard SSM given in ISO/TS 15066. By using the semantic information and the robot's motion parameters, we can calculate the robot's dynamic safe area according to different danger degrees, and then adjust the robot's motion speed according to the position information of the safe area and the dangerous target. It is always guaranteed that at the current moment, the robot can stop before colliding with the dangerous target through emergency braking. The following is the pseudo-code of the algorithm, if the dangerous target keeps approaching, the speed of the robot decreases continuously, and finally decreases to zero. When the dangerous target is far away, the speed will gradually recover.

Algorithm 1: M-SSM

Input: *RGB, Depth, Temperature*

Output: V_{safe}

While robot start do
  $Features \leftarrow Con(RGB, Depth, Temperature)$
  $classes, regions \leftarrow Detect(Features)$
  $\sigma_i^{t_0} \leftarrow Evaluate(classes, regions)$
  $\sigma^t = \sum_{i=0}^{n}((\sigma_i^{t_0} * m) + 1)\sqrt{a^2 + b^2}$
  $D_{safe} = Safe\_distance(\sigma_i^{t_0})$
  If $D_{safe} < D_{ture}$
  Robot deceleration
    V_{safe} *is calculated from* D_{ture}
    Return V_{safe}
  Else
    Robot running normolly

According to the two-dimensional dynamic speed and separation monitoring algorithm proposed by Christoph Byner [7] based on the standard SSM, we propose a dynamic three-dimensional model to calculate the minimum safe distance. From time t_0, the robot starts to brake until the speed of the robot drops to zero, and the end of the robot just contacts the dangerous target. Then, the geometric relationship of the robot motion from time t_0 is shown in Fig. 6.

It can be seen from Fig. 6 that α is the angle between the robot's motion direction and the initial position line of the human-robot at time t_0. $R_i(t_0)$ and $H_j(t_0)$; are the positions of the robot and the human at time t_0 respectively. β is the angle between S_s and $R_i(t_0 + t_r)H_j(t_0)$; and θ is the angle between the human path and $R_i(t_0 + t_r)H_j(t_0)$; r_i is the geometric radius of the robot end; D is the effective safe distance, and σ^{t_0} is the risk factor at time t_0. In order to simplify the motion model and establish the above geometric relationship, according to reference [7], we need to make the following assumptions: (1) in the system reaction time T_r, the speed of the robot is constant, and the running direction of the robot is determined by the position of the nearest sampling point. Because we will decelerate the robot after the dangerous target is detected, and the application scene of this algorithm is large, this assumption has little impact on the accuracy. (2) The stopping distance S_s

