2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Enhanced Robotic Assistance for Human Activities through Human-Object Interaction Segment Prediction

Yuankai Wu[1], Rayene Messaoud[1], Arne-Christoph Hildebrandt[2], Marco Baldini[2], Driton Salihu[1], Constantin Patsch[1], Eckehard Steinbach[1]

***Abstract*— Robotic assistance is a current research topic with high application value and multiple challenges. Assistive robots are used in various scenarios, such as production lines, operating tables, and elderly care. While providing effective assistance, most of the assistance tasks that current robots can perform are limited to predefined tasks. This limitation arises from the insufficiency of the current robot perception system to forecast future human activities. To address this issue, we propose a novel 2-stage robotic assistant for human activities through future human-object interaction (HOI) segment prediction. Unlike previous work focusing on predefined or short-term tasks, our robotic assistant can make predictions for future assistance according to human habits. In the first stage, we propose a visual-based human-object interaction segment prediction method to predict human activities, which enables the robotic system to infer human intention. Moreover, we define the robotic executable tasks as an interactive tuple to keep the robotic assistance normatively consistent with human activity. Meanwhile, a graph convolutional network with geometric features that can predict human-object interaction segments is proposed to provide target manipulation and target object for the assistive robot. In the second stage, we present a mobile task completion process including visual navigation, object localization and grasping. The perception stage is evaluated on the MPHOI dataset and custom-collected SPHOI dataset. Finally, we evaluate our comprehensive framework through real-time experimentation.**

I. Introduction

Assistive robots play a significant role in the contemporary domain of robotics. It focuses on improving the interaction between humans and robots, rather than only traditional physical interaction [1], [2]. This brings additional challenges when applying daily life or industry scenarios as the research context[3], [4], [5]. In order to provide more comfortable assistance, the assistive robot needs to observe the current human activities and accomplish the desired human tasks for the future [6]. Furthermore, the input information of the robot changes from physical signals of human-robot contact [7], [8] to multi-dimensional signals collected by different sensors [9], [10]. In this paper, we only use an RGB-D camera as input instead of wearable sensors for human activity analysis, e.g., inertial measurement unit [11]. The main reason is to consider that wearable sensors are more uncomfortable for people in daily life scenarios [12]. Besides, the vast majority of current assistive robots have integrated vision sensors [13], [14]. This provides the prerequisite for a vision-based perception system to be used in daily life and industry scenarios.

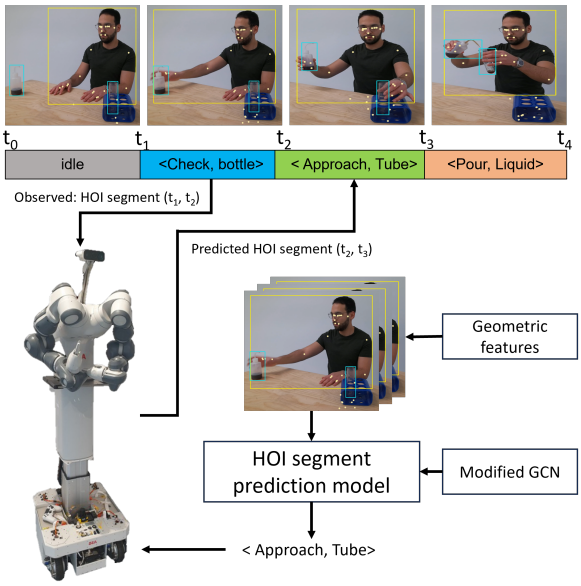

Fig. 1: Human-object interaction segment prediction for the mobile robotic assistant YuMi. The robot observes the visual information at anytime during the segment from t_1 to t_2 through the camera. And our HOI segment prediction model further predicts the next human-object interact segment from t_2 to t_3. The results obtained are sent back to the robot for performing the auxiliary tasks.

Previous works mostly focus on physical interaction. The traditional learning from demonstration approach plans the robotic execution path through the physical manipulation of the robot by a human [15], [16]. However, this approach is often limited to tasks that can be performed within a short period of time. Jain et al. [17] proposed to use the recursive Bayesian filtering approach to make predictions about human intention. However, this shared autonomy method requires human participation for initial remote control. In this work, we try to explore how robotic assistants can provide fully automated assistance without physical human interaction with the robot. Meanwhile, it should allow robot assistant to make future decisions, which is more compatible with the logical needs of humans. As shown in Figure 1, we propose

[1]Authors are at the Chair of Media Technology and Munich Institute of Robotics and Machine Intelligence, School of Computation, Information and Technology, Technical University of Munich, Germany. {yuankai.wu, rayene.messaoud, driton.salihu, constantin.patsch, eckehard.steinbach}@tum.de

