ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Procedia CIRP 107 (2022) 943–948

www.elsevier.com/locate/procedia

55th CIRP Conference on Manufacturing Systems

Dynamic Scene Graph for Mutual-Cognition Generation in Proactive Human-Robot Collaboration

Shufei Li[a], Pai Zheng[a,*], Zuoxu Wang[b], Junming Fan[a], Lihui Wang[c]

[a] *Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hong Kong, China*
[b] *Department of Industrial and Manufacturing Systems Engineering, Beihang University, Beijing, China*
[c] *Department of Production Engineering, KTH Royal Institute of Technology, Stockholm, Sweden*

Abstract

Human-robot collaboration (HRC) plays a crucial role in agile, flexible, and human-centric manufacturing towards the mass personalization transition. Nevertheless, in today's HRC tasks, either humans or robots need to follow the partners' commands and instructions along collaborative activities progressing, instead of proactive, mutual engagement. The non-semantic perception of HRC scenarios impedes mutually needed, proactive planning and high-cognitive capabilities in existing HRC systems. To overcome the bottleneck, this research explores a dynamic scene graph-based method for mutual-cognition generation in Proactive HRC applications. Firstly, a spatial-attention object detector is utilized to dynamically perceive objects in industrial settings. Secondly, a linking prediction module is leveraged to construct HRC scene graphs. An attentional graph convolutional network (GCN) is utilized to capture relations between industrial parts, human operators, and robot operations and reason structural connections of human-robot collaborative processing as graph embedding, which links to mutual planners for human operation supports and robot proactive instructions. Lastly, the Proactive HRC implementation is demonstrated on disassembly tasks of aging electronic vehicle batteries (EVBs) and evaluate its mutual-cognition capabilities.

Keywords: Cognitive computing; scene graph; human-robot collaboration; human-centric manufacturing;

1. Introduction

Among the advanced manufacturing transition to Industry 5.0 [1] and reindustrialization, human operators play a central role in the production process [2]. For one side, the mass personalized production tendency raises ever-increasing flexible manufacturing requirements for modern factories, which remain unattained and rely on human manually agile operations. Besides, to achieve sustainability and resilience principles of Industry 5.0, the re-use, re-purpose and recycle processes of products demand high-level flexible and adaptable automation technologies. Motivated by this situation, human-robot collaboration (HRC) [3] has elicited particular interest in flexiable automation tasks, which combines human and robotic complementing competencies for improved productivity [4].

Ever since the HRC systems emerge, numerous research efforts have been investigated to place the best interplay of human skills and robot manipulation in industrial settings [5]. For instance, the perception of human operators' actions [6] and 6-DoF [7] poses of workpieces was explored, which provide a prerequisite for further adaptive robot decision-making and reasonable reactions. Based on the perceptual results, the HRC system which consists of accurate robot control [8] and human safety mechanism [9] was implemented on manufacturing tasks. Despite the above exploration, the context-aware capabilities among HRC systems are still stuck into the non-semantic perception of surrounding environments, rather than mutual-cognition of proactive guidance and path planning for human and robotic agents among the execution loop. The mutual-cognition remains a critical issue when implementing Proactive HRC systems [10], as few works discuss about how to advance the perceptual results to scene graphs containing knowledge of human-robot mutually co-work.

Aiming to fill this research gap, this paper proposes a dynamic scene graph approach for mutual-cognitive capabilities

[*] Corresponding author. *E-mail address:* pai.zheng@polyu.edu.hk

2212-8271 © 2022 The Authors. Published by Elsevier B.V.

10.1016/j.procir.2022.05.089

of Proactive HRC systems. First, a spatial-attention pyramid network [11] is leveraged to detect objects (including workpieces, robot arms, and human hands) from HRC environments as scene graph nodes. Then, the link prediction algorithm is leveraged to distill contextual representation between perceived nodes and their relations. Meanwhile, an attentional graph convolutional network (GCN) [12] is introduced to learn the representation of the scene graph embedding, which contains knowledge of HRC task operations and links to mutual-cognition for human support and robot instructions. Lastly, with the disassembly task of aging electronic vehicle batteries (EVBs), the proposed approach is evaluated for the Proactive HRC demonstration.