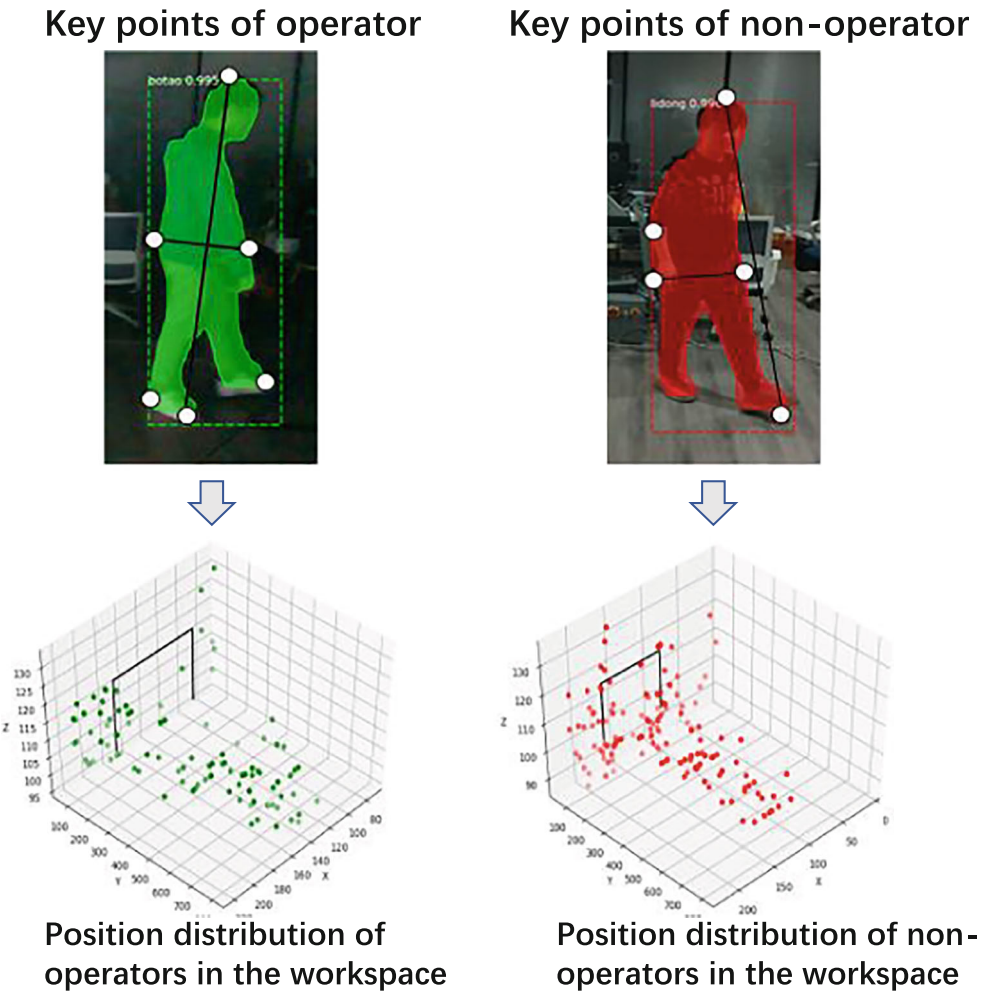

Fig. 5 Distribution of key points on the edge of human body

of the robot is directly towards the direction of the human. This ensures that in the worst case, the robot can also reduce its speed to zero before colliding with a dangerous target.

From Fig. 6, we can obtain the relationship between the running speed v_r of the robot and the safe distance S_p, as shown in (6):

$$\begin{aligned}&((\mathrm{S_h}+\mathrm{S_s})\cos\beta+(C+r_i)\cos\theta)^2\\&=(v_rT_r)^2+\left(S_p\sigma^{t_0}\right)^2-2S_p\sigma^{t_0}v_rT_r\cos\alpha\end{aligned}\tag{6}$$

According to (6) and Fig. 6 the relationship between the maximum speed of the robot and the safe distance is obtained as Eq. (7).

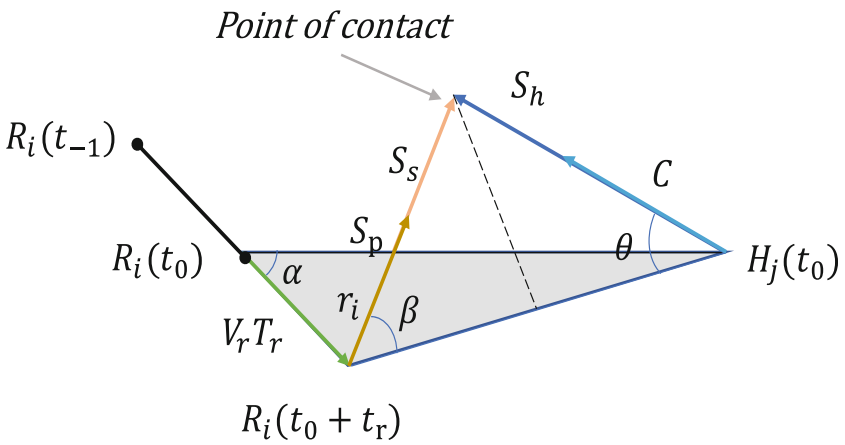

Fig. 6 Kinematic geometry diagram

$$0=\frac{\cos^2\beta}{4a_s^2}v_r^4+\frac{\cos\beta\cos\theta}{a_s^2}v_r^3+\left(\frac{\cos^2\theta v_h^2}{a_s^2}+\frac{\cos^2\beta r_i+\cos\beta\cos\theta v_h T_r+\cos\beta\cos\theta C}{a_s}-T_r^2\right)v_r^2+\left(2DT_r\cos\alpha+\frac{2\cos\beta\cos\theta v_h r_i}{a_s}\right)v_r$$
$$+(\cos\beta r_i+\cos\theta v_h T_r)^2+\cos^2\theta C^2+2\cos\beta\cos\theta C r_i-D^2 \tag{7}$$

Where $D=S_p/\sigma^{t_0}$.