[2]Authors are with the mobile robotics department at Asea Brown Boveri (ABB) Group. {arne-christoph.hildebrandt, baldini.marco}@de.abb.com

a novel human-object interaction segment prediction method to anticipate future human activity via current observation. This is mainly because if we only employ the action detection method [18], [19], the robot will perform the hard-coded reaction with the same activity as the current human action, which causes conflict during assistance. Moreover, the general applicability of the robot would be constrained if it relies on a predetermined approach for the subsequent tasks. To this end, we propose a two-stage structure with a perception stage and an execution stage. The first stage utilizes the modified graph neural network approach using geometric features for future human-object interaction segment prediction. The second stage constructs a task execution sequence that processes the output from the first stage. Furthermore, the complete robotic assistant pipeline is comprehensively tested using suitable real-world scenarios to ensure its efficacy. The summary of our contributions are as follows:

- We propose a novel 2-stage enhanced robotic assistance using HOI segment prediction on the mobile YuMi robot that can automatically provide human assistance.
- In the 2-stage assistance structure, we employ an enhanced graph convolutional network with geometric features for future HOI prediction using a contractive tuple. It is associated with a task execution pipeline including visual navigation, object localization and grasping by behavior tree.
- We conduct and evaluate our proposed HOI segment prediction method on the MPHOI dataset and our own collected SPHOI dataset. To the best of our knowledge, we are the first work to present HOI segment prediction to complete real-time testing in an industrial laboratory scenario.

II. Related Work

A. Robotic assistance

Robotic assistance aims to provide convenient assistance to humans in environments shared by humans and robots. Previous work by Hu et al. [20] proposed a robotic nursing assistant for patient care. The assistance can be provided through telepresence control and video communication by the nurse. However, this work does not involve human-object interaction. In later work, [21] proposed a robotic assistant that automatically assists a person in drinking water. [22], [23] proposed the use of MOCA robotic assistance to perform heavy or prolonged manipulation tasks. However, these tasks are relatively unitary and cannot satisfy the application scenarios under long task sequences. In the recent work, Massardi et al. [24] proposed a robotic assistant with human activity recognition and a predefined task plan. In this work, the robot assistant needs to recognize the current activity of the human and then return to the task list to find the next task that needs to be performed. Unlike previous work, we transform the coarse recognition of human activities into the prediction of human-object interactions. By predicting the segment of human-object interactions at the next moment, we effectively provide the robot directly with the actions to be performed and the target object to be manipulated.

B. Human-object interaction understanding

Human-object interaction (HOI) understanding is an important and evolving area of computer vision research focusing on understanding how humans interact with objects in a shared environment. Previous work by Dreher et al. [25] proposed a scene graph based method to learn human-object interactions using a graph neural network. However, the classification performance of HOIs using a scene graph generated by 3D locations in the environment is limited. Morais et al. [26] later proposed ASSIGN to detect human-object interactions in video. This work gives a better performance in the understanding of HOI by analyzing the association between human subjects and objects and object affordance. The recent work by Qiao et al. [27] proposed an HOI recognition approach using a graph convolutional network with geometric features. This work makes good use of the geometric information of humans and objects in 3D scenes and obtains a significant performance improvement. But these works take the human action and the object as a single unit, which does not allow the robot to accurately discriminate between the action and the object that needs to be performed. Inspired by [28], [29], [30], we separated the human action and object and composed a $< action, object >$ tuple to provide task input information to the robot assistant. Different from the above mentioned work, we do not perform only HOI detection. We further extend the network into the prediction of the next HOI segment for humans based on 2G-GCN [27]. And having the robot perform the prediction task is more in accordance with the auxiliary logic of long task sequences for human.

III. Two-stage mobile assistant

In order to address robotic assistance for humans, we propose a two-stage mobile assistant system. The first stage aims to predict the future human-object interaction segment. The mobile YuMi robot receives RGB-D information from the camera and further generates geometric features via object detection and human skeleton detection. The generated geometric features are then passed to the modified HOI prediction network. As a result, a tuple is produced and passed to the second stage. In the second stage, the tuple output from the previous stage acts as an input. The activity and object are then parsed and post-processed. After that, the mobile manipulation application is started. Furthermore, a behavioral tree is constructed to help guide the robot into fetching the object needed from the predicted interaction. It also uses the detection pipelines that are mentioned in the first stage for accurate navigation.

A. Human-object interaction segment predction

In this section, we discuss the first stage of the two-stage mobile assistant, which is centered around the domain of human activity understanding. We will start from extracting geometric features and further explain how 2G-GCN [27] can be modified and achieve the ability to make predictions about HOI segments.

