Preliminary experiments for behavior log generation considering spatial relationship between actions and objects

Masae Yokota[1], Sarthak Pathak[2][0000−0002−5271−1782], Mihoko Niitsuma[2][0009−0008−3800−7319], and Kazunori Umeda[2][0000−0002−4458−4648]

[1] Course of Precision Engineering, Graduate School of Science and Engineering, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo, Japan
yokota@sensor.mech.chuo-u.ac.jp
[2] Department of Precision Mechanics, Faculty of Science and Engineering, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo, Japan
{pathak, niitsuma, umeda}@mech.chuo-u.ac.jp

Abstract. This paper describes a method for creating human action sentences from video information, which aims to infer the purpose of human actions from video information. To develop a flexible human-robot communication method that enables robots to infer their own roles from the appearance of people, it is necessary to observe and convert human behavior into data. Therefore, we have created sentences describing human behavior by combining an action recognition model and an object detection algorithm, and we have organized the spatial relationships between objects and people. In Verification 1, a simple one-sentence explanation of the behavior was provided based on the distance between the two-dimensional (2D) coordinates of the person and the object taking the action from the video image information. Through the verification, we have identified issues in the generation of behavioral sentences. In Verification 2, we spatially organized their relationship based on the change in the three-dimensional (3D) coordinates of the object and the person. Through this validation, we were able to obtain spatial information on human behavior and object access records. In the future, by integrating these methods, we aim to generate an in-spatial action log that links human behavior and spatial information.

Keywords: Human-object Intention · Intelligent Room · Action recognition.

1 Introduction

In recent years, flexible human-robot communication methods have been studied to improve the efficiency of human-robot communication and reduce the burden on humans [1] - [3]. In human-robot collaborative work, a person naturally understands the division of roles by observing the behavior and actions of the other person. However, in human-robot collaborative work, people often give explicit

instructions to the robot, which increases the burden on the person during the work and prevents flexible communication. Therefore, methods to help robots understand the intent of human actions by observing and contextually interpreting records of human actions and access to objects have been studied [4]. In this paper, we attempted to generate action logs from spatial information of human actions and accessed objects in the observation space in order to make robots understand the potential purpose of human behaviors and to generate human action sentences from video image information. In Verification 1, the distance between the bounding boxes of detected actions and objects was observed, and the method of identifying the object that is the target of the action was discussed. In Validation 2, we organized the location of people and objects in the overhead view. From these results, we discuss and identify issues for the integration of the method and the automatic generation of human action sentences.

2 Related works

Research is being conducted with the aim of improving the effectiveness of human-robot collaboration by building natural and intuitive communication methods using image information [5]. Among these methods, research is being conducted to understand human intention by observing human behavior, and robots are actively acting to assist human tasks [6]. In our study, based on the seven-step action model proposed by Norman [7], we believe that clear intentions are latent in a person's sequence of actions. For example, when the intention to go shopping is formed, a person plans and performs detailed actions such as "picking up a bag," "putting a purse in a bag," "putting on a jacket," "going to the entrance," "putting on shoes," and "opening the door." By contextually inferring a person's intentions from the history of the multiple actions performed, it is possible to predict the actions that will be performed after the inferred point in time. Thus, it is necessary to represent actions as data in order to understand and infer people's intentions, which is the ultimate goal of this research. Here, it is important to consider time series and context in order to understand the intention of a person's movement from the observation results of the person's behavior. Research is also being conducted to generate explanatory text from images, taking into account the interaction between people and objects [8].

