2025 IEEE/SICE International Symposium on System Integration (SII)
January 21-24, 2025, Munich, Germany

Preliminary Experiments of Inferring Human Intention by Analyzing Time-Series Images from Multiple Views

Masae Yokota[1], Sarthak Pathak[2], Mihoko Niitsuma[2] and Kazunori Umeda[2]

Abstract— **The objective of this research is to construct an intelligent human-robot environment that can infer human behavorial intentions and adjust the space accordingly. In this research, we perform preliminary studies and verify whether inferring of human behavorial intention can be done from image information alone. First, the vision and Language Model (VLM) and object detection methods are used to infer possible human actions for each object detected in images. Differences between inference results and actual behavior are identified and methods needed for more accurate inference are discussed. The spatial relationship between the skeletal points and the object by observation reveals which skeletal points to focus on in order to predict the behavior. We confirmed that it is possible to predict behaviors by focusing on the neck point for actions performed with the clear intention of sitting on or passing by a chair. Parameters for the neck skeletal points are selected and each behavior is predicted by a Temporal Convolutional Network (TCN) with 91% performance. Through preliminary experiments, we discuss the methods necessary for inferring human behavioral intentions from images.**

I. INTRODUCTION

In recent years, robot tasks have become more complex and sophisticated in automation. In highly uncertain real-world environments, robots are not able to complete tasks alone, but adapt by working in complementary cooperation between humans and robots. Thus, it is necessary to realize efficient symbiosis between robots and humans, and human-robot communication methods that are flexible and less burdensome on humans are being studied[1]-[3]. In this symbiosis, it is desirable for robots to move actively to achieve efficiency through natural coordination like humans do with each other, in order to avoid adding new workloads to humans. Therefore, various approaches have been studied to solve the problem of robots acquiring and interpreting information about people and the surrounding environment[3]-[6]. Rather than waiting for explicit instructions from the person, the robot understands the intent and context of the action from the person's movements, understands its own role, and anticipates the person's actions, thereby reducing the workloads on the person. For example, when a person is walking toward a door, the robot should estimate the person's intention to go out based on and take proactive actions such as going to open the door ahead of the person. Here, the intention of action refers to the sequence of actions that the person will take in the future in order to solve their needs. The goal of this study is to realize more flexible human-robot communication by inferring this action sequence.

In this research, we focus on estimating "intention". Estimating intention is important because a person can perform multiple actions towards the same intention. For example, "exiting a room" involves a person getting up, moving towards a door, opening the door, and stepping out. Based on actions alone, it is difficult to predict human behavior. However, by estimating the "intention" of exiting a room, several actions can be tied together and predicted in advance. Therefore, "intention" estimation is important. Predicting "intention" as opposed to "actions" alone will lead to a more robust and accurate prediction of human behavior for human-robot interaction.

The estimation of intentions requires consideration of intentions that are feasible under the environment and multiple actions to achieve them. It is necessary to be able to reliably predict the feasible actions in a certain environment. Thus, we want to confirm whether it is possible to predict the actions that people take toward objects from time-series data of spatial relationships between person and objects.

In this paper, as preliminary experiments for the inference of action sequences, we aim at the following three points:

- To interpret the possible actions of a person from object information in images.
- To clarify the information necessary for predicting the actions that a person takes to sit on a chair and the actions that a person passes by without sitting, using skeletal point information and object information considering time-series, and to confirm the contribution to the estimation of action intention.
- To discuss from these preliminary experiments and conduct an initial study of a framework for inferring the intention of human behavior from image information.

In addition, among various in which robots and humans coexist, we assume daily life in this paper.

II. RELATED WORKS

Various sensing methods for collaboration of humans and robots have been studied to help robots understand human behaviors, purposes and tendencies. In the field of computer vision, the effectiveness of collaboration has been improved by facilitating the exchange of information between humans and robots and by developing natural and intuitive communication methods[7][8].

