Sensor Glove Approach for Japanese Fingerspelling Recognition System Using Convolutional Neural Networks

Tomohiko Tsuchiya*, Akihisa Shitara†, Fumio Yoneyama*, Nobuko Kato* and Yuhki Shiraishi*
*Faculty of Industrial Technology, Tsukuba University of Technology, Japan
Email: {a193102, yonefumi, nobuko, yuhkis}@a.tsukuba-tech.ac.jp
†Graduate School of Library, Information, and Media Studies, University of Tsukuba, Japan
Email: theta-akihisa@digitalnature.slis.tsukuba.ac.jp

***Abstract*—We have developed a Japanese fingerspelling recognition system based on a sensor glove, using deep learning, to achieve smooth communication between the deaf and hard-of-hearing, and hearing people. In this study, we conducted evaluation experiments using a convolutional neural network to recognize 76 characters of Japanese fingerspelling. In the developed system, we have adopted a sensor glove that is light and cheap. Additionally, the target Japanese fingerspelling alphabet includes 35 characters for dynamic fingerspelling, which require both finger and wrist movement. The experimental results demonstrated that the average recognition rate of the developed system was approximately 70.0%. Based on these results, we have discussed the peculiarity of Japanese fingerspelling and potential improvements to sensor gloves and algorithms.**

***Keywords*–Sign language; Japanese fingerspelling; Sensor glove; Recognition; Convolutional neural network.**

I. Introduction

In recent years, there has been an increased interest in research on speech recognition and information technology devices with voice input functions. Various applications, such as KoeTra [1] and UDtalk [2], as well as cloud-speech-to-text services [3], have been released to provide information accessibility to the Deaf and Hard-of-Hearing (DHH) based on speech recognition. As a result, the DHH can read text corresponding to vocalizations.

However, it is difficult for hearing people to read sign language. Some research on information accessibility systems for sign language recognition has been reported [4]–[11]. However, compared to information accessibility systems based on speech recognition, the development of a practical sign language recognition system is still in progress.

As a primary communication method, sign language is used in everyday conversations among the DHH. Sign language recognition has different characteristics than speech recognition. It is difficult for hearing people to learn and read sign languages. Therefore, a system for converting sign language into voice information or text information (i.e., a sign language recognition system) is necessary (see Figure 1).

A sign language recognition system must recognize the position, direction, and shape of the hands, as well as motion. Methods for recognizing sign language can be roughly classified into recognition using cameras [4] [5] [9], which are non-contact-type sensors, and recognition using sensor gloves, which are contact-type sensors [6] [7] [10] [11]. Luzhnica et al. [6] reported a recognition accuracy of 98.5% for sign language using a sensor glove; however, they only considered approximately 30 recognition candidate classes, which is insufficient for practical use.

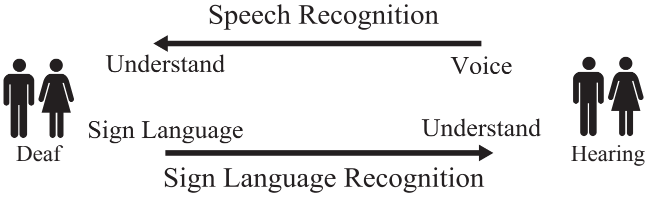


Figure 1. Information accessibility system.

TABLE I. Numbers of fingerspelling characters in different countries.

Language	*Dynamic*	*Static*	*Sum*
American	2	24	26
French	3	23	26
Japanese	35	41	76

In recent years, technologies based on deep learning have attracted significant attention. Deep learning, which increases the number of hidden layers in a neural network, is a type of machine learning that can contribute to improving recognition rates. For example, to improve hand gesture recognition accuracy based on image recognition, various techniques for applying deep learning have been reported [4].

In this study, as a first step toward sign language recognition to facilitate communication with the DHH, we attempted to recognize the Japanese FingerSpelling (JFS) recognition system. JFS is composed of representations of Japanese characters, using only the fingers.

A camera, which is a non-contact-type sensor, is difficult to use for sign language recognition in everyday life because hands must be captured by the camera. Additionally, cameras are easily affected by environmental factors. In contrast, hand shape recognition using contact sensors, such as sensor gloves, is easy to perform because sensors can be attached directly to the hands.

