Robust Gesture-based Appliance Control via Operator Identification and Tracking

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

***Abstract*— In this paper, we improve the robustness of a multi-camera gesture recognition system in multi-person situations by identifying and tracking the operator. This system is meant as an intelligent room to operate and interact with surrounding devices based on pointing gestures. In the method, we identify the operator by a hand-raising gesture, followed by tracking using the coordinates of the center of the operator's head and extracting only the operator's whole body. From the experimental results, we confirmed that highly accurate tracking could be performed in a multi-person situation of 2 to 5 persons, and that the success rate of extracting images of the operator's whole body was more than 70%. We also clarified issues in the operator identification process and the extraction process.**

I. INTRODUCTION

Currently, the main means of operating devices in daily life is by using a remote control corresponding to the device. Recently, more intelligent and convenient alternatives for device interaction have been developed. Although voice recognition is widely used to operate devices from any position, the degree of freedom is limited to only verbal instructions and does not support spatial instructions. Therefore, in recent years, research has been conducted on operation methods based on gesture recognition in order to convey spatial instructions to devices intuitively.

Various methods have been studied for device operation with gesture recognition[1]. Many studies of device operation by gesture recognition tie specific gestures to the content of the operation, as in the literature [2]. It is necessary to appropriately set the gestures that are associated with the operation[3], because some gestures have different meanings depending on the culture and context. Thus, some studies have employed gestures in which a user points to objects, which is intuitive and has the same meaning in many situations and cultures[4]-[6].

In addition, a system that assumes only the situation where the user is in front of a device or camera cannot handle various postures and situations in daily life or deal with multiple devices. For this reason, it is effective to be able to operate devices whose location information is unknown in an environment where images of the entire room can be acquired via multiple cameras. To this end, we introduced in our paper [7] a method that combines object detection and skeletal point detection to construct a system that can be operated by arm-pointing at an arbitrarily located home appliance without the need to measure the home appliance or the operator's location in advance. This method achieved good results in terms of recognition rate and usability. However, because it does not define or identify a particular operator, it is vulnerable to gesture recognition and false detection of skeletal points in multi-person situations. Any person can trigger a gesture by mistake. The most important aspect of systems research for human-machine interface is to enable people to trust the system. To achieve this, it is important to achieve both robustness and prevent false operations, and responsiveness by ensuring that recognized operations are performed promptly.

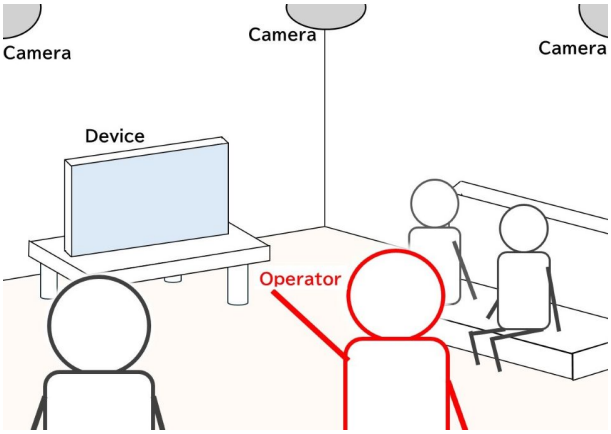

Fig. 1: Conceptual image of device operation by gesture in a multi-person situation. The operator is shown in red.

In this study, we solve the problem of our previous study [7] by identifying and tracking an operator and extracting only the operator's gestures from the tracking results, as shown in Fig. 1. Many studies have been conducted to track a specific person using multiple cameras[8]. Most of them use color and appearance information, which makes it very difficult to distinguish and track people who appear similar. Therefore, it is important to identify and track operators without being affected by appearance information. Although [9] have used background subtraction for target tracking, it cannot deal with situations in which multiple people are moving. [10] used head detection by segmentation to track all persons' trajectory in a dense environment via head tracking. However, this cannot be used to identify a particular operator, which requires performing a gesture to let the system identify the oeprator.