The rest of the paper is organized as follows. Section 2 introduces related works of Proactive HRC, object detection and scene graphs. The proposed methods which contain object detection for node generation, link prediction and scene graph embedding are depicted in Section 3. Experimental results of mutual-cognition generation and its application on the disassembly process of EVBs are described in Section 4. Conclusions and future works are summarized in Section 5.

2. Related work

This section gives a comprehensive review of the rising applications of Proactive HRC and related techniques including scenario perception and scene graph-based visual reasoning.

2.1. Proactive human-robot collaboration

HRC plays a crucial role in flexible and changeable manufacturing [13], such as producing new types of products and agile operation for complicated components. For example, in the production line of a paint factory, Automated Guided Vehicles (AGV) were introduced to pick and transport raw materials to mixing tanks, while a human operator published transaction orders and conducted dexterous manual operations [14]. With wearable devices [15] and real-time perception of the shopfloor, a mobile robot can anatomically navigate the trajector and act assisted operations to human operators [16].

To further enhance human-robot bidirectional teamwork skills, Proactive HRC was proposed for cognitive co-work between human operators and robots [10]. To enable robot decision-making in advance and proactive motion planning, a multimodal action prediction approach was proposed to estimate the human operator's intention with partly observation of video streams [6]. Besides, Pulikottil et. al [17] predicted human intention and generated robot path planning with knowledge of task temporal constraints. The cognitive co-work strategies in Proactive HRC allow robots to operate following time-varying position and orientation desired by humans, while humans manipulate consciously to decrease robot idle time.

2.2. Scenario perception via object detection

Scenario perception is the precondition for geometric and semantic knowledge interpretation of the human-robot and surrounding environment, which includes detection of static and dynamic objects and their spatial pose estimation [18]. In specfic, Rosenberger et. al utilized a deep Convolutional Neural Network (CNN) series model (YOLOv3) to detect industrial components from a heavily cluttered background [19]. For higher localization accuracy, Lee et. al introduced object segmentation methods to obtain fine shape information of industrial parts during the electric motor assembly process [20].

For spatial pose estimation, Franceschi et. al adopted Oriented Bounding Box (OBB) to obtain rough results and further leveraged the Iterative Closest Point (ICP) algorithm for refinement [21]. For surrounding environment construction, Moon et. al proposed to scene description generation by extracting local features from a 3D semantic graph map via GCN [22]. With the manufacturing task proceeding, the scenario perceptual results provide real-time information support for adaptive robot control and intuitive human operation.

2.3. Scene graph-based visual reasoning

Visual reasoning aims to learn context structure of perceived scene objects, which open the door to relationship understanding between human, robot and surrounding environment. In HRC systems, Ahn et. al [23] proposed a Text2Pickup network to allow for cognitive interaction based on workspace visual observations and human language commands. The visual reasoning network located desired objects in images and generated interactive questions for humans to clarify and correct their intention. Similarly, Venkatesh et. al [24] developed robot cognitive skills which reason about picking and placing coordinates of objects from language and image cues, uniting language features and visual features as semantic knowledge.

These above approaches inflexibly depend on secondary input of language questions to generate a reasonable answer as the interpretation of perceived scenes, which deviates from demands of direct cognition generation for Proactive HRC systems. The currently advanced scene graph methods facilitate efficient visual reasoning, beyond perception. For example, to reason explainable and explicit interpretation of scenes, Shi et. al [25] defined objects as nodes and pairwise relationships as edges in a scene graph with structured knowledge. These cutting-edge techniques directly extract scene context representations from visual images, which show the potential of generating mutual-cognition for human-robot co-work.

From the literature, one can find that existing research efforts on context awareness of HRC systems can well tackle perception issues of human-robot and surrounding workspaces, whereas cognitive capabilities lack efficient solutions, which impede the widespread applications of HRC. Motivated by this, a novel scene graph-based approach to deriving co-work cognition for implementation of Proactive HRC deserves exploration.