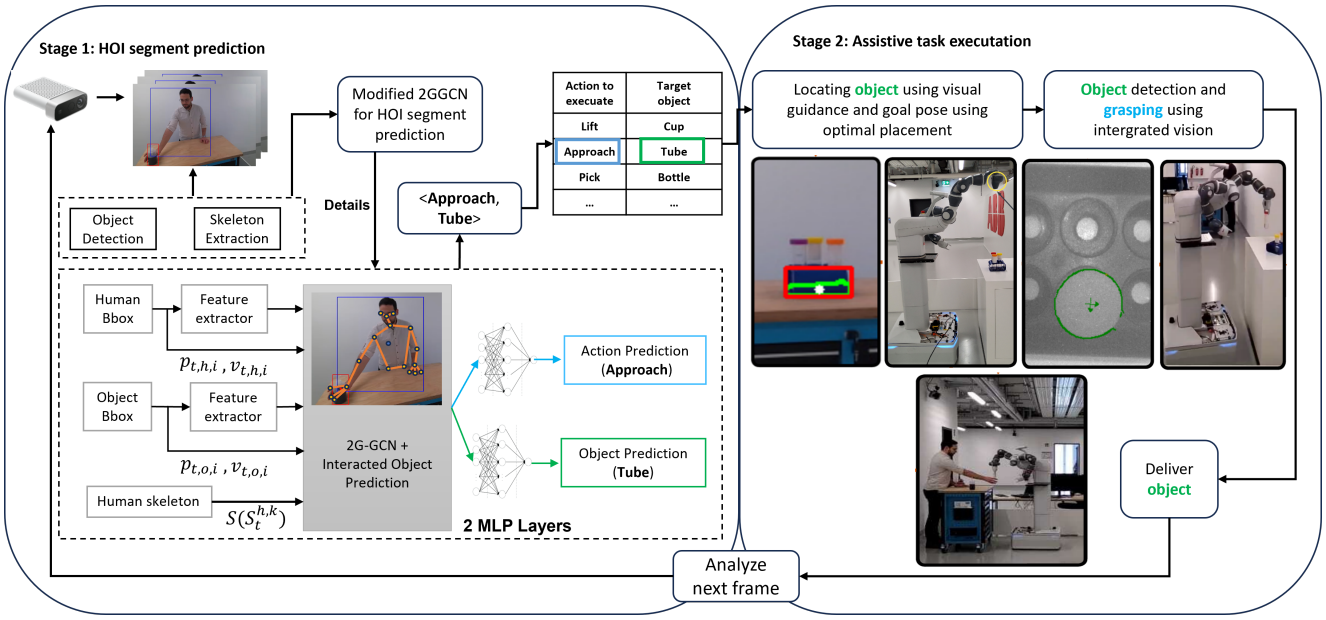

Fig. 2: Overview structure of the mobile robotic assistant using human-object interaction segment prediction. In Stage 1, the modified 2G-GCN network uses geometric information about the position, velocity, and skeleton of the human and object, as well as visual information extracted by the feature extractor, to predict the future desired action and object of the human. In stage 2, the mobile robot accomplish human assistance by executing predicted action and object.

1) Geometric features extraction: In order to generate the geometric features, object detection and skeleton detection are utilized. In this work, we make use of multiple geometric features as follows:

- The bounding box (Bbox) of the objects is defined by the pixel coordinates of the point (x_{min}, y_{min}, x_{max}, y_{max}).
- The Bbox of the human in the scene is defined in the same way as the bounding box of the object.
- The two-dimensional skeleton pose of the human is represented by the pixel coordinates (x, y) of the joints.

For the purpose of object detection, we make use of the YOLOv5 framework [31], which proved to be real-time and highly precise. We train our custom YOLOv5 model to detect humans as well as a set of predefined objects that are important for our tasks. As for the two-dimensional skeleton pose detection of the human, we make use of the Mediapipe [32] framework from Google. Mediapipe enables us to detect up to 33 joints for each detected skeleton. Following the steps and methodology used in creating the 2G-GCN network, we decide to use a sample of the upper body of the extracted joints. Because the upper part of the body is the most involved in performing the specific HOIs. The activity of pouring liquids from a bottle to a cup involves many of the joints that are in the upper part of the skeleton (arms, hands, and torso). Another reason for focusing on the upper body of the human is that these joints are more likely to be visible even if the person is seated during the activity. To further process the extracted geometric features, a specific set of joints of the human skeleton is used. These are represented as $S = \{S_t^{h,k}\}_{t=1,h=1,k=1}^{T,H,K}$, where $S_t^{h,k}$ denotes the body joint k of human h at time t. The position $p_{t,o,i}$ and velocity $v_{t,o,i}$ of objects O_i are also calculated and added as additional geometric information. These are extracted from the bounding box coordinates (x_{min}, y_{min}, x_{max}, y_{max}) of the humans and objects in the scene. The position of the joint is defined as $p_{t,h,k} = (x_{t,h,k}, y_{t,h,k})$ in 2D and the framewise velocity as $v_{t,h,k} = p_{t+1,h,k} - p_{t,h,k}$. Besides object detection and skeleton extraction, we also used Faster R-CNN [33] backbone as a feature extractor to extract the visual information in the bounding box. This is because in addition to the semantic information, we also want to utilize the pixel information in the frame to augment the HOI segment prediction model.

