*electronics*

*Article*

# Sign Language Recognition with Multimodal Sensors and Deep Learning Methods

**Chenghong Lu, Misaki Kozakai and Lei Jing ***

Graduate School of Computer Science and Engineering, University of Aizu, Tsuruga, Ikki-machi, Aizuwakamatsu 965-8580, Japan
* Correspondence: leijing@u-aizu.ac.jp

**Abstract:** Sign language recognition is essential in hearing-impaired people's communication. Wearable data gloves and computer vision are partially complementary solutions. However, sign language recognition using a general monocular camera suffers from occlusion and recognition accuracy issues. In this research, we aim to improve accuracy through data fusion of 2-axis bending sensors and computer vision. We obtain the hand key point information of sign language movements captured by a monocular RGB camera and use key points to calculate hand joint angles. The system achieves higher recognition accuracy by fusing multimodal data of the skeleton, joint angles, and finger curvature. In order to effectively fuse data, we spliced multimodal data and used CNN-BiLSTM to extract effective features for sign language recognition. CNN is a method that can learn spatial information, and BiLSTM can learn time series data. We built a data collection system with bending sensor data gloves and cameras. A dataset was collected that contains 32 Japanese sign language movements of seven people, including 27 static movements and 5 dynamic movements. Each movement is repeated 10 times, totaling about 112 min. In particular, we obtained data containing occlusions. Experimental results show that our system can fuse multimodal information and perform better than using only skeletal information, with the accuracy increasing from 68.34% to 84.13%.

**Keywords:** sign language recognition; sensor fusion; deep learning

check for updates

**Citation:** Lu, C.; Kozakai, M.; Jing, L Sign Language Recognition with Multimodal Sensors and Deep Learning Methods. *Electronics* **2023**, *12*, 4827. https://doi.org/10.3390/electronics12234827

Academic Editor: Zhenhua Guo

Received: 18 September 2023
Revised: 22 November 2023
Accepted: 23 November 2023
Published: 29 November 2023

## 1. Introduction

Recognition of hand motion capture is an interesting topic. Hand motion can represent many gestures. In particular, sign language plays an important role in the daily lives of hearing-impaired people. About 2.5 billion people are expected to have some degree of hearing loss by 2050, according to the WHO. Additionally, more than 1 billion young people are at risk of permanent hearing loss [1]. Moreover, due to the impact of infectious diseases in recent years, online communication has become important. Sign and language recognition can assist individuals with speech or hearing impairments by translating their sign language into text or speech, making communication with others more accessible [2]. In Human–Computer Interaction (HCI), sign recognition can be used for gesture-based control of computers, smartphones, or other devices, allowing users to interact with technology more naturally [3]. Facilitating communication between sign language users and non-users via video calls remains a pertinent research focus. However, the intricate nature of sign language gestures presents challenges to achieving optimal recognition solely through wearable data gloves or camera-based systems.

Both wearable data gloves and camera-based systems have been extensively explored for sign language recognition. Bending sensor gloves only focus on the degree of finger bending. Consequently, several sign language words exhibiting similar curvature patterns become indistinguishable. This limitation curtails the utility of such devices. Given the significance of hand and arm gestures in sign language, it is imperative for vision-based approaches to prioritize the extraction of keypoints data from the hands, thereby reducing interference from extraneous background elements. Occlusion presents a significant challenge to vision-based methodologies. During the acquisition of hand keypoints, monocular cameras may fail to capture certain spatial information due to inter-finger occlusions. Such occlusions often act as impediments, constraining the potential for enhancement in recognition accuracy. In gesture recognition, fingers can easily block each other, objects can block hands, or parts can become nearly unrecognizable due to being overexposed or too dark. As shown in Figure 1, occlusion problems significantly hinder the effective acquisition of keypoints. Integration with bending sensors offers a solution, enabling precise measurement of finger angles, even in regions overlapped by external entities.

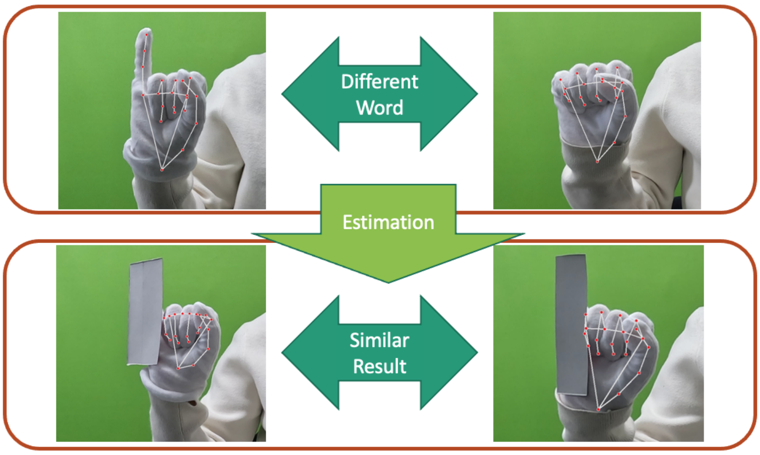

**Figure 1.** Occlusion problem in hand sign language.

We hope to improve the stability of sign language recognition, verify the information complementarity of data gloves and cameras, and study appropriate data fusion methods.

