Zheng, P., Li, S., Xia, L., Wang, L., & Nassehi, A. (2022). A visual reasoning-based approach for mutual-cognitive human-robot collaboration. *CIRP Annals*, *71*(1), 377-380. https://doi.org/10.1016/j.cirp.2022.04.016

Publisher's PDF, also known as Version of record

License (if available):
CC BY

Link to published version (if available):
10.1016/j.cirp.2022.04.016

Link to publication record in Explore Bristol Research
PDF-document

This is the final published version of the article (version of record). It first appeared online via Elsevier at https://doi.org/10.1016/j.cirp.2022.04.016 .Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research

General rights

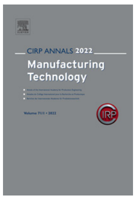
This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available: http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/

CIRP Annals - Manufacturing Technology 71 (2022) 377–380

Contents lists available at ScienceDirect

CIRP Annals - Manufacturing Technology

journal homepage: https://www.editorialmanager.com/CIRP/default.aspx



A visual reasoning-based approach for mutual-cognitive human-robot collaboration

Pai Zheng[a], Shufei Li[a], Liqiao Xia[a], Lihui Wang (1)[b,*], Aydin Nassehi (1)[c]

[a] Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hong Kong Special Administrative Region, China
[b] Department of Production Engineering, KTH Royal Institute of Technology, Sweden
[c] Department of Mechanical Engineering, University of Bristol, UK

ARTICLE INFO

Article history:
Available online 26 April 2022

Keywords:
Human robot collaboration
Manufacturing system
Visual reasoning

ABSTRACT

Human-robot collaboration (HRC) allows seamless communication and collaboration between humans and robots to fulfil flexible manufacturing tasks in a shared workspace. Nevertheless, existing HRC systems lack an efficient integration of robotic and human cognitions. Empowered by advanced cognitive computing, this paper proposes a visual reasoning-based approach for mutual-cognitive HRC. Firstly, a domain-specific HRC knowledge graph is established. Next, the holistic manufacturing scene is perceived by visual sensors as a temporal graph. Then, a collaborative mode with similar instructions can be inferred by graph embedding. Lastly, mutual-cognitive decisions are immersed into the Augmented Reality execution loop for intuitive HRC support.

© 2022 The Author(s). Published by Elsevier Ltd on behalf of CIRP. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

1. Introduction

In modern factories, personalized production of many complicated mechanical products relies on both robots' precision manipulation and human operators' agile operations. In this context, human-robot collaboration (HRC) has attracted much interest from the industry and academia, which leverages humans' high flexibility and robots' high efficiency and reliability [1]. Human and robotic agents have complementary operation goals and capabilities, and collaboratively conduct manufacturing tasks in a shared workspace. To date, numerous research efforts on HRC solutions have emerged, such as human safety [2], accurate robot control [3], multimodal communication [4], task allocation [5,6], and context awareness [7,8], of which the cognitive capability [9] is critical to achieve flexible automation in the dynamic manufacturing process.

However, the mutual-cognitive capabilities in HRC systems remain unsolved, especially when facing similar but new subtasks in real cases. Three major limitations exist: 1) Context-aware capabilities of existing HRC systems focus on the perception of the working environment, rather than a human-like understanding of the task process. 2) Current HRC systems directly transmit perceptual results into reactive control, and seldom consider knowledge learning of operation rules for proactive path planning and intuitive support. 3) The planner for robot execution and human operations is normally predefined, lacking on-the-fly adjustment capabilities during task fulfilment progress.

To address these issues, it is assumed that a mutual-cognitive HRC should embrace human-like intelligence and capabilities, which percept holistic surrounding scenes, recognize suitable collaborative modes, assign reconfigurable operation sequences, and transmit intuitive instructions to proactively support long-duration co-work between humans and robots. Hence, the main characteristics of mutual-cognitive HRC include 1) proactive operations of human and robotic agents, desired by partners and required by the actual situation in task progressing; 2) collaborative cognition derived from the holistic scene understanding of human, robot and environment, and knowledge learning of manufacturing information; 3) spatio-temporal subtasks fulfilment conducted in a shared workspace, under the same goal; and 4) mixed reality-enabled execution loop, where digital twins of the entire scene, along with their mutual cognitions are immersed into the augmented reality (AR) execution loop for human operation support and robot planning feedback [10].

