Page Proof Instructions and Queries

Journal Title: PIB
Article Number: 1223783

Thank you for choosing to publish with us. This is your final opportunity to ensure your article will be accurate at publication. Please review your proof carefully and respond to the queries using the circled tools in the image below, which are available in Adobe Reader DC* by clicking **Tools** from the top menu, then clicking **Comment**.

Please use *only* the tools circled in the image, as edits via other tools/methods can be lost during file conversion. For comments, questions, or formatting requests, please use T. Please do *not* use comment bubbles/sticky notes.

*If you do not see these tools, please ensure you have opened this file with **Adobe Reader DC**, available for free at get.adobe.com/reader or by going to Help >Check for Updates within other versions of Reader. For more detailed instructions, please see us.sagepub.com/ReaderXProofs.

Sl. No.	Query
	Please note that we cannot add/amend orcid ids for any article at the proof stage. following orcid's guidelines, the publisher can include only orcid ids that the authors have specifically validated for each manuscript prior to official acceptance for publication.
	Please confirm that all author information, including names, affiliations, sequence, and contact details, is correct.
	Please review the entire document for typographical errors, mathematical errors, and any other necessary corrections; check headings, tables, and figures.
	Please ensure that you have obtained and enclosed all necessary permissions for the reproduction of art works (e.g. illustrations, photographs, charts, maps, other visual material, etc.) not owned by yourself. please refer to your publishing agreement for further information.
	Please note that this proof represents your final opportunity to review your article prior to publication, so please do send all of your changes now.
	Please confirm that the funding and conflict of interest statements are accurate.
1	Please Confirm Corresponding Author Email Id
2	Please approve the citation provided for Table 1.
3	Please provide publisher location for reference "4."
4	Please provide place, date, publisher name, and publisher location for reference "6."
5	Please provide place, date, publisher name, and publisher location for reference "8."
6	Please provide place, date, and publisher location for reference "10."
7	Please provide publisher location for reference "12."
8	Please provide place, date, and publisher location for reference "20."
9	Please provide volume and page range for reference "24."
10	Please provide place, date, and publisher location for reference "26."
11	Please provide volume and page range for reference "29."
12	Please provide place, date, and publisher location for reference "33."
13	Please provide editor name, publisher name, and publisher location for reference "34."
14	Please provide publisher location for reference "36."
15	Please provide place, date, and publisher location for reference "38."
16	Please provide place, date, publisher name, and publisher location for reference "39."
17	Please provide place, date, publisher name, and publisher location for reference "41."

Original Article

Institution of MECHANICAL ENGINEERS
Journal of ENGINEERING MANUFACTURE

Proc IMechE Part B:
J Engineering Manufacture
1–15

DOI: 10.1177/09544054231223783
journals.sagepub.com/home/pib
Sage

Digital twins for hand gesture-guided human-robot collaboration systems

Ao Liu[1], Yifan Zhang[2] and Yuan Yao[2,3]

Abstract

Gesture control is one of the effective and flexible communication method between humans and robots. However, it always depends on complex hardware and configurations in human-robot collaboration systems. Simplifying the design of gesture-interaction systems and avoiding miscommunication are challenging problems. In this paper, we proposed a method that utilizes an RGB sensor to realize spatial human-robot collaboration. A random forest based depth estimator is presented to supplement the additional spatial information for hand gesture recognition. Additionally, we demonstrate the construction of secure human-robot collaboration scenarios in Unity and validate our approach in real-world settings, based on which a digital twin system oriented to human-machine collaboration is constructed to realize rapid human-machine task simulation, safety specification testing, and real-scene applications development.

Keywords

Human-robot collaboration, random forest depth estimation, digital twin system

Date received: 3 July 2023; accepted: 5 December 2023

Introduction

Human-robot collaboration (HRC) is consistent with the developmental trends in modern manufacturing flexibility. Numerous collaborative robots have been designed to efficiently assist human workers in various production tasks.[1] When dealing with complex production lines, it is crucial to assess the feasibility of utilizing robots for adaptable selection by considering changing environments.[2] Simulation models cannot capture the full real-world environment of an industry because of the uncertain and highly dynamic nature of the production workflows. Under such dynamic conditions, it is difficult to improve the dynamic adaptability of the robots. Despite the rapid development of machine learning techniques, constructing massive datasets that simulate the production time remains a very difficult task. The challenges associated with dynamic robot training and the execution of different tasks can be significantly alleviated by deploying a collaborative environment in which robots and humans work together on the production line.[3]

From another point of view, although automation allows robots to perform repetitive and unergonomic works in the industrial sector, human operators are still required to perform precise works that require dexterity and flexibility, thereby fully utilizing both robots and humans.[4] Thus, the complementarity between robots and human capabilities can be fully utilized to achieve the goals of intelligent and flexible manufacturing.

In modern HRC systems, there are multiple ways of interaction, and the most widely used method is gesture.[5] By utilizing on the method of gesture, robots can interpret human intent through technologies such as sensors based on electromagnetic effects,[6] gloves equipped with angle sensors,[7] exoskeleton systems,[8] or visual recognition techniques.[9]

However, traditional interaction methods in HRC often require complex configurations in practical application systems and sophisticated sensors, which means high cost and cumbersome configuration process.[10] Moreover, the existing sensor configurations have limited applicability and strict environmental requirements, such as limitations on penetration depth when using a sub-THz imaging camera[11] and the lack of

[1]Sino European School of Technology, Shanghai University, Shanghai, China
[2]Engineering Training Center, Shanghai University, Shanghai, China
[3]School of Mechatronic Engineering and Automation, Shanghai University, Shanghai, China

Corresponding author:
Yuan Yao, Sunrise School of Mechatronic Engineering and Automation, Shanghai University, 99 Road Shangda, P.O. Box 113, Shanghai, 200444, China.
Email: yaoyuan@shu.edu.cn[AQ: 1]

infrared sensors in bright sunlight.[12] These limitations emphasize the need for alternative solutions that are both cost effective and adaptable to diverse industrial environments.

Related works

Existing solutions focus on different gesture recognition methods and digital twin system constructions in the HRC system.

Gesture in HRC system

Collaborative robots are specifically designed to interact with humans in shared environments and offer the advantage of quick and cost-effective layout changes.[13] In the field of HRC, safety concerns have been addressed through various strategies such as workspace partitioning and active human engagement with robots.[14] By leveraging AI and deep learning, predictive capabilities can be employed to anticipate potential issues or failures in the manufacturing process before they manifest themselves in real-life scenarios.[15] The integration of manufacturing and business processes through Cyber-Physical Systems in intelligent factories brings numerous benefits in terms of quality, time, resources, and cost.[16]

Several notable contributions have been made in gesture-guided HRC systems. Su et al.[17] introduced the use of electromyography signals for accurate gesture classification. Ding and Su[18] used Hu moment invariants from depth images to establish interaction patterns with manipulator robots. Che and Qi[19] proposed a novel method for real-time 3D hand tracking using depth images. Another approach involves employing a Microsoft Kinect device to provide guidance to a collaborative robot.[20] Rautiainen et al.[21] proposed a multimodal offline and online programming framework that allows the use of custom gestures in collaborative work.

Maintaining a safe distance between humans and robots is a critical aspect of an HRC system. Kianoush et al.[22] utilized IoT platforms and multi-sensor data fusion to ensure effective human-robot distancing. In the realm of robot motion planning algorithms, Liu et al.[23] enhanced the algorithm by conducting quantitative research on the impact of human psychological responses. Conversely, Liu et al.[24] developed a neural model that predicts and quantifies the uncertainty of human motion in collaborative work. Trinh et al.[25] ensured interaction distance by implementing a framework that combines Behavior Trees and Computer Vision. Jantos et al.[26] introduced a transformer-based approach that leverages RGB images as an input to predict 6D poses of objects within the images.

Digital twin system in HRC

Utilization of the virtual environment provided by digital twin technology enables more efficient safety validation processes.[27] Douthwaite et al.[28] introduced a novel modular digital twinning framework specifically designed for safety validation in HRC manufacturing processes. Ye et al.[29] used digital twin-verified human-machine collaboration to achieve a more efficient building assembly. To address the integration of human factors in contemporary scenarios, Kuts et al.[30] conducted comprehensive performance tests on a digital twin virtual reality operation interface based on production units. These efforts contributed to enhancing safety measures and improving the overall performance of the HRC.

