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

Lightweight Hand-Waving Action Recognition Using Reservoir Computing in a Cafeteria Environment*

Kosei Isomoto[1], Soma Fumoto[2], Ryohei Kobayashi[1], Yuichiro Tanaka[1,3] and Hakaru Tamukoh[1,3]

***Abstract*— Owing to the global labor shortage and increasing need for operational efficiency, the adoption of service robots is advancing rapidly. These robots must recognize human action to understand human intention and respond appropriately based on that understanding. The action recognition systems embedded in robots need to be lightweight to operate efficiently with limited computational resources. Reservoir computing (RC) is one of the solutions for lightweight action recognition systems. Yamaguchi et al. proposed an RC-based hand-waving recognition system; however, the system cannot process multiple persons simultaneously and works only when one person is in the image. Therefore, this study proposes a lightweight hand-waving recognition system that integrates OpenPose, StrongSORT, and RC to work in complex environments with multiple individuals. Experimental results demonstrated the effectiveness of the proposed system in processing multiple people simultaneously in a crowded environment and accurately recognizing hand-waving actions with 90.75% accuracy. We also confirmed that the proposed system can process data at 24-26 FPS. We demonstrated that the proposed system can perform real-time processing. In addition, the robot with the proposed system recognized hand-waving actions in the "Restaurant" task of RoboCup@Home 2024 and obtained the second-place score.**

Fig. 1: "Restaurant" task in Robocup@Home

I. Introduction

Service robots are gaining attention as a solution to the global labor shortage and the need for increased work efficiency. In particular, service robots are rapidly expanding in the medical and food industries. Consequently, the global market for service robots continues to grow, already exceeding 2 trillion yen in 2023, and is expected to reach 4.71 trillion yen by 2030 [1].

In addition to the rapid introduction into commerce, home service robots, which assist in daily life in the home are also gaining attention. They are expected to perform various household tasks efficiently, such as caring for the elderly, supporting children, and helping with housework.

RoboCup@Home [2], an international competition for the research and development of home service robots, involves various tasks in settings that simulate a home environment. In the "Restaurant" task of this competition, the robot takes and serves the orders of several customers in a restaurant. The robot recognizes the calls and hand-waving actions of the customer, goes to the table, and takes their order. Additionally, it carries the ordered items from the kitchen. Fig. 1 shows the flow of tasks.

In particular, human action recognition in complex environments is challenging and requires advanced recognition skills. One action recognition based on deep learning is a method using convolutional neural networks and long short term memory [3], which has achieved high recognition accuracy. However, such deep learning-based approaches are computationally expensive and require processing on a graphic processing unit (GPU). Using GPU improves the computational speed, however, it increases the battery consumption of the robot and limits the running time.

In contrast, action recognition methods [4] that combine lightweight human pose estimation [5], Deep SORT [6], and ST-GCN [7] reduce model weight by using a skeletal estimation model as a base. However, ST-GCN performs both graph and temporal convolutions and uses a multilayered structure for inference; thus, the problem of using many

*This research is based on results obtained from a project, JPNP16007, commissioned by the New Energy and Industrial Technology Development Organization. This work received support from JSPS KAKENHI Grant Numbers 23H03468, as well as from JST ALCA-Next Grant Number JPMJAN23F3.

[1]Kosei Isomoto, Ryohei Kobayashi, Yuichiro Tanaka, and Hakaru Tamukoh are with Graduate School of Life Science and Systems Engineering, Kyushu Institute of Technology, 2-4 Hibikino, Wakamatsu, Kitakyushu 808-0196, Japan isomoto.kosei778@mail.kyutech.jp, kobayashi.ryohei621@mail.kyutech.jp, tanaka-yuichiro@brain.kyutech.ac.jp, tamukoh@brain.kyutech.ac.jp.

[2]Soma Fumoto is with Faculty of Environmental Engineering, University of Kitakyushu, 1-1 Hibikino, Wakamatsu, Kitakyushu 808-0135, Japan c1531049@eng.kitakyu-u.ac.jp.