We were motivated by the goal of improving recognition accuracy by adopting conductive fiber weaving technology [12], which can reduce the weight and cost of sensor gloves and simplify hand movements (see Figure 2).

In our experiments, we evaluated our developed system by classifying 76 JFS characters, including dynamic (non-static) fingerspelling characters, which are a unique feature of JFS compared to other fingerspelling systems, as shown in Table I. Evaluation experiments were conducted using a Convolutional

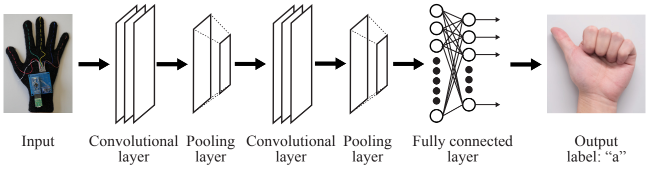


Figure 2. Recognition diagram.

Neural Network (CNN) as a learning model (this type of model performed best in previous studies) to perform data reduction by calculating moving averages of the data acquired from gyro sensors. In these evaluation experiments, all 76 characters of JFS were included as recognition targets, as well as dullness, semi-voiced sounds, diphthongs, and long vowels. Evaluation experiments were conducted using all collected data under a variety of experimental conditions.

In Section II, we present the related works. In Section III, our developed system is detailed. In Section IV, the experimental method is described. In Section V, experimental results are presented, and their implications and limitations are discussed. In Section VI, the conclusions are provided.

II. Related Work

In past research on fingerspelling recognition, two main types of sensors have been proposed to recognize a series of operations in fingerspelling: contact-type sensor gloves and non-contact-type cameras for image recognition.

A. Image recognition

Several methods for recognizing hand shapes based on processing images of fingerspelling captured by cameras have been proposed. Mukai et al. [8] reported that fingerspelling recognition targeting 41 characters without movement in Japanese sign language resulted in an average recognition accuracy of 86%. They used a classification tree and machine learning based on a support vector machine to classify individual images. Hosoe et al. [9] employed deep learning to perform recognition and achieved a recognition rate of 93%, but only for static fingerspelling. Jalal et al. [5] reported a recognition rate for American Sign Language (ASL) images of 99% based on a deep learning algorithm, but only for static fingerspelling (i.e., excluding "J" and "Z"). Therefore, recognition accuracy cannot be considered sufficient for the practical recognition of JFS. Additionally, very few recognition results for dynamic fingerspelling (i.e., the fingers move when expressing a character) have been reported.

B. Sensor glove recognition

Several methods for recognizing hand shapes based on measurement data acquired by contact-type sensor gloves have been proposed. This method can be used to measure finger bending, hand position, and directional data. The measurement data are then sent to a personal computer and a classification algorithm is used to recognize hand shapes. Cabrera et al. [10] paired the Data Glove 5 Ultra [13] sensor glove with an acceleration sensor and acquired information regarding the degree of flexion of each finger, as well as wrist direction. They conducted test classification using 24 static fingerspelling characters in ASL, excluding "J" and "Z." Their neural network was trained using 5 300 patterns and achieved a recognition rate of 94.07% for 1 200 test patterns. Mummadi et al. [11] proposed a sensor glove prototype with multiple embedded small inertial sensors. They collected French sign language fingerspelling data from 57 people and achieved an average recognition rate of 92% with an F value of 91%. Among various methods for performing JFS recognition, the conductive fiber braid method [12] uses gloves woven with conductive fibers instead of bending sensors. Additionally, such gloves can recognize hand shapes and hand movements by incorporating directional gyro sensor. However, the recognition rate for JFS ("a", "i", "u", "e", "o") based on Euclidean distance has been reported to be only 60%.