3. Methodology

In this paper, a dynamic scene graph approach is proposed to generate mutual-cognition for Proactive HRC. The dynamic

This is a resupply of March 2023 as the template used in the publication of the original article contained errors. The content of the article has remained unaffected.

scene graph contains the geometric, contextual interpretation of human-robot-workspace relationships among the manufacturing task process, i.e., ”what is human and robotic agents doing and how to do”. With three stepwise procedures, namely, object detection, linking prediction and graph embedding, the scene graph is dynamically constructed with task progressing. Then, the learned scene graph embedding triggering by temporal visual perceptual results is linked to the mutual-cognition, which is immersed into the execution loop of Proactive HRC systems to provide domain knowledge support for human operators and proactive robot motion planning among the co-work. As the HRC scene graph is up to both linking prediction and graph embedding, in this section, the detailed methods of object detector are introduced firstly, followed by the scene graph construction and embedding.

3.1. Spatial-attention object detector

A spatial-attention pyramid network is leveraged to locate coordinates of objects from images and classify their types. The objects in HRC scenarios consist of robot arms, human hands, and working-in-progress workpieces. The output of the network for each object is represented by a spatial location of the bounding box $v_i^o = [x_i, y_i, w_i, h_i]$ and a label of estimated categories $c_i^o \in \{1, \cdots, k\}$. The architecture of the spatial-attention pyramid network is shown in Fig. 1, which consists of feature extractor stem, feature pyramid net and output layer.

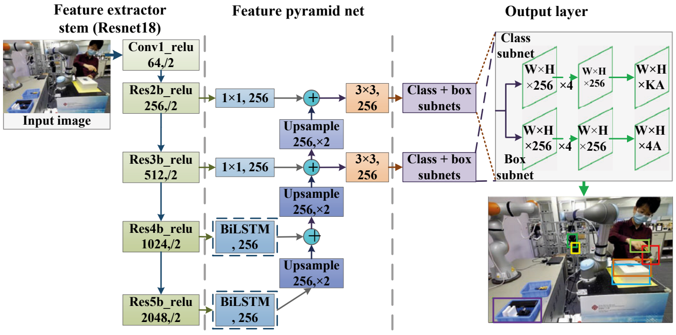

Fig. 1. Spatial-attention pyramid network for node generation of scene graphs.

For the feature extractor stem, Resnet18, a light CNN variant, is utilized to extract visual features of images, i.e., from low-level geometrical patterns to high-level aggregate representations. The initial images go through four operations of downsampling, of which the former two extracted feature maps are connected with a 1×1 convolutional filter and the other two are stacked to BiLSTM components. The two feature maps connected with BiLSTM operations can capture spatial relations of objects on images, as industrial parts normally connect with others in distribution. Then, in the feature pyramid net, these extracted feature maps undergo upsampling operations and concatenate values with an upper feature map. Finally, two 3×3 convolutional filters are stacked to the first upper two feature maps, which are utilized to generate output. In the output layer, the class subnet predicts the probability of object labels over K classes, while the box subnet estimates the spatial region of the object with four coordinates.

With the spatial-attention pyramid network estimating object locations and labels from images along the time, the scene graph dynamically updates its node attributes from the output. Given n object proposals, vectors of the scene graph node are denoted as the matrices $v^o \in \mathbb{R}^{n \times 4}$, and $c^o \in \mathbb{R}^{n \times k}$.

3.2. HRC scene graph construction and embedding

The scene graph for mutual-cognition generation among Proactive HRC includes three parts, 1) temporal node updating, 2) dynamic graph connection, and 3) graph embedding mapping, as presented in Fig. 2. Temporal node updating procedure searches for nodes appearing in the HRC scenarios along the time. With the above object detectors perceived results, nodes of detected objects are activated and update their attributes, whereas the other nodes remain inactive (i.e., the gray circle blocks). The dynamic graph connection aims for edge relationship pruning and scene graph construction. In detail, the perceptual objects are linked with edges to construct a dynamic scene graph, which embodies object relations and HRC setting knowledge. The graph embedding mapping denotes the explainable interpretation of scenes, which are used to link to reasonable robot commands and human operator reminders for mutual cognition. The interpretation is learned by distilling the entire scene graph representation, and paying specific attention to critical nodes and edges of subregions of the graph. To stepwise achieve this, two modules of link prediction and graph embedding are essential, which are described accordingly.