Hence, the purpose of this research is to make the arm-

[1]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]The Department of Precision Mechanics, Faculty of Science and Engineering, Chuo University.

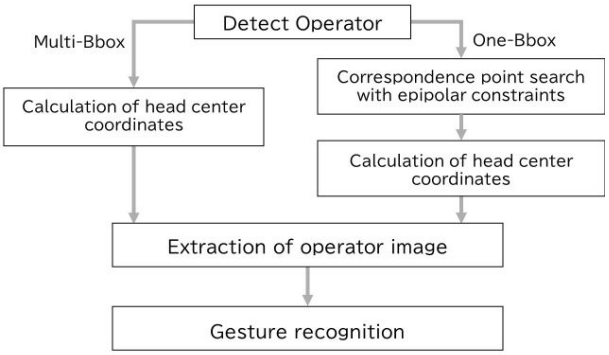

Fig. 2: The flow of this system

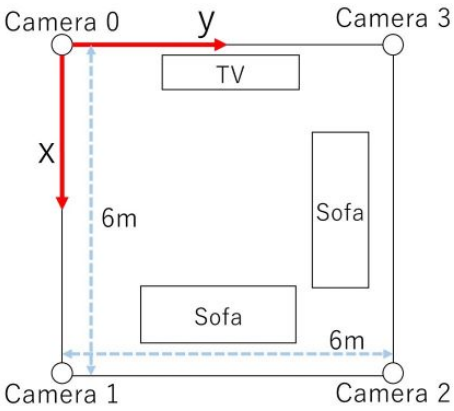

Fig. 3: Layout of the environment

pointing home appliance operation system more robust by identifying and tracking the operator via multi-camera and making it possible to recognize only the operator's gestures. We propose an "Operator Identification Gesture" to select one person in a multi-person situation and track them till the operation is completed. We do not use color information for identification and tracking. Fig. 1 shows a conceptual diagram of this study which shows a person identified as the operator in a multi-person situation.

II. PROPOSED METHOD

A. System Overview

The system environment is captured from cameras installed in the four corners of the ceiling as shown in Fig. 3. The proposed method focuses only on the operator's gestures even in multi-person situations, as shown in Fig. 1. The internal and external parameters of these cameras are known, and the projection matrix is obtained by calibration when the system is started. First, the system detects a hand-raising gesture as an "Operator Identification Gesture" and identifies the operator in multiple camera images via bounding box fusion. Then, the 3D coordinates of the identified operator's head center are calculated, and the operator is tracked using this point. At the same time, the system extracts images of the operator's entire body using this point as a reference. Finally, a skeleton point is extracted from the extracted full-body image of the operator, and the 3D coordinates of the calculated skeleton point are used to recognize gestures. The brief flow of the proposed method is shown in Fig. 2.

YOLO[11] is used to detect operators via hand-raising gestures, people's full bodies, their heads, and devices, followed by fusing the detection results. The operator is identified by detecting the hand-raising gesture for several frames in a row, after which they can start operating the home appliance. The detected heads are used for Kalman filter based tracking. Finally, the system uses skeletal point extraction with OpenPose[12] to operate the home appliance by recognizing the arm gesture.

B. Object Detection

Our system detects the following:

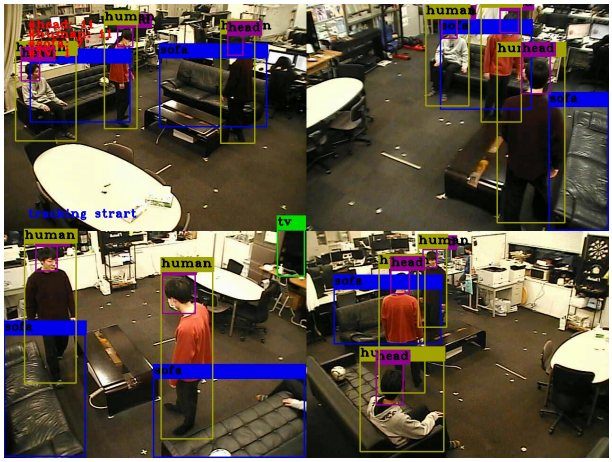

Fig. 4: Object detection results: bounding boxes

1) "operator" detects the hand raising gesture to identify the operator bounding box
2) "human" detects all the humans in the environment
3) "head" detects all the humans' heads for Kalman filter-based tracking, as the head is usually always visible.