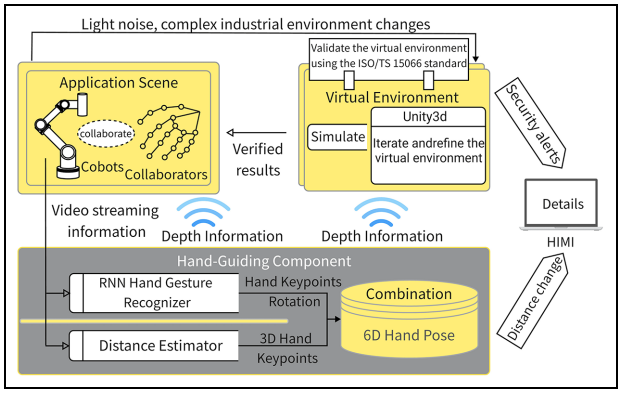

Figure 1. Hand gesture guided digital twin for HRC.

Upon iterative validation of the virtual environment framework, several proposed safety algorithms can be utilized to calibrate different safety scenarios. Oyekan et al.[31] introduced an innovative approach that utilizes virtual reality digital twins of physical layouts to study human reactions to predictable and unpredictable robotic motions. Du et al.[32] use augmented reality and multi-sensory feedback to make human-robot collaboration more natural. Che and Qi[33] proposed a quick method for simulating the interaction between human hands and virtual objects. Antakli et al.[34] proposed a 3D simulation framework for mixed teams in production scenarios based on an agent-based framework. Another significant research achievement is the unified framework developed by Malik et al.,[35] which integrates human-robot simulation with virtual reality.

The digital twin system is always a critical component of the HRC system, as its use provides collaborators with additional information to mitigate safety errors.

Method

To simplify the configuration of HRC systems in various industrial environments, we propose an HRC framework that is easy to deploy, as shown in Figure 1.

This framework contains three parts: hand-guiding component, virtual environment, and application scene.

Hand-guiding component

In the hand-guiding component, the hand-tracking method uses only a monocular egocentric RGB camera

Table 1. Comparison of proposed research with existing literature.

Issues in existing literature	Existing solutions	Proposed research
Gesture recognition methods have limitations in accuracy and real-time performance	Use different sensors and algorithms for gesture recognition, for example, EMG, depth cameras, 3D hand tracking	Propose a new CNN-based method with higher accuracy and real-time performance
Digital twin systems lack integration with gesture control and focus on simulation	Develop modular frameworks and simulations for virtual validation of HRC safety	Integrate gesture control with digital twin system for closed-loop validation and control
Safety distance keeping lacks real-time adaptive adjustment	Static distance thresholds or model-based predictions	Real-time distance adjustment based on gesture recognition and risk assessment
Motion planning lacks psychological adaptation	Some work on modeling human discomfort	Integrate perceived safety into motion planning

to obtain the three-dimensional (3D) information. To acquire real-time six-degree (6D) information of the hand from the RGB image, depth information plays a pivotal role. Relying solely on RGB images to obtain depth information is undoubtedly an effective approach. The video streaming information obtained from the application scene is sent to the Recurrent Neural Network (RNN) hand gesture recognizer and distance estimator to extract features. Subsequently, the three-dimensional hand pose, which is based on the coordinates of the camera center, is obtained from a combination of hand gesture and depth information.

Virtual environment

By analyzing the information of 6D hand pose in a virtual environment, the digital twin system can simulate various possible industrial environments and validate the safety of interactions based on predefined safety strategies to ensure the effectiveness of our safety strategies. To prioritize safety in a virtual environment, we developed an HRC digital twin system that adhered to the ISO 15066[36] standard. When receiving hand-pose information from the hand-guiding component, the virtual environment simulates the interaction under multiple situations that are preset to validate whether it is safe. In our HRC system, if the validation results indicate that the interaction is safe, the collaborative robots are allowed to operate as intended. However, if the interaction is deemed unsafe, corresponding emergency strategies are triggered for the robot, and the system projects safety alerts onto external displays such as HMI or uses flashing lights to alert personnel of potential hazards. For the digital twin system to work properly, the coordinate transformation between these three parts adopts an efficient communication system, as introduced in Section (Table 1).[AQ: 2]

Application scene

Adopting the HRC framework in the real-life interaction scenarios, Every robot's motion was driven by the verified results to collaborate with humans. This scenario reflects HRC framework in actual collaboration environments, where the robot works closely with human operators to accomplish tasks together. In this interaction scenario, the actions of the collaborative robot must adapt flexibly to the guidance of the human and be able to stop or adjust immediately when necessary to ensure the operator's safety. Therefore, having a very high corresponding speed will greatly improve the stability and safety of the HRC framework. The errors generated by this method are allowed in many human-robot collaboration scenarios such as hybrid order picking.[37]

Hand-guiding component

In the hand-guiding component, dividing depth information capture and hand gesture recognition from hand tracking results in greater efficiency; details of the process are shown in Figure 2.

There are two subcomponents to receive the video streaming information:

- RNN Hand Gesture Recognizer: This component mainly used to gain the rotation of hand keypoints.
- Distance Estimator: The main objective of this component is to estimate the depth information from the RGB image. And Depth is used to expand the dimension based on 2D information.

A considerable amount of research has been conducted on RNN hand-gesture recognition. The approach outlined by Chen et al.[38] was employed to construct an RNN model specifically for dynamic hand gesture recognition in this particular subcomponent. This approach can be used to obtain information regarding hand keypoint rotation. The use of an end-to-end framework ensures the accuracy of the gesture 3D estimation under occlusion.[39] Depth plays an important role in obtaining 3D hand keypoint information based on RGB 2D image. Therefore, the primary focus of this study is distance estimation.

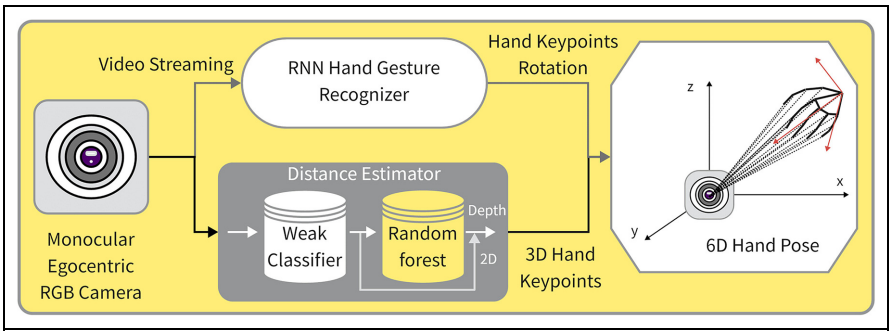

Figure 2. Hand-guiding component.

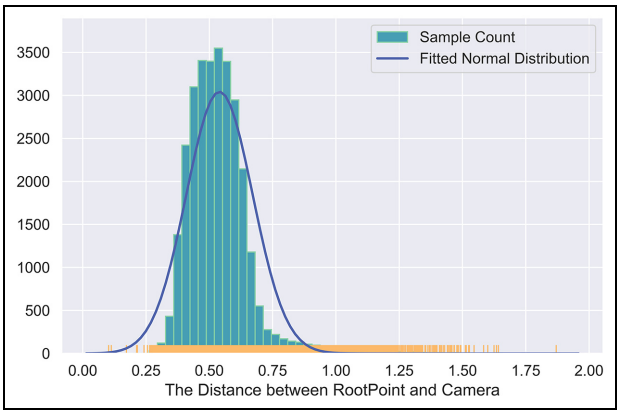

Figure 3. Sample distribution.

To further reduce the complexity of construction and deployment of the estimator, we built a random forest model based on the concept of federated learning. The initial processing involves transforming video streaming information through weak classifier recognition into hand keypoint data. Subsequently, the random forest distance estimator was employed to process the recognition results and generate depth information for the hand.

The federated learning approach has been adopted to construct a distance estimator. This solution consists of a machine learning pipeline comprising two models:

- Weak classifier: It extracts hand keypoints from the raw video stream 2D images.
- Random Forest Estimator: The capability of estimation is derived from the learned complex relationship between the hand keypoints and distance, based on the RGB-D dataset.

Through this approach, we could deploy the estimator without any settings, based solely on experience, as humans estimate 3D information in real life.

Once the depth information for each hand key point is obtained, the three-dimensional hand pose is calculated using vector calculations. By combining rotation and location, precise hand-guiding information is acquired and subsequently transmitted to the next component within our HRC framework.

Model selection and construction for distance estimator

By employing the federated learning approach, we selected the RNN + CNN combination[40] as the weak classifier module for our distance estimation system. To enhance the generalization capabilities, we opted to use the Random Forest algorithm. Additionally, using a smaller model in practical applications can still yield satisfactory results.

Preprocessing and data cleaning of the training dataset. We utilized the Rendered Hand Pose (RHD) dataset[41] for training and evaluating our model. This dataset comprised 41,258 training samples and 2728 test samples, each containing an RGB image capturing the gesture, a corresponding depth image, and a mask image indicating the hand region.