[3]Yuichiro Tanaka and Hakaru Tamukoh are with Research Center for Neuromorphic AI Hardware, Kyushu Institute of Technology, 2-4 Hibikino, Wakamatsu, Kitakyushu 808-0196, Japan tanaka-yuichiro@brain.kyutech.ac.jp, tamukoh@brain.kyutech.ac.jp.

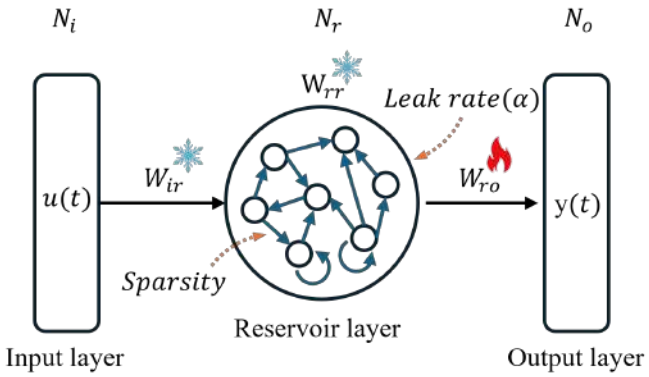

Fig. 2: Structure of reservoir computing

computational resources has not been solved. Therefore, more memory-efficient and lightweight models are required for implementation in robotic systems.

In consideration of the aforementioned issues, this study aims to realize a lightweight and accurate action recognition system using reservoir computing (RC) [8]. In particular, we propose a lightweight hand-waving recognition system that can work efficiently in complex environments such as a restaurant where multiple people are performing various actions.

II. Related works

RC is a computational framework specialized for processing time-series data, and models such as echo state networks (ESNs) [9] and liquid state machines [10] are typical examples of RC models. As shown in Fig. 2, RC has input, reservoir, and output layers. The reservoir layer consists of sparsely connected nodes that map input data to high dimensional spaces nonlinearly. The output layer analyzes patterns from the high dimensional state. The connection weights from the input layer to the reservoir layer ($W_{ir} \in \Re^{N_u \times N_r}$) and within the reservoir layer ($W_{rr} \in \Re^{N_r \times N_r}$) are randomly generated and fixed even during the training. Equation (1) represents an update of a reservoir state ($\mathbf{s}(t+1)$), and the leak rate (α) determines how well the state computed from the new input affects the previous state.

$$\mathbf{s}(t+1) = (1-\alpha)\mathbf{s}(t) + \alpha \tanh(W_{rr}\mathbf{s}(t) + W_{ir}\mathbf{u}(t)) \quad (1)$$

Only the weights ($W_{ro} \in \Re^{N_r \times N_o}$) from the reservoir layer to the output layer are updated during the training.

The overall structure is simple, and the training does not require much computational resources, making it lightweight and easy to use even in environments with limited computational resources.

Yamaguchi et al. proposed a hand-waving recognition system using an ESN [11]. They extracted skeletal information in real-time using MediaPipe [12] and input the difference between the x-coordinates of the left and right wrists and the x-coordinate of the nose, divided by the difference between the x-coordinates of the left and right shoulders into the ESN. The ESN extracts hand movement features by capturing temporal patterns in these inputs and then classifies the hand-waving or not-waving.

The system does not process multiple people simultaneously and only works when one person is in the image because the system lacks the ability to identify individuals across time. When normalizing utilizing the difference in the x-coordinates of the left and right shoulders, the shoulder width visible to the camera changes if the target person's orientation is at an angle, which can also decrease the accuracy.

III. Proposed method

The method proposed by Yamaguchi et al. does not support simultaneous processing for multiple people gestures and works only when one person is in the image. So, this study proposes a multi-person hand-waving recognition system using RC to recognize diverse behaviors in a cafeteria environment.

Fig. 3 shows an overview of our proposed system using OpenPose [13], StrongSORT [14], and RC. OpenPose extracts human joint points (keypoints) from an input RGB image. These keypoints include joints such as the nose, shoulders, and wrists. Based on these coordinates information, we obtain the minimum and maximum x and y coordinates of the keypoints and form a bounding box around the person.

