Classification in Japanese Sign Language Based on Dynamic Facial Expressions

Yui Tatsumi
School of FSE,
Waseda University
Tokyo, Japan

Shoko Tanaka
School of FSE,
Waseda University
Tokyo, Japan

Shunsuke Akamatsu
Graduate School of FSE,
Waseda University
Tokyo, Japan

Takahiro Shindo
Graduate School of FSE,
Waseda University
Tokyo, Japan

Hiroshi Watanabe
Graduate School of FSE,
Waseda University
Tokyo, Japan

***Abstract*—Sign language is a visual language expressed through hand movements and non-manual markers. Non-manual markers include facial expressions and head movements. These expressions vary across different nations. Therefore, specialized analysis methods for each sign language are necessary. However, research on Japanese Sign Language (JSL) recognition is limited due to a lack of datasets. The development of recognition models that consider both manual and non-manual features of JSL is crucial for precise and smooth communication with deaf individuals. In JSL, sentence types such as affirmative statements and questions are distinguished by facial expressions. In this paper, we propose a JSL recognition method that focuses on facial expressions. Our proposed method utilizes a neural network to analyze facial features and classify sentence types. Through the experiments, we confirm our method's effectiveness by achieving a classification accuracy of 96.05%.**

***Index Terms*—Japanese Sign Language, sign language, facial expressions, pose estimation**

I. Introduction

In Japan, communication methods that rely on knowledge of the Japanese language are frequently used between deaf and hearing individuals. For instance, there are tools such as written communication and speech-to-text applications. However, many deaf individuals struggle with communicating in Japanese because Japanese Sign Language (JSL) has its unique vocabulary and grammar, separate from Japanese. Furthermore, many hearing individuals are not familiar with JSL. The development of JSL recognition methods is required in order to ensure precise and smooth communication between deaf and hearing.

Previous research on sign language recognition has primarily focused on American, German, and Chinese sign languages. Moreover, these studies often concentrate on hand movements and apply hand pose estimation techniques for recognition. However, non-manual markers such as facial expressions and body orientation are indispensable for a complete understanding of sign language sentences.

In this paper, we propose a recognition method for JSL that focuses on non-manual markers. These markers have significant impact on syntactic and semantic information. For example, when hand gestures expressed in an affirmative statement are accompanied by facial expressions such as wide-open eyes, raised eyebrows, and a tucked chin, the sentence transforms into a Yes/No-question. When paired with repeated weak head shakes and furrowed eyebrows, it is classified as a WH-question. Using a neural network, we analyze facial expressions to distinguish the affirmative sentence, Yes/No-question, and WH-question.

II. Related Work

A. Datasets

In sign language recognition research, extensive publicly available datasets for American, German, and Chinese sign languages are widely used [1]. These datasets include videos of individuals using sign language, as well as the corresponding translations. However, datasets for JSL are limited in number; as a result, there are few studies on the topic. To address this issue, in this paper, we create JSL video datasets and apply effective data augmentation.

B. Methods

Numerous previous studies on sign language recognition utilize hand poses. Wang *et al.* combine object detection and hand pose estimation to detect hand shapes and recognize the alphabet and numerals in American Sign Language (ASL) [2]. Similarly, Chu *et al.* and Wu *et al.* focus on hand poses to recognize Japanese and Chinese Sign Language, respectively [3], [4]. However, non-manual markers such as facial expressions are also important in sign language recognition and should be focused on. A study on ASL proposes a recognition method employing facial expressions analysis [5]. In their study, the exact location of facial feature points is reconstructed by correcting the tracked points based on learned face shape subspaces. The extracted data are then analyzed by a recognition system to identify six non-manual markers in ASL.

979-8-3503-5506-2/24/$31.00 ©2024 IEEE

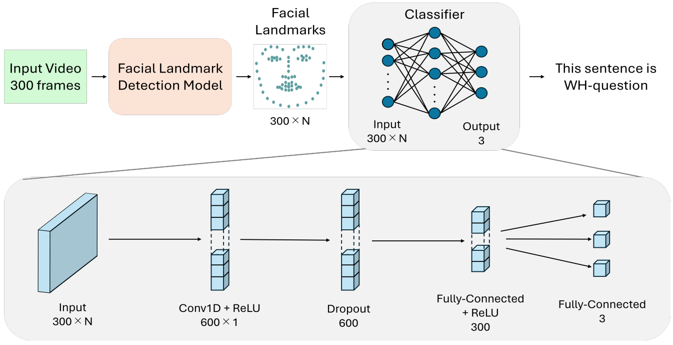

Fig. 1. An overview of our proposed method. The variable N indicates the number of facial landmarks extracted from OpenPose, MediaPipe, and Dlib models, which are 140, 936, and 136, respectively.