For preprocessing the RGB gesture images, we employed the hand gesture recognition model provided by Mediapipe as a weak classifier, enabling us to extract the coordinates for each hand keypoint. By locating the corresponding keypoints in the depth images, we recorded their respective depth values. After applying the weak classifier to all images, we obtained a total of 30,642 training samples with hand keypoints. Following data cleaning procedures, we were left with 30,100 usable training samples.

In order to ensure the generalization ability of the random forest distance estimator across different cameras, it was essential to normalize the input features and carefully select the appropriate features.

Analysis of feature engineering. A normal distribution analysis was conducted on the cleaned dataset considering the distance from the camera to the wrist root node as a random variable. Figure 3 illustrates the sample distribution, indicating an approximately normal-distribution pattern. The majority of samples were found within the range of 0.40–0.65 m, representing the

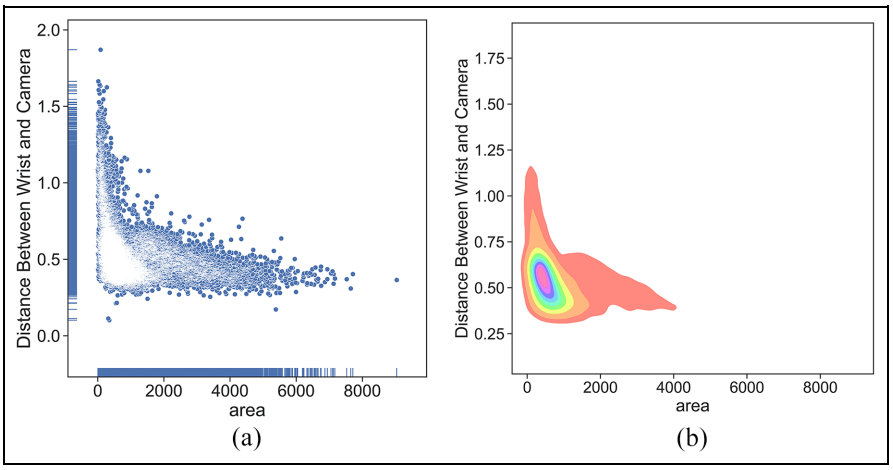

Figure 4. Palm area characteristics. (a) scatter plot. (b) Kernel density estimation.

most stable distance interval for optimal depth camera performance. Based on this distribution and the typical characteristics of random forest, it can be inferred that the model is expected to perform well within this interval. However, in other intervals, the predictive capabilities of the model may be limited because of insufficient sample size.

Post-processing was performed on the 3D mapping of the hand obtained from week classifier's recognition to identify significant features for the training set input of the random forest model. Through observation of hand poses, it was noted that the palm area remained relatively stable across most physiological states. Consequently, selecting this feature can yield the desirable inputs. A bivariate distribution plot, depicted in Figure 4, was created to demonstrate the relationship between the palm area calculated by weak classifier and the distance from the camera to the wrist root node.

Figure 4(a) shows the sample distribution plot on the left, while Figure 4(b) shows the kernel density plot on the right. Based on the kernel density plot, it is evident that the majority of the samples are concentrated in a specific region, with the density gradually decreasing towards the surrounding areas. By analyzing the distribution of the samples, it becomes apparent that the relationship between the root node and palm area follows a nonlinear distribution.

By examining the palm area, it is apparent that relying solely on its features for regression processing is inadequate. To address this limitation, we must treat it as a higher-dimensional regression problem, incorporating more features to improve predictive capabilities. As a result, we decided to introduce the length of each finger bone in the pixel mapping as an additional feature. Following the same methodology as that for the palm area, we generated a bivariate distribution plot, as depicted in Figure 5.

Twenty skeletal length features were calculated based on the coordinates of 21 hand keypoints obtained through weak classifier. Among these features, the first row of five bivariate distribution plots illustrates the distances from the hand's root node to each finger's metacarpophalangeal joint (also known as the palm bone). The second row of the five bivariate distribution plots represents additional manually added features that indicate the relative distances between the five metacarpophalangeal joints. These positions exhibit stable variations in relative length during actual motion postures, making them effective in reflecting the hand posture and positional characteristics. The third row of five bivariate distribution plots represents the relative lengths of the proximal phalanges of each finger, and the fourth and fifth rows represent the relative distances of the middle and distal phalanges.

It is worth noting that these 20 features are very similar to the hand area feature. Therefore, inputting only these features did not significantly improve the regression mapping capability of the random forest for distance estimation.

After re-evaluating the raw data, it was discovered that the angle information between the bones was overlooked during the extraction of skeletal length features. Consequently, the unit vectors of each bone in the weak classifier recognition mapping were recalculated. This adjustment aimed to increase the number of input features into the random forest, leading to a better understanding of the hand posture and more stable distance estimation outputs.

Scatter plots, as shown in Figure 6, were created for each bone in a 3D space to visually depict the relationship between the features and predicted values. In these plots, the color was used to represent the corresponding distances from the wrist and the camera.

The relationship between the feature vector of an individual bone and the predicted value exhibits nonlinearity, which necessitates the consideration of multiple features. In practical model training, the predictive capability of the random forest did not significantly improve, with the R2 value increasing from 0.55,

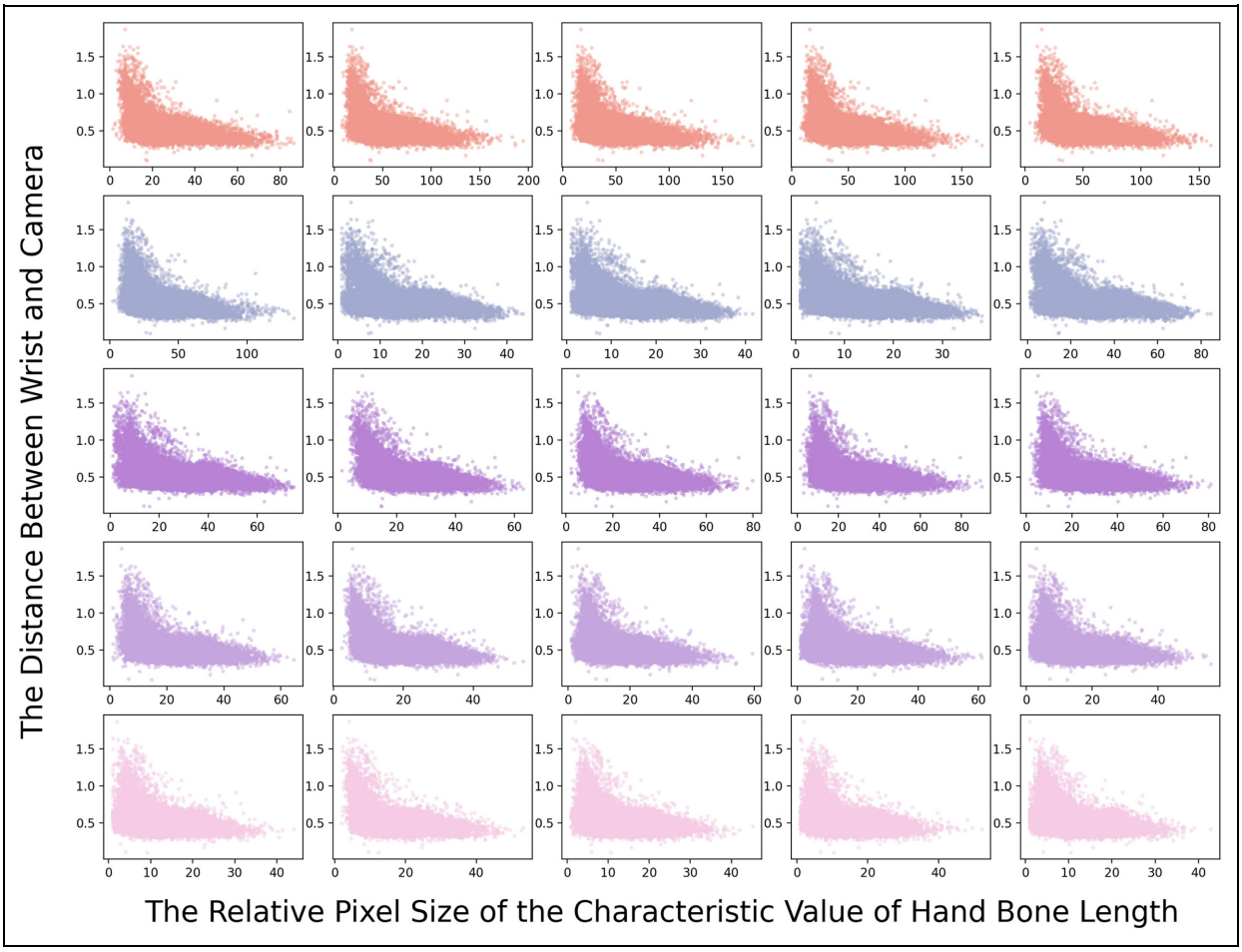

Figure 5. All hand bone length features.