In this research, we integrate a wearable-sensor-based system with a camera-based approach to enhance the precision of hand sign language capture. One inherent challenge in extracting skeletal information for sign language is addressing occlusions among fingers and accessing spatial data that is unattainable by standalone camera systems.

To address this, our proposed system leverages hand skeletons as delineated by MediaPipe for sign language prediction. We adopt a hybrid methodology, intertwining Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) models, to bolster our sign language recognition capabilities. CNN is good at extracting relationships between features, and BiLSTM models are adept at temporal data feature comprehension, rendering them ideal for action-oriented tasks such as sign language interpretation. Through this CNN + BiLSTM amalgamation, we have achieved superior recognition accuracy compared to single-sensor solutions.

The contributions of this research are itemized below:

1. To the best of our knowledge, this is the first time that the data from the bending sensor and the keypoint data calculated by the camera have been fused to study sign language recognition.
2. Our devised system integrates visual and bending sensor inputs. Visual data are utilized to extract essential keypoints and joint angles while eliminating redundancy. This approach mitigates the influence of background and lighting variations, enhancing the system's generalizability and data efficiency. The flex sensor captures finger flexion patterns, enabling adaptability across diverse environments.
3. We amalgamated keypoint coordinates, finger joint angles, and curvature features, strategically combining multifaceted information at the feature level. This integration forms the foundation for our CNN–BiLSTM model, facilitating information synergy and effectively enhancing recognition rates.

Existing works based on camera systems have achieved high recognition rates, but there are problems such as light and shadow, occlusion, and so on. For the first time, we propose the use of data gloves in combination with cameras to utilize information redundancy and improve the stability of the system.

This paper consists of six sections. Section 1 explains sign language recognition, outlines the goals, issues, solutions, and contributions of this research. Section 2 introduces related works. In this section, we will introduce papers on sign language recognition and hand skeleton prediction, and clarify the purpose of this research. Section 3 describes the method of this research. It contains the overall system structure, including the bending sensor, hand keypoint estimation method, and recognition method, introduced respectively. Section 4 describes the implementation of this study, including the implementation of curved sensor data gloves, sign language dataset collection, and data fusion. Section 5 describes the experiments and evaluation of this study. The experiments compare methods using only image data and methods that fuse images and bending sensors. Section 6 presents the discussion and conclusions of this work, along with the current problems in research and some directions for future research.

## 2. Related Works

### *2.1. Data Gloves System*

The main research directions in sign language recognition include computer vision systems and systems based on data gloves. In recent years, the evolution of wearable hand measurement devices has been evident, predominantly driven by miniaturization processes and advancements in algorithms. Notably, data gloves [4,5], including IMU [6] and bending sensors [7,8], have demonstrated significant advancements in wearability, accuracy, and stability metrics. Such advancements have consequently led to marked enhancements in the results of sign language recognition leveraging these measurement apparatus. The application model for sign language recognition based on data gloves is shown in Figure 2.