In computer vision, research is being conducted to understand the intent of human movement by observing hu-

[1]Masae Yokota is with the Precision Engineering Course, Graduate School of Science and Engineering, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo, Japan. (Corresponding author: yokota@sensor.mech.chuo-u.ac.jp)

[2]Sarthak Pathak and Kazunori Umeda are with the Department of Precision Mechanics, Faculty of Science and Engineering, Chuo University, Tokyo, Japan

man behavior for making robots to assist human tasks actively[10]. Therefore, in this paper, we develop a behavior prediction method that observes behavior in two ways based on person's explicit intentions and aims to develop a method for estimating behavioral intentions.

The two methods are as follows:

- Focusing on each object from an image and inferring human behavior in Section III
- Identifying the characteristics of the behavior that differ depending on the intention in Section IV
- Predicting behavior by parameters obtained from observations in Section IV

III. Inference of Behavioral Intention from Images using VLM

A. Method

It is essential to link the meaning of objects and actions in order to infer the intent of a person's actions. Thus, this paper tried to list possible actions that a person may take in response to objects detected in an image by using a vision and Language Model (VLM)[9]. VLM is a model that aims to process and understand textual and image information in an integrated manner. First, we tried to predict actions from videos using only CLIP[11]. This is because we expected that the VLM would allow us to infer what aims people have for their actions from multiple consecutive actions. However, the accuracy of the inference was not good, so we combined it with object detection using YOLOv8. YOLOv8 is the latest in YOLO series of object detection algorithms, open-sourced by Ultralytics on January 10, 2023[12]. We thought that the combination of object detection and YOLOv8 would allow us to infer the possible actions of people for each object.

There is not necessarily one object that a person works with simultaneously in daily life, such as drinking a drink during reading a book. Therefore, it is necessary to pick up the actions that a person may take for each object. By organizing the related human actions for each object, we thought it would be easier to infer the aim behind the actions when analyzing a time-series of consecutive actions.

B. Experiment 1

The purpose of this section is to see if it is possible to actually match the possible actions of a person for each object as shown in Fig. 1. First, a list of possible actions is prepared for each object. Second, YOLOv8 detects an object, and an area slightly larger than the bounding box is croped each object. After the cropping process, CLIP is used to predict the most likely action from the action list.

In Video 1, a person with a book approaches a table with a cup on it, sits down to read, and drinks tea from the cup while reading. In Video 2, the person reading in the chair closes the book and walks away with the cup and the book.

When we checked the correctness of the predicted behavior against the behavior performed by the person afterward, the results were as shown in Table I and Table II.

Fig. 1: A scene of videoes: object detection result is shown as green bounding box

TABLE I: Results: actual behaviors and predicted behaviors for a book

	Actual behaviors	Predicted behaviors
Video 1	Holding	Throwing
	Reading	Holding
Video 2	Reading	Holding
	Holding	Throwing

IV. Investigation of Behavior Prediction using Time-Series Spatial Relationships between People and Objects

A. Method

In this paper, we conducted a preliminary experiment for action prediction aiming at action-sequence inference. We observed two types of actions with definite intentions, seating and passing, and analyzed time-series data of spatial relationships between people and objects to predict actions.

The 3D coordinates of human skeletal points were obtained and observed using Intel RealSense Depth Camera D455 from Intel Corporation[13] and LIPSense 3D Body Pose SDK from LIPS Corporation[14] for skeletal point detection. Temporal Convolutional Network (TCN)[15] was used as a prediction method based on time series information. Various networks are used for behavior prediction, but TCNs have attracted attention for natural language processing and processing time-series information[16][17]. In addition, TCN has been used to predict point motions with intense and flexible movements in [18], and we thought that TCN would be useful for human actions in real environments. Thus, we use TCN to predict the action of a person toward a chair based on the spatial relationship between the person's skeletal point and the chair.

We considered that using the actual coordinate values as they are in the TCN process would be affected by measurement errors and would depend on the initial positions of the person and chair. Therefore, it is necessary to generate parameters using the positional relationship between the person and the chair. Therefore, in Experiment 2, we obtained skeletal point information from various subjects, observed

TABLE II: Results: actual behaviors and predicted behaviors for a cup

	Actual behaviors	Predicted behaviors
Video 1	Holding	Holding
	Drinking	Holding
	Holding	Holding
Video 2	Holding	Holding

TABLE III: Interpretation of feature values

Value of each feature	Interpretation
$d_f < d_{f-1}$	Approaching a chair
$d_f \geq d_{f-1}$	Move away from a chair
$d_f \leq 50mm$	Reach a chair
$m_f > 30mm$	High movement speed
$10 < m_f \leq 30mm$	Low movement speed
$m_f \leq 10mm$	Almost stationary

time-series changes in the positional relationship between the subject and the chair, and examined the parameters. In Experiment 3, we used the parameters selected in Experiment 2 to predict the behavior of the subject in the TCN, and observed the results.