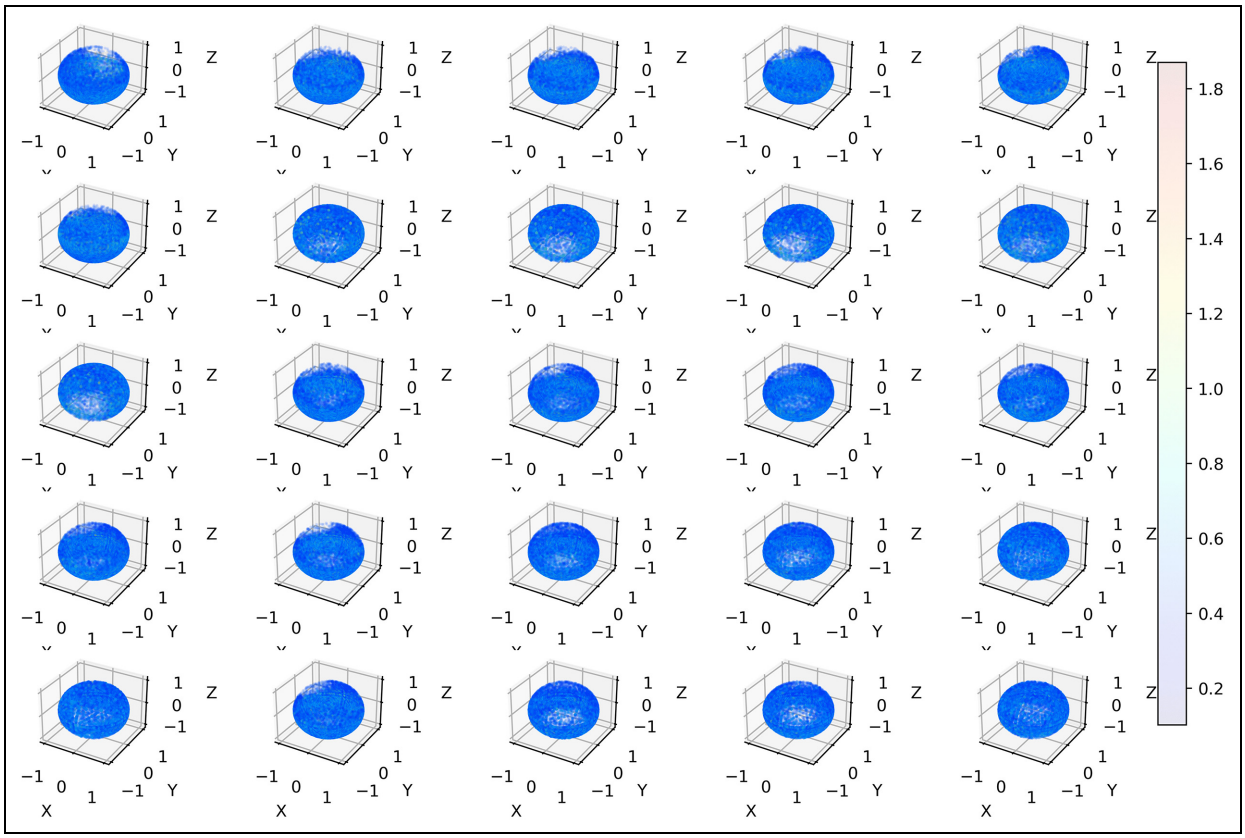

Figure 6. Bone feature vector distribution.

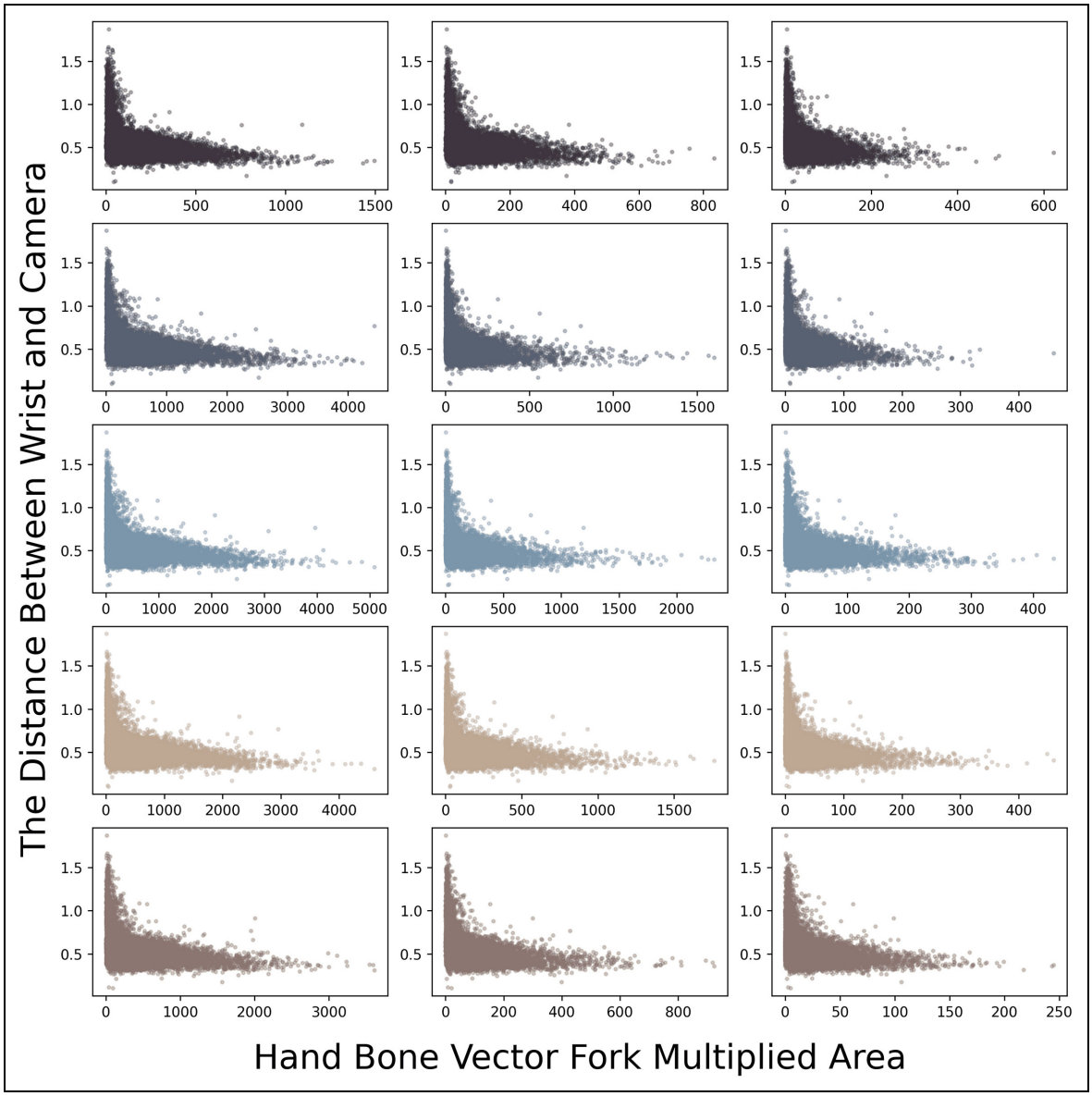

Figure 7. Hand bone vector fork multiplication distribution.

achieved when training with only skeletal length features, to 0.62. To represent the vector relationship between features more explicitly, improvements are required in the input of the features.

Simply inputting individual unit vectors does not sufficiently capture the relationship between bones; they merely indicate the orientation of the bones. To address this limitation and emphasize the preservation of 3D spatial features, an alternative approach was employed. Specifically, the cross-product of bone vectors was utilized as a feature to represent the interrelation between bones, enabling a more accurate depiction of their relationships. This process is visually depicted in Figure 7. A comparison between Figure 7 and Figure 6 facilitates the observation and analysis of how the incorporation of inter-bone relationships as additional features in the random forest input leads to more prominent features.

Building random forest distance estimator. Through the thorough feature-engineering analysis, a comprehensive set of 120 features was derived. These features encompass various aspects, such as skeletal length, bone unit vectors, and the cross-product area of the bone vectors. To facilitate the utilization of these features, they were flattened into a one-dimensional vector that served as the input for the random forest model. The primary objective of this model is to estimate the distance from the root node of the wrist to the camera.

