A Low Computational Cost Hand Waving Action Recognition System with Echo State Network for Home Service Robots

Hiromasa Yamaguchi
School of Computer Science and Systems Engineering, Kyushu Institute of Technology,
680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan

Akinobu Mizutani
Graduate School of Life Science and Systems Engineering, Kyushu Institute of Technology,
2-4 Hibikino, Wakamatsu, Kitakyushu 808-0196, Japan

Arie Rachmad Syulistyo
Graduate School of Life Science and Systems Engineering, Kyushu Institute of Technology,
2-4 Hibikino, Wakamatsu, Kitakyushu 808-0196, Japan

Yuichiro Tanaka
Research Center for Neuromorphic AI Hardware, Kyushu Institute of Technology,
2-4 Hibikino, Wakamatsu, Kitakyushu, 808-0196, Japan

Hakaru Tamukoh
Graduate School of Life Science and Systems Engineering, Kyushu Institute of Technology,
2-4 Hibikino, Wakamatsu, Kitakyushu 808-0196, Japan

E-mail: yamaguchi.hiromasa512@mail.kyutech.jp
https://www.brain.kyutech.ac.jp/~tamukoh/

Abstract

This study proposes a low computational cost hand-waving action recognition system for non-verbal communication in home service robots. The system is based on an echo state network, which requires lower computational costs than that of deep neural networks (DNNs), and processes time-series data of skeletal coordinates of humans to recognize hand-waving actions. Additionally, this study proposes and compares two types of Preprocessing ing methods of the skeletal coordinates to ensure the robustness of the human positions on the frame: one method extracts shoulder and arm angles, which are invariable regardless of the humans' positions and the other normalizes the skeletal coordinates. The experimental result shows that the proposed system has competitive accuracy and is robust to varying human positions.

Keywords: Action recognition, Home service robot, Low computational cost, Echo state network

1. Introduction

The declining birthrate and aging population have become social problems in many countries, not only Japan, and one of the solutions is the use of home service robots [1]. For this reason, research on home service robots has been active [2], [3], [4]. Home service robots can expand the range of their use by acquiring not only verbal but also non-verbal information such as human facial expressions and movements. In addition, since home service robots are required to judge situations and perform actions in real-time, computers for their information processing should be mounted on the home service robot. Therefore, the resources in the computers are limited and the computational cost of the process must be low.

In this study, we focus on hand-waving action recognition, which indicates that a human is calling the home service robot and aim to construct a system with low computational cost. The proposed system recognizes hand-waving action by providing skeletal information extracted from videos using MediaPipe [5] to an echo state network (ESN) [6], a lightweight machine learning model.

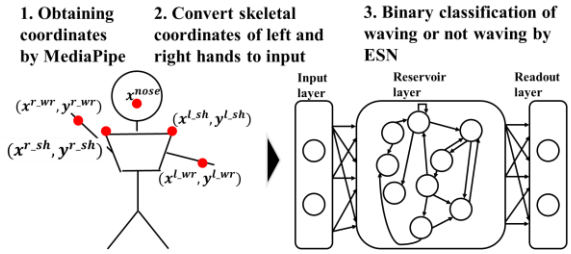

Fig. 1. Process flow of the proposed hand-waving action recognition.

2. Related Works

MMAction2 [7], provided by OpenMMLab, is a library for action recognition. This library estimates human behaviors such as handshaking, hugging, hand waving, etc. The tool offers powerful deep neural network (DNN)-based models for action recognition such as SlowFast [8], UniFormerV2 [9], and VideoMAE V2 [10], however, the three models listed as examples have a huge number of parameters (more than 20M) and are computationally expensive.

While the methods [8], [9], [10] directly feed input images into DNNs, Doan proposed an action recognition system with a long short-term memory model that processes skeletal information [11]. Compared to the end-to-end processing models, the method is more lightweight but still requires high computational costs because it is based on the DNN.

3. Proposed Methods

We propose a lightweight hand-waving action recognition system with MediaPipe [5] and an ESN [6]. The processing flow of the proposed method is shown in Fig. 1. For each video frame, the system extracts skeletal coordinates by using MediaPipe The skeletal coordinates are then given to an ESN, which performs a binary classification of whether the hand is waving or not. Before feeding the coordinates to an ESN, we apply preprocessing to skeletal coordinates to classify the action not depending on the human position on the frame. We propose and compare two types of preprocessing in this study.

Preprocessing A extracts the angles of the shoulders and arms. Preprocessing A shown in Eq. (1) and (2) use the *x*-coordinate of the left and right wrists (x^{l_wr}, x^{r_wr}), the left and right shoulder (x^{l_sh}, x^{r_sh}), and the y-coordinate of the left and right wrists (y^{l_wr}, y^{r_wr}), and the left and right shoulder (y^{l_sh}, y^{r_sh}) respectively. $\boldsymbol{X^{in}}$ in Eq. (3) is the input to an ESN.