We implement YOLOv4[13] detectors: Detector A, which detects "operator", and Detector B, which detects "human" and "head". Detector B also includes"TV" and "sofa" for detecting environmental objects and interactive devices. In this research, we chose a "TV" as a device to be interacted with. Detector A was annotated with photographs of the system's space, and the data was expanded to obtain a dataset of 1,235 pairs. Of these, 987 pairs were used for training, with an average fit rate of 91.5 percent. Detector B was annotated with 122 photos taken in the system's space, and the data was expanded to obtain 1128 datasets, of which 902 were used for training. Of these, 902 were used for training and 226 for validation. The average fit rates for "human", "head", "TV", and "sofa" were 89.6, 98.3, 97.8, and 98.6 percent, respectively. Using these detectors, bounding boxes are detected as shown in Fig. 4.

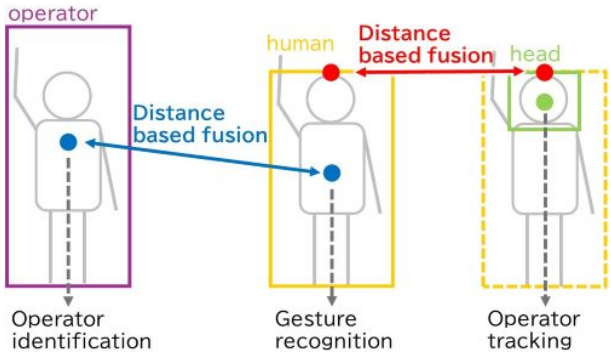

Fig. 5: Operator identification via bounding box fusion

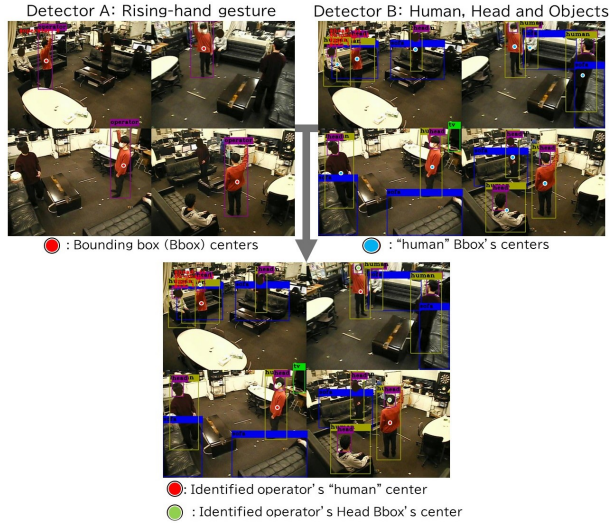

Fig. 6: The fusion of detection results

C. Identification and Tracking of Operators via Bounding Box fusion

As shown in Fig. 6, the operator is identified and by fusing "operator" and "human", followed by fusion with "head" for operator tracking.

The operator is determined when the "operator" bounding box center does not move and the detection is maintained for 9 consecutive frames.

2D distances between "operator" and "human" are calculated, and the "human" bounding box with the closest distance is identified as matching the"operator"'s bounding box. Following this, "human" and "head" are fused by calculating the horizontal distances between the upper edge centers of "human" and upper edge centers of "head" bounding boxes and choosing the one with smallest distance as the operator's head. This gives us the uniquely identified operator and the coordinates of their head. An overview of operator identification using bounding box fusion is shown in Fig. 5

For tracking, we triangulate the 3D coordinates of the operator's head from multiple cameras [14]. By doing so, we can restore these 3D coordinates to 2D coordinates in each camera image using the perspective projection matrix