After conducting hyperparameter tuning, the random forest model was trained using the following parameters:

n_estimators = 1500,
random_state = 0,
max_features = "log2"

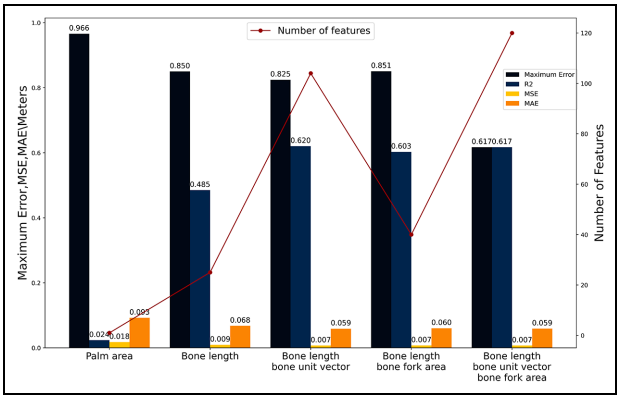

Figure 8. Evaluation of feature engineering model parameters.

Once trained, the random forest model was adept at receiving hand keypoints recognized by weak classifier as input and making predictions regarding the depth of the root node of the hand. By leveraging the 2D hand keypoints coordinates provided by weak classifier, straightforward vector computations enable the calculation of the spatial position relationship between each hand node and the camera.

Model evaluation based on feature engineering analysis

To validate the conclusions derived from the feature engineering analysis described earlier, random forest models were trained using various inputs. Figure 8 illustrates the performance of the random forest model on the test set as the number of features gradually increases.

Through a comparison of the model evaluations conducted on the training set, it was observed that several models exhibited similar metrics including R2, MSE, and MAE. However, the final selection was a random forest model that incorporated skeletal length, skeletal unit vectors, and skeletal cross-product area as input features. Among the models with comparable performance across all metrics, this particular model not only showed the smallest maximum error, but also encompassed the widest combination of input features. The rationale behind selecting the combination with the maximum number of input features is to theoretically achieve the smallest error and ensure the most stable performance.

Digital twin HRC system

HRC's digital twin consists of manually guided components, virtual environments, and application scenarios. The latter two are both driven by the gesture component.

In the virtual environment of the HRC framework, the ability to iterate and conduct rapid testing allows for the simulation and study of human-robot interactions in diverse industrial settings and challenging scenarios that are difficult to replicate in reality. This approach enables the precise control and manipulation of variables, including collaborative human characteristics and workload requirements of collaborative robots.

To ensure smooth operation and safety in application scenarios, it is necessary to create various typical industrial scenarios and extensively test them in the virtual environment. When constructing these scenarios, we need to define and create various interactive actions and possible collision situations in advance to simulate various situations in actual industrial environments.

Coordinate transformation for three parts

As shown in Figure 9, there are three modules that play indispensable roles in HRC system. The first module is the real interaction scene system, which represents the actual physical environment in which interactions between humans and robots occur. It provides real-world data and context as the foundation for our research. The second module is the interaction computation scene system that revolves around the collaborative robot and serves as an intermediate layer. It converts data from the real scene into a coordinate system relative to the collaborative robot, enabling subsequent transfer and decision making for the robot. The last module is the virtual environment iteration, which validates the safety strategy. It offers a virtual environment for conducting various experiments and simulations, allowing the evaluation of safety strategies under different scenarios, and providing validation results to the real environment to ensure collaboration safety. These three modules complement each other and form the essential components of our study.

To achieve coordinate transformation between these three modules, we employed the following methods and processes: First, we utilized a weak classifier for hand gesture recognition, which provided the camera-plane coordinates of the hand keypoints. Subsequently, the video stream information is processed using the distance estimator in the hand-guiding component. This step converts the camera plane coordinates into a 3D world coordinate system, using the camera as the origin. The 3D world coordinate system represents the absolute positions of hand keypoints in a real scene. In addition, we determined the position of the collaborative robot and calculated the relative positions of the joints of the hand with respect to the robot. By performing vector operations and coordinate transformations on the data in the world coordinate system, we can send all the relevant information to the virtual environment. Within the virtual environment, we utilized a coordinate system with the Tool Center Point (TCP) as the origin to facilitate further simulation and experimentation. Through iterative exploration of different scenarios and feedback of safety information from the virtual environment to the real environment, we can

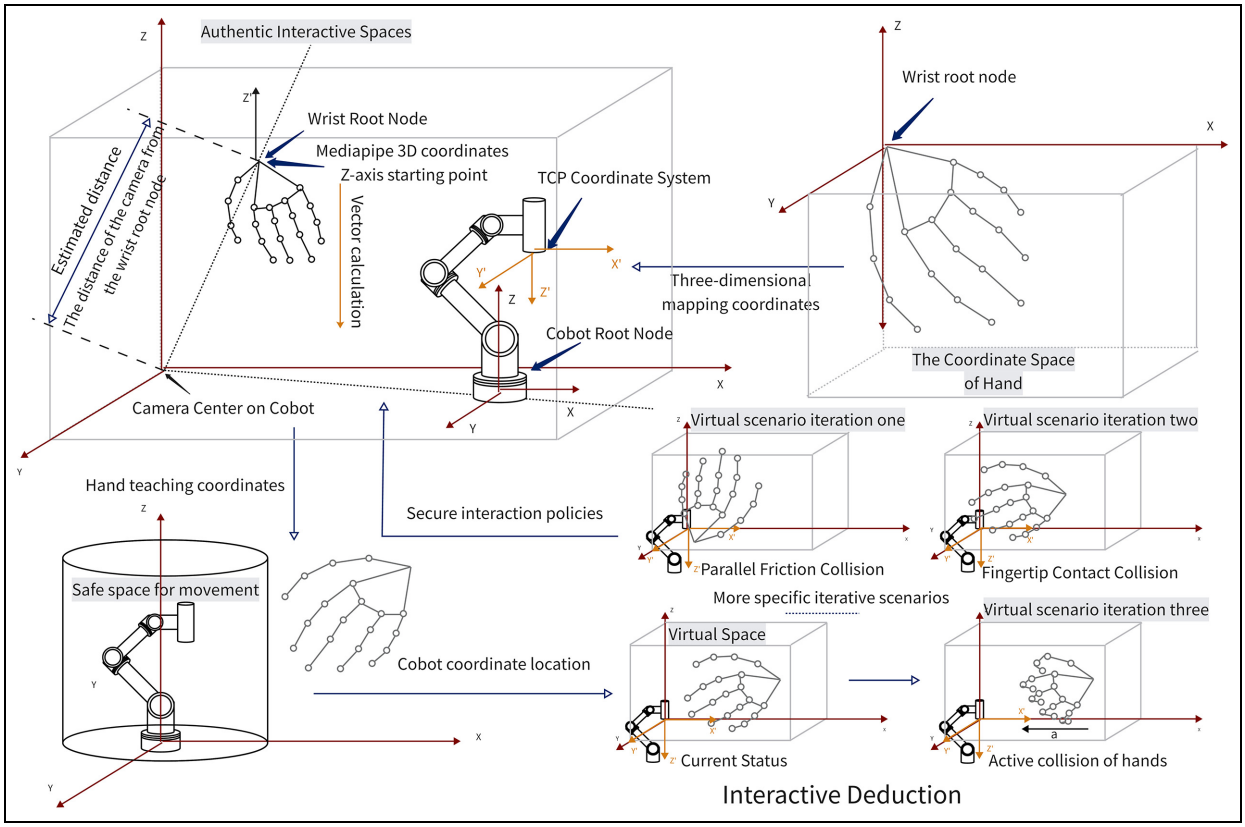

Figure 9. Transformation for HRC.

validate and adjust safety strategies to ensure collaboration safety.

This method and the process of coordinate transformation enable data transfer and decision-making between different systems, enabling effective human-robot collaboration. This allows us to acquire data from real scenes and transform it into a coordinate system relative to the collaborative robot, facilitating analysis and decision-making within the virtual environment.

Intercommunication for HRC system

Improving the performance of the HRC framework requires achieving effective communication among its three modules. To achieve this, a communication network was designed, as illustrated in Figure 10, based on the principles of fast intercommunication and rapid response.