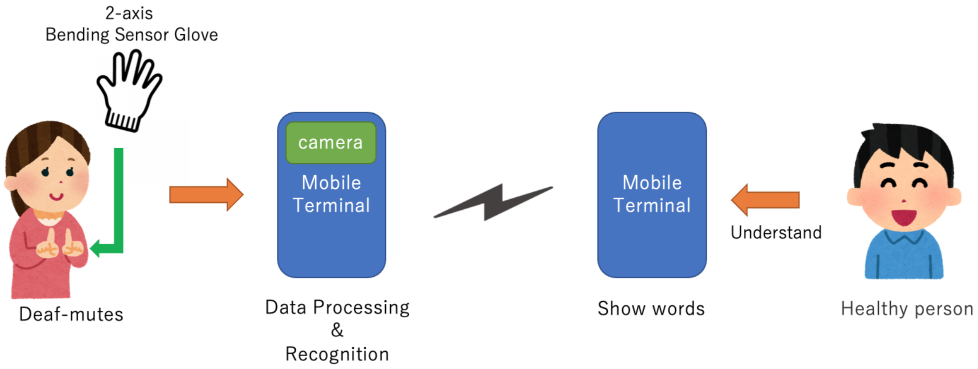

**Figure 2.** Application model.

### *2.2. Vision-Based Techniques*

There are many studies on sign language recognition solutions based on computer vision [9,10]. With the evolution of deep learning algorithms, the extraction and analysis of features from visual data, including bone keypoint prediction [11], have substantially improved. While sign language recognition has experienced significant advancements, occlusions in images remain a notable challenge in computer vision. Himanshu and Sonia's review discusses the effects of occlusion on the visual system [12]. There are ways to avoid occlusion problems by using a depth camera, multiple cameras, or labeling invisible objects. There are also methods to detect occlusion, such as using shadows of objects and learning information before and after occlusion using time series data. Although motion capture using a special device such as Kinect [13] and Leap Motion Controller (LMC) [14] exist, sign language recognition using a monocular camera is superior in that it can use a common camera.

Many vision-based studies based on deep learning methods have been proposed. Deep Rameshbhai et al. [15] proposed Deepsign to recognize isolated Indian Sign Language in video frames. The method combined LSTM and GRU and achieved approximately 97% accuracy on 11 different signs.

Arun Singh et al. [16] proposed a model based on sign language recognition (SLR) of dynamic signs using Convolutional Neural Network (CNN), achieving a training accuracy of 70%. Avola et al. [17] used the SHREC dataset to perform sign language recognition. SHREC is a dataset that uses a depth camera to acquire gesture skeletons. DLSTM, a deep LSTM, is used for sign language recognition. In their method, SHREC is utilized, wherein the angles formed by the fingers of the human hand, calculated from the predicted skeleton, are used as features. The training using SHREC and DLSTM enables highly accurate sign language recognition.

Existing work in hand pose estimation includes the following. Liuhao Ge et al. [18] proposed Hand PointNet. This method directly processes 3D point cloud data representing the hand's visible surface for pose regression. It incorporates a fingertip refinement network, surpasses existing CNN-based methods, and achieves superior performance in 3D hand pose estimation. Nicholas Santavas et al. [19] introduced a lightweight Convolutional Neural Network architecture with a Self-Attention module suitable for deployment on embedded systems, offering a non-invasive vision-based human pose estimation technology for various applications in Human-Computer Interaction with minimal specialized equipment requirements. Liuhao Ge et al. [20] explained the prediction of the skeleton of the hand from image recognition. It estimates the complete 3D hand shape and poses from a monocular RGB image.

Multimodal sensor data fusion methods are crucial in systems that combine bending sensors and vision. CNN [21] and BiLSTM [22] methods can obtain information from spatial and time series data, respectively. The fusion of CNN and BiLSTM [23,24] has been used in the field of Natural language processing. Moreover, the skeleton of the hand is extracted from videos using a method called MediaPipe [25]. In addition, by using the sensor, we can expect to measure the angle of the finger more accurately even in the part that overlaps other objects. Therefore, combining sensor data with sign language recognition will make it possible to accurately predict hand movements.

A comparison of related work is shown in Table 1. Sign language recognition mainly includes two types: data-glove-based and camera-based. Systems based on data gloves generally use bending sensors and IMUs to obtain key point information on the hand skeleton, and the amount of information is less than that of camera systems. The camera system's recognition rate will decrease due to line-of-sight occlusion, darkness, or overexposure. Therefore combining cameras and data gloves is a potential solution.

**Table 1.** Comparison of related research.

| Researches | Sensor | Input Features | Fusion Algorithm | Occlusion Data |
|---|---|---|---|---|
| Our | Camera and Bending | Hand Landmarks,<br>Finger Bending | CNN-BiLSTM | ○ |
| Chu et al. [26] | Bending sensor | Finger Bending | DTW | Unnecessary |
| Clement et al. [27] | IMU, Bending sensor | Orientation<br>Finger Bending | HMM | Unnecessary |
| Samaan et al. [11] | Camera | Hand Landmarks | Bi-LSTM | × |
| Rao et al. [28] | Camera | Hand Landmarks<br>Face Landmarks | LSTM | × |
| Kothadiya et al. [15] | Camera | Images | LSTM and GRU | × |
| Mohammed et al. [9] | Camera | Images | EfficientNetB4 | × |

## 3. Method

The system simultaneously acquires data from bending sensors and vision and uses deep learning methods to fuse the data for sign language recognition.

### 3.1. System Design

The structure of the system is shown in Figure 3. The system comprises two inputs: video collected by the camera and sensor data collected by the bending data glove. Camera data are used to obtain the keypoints of the hand through MediaPipe, and the joint angles of the fingers are obtained through the keypoints. Subsequently, the joint angle data from the keypoints and the finger bending angles from the sensor are combined. The semantics of the sign language are then obtained through CNN + BiLSTM recognition.