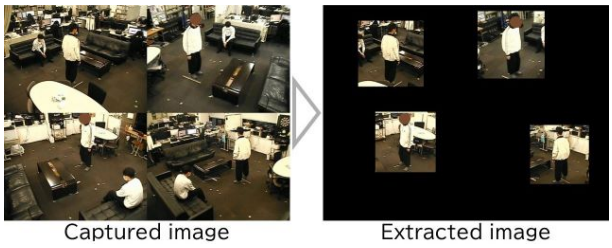

Fig. 7: Operator extraction results

and locate the operator's head even in camera images where the "operator" could not be detected or was occluded. This allows for highly accurate tracking. On the other hand, if the "operator" is detected in only one camera image, the corresponding "head" bounding box is identified in a corresponding camera image using epipolar constraints and is considered to be the operator's head. We choose corresponding cameras as Camera 0 and Camera 2, Camera 3 and Camera 1 (Fig. 4.

After the operator is identified, the "head" bounding box center 2D coordinates are used for operator tracking. In order to deal with situation where the operator and other people pass each other, a Kalman filter is used once every 5 frames to correct the predicted values.

D. Correction by Kalman filter

The 3D coordinates of the head center calculated based on the principle of stereo vision are stored and the coordinates after the movement are predicted using a Kalman filter. The predictions obtained by the Kalman filter are always restored to the 2D coordinates in each camera using pre-calibrated projection matrices, and once every five frames, the restored predictions are compared with the values detected in the acquired images. If the difference exceeds a threshold value, the predicted value is given priority as the true value of the operator's head position in that frame. Empirically, this threshold is set to 40 pixels. This correction method is expected to maintain robust tracking when the operator and other people pass in front of each other and to eliminate large noise due to false positives and measurement errors.

E. Recognition of Gestures

Once the operator is properly tracked and identified, we detect their skeletal points by extracting the operator alone from the acquired image as shown in Fig. 7. To enable gesture recognition even when the operator's outstretched arm is outside the bounding box, the extraction is performed in a square region with a side length equalling the height of the operator's "human" bounding box.

This study uses gesture recognition with skeletal point detection by OpenPose[12]. The algorithm is used to obtain the 2D coordinates of the skeltal points of the elbow and wrist from an image showing only the operator, as shown in Fig. 8. OpenPose is a CNN-based algorithm that performs person pose estimation by cascading heatmaps and Part Affinity Fields. In this system, skeleton points with a confidence

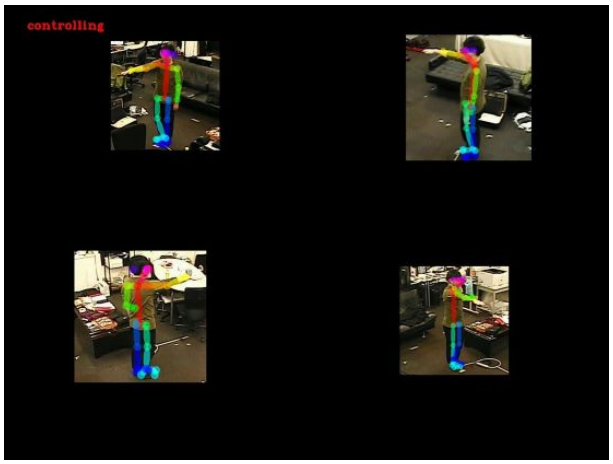

Fig. 8: Operator gesture recognition

level of 0.60 or higher were adopted to prevent inaccurate coordinates with a low confidence level from being used in the calculation.

The gesture recognition method is the same as [7]. As shown in Fig. 4, multiple 2D coordinates of the bounding box center are gotten for the same device for each camera. The 3D coordinates of devices' center are calculated by triangulating the 2D center points of these bounding boxes. The 3D point is considered as the device's 3D center and is denoted (x_a, y_a, z_a). In addition, the 3D coordinates of the right elbow and right wrist are calculated by triangulating using the 2D coordinates of them on the extracted image. The 3D points of right elbow and right wrist are denoted by (x_e, y_e, z_e) and (x_w, y_w, z_w).