The Hand-guiding component comprises two threads: one thread is dedicated to receiving video stream data from the camera, processing it for distance estimation, and packaging collaborative robot coordinate information. This information is then transmitted to a virtual environment. The other thread is responsible for receiving safety information from the virtual environment and controlling the collaborative motion

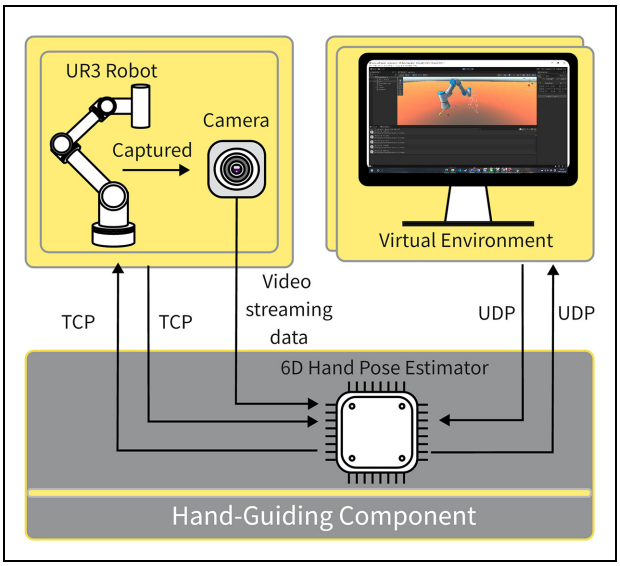

Figure 10. Intercommunication for HRC system.

of the robot. To enable rapid data transmission, communication between the module and virtual environment utilizes the UDP protocol. Additionally, communication between this module and the UR3 robot was established using the TCP protocol, ensuring

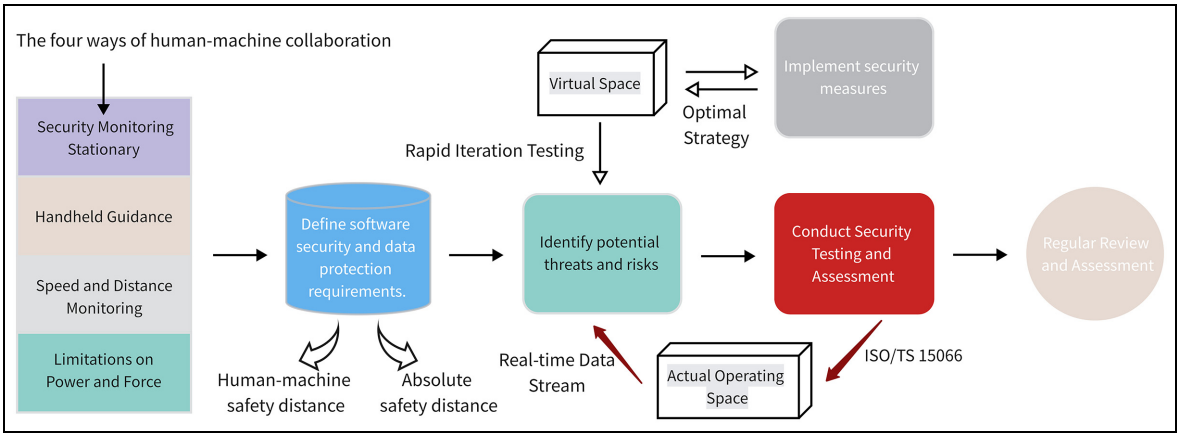

Figure 11. Verification process.

secure transmission of the safety information computed by the virtual environment.

HRC system verification process

Drawing upon the ISO/TS 15066[36] standard, a comprehensive evaluation process has been devised in the virtual environment to accommodate various industrial interaction scenarios. This verification process encompasses multiple steps, depicted in Figure 11.

Definition of virtual environment security information and data protection requirements: Our objective is to guarantee the reliability of the human-robot safety distance by dynamically adjusting its range based on real-time iterative updates of the spatial position of the human hand. This dynamic adjustment is crucial for maintaining a continuously safe interaction between the collaborative robot and the human operator.

Identification of potential threats and risks: By conducting iterative tests using predefined scenarios in a virtual environment, potential threats and risks can be proactively identified. This allows us to develop contingency plans that effectively safeguard personnel safety in practical applications.

Implementation of security measures: Tailored contingency solutions have been developed to address different hazards encountered in various industrial scenarios. Timely execution of optimal strategies is essential to maximize the safety of collaborators.

Conducting security testing and evaluation: Rigorous testing and evaluation are performed on emergency strategies and real-time collaborative distances to ensure compliance with the safety requirements. This ensured that the working distance and intensity requirements of the collaborative robot aligned with the ISO/TS 15066[36] standard in practical applications.

Regular review and assessment: The reliability of the collaborative contingency plans was reviewed and assessed regularly. Trigger events were documented, and contingency plans were improved based on statistical results to further enhance safety and reliability.

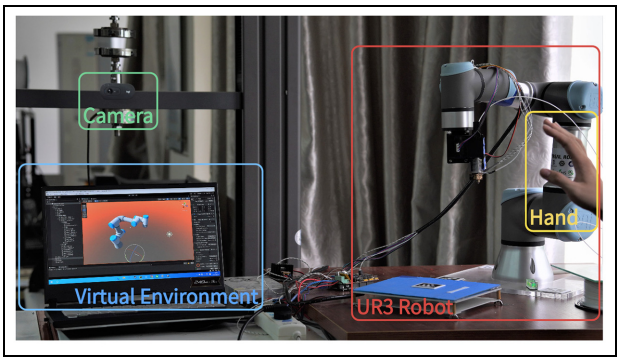

Figure 12. Collaboration space demonstration.

By following this verification process, different industrial interaction scenarios can be validated in virtual environments, thereby ensuring the safety of collaborative interactions between robots and human operators. Corresponding contingency plans were devised to minimize personnel safety risks.

Experiments and evaluation

Figure 12 illustrates the digital twin HRC system employed in collaborative tasks. Unity3d was selected as the platform for building the virtual environment, and the experiment was conducted using the UR3 robot. In this demonstration, a camera was strategically positioned to accurately estimate the 6D hand pose and to ensure collaboration safety.

The parameters of hardware devices to conduct this Demo experiment as shown in Table 2.

There are two experiments has been conducted:

- Estimator performance testing: To evaluate the performance of random forest distance estimator, another estimator that utilized focal length method has been testing together.

Table 2. Hardware parameters.

Hardware	Parameters
RGB camera	Logitech C270i
CPU	Intel Core i7-10875H
GPU	NVIDIA GeForce RTX 2060
RAM	16GB
SSD	512GB
Software	Parameters
OS	Windows 10
Unity version	2021.3.10f1c1

- Safety demonstration of HRC framework: The most common situation, which named collision, was selected to conduct this Demno experiment to verify the usability of the HRC framework.

Distance estimation and accuracy evaluation

The Demo experiment conducted with type of camera that was shown in Table 2. Following the method described in the previous section, we constructed a random forest distance estimator and evaluated its ranging accuracy across different distance intervals to measure its accuracy and stability.

A focal-length estimator was employed to compare the results of the two approaches. We tested the estimator by hand at various distance intervals and recorded the distance estimation results obtained from both methods.

Within the experimental space, we calibrated the spatial distances ranging from 30 to 80 cm, with calibrations performed at 5 cm intervals using the MAE metric as the spatial distance calibration standard. We tested these ten spatial distances and compared the accuracy of distance estimation using the random forest predictor and the focal length method.

During the testing process, participants performed a 2-min continuous circular motion in both clockwise and counterclockwise directions, with their wrists as the center and their hands naturally open. Figure 13 illustrates the bar chart depicting the accuracy of distance estimation obtained using the random forest predictor and the focal length method within these ten intervals. By simulating Gaussian noise and using a standard deviation of 25, we tested these two approaches. The results are shown in Figure 13.

This experimental evaluation aims to compare the performance differences in spatial accuracy between the random forest predictor and the focal length method. By comparing the test results, we can assess the predictive accuracy of these two distance measurement models within different intervals and draw conclusions to guide selection and decision-making in practical applications.