2. Visual reasoning-based mutual-cognitive HRC system

To achieve it, the overall architecture of the proposed visual reasoning-based mutual-cognitive HRC system is shown in Fig. 1.

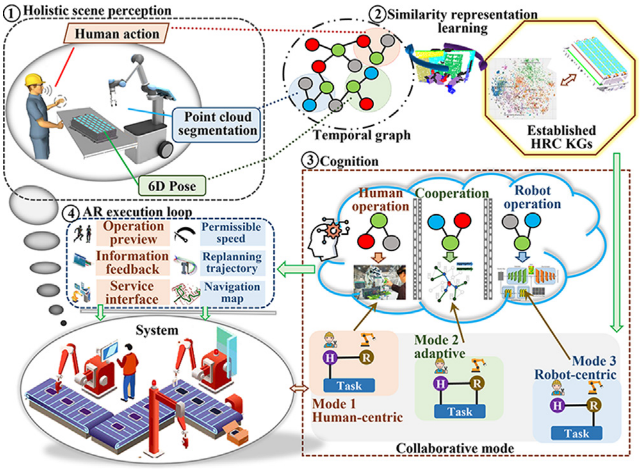


Fig. 1. Visual reasoning-based mutual-cognitive HRC system.

[*] Corresponding author.
E-mail address: lihui.wang@iip.kth.se (L. Wang).

https://doi.org/10.1016/j.cirp.2022.04.016
0007-8506/© 2022 The Author(s). Published by Elsevier Ltd on behalf of CIRP. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

Firstly, along the manufacturing task progressing, visual sensors sample live video streams and depth images of the entire workspace for holistic scene perception [11], including (1) human intention recognition, (2) object 6 Degree of Freedom (DoF) estimation, and (3) robot perception of the working environment. The perceptual results are then constructed as a temporal graph, which depicts relations of perceived objects, such as ‘human hand’ holding a ‘screwdriver’, ‘robot’ picking up a ‘toolbox’, etc.

Followed by the representation learning step, which captures the intrinsic similarity of perceptual results with previously seen manufacturing scenarios in the domain-specific HRC knowledge graphs (KGs). Meanwhile, cognitive intelligence is generated by injecting dynamism [12] into the knowledge representation of task allocation and updating sequential operation planning for humans and robots. Hence, a specific collaborative mode between a human and a robot can be inferred by graph embedding calculations based on extracted similarity of a new task, including:

(1) Human-centric operation represents that a human plays an active role with robot support, where the KG can search linked entities as instructions and guidance for human support, and the robot conducts assistive operations. With the information guide, a human can take mental-free operations assisted by the robot.
(2) Adaptive operation requires both the human and robot are actively engaged in a bidirectional manner, whereas KG ensures the dynamism of the graph sequence to evolve its link prediction. The linked entities are mutual planners which generate robot adaptive execution and visualize human required operations. The relationship between them is closer with mutual understanding.
(3) Robot-centric operation refers to the robot active, human supportive manner, where the most repetitive workload is carried out by the robot. The KG constructs production strategies as cognitive support for robotic operations in tasks, whereas human operators may act under supervisory control and plan for robots. The performance of robots is improved by rationale-based motion planning enabled by the human and task cognitions.

Lastly, the generated mutual cognition is deployed in the AR execution loop as a strategy assignment [13]. The co-work strategy contains on-demand support for human operators, such as the preview of complicated operations and information feedback. Meanwhile, considering the permissible speed, the robot arm and base are respectively controlled by the re-planned trajectory and a navigation map from the strategy for safe co-work with humans. Hence, both humans and robots can receive cognitive support and services on the fly in the execution loop as elaborated in Section 3.

3. The proposed stepwise visual reasoning approach

3.1. HRC knowledge graph construction

To describe the HRC process in a hierarchical and systematic manner, a domain-specific HRC KG should be established, based on accumulated expertise in HRC task allocation and planning, which can be divided into 5 node types with 3 layers, as shown in Fig. 2.