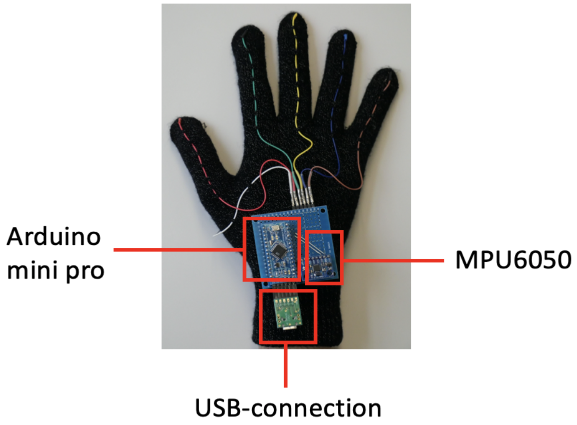


Figure 3. Prototype sensor glove.

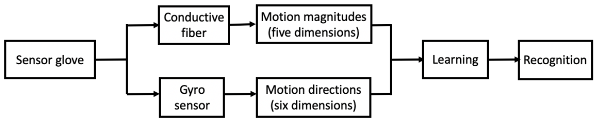

Sensor glove
Conductive fiber
Gyro sensor
Motion magnitudes (five dimensions)
Motion directions (six dimensions)
Learning
Recognition

Figure 4. Software structure.

III. System Development

In this study, to achieve smooth communication in real-world environments, we designed a system for communicating information using lightweight and comfortable sensor gloves to recognize fingerspelling with high accuracy in real time. The developed system consists of a sensor value measurement unit and recognition unit. Figure 3 presents the JFS recognition system developed in this study. Figure 4 presents the corresponding software architecture.

A. Sensor glove

To recognize fingerspelling efficiently based on hand, finger, and wrist data, it is necessary to detect motion magnitudes and directions using a sensor glove. In this study, we adopted a hand shape recognition technique using conductive fiber sensor gloves, which are more comfortable, less expensive, and lighter than traditional sensor gloves. Motion direction is detected using a gyro sensor. Motion magnitudes are detected based on resistance changes in the conductive fibers in the

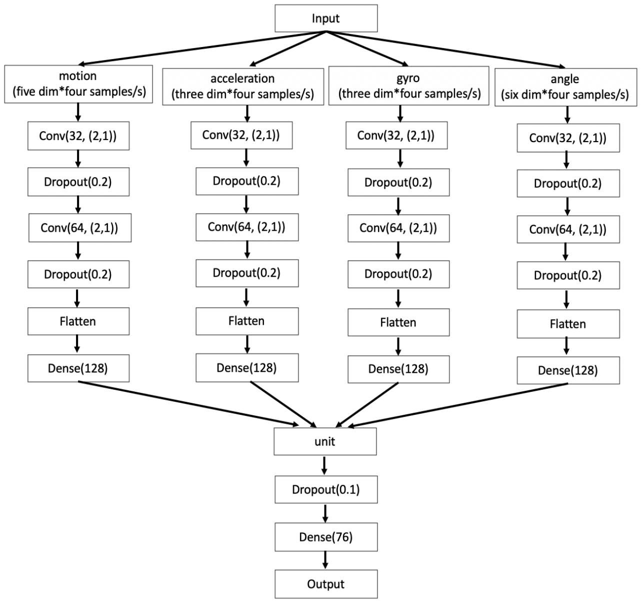

Input
motion (five dim*four samples/s)
acceleration (three dim*four samples/s)
gyro (three dim*four samples/s)
angle (six dim*four samples/s)
Conv(32, (2,1))
Dropout(0.2)
Conv(64, (2,1))
Dropout(0.2)
Flatten
Dense(128)
unit
Dropout(0.1)
Dense(76)
Output

Figure 5. Architecture of the convolutional neural network.

gloves. The motion detection board is an Arduino board and the measurement values from the sensor glove are transferred from the detection board to a PC, where they are saved in comma-separated-value format. Machine learning and motion recognition are performed using Python implementations on a PC. Sensor readings for JFS motion from the data gloves have different scales depending on the wearer. Therefore, the data are subjected to linear normalization in consideration for differences in movement. Additionally, because the activation function and likelihood function of the proposed system are based on probabilities, as a pretreatment for network inputs, we perform scale conversion to a range of zero to one.