3.2.1. Link prediction for scene graph construction

To construct an accurate scene graph of HRC settings, two multi-layer perceptrons (MLPs) are leveraged to prune spurious connections and retain object pairs with the most relatedness. The link prediction for object nodes is capable of efficiently sparsifying node connections, which allows the HRC scene graph to pay attention to likely object relations.

In specific, the vectors of perceived object nodes z^o include predicted class distributions c^o and estimated spatial regions v^o. The relatedness of object pairs is learned by the node vector z^o, as classes c^o indicate likeliness interacting with neighbor nodes, while spatial regions v^o represent their connection types. Given object pairs $[z_i^o, z_j^o]$, the relatedness s_{ij} is computed as follows,

$$s_{ij} = f(z_i^o, z_j^o) = \left\langle \phi(z_i^o), \psi(z_j^o) \right\rangle, i \neq j \tag{1}$$

where $f(\cdot, \cdot)$ is a learned relatedness function, which can be calculated by kernel functions. As shown in the bottom left corner of Fig. 2, $\phi(\cdot)$ and $\psi(\cdot)$ are project functions for the *subject* and the *object* in object pairs, respectively. The projection processes for $\phi(\cdot)$ and $\psi(\cdot)$ are leveraged two MLPs with the same structure but different learning parameters. Then, matrix multiplication is applied for concatenation $\langle \phi(\cdot), \psi(\cdot) \rangle$ and to calculate the score matrix $S = \{s_{ij}\}^{n \times n}$. Followed by a sigmoid function, it is utilized to output the value of relatedness scores ranging from 0 to 1. Lastly, the relatedness scores are sorted in

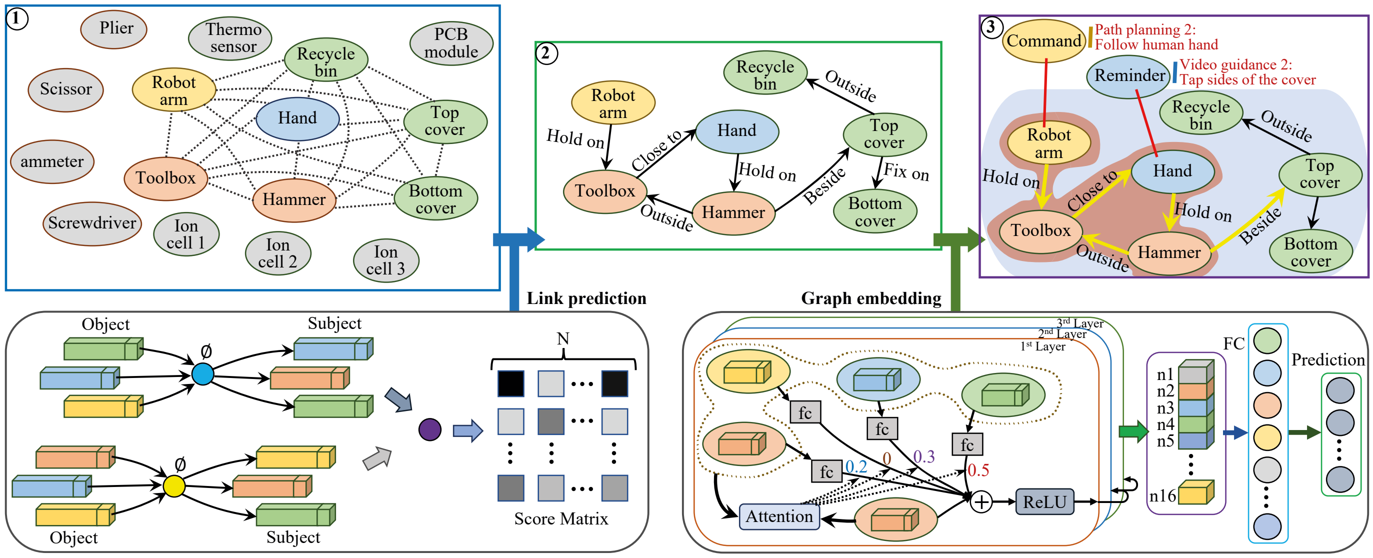

Fig. 2. HRC scene graph derived from link prediction and graph embedding.

descending order, and top C pairs are remained, of which object pairs whose spatial regions half overlapping with others are filtered out.