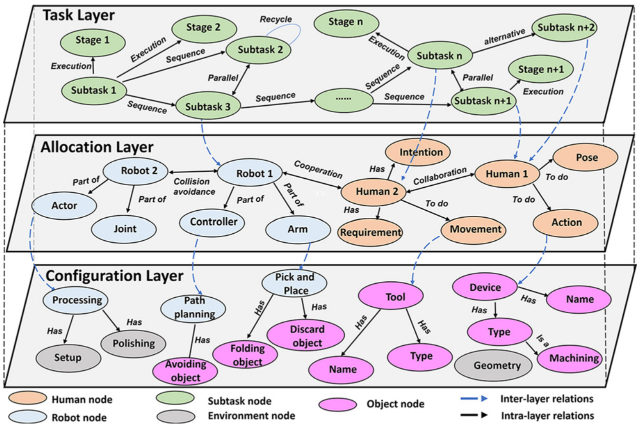


Fig. 2. The proposed HRC KG schema.

The HRC task layer triggers the location of new tasks and describes the sequential order, the complementary relationship among different subtasks, and corresponding executive stages. It can provide a sequential-based task-oriented configuration search to avoid any unsafe issues. Meanwhile, the allocation layer plays a critical role to associate various HRC subtasks with humans and robots. It indicates the involved humans and robots in subtasks through the inter-layer relations, and also introduces various relationships between them (e.g., human-human, robot-robot, and human-robot interactions). The action and movement elicitation of humans and robots are elaborated by the relationship (edge) ‘part of’, ‘has’, ‘to do’, etc. Moreover, with the extension of robots’ structure and the description of human action [14], the configuration layer offers comprehensive solutions to the humans and robots for similar task fulfilment. Hence, HRC subtasks can be allocated to various humans/robots, effectively.

3.2. HRC holistic scene perception

Human intention recognition. In HRC tasks, humans’ actions represent their operation intentions, i.e., “what the worker is doing”. Based on our previous work [7], a transfer learning-based Spatial-Temporal Graph Convolutional Network (ST-GCN) is proposed for action recognition in Fig. 3. It includes two parts: (1) pre-processing for human pose estimation, and (2) action recognition for skeleton patterns extraction and classification.

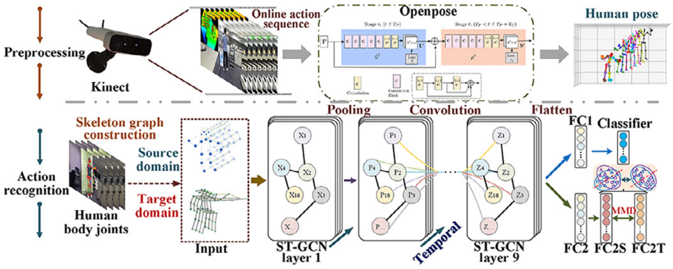


Fig. 3. Human action recognition by transfer learning-based ST-GCN.

For human pose estimation, Azure Kinect is utilised to capture live video streams from the workspace, which contain human operators’ action sequences. Then, with Openpose Toolbox, human skeleton joints of poses can be estimated from the videos.

For action recognition, the estimated human body joints are firstly constructed into skeleton graphs, including the natural connection of skeleton joints (spatial graph) and linking of the same joints across sequential frames (temporal graph). Followed by nine ST-GCN layers and one classifier, these stacked neural networks are used to extract patterns of human skeleton topology and classify pattern presentations. During the training process, the maximum mean discrepancy (MMD) is introduced to align action patterns between the source and the target datasets.

Object 6-DoF estimation. Precise 6-DoF pose estimation helps to recognise the 3D location and posture of target workpieces or tools, i.e., “where and what is the object”, even with partial occlusions in the HRC environment. As presented in Fig. 4, a modified high-resolution network (HRNet)-based approach [11] is proposed to estimate 6-DoF poses of industrial parts, even in close range HRC with partial occlusions. The high-resolution features of observed visual images are extracted, together with a mask-guided attention mechanism leveraged to model the occlusion area (e.g., hand).

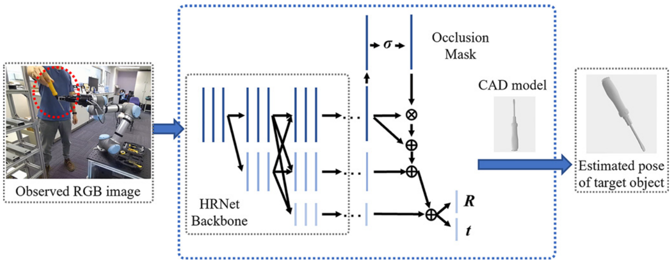


Fig. 4. Object 6-DoF estimation from occluded observations in HRC.