We then input the formed bounding box and RGB image into StrongSORT, an object-tracking algorithm that assigns a unique tracking ID to each person in each frame. This allows us to map keypoints based on the tracking ID, making it possible to track the same person from frame to frame continuously. We store the tracking ID, two-dimensional coordinates of the keypoints, and distance to each keypoint obtained from the depth image in the tracking table.

For hand-waving recognition, we use the input information ($\mathbf{u}(t)$) calculated according to Equation (2), where we perform the normalization by multiplying the difference between the x-coordinates of the nose (x_{nose}) and the left and right wrists (x_{l_wrist}, x_{r_wrist}) by the distance from the camera to the keypoints ($Distance$). Because light reflection may prevent us from obtaining the distance of the keypoints, we prioritize using the distance of the nose, shoulders, and wrists. If there are no keypoints with distance, we use the distance from the previous frame. We expect this method to form the input information independently of the person's orientation.

$$\mathbf{u}(t) = \begin{bmatrix} (x_{r\_wrist} - x_{nose}) * Distance \\ (x_{l\_wrist} - x_{nose}) * Distance \end{bmatrix} \quad (2)$$

Using the normalized input information, the hand-waving recognition module, which is an RC model, performs a binary classification of whether the hand is waving or not. Finally, the proposed system stacks the classification results in a result list.

A *likelihood* of hand-waving output by the proposed system is calculated by dividing the number of "true" (N_{true})

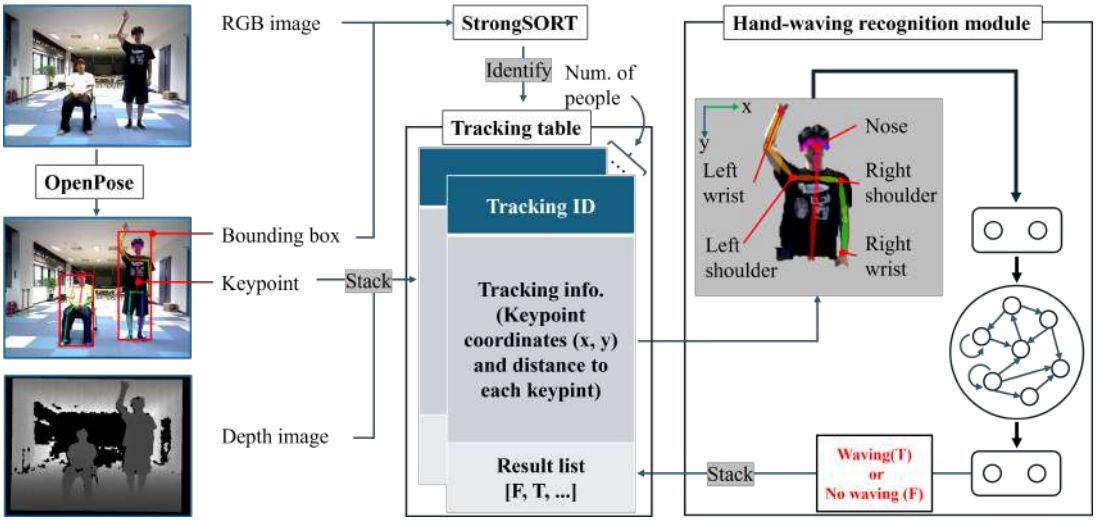

Fig. 3: Overview of the proposed system

outcomes by the total number of items in the result list (N_{length}), as shown in Equation (3). If the likelihood output by the proposed system exceeded the threshold value of 0.5, the individual is classified as hand-waving.

$$likelihood = \frac{N_{true}}{N_{length}} \tag{3}$$

IV. Experimental Setting

We set five experiments to evaluate the recognition performance of the proposed system in static and dynamic environments as well as its processing time. We developed datasets consisting of three different distances. We evaluated the impact of the normalization methods of input data and the distance of the dataset on recognition performance in Experiments (I) and (II). In Experiment (III), we tested whether hand-waving recognition can be performed accurately in a crowded environment with multiple people. In Experiment (IV), we measured the processing time per frame to verify if real-time processing is feasible. Finally, in Experiment (V), we used the proposed system in a competition to verify its effectiveness.