3.2.2. Graph embedding for mutual-cognition mapping

With the constructed scene graph, the next step is to learn the graph embedding which is utilized to map the mutual-cognition for Proactive HRC systems. The HRC scene graph contains numerous types of connections, i.e., from *subject* to *predicate*, from *predicate* to *object*, and from *subject* to *object*. GCN is introduced to infer contextual information of the scene graph structure, which consists of a linear transformation w, preset weights α, and a non-linear function (ReLU) σ. As presented in the bottom right corner of Fig. 2, with the information flow from node a to node b representing as $(\cdot)^{ab}$, the propagation of object features Z^o from layer l to layer $(l+1)$ in GCN is denoted as,

$$Z^o_{(l+1)} = \sigma(\overbrace{W^{so}Z^o_{(l)}\alpha^{so}}^{\text{Nodes}} + \overbrace{W^{sp}Z^p_{(l)}\alpha^{(sp)} + W^{op}Z^p_{(l)}\alpha^{(op)}}^{\text{Edges}}) \qquad (2)$$

where s=*subject*, p=*predicate*, and o=*object*. The first part in Eq. 2 is node representations from neighboring nodes, while the other one is relationship features from edges. Similarly, the representations of node connection are updated as,

$$Z^p_{(l+1)} = \sigma(\overbrace{Z^p_{(l)}}^{\text{Edges}} + \overbrace{W^{ps}Z^o_{(l)}\alpha^{(ps)} + W^{po}Z^o_{(l)}\alpha^{(po)}}^{\text{Nodes}}) \qquad (3)$$

In traditional GCN architecture, the weights α are predetermined in terms of the adjacency matrix of node features. To allow HRC scene graphs to capture key information of subregions, see the upper left corner of Fig. 2, an attentional mechanism is applied to adjust vectors α at each iteration. With a target node z_i and its neighboring nodes z_j, the attention is obtained by,

$$\begin{aligned} u_{ij} &= w_h^T\sigma(W_a[z_i, z_j]) \\ \alpha_i &= \text{Softmax}(u_i) \end{aligned} \qquad (4)$$

where w_h and W_a are parameters learning and updating from iteration results. In this way, weights α in Eq. 2 and Eq. 3 adjust after each iteration, as shown in the bottom right corner of Fig. 2. Three attentional GCN layers are stacked to learn the scene graph embedding, followed by a fully connected (FC) layer, which linear transforms extracted node features. Lastly, a Softmax function is utilized to classify the extracted features and to link the graph embedding to corresponding co-work cognition among HRC systems.

4. Case study and experiment results

To evaluate the significance of the proposed approach, the HRC scene graph is demonstrated in disassembly tasks of EVBs to generate mutual-cognition for human support and robot control among the co-work. In this context, experiment results of object detection, graph construction, and embedding in HRC scenarios are depicted in this section.

4.1. HRC scene graph for disassembly of EVBs

The disassembly task of EVBs depends on human-robot mutual operations. For one side, some flexible subtasks, such as cutting wires and removing glue, desires human manual operations. On the other hand, robot manipulation is necessary, especially for risk operations, such as picking-and-placing tasks of battery cells. As a demonstration of EVBs disassembly, the entire graph of human-robot co-work settings is shown in Fig. 3. The graph implying knowledge of disassembly processes consists of six different kinds of nodes, including *Tool*, *Workpiece*,

This is a resupply of March 2023 as the template used in the publication of the original article contained errors. The content of the article has remained unaffected.

Table 1. Accuracy of scene graph (SG) construction and embedding.

HRC SG	SG1	SG2	SG3	SG4	SG5	SG6	SG7	SG8	SG9	SG10	SG11
SGGen+	13/14	14/14	21/21	20/21	34/34	32/34	34/34	33/34	34/34	33/34	32/34
Precision	91.67	95.00	96.36	91.67	94.44	95.74	96.30	92.86	96.00	93.75	92.31

Human, *Reminder* for manual operation, *Robot* and its *Command*. According to the label and spatial location of nodes, there are eight relations connecting them. In specific, across the fulfillment of the disassembly task, 11 kinds of scene graphs can be constructed with various nodes and edges activated by the temporal visual input. These dynamic scene graphs link to mutual-cognition support for human reminders and robot commands, from *unscrewing*, *loose coupling*, *unfolding cover*, *electric test*, *cutting wires*, *removing glue*, *picking-and-placing of components* (including tools and workpieces) in HRC tasks.