Motion magnitudes are detected based on resistance changes in the conductive fibers during flexion and extension of the fingers. We use partial pressure values to calculate input voltages based on (1).

$$V_{in} = \frac{R_1}{R_1 + R_2} * V_{out} \tag{1}$$

In this equation, V_{in} is the estimated motion magnitude, V_{out} is the reference voltage, R_1 is the variable resistance of the conductive fibers, R_2 is a fixed resistance. When a finger is stretched, the resistance value of the conductive fiber increases. When a finger is bowed, the resistance value of the fiber decreases.

B. Recognition algorithm

In this study, we adopted a CNN. This type of network has achieved high recognition rates in previous studies. The CNN and k-fold cross validation were implemented using open-source libraries called TensorFlow [14] and scikit-learn [15]. We adopted the RMSprop training algorithm [16]. The activation function is a rectified linear unit, as shown in (2). The error function is the cross-entropy function shown in (3), where t_k is the correct label (one-hot expression) and y_k expresses the network output.

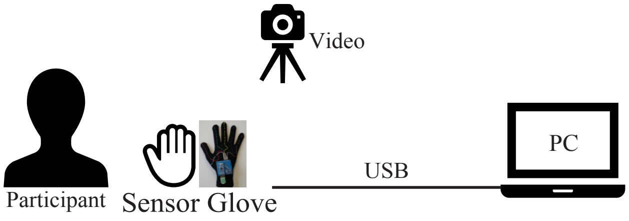


Figure 6. Data acquisition experiment.

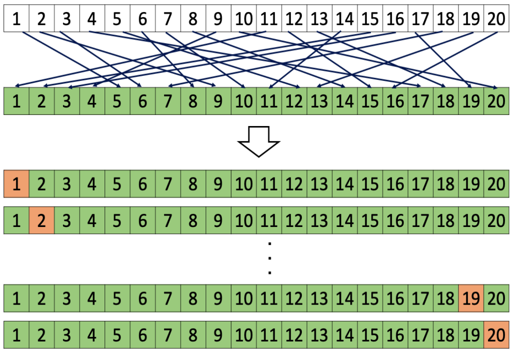


Figure 7. Twenty-fold cross validation by shuffling data.

$$f(u) = max(u, 0) \tag{2}$$

$$E = -\sum_{k} t_k \log y_k \tag{3}$$

CNNs are often used for image recognition and can generally achieve high recognition rates. Convolutional layers and pooling layers are the main features of CNNs. These layers are updated as their feature values are extracted during the training process. We transform the measurement data acquired by the sensor glove into two dimensions based on training and evaluation trials. The motion magnitudes, accelerations, and gyro readings are branched at the time of input. Through the CNN (typical layer size of 32 to 64 nodes), these data are coupled using “Flatten” and “Dense” operations (128 nodes). Finally, by using an additional Dense operation (76 nodes) corresponding to the number of JFS characters, outputs are generated. Figure 5 presents a system overview of the CNN. In the CNN, inputs are initially separated based on the physical meanings of each signal. The separated signals are eventually combined to recognize JFS characters.

IV. Experimental Method

A. Data collection

To target 76 JFS characters, we recruited 20 participants (from 20 to 27 years old). In our experiments, each participant wore a sensor glove and performed the motions of finger-spelling characters in sequence for 1 s at a time according to directions provided by a moderator. As shown in Figure 6, video was also recorded to capture the motions of the wrists and fingers of the participants. For each 1 s motion, at a