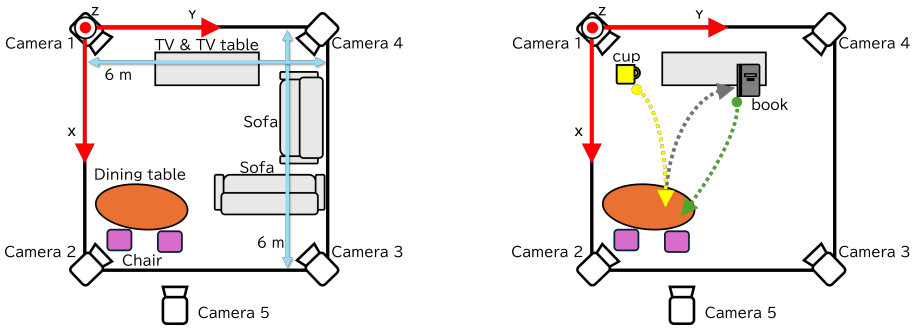
However, it is difficult to construct a model from scratch that explains the meaning of each action and each accessed object, because it requires a huge amount of data sets and time. Therefore, this paper aims to organize human-object interaction from video information and represent it as data, and conducted the following two types of verification.

- Verification 1: Verification of a Method for Searching Objects for Action by 2D Distance Relationship between Action Detection Bounding Box and Object Detection Bounding Box
- Verification 2: Verification to identify the conditions necessary to identify the target object for action by observing the 3D position of a person and the 3D position of an object

3 Preliminary verifications

3.1 Overview

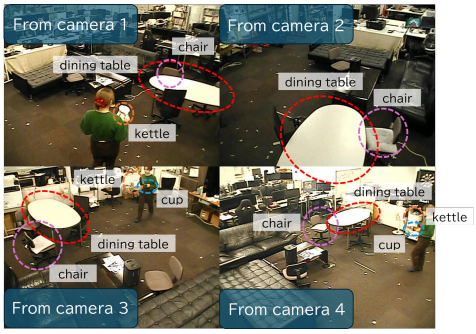
Two types of elemental validation were performed to generate human action sentences from video image information. Both validations used MMAction2 [9] as the action detection algorithm and YOLO [10] as the object detection algorithm. A single video recorded in the environment shown in Fig. 1a was used for verification. The path of movement of the person is shown in Fig. 1b. The combined image taken from camera 1 to 4 is shown in Fig. 2a. The image taken from camera 5 is shown in Fig. 2b. The "chair" indicated by the purple triangle in Fig. 1b is the chair in Fig. 2a and Fig. 2b. The images were recorded at the same time. First, a person with a cup and a kettle approaches the dining table from the far left of the screen and sets them down, then goes to the far right to retrieve a book, returns to the dining table, sits on the chair, and finally drinks a drink while reading the book. In Verification 1, we used videos taken as shown in Fig. 2b. We verified the possible to extract the object of the action from the 2D distance between the center of the bounding box acquired when the action was detected and the center of the bounding box acquired when the object was detected. Then, in Verification 2, we used the video acquired in Fig. 2a. The location of the objects accessed by the person and the path of the person's movement were organized in an overhead view and compared to the timing at which the action occurred.



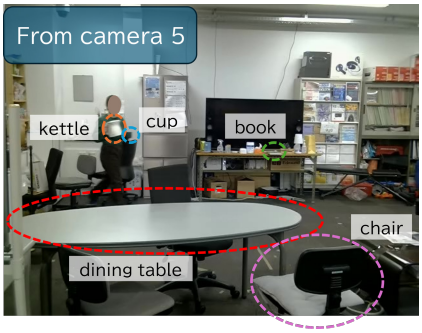
(a) Overead view of the used environment

(b) Overhead view of a simplified representation of the behavior in the video used for verification