B. Human-object interaction future prediction

For the purpose of allowing mobile YuMi to assist humans, we need a tuple input consisting of $< action, object >$ from the output of 2G-GCN network. Based on the inference process of 2G-GCN, we first build the graph G_t at frame t. Each node n_t in G_t represents the geometric features that we proposed previously. Different from the 2G-GCN model, we modified the node detention D_n to K (body joints) + i (objects) in case the robot can only offer assistance to a single person. After embedding two fully connected layer for n_t to enhance the feature correlation between human and objects, we got optimized node $\tilde{n}_t$:

$$\tilde{n}_t = \sigma(W_2(\sigma(W_1\mathbf{n}_t + \mathbf{b}_1)) + \mathbf{b}_2) \quad (1)$$

where W denotes the weights and b denotes the bias of the fully connected layer. σ denotes the ReLU activation function.

We further define the adjacency matrix A using the same approach as in 2G-GCN by θ and ϕ, which are transformation functions represented by 1×1 CNN layers:

$$A_t(D_1, D_2) = \theta(\tilde{n}_{t,D_1})^T \phi(\tilde{n}_{t,D_2}) \quad (2)$$

After the geometric features pass through the entire GCN network, we can obtain the feature map M_t before making the final prediction:

$$M_t = A_t \tilde{G}_t W_n \quad (3)$$

where $G_t = (\tilde{n_1}, \tilde{n_2}, ..., \tilde{n_D})$, W_n is the weight matrix of GCN. Since the original network is constrained only to detect the action that is occurring, whereas we need to know the object that the person is interacting with at the same time. Therefore, we improve the classification layer with two branches so that the network structure can recognize both action and object:

$$Output_{a,t} = Softmax(\sigma(W_{a_2}(\sigma(W_{a_1} M_t)))) \quad (4)$$

$$Output_{o,t} = Softmax(\sigma(W_{o_2}(\sigma(W_{o_1} M_t)))) \quad (5)$$

However, the current network still can not perform the next human-object interaction segment. To provide more convenient assistance for human, the robot system needs to predict future human activities rather than recognize the current occurring activity by human. For this purpose, we introduce a label shift approach via training as shown in Figure 3. We shift the ground truth (GT) labels of the actions and objects of the interaction to the previous time step for training. This then ensures that the joint geometric and visual features of a HOI segment at a time step hoi_i reflect the HOI segment at time point hoi_{i+1} in the video.

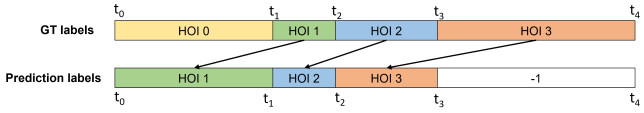

Fig. 3: The labeling technique used for transforming the network into predicting future HOI.

After the training of our model is finished, we can perform further prediction for the HOI segment. The final prediction results for the HOI segment are as follow:

$$Output_{a,hoi_{i+1}} = Softmax(\sigma(W_{a_2}(\sigma(W_{a_1} M_{hoi_i})))) \quad (6)$$

$$Output_{o,hoi_{i+1}} = Softmax(\sigma(W_{o_2}(\sigma(W_{o_1} M_{hoi_i})))) \quad (7)$$

In addition to the modifications, we design a new loss function to simultaneously backpropagate the incorrect prediction for action and object. We use the shifted segment labels $l_{hoi_{i+1}}$ and the binary cross entropy loss [34] for the final loss function:

$$\mathcal{L}_{seg} = \frac{1}{T} \sum_{t=1}^{T} \left[\frac{1}{N} \sum_{i=1}^{N} BCE\left(\hat{l}_{hoi_{i+1}}, l_{hoi_{i+1}}\right) \right] \quad (8)$$

C. The behavior tree for assistive task executation

The whole mobile manipulation application is implemented as a structured behavioral tree [35] for reference. This tree is highly reconfigurable and can be adapted and modified for any specific application. Its strength comes from the ability to create complex tasks composed of simpler ones. In our case, as discussed before, the activity of fetching an object is implemented. Figure 4 denotes the structure of the behavioral tree we use.

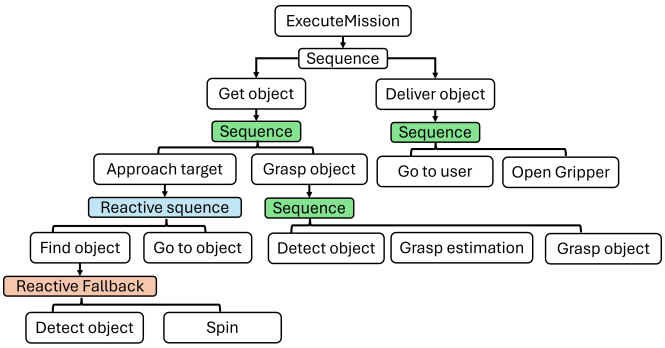

Fig. 4: The behavior tree implementation of the mobile manipulation task for YuMi. Assistive robots can perform object localization, navigation, grasping the target object and delivery tasks with this architecture. When the target object is lost, YuMi can also spin for a wider search.