According to ISO/TS15066, the risk factors are divided into three categories: robot related, system related, and application related. The five dangerous targets we selected covered these three types of risk factors, which can reflect the risk of the scene. According to the risk matrix [29] used by the National Institute of Standards and Technology's engineering laboratory, the risk assessment of these objectives is summarized in Table 1. The risk probability from low to high is: impossible, slight, accidental, possible and frequent. The severity of the danger from low to high is: slight, medium, serious, and disaster. The safety description method of collaborative robot proposed by Marvel [12] is used for quantitative evaluation, and the final risk factor is obtained.

After getting the dangerous target information at time t_0, we need to make a real-time evaluation of the danger degree of the scene at timet_0. the evaluation algorithm is shown in (8), (9).

$$\sigma^{\mathrm{t}}=\sum_{i=0}^{n}\left(\left(\sigma_i^{t_0}*m\right)+1\right) \tag{8}$$

$$\sigma_{\mathrm{i}}^{t_0}=\{\sigma_1,\sigma_2,\sigma_3,\sigma_4,\sigma_5\} \tag{9}$$

According to Eqs. 6, 7 and 9, the danger range diagram is drawn, and the system parameters $C=0.32m$ $v_h=1.6m/s$ $T_r=210ms$and$a_s=10m/s^2$.The maximum running distance is shown in Fig. 7, where the vertical axis represents the running speed of the robot, and the horizontal axis represents the corresponding safe distance. The five colors in Fig. 7 represent five scenes with different degrees of danger, and yellow, green, blue, purple, and red represent scenes with risk factors for 1.0, 1.1, 1.2, 1.3, and 1.4, respectively. It can be seen from Fig. 7 that in five different states, even at the same speed, the size of the dangerous area is different. The greater the risk factor for the target in the monitoring area, the greater the safety distance.

Table 1 Dangerous target and its dangerous coefficient

Target name	Possibility	Seriousness	$\sigma_i^{t_0}$
robot	slight	Serious	0.1
three-finger gripper	probably	slight	0.1
two-finger gripper	By chance	slight	0.05
operator	By chance	Serious	0.05
Non-operator	Frequent	Serious	0.15

4 Experiments

The experimental platform includes a 3.3GHz CPU, rtx2080ti GPU, two UR5 robots, a Realsense l515 depth camera, and a Haikang DS-2TA03-10AUF thermal imager. Two robots are equipped with two different kinds of grippers, gripper2 is a three finger flexible gripper, and the other is an ordinary two-finger gripper. They correspond to different dangerous end tools. The experimental scene is shown in Fig. 8.

4.1 Experiment on Perception Effect of Dangerous Target

In order to verify the effect of network detection, the following experiments are carried out. In the robot workspace, there are five kinds of targets that affect safety. They are: robot, two-finger mechanical gripper(gripper1), three-finger mechanical gripper(gripper2), operators and non-operators. A total of 3000 groups of photos taken in different lighting environments are used for training. The shooting environment of the training set includes backlight, dark light, and normal light. The RGB map and depth map required for training are collected by KinectV2, and the thermal imaging map is collected by Haikang DS-2TA03-10AUF. The three kinds of images have been registered, and the depth map and thermal imaging map have been weighted and fused. The fusion method is shown in Formula 1. The CPU used in the training is i9-9940x, the clock speed is 3.3GHz, and the GPU is RTX2080Ti.