A. Dataset

We created a video dataset for a cafeteria environment, including waving and other actions. The hand-waving action dataset included videos of 32 types of actions per person based on a combination of hand-waving direction (right or left), hand-waving width (large or small), hand-waving speed (fast or slow), and body orientation (front, right, left, or back) as shown in Fig. 4.

For the no-hand-waving actions, we defined four types of postures: hands up, hands on the head, hands down, and walking for four types of body orientations. Based on the combination of these postures and walking movements (left-to-right and right-to-left), the dataset consisted of 28 videos per person, as shown in Fig. 5.

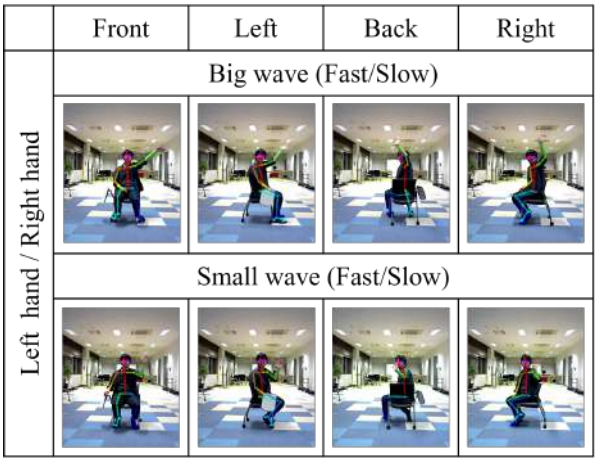

Fig. 4: Dataset with hand-waving

We acquired these data at distances of 3, 5, and 7 m, resulting in 180 videos in the dataset. Each data is approximately 5 s (24–26 FPS).

B. Experiment (I): Evaluation of hand-waving recognition with various normalization methods using multi-distance dataset

We conducted an experiment to evaluate whether the proposed system could recognize hand-waving. We compared the accuracy of the method without any normalization for the inputs, the normalization methods based on shoulder width (proposed by Yamaguchi et al. [8]), and the normalization methods based on distance (ours).

The proposed system utilized an ESN for the hand-waving recognition module. The experiment was conducted according to the following procedure:

1) Train the proposed system using variations of 3, 5, and 7 m from the dataset of three individuals.
2) Validate the trained system using 3, 5, and 7 m from the dataset of one individual that is different from those

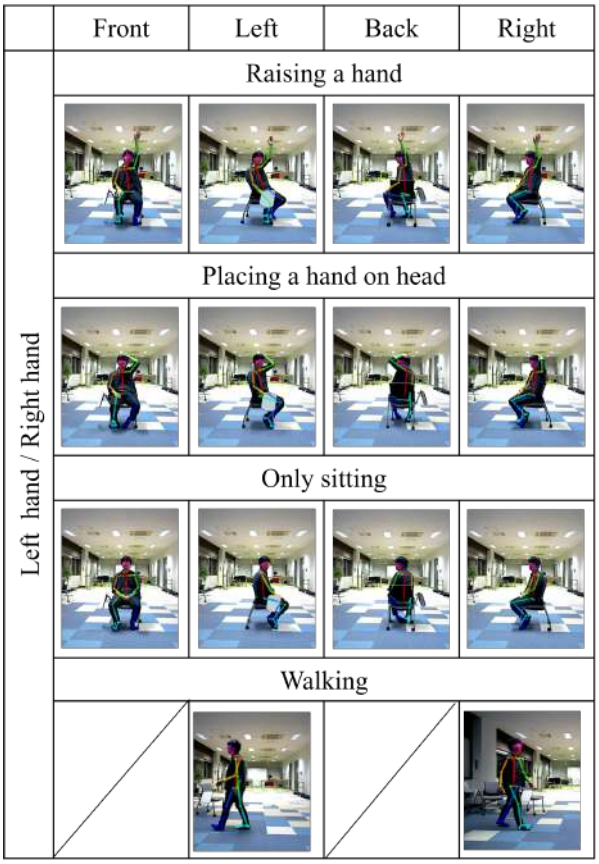

Fig. 5: Dataset without hand-waving

used for training.