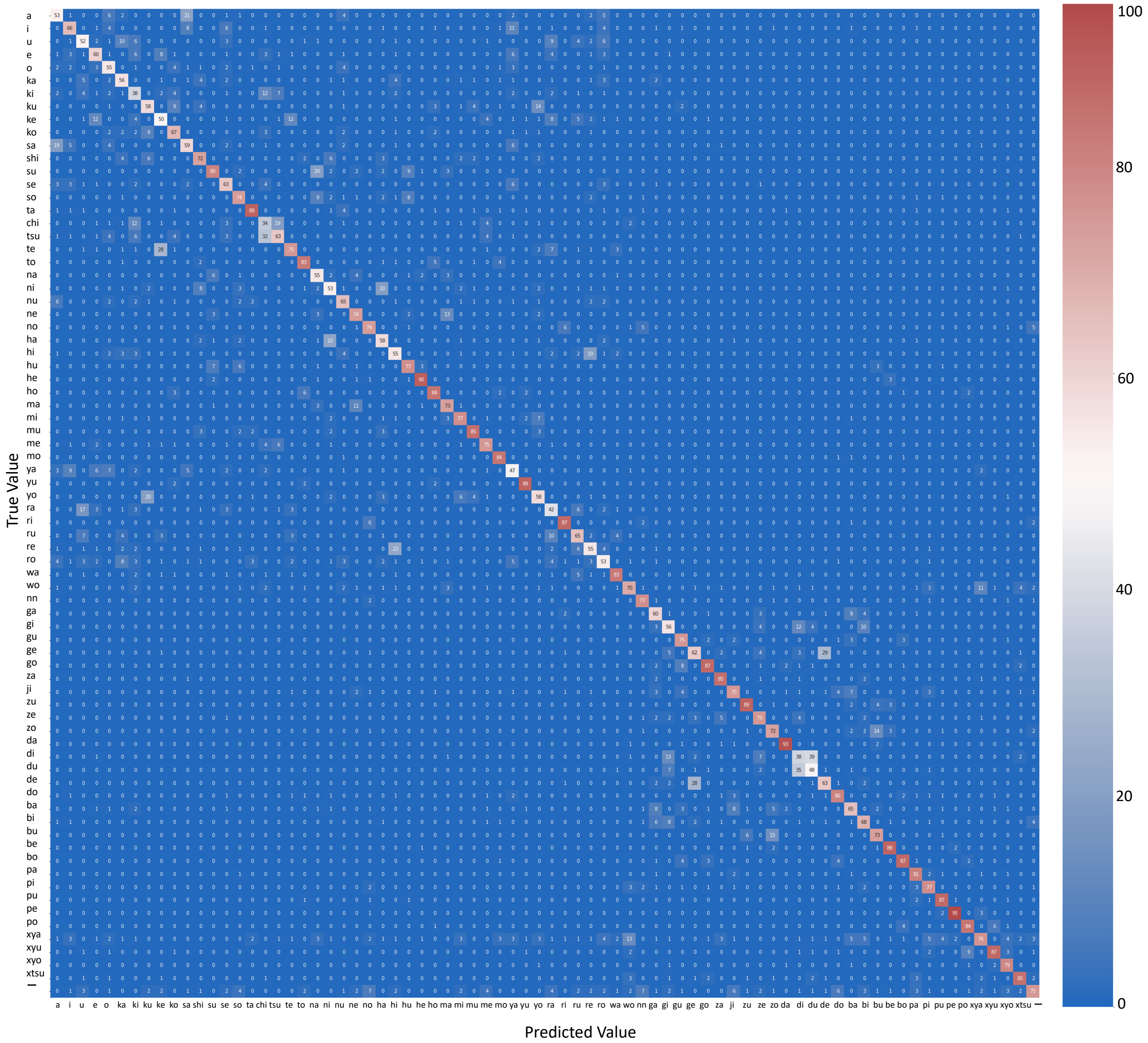


Figure 8. Confusion matrix.

rate of 200 samples per second, the sensor gloves captured five dimensions of motion magnitude data, three dimensions of acceleration data, and three dimensions of gyro data for a total of 11 dimensions. Data labeling was conducted manually at the same time as data collection. This series of motions was repeated five times. Therefore, with five repetitions per participant, 76 JFS characters, and 200 samples per second for 1 s, a total of 76 000 motion measurement data were collected for each participant. We were able to collect a total of 1 520 000 data samples for all 20 participants. These experiments were conducted with approval from the Tsukuba University of Technology Research Ethics Committee (Approval number: H30-17).