In this way, the extracted feature map is aware of the occlusion area. Then, the model predicts the translation parameters $\boldsymbol{t} = (t_x, t_y, t_z)^T$ and the rotation parameters $\boldsymbol{R} = (R_{roll}, R_{pitch}, R_{yaw})^T$. During the training process, the loss function constrains the predicted posed parameters, which are close to the ground truth ones. Hence, estimated workpiece posture can be obtained by applying the 6-DoF pose parameters to the 3D CAD model.

Robot perception of the working environment. Environment parsing equips the robot with the skills to perceive and model geometric interpretation of the entire workspace, where the 3D point cloud is considered as the 3D representation for robot perception, as shown in Fig. 5. Specifically, the captured visual and depth images (RGB-D data) are transformed into the point cloud with n × 3 shape, of which each point is represented by a 3D coordinate and n denotes the number of points. The PointNet model is then utilized to extract features of the point cloud. The original features of the 3D point cloud are encoded into vectors with a length of 1024 and decoded sequentially to output an n × m matrix, where m represents categories for each point. The category with the maximum confidence value is the correct classification result. By stacking a softmax operation to the output, 3D point cloud segmentation results are obtained with a label for each point.

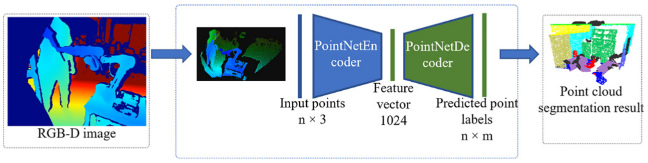


Fig. 5. 3D point cloud semantic segmentation of working environment.

3.3. Graph embedding-based HRC task planning

The information obtained from the holistic scene perception in Section 3.2 is isolated, which should be further aggregated through an Encoder model, as denoted below:

$$X_e = \sigma(W_h X_h + b_h) \tag{1}$$

where X_h = {H, O, R, E} is the concatenate of the holistic scene result. H represents the human intention, O is the object 6-DoF pose, R refers to the robot status, and E denotes the environment parsing. Meanwhile, X_e is the Encoder representation, and W_h and b_h denote the weight matrix and bias, respectively. Eq. (1) aims to transfer the perception result to node-level embedding. Based on it, a new subtask can be linked with an existing subtask (dashed line in Fig. 6), denoted as a Subtask node in the KG by calculating the most similar pair. Then, relevant edges in the existing KG can be predicted, as feasible configurations, to provide context-based instructions to humans and robots in the HRC process.

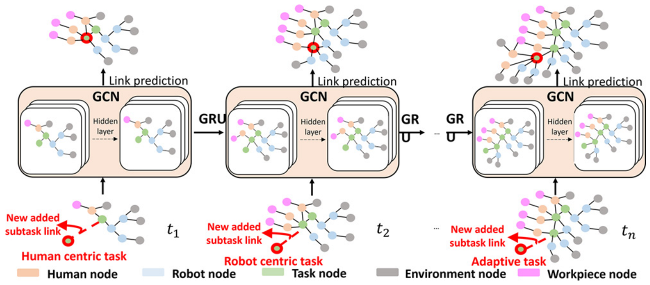


Fig. 6. The structure of EvolveGCN method for various HRC modes.

To achieve it, an EvolveGCN approach, combining the GCN with Gated Recurrent Unit (GRU) for link prediction in the dynamic changing human-robot roles in subtasks is introduced, as shown in Fig. 6. In timestamp t, the aggregation of GCN is iterated by:

$$H_t^{(l+1)} = \sigma\left(\hat{A}_t H_t^{(l)} W_t^{(l)}\right) \tag{2}$$

where $\hat{A}_t = \tilde{D}^{-\frac{1}{2}} \tilde{A}_t \tilde{D}^{-\frac{1}{2}}$ is the normalization of an adjacency matrix A_t, $\tilde{A}_t = A_t + I$, $\tilde{D} = diag\left(\sum_j \tilde{A}_{t\,ij}\right)$, I is an identity matrix, $\tilde{D}$ is the degree matrix, $W_t^{(l)}$ is the weight matrix in layer l, and $H_t^{(l)}$ is the node embedding matrix in layer l, where the initial embedding matrix comes from node features. Algorithm 1 is the pseudocode of the graph embedding-based dynamic HRC task planning.