B. Experiment 2: Parameter Consideration

1) Overview: The purpose of this section is to examine the appropriate parameters about the time-series changes in the positional relationship with the chair to be used for behavior prediction in TCN by observing skeletal point information. The coordinate system of the depth camera is shown in Fig. 2a.

LIPSense 3D Body Pose SDK can detect 18 skeletal points. The neck skeletal point of them was extracted and observed in this paper. The neck skeletal point was selected for two reasons. First, in order to clarify the positional relationship between a person and an object, noise caused by body motion and stable detection were taken into consideration. Since noise due to body motion is considered to be greater at the ends of the body, skeletal points in the center of the body or close to the body axis are desirable. Second, there are only a few skeletal points that can be detected stably independent of the positional relationship between the camera and the subject. For these reasons, we decided to focus on skeletal points of the neck to observe the movement trajectory of a person.

2) Results: The time-series information of the skeletal points actually obtained was then organized. The experimental environment in which the data was obtained is shown in Fig. 2b. Only seating is considered as the only intention for a person to approach a chair. Subjects were asked to perform two types of actions, seating and passing, without specifying the details of the route. This experiment was conducted on nine adult subjects, all of whom gave permission for measurement.

An example of the trajectory of the skeletal point of the neck in the xz-plane, which is parallel to the floor, is shown in Fig. 3. This figure shows that the trajectory was generally a straight line rather than an arc, although there were individual differences in the movement of the left and right toward the direction of motion when seated. In both cases of sitting with the body sliding sideways and stopping once in front of the chair before sitting down, the neck position was found to show a movement that passed slightly in front of the chair once and then returned.

Then, the time-series of the distance values to the object at the neck skeleton point changed as shown in Fig. 4. In Fig. 4a, the distance value became constant near the chair and then dropped sharply. In Fig. 4b, the subject approached the chair at an almost constant speed, but never closer than about 500 mm in any case.

Next, the time-series change in the amount of movement between frames in the xz-plane is shown in Fig. 5. In Fig. 5a, the displacement becomes smaller as it approaches the chair, and immediately after the distance from the chair becomes constant in Fig. 4a, the displacement increases rapidly and then decreases rapidly. On the other hand, in Fig. 5b, the amount of movement does not change significantly as the distance from the chair gets closer.

These results suggest that attention to the skeletal points of the neck can distinguish when a person is about to sit in a chair from when he or she is not.

C. Experiment 3: Predicting Behavior via TCN

From Section IV-B, the following two features are used to determine seating and passing. The f denotes the current frame point in time. The features are interpreted as shown in the table III.

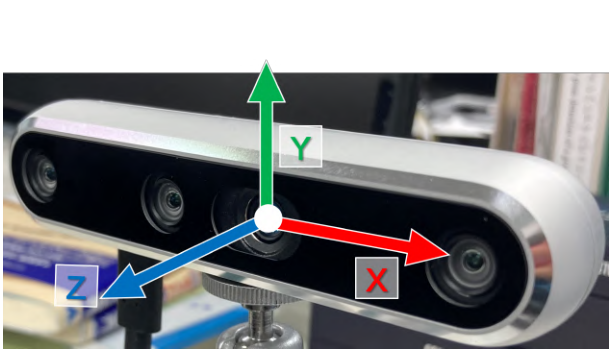
- Distance between the neck skeleton point and the chair on the xz-plane: d_f
- Frame-to-frame displacement of the neck skeleton point on the xz-plane: m_f

Referring to [16], we created a TCN model with two convolutional layers for each block (with 25, 50, and 100 channels in order), a kernel size of 2, a batch size of 64, and a window size of 10. Based on these interpretations, we used the TCN model to predict seating or passing from the data of several people, and obtained a model with 91% accuracy.