To overcome several of the issues in previous works, we performed extensive data cleaning and feature selection operations. To prevent gyro drift, we used Madgwick filters [17], which calculate angles from the values of acceleration and gyro sensors in real time. This allowed us to calculate three angle dimensions from the acceleration and gyro data. To clarify hand directions, the angles were converted into sine and cosine data. The resulting six dimensions were combined with the motion magnitudes (five dimensions) and motion directions (six dimensions) mentioned above to generate a total of 17 dimensions. Next, we conducted a review of the sampling frequency. Although 200 samples per second can be acquired without leakage, noise and training time are included

TABLE II. TWENTY-FOLD CROSS VALIDATION RESULTS.

k	Learning data (%)	Validation data (%)
1	93.6	65.0
2	94.1	75.5
3	94.8	68.7
4	93.1	69.7
5	94.2	66.3
6	93.9	73.2
7	92.9	67.9
8	93.5	71.1
9	93.0	67.4
10	94.6	70.5
11	93.4	71.6
12	93.0	66.1
13	94.6	68.9
14	94.3	70.3
15	93.0	69.7
16	93.4	68.4
17	92.9	71.3
18	93.1	71.1
19	94.5	74.2
20	94.5	72.4
Average	93.7	70.0

TABLE III. MISRECOGNITION PATTERNS.

Teacher	a	sa	ku	yo	ke	te	ki	chi	chi
Prediction	sa	a	yo	ku	te	ke	chi	ki	tsu
Rate (%)	21.0	19.0	14.0	20.0	12.0	28.0	12.0	12.0	34.0
Teacher	tsu	ni	ha	ne	ma	hi	re	wo	xya
Prediction	chi	ha	ni	ma	ne	re	hi	xya	wo
Rate (%)	32.0	20.0	22.0	13.0	11.0	19.0	23.0	11.0	13.0
Teacher	gi	di	ge	de	di	du	zo	bu	
Prediction	di	gi	de	ge	du	di	bu	zo	
Rate (%)	12.0	13.0	29.0	20.0	39.0	35.0	14.0	15.0	

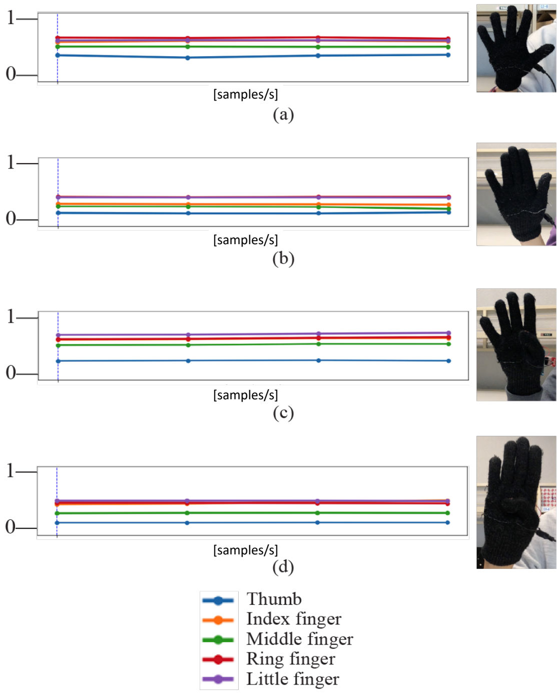


Figure 9. Example input data (only five dimensions): (a) predict "te" as "te" correctly, (b) predict "te" as "ke" incorrectly, (c) predict "ke" as "te" incorrectly, (d) predict "ke" as "ke" correctly.

in these samples. Therefore, the number of data was reduced by calculating a moving average to achieve a final value of 4 samples/s.

B. Evaluation experiments

The collected data were evaluated using a CNN (Figure 5) and k-fold cross validation (k = 20). In our evaluation experiments, data shuffling was performed using Google Colaboratory [18]. The number of folds for k-fold cross validation was set to 20 according to the number of participants. Additionally, confusion matrices and accuracy rates were generated using 20-fold cross validation of all data shuffling evaluations (see Figure 7).

V. RESULTS AND DISCUSSION

The experimental results of 20-fold cross-validation are listed in Table II. This table reveals an average recognition rate of approximately 70.0%.

As shown in Figure 8 and Table III, various misrecognition patterns occurred. We believe these patterns occurred because the conductive fibers are firmly attached to the sensor gloves. We confirmed that the hand directions for "ha" and "ni, " which are JFS characters, varied among participants. Additionally, "ne" and "ma" appear to be confused based on both hand bending and finger bending.

Figure 9 presents sample input data leading to misrecognition for the JFS characters "te" and "ke". By analyzing the data, it was confirmed that close contact between the fingers caused these errors. Notably, the thumb sometimes contacted the forefinger. Additionally, depending on the participant, the hand may be widely opened or the fingers may be in close contact.