Based on the different link prediction results in KG, the system can easily recognize the specific HRC mode, such as human-centric, adaptive, or robot-centric one, as shown in Fig. 6. For instance, the new subtask in t_1 is recognized as a human-centric operation, since most of the linked nodes are human nodes, where human operators take the leading role. Similarly, robot-centric operation (t_2) and adaptive operation (t_3) can also be discovered by the amount and types of linked nodes, respectively. Therefore, based on the specific mode, the proposed mutual-cognitive HRC system not only can provide essential instructions to humans and robots, but also regulate each party's permissions to avoid any conflicts (e.g., misleading orders by humans).

Algorithm 1. Pseudocode of graph embedding-based HRC task planning

```
Notations: ST_Num represents the total number of subtask nodes, where $ST\_E_n$
denotes the $n^{th}$ node; $l_{EG}$ is the layer number in EvolveGCN; $\psi_{ij}$ is the edge vector
between nodes i and j; MLP is the multilayer perception model.
Input: Holistic scene perception $X_h$, existing KG
Output: Subtask Instructions, updated KG
Methods: Subtask Encoding and EvolveGCN-based link prediction
$X_e$ = Encoder($X_h$)                                   //Encoder representation of $X_h$
Similar_Edge = $\max_{n \epsilon ST\_Num}(Similarity(ST\_E_n, X_e))$
KG ← KG_add_edge(Similar_Edge)                   //Add subtask as a node in KG
For $l \in l_{EG}$ do                                   //Update graph-embedding
    $W_t^{(l)} = GRU\left(H_t^{(l-1)}, W_{t-1}^{(l-1)}\right)$             //Weight matrix at timestamp t
    $H_t^{(l)} = GCN\_layer\left(H_t^{(l-1)}, W_t^{(l)}\right)$          //Embedding at timestamp t
End for
$P = MLP(\psi)$                                          //Calculate the probability of each edge in $\psi$
Predicted_edges = $Filter(P)$                          //Find the most relevant edges
KG ← KG_add_edge(Predicted_edges) //KG complement
Subtask Instructions = Parser(Predicted_edges)
        //Convert the predicted result into exercisable instructions
Return Subtask instructions, updated KG
```

4. System deployment

The mutual-cognitive HRC system is deployed in the Industrial Internet environment, including high-performance computing in the INDICS cloud, and Kinect, HoloLens, ROS, robot controller and GPU server in the edge plane, as shown in Fig. 7.

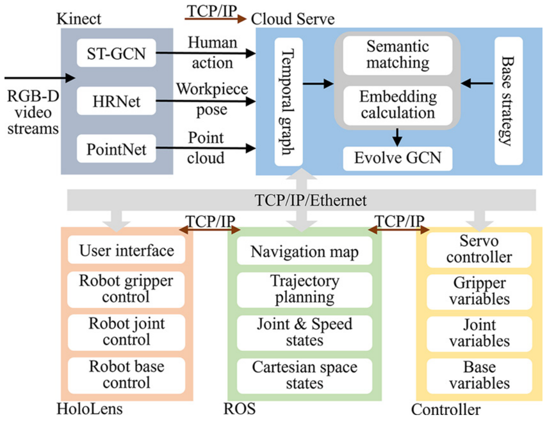


Fig. 7. The mutual-cognitive HRC system deployment structure.

RGB-D video streams are first input to the visual perception models, and then perceptual results of the holistic scene are transmitted to the cloud server via TCP/IP protocol. In the cloud plane, reasonable collaboration modes are first generated by semantic matching and embedding between the temporal graph and the established HRC KG. Then, detailed mutual-cognitive knowledge is learnt by EvolveGCN models, and the generated cognitions are transferred to HoloLens, ROS, and robot controller in the edge. The HoloLens device provides an intuitive user interface with robot control modules, to augment human operators' skills in HRC tasks. Meanwhile, these user commands are delivered to the ROS, where suitable navigation maps and robot trajectories can be generated with dynamics and kinetics library, while robot joint, speed, and the cartesian space states are monitored. Supported by ROS control commands and retrieved knowledge, the robot controller dynamically changes gripper, joint, and base variables in the servo controller for feasible execution.