Fig. 1: Environment under which the video used for verification was recorded

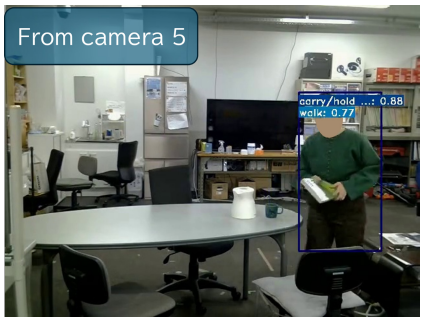


(a) The combined image taken from camera 1 to 4

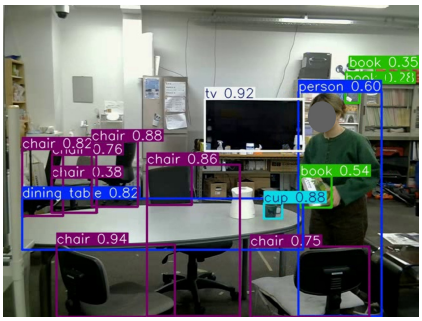


(b) The image taken from camera 5

Fig. 2: The video images



(a) The action detection result in camera 5



(b) The object detection result in camera 5

Fig. 3: The detection results

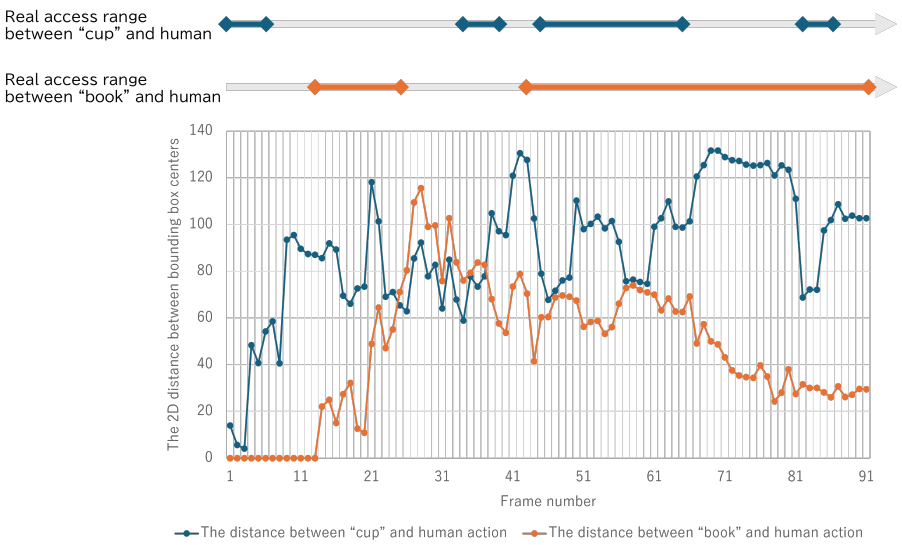
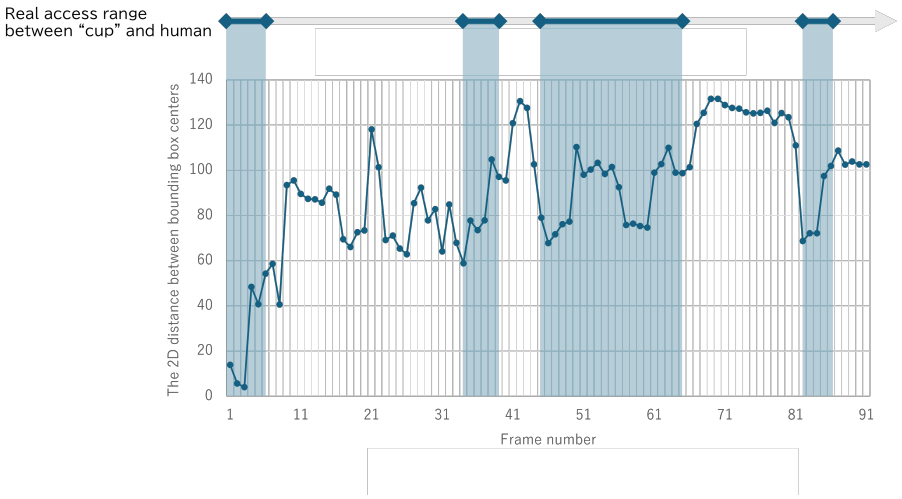


Fig. 4: The graph: Real access range and distance between the bounding box centers of each object and human action