The behavior tree implementation is composed of three main branches as follows:

- The *ApproachTarget* branch which is a child of the GetObject action. Here we use a visual guidance algorithm for navigating the robot to the object using the camera and the Nav2 stack of ROS2. A recovery behavior is also integrated with the branch. Specifically, if the goal object is not detected in the current frame, the robot spins. When the detection is successful, an optimal placement algorithm is used to generate a goal pose.
- The *GraspObject* branch which is also a child of the GetObject action. Here the robot has already navigated to the object position. An optimal grasp pose is generated using the YOLOv5 framework and a grasp detection algorithm. The robot then tries to drive one of his arms to the corresponding pose and grasps the object. This branch also comes with a recovery behavior where the robot tries to use its other arm if the first one meets a joint limit or a singularity.
- The last action branch is a deliver object action. Here the robot spins to locate the human user, calculates its 3D pose using the visual guidance algorithm, and uses the Nav2 stack to plan and drive back to the subject.

IV. Experimental Results

In this chapter, we discuss the experiments of the 2-stage mobile assistant structure. We evaluated the HOI detection and HOI segment prediction in the first stage. Furthermore, we also evaluate the execution time of the whole pipeline.

Fig. 5: The SPHOI dataset activities we used. From left to right these are: 1) pouring liquid from a bottle to a tube. 2) Pouring liquid from a bottle into a cup. 3) Cutting hair and drying it. 4) Working on a laptop and solving problems on a notebook.

TABLE I: The activities in the MPHOI and SPHOI datasets along with the corresponding actions and objects involved.

MPHOI	Actions	Objects
Cheering activity	Approach, Retreat, Lift, Place, Cheer, Drink, Pour	Cup Bottle
Hair cutting activity	Approach, Retreat, Lift, Place, Cut, Dry, Sit	Scissors Hair-dryer
Working activity	Work Solve, Ask	Laptop, Mouse Keyboard
SPHOI	Actions	Objects
Working with tubes activity	Approach, Retreat, Lift, Place, Check, Pour	Bottle Tube

A. Dataset

In order to demonstrate the feasibility of our method and to evaluate the modified model, we have conducted experiments using MPHOI [27] dataset. It mainly focuses on multi-person activity understanding. It is composed of 72 videos with an average length of 15 seconds, containing 13 sub-activities and 2 humans. In order to satisfy our need for output in the form of $< action, object >$, the labels of the original dataset are split into the required forms in Table I (MPHOI). But it can not satisfy the real application of our model for the robot assistant in real scenarios. To this end, we created a simple single-person human-object interaction (SPHOI) dataset to fulfill the requirements for the mobile robotic assistant YuMi.

Specifically, our focus is on single human assistance and object grasping. We have selected 4 different activities that are relevant to both industrial and daily living tasks. Most of these activities are motivated by the MPHOI dataset. We have also defined a set of target objects to be handled by the assisting robot. Table I (SPHOI) provides a detailed description of the new activity that we added on top of the MPHOI dataset and its corresponding related objects and sub-activities. Figure 5 also gives an overview of the activities we used in the dataset.

B. Experiments setup

We use a NVIDIA Quadro T1200 GPU for model training and inference. We train 40 epochs for each model and evaluate them using the standard test dataset developed on both MPHOI and SPHOI datasets. To validate the performance of the model, we use the standard F1@k form of evaluation as [37], [27]. We evaluated both the detection of the current HOI and the prediction of the next HOI segment.

TABLE II: HOI tuple detection evaluation of the ASSIGN, 2G-GCN and the modified 2G-GCN on the MPHOI dataset.

Model	F1@k score		
	F1@10	F1@25	F1@50
ASSIGN [36]	59.10	51.00	33.20
2G-GCN [27]	68.50	60.20	45.30
Modified 2G-GCN	**69.33**	**62.81**	**46.31**

TABLE III: Comparison of the F1@k score using the modified model for HOI prediction on the SPHOI and MPHOI datasets.