5. Case study

Motivated by Wang [15], a case study on the quality checking of electronic vehicle batteries (EVBs) is conducted, to demonstrate the feasibility and effectiveness of the proposed HRC system in the lab environment, containing visual sensors (Azure Kinect and Intel D435), AR device (HoloLens 2), a GPU server (RTX 3080), and a mobile robot (UR5 and MiR100), as shown in Fig. 8.

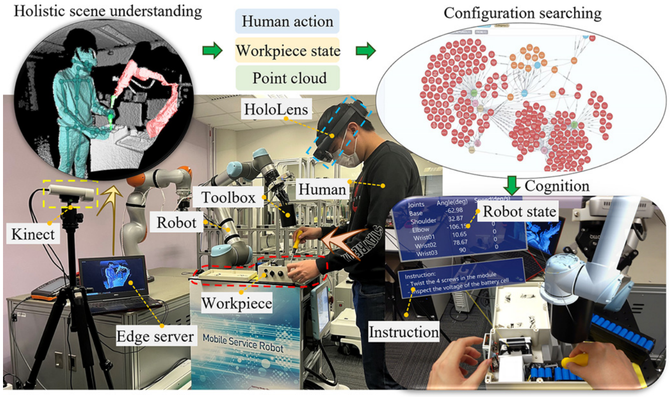
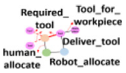
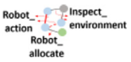
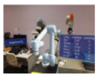
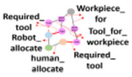


Fig. 8. Prototype system setup for EVBs quality checking.

The EVBs-based KG has been established, which consists of 341 entities (nodes) in 5 types, and 638 edges. For demonstration simplicity, the three typical visual reasoning based HRC operations are performed in querying similar subtasks to indicate their effectiveness for cognitive support, as listed in Table 1. For instance, by observing the scene of a screwdriver and battery shells, the HRC KG firstly links it to a most similar subtask node ‘unscrew’ (p_1=0.575). Based on the new subtask node, the ‘required tool’ edge corresponding to different target nodes can be predicted by the EvolveGCN model, representing a human-centric operation, and then infer knowledge of the unscrewing process of EVBs for mutual cognition. The result shows that ‘screwdriver’ is the most probable target node (p_2=0.753), denoting that operator requires a screwdriver in this subtask in the AR environment for intuitive unscrewing guidance. Similarly, in the robot-centric operation, the HRC KG infers suitable inspection manners and feedback control commands to the robot and provides a preview digital twin of the robot path planning in advance via AR devices, ensuring human safety in the collaboration. Lastly, for the adaptive HRC operation, robot path planning preview and human intuitive guidance are essential for effective collaboration, which dynamically generate interactive HRC task planning via AR for human and robotic agents.

Table 1
The preliminary experimental results of different HRC modes.

HRC mode	Top 3 similar subtasks	Probability (p_1)	Predicted link in most similar subtask	Target node	Probability (p_2)	AR scene
Human-centric operation	Unscrew Current test Screw	**0.575** 0.458 0.162	 Required_ tool Tool_for_ workpiece Deliver_tool human_ allocate Robot_allocate	Screwdriver Spanner Electric drill	**0.753** 0.679 0.597	
Robot-centric operation	Inspect Pick Handover	**0.555** 0.381 0.250		Snapshot Probing –	**0.938** 0.917 -	
Adaptive operation	Uninstall Package Install	**0.489** 0.324 0.142	 Required_ tool Robot_ allocate human_ allocate Workpiece_ for Tool_for_ workpiece Required_ tool	Gripper Sucker Screwdriver	**0.660** 0.612 0.592	

Meanwhile, the response time among various hardware devices and software systems is evaluated, as listed in Table 2. Among 10 random trials, the average communication interval between the AR glasses and the system with ROS is 35.068 ms via UDP protocol, resulting in a large deviation due to the shared WIFI channel in the lab environment. The average ROS-enabled path-planning costs 60.57 ms, where target positions of robot are uniformly sampled in the workspace. Meanwhile, the control latency from ROS to the UR5 robot in average is 2.921 ms, which is quite stable via ethernet cable connection. In summary, the maximum overall latency of the vision-based mutual cognitive HRC system is less than 0.4 s, which can be well adapted for onsite collaborative operations efficiently.

Table 2
Response time of each stage.