Table 1. Experimental results

	Ours (A)	Ours (B)	Slow Only	UniFormer V2
Accuracy	0.667	**0.771**	0.625	0.646
Precision	0.875	**0.928**	0.667	0.622
Recall	0.500	0.683	0.714	**1.000**
processing time[s]	3.445	**3.446**	20.519	29.607

Table 2. Accuracy per subject for the proposed method (Preprocessing B)

Subject	S1	S2	S3	S4
Accuracy	0.813	0.863	0.775	0.662

$$X^{l\_in} = \tan^{-1}\frac{(y^{l\_wr} - y^{l\_sh})}{(x^{l\_wr} - x^{l\_sh})} \quad (1)$$

$$X^{r\_in} = \tan^{-1}\frac{(y^{r\_wr} - y^{r\_sh})}{(x^{r\_wr} - x^{r\_sh})} \quad (2)$$

$$\boldsymbol{X^{in}} = \begin{pmatrix} X^{l\_in} \\ X^{r\_in} \end{pmatrix} \quad (3)$$

Preprocessing B shown in Eq. (4) and (5) use the *x*-coordinate of the left and right wrists (x^{l_wr}, x^{r_wr}), the nose (x^{nose}), and the left and right shoulder (x^{l_sh}, x^{r_sh}), respectively. The coordinates of the wrists are relative to the coordinates of the nose to ensure robustness against parallel movement of the recognition target. In addition, the wrist coordinates are normalized by the length of the shoulder. The robustness against human size in the frame is ensured by normalizing the length of the shoulder. $\boldsymbol{X^{in}}$ in Eq. (6) is the input to an ESN.

$$X^{l\_in} = \frac{(x^{nose} - x^{l\_wr})}{|x^{l\_sh} - x^{r\_sh}|} \quad (4)$$

$$X^{r_{in}} = \frac{(x^{nose} - x^{r\_wr})}{|x^{l\_sh} - x^{r\_sh}|} \quad (5)$$

$$\boldsymbol{X^{in}} = \begin{pmatrix} X^{l\_in} \\ X^{r\_in} \end{pmatrix} \quad (6)$$

4. Experiments

4.1 Setup

We collected a total of 360 video files of actions from six persons, of which 240 data points were used as training data from one person, 24 data points were used as validation data from another person, and 96 data points were used as test data from four persons. The CPU of the computer that performed the processing was a CPU (Intel (R) Core (TM) i7-6700K).

To verify whether the proposed method can be a substitute for the prior cases, we conducted similar tests using the behavior recognition models of the prior cases (however, these models can recognize multiple

behaviors). We used the SlowOnly [8] and the UniFormerV2 [9] models provided by the MMAction2 model trained on the Kinetics-700 dataset [12] of 700 behaviors to predict the behaviors of the test data.

4.2 Result

The averages of the accuracy, precision, recall rates, and processing time for the ten estimation results for the test data are shown in Table 1. The proposed method has a higher accuracy rate in hand-waving action recognition than the other two methods. In addition, we can see that Preprocessing B is superior to A in all indices. The proposed method also has a shorter processing time than the other two methods.

Table 2 shows the accuracy for each subject using the proposed method with Preprocessing B, which had the highest accuracy. These results show that the accuracy varies greatly depending on the subject.

5. Discussion

We consider the Preprocessing B shown in Eq. (4), (5) and (6) with normalization to the wrist coordinates to be more accurate than the Preprocessing A using the angles shown in Eq. (1), (2) and (3), because the rate of fluctuation of the values is larger. In particular, the rate of fluctuation of the values for Preprocessing A is small when the elbow is bent while hand waving. Therefore, we consider that Preprocessing B is superior.

The proposed method has a low recall for Preprocessing B. In Preprocessing B, we consider the influence of the data from the wrist-fixed hand wave to be significant. Preprocessing B makes predictions based on fluctuation in wrist coordinates. Therefore, if the hand is waving without wrist fluctuation, it cannot be recognized. Therefore, it is necessary to devise a system that allows recognition without wrist fluctuation.

As shown in Table 2, variations in accuracy were observed among the subjects. This is expected to be caused by the body size and waving habits of the subjects. It is necessary to verify whether this problem can be solved by using data from multiple subjects in the training data in the future.

6. Conclusion

In this study, we proposed a hand-waving action recognition system based on an ESN to construct a low-computational-cost hand-waving action recognition system to be introduced into home service robots. The results show that the accuracy of the lightweight model is comparable to that of deep learning-based multiple-action recognition models.