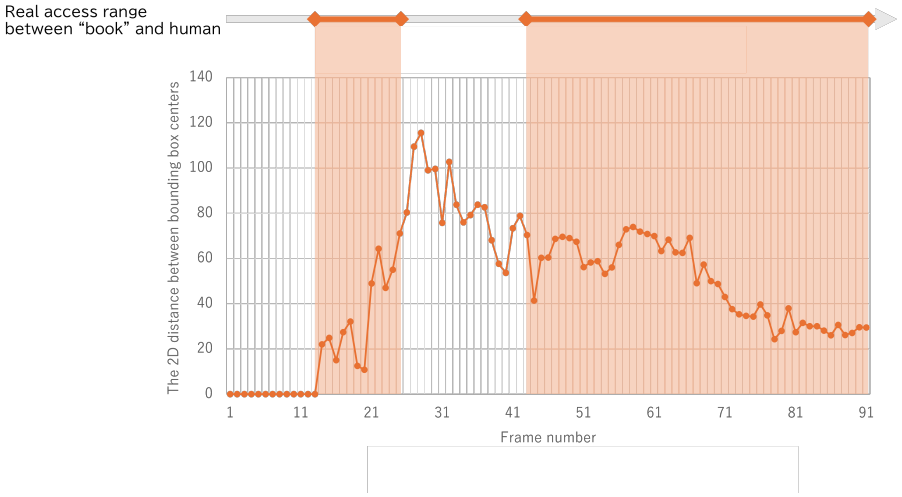
3.2 Verification 1: Identification of objects for action based on the two-dimensional distance between the action and the object's bounding box center

In this verification, we aimed at automatic generation of action statements and identified objects to be acted upon by integrating the results of action detection and object detection. Using video images as shown in Fig. 2b, taken at the positions shown in Fig. 1, the results of action detection and object detection were integrated to identify the object to be acted upon. MMAction2 [9] developed by OpenMMlab was used for action detection. This detected the actions of a person in the video every 8 frames, as shown in Fig. 3a, and saved the action label name and the center coordinates of the bounding box at the time of action detection. YOLOv8 [10] was used for object detection, and the object name and the center coordinates of the bounding box detected in each frame were saved as shown in Fig. 3b.

The 2D distance between the bounding box centers of the detected action and the detected object was calculated. We focused only on the action “carry/hold (an object)” and assumed that the two possible objects for this behavior are “cup” and “book.” The distance between the bounding box center in each frame and the true or false of the real access as a combination of actual actions and objects are shown in Fig. 4. The graphs are divided by attention to each object in Fig. 5.



(a) The graph: Real access range and distance between the bounding box centers of “cup” and human action



(b) The graph: Real access range and distance between the bounding box centers of “book” and human action

Fig. 5: The divided graph

The straight line above the line graph indicates the extent of actual contact between the object and the person. The gray arrows indicate the time series, and the blue line indicates the time between actual accesses between the "cup" and the person. The orange line indicates the time between the actual accesses between "book" and the person. Here, the cases when the object is being grasped and when it is being let go are excluded. The line graphs show the distance between the human action bounding box center and the respective object bounding box center at each frame point. In the initial frames, the bounding box center of the "book" could not be measured due to occlusion, so the distance is shown as 0.

3.3 Verification 2: Observation of three-dimensional positional relationships between objects and people for clarification of conditions for identification of objects to be acted upon