III. Proposed Method

We propose a JSL recognition method focusing on its facial expressions. The overview of our proposed method is shown in Fig. 1. This method classifies JSL videos into three classes: (1) affirmative sentences, (2) Yes/No questions, or (3) WH-questions, based on the facial expressions of the signers. Initially, sign language videos are input into a model for detecting facial landmarks. In our experiments, OpenPose [6], MediaPipe [7], and Dlib [8] are utilized as the model for comparison. The outputs from the models include x and y coordinates of 70, 468, and 68 estimated facial landmarks, respectively. Then, these landmarks are utilized by a neural network based classifier to classify JSL sentences. The classifier consists of a convolutional layer, an activation function (ReLU), and two fully-connected layers. Cross Entropy Loss is used as the loss function.

IV. Experiment

A. Datasets

A total of 378 JSL videos are collected as our dataset. Among these, 12 videos are created by Morita *et al.* [9], and 39 videos by Oka *et al.* [10]. These videos are performed by three native JSL signers who are deaf. The remaining 327 videos are created by us. The signers are four individuals from Waseda University Sign Language Club, including a deaf student. Each of the 378 videos captures a single participant expressing a brief sentence in JSL within 10 seconds. The frame rate of the videos is 30 fps. Each sentence is labeled according to three categories: affirmative, Yes/No-question, and WH-question. 302 videos are used for training and 76 videos are employed for validation.

B. Training and Validation

For training, initially, each video is uniformly extended to 300 frames by padding with the value from its final frame. Secondly, facial landmarks are extracted using the facial landmark detection model. These landmarks are normalized so that the average distance from the nose tip to the other coordinates is one. In addition, permutation is applied as a data augmentation strategy [11]. Finally, the classifier is trained using normalized data of 302 videos.

TABLE I
Classification Results Using Various Detection Models

Facial Landmark Detection Model	Accuracy(%)	Precision(%)	Recall(%)	F1 Score(%)
OpenPose [6]	**96.05**	**96.25**	**96.05**	**96.12**
MediaPipe [7]	88.16	88.48	88.25	88.34
Dlib [8]	82.89	83.59	82.91	83.10

For validation, 76 videos are employed. The validation metrics include accuracy, precision, recall, and F1 score. Table I shows the results. The classification accuracy achieved using OpenPose demonstrates the validity of our dataset and data augmentation, as well as the effectiveness of our proposed method. Additionally, employing OpenPose as the detection model results in higher accuracy compared to both MediaPipe and Dlib. This superiority stems from OpenPose's robust ability to detect facial landmarks in videos with cluttered backgrounds and sudden movements, such as head shakes.

V. Conclusion

In this paper, we propose a JSL recognition technique that focuses on its facial expressions. We construct a classifier that categorizes JSL videos into affirmative statements, Yes/No-questions, or WH-questions based on various facial features. Experimental results present the effectiveness of our proposed method. Future work is required to achieve comprehensive recognition of JSL by combining hand pose estimation techniques.

References

[1] N. Adaloglou *et al.*, "A Comprehensive Study on Deep Learning-Based Methods for Sign Language Recognition," IEEE Transaction on Multimedia, Vol. 24, pp. 1750-1762, Apr. 2022.

[2] Z. Wang *et al.*, "American Sign Language Alphabet Recognition with YOLOv5 Enhanced by MediaPipe Hands," 8th International Conference on Instrumentation, Control, and Automation, pp. 103-108, Aug. 2023.

[3] X. Chu *et al.*, "A Sensor-Based Hand Gesture Recognition System for Japanese Sign Language," IEEE 3rd Global Conference on Life Sciences and Technologies, pp. 311-312, Mar. 2021.

[4] B. Wu *et al.*, "A Modified LSTM Model for Chinese Sign Language Recognition Using Leap Motion," IEEE International Conference on Systems, Man, and Cybernetics, pp. 1612-1617, Oct. 2022.

[5] T. D. Nguyen *et al.*, "Facial expressions in American sign language: Tracking and recognition," Pattern Recognition, Vol. 45, No. 5, pp. 1877-1891, May 2012.

[6] Z. Cao *et al.*, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 43, No. 1, pp. 172-186, Jan. 2021.

[7] C. Lugaresi *et al.*, "Mediapipe: A framework for perceiving and processing reality," Third workshop on computer vision for AR/VR at IEEE CVPR, Vol. 2019, June 2019.

[8] D. E. King, "Dlib-ml: A Machine Learning Toolkit," Journal of Machine Learning Research, Vol. 10, pp. 1755-1758, Dec. 2009.

[9] A. Morita *et al.*, "Know, Learn, Teach Japanese Sign Language Meisei Gakuen Method," (in Japanese), Gakuji Shuppan, Aug. 2023.

[10] N. Oka *et al.*, "Japanese Sign Language System Practice Book," (in Japanese), Taishukan, Apr. 2016.

[11] T. T. Um *et al.*, "Data augmentation of wearable sensor data for parkinson ’ s disease monitoring using convolutional neural networks," 19th ACM International Conference on Multimodal Interaction, pp. 216-220, Nov. 2017.