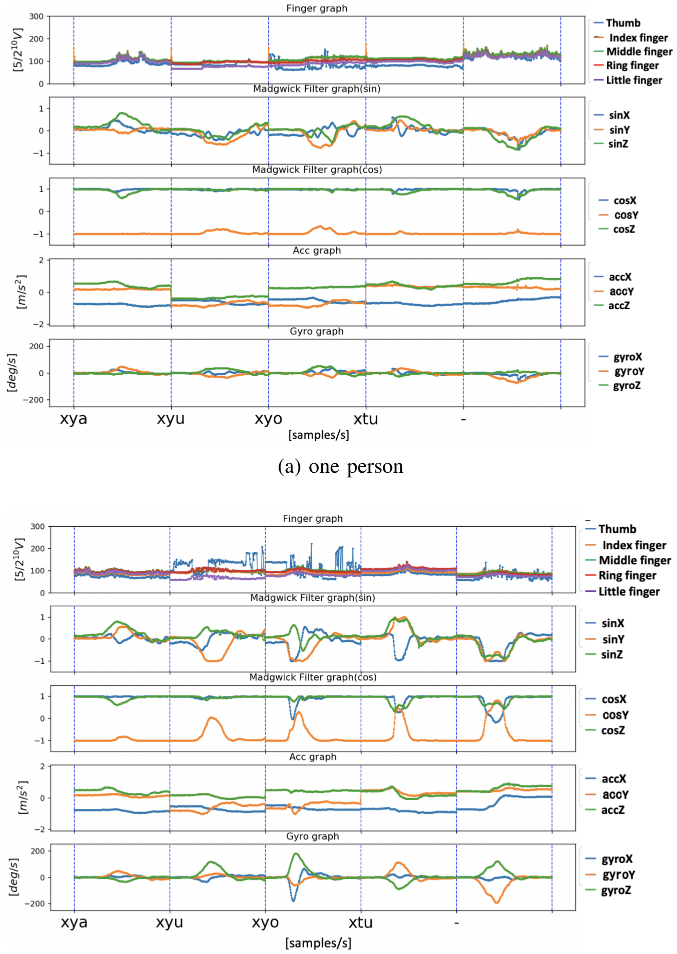
Figure 10 presents examples of acquiring data from two participants using the sensor glove for dynamic fingerspelling. This figure clearly highlights the individual differences in fingerspelling between participants, particularly in the strength of finger bending (including noisy signals), timing of hand move movement, and shape of the fingers. Therefore, it is necessary to improve recognition algorithms and data glove devices (e.g., detecting hand movement periods and constructing more robust glove devices).

Based on the aforedescribed results, we determined that recognition errors largely occurred based on variance in the flexion and direction of the fingers. We also confirmed that finger expressions vary based on individual differences, which can be attributed to different home and social environments, making recognition more difficult.

However, JFS is widely used for displaying proper names and technical terms. Therefore, the recognition of JFS is essential for realizing a Japanese sign language recognition system.

VI. CONCLUSIONS AND FUTURE WORK

In this study, to realize smooth communication between the DHH and hearing people, we adopted a lightweight sensor glove, developed an effective convolutional neural network



(a) one person

(b) another person

Figure 10. Example of acquiring data.

model, implemented a JFS recognition system, and evaluated the performance of the developed system. JFS data collection experiments with 20 participants and 76 target JFS characters were repeated five times. Data were acquired at a rate of 200 samples per second for 11 input dimensions. Angle data were then transformed by applying a Madgwick filter to gyro readings and converted into the sine and cosine space, which increased the total number of input dimensions to 17. However, the data acquired at 200 samples per second contained various issues, including noisy signals. To solve this problem, we calculated moving averages to reduce the frequency to 4 samples/s.

Finally, a 20-fold cross validation evaluation experiment was conducted. The average recognition rate was approximately 70.0% and the maximum recognition rate was approximately 75.5%. It was determined that the firm attachment of conductive fibers was a significant cause of misrecognition.

In future work, we will construct improved sensor gloves and investigate methods to handle various problems, such as individual differences and hand movement detection. To this end, we are planning additional experiments for data collection under more controlled conditions. Additionally, we will conduct continuous fingerspelling recognition experiments.