Response time (ms)	Holo Lens to ROS	ROS planning	ROS to controller	KG link prediction	Overall latency
Max	207	109	11.639	31.913	359.552
Min	4.64	49	1.659	24.934	55.299
Average	35.068	60.57	2.921	27.387	125.946

6. Conclusions and future work

To ensure a more efficient and intuitive HRC process, this research introduced a visual reasoning-based approach for establishing the mutual-cognitive HRC system, by considering the human action, workpiece 6D pose, working environment, and manufacturing task information integrally. It specified, realised and validated several aspects advancing the state of the art: (1) the system structure with elaborated steps for achieving mutual-cognitive HRC, (2) the holistic scene perception method, (3) the vision-based HRC KG querying for link prediction, and (4) the AR execution system deployment. Their feasibility and effectiveness were further demonstrated in the quality inspection process of an aging EVB module with some preliminary results. In the future, the mutual cognition for multiple HRC, and computational accuracy and robustness of the proposed system can be further improved with advanced AR techniques, more comprehensive HRC KG, and complex operations and manufacturing scenarios.

Declaration of Competing 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.

References

[1] Wang L, Gao R, Váncza J, Krüger J, Wang XV, Makris S, Chryssolouris G (2019) Symbiotic Human-robot Collaborative Assembly. *CIRP Annals* 68:701–726.

[2] Liu H, Wang L (2021) Collision-free Human-robot Collaboration Based on Context Awareness. *Robotics and Computer-Integrated Manufacturing* 67:101997.

[3] Wang XV, Wang L, Lei M, Zhao Y (2020) Closed-loop Augmented Reality Towards Accurate Human-Robot Collaboration. *CIRP Annals* 69:425–428.

[4] Liu S, Wang L, Wang XV (2020) Symbiotic Human-robot Collaboration: Multimodal Control Using Function Blocks. *Procedia CIRP* 93:1188–1193.

[5] Johannsmeier L, Haddadin S (2017) A Hierarchical Human-Robot Interaction-planning Framework For Task Allocation in Collaborative Industrial Assembly Processes. *IEEE Robotics and Automation Letters* 2:41–48.

[6] Buehler MC, Weisswange TH (2018) Online Inference of Human Belief for Cooperative Robots. In: *Proceedings of the IEEE International Workshop on Intelligent Robots and Systems (IROS)*, 409–415.

[7] Li S, Zheng P, Fan J, Wang L (2022) Towards Proactive Human Robot Collaborative Assembly: A Multimodal Transfer Learning-enabled Action Prediction Approach. *IEEE Transactions on Industrial Electronics* 69:8579–8588.

[8] Wang P, Liu H, Wang L, Gao RX (2018) Deep Learning-Based Human Motion Recognition for Predictive Context-aware Human-Robot Collaboration. *CIRP Annals* 67:17–20.

[9] Li S, Wang R, Zheng P, Wang L (2021) Towards Proactive Human−robot Collaboration: A Foreseeable Cognitive Manufacturing Paradigm. *Journal of Manufacturing Systems* 60:547–552.

[10] Jones D, Snider C, Nassehi A, Yon J, Hicks B (2020) Characterising The Digital Twin: A Systematic Literature Review. *CIRP Journal of Manufacturing Science and Technology* 29:36–52.

[11] Fan J, Zheng P, Li S (2022) Vision-based Holistic Scene Understanding Towards Proactive Human-robot Collaboration: A Survey. *Robotics and Computer-Integrated Manufacturing* 75:102304.

[12] Pareja A, Domeniconi G, Chen J, Ma T, Suzumura T, Kanezashi H, Kaler T, Schardl TB, Leiserson CE (2020) Evolvegcn: Evolving Graph Convolutional Networks for Dynamic Graphs. In: *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, 5363–5370.

[13] Li S, Zheng P, Zheng L (2021) An Ar-Assisted Deep Learning-Based Approach for Automatic Inspection of Aviation Connectors. *IEEE Transactions on Industrial Informatics* 17:1721–1731.

[14] Zhang J, Liu H, Chang Q, Wang L, Gao RX (2020) Recurrent Neural Network For Motion Trajectory Prediction in Human-robot Collaborative Assembly. *CIRP Annals* 69:9–12.

[15] Wang L (2021) A Futuristic Perspective on Human-Centric Assembly A Futuristic Perspective On Human-Centric Assembly. *Journal of Manufacturing Systems* 62:199–202.