3) Repeat steps 1) and 2) 500 times to search for the optimal parameters. The optimization was performed using Optuna [15] to search for the ESN parameters listed in Table 1, where the parameter determining the connection weight density in the reservoir layer was defined as $Sparsity$.
4) Test the system trained with the optimal parameters using 3, 5, and 7 m from the dataset of one individual that is different from those used for validation and training.

We used accuracy, recall, precision, and F-measure during this experiment to evaluate the proposed system.

C. Experiment (II): Evaluation of hand-waving recognition with various normalization methods using single-distance dataset

In this experiment, we verified the effectiveness of the normalization method by including test data with distances different from those used during training and validation. We trained and validated the proposed system using only 3 m data from the same person dataset used for training and validation in Experimental condition (I). Subsequently, we tested the optimized system with data at distances of tested with data from 3, 5, and 7 m distances. Training and parameter exploration were performed using the same procedure as in Experimental condition (I), and accuracy, recall, precision, and F-measure were evaluated.

Ex. 1) A hand-waving person

Ex. 2) Two hand-waving people

Fig. 6: Samples of hand-waving test data (Mark each person waving hand with a red rectangle)

D. Experiment (III): Testing hand-waving recognition in crowded environments

To evaluate whether the proposed system with the normalization method using shoulder and the normalization method using distance could recognize hand-waving actions in a crowded environment, we prepared test data in an environment where the "Restaurant" task of RoboCup@Home 2024 [16] was performed. As shown in Fig. 6, the recordings included various people, with a minimum of one person waving their hand while others were engaged in random activities. The videos were split into 5-s clips to generate 28 test data.

The videos had a frame rate of 11 to 15 FPS and contained between 55 and 75 frames. Within each video, we defined the people who could be detected in at least 20 frames using OpenPose and StrongSORT as the individuals appearing in the video and counted them as the total number of attendees.

TABLE I: Parameter search range of ESN

Parameter	Lower	Upper	Resolution
Leak rate of reservoir layer (α)	0.7	1.0	0.01
Scale of input connection weights (W_{ir})	0.001	1.0	0.001
Scale of recurrent connection weights (W_{rr})	0.01	2.0	0.001
Sparsity of reservoir layer ($Sparsity$)	0.01	0.5	0.01

TABLE II: Experimental result (I): Performance comparison of various normalization methods using the multi-distance dataset (Train: 3, 5, and 7 m data, Test: 3, 5, and 7 m data)

Normalization	Accuracy (Validation)	Accuracy (Test)	Recall	Precision	F-meature
None	**0.840**	0.723	**0.771**	0.733	0.751
Shoulder	0.822	**0.739**	0.760	**0.753**	**0.757**
Distance (Proposed)	0.800	0.728	0.740	0.747	0.744

TABLE III: Experimental result (II): Performance comparison of various normalization methods using the single-distance dataset (Train: 3 m data, Test: 3, 5, and 7 m data)

Normalization	Accuracy (Validation)	Accuracy (Test)	Recall	Precision	F-meature
None	0.761	0.667	0.656	0.700	0.677
Shoulder	**0.833**	**0.733**	**0.740**	0.757	**0.747**
Distance (Proposed)	**0.833**	0.717	0.688	**0.759**	0.721

TABLE IV: Results of hand-waving recognition in an environment with multiple people

			Shoulder		Distance (Proposed)	
	Number of people in videos	Number of people waving hand	Number of correctly recognized	Number of incorrectly recognized	Number of correctly recognized	Number of incorrectly recognized
Total	227	34	23	11	21	8

We classified attendees as a positive case with a likelihood exceeding 0.5 and evaluated it based on the number of individuals correctly and incorrectly identified as waving.