Acknowledgment

This work was partially supported by JSPS KAKENHI Grant Number #19K11411 and the Promotional Projects for Advanced Education and Research in NTUT. We would like to thank Editage (www.editage.com) for English language editing.

References

[1] “KoeTra,” 2015, URL: https://www.koetra.jp/en/ [retrieved: February,2020].

[2] “UDtalk,” 2015, URL: https://udtalk.jp/ [retrieved: February,2020].

[3] “speech-to-text,” 2019, URL: https://cloud.google.com/speech-to-text [retrieved: February,2020].

[4] S. Gattupalli, A. Ghaderi, and V. Athitsos, “Evaluation of deep learning based pose estimation for sign language recognition,” in Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments, 2016, pp. 1–7.

[5] M. A. Jalal, R. Chen, R. K. Moore, and L. Mihaylova, “American sign language posture understanding with deep neural networks,” in 2018 21st International Conference on Information Fusion (FUSION). New York, NY, USA: IEEE (Institute of Electrical and Electronics Engineers), July 2018, pp. 573–579.

[6] G. Luzhnica, J. Simon, E. Lex, and V. Pammer, “A sliding window approach to natural hand gesture recognition using a custom data glove,” in 2016 IEEE Symposium on 3D User Interfaces (3DUI). New York, NY, USA: IEEE (Institute of Electrical and Electronics Engineers), March 2016, pp. 81–90.

[7] K. Murakami and H. Taguchi, “Gesture recognition using recurrent neural networks,” in Proceedings of the SIGCHI conference on Human factors in computing systems, 1991, pp. 237–242.

[8] N. Mukai, N. Harada, and Y. Chang, “Japanese fingerspelling recognition based on classification tree and machine learning,” in 2017 Nicograph International (NicoInt). New York, NY, USA: IEEE (Institute of Electrical and Electronics Engineers), June 2017, pp. 19–24.

[9] H. Hosoe, S. Sako, and B. Kwolek, “Recognition of jsl finger spelling using convolutional neural networks,” 05 2017, pp. 85–88.

[10] M. E. Cabrera, J. M. Bogado, L. Fermin, R. Acuna, and D. Ralev, “Glove-based gesture recognition system,” in Adaptive Mobile Robotics. World Scientific, 2012, pp. 747–753.

[11] C. K. Mummadi, F. P. P. Leo, K. D. Verma, S. Kasireddy, P. M. Scholl, and K. Van Laerhoven, “Real-time embedded recognition of sign language alphabet fingerspelling in an imu-based glove,” in Proceedings of the 4th International Workshop on Sensor-Based Activity Recognition and Interaction, ser. iWOAR ’17. New York, NY, USA: Association for Computing Machinery, 2017, pp. 1–6. [Online]. Available: https://doi.org/10.1145/3134230.3134236

[12] R. Takada, J. Kadomoto, and B. Shizuki, “A sensing technique for data glove using conductive fiber,” in Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, ser. CHI EA ’19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 1–4. [Online]. Available: https://doi.org/10.1145/3290607.3313260

[13] “5DT Data Glove 5 Ultra,” 2019, URL: https://5dt.com/ [retrieved: February,2020].

[14] “TensorFlow,” 2019, URL: https://www.tensorflow.org [retrieved: February,2020].

[15] “scikit-learn,” 2019, URL: https://scikit-learn.org/stable/index.html [retrieved: February,2020].

[16] G. Hinton, N. Srivastava, and K. Swersky, “Lecture 6e-rmsprop: Divide the gradient by a running average of its recent magnitude. coursera neural networks mach learn. 2012.”

[17] S. O. H. Madgwick, A. J. L. Harrison, and R. Vaidyanathan, “Estimation of imu and marg orientation using a gradient descent algorithm,” in 2011 IEEE International Conference on Rehabilitation Robotics. New York, NY, USA: IEEE (Institute of Electrical and Electronics Engineers), June 2011, pp. 1–7.

[18] E. Bisong, “Google colaboratory,” in Building Machine Learning and Deep Learning Models on Google Cloud Platform. Springer, 2019, pp. 59–64.