Dataset	Model	F1@k score		
		F1@10	F1@25	F1@50
MPHOI	HOI detection	69.33	62.81	46.31
	HOI prediction	73.75	68.36	59.5
SPHOI	**HOI detection**	**84.42**	**80.04**	**70.41**
	HOI prediction	**80.15**	**71.98**	**64.98**

C. Comparison with visual feature based networks

We first compared the model to the current SOTA models evaluated on the MPHOI dataset. The purpose of this comparison was primarily to demonstrate the validity of our current model and to select an effective model for the next work. The results are shown in Table II. The geometric feature-informed 2G-GCN model performs better than the visual-based ASSIGN model by a considerable gap. The 2G-GCN score is higher in every F1 configuration, reaching an F1@10 score of 68.5%, which is 9.5% higher than ASSIGN, and an F1@50 score reaching 45.3% marking a 12% increase. Our modified 2G-GCN model has a slight improvement with respect to the original model, which is due to the correction of the model in the loss function as a result of object recognition.

D. HOI detection and segment prediction evaluation

In this section, the modified 2G-GCN model is evaluated on the ability to predict future $< activity, object >$ pairs. To predict future HOI segment, we utilize the proposed HOI segment label shift approach. We conducted separate evaluations on the MPHOI and SPHOI, and the results are shown in Table III. The modified 2G-GCN model not only performs stably on detection but also can achieve stable performance on prediction. Comparing the results to the MPHOI dataset, the F1@k score with the SPHOI dataset increased. The F1@10 score reaches 80.15% and an F1@50 scores 64.98%. In general, a strong performance is indicated, surpassing even the training on the MPHOI dataset. However, this outcome could potentially be attributed to the fact that the SPHOI dataset comprises fewer videos and thus possesses smaller training data compared to the MPHOI dataset. The predicted outcomes are subsequently employed for real-world testing on the Mobile YuMi robot.

E. Qualitative evaluation of the HOI segment prediction

We also perform the qualitative evaluation of the first stage of the pipeline. The results are evaluated on the MPHOI and SPHOI datasets separately.

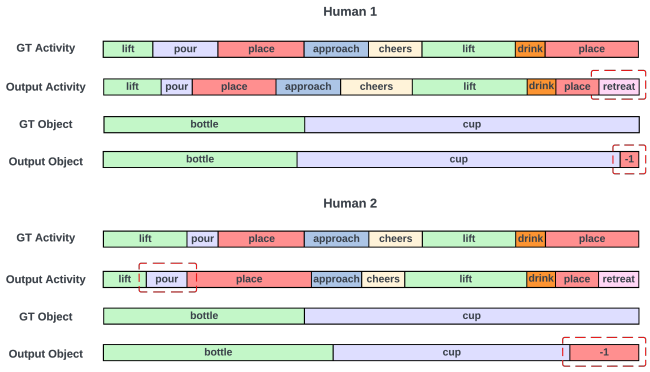

Fig. 6: The modified 2G-GCN HOI segment prediction evaluation results for the activity *Cheering* of the MPHOI dataset. The dashed squares show predictions where the model is still deficient.

1) HOI prediction evaluation on the MPHOI dataset: After retraining the model for prediction, its functioning is tested on the same *Cheering* activity that was used to evaluate the original model. The segments for the ground truth and the predicted $< activity, object >$ pairs are shown in Figure 6. The results of the modified 2G-GCN model generally show great potential in predicting future sub-activities of the *Cheering* task. The segments of the prediction overlap to a point with their ground truth counterpart. As for the prediction of the objects of the HOI, the result is also fairly accurate for both humans. The only exception is the prediction in the last frames, where the *Null* object is detected. It means that no object is used with that sub-activity. This resembles the behavior observed in the prediction of activities. The cause of such prediction can be the missing information and labels of the last frames when training the network. This occurs due to the shifting of the ground truth HOI tuple labels to an earlier point before training. Therefore, the last frames of the videos are not annotated and are ignored later in the training process.

2) HOI prediction evaluation on the SPHOI dataset: The model inference on the proposed SPHOI dataset results are shown in 7. The future action and interacted object in the HOI are accurately predicted. However, there are still certain inaccuracies in action prediction. Specifically, the *approach* action is under-segmented, while the *lift* action is over-segmented. Fortunately, these issues do not negatively impact the correct sequencing of the action sequence. It is important to note that no activities are missed altogether. Additionally, the object prediction closely aligns with its actual ground truth, displaying accurate labeling without errors. These findings will be instrumental in the practical implementation of our robotic assistance application.

F. Evaluation of real-time robotic assistance pipeline

In the final stage, we assess the functionality of the complete pipeline. Averaging across 20 experiments, we record the mean duration taken by the entire pipeline, starting from predicting a task to be performed by a human, up to the completion of assistance provided by the mobile assistive robot. The outcomes are illustrated in Figure 8. For the specific activity involving working with tubes, the average time taken for providing assistance reached 76 seconds, accompanied by a standard deviation of 16.2 seconds. It is important to note that this time can vary significantly based on factors such as the chosen path of robot to the target object and its navigation and manipulation speeds.