The vector $\vec{v_p}$ from the right elbow to the right wrist and the vector $\vec{v_a}$ from the right elbow (x_e, y_e, z_e) to the device's center (x_a, y_a, z_a) can be estimated by (1).

$$\vec{v_p} = \begin{bmatrix} x_e - x_w \\ y_e - y_w \\ z_e - z_w \end{bmatrix}, \quad \vec{v_a} = \begin{bmatrix} x_a - x_e \\ y_a - y_e \\ z_a - z_e \end{bmatrix} \tag{1}$$

The angle θ between $\vec{v_p}$ and $\vec{v_a}$ is used to judge device control. If θ is less than or equal to the threshold value θ_{th} and the instruction state is maintained for 6 frames continuously, the device is turned on. θ is calculated by (3). In addition, it is considered that the ease of pointing correctly to the center of the device changes by the distance between the operator and the device. Therefore, the threshold θ_{th} is varied depending on the distance $|\vec{v_a}|$.

$$\theta = \arccos\left(\frac{\vec{v_p} \cdot \vec{v_a}}{|\vec{v_p}||\vec{v_a}|}\right) \tag{3}$$

Once the pointing gesture is identified, the device on/off switch is toggled using a Nature Remo smart remote control device and operation is considered complete. The operator identification is also reset.

III. EVALUATION

Experiments were conducted on operator identification, tracking, and operator image extraction to confirm the usefulness of the method in multiple-person situations. Experiments were conducted in an offline environment. The operator's and non-operators' standing positions were not specified to create a natural multi-person situation. The operator was asked to raise their hand to signal the operator identification gesture. Following this, they put their hand down and walked around in the experimental environment. No time was specified for the hand raising gesture. We evaluated the system via five recorded videos. Permission was obtained from the subjects for recording in this evaluation experiment.

First, in Experiment 1, we evaluated the accuracy of identifying the operator in multiple-cameras by the hand-raising gesture. Next, in Experiment 2, we used the videos from Experiment 1 in which the operator was correctly identified to evaluate tracking. In addition, we confirmed the accuracy with which the operator's entire body was correctly extracted by the operator head tracking method.

A. Experiment 1: Identification of operator

This experiment evaluates the accuracy of the operator identification method based on the bounding box fusion method. When a hand-raising gesture was not recognized by any of the cameras, a time period was set to wait for the judgment to be reset. If this time was too short, the operator could not be identified at all, so we set it to 2 seconds in this experiment.

As shown in Table I, four of the five videos succeeded and one failed. Of these successful results, some non-operators were dressed in clothing of the same color as the operator's clothing, both top and bottom, but this did not affect the identification.

TABLE I: Success (S) or failure (F) of operator identification

Number of persons in room	2	3	3	5	5
Operator identification	S	S	S	S	F

In the video in which the operator identification failed, the search for the correspondence point at the center of the head failed as shown in Fig. 9. It can be seen that the operator's head recognized by Camera 0 was not captured by Camera 2, indicating that the search for the correspondence point failed. Therefore, when there is only one camera in which the operator is detected, instead of searching for correspondence points from cameras in a correspondence relationship, it is necessary to try the method of searching for correspondence points from the camera with the most "head" bounding boxes.

When an operator cannot be identified in other cameras even though the one is detected by one camera, interaction between the operator and the system may solve the problem. It is considered that the system can identify the operator