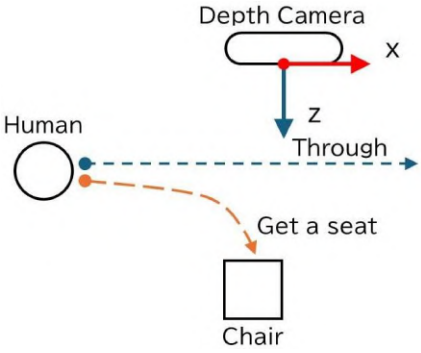
In this experiment, a simple TCN model was used to predict behavior. The loss function increased after about 100 epochs when the number of convolution layers was increased, so two layers were used because they were the least likely to increase. In addition, the loss function did not decrease sufficiently when the window size was changed, and it increased after about 120 epochs when the window size was increased.

V. DISCUSSION

In Experiment 1, YOLOv8 and CLIP were used to select the most likely behavior of a person for each object through 2 videoes from a predefined list of actions. By collecting

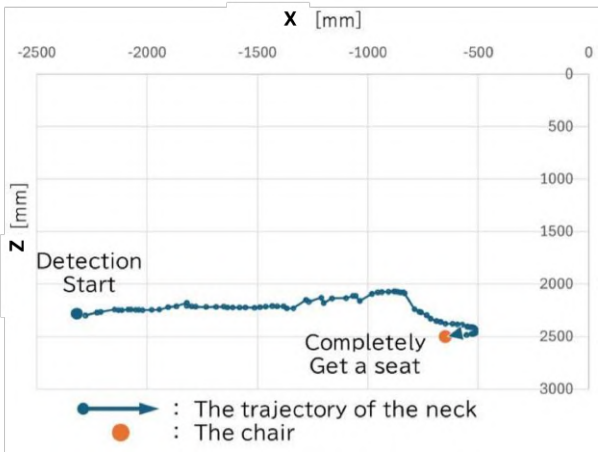


(a) Depth camera coordinate system

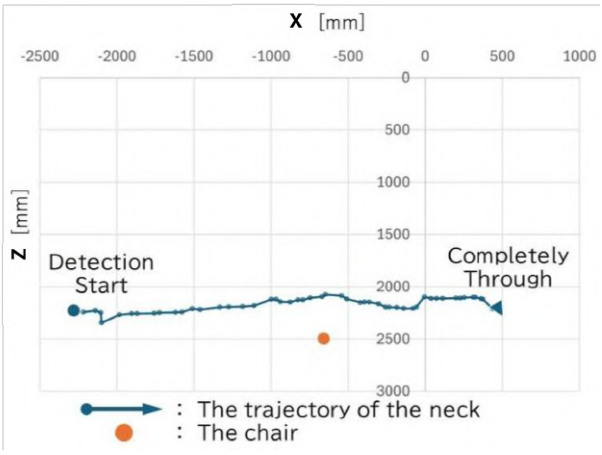


(b) Experiment environment (overhead view)