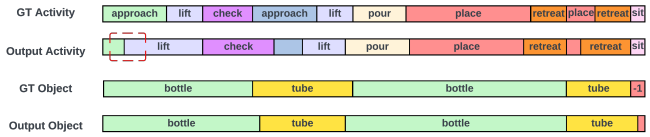

Fig. 7: The modified 2G-GCN HOI segment prediction evaluation results for the activity *working with tubes* of the SPHOI dataset. The dashed squares show predictions where the model is still deficient.

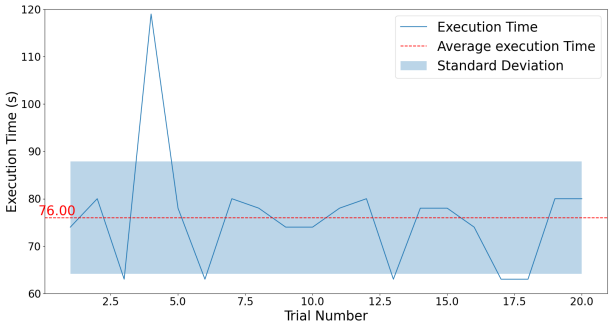

Fig. 8: Evaluation of the execution time of the whole mobile manipulation pipeline.

V. Conclusion

In this work, we propose a novel HOI segment prediction approach for the future decision making of the mobile robotic assistant YuMi. Furthermore, a two-stage assistive structure is constructed to perform the tasks of the entire assistance. The proposed HOI segment prediction model successfully provides robotic assistance with execution inputs and completes the entire process of supporting human. We also evaluate the perceptual HOI segment prediction approach on two datasets. The entire mobile robotic assistance framework is evaluated in real time with an industry scenario. The application serves its purpose and has a high success rate. Therefore, by understanding human activities and predicting the HOIs segment, assistive robots can provide effective assistance in real-world applications.

Acknowledgment

We gratefully acknowledge the funding of the Lighthouse Initiative Geriatronics by StMWi Bayern (Project X, grant no. 5140951) and LongLeif GaPa GmbH (Project Y, grant no. 5140953).

REFERENCES

[1] P. Maurice, M. E. Huber, N. Hogan, and D. Sternad, "Velocity-curvature patterns limit human–robot physical interaction," *IEEE Robotics and Automation Letters*, vol. 3, no. 1, pp. 249–256, 2018.

[2] K. M. Lynch, C. Liu, A. Sørensen, S. Kim, M. Peshkin, J. E. Colgate, T. Tickel, D. Hannon, and K. Shiels, "Motion guides for assisted manipulation," *The International Journal of Robotics Research*, vol. 21, no. 1, pp. 27–43, 2002.

[3] H.-M. Gross, S. Mueller, C. Schroeter, M. Volkhardt, A. Scheidig, K. Debes, K. Richter, and N. Doering, "Robot companion for domestic health assistance: Implementation, test and case study under everyday conditions in private apartments," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015, pp. 5992–5999.

[4] S. Jain and B. Argall, "Grasp detection for assistive robotic manipulation," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, 2016, pp. 2015–2021.

[5] M. Shomin, J. Forlizzi, and R. Hollis, "Sit-to-stand assistance with a balancing mobile robot," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 2015, pp. 3795–3800.

[6] H. S. Koppula and A. Saxena, "Anticipating human activities using object affordances for reactive robotic response," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 1, pp. 14–29, 2015.

[7] P. Wang, J. Liu, F. Hou, D. Chen, Z. Xia, and S. Guo, "Organization and understanding of a tactile information dataset tacact for physical human-robot interaction," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 7328–7333.

[8] J. Woo, S. Inoue, Y. Ohyama, and N. Kubota, "Physical contact interaction based on touch sensory information for robot partners," in *2020 13th International Conference on Human System Interaction (HSI)*, 2020, pp. 100–105.

[9] L. Fortini, M. Leonori, J. M. Gandarias, E. De Momi, and A. Ajoudani, "Open-vico: An open-source gazebo toolkit for vision-based skeleton tracking in human-robot collaboration," in *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2022, pp. 511–517.

[10] E. Merlo, E. Lamon, F. Fusaro, M. Lorenzini, A. Carfi, F. Mastrogiovanni, and A. Ajoudani, "Dynamic human-robot role allocation based on human ergonomics risk prediction and robot actions adaptation," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 2825–2831.

[11] W. Jiang and Z. Yin, "Human activity recognition using wearable sensors by deep convolutional neural networks," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1307–1310.

[12] Y. Matsumoto, K. Ogata, I. Kajitani, K. Homma, and Y. Wakita, "Evaluating robotic devices of non-wearable transferring aids using whole-body robotic simulator of the elderly," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 1–9.

[13] M. Tröbinger, C. Jähne, Z. Qu, J. Elsner, A. Reindl, S. Getz, T. Goll, B. Loinger, T. Loibl, C. Kugler, C. Calafell, M. Sabaghian, T. Ende, D. Wahrmann, S. Parusel, S. Haddadin, and S. Haddadin, "Introducing garmi - a service robotics platform to support the elderly at home: Design philosophy, system overview and first results," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 5857–5864, 2021.