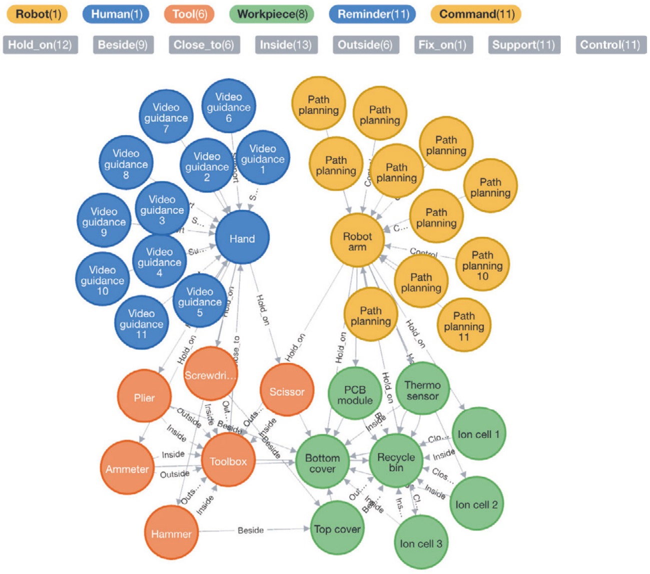

Fig. 3. HRC scene graph for disassembly of EVBs.

4.2. Object detection results in HRC scenarios

To map the graph embedding to mutual-cognition, accurate object detection in HRC scenarios is the prerequisite for node generation and dynamic scene graphs construction. An EVBs disassembly dataset which contains 779 RGB images is captured by Azure Kinect. The dataset is developed to demonstrate the HRC task fulfillment covering the disassembly process of EVBs in a laboratory environment. There are 14 different objects in the dataset, namely, Ion cell, PCB module, Thermo sensor, Top cover, Bottom cover, Recycle bin, Toolbox, Screwdriver, Scissor, Plier, Ammeter, Hammer, Hand, and Robot arm.

For experiment settings, 467 samples of the EVBs disassembly dataset are used in the training process, while the remaining 312 images are testing data. The object class K is set as 14. The spatial-attention pyramid network is employed on a Tesla V100 GPU (16G) for training the object detector, with the optimizer of the stochastic gradient descent (SGD). For objects in HRC scenes, the perceived results of the trained model are depicted in the left part of Fig. 4. With temporal visual input, the object detector can locate various objects and classify their categories for dynamic node generation.

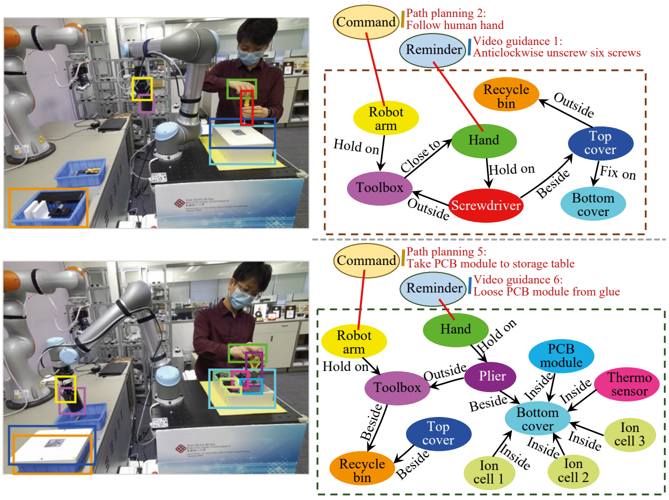

Fig. 4. Examples of object detection, graph construction and embedding.