Fig. 2: Environment for data acquisition

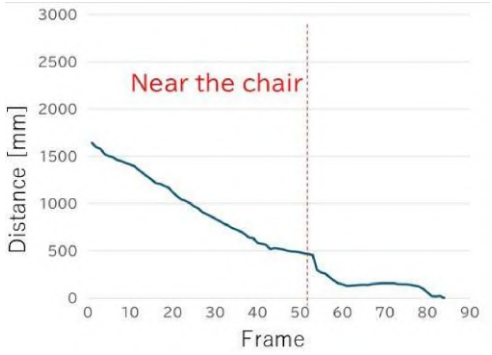


(a) In case of sitting action

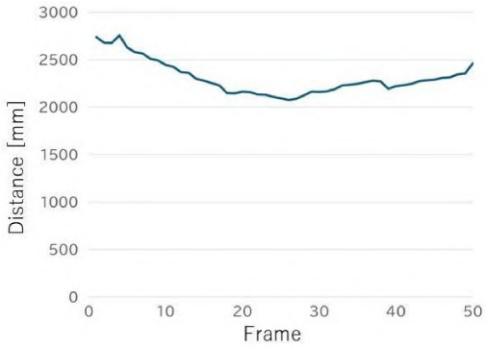


(b) In case of passing action

Fig. 3: Overhead view of the trajectory of the neck point



(a) In case of sitting action

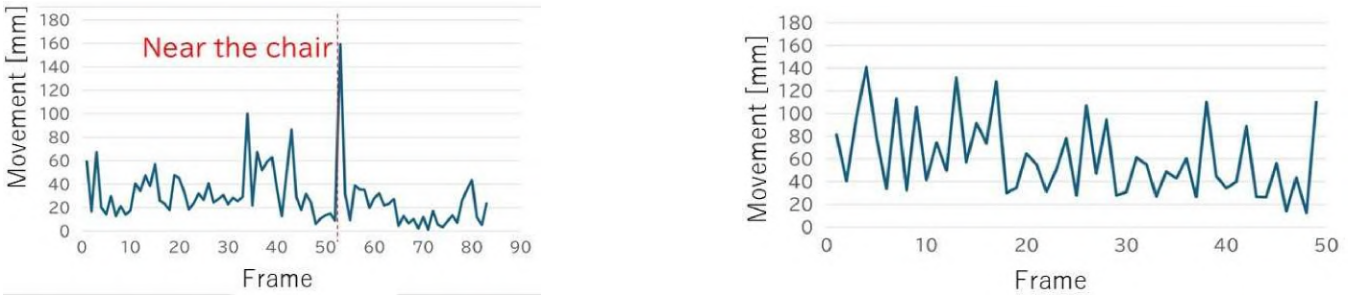


(b) In case of passing action

Fig. 4: Time-series changes in the distance between the neck point and the chair

actions taken by people for each object and organizing them chronologically, we expect to understand the aim that people want to achieve from multiple consecutive actions in the future.

However, as can be seen from Tables I and II, the guesses was poor. There are three possible causes for this issue. First, the cut-out area was too small. The scene of a person working on an object could not be fully interpreted, resulting in incorrect inferences. Second, the list of objects and corresponding actions needs to be optimized. Since inference results vary depending on the prompt, it is important to try various prompts and find the appropriate one in order to obtain better results. Finally, there is a possible lack of chronological understanding. In this method, inferences are



(a) In case of seated action

(b) In case of passing action

Fig. 5: Time-series variation of the shift of the neck point on the xz-plane

made only for a single frame, and not for action inferences that take into account the relationship with the immediately preceding action. For these reasons, we considered that the association between objects and actions by VLM was not successful. In the future, we plan to combine object detection and Large Language Models (LLM) to test methods such as [19] that contextually understand how people interact with objects.

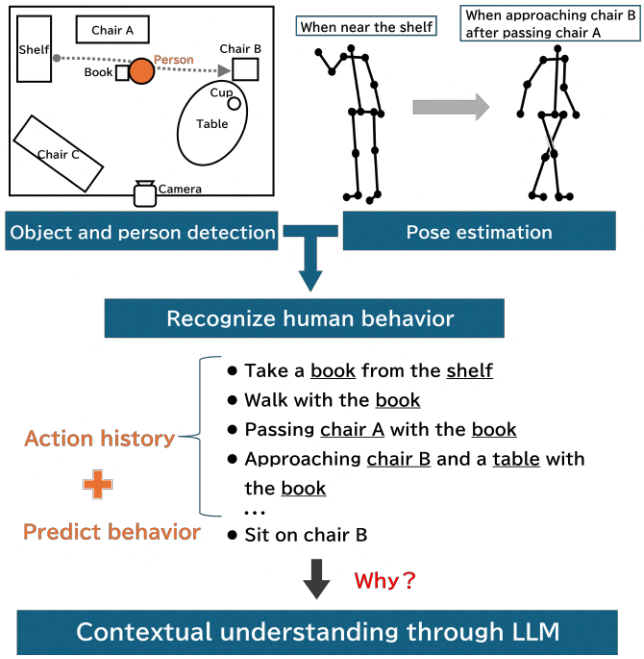


Fig. 6: Framework in future work

In Experiment 2, we focused only on the skeletal point of the neck, which is not easily affected by body movements, and thus we were able to observe how a person approaches an object when he/she has a clear intention to act on it. First, it was found that in Fig. 3a, the person approached the chair further in the vicinity of the chair and drew a trajectory that seemed to go around the chair. In Fig. 4a, it was observed that the motion stopped near the chair and then rapidly approached the chair. In Fig. 5a, it was found that a large difference in the amount of movement between frames tends to be observed when the subject has a clear intention to sit on the chair.