[14] T. Asfour, L. Kaul, M. Wächter, S. Ottenhaus, P. Weiner, S. Rader, R. Grimm, Y. Zhou, M. Grotz, F. Paus, D. Shingarey, and H. Haubert, "Armar-6: A collaborative humanoid robot for industrial environments," in *2018 IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids)*, 2018, pp. 447–454.

[15] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robotics and autonomous systems*, vol. 57, no. 5, pp. 469–483, 2009.

[16] B. Hertel and S. R. Ahmadzadeh, "Learning from successful and failed demonstrations via optimization," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 7807–7812.

[17] S. Jain and B. Argall, "Probabilistic human intent recognition for shared autonomy in assistive robotics," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 9, no. 1, pp. 1–23, 2019.

[18] Z. Zhuang, Y. Ben-Shabat, J. Zhang, S. Gould, and R. Mahony, "Goferbot: A visual guided human-robot collaborative assembly system," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 8910–8917.

[19] Y. Wu, X. Su, D. Salihu, H. Xing, M. Zakour, and C. Patsch, "Modeling action spatiotemporal relationships using graph-based class-level attention network for long-term action detection," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 6719–6726.

[20] J. Hu, A. Edsinger, Y.-J. Lim, N. Donaldson, M. Solano, A. Solochek, and R. Marchessault, "An advanced medical robotic system augmenting healthcare capabilities - robotic nursing assistant," in *2011 IEEE International Conference on Robotics and Automation*, 2011, pp. 6264–6269.

[21] S. Schröer, I. Killmann, B. Frank, M. Völker, L. Fiederer, T. Ball, and W. Burgard, "An autonomous robotic assistant for drinking," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 2015, pp. 6482–6487.

[22] W. Kim, P. Balatti, E. Lamon, and A. Ajoudani, "Moca-man: A mobile and reconfigurable collaborative robot assistant for conjoined human-robot actions," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 10 191–10 197.

[23] E. Lamon, F. Fusaro, P. Balatti, W. Kim, and A. Ajoudani, "A visuo-haptic guidance interface for mobile collaborative robotic assistant (moca)," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 11 253–11 260.

[24] J. Massardi, M. Gravel, and Beaudry, "Parc: A plan and activity recognition component for assistive robots," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 3025–3031.

[25] C. R. G. Dreher, M. Wächter, and T. Asfour, "Learning object-action relations from bimanual human demonstration using graph networks," *IEEE Robotics and Automation Letters*, vol. 5, no. 1, pp. 187–194, 2020.

[26] R. Morais, V. Le, S. Venkatesh, and T. Tran, "Learning asynchronous and sparse human-object interaction in videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 16 041–16 050.

[27] T. Qiao, Q. Men, F. W. Li, Y. Kubotani, S. Morishima, and H. P. Shum, "Geometric features informed multi-person human-object interaction recognition in videos," in *European Conference on Computer Vision*. Springer, 2022, pp. 474–491.

[28] J. Lim, V. M. Baskaran, J. M.-Y. Lim, K. Wong, J. See, and M. Tistarelli, "Ernet: An efficient and reliable human-object interaction detection network," *IEEE Transactions on Image Processing*, vol. 32, pp. 964–979, 2023.

[29] Y.-W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng, "Hico: A benchmark for recognizing human-object interactions in images," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1017–1025.

[30] Y.-L. Li, S. Zhou, X. Huang, L. Xu, Z. Ma, H.-S. Fang, Y. Wang, and C. Lu, "Transferable interactiveness knowledge for human-object interaction detection," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3580–3589.

[31] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, NanoCode012, Y. Kwon, TaoXie, K. Michael, J. Fang, imyhxy, and et al., "ultralytics/yolov5: v6.2 - yolov5 classification models, apple m1, reproducibility, clearml and deci.ai integrations," Aug 2022.

[32] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee *et al.*, "Mediapipe: A framework for building perception pipelines," *arXiv preprint arXiv:1906.08172*, 2019.

[33] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[34] J. Nam, J. Kim, E. Loza Mencía, I. Gurevych, and J. Fürnkranz, "Large-scale multi-label text classification—revisiting neural networks," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part II 14*. Springer, 2014, pp. 437–452.

[35] M. Colledanchise and P. Ögren, *Behavior trees in robotics and AI: An introduction*. CRC Press, 2018.

[36] R. Morais, V. Le, S. Venkatesh, and T. Tran, "Learning asynchronous and sparse human-object interaction in videos," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 16 041–16 050.

[37] O. Sener and A. Saxena, "rcrf: Recursive belief estimation over crfs in rgb-d activity videos." in *Robotics: Science and systems*, 2015.