4.3. Graph construction and embedding results

With locations and labels of above perceived object, the link prediction procedure learns their relations and connects edges among these nodes. During the training, the number of remained layers C is set to 128 in the two-layer MLPs. Besides, object pairs whose spatial regions overlap with others more than 0.7 are filtered. The attentional GCN is optimized by the SGD with a learning rate of 0.01. Along the disassembly process of EVBs, there are 11 scene graphs dynamically triggering with different visual inputs. These constructed scene graphs link to various co-work strategies for human and robotic agents in HRC systems, including domain knowledge reminders and control commands. With the trained models, the cognitive support of task strategies can be generated with around 185ms time consuming, which satisfy real-time requirements.

As demonstrated in the upper right corner of Fig. 4, the dynamic scene graph is generated for the unscrewing process of the top cover. The entire graph embedding of activated nodes and edges are learned and mapped to cognitive support in the HRC system, which guidance human unscrewing operations, while the robot obtains commands of holding the toolbox and following the human aside as assistant. For the bottom right corner of Fig. 4, the HRC scene graph is dynamically constructed after removing the top cover, which reminds the human operator to loose the PCB module from glue and control the robot to take the module to the storage table at the same time.

Except for the demonstration, the SGGen+ [12] metric is utilized to evaluate the performance of scene graph construc-

This is a resupply of March 2023 as the template used in the publication of the original article contained errors. The content of the article has remained unaffected.

tion, which calculates the number of nodes, edges, and their triplet accurately generated. The prediction precision of graph embedding which maps to mutual-cognition is also evaluated on 11 scene graphs. As presented in Table 1, the link prediction approach can correctly connect most of the relations between nodes. For the X/Y of the first row, X is the accuracy linked edges out of Y relations. Based on the constructed scene graphs, the performance of mapping processes from graph embedding to HRC mutual-cognition can achieve high accuracy.

5. Conclusion

This paper proposed a dynamic scene graph method for mutual-cognition generation in Proactive HRC systems. A spatial-attention pyramid network was utilized to detect objects arranging with spatial relations in HRC scenes. With MLPs-based link prediction, dynamic scene graphs are triggered by visual input along with the task fulfillment. The attentional GCN is leveraged to learn the graph embedding and map the representation to robot commands and human reminders, which feedback as mutual-cognition among HRC systems. This research explores the cognitive capabilities for HRC systems, from the explainable knowledge graph base.

Apart from the above advantages, the performance of SG construction and embedding can be further improved by data augmented methods. Besides, the dynamic scene graph approach can be deployed on more HRC tasks to generate cognitive co-work strategies. Meanwhile, potential future research directions are highlighted here, including 1) Mixed Reality-based deployment of the proactive HRC system for intuitive human support and proactive robot control, and 2) the mutual-cognition generation when facing similar but different HRC tasks. It is hoped that this work improves the cognitive capabilities of HRC systems in line with human-centric manufacturing.

Acknowledgements

This research work was partially supported by the grants from the National Natural Science Foundation of China (No. 52005424), and Research Committee of The Hong Kong Polytechnic University under Research Student Attachment Programme 2021/22 and Collaborative Departmental General Research Fund (G-UAMS) from the Hong Kong Polytechnic University, Hong Kong SAR, China.

References

[1] X. Xu, Y. Lu, B. Vogel-Heuser, L. Wang, Industry 4.0 and industry 5.0—inception, conception and perception, Journal of Manufacturing Systems 61 (2021) 530–535.

[2] L. Wang, A futuristic perspective on human-centric assembly, Journal of Manufacturing Systems 62 (2022) 199–201.

[3] L. Wang, R. Gao, J. Váncza, J. Krüger, X. V. Wang, S. Makris, G. Chryssolouris, Symbiotic human-robot collaborative assembly, CIRP annals 68 (2019) 701–726.

[4] S. Li, J. Fan, P. Zheng, L. Wang, Transfer learning-enabled action recognition for human-robot collaborative assembly, Procedia CIRP 104 (2021) 1795–1800.

[5] S. Liu, L. Wang, X. V. Wang, Symbiotic human-robot collaboration: multimodal control using function blocks, Procedia CIRP 93 (2020) 1188–1193.

[6] S. Li, P. Zheng, J. Fan, L. Wang, Towards proactive human robot collaborative assembly: A multimodal transfer learning-enabled action prediction approach, IEEE Transactions on Industrial Electronics (2021).