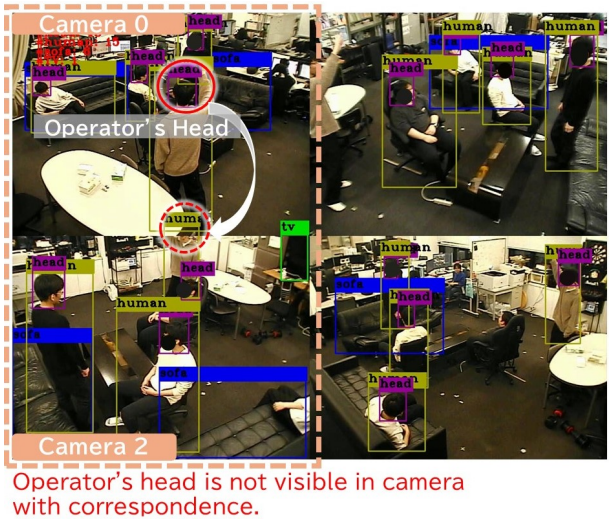

Fig. 9: Failure to identify operator: operator's head is not visible in Camera 2

more reliably by providing feedback to the operator on the recognition status on the system side. When the system is confused, it is easiest and most effective to first confirm with the operator about the intention to operate. If the operator does intend to operate the system, the system can inform the operator that the hand-raising gesture detection is not working well, and the operator can cooperate by changing the their posture to make it easier to recognize the gesture.

B. Experiment 2: Tracking and extraction of operator

Experiment 2 was conducted on the four videos from Experiment 1 in which operator identification was successful in multiple cameras. A bird's-eye view of the trajectory centered on the operator's head is shown in Figs. 10 to 13. The blue line is the trajectory obtained by triangulation based on the detection results, and the orange line is the trajectory predicted by the Kalman filter. It can be seen that no false trajectories were observed in which a person other than the operator was mistakenly identified as the operator. The operator could be consecutively tracked even when the operator changed the direction of movement or passed another person. The tracking accuracy was not affected by the presence of a person wearing the same color clothing on both the upper and lower sides, as no color information was used for tracking. In addition, we confirmed that the Kalman filter correction was able to cope with changes in the speed of the operator's movement. One of the reasons for the good tracking results was the use of the head as a reference point. The head has a small area among the body parts, and the bounding box center does not easily change significantly, which is thought to have contributed to the successful tracking. In addition, the Kalman filter correction is considered to have enabled the tracking to cope with the operator/people passing each other.

Furthermore, the operator's whole body was extracted based on the one's head tracking results. Here, the operator iwas extracted from a square region with the height of the operator's "human" bounding box as one of its sides. The 2D coordinates of the center of the operator's head were compared with the 2D coordinates of the center of the "human" bounding box of all persons, and the "human" bounding box with the closest x-axis distance was considered to be the operator's "human" bounding box. In the extracted images shown in Fig. 7, a failed frame was defined as one in which one or more cameras extracted a person other than the operator. The number of failed frames was counted by visually checking each extracted image frame-by-frame. The number of successful frames was calculated by subtracting the number of failed frames from the total number of frames in the video after the operator was identified, and the percentage of successful frames was defined as the success rate of extraction. The success rate of extracting the operator's region in the successfully operator identification videos is shown in Table II. The failed frames out of all frames are shown in red in the bars on the top in Figs. 10 to 13, together with an overhead view of the tracking results.

We expect that the success rate of extracting only the operator's body would decrease as the frequency of passing each other increases with the number of people present around the operator. Therefore, it is reasonable to assume that the extraction accuracy will decrease as the number of people in the room increases, as can be seen in Table II. However, even the lowest extraction accuracy was about 70%. In addition, the triangulation of the skeletal points was feasible because the operator area was successfully extracted by more than one camera in every failed frame. In order to improve the accuracy of gesture recognition in the future, two approaches can be considered: (1) improving the extraction accuracy and (2) selecting camera pairs with higher accuracy for 3D triangulation.

In order to improve the success rate of extraction, the mapping conditions between the operator's "head" bounding box and the "human" bounding box must be reviewed. In this experiment, the mapping was based on the horizontal distances on the image. In the future, it is expected that the use of 3D coordinates information will enable more robust mapping conditions to be established.

It is also important to set conditions for the skeleton points to be employed. If all the cameras produce correct extraction results, the confidence level of the skeleton points can be used to perform gesture recognition with high accuracy. It will be effective to perform 3D matching only among the skeletal points with the highest confidence level in the same region.