E. Experiment (IV): Measuring real-time performance of the proposed system

To evaluate the real-time performance of the proposed system, we measured the processing time per frame. Data from the dataset created in subsection IV-A. Dataset was used, and the processing time was determined by averaging the time of 20 hand-waving recognitions. The processing of OpenPose and StrongSORT was performed on a GPU, while the processing of RC was executed on a central processing unit (13th Gen Intel Core i9-13900H).

F. Experiment (V): RoboCup@Home 2024 "Restaurant"

We participated in RoboCup@Home 2024 held in Eindhoven, the Netherlands, in July 2024 and operated and evaluated the proposed system in a "Restaurant" task. In this competition, we used the Human Support Robot [17] developed by Toyota Motor Corporation, which was adopted as the standard machine.

The "Restaurant" task consists of the following cycles: find a customer, take an order, tell the order, and serve the order. The evaluation is based on the score after two cycles.

V. Experimental result

A. Experiment (I): Evaluation of hand-waving recognition with various normalization methods using multi-distance dataset

Table 2 summarizes the evaluation result of hand-waving recognition using various normalization methods on the multi-distance dataset. The result indicates that the hand-waving recognition system, which normalizes the input of the proposed system by distance and shoulder width, is slightly more accurate than the non-normalized input.

B. Experiment (II): Evaluation of hand-waving recognition with various normalization methods using single-distance dataset

Table 3 summarizes the evaluation result of hand-waving recognition using various normalization methods on the single-distance dataset. The result indicates that the hand-waving recognition system using normalized input achieved high accuracy. Scores when normalizing the input are more than at least five points higher than that of the non-normalized input.

C. Experiment (III): Testing hand-waving recognition in crowded environments

Table 4 summarizes the experimental results of the test data when shoulder width was used for normalization and when distance was used, showing the number of correctly and incorrectly recognized hand-waving. The experiment results indicate that 227 people appeared in all the videos, and 32 people waved their hands.

When shoulder width was used for normalization, the number of correctly detected hand-waving was 23, and the number of incorrectly detected hand-waving was 11. The accuracy, recall, precision, and F-measure are 90.31%, 67.65%, 67.65%, and 67.65%, respectively.

Conversely, hand-waving was detected correctly 21 times when using distance, whereas hand-waving was detected

TABLE V: Experimental result (V): Top 3 teams of RoboCup@Home 2024 "Restaurant"

Team	Point
eR@sers	1060
Hibikino-Musashi@Home (Ours)	**995**
Tech United Eindhoven	615

incorrectly 8 times. The accuracy, recall, precision, and F-measure are 90.75%, 65.62%, 72.41%, and 68.85%, respectively.

The normalization method using distance shows slightly higher accuracy than the normalization method using shoulder width with fewer false positives.

D. Experiment (IV): Measuring real-time performance of the proposed system

The measurement resulted in average processing times of 27.05 ms for OpenPose and 11.42 ms for StrongSORT, and the average time required for hand-waving recognition using RC was 0.0665 ms per frame.

E. Experiment (V): RoboCup@Home 2024 "Restaurant"

In the task, our team succeeded in finding a person waving twice. Consequently, we obtained second place in Restaurant at RoboCup@Home 2024, as illustrated in Table 5, demonstrating that the proposed system works well under real-world conditions.

VI. Discussion

In Experiment (I), because the training and test conditions regarding distances from persons were the same, the property difference between the training and test data was slight, resulting in a high accuracy even without normalization.

In Experiment (II), we used only the 3 m data from the dataset for training. The accuracy, without normalization, is six points lower than the Experiment (I) results. However, normalization using shoulder width and distance information slightly decreased the recognition rate compared to the input without normalization.

These findings suggest that normalization can significantly enhance the adaptability to changes in environmental and condition factors, even with a limited dataset. This potential for increased adaptability can reduce the cost of data collection and improve the general flexibility and utility for a wide variety of applications.