[7] J. Fan, S. Li, P. Zheng, C. K. Lee, A high-resolution network-based approach for 6d pose estimation of industrial parts, in: 2021 IEEE 17th International Conference on Automation Science and Engineering (CASE), IEEE, 2021, pp. 1452–1457.

[8] X. V. Wang, L. Wang, M. Lei, Y. Zhao, Closed-loop augmented reality towards accurate human-robot collaboration, CIRP annals 69 (2020) 425–428.

[9] H. Liu, L. Wang, Collision-free human-robot collaboration based on context awareness, Robotics and Computer-Integrated Manufacturing 67 (2021) 101997.

[10] S. Li, R. Wang, P. Zheng, L. Wang, Towards proactive human–robot collaboration: A foreseeable cognitive manufacturing paradigm, Journal of Manufacturing Systems 60 (2021) 547–552.

[11] S. Li, P. Zheng, L. Zheng, An ar-assisted deep learning-based approach for automatic inspection of aviation connectors, IEEE Transactions on Industrial Informatics 17 (2020) 1721–1731.

[12] J. Yang, J. Lu, S. Lee, D. Batra, D. Parikh, Graph r-cnn for scene graph generation, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 670–685.

[13] J. Krüger, T. K. Lien, A. Verl, Cooperation of human and machines in assembly lines, CIRP annals 58 (2009) 628–646.

[14] R. Rey, J. A. Cobano, M. Corzetto, L. Merino, P. Alvito, F. Caballero, A novel robot co-worker system for paint factories without the need of existing robotic infrastructure, Robotics and Computer-Integrated Manufacturing 70 (2021) 102122.

[15] G. Michalos, N. Kousi, P. Karagiannis, C. Gkournelos, K. Dimoulas, S. Koukas, K. Mparis, A. Papavasileiou, S. Makris, Seamless human robot collaborative assembly–an automotive case study, Mechatronics 55 (2018) 194–211.

[16] A. C. Bavelos, N. Kousi, C. Gkournelos, K. Lotsaris, S. Aivaliotis, G. Michalos, S. Makris, Enabling flexibility in manufacturing by integrating shopfloor and process perception for mobile robot workers, Applied Sciences 11 (2021) 3985.

[17] T. B. Pulikottil, S. Pellegrinelli, N. Pedrocchi, A software tool for human-robot shared-workspace collaboration with task precedence constraints, Robotics and Computer-Integrated Manufacturing 67 (2021) 102051.

[18] F. Junming, Z. Pai, L. Shufei, Vision-based holistic scene understanding towards proactive human-robot collaboration, Robotics and Computer-Integrated Manufacturing Accepted (2022).

[19] P. Rosenberger, A. Cosgun, R. Newbury, J. Kwan, V. Ortenzi, P. Corke, M. Grafinger, Object-independent human-to-robot handovers using real time robotic vision, IEEE Robotics and Automation Letters 6 (2020) 17–23.

[20] H. Lee, Y. Liau, S. Kim, K. Ryu, A framework for process model based human-robot collaboration system using augmented reality, in: IFIP International Conference on Advances in Production Management Systems, Springer, 2018, pp. 482–489.

[21] P. Franceschi, N. Castaman, S. Ghidoni, N. Pedrocchi, Precise robotic manipulation of bulky components, IEEE Access 8 (2020) 222476–222485.

[22] J. Moon, B. Lee, Scene understanding using natural language description based on 3d semantic graph map, Intelligent Service Robotics 11 (2018) 347–354.

[23] H. Ahn, S. Choi, N. Kim, G. Cha, S. Oh, Interactive text2pickup networks for natural language-based human–robot collaboration, IEEE Robotics and Automation Letters 3 (2018) 3308–3315.

[24] S. G. Venkatesh, A. Biswas, R. Upadrashta, V. Srinivasan, P. Talukdar, B. Amrutur, Spatial reasoning from natural language instructions for robot manipulation, arXiv preprint arXiv:2012.13693 (2020).

[25] J. Shi, H. Zhang, J. Li, Explainable and explicit visual reasoning over scene graphs, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 8376–8384.

This is a resupply of March 2023 as the template used in the publication of the original article contained errors. The content of the article has remained unaffected.