TABLE II: Operator extraction success rate [%]

Number of persons in room	2	3	3	5
Extraction accuracy	93.8	89.1	88.4	70.7

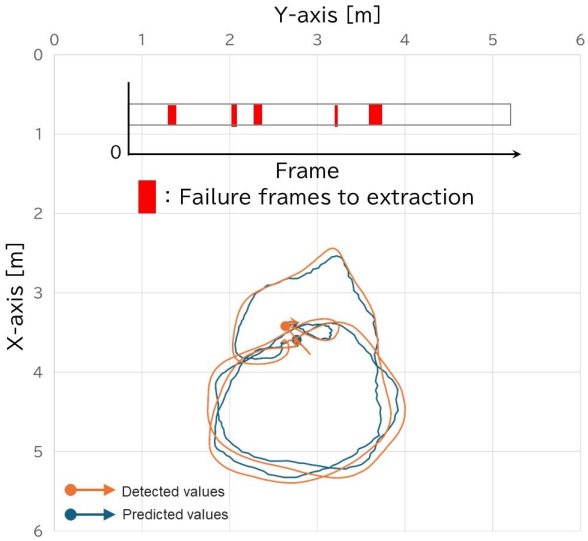

Fig. 10: 2 people situation

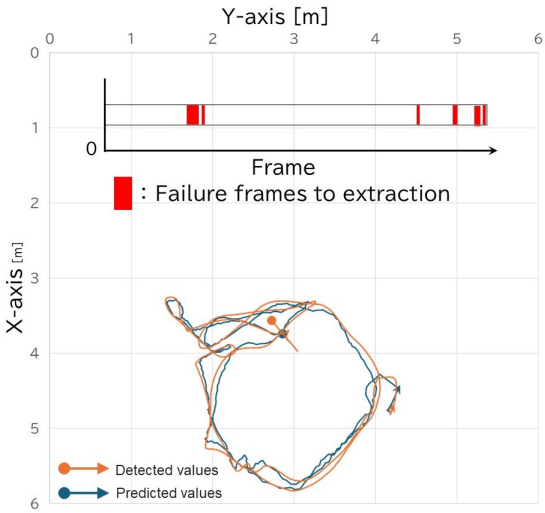

Fig. 11: 3 people situation

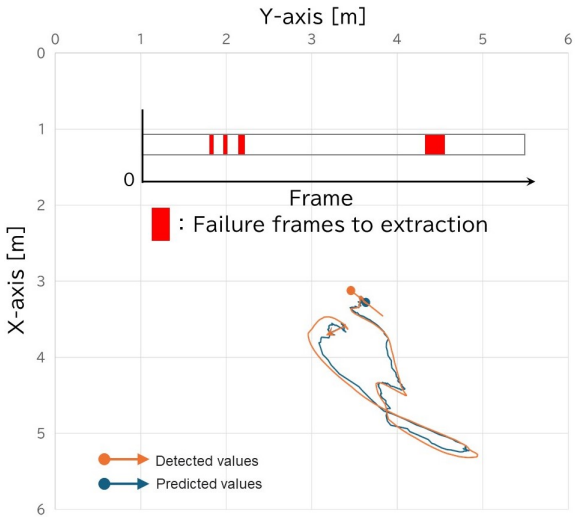

Fig. 12: 3 people situation

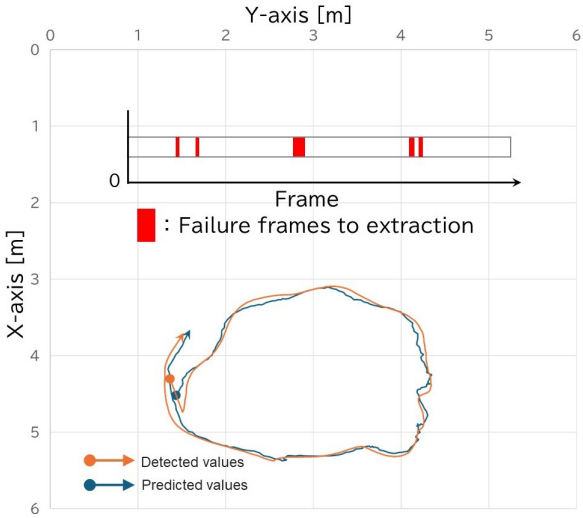

Fig. 13: 5 people situation

IV. CONCLUSIONS

In order to achieve robustness for pointing gesture based device interaction systems in multi-person environments, this study solved the problem of identifying and tracking the operator. The operator was identified using a hand raising gesture and tracked by fusion of bounding box detection results. Experiments showed that the proposed method was able to identify and track the operator with high accuracy except in situations where the operator's head could not be detected. We also confirmed that the proposed method can extract operator images with more than 70% accuracy in a multi-person environment of 2 to 5 persons.