We had thought that it would be difficult to obtain the amount of information necessary for behavior prediction because the skeletal points are not easily affected by body movements, but we were able to observe sufficient differences in two types of behaviors, seating and passing, from the neck skeletal points. On the other hand, we could not read the orientation of the face and body only from the skeletal point information of the neck. Therefore, it is necessary to consider employing other skeletal points in order to estimate the intention to approach a variety of objects.

In Experiment 3, we obtained a model that predicted seating and passing with 91% accuracy. This accuracy value indicates that we were able to employ features suitable for predicting seating and passing behavior. However, only two types of actions were treated in this experiment: seating and passing, and the object focused on was a chair that is used only for sitting. Therefore, it is considered that a feature set with a small effect of body motion, such as a neck skeletal point, was sufficient for predicting simple actions in which body motion is inevitably large. In order to predict the intention of various human actions, it is necessary to take into account the diversity of possible actions from a person to an object, and to select appropriate features based on the observation of such actions.

In Experiment 2 and Experiment 3, the scenario involved a person walking in the direction of a chair and then sitting down, but since people usually interact with a variety of objects in their daily lives, it is not possible to accurately capture the intent of the action by simply looking at the relationship between a single object and a person. Therefore, it is necessary to know which object the person was acting on before acting on an object and how he/she was acting on it. We believe that it is possible to understand and predict the intention of human actions by not only organizing the actions of people to various objects by object, but also by clarifying the relationships among these actions.

In summary, the conclusions of the experiment are as follows:

1) The most likely behavior of a person for each object was selected from the prepared list to predict a behavior in Experiment 1. For improving the prediction accuracy, we plan to test the combined method that

contextually understand how people interact with objects.

2) We were able to observe how a person approaches an object when a person has a clear intention to act on it by focusing on the neck point in Experiment 2. In the future we intend to observe and use other skeletal points.
3) Based on the observation of time-series changes in the spatial positional relationship between the neck point and the object, we identified the features necessary to predict the behaviors of seating and passing, and predicted the behaviors using TCN in Experiment 3. We plan to observe the relationship of each of the other skeletal points and objects to understand the human intention contextually.

Through these preliminary verifications, we are considering the method shown in Fig. 6 for future research.

1) Detects objects and persons and calculates their positions.
2) Detects skeletal points and recognizes posture.
3) Recognize human behavior based on the positions and postures of objects and persons.
4) Add to the action history.
5) Predicts the next object to be accessed and possible actions based on the person's body movements.
6) Generate sentences from the action history and predicted actions, and infer the person's desires and intentions to achieve them by understanding the context using LLM.

VI. CONCLUSION

In this paper, we conducted preliminary experimentss of elemental methods for inferring the intent of human behavior.

In Section III, we tried to recognize and predict possible human actions for each object using VLM. This verification allowed us to confirm the issues involved in organizing human behavior by objects from image information.

In Section IV, we examined the basis of time-series information for predicting the intention of a person to act on an object by using TCN to predict sitting and passing behavior based on time-series information considering the positional relationship between the neck skeleton point and the chair and the frame-to-frame movement of the neck skeleton point. However, since the environment and the behaviors focused on in the preliminary experiments were extremely simple, it is desirable to be able to respond to the diversification of objects and the complexity of human behavior and intentions. Therefore, the following two points should be considered in both the analysis and the study of methods for the basis of time-series information.

- By analizing the relationship between the various actions from a person to each object and between objects.
- By treating the time-series information of other skeletal points selected for each object use as features, a more versatile model can be obtained.

From the above discussion, we devised a framework for future research.