In the future, it will be necessary to construct a system that can recognize a waving person even if there are multiple people in the frame because there is a possibility that multiple people may be in the image in the real world.

Acknowledgment

This research is based on results obtained from a project, JPNP16007, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

References

1. 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. 4, 2019.
2. K. Isomoto, Y. Yano, Y. Tanaka, and H. Tamukoh, Robust trash can lid opening system, Proceedings of the 2023 International Workshop on Smart Info-Media Systems in Asia (SISA), 2023.
3. T. Shiba, T. Ono and H. Tamukoh, Object Search and Empty Space Detection System for Home Service Robot, Proceedings of the 2023 International Conference on Artificial Life and Robotics (ICAROB2023), 2023.
4. A. Mizutani, Y. Tanaka, H. Tamukoh, K. Tateno, O. Nomura and T. Morie, A knowledge acquisition system with a large language model and a hippocampus model for home service robots, Proceeding of the Institute of Electronics, Information and Communication Engineers (IEICE) Rep. vol. 123, no. 208, 2023, pp. 13–18.
5. C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C. Chang, M. G. Yong, J. Lee, Wan-Teh Chang, W. Hua, M. Georg and M. Grundmann, MediaPipe: a framework for building perception pipelines, 2019.
6. H. Jaeger: Short term memory in echo state networks, GMD Report 152, 2002.
7. MMAction2 Contributors, OpenMMLab's Next Generation Video Understanding Toolbox and Benchmark, https://github.com/open-mmlab/mmaction2, 2020.
8. C. Feichtenhofer, H. Fan, J. Malik and K. He, SlowFast Networks for Video Recognition, Proceedings of the IEEE international conference on computer vision, 2019, pp. 6202-6211.
9. K. Li, Y. Wang, Y. He, Y. Li, Y. Wang, L. Wang and Y. Qiao, UniFormerV2: Spatiotemporal Learning by Arming Image ViTs with Video UniFormer, 2022.
10. L. Wang, B. Huang, Z. Zhao, Z. Tong, Y. He, Y. Wang, Y. Wang, Y. Qiao, VideoMAE V2: Scaling Video Masked Autoencoders with Dual Masking, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp.14549-14560.
11. T. Doan, An Efficient Patient Activity Recognition using LSTM Network and High-Fidelity Body Pose Tracking, International Journal of Advanced Computer Science and Applications (IJACSA), VoL 13, 2022.
12. L. Smaira, J. Carreira, E. Noland, E. Clancy, A. Wu and A. Zisserman, A Short Note on the Kinetics-700-2020 Human Action Dataset, 2019.

Authors Introduction

Mr. Hiromasa Yamaguchi

He is currently an undergraduate student at Kyushu Institute of Technology, Japan. His research interests include reservoir computing and home service robots.

Mr. Akinobu Mizutani

He received the B.Eng. and the M.Eng. degree from the Kyushu Institute of Technology, Japan, in 2021 and 2023, respectively. He is currently a Ph.D. student at the Graduate School of Life Science and Systems Engineering, Kyushu Institute of Technology, Japan. His research interests include brain-inspired artificial intelligence and its robot application.

Mr. Arie Rachmad Syulistyo

He received the bachelor degree in computer science from the University of Brawijaya, Indonesia, in 2011 and the master degree in computer science from the University of Indonesia, Indonesia, in 2015. He is currently pursuing a Ph.D. degree in life science and systems engineering at Kyushu Institute of Technology, Wakamatsu, Japan. His research interests include computer vision and machine learning.

Dr. Yuichiro Tanaka

He received the B.Eng., M.Eng., and Ph.D. degrees from the Kyushu Institute of Technology in 2016, 2018, and 2021, respectively. He has also been a research fellow at the Japan Society for the Promotion of Science (JSPS) from 2019 to 2021. He has been an assistant professor at the Research Center for Neuromorphic AI Hardware of the Kyushu Institute of Technology since 2021. His research interests include neural networks, home service robots, etc. He is a member of IEEE, IEICE, and JNNS.

Prof. Hakaru Tamukoh

He received the B.Eng. degree from Miyazaki University, Japan, in 2001. He received the M.Eng and the Ph.D. degree from Kyushu Institute of Technology, Japan, in 2003 and 2006, respectively. He was a postdoctoral research fellow of 21st century center of excellent program at Kyushu Institute of Technology, from April 2006 to September 2007. He was an assistant professor of Tokyo University of Agriculture and Technology, from October 2007 to January 2013. He had been an associate professor from February 2013 to March 2021 and is currrectly a professor in the graduate school of Life Science and System Engineering, Kyushu Institute of Technology, Japan. His research interest includes hardware/software complex system, digital hardware etc. He is a member of IEICE, SOFT, JNNS, IEEE, JSAI and RSJ.