First, we use RGB and depth images, RGB and temperature images, RGB, and DT fusion images as the input of the two branches of the network for three times, and only use RGB images for training. The original Mask-RCNN network is trained. For the four cases of networks in the training set of training loss, the loss value represents the degree of network convergence, ACC represents the degree of network convergence, and MIOU represents the coincidence degree between the detected target mask and the real target, as shown in (10), (11),and (12).

$$\mathrm{L}=\mathrm{L}_{cls}+\mathrm{L}_{box}+\mathrm{L}_{mask} \tag{10}$$

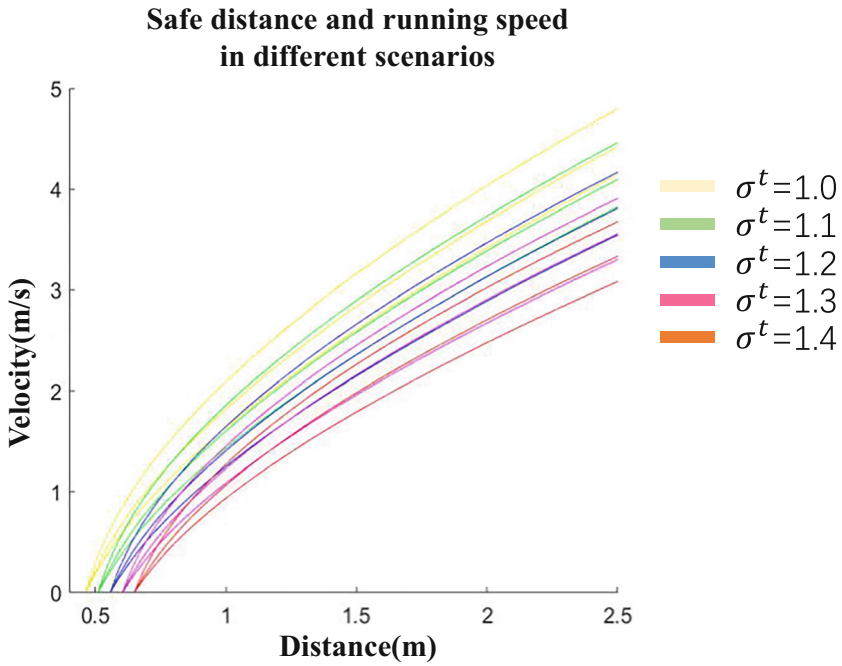

Fig. 7 The relationship between safe distance and running speed in three scenarios with different degrees of danger

$$\mathrm{A}cc = \frac{TP + \mathrm{TN}}{TP + FN + \mathrm{FP} + \mathrm{FN}} \tag{11}$$

$$\mathrm{M}IoU = \frac{1}{N+1} \sum_{i=1}^{N} \frac{TP_i}{TP_i + FP_i + FN_i} \tag{12}$$

The loss value represents the convergence degree of the network. It can be seen from Fig. 9 that due to the image light quality and other reasons, the loss value of the Mask-RCNN network eventually fluctuates around 0.2, while the loss value of MSNET eventually drops below 0.05. This is because our dataset has pictures taken under different lighting. The Mask-RCNN network has good detection results for pictures taken under normal lighting, but it is difficult to accurately detect images taken under low light. Its detection effect is affected by light, and the detection effect is unstable, resulting in a higher final loss value. The illumination has little effect on the depth image and the temperature image. The addition of these two kinds of information enables the MSNET network to have a relatively stable effect under different illumination conditions, so the final loss value is lower.

As can be seen from Fig. 10, the detection of human body contours by the original network in dark light is inaccurate, and the judgment of gripper sometimes makes mistakes. After fusing DT information, the effect is significantly improved. It can be seen from Table 2 that the detection effect of three targets is greatly improved by adding temperature images for robots, operators and non-operators, with the detection accuracy increased by more than 3% and IOU increased by 0.07, while the detection effect of two targets is greatly improved by adding depth images for gripper 1 and gripper 2, with the detection accuracy increased by more than 3% and IOU increased by 0.07. This phenomenon occurs because these targets have different characteristics. The heating of the robot body and the temperature of the human body lead to