Through comparative experimental testing, we evaluated the two methods and made the following assessments for each aspect:

- Random Forest Distance Estimator: Within the distance range of 0.45–0.60 m, the random forest method exhibits greater stability. Even under larger motion amplitudes and Gaussian noise, this method yields relatively small errors. This indicates that the random forest method performs well for distance estimation within this range and has an ability to resist noise. In addition, the random forest distance estimator demonstrated stronger generalization capabilities. Because it relies on the normalized results obtained from weak classifier recognition, the camera parameters have a minimal influence on

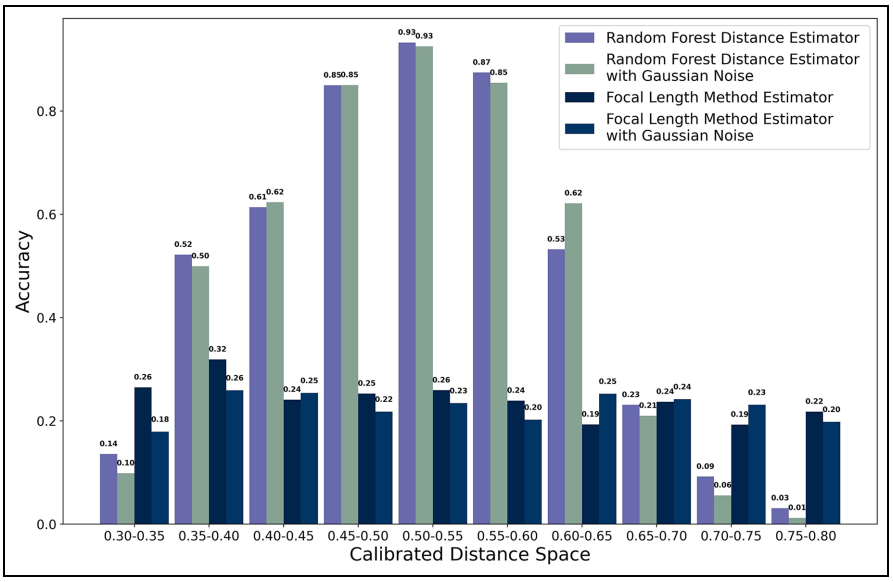

Figure 13. Test results of spatial accuracy using two different distance measurement model parameters.

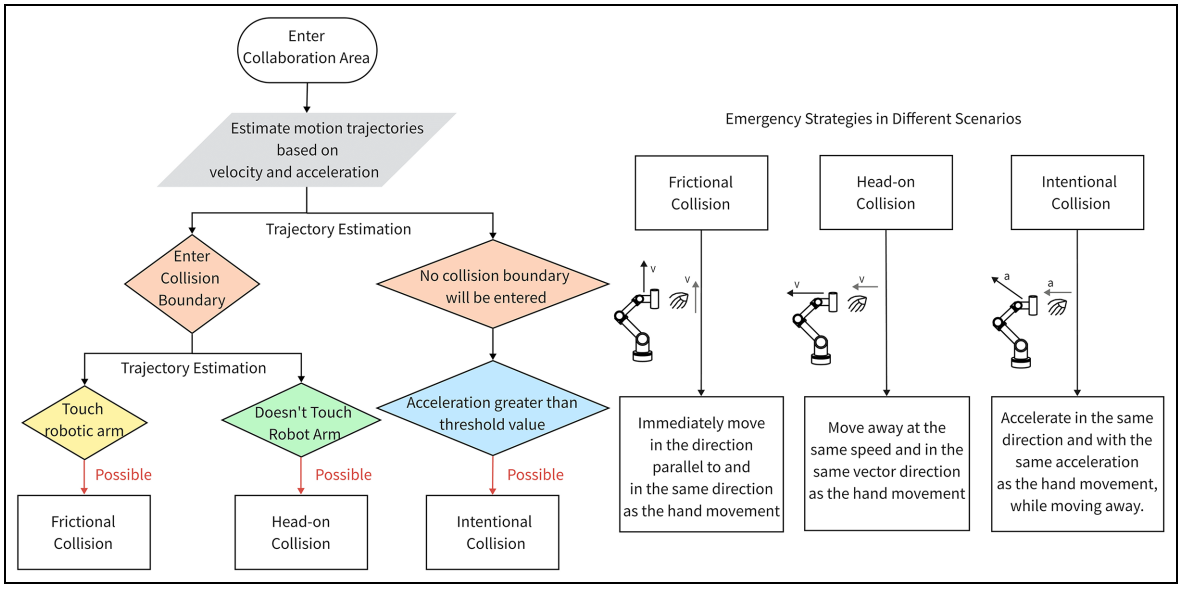

Figure 14. Collision classification decision tree and response strategies.

the final distance estimation. This implies that the random forest method can be adapted to different cameras without the need for prior camera parameter calibration.

- Focal Length Method Measurement: The distance estimation values obtained using the focal length method exhibited greater stability across all the intervals. This suggests that the focal length method consistently performs distance estimations within different distance ranges. However, the focal length method performed poorly in distance estimation for hand movements under different motion states, indicating lower generalization capabilities. This implies that the focal length method has limitations in predicting distances for larger motion amplitudes or hands in different postures.

In conclusion, we consider the random forest distance estimator advantageous in practical applications. It can achieve stable distance estimation within the range of 0.45–0.60 m and exhibits stronger generalization capabilities without the need for prior camera parameter calibration. In contrast, although the focal length method demonstrates stability across different intervals, its performance in distance estimation under different motion states is poor and its generalization capabilities are lower. Therefore, the random forest distance estimator is more suitable for practical applications because it can better adapt to various cameras without the need for prior calibration.

Assess HRC framework

To demonstrate the ability of the framework, we chose the collision scenario to verify its security. It can be roughly divided into three types, as shown in the right half of Figure 14. For the first type, we focus on the frictional collision between the hand and cobot during parallel motion, which is common in close collaboration environments. One of the most striking features of this scenario is that the movement of the hand and robot is parallel, but there is friction due to the difference in speed. For the second type, we pay special attention to contact collisions, which occur in the form of nearly head-on collisions during motion. The most important feature of the contact collisions is their positive impact. To minimize the magnitude of such positive impacts, the HRC system determines whether the hand has entered the previously calibrated danger zone after obtaining the hand depth information, and if so, maintaining a cooperative distance becomes the highest-priority task of the current robot. Finally, we focus on the case in which the hand actively collides with the robot, as this situation may involve operational errors or intentional human intervention. Such collisions are caused by the sudden physiological movement of the collaborator, and the collaborator's hand has a certain acceleration at which time the danger zone of the HRC system expands in all directions according to the acceleration, and the cooperation distance is always safe. In actual interactions, when distance and data flow related to collaborative gestures are input and tested, the real-time predefined iterative system safety strategy, the virtual environment receives input with time series, making it sensitive to the speed and acceleration changes of the hand, and can provide more personalized contingency plans.

The Box Collider component in Unity3d used for calibrating the dynamic safety distance between the robotic arm's motion and the human hand. We created

predefined safe distance spatial range colliders to simulate whether the collaborative distance is smaller than the predetermined safety standard.

After determining the dynamic safety space between the robotic arm's motion and the human hand, we further enhance the classification and decision-making for collision situations to validate the robot's correct response capability. Due to the limited field of view of the camera, the collaboration decision-making system is only activated when the human enters a certain area. As shown in the left half of Figure 14, we designed a collision classification decision tree to differentiate between three different collision scenarios. In the right half, we provide corresponding response strategies for each of the three scenarios.

Based on the criteria and response strategies defined in this decision tree, we conducted a statistical analysis of the success rate of the safety strategy responses for the three scenarios. We performed 100 test cases for each scenario at different speeds; the statistical test results are shown in Figure 15.

Based on the data presented in Figure 15, we can observe a high success rate of the safety strategy in virtual environment testing. This demonstrates that the implemented safety strategies performed well in a virtual environment. Based on the methodology outlined above, additional safety scenarios can be defined to facilitate iterative testing and ensure adaptability to different industrial collaboration situations.

For the failed tests that could not be responded to in a timely manner, we found that it may be due to fluctuations in the random forest distance estimator or frame rate variations in Unity3d. To address this issue, we employed a linear interpolation algorithm to stabilize the hand coordinates, which resulted in improved stability in the virtual environment. However, this approach introduces a time delay, which reduces the ability to detect rapid hand movements. Considering the cost and need for rapid deployment, this solution offers a certain level of usability.

Overall, the statistical analysis of the performance of the safety strategy and the implementation of stability-enhancing algorithms provide insights into the effectiveness and limitations of the system in a virtual environment. These findings contribute to the continuous refinement and optimization of safety measures, enabling better adaptability to various industrial collaboration scenarios.

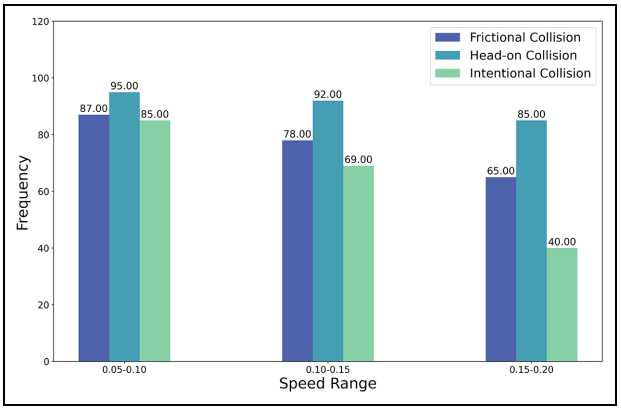

Figure 15. The number of correct responses in three different scenaiors.