We initially hypothesized that normalization using shoulder width could negatively affect recognition accuracy because the values could change depending on the orientation of a target. When comparing the results of Experiments (I) and (II), no significant difference between normalizations using shoulder width and distance was observed. In this hand-waving recognition, whether the waving hand is being made or not is determined by the change in the x-coordinate of the wrist centered on the x-coordinate of the nose, and the normalization using shoulder width plays a role in suppressing the attenuation of the amplitude of changes due to distance. Therefore, the more the body is turned sideways and the smaller the shoulder width becomes, the larger the amplitude becomes after the normalization, making the difference from the state where the hand was not waving more distinct. Therefore, changes in shoulder width were not considered to affect recognition negatively.

Conversely, the normalization using distance was a challenge because acquiring keypoint distance information can vary depending on the lighting conditions. In the proposed system, keypoints were searched in the order of nose, shoulder, and wrist to use distance information. However, if the keypoints used to determine the distance change, the normalized values also change, potentially negatively affecting recognition.

Experiment (III) indicates that the proposed system can track people and recognize hand-waving in environments where an unspecified number of people are present by adding an object-tracking algorithm to the system that estimates the human skeleton. Therefore, the proposed system is an effective hand-waving recognition system in environments with multiple people.

The proposed system recognized non-hand-waving actions as hand-waving in some cases owing to incorrect keypoint estimations by OpenPose, as illustrated in Fig. 7. The mistakes in keypoint estimation were caused by the lighting environment and the distance from the target and were mainly found at the ends, such as the elbows and wrists, but up to the shoulders were estimated correctly. Additionally, the incorrectly estimated keypoint for the wrist was located in background areas; therefore, the distance value of the keypoint was bigger than the actual. Because the keypoint estimation error sometimes and repeatedly happened, fluctuations in the x-coordinates and distance of the elbow and wrist were observed even if people did not wave their hands.

In this case, the normalization using shoulder width gives the exact shoulder-width information and the fluctuated x-coordinate to the ESN. This fluctuation in x-coordinates was similar to a small hand-waving in the trained dataset; therefore, the ESN recognized it as hand-waving.

In contrast, the normalization using distance provides the ESN with fluctuated distance in addition to the x-coordinate. From Equation 2, we considered that the fluctuation in the x-coordinate multiplied by the fluctuated distance produced an input pattern different from the learned hand-waving and that the ESN did not recognize this as hand-waving. Therefore, the normalization using distance had a higher precision compared with the normalization using shoulder width.

In this experiment, we defined N_{length} as the number of detections for each person in the video within a given period. In actual operation, the video involves real-time streaming, and the video length can be considered infinite. To accommodate this, we fix N_{length} and detect hand-waving at regular intervals. After inference, the system resets the internal state of the reservoir to prevent accumulated errors from affecting subsequent detections.

In Experiment (IV), the average total processing time per frame was 38.54 ms. The frame rate of the dataset used

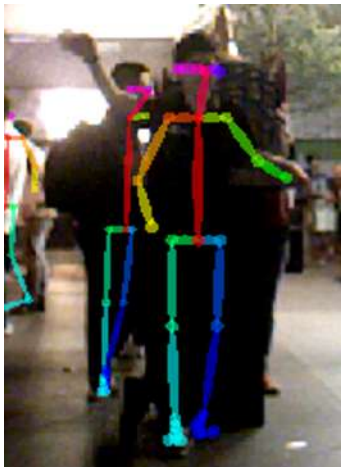

Fig. 7: Sample with incorrect keypoint estimation for the left arm

in this experiment is 24–26 FPS, corresponding to a cycle time of approximately 38.5–41.7 ms. Real-time processing is feasible because the processing time per frame is comparable to the cycle time.

Because OpenPose accounts for more than half of the processing time, further real-time performance could be achieved by moving to a lighter model or applying quantization. Furthermore, because the proposed hand-waving recognition module performs sequential processing each time a frame is accumulated for each person, the time required for hand-waving recognition increases monotonically as the number of subjects increases. Integration of batch processing is necessary to maintain real-time performance.

VII. Conclusion & Future work

In this study, we proposed a lightweight hand-waving recognition system using OpenPose, StrongSORT, and RC to recognize actions in complex environments with multiple people.