Fig. 8 Experimental scene

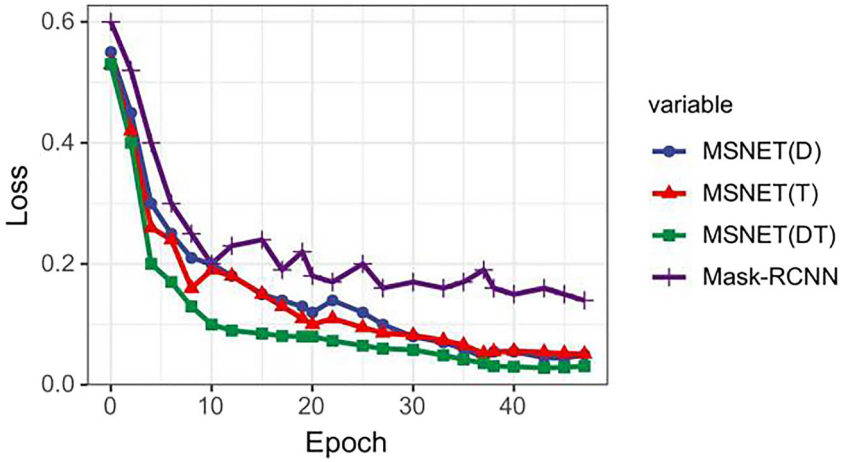

Fig. 9 Comparison of loss changes

significant differences between these three targets and the surrounding temperature, and the spatial position of the two claws is quite different from the surrounding environment. Therefore, the detection effect of temperature information on robot, operator and non-operator is significantly improved, while the detection effect of depth information on two kinds of grippers is greatly improved. DT image combine the characteristics of the two kinds of information and have a good effect on the detection of all targets.

4.2 Safety Monitoring Effect Experiment

The M-SSM proposed in this paper will adjust the safety threshold through the risk factor according to the number of dangerous targets and the types of tasks in the robot working environment; to intelligently adjust the running state of the robot. In order to verify the effectiveness of the dynamic risk factor, we set up four scenarios with different risk factors, in which the risk targets are: gripper1 + operator, gripper2 + operator, gripper1 + non-operator,gripper2 + non-operator (robot omitted). The risk factors are 1.1, 1.15, 1.2 and 1.25 respectively. In the first group of experiments, we carry out the grasping task in these four scenarios respectively. In the process of each grasping task, there will be operators or non-operators approaching the position that the robot will reach with a fixed trajectory. In this experiment, this position is the middle point of the path that the robot moves in a straight line after grasping. After the robot catches the object, the man begins to move. Because operators are familiar with the trajectory of the robot, they will eventually stop at about 10 cm around the trajectory of the robot, while non-operators will directly invade into the motion path. If the robot stops before colliding with the human body, it will be regarded as successful. If there is a collision or the speed does not drop to 0, it will be regarded as a failure. The standard SSM algorithm without risk factors and the M-SSM algorithm proposed in this paper are used to conduct 100 experiments in four scenarios respectively, and the success rate of monitoring and the completion time of tasks are recorded.

The second group of experiments set up a scene in which the risk factors changed dynamically. In this scene, the risk target was: Gripper1 + Operators + Non-operators. The experimental method is to let the operator perform the same intrusion as the previous group of experiments. After the robot