In the future, by inferring from a person's actions the intentions that the person hopes to achieve, it is possible to predict the multiple actions that will take place and the work required to achieve the intentions. Based on the results of these predictions, the robot will be able to proactively assist the person's tasks ahead of time, making the collaboration between the person and the robot more flexible.

REFERENCES

[1] D. Mukherjee, K. Gupta, L. H. Chang and H. Najjaran: "A Survey of Robot Learning Strategies for Human-Robot Collaboration in Industrial Settings," Robotics and Computer-Integrated Manufacturing, Volume 73, 102231, 2022.

[2] D. Wei, L. Chen, L. Zhao, H. Zhou and B. Huang: "A Vision-Based Measure of Environmental Effects on Inferring Human Intention During Human Robot Interaction," IEEE Sensors Journal, vol. 22, no. 5, pp. 4246-4256, 1 March, 2022.

[3] Y. Zhang and T. Doyle: "Integrating intention-based systems in human-robot interaction: a scoping review of sensors, algorithms, and trust," Frontiers in Robotics and AI, Volume 10, 2023.

[4] W. Wang, R. Li, Y. Chen, Y. Sun and Y. Jia: "Predicting Human Intentions in Human–Robot Hand-Over Tasks Through Multimodal Learning," IEEE Transactions on Automation Science and Engineering, vol. 19, no. 3, pp. 2339-2353, July 2022.

[5] I. Jacoby, J. Parron and W. Wang, "Understanding Dynamic Human Intentions to Enhance Collaboration Performance for Human-Robot Partnerships," 2023 IEEE MIT Undergraduate Research Technology Conference (URTC), Cambridge, MA, USA, pp. 1-6, 2023.

[6] N. K. Gajjar, K. Rekik, A. Kanso, R. Müller: "Human intention and workspace recognition for collaborative assembly," IFAC-PapersOnLine, Volume 55, Issue 10, 2022.

[7] N. Robinson, B. Tidd, D. Campbell, D. Kulić, and P. Corke: "Robotic vision for human-robot interaction and collaboration: A survey and systematic review," J. Hum.-Robot Interact., 12(1), February, 2023.

[8] C. Gao, Y. Zou and J.-B. Huang: "iCAN: Instance-Centric Attention Network for Human-Object Interaction Detection," arXiv, 2018. https://arxiv.org/abs/1808.10437.

[9] J. Zhang, J. Huang, S. Jin and S. Lu: "Vision-Language Models for Vision Tasks: A Survey," arXiv, 2024. https://arxiv.org/abs/2304.00685.

[10] E. V. Mascaro, D. Sliwowski,D. Lee: "HOI4ABOT: Human-Object Interaction Anticipation for Human Intention Reading Collaborative roBOTs," Proceedings in Conference on Robot Learning, 2023.

[11] Radford, Alec, et al.: "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.

[12] J. R. Terven and 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, volume 5, pp. 1680-1716, November, 2023.

[13] Intel RealSense Depth Camera D455: https://www.intel.com/content/www/us/en/products/sku/205847/intel-realsense-depth-camera-d455/specifications.html.

[14] LIPSense 3D Body Pose SDK: https://www.lips-hci.com/ja/3d-body-pose-sdk.

[15] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, H. Austin and G. D. Hager: "Temporal convolutional networks for action segmentation and detection," IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[16] S. Bai, J. Z. Kolter, V. Koltun: "An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling," arXiv, 2018. https://arxiv.org/abs/1803.01271.

[17] K. Lyu, H. Chen, Z. Liu, B. Zhang and R. Wang: "3D Human Motion Prediction: A Survey," Neurocomputing, 489, pp.345-365, 2022.

[18] T. Sawahata, A. Moro, S. Pathak and K. Umeda, "Instance Segmentation-Based Markerless Tracking of Fencing Sword Tips," 2024 IEEE/SICE International Symposium on System Integration (SII), pp. 472-477, 2024.

[19] S. Wang, K. -H. Yap, H. Ding, J. Wu, J. Yuan and Y. -P. Tan: "Discovering Human Interactions with Large-Vocabulary Objects via Query and Multi-Scale Detection," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 13455-13464, 2021.