Experiments (I) and (II) indicated that incorporating normalization techniques can achieve robust recognition performance across test data at various distances. Experiment (III) demonstrated that the proposed system achieves high accuracy in recognizing hand-waving actions, even in environments like cafeterias with numerous people. Experiment (IV) confirmed that the proposed method can process data at the frame rate of the dataset created for the experiment. We used the proposed system in RoboCup@Home "Restaurant" task during Experiment (V), confirming its effectiveness.

Future research will explore several approaches to improve the recognition accuracy of the system further. First, there is a need to investigate methods to enhance the accuracy of keypoint estimation by OpenPose, mainly by introducing algorithms that are more robust to lighting conditions and visual interference. Additionally, it is essential to evaluate the generalizability by testing it with different datasets and assessing its adaptability to various scenarios.

Moreover, continued efforts to optimize the computational resources and reduce computational load will be crucial for improving real-time processing capabilities. These enhancements could broaden the application range of service robots, enabling them to handle a more comprehensive array of real-world tasks.

References

[1] Fuji Keizai Group, "Survey of the Global Service Robot Market," https://www.fuji-keizai.co.jp/press/detail.html?cid=24010, 2024.08.22 (Accessed).

[2] D. Holz, J. R. del- Solar, K. Sugiura, and S. Wachsmuth, "On RoboCup@Home – Past, Present and Future of a Scientific Competition for Service Robots," RoboCup 2014: Robot World Cup XVIII, pp. 686-697, 2015.

[3] S. Ji, W. Xu, M. Yang, and K. Yu, "3D Convolutional Neural Networks for Human Action Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013.

[4] T. Ono, et al., "Solution of world robot challenge 2020 partner robot challenge (Real Space)," Advanced Robotics, pp. 870–889, 2022.

[5] D. Osokin, "Real-time 2D Multi-Person Pose Estimation on CPU: Lightweight OpenPose," International Conference on Pattern Recognition Applications and Methods, 2018.

[6] N. Wojke and A. Bewley, "Deep Cosine Metric Learning for Person Re-identification," 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 748-756, 2018.

[7] S. Yan, Y. Xiong, and D. Lin, "Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition, " Association for the Advancement of Artificial Intelligence, 2018.

[8] H. Jaeger, "The "echo state" approach to analyzing and training recurrent neural networks-with an erratum note," Bonn, Germany: German National Research Center for Information Technology GMD Technical Report, vol. 148, no. 34, 2001.

[9] H. Jaeger and H. Hass, "Harnessing nonlinearity: predicting chaotic systems and saving energy in wireless communication," Science, vol. 304, pp. 78-80, 2004.

[10] W. Maass, M. Henny, "On the Computational Power of Recurrent Circuits of Spiking Neurons," Journal of Computer and System Sciences, vol. 69, pp. 593-616, 2004.

[11] H. Yamaguchi et al., "A Low Computational Cost Hand Waving Action Recognition System with Echo State Network for Home Service Robots," Proceedings of International Conference on Artificial Life & Robotics (ICAROB2024), 2024.

[12] C. Lugaresi et al., "MediaPipe: A Framework for Perceiving and Processing Reality," Computer Vision and Pattern Recognition (CVPR), 2019.

[13] Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. A. Sheikh, "Open Pose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, pp. 172-186, 2021.

[14] Yunhao Du, Z. Zhao, Y. Song, Y. Zhao, F. Su, T. Gong, and H. Meng, "StrongSORT: Make DeepSORT Great Again," IEEE Transactions on Multimedia, vol. 25, pp. 8725-8737, 2023.

[15] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," Proceedings of the 25th International Conference on Knowledge Discovery and Data Mining, 2019.

[16] RoboCup 2024, "RoboCup@Home," https://2024.robocup.org/leagues/robocuphome/, 2024.08.29 (Accessed).

[17] T. Yamamoto, K.Terada, A. Ochiai, F. Saito, Y. Asahara, and K. Murase, "Development of Human Support Robot as the research platform of a domestic mobile manipulator," ROBOMECH Journal, vol. 6, no. 1, pp. 1–15, 2019.