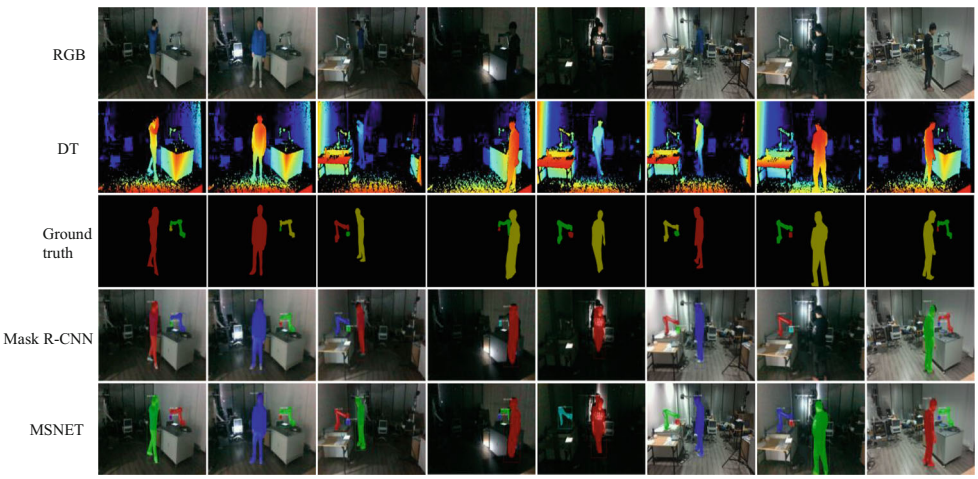

Fig. 10 Detection effect diagram

Table 2 Comparison of network effects

Method	Robot		Gripper1		Gripper2		Operator		Non-operator	
	Acc	mIOU	Acc	mIOU	Acc	mIOU	Acc	mIOU	Acc	mIOU
Mask-RCNN	0.932	0.769	0.836	0.601	0.824	0.598	0.943	0.713	0.933	0.692
MSNET(D)	0.947	0.787	0.844	0.663	0.851	**0.629**	0.956	0.762	0.948	0.745
MSNET(T)	**0.964**	0.843	0.837	0.608	0.828	0.603	0.977	**0.795**	0.982	0.759
MSNET(DT)	0.963	**0.852**	**0.849**	**0.669**	**0.855**	0.628	**0.983**	0.791	**0.988**	**0.767**

stops, the operator exits immediately, and then the non-operator immediately enters to perform the same intrusion as the previous group. Only if the two times are successful can it be recorded as successful. To verify the impact of dynamic risk factors, we manually solidify the risk factors of the M-SSM algorithm into the fixed value when operators enter for experiments, and compare it with the standard SSM without risk factors and the dynamic SSM algorithm(ours) with a real-time change in risk factors, recording the success rate and task completion time.

As can be seen from Table 3, the success rate of standard SSM will decrease significantly when non-operators enter, compared with the scenario where operators enter. This is because operators are familiar with the movement of the robot. They will actively stop outside the path, only occasional improper operation may lead to collision, so the robot is easy to avoid. The non-operator is not familiar with the motion of the robot and often invades the motion path of the robot, which makes it difficult for the robot to avoid. Compared with the traditional algorithm, M-SSM algorithm will adjust the speed of the robot according to the danger degree of the scene. In the dangerous scene entered by non-operators, the speed of the robot will be reduced to the safe speed in advance, which makes it easier for the robot to avoid dangerous targets and improve the safety of the robot system. Therefore, in the two scenarios entered by non-operators, the monitoring success rate increased by 12.3% and 11.2% respectively, and the completion time increased by 3.4 s and 4.7 s respectively.

Table 3 Comparison of monitoring success rate

Scene	Risk factor	Success rate	Average time
Gripper1+ Operator	none(SSM)	0.955	25.3 s
	σ'=1.1(ours)	0.985	27.4
Gripper2+ Operator	none(SSM)	0.938	25.6
	σ'=1.2(ours)	0.993	26.8
Gripper1+ Non-operator	none(SSM)	0.851	25.7
	σ'=1.3(ours)	0.974	29.1
Gripper2+ Non-operator	none(SSM)	0.867	26.2
	σ'=1.4(ours)	0.979	30.9

In the workspace, the risk factors within the monitoring range will change at any time. The following dynamic experiments show the effect in this case. When operators and non-operators enter alternately, the success rate of standard SSM algorithm without risk factors is 81.5%; When the risk factor was fixed at 1.1, the success rate was 91.8%; When the M-SSM algorithm with risk factors changes, the success rate is 96.7%. It can be seen that when the risk factors change dynamically for different scenarios, the success rate is greatly improved compared with the first two. This is because the risk degree in different scenarios will affect the timing and speed of robot adjustment. When the risk degree of the scene changes during the operation of the robot, the risk coefficient should also be adjusted in time. When the risk coefficient is fixed at 1.1, the risk coefficient does not match the scene, and the speed of the robot is not adjusted in time, which will reduce the success rate. Due to early deceleration, the completion time of dynamic SSM algorithm increases slightly, but in practical work, the number of intrusion is greatly reduced compared with the experimental environment, so it has little impact on the overall efficiency of the robot. In short, when M-SSM algorithm is used, lower efficiency can be reduced in exchange for significant improvement of robot safety Table 4.