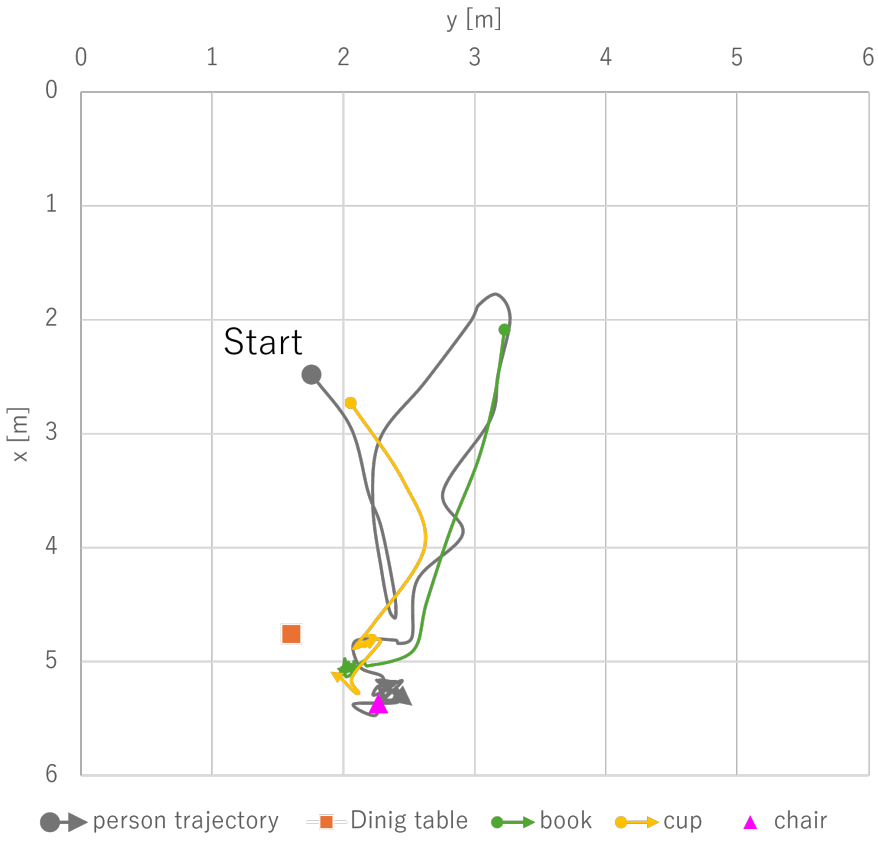


Fig. 6: Trajectory

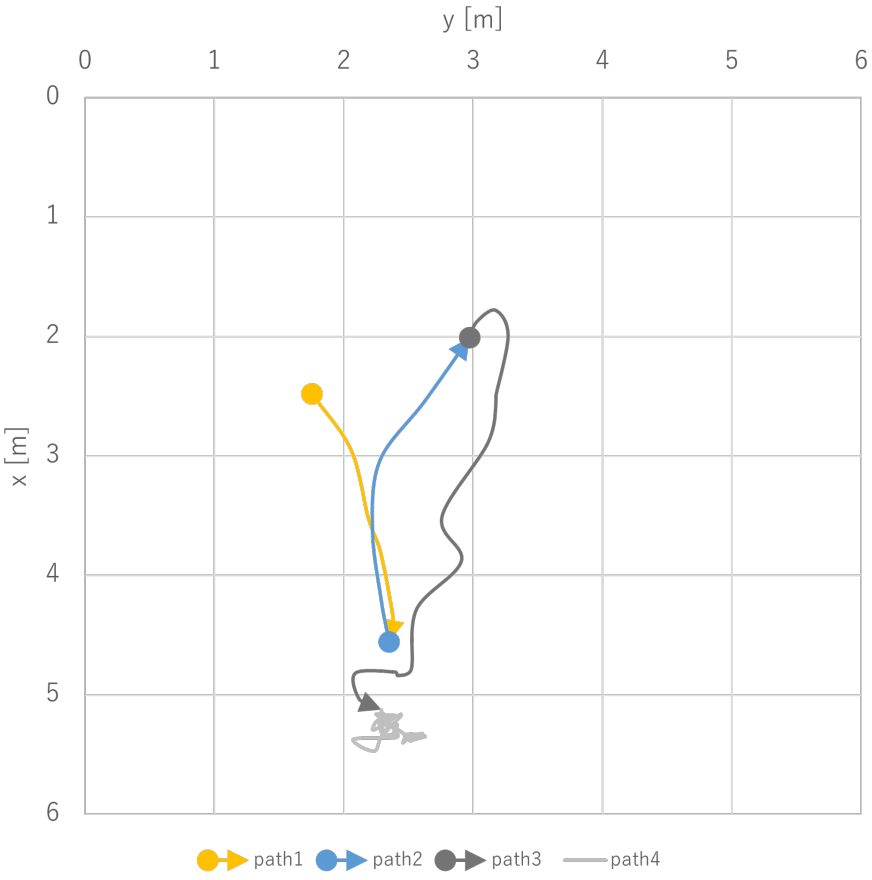


Fig. 7: The divided trajectory

In this verification, the spatial relationship between objects and people was organized using video images recorded from different angles of the video images used in Verification 1. Images obtained from 1 to 4 cameras in the environment shown in Fig. 1a were combined as shown in Fig. 2a, and object detection using YOLO was performed on these moving images to organize the relationship between the trajectory of the person and the position of the object. In this process, since the same point cannot be observed stably at the bounding box center of a person's whole body, we focused on the head, which can be approximated as a sphere, and used the path of the bounding box center of the head as the person's movement path.

As in Verification 1, we focused only on "carry/hold (an object)" when considering access to objects, and "cup" and "book" as objects that could cause this action. The movement paths of people and the changes in the positions of "cup"

and "book" are plotted in Fig. 6. The positions of the dining table and chairs are also plotted to improve visibility as movement paths. The color-coded pathways based on the model diagram in Fig. 1b are shown in Fig. 7. Path1 is to walk up to "dining table" with "cup" and place it there. Path2 is to walk up to "dining table" and place it there. path2 shows the path from "dining table" to "book." Path3 is the path from picking up "book" to walking to "dining table" and setting it down. After that, the position where the user is sitting on "chair" reading a book is path4.

4 Discussion

4.1 Discussion from verification 1

In Verification 1, we tried to identify objects by the distance between bounding box centers. Initially, we considered identifying which object a person is holding by determining the distance between the bounding box centers by a threshold value. A plot of the distance between the bounding box center of "cup" and "book" and the bounding box center of the human action is shown in Fig. 4. For each object, Fig. 5 shows the range of frames in which the person is actually holding the object and the distance between the action bounding box center and the object bounding box center in each frame.

From Fig. 5, it can be seen that in many cases the 2D distance is less than 100 when the object is actually held. Initially, we considered identifying which object a person is holding by judging the distance between bounding box centers by a threshold value. However, it is difficult to identify the target of the action by the threshold value because there are cases where the distance is greater than 100 even though the person is actually holding