Conclusion

The HRC framework proposed in this paper provides an economically feasible and convenient solution for deployment in collaborative tasks that do not require extreme precision. We first integrated the random forest distance estimator in the gesture component, avoiding complex pre-settings and the need for high-cost sensors, and did not even need to calibrate the RGB camera, which significantly improves the efficiency and adaptability of HRC system development and deployment flexibility, making it a user-friendly and cost-effective method for a variety of environments and collaboration scenarios. Second, in addition to gesture estimation, we provide two modules of virtual and real scenes to form an HRC-oriented digital twin system, which can effectively debug and monitor human-machine collaboration, speed up system development, deployment, and real-time monitoring, and can effectively improve the HRC system security.

In future works, we will further increase the size of the depth gesture training data set, improve the stability and accuracy of gesture depth estimation, and attempt to apply the HRC-oriented digital twin system to actual industrial scenarios for extensive testing.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Ao Liu https://orcid.org/0009-0008-9138-0059
Yuan Yao https://orcid.org/0000-0003-3302-4909

References

1. Li S, Wang R, Zheng P, et al. Towards proactive human–robot collaboration: A foreseeable cognitive manufacturing paradigm. *J Manuf Syst* 2021; 60: 547–552.
2. Meziane R, Otis MJD and Ezzaidi H. Human-robot collaboration while sharing production activities in dynamic environment: Spader system. *Robot Comput Integr Manuf* 2017; 48: 243–253.

3. Rozo L, Silverio J, Calinon S, et al. Learning controllers for reactive and proactive behaviors in human–robot collaboration. *Front Robot AI* 2016; 3: 30.
4. Nayyar A and Kumar A. *A Roadmap to Industry 4.0: Smart Production, Sharp Business and Sustainable Development*. Springer, 2020. [AQ: 3]
5. El Zaatari S, Marei M, Li W, et al. Cobot programming for collaborative industrial tasks: An overview. *Rob Auton Syst* 2019; 116: 162–180.
6. Sluÿters A, Lambot S and Vanderdonckt J. Hand gesture recognition for an off-the-shelf radar by electromagnetic modeling and inversion. In: *27th International conference on intelligent user interfaces*. pp. 506–522. [AQ: 4]
7. Nishiyama M and Watanabe K. Wearable sensing glove with embedded hetero-core fiber-optic nerves for unconstrained hand motion capture. *IEEE Trans Instrum Meas* 2009; 58(12): 3995–4000.
8. Gu X, Zhang Y, Sun W, et al. Dexmo: An inexpensive and lightweight mechanical exoskeleton for motion capture and force feedback in vr. In: *Proceedings of the 2016 CHI conference on human factors in computing systems*. pp. 1991–1995. [AQ: 5]
9. Oudah M, Al-Naji A and Chahl J. Hand gesture recognition based on computer vision: a review of techniques. *J Imaging* 2020; 6(8): 73.
10. Mainprice J and Berenson D. Human-robot collaborative manipulation planning using early prediction of human motion. In: *2013 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, pp. 299–306. [AQ: 6]
11. Shin SH and Lucyszyn S. Benchmarking a commercial (sub-) thz focal plane array against a custom-built millimeter-wave single-pixel camera. *IEEE Access* 2020; 8: 191174–191190.
12. Miller JL. *Principles of infrared technology*. Springer, 1994. [AQ: 7]
13. Matheson E, Minto R, Zampieri EG, et al. Human–robot collaboration in manufacturing applications: a review. *Robotics* 2019; 8(4): 100.
14. Vysocky A and Novak P. Human-robot collaboration in industry. *MM Sci J* 2016; 9(2): 903–906.
15. Ayvaz S and Alpay K. Predictive maintenance system for production lines in manufacturing: a machine learning approach using iot data in real-time. *Expert Syst Appl* 2021; 173: 114598.
16. Evjemo LD, Gjerstad T, Grøtli EI, et al. Trends in smart manufacturing: role of humans and industrial robots in smart factories. *Curr Robot Rep* 2020; 1: 35–41.
17. Su H, Ovur SE, Zhou X, et al. Depth vision guided hand gesture recognition using electromyographic signals. *Adv Robot* 2020; 34(15): 985–997.
18. Ding IJ and Su JL. Designs of human–robot interaction using depth sensor-based hand gesture communication for smart material-handling robot operations. *J Eng Manuf* 2023; 237(3): 392–413.
19. Che Y and Qi Y. Embedding gesture prior to joint shape optimization based real-time 3d hand tracking. *IEEE Access* 2020; 8: 34204–34214.
20. Li Q and Yang P. Keep up with me: A gesture guided moving robot with microsoft kinect. In: *2013 IEEE 10th international conference on mobile ad-hoc and sensor systems*. IEEE, pp. 435–436. [AQ: 8]
21. Rautiainen S, Pantano M, Traganos K, et al. Multimodal interface for human–robot collaboration. *Machines* 2022; 10(10): 957.
22. Kianoush S, Savazzi S, Beschi M, et al. A multisensory edge-cloud platform for opportunistic radio sensing in cobot environments. *IEEE Internet Things J* 2020; 8(2): 1154–1168.
23. Liu B, Fu W, Wang W, et al. Cobot motion planning algorithm for ensuring human safety based on behavioral dynamics. *Sensors* 2022; 22(12): 4376.
24. Liu W, Liang X and Zheng M. Task-constrained motion planning considering uncertainty-informed human motion prediction for human–robot collaborative disassembly. *IEEE/ASME Trans Mechatron* 2023. [AQ: 9]
25. Trinh M, Kötter D, Chu A, et al. Safe and flexible planning of collaborative assembly processes using behavior trees and computer vision. *Intell Hum Syst Integr (IHSI)* 2023; 69: 869–879.
26. Jantos TG, Hamdad MA and Granig W et al. Poet: Pose estimation transformer for single-view, multi-object 6d pose estimation. In: *Conference on robot learning*. PMLR, pp. 1060–1070. [AQ: 10]
27. Agnusdei GP, Elia V and Gnoni MG. A classification proposal of digital twin applications in the safety domain. *Comput Ind Eng* 2021; 154: 107137.
28. Douthwaite JA, Lesage B, Gleirscher M, et al. A modular digital twinning framework for safety assurance of collaborative robotics. *Front Robot AI* 2021; 8: 758099.
29. Ye Z, Jingyu L and Hongwei Y. A digital twin-based human-robot collaborative system for the assembly of complex-shaped architectures. *J Eng Manuf* 2022. [AQ: 11]
30. Kuts V, Marvel JA, Aksu M, et al. Digital twin as industrial robots manipulation validation tool. *Robotics* 2022; 11(5): 113.
31. Oyekan JO, Hutabarat W, Tiwari A, et al. The effectiveness of virtual environments in developing collaborative strategies between industrial robots and humans. *Robot Comput Integr Manuf* 2019; 55: 41–54.
32. Du G, Han R, Yao G, et al. A gesture-and speech-guided robot teleoperation method based on mobile interaction with unrestricted force feedback. *IEEE/ASME Trans Mechatron* 2021; 27(1): 360–371.
33. Che Y and Qi Y. Fast hand-object interaction using gesture guide optimization. In *2020 IEEE conference on virtual reality and 3D user interfaces abstracts and workshops (VRW)*. IEEE, pp. 792–793. [AQ: 12]
34. Antakli A, Spieldenner T, Rubinstein D, et al. Agent-based web supported simulation of human-robot collaboration. In: *WEBIST*, pp. 88–99. [AQ: 13]
35. Malik AA, Masood T and Bilberg A. Virtual reality in manufacturing: immersive and collaborative artificial-reality in design of human-robot workspace. *Int J Comput Integr Manuf* 2020; 33(1): 22–37.
36. Normalización OId. *ISO-TS 15066: Robots and Robotic Devices: Collaborative Robots*. ISO, 2016. [AQ: 14]
37. Winkelhaus S, Zhang M, Grosse EH, et al. Hybrid order picking: a simulation model of a joint manual and autonomous order picking system. *Comput Ind Eng* 2022; 167: 107981.
38. Chen X, Guo H, Wang G, et al. Motion feature augmented recurrent neural network for skeleton-based dynamic

hand gesture recognition. In *2017 IEEE international conference on image processing (ICIP)*. IEEE, pp. 2881–2885. [AQ: 15]

39. Zhang X, Li Q and Mo H et al. End-to-end hand mesh recovery from a monocular rgb image. In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2354–2364. [AQ: 16]
40. Prakash KB, Eluri RK and Naidu NB et al. Accurate hand gesture recognition using cnn and rnn approaches. *Int J Adv Trends Comp Sci Eng* 2020; 9(3): 3216–3222.
41. Zimmermann C and Brox T. Learning to estimate 3d hand pose from single rgb images. In: *Proceedings of the IEEE international conference on computer vision*. pp. 4903–4911. [AQ: 17]