5 Conclusion

In this paper, scene semantic information is employed in dynamic speed and separation monitoring for HRCs. First, a risk information fusion perception network MSNET is designed, which combines depth and temperature information with RGB images. MSNET uses the feature pyramid structure to extract the features of dangerous targets in dark or backlit

Table 4 Effect of dynamic risk factors

Scene	Risk factor	Success rate	Average time
Gripper1 + Operator+ Non-operator	none(SSM)	0.815	50.1 s
	1.1(ours)	0.918	53.9 s
	Dynamic(ours)	0.967	55.2 s

work areas to accurately obtain semantic information in the work scene. Then, the dynamic conversion of scene semantic information to risk factors is realized through the dynamic risk information database. Finally, the M-SSM algorithm is proposed, which can dynamically adjust the state of the robot system according to the degree of danger in the scene to avoid collisions. Experiments have shown that safety is increased by more than 15% with little reduction in efficiency. The algorithm relies on the neural network to accurately extract the semantic information of the scene, and the training of the neural network requires a large amount of data, which has high requirements for data collection and labeling. In the future, we need to refine and classify various complex scenes and study the reaction strategies of robots in different scenes.

Acknowledgement Research was supported by the National Key Research and Development Program of China under Grant 2019YFB1310200.

Code Availability All data and materials as well as software application or custom code support our published claims and comply with field standards.

Authors' Contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Botao Yang, Shuxin Xie, Guodong Chen, Zihao Ding, and Zhenhua Wang. The first draft of the manuscript was written by Botao Yang and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript. Botao Yang and Shuxin Xie are contributes equally to this work.

Funding Partial financial support was received from the National Key Research and Development Program of China under Grant 2019YFB1310200.

Declarations

Conflict of Interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethics Approval Approval was obtained from the ethics committee of Soochow University.

Consent to Participate Informed consent was obtained from all individual participants included in the study.

Consent for Publication The participant has consented to the submission of the research manuscript to the journ.

References

1. Ajoudani, A., Zanchettin, A.M., Ivaldi, S., Albu-Schäffer, A., Kosuge, K., Khatib, O.: Progress and prospects of the human—robot collaboration. Auton. Robot. **42**(5), 957–975 (2018)
2. Bdiwi, M.: Integrated sensors system for human safety during cooperating with industrial robots for handing-over and assembling tasks - ScienceDirect. Proc. CIRP. **23**, 65–70 (2014)
3. Matthias, B., Reisinger, T.: Example Application of ISO/TS 15066 to a Collaborative Assembly Scenario. In: Isr: International Symposium on Robotics (2016)
4. Marvel, J.A.: Performance metrics of speed and separation monitoring in shared workspaces. IEEE Trans. Autom. Sci. Eng. **10**(2), 405–414 (2013)
5. Zanchettin, A.M., Ceriani, N.M., Rocco, P., Ding, H., Matthias, B.: Safety in human-robot collaborative manufacturing environments: metrics and control. IEEE Trans. Autom. Sci. Eng. **13**(2), 882–893 (2015)
6. Shin, H., K. Seo, and S. Rhim. Allowable Maximum Safe Velocity Control Based on Human-Robot Distance for Collaborative Robot. 2018
7. Byner, C., Matthias, B., Ding, H.: Dynamic speed and separation monitoring for collaborative robot applications - Concepts and performance. Robot. Comput. Integr. Manuf. **58**(AUG.), 239–252 (2019)
8. Cai, K., Wang, C., Song, S., Chen, H., Meng, M.Q.H.: Risk-aware path planning under uncertainty in dynamic environments. J. Intell. Robot. Syst. **101**(3), 47 (2021)
9. Tarbouriech, S., Suleiman, W.: Bi-objective motion planning approach for safe motions: application to a collaborative robot. J. Intell. Robot. Syst. **99**(1), 45–63 (2020)
10. Chen, J.H., Song, K.T.: Collision-free motion planning for human-robot collaborative safety under Cartesian constraint. In: 2018 IEEE international conference on robotics and automation (ICRA) (2018)
11. Wei, Q., Zha, D., Jie, Z.: An effective approach for causal variables analysis in diesel engine production by using mutual information and network deconvolution. J. Intell. Manuf. **9**, 1–11 (2018)
12. Marvel, J.A., Falco, J., Marstio, I.: Characterizing task-based human–robot collaboration safety in manufacturing. IEEE Trans. Syst. Man Cybern. Syst. **45**(2), 260–275 (2015)
13. Lucci, N., Lacevic, B., Zanchettin, A.M., Rocco, P.: Combining speed and separation monitoring with power and force limiting for safe collaborative robotics applications. IEEE Robot. Autom. Lett. **5**(4), 6121–6128 (2020)
14. Kim, E., Kirschner, R., Yamada, Y., Okamoto, S.: Estimating probability of human hand intrusion for speed and separation monitoring using interference theory. Robot. Comput. Integr. Manuf. **61**, 101819.1–101819.7 (2020)
15. Kumar, S., Arora, S., Sahin, F.: Speed and Separation Monitoring using on-robot Time–of–Flight laser–ranging sensor arrays. In: 2019 IEEE 15th International Conference on Automation Science and Engineering (CASE) (2019)
16. Mazhar, O., Navarro, B., Ramdani, S., Passama, R., Cherubini, A.: A real-time human-robot interaction framework with robust background invariant hand gesture detection. Robot. Comput. Integr. Manuf. **60**, 34–48 (2019)
17. Aliev, K., Antonelli, D.: Proposal of a monitoring system for collaborative robots to predict outages and to assess reliability factors exploiting machine learning. Appl. Sci. **11**(4), 1621 (2021)
18. Wu, Q., Ding, K., Huang, B.: Approach for fault prognosis using recurrent neural network. J. Intell. Manuf. **31**(7), 1621–1633 (2020)
19. Sun, Y., Zuo, W., Liu, M.: RTFNet: RGB-thermal fusion network for semantic segmentation of urban scenes. IEEE Robot. Autom. Lett. **4**(3), 2576–2583 (2019)
20. Qi, C.R., et al.: PointNet: deep learning on point sets for 3D classification and segmentation. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR) (2017)
21. Lecun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature. **521**(7553), 436–444 (2015)
22. Haddadin, S., Luca, A., Albu-Schäffer, A.: Robot collisions: a survey on detection, isolation, and identification. IEEE Trans. Robot. **33**(6), 1292–1312 (2017)