an object, and there are cases where the distance is less than 100 even when the person is not actually holding an object. This may be due to the fact that the size of each bounding box is not taken into account. In addition, it was difficult to distinguish the difference between what a person was touching and what was behind the person because of only 2D image. Therefore, we considered it necessary to consider the following situations in order to identify the target of a person's behavior.

- Taking into account the 3D distance between a person and an object
- Threshold setting considering the size of each bounding box of human actions and object bounding boxes

4.2 Discussion from verification 2

Comparing the actual human path in Fig. 7 with the path model in Fig. 1b, it can be said that we were able to record human movements almost accurately. Looking at the paths of "cup" and "book," respectively, we can see that they follow the human motion. On the other hand, the trajectory of "cup" deviates significantly from the human's trajectory in the middle. This is thought to be due to the bending and stretching of the person's arms. Therefore, it is necessary

to take into account the size of the person's body and the length of the arms and legs in order to recognize this kind of behavior. However, while trajectories are easy to observe when a person is moving with an object, changes in the 3D position of an object when a person is sitting in a chair are more minute than when moving.

Therefore, in order to observe the interaction between a person and an object when the person is stationary, it is necessary to consider the object to be focused on for each action, rather than only the coordinate change.

4.3 Discussion for the future work

Through these verifications and discussions, the following three points can be identified as future issues to be addressed in order to automatically generate human action sentences, focusing also on the object to be acted upon.

- Consideration of the respective sizes of the human behavior bounding box and object bounding box
- Identification of the target object of the action considering the nature of the action and the meaning of the object
- Changes in sentence expression depending on the combination of actions and objects

First, the search for an object to be acted upon should be performed by considering human behavior and the size of the object's bounding box. Even if the distance between the bounding box centers is large, the situation may be that the person is holding an object with arms extended. In the future, increasing the number of behaviors to be observed will improve the accuracy of the retrieval of objects to be acted upon, taking the size of the bounding box into consideration.