In order to make the system more robust and sensitive, it is important to design it with the cooperation of the operator in mind. When the detection of hand-raising gestures fails in the process of operator identification, it is necessary for the system side to provide feedback to the operator on his/her own situation. This feedback will allow the operator to change the one's posture and improve the accuracy of operator identification. There is also a need to improve the gesture recognition rate in extracting whole body regions.

Future prospects also include the development of research that assigns tasks to moving devices such as robot vacuum cleaners by directing them to an area using a pointing gesture. By identifying the operator, it will be possible to operate personal devices.

REFERENCES

[1] R. Z. Khan and N. A. Ibraheem, "Survey on Gesture Recognition for Hand Image Postures", International Journal of Computer And Information Science, vol. 5, no. 3, 2012.

[2] Y. Muranaka, M. Al-Sada and T. Nakajima, "A Home Appliance Control System with Hand Gesture based on Pose Estimation," 2020 IEEE 9th Global Conference on Consumer Electronics (GCCE), Kobe, Japan, pp. 752-755, 2020.

[3] K. Deshpande, V. Mashalkar, K. Mhaisekar, A. Naikwadi and A. Ghotkar, “Study and Survey on Gesture Recognition Systems,” 2023 7th International Conference On Computing, Communication, Control And Automation, India, 2023.

[4] J. R. B. Bodollo, J. Daniel V. Cortez, E. R. P. Maraya, E. V. Navarro, R. Q. L. Saquing and R. E. Tolentino, “Selection of Appliance Using Skeletal Tracking and 3D Face Tracking for Gesture Control Home Automation,” 2019 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE), Bangalore, India, pp. 1-7, 2019.

[5] A. I. D. Viaje, P. S. Bernardo, K. N. Manuel, G. M. Pacheco, K. - R. C. Barroma and R. E. Tolentino, “Selection of Appliance Using Skeletal Tracking of Hand to Hand-tip for a Gesture Controlled Home Automation,” 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, pp. 575-580, 2020.

[6] M. A. Iqbal, S. K. Asrafuzzaman, M. M. Arifin and S. K. A. Hossain, “Smart home appliance control system for physically disabled people using kinect and X10,” 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV), Dhaka, Bangladesh, pp. 891-896, 2016.

[7] M. Yokota, S. Majima, S. Pathak and K. Umeda, “Intuitive Arm-Pointing based Home-Appliance Control from Multiple Camera Views,” 2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), 2023.

[8] R. Iguernaissi, D. Merad, K. Aziz, P. Drap, “People tracking in multi-camera systems: a review,” Multimedia Tools and Applications, 78, 10773–10793, 2019.

[9] T. T. Santos, and C. H. Morimoto, “Multiple camera people detection and tracking using support integration,” Pattern Recognition Letters 32.1, pp. 47-55, 2011.

[10] R. Eshel, and Y. Moses, “Tracking in a Dense Crowd Using Multiple Cameras,” International Journal of Computer Vision, 88, pp. 129–143, 2010.

[11] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 779-788, 2016.

[12] Z. Cao, G. Hidalgo, T. Simon, S. -E. Wei and Y. Sheikh, “OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields,” in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 1, pp. 172-186, 2021.

[13] A. Bochkovskiy, C. Wang, H. Mark Liao, “YOLOv4: Optimal Speed and Accuracy of Object Detection,” arXiv preprint arXiv:2004.10934, 2004.

[14] R. Hartley, A. Zisserman, “Multiple View Geometry in Computer Vision,” Cambridge University Press, second edition, 2004.