23. Albini, A., Cannata, G.: Pressure distribution classification and segmentation of human hands in contact with the robot body. Int. J. Robot. Res. **39**(6), 668–687 (2020)
24. ArkinJacob, et al.: Multimodal estimation and communication of latent semantic knowledge for robust execution of robot instructions. Int. J. Robot. Res. **39**(10–11), 1279–1304 (2020)
25. Zhou, T., Wachs, J.P.: Spiking neural networks for early prediction in human–robot collaboration. Int. J. Robot. Res. **38**(14), 1619–1643 (2019)
26. Bouvrie, J.: Notes on Convolutional Neural Networks. neural nets (2006)
27. He, K., et al.: Mask R-CNN. In: 2017 IEEE International Conference on Computer Vision (ICCV) (2017)
28. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (2017)
29. The American Society of Safety Engineers: ANSI Z10, Occupational health and safety management systems. The American National Standards Institute (2012)

Botao Yang was born in 1997. He received the B.Sc. degree from Hunan University of Technology, China, in 2019, where he is currently pursuing the M.A. degree in School of Mechanical and Electric Engineering, Soochow University. His research interests include computer vision and robotics.

Shuxin Xie was born in 1980. He received the B.E. degree from the School of Material Engineering and the M.S. degree from the School of Mechanical and Electrical Engineering, Soochow University, Suzhou, China, in 2003 and 2010, respectively. He is currently pursuing the Ph.D. degree in intelligent robot technology. His research interests include robot vision and motion planning.

Guodong Chen was born in 1983. He received the Ph.D. degree from the Harbin Institute of Technology, Harbin, China, in 2011. He is currently an Associate Professor with Soochow University. His research interests include robot vision and intelligent industrial robots.

Zhihao Ding was born in 1996. He received the B.Sc. degree from Wuhan Textile University, China, in 2017, where he is currently pursuing the Ph.D. degree in School of Mechanical and Electric Engineering, Soochow University. His research interests include robot vision and robotics.

Zhenhua Wang was born in 1974. He received Ph.D. degree from the Harbin Institute of Technology, Harbin, China. He is young and middle-aged academic leaders of the "Blue Project" in Jiangsu Province. His research interests include industrial robots and intelligent automation equipment.