Next, it is necessary to consider the nature of the action and the meaning of the object. For example, in the case of "carry/hold," only objects that can be held in the hand may be the target of the action. For example, in the case of "carry/hold," only objects that can be carried by hand may be the target of the action, and in the case of "sit," a chair or sofa may be the target. By calculating the likelihood of each object in a person's surroundings being the target of each of these types of actions, it is possible to identify the target object of the action in combination with location information.

Finally, after identifying the target object, it is important to consider the combination of action expressions and objects in order to generate action sentences as natural expressions. For example, when combining "sit" with an object name to generate an action sentence, "on" is expected to be added in front of the object name. On the other hand, when multiple actions are being performed simultaneously, it is difficult to identify the target object and automatically generate natural sentences. In the future, we will observe various types of multiple actions being performed at the same time and search for a method to fuse location information with the results of action detection.

5 Conclusion

In this paper, we conducted a series of validations to fuse the results of action detection and object detection for the purpose of automatic generation of human action sentences. In Verification 1, the 2D distance between the center of the action detection bounding box and the center of the object detection bounding box was calculated, and the difference in distance was observed when the actual object was being accessed and when it was not. In Verification 2, we observed the 3D position coordinates of the person and the object in an overhead view to confirm the positional relationship in the video with the person holding the object in his/her hand. Through these verifications, we clarified issues for automatic generation of human action sentences by fusing the results of action and object detection. In the future, we aim to automatically generate human action sentences with the following points in mind, and to estimate a person's action intention based on the action history sentences.

- Search for objects for action considering the size of the action detection bounding box and object detection bounding box
- Identification of objects subject to action based on the nature of the action and the meaning of the object and adjustment of the expression of the behavior sequence

References

1. D. Mukherjee., K. Gupta., L. H. Chang., H. Najjaran.: A Survey of Robot Learning Strategies for Human-Robot Collaboration in Industrial Settings. Robotics and Computer-Integrated Manufacturing **73**, 102231 (2022).
2. D. Wei., L. Chen., L. Zhao., H. Zhou., B. Huang.: A Vision-Based Measure of Environmental Effects on Inferring Human Intention During Human Robot Interaction. IEEE Sensors Journal **22**, (5), 4246–4256 (2022).
3. Y. Zhang., T. Doyle: Integrating intention-based systems in human-robot interaction: a scoping review of sensors, algorithms, and trust. Frontiers in Robotics and AI **10**, (2023).
4. Qiu. Y., Y. Nagasaki., K. Hara., H. Kataoka., R. Suzuki., K. Iwata.: VirtualHome Action Genome: A Simulated Spatio-Temporal Scene Graph Dataset with Consistent Relationship Labels. In: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 3340-3349, USA(2023).
5. R. Nicole., T. Brendan., C. Dylan., K. Dana., C. Peter.: Robotic Vision for Human-Robot Interaction and Collaboration: A Survey and Systematic Review. ACM Transactions on Human-Robot Interaction **12**(1), 1–66 (2023)
6. E. V. Mazcaro., D. Sliwowski., D. Lee.: HOI4ABOT: Human-Object Interaction Anticipation for Human Intention Reading Collaborative roBOTs. In: 7th Conference on Robot Learning, pp. 1111–1130. PMLR (2023)
7. Norman, D.: The Design of Everyday Things. (Revised and Expanded Edition). Cambridge, MA London (2013)
8. H. Zhang., W. Zhang., H. Qu., T. Liu.: Enhancing Human-Centered Dynamic Scene Understanding via Multiple LLMs Collaborated Reasoning. arXiv **2403.10107**, (2024)

9. MMAction2 Contributors.: OpenMMLab's Next Generation Video Understanding Toolbox and Benchmark. https://github.com/open-mmlab/mmaction2 (2020)
10. J. R. Terven., D. M. Córdova-Esparza.: A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS. Machine Learning and Knowledge Extraction, **5**(4), 1680–1716 (2023)