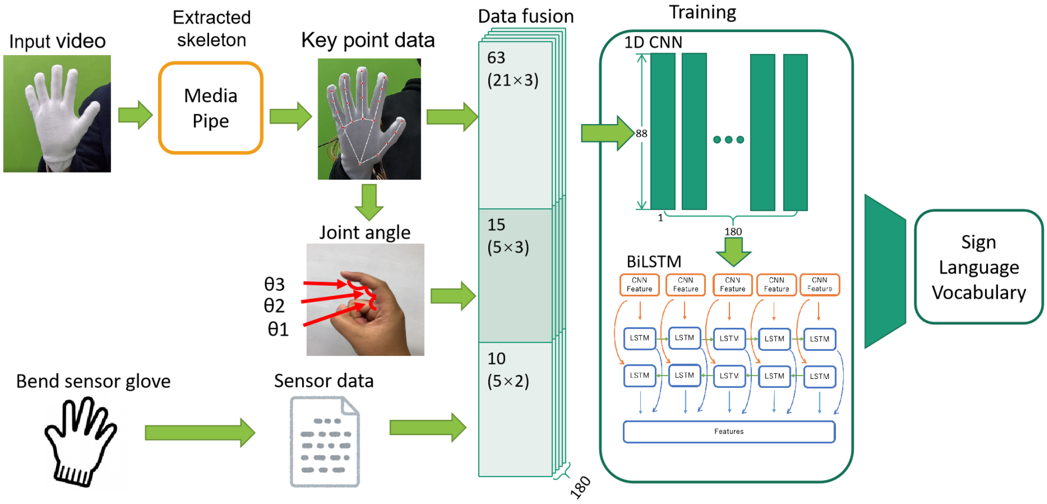

**Figure 3.** Method structure: data collection and training.

### 3.2. MediaPipe

We use MediaPipe to predict skeletons from images. MediaPipe can predict face, posture, and hand skeletons with high accuracy. This method is intended for use with GPUs for real-time inference. However, there are also lighter and heavier versions of the model to deal with CPU inference on mobile devices, which is less accurate than running on desktop computers [29]. Figure 4 shows the output of MediaPipe hand skeleton data. In (a), the predicted 21 keypoint positions are shown. In (b), the points in (a) correspond to the numbers. In (c), an example of using MediaPipe is presented. In this research, the 21 keypoints, indicated by red dots, are utilized as skeleton data for the dataset.

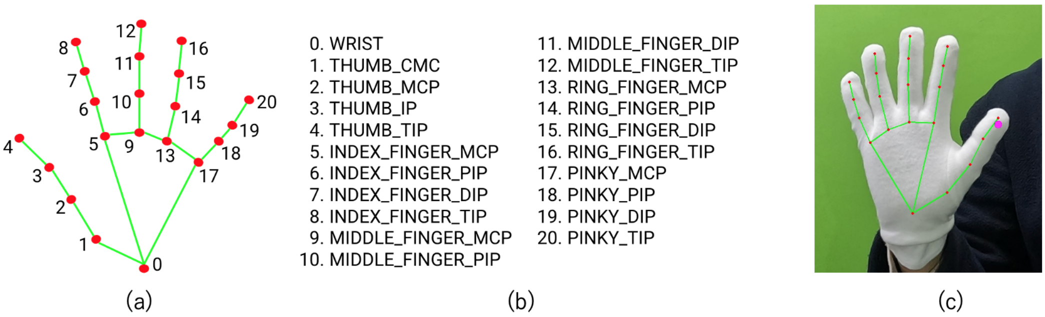

**Figure 4.** Skeleton and bending sensor data fusion. (**a**) Definition of key points of the hand (**b**) Name of each key point. (**c**) Results of real hand recognition of key points.

### 3.3. CNN + BiLSTM

Since video data is used for sign language recognition, a method that processes both spatial information and time series data is effective. Spatial information is learned using

CNN, and time series information is learned using BiLSTM. First, a sign language dataset is input to MediaPipe. MediaPipe outputs the keypoint data of the sign language, which is used as skeleton data. This skeleton data is then input to the CNN to extract spatial information, and then temporal information is extracted by BiLSTM. The spatial and temporal information are then integrated and used to train the model. By combining CNN and BiLSTM, we achieve higher recognition accuracy, as this approach learns both spatial and temporal features more effectively than using either method alone. We use Keras (https://keras.io/) in Python to build our deep-learning network.

Our network structure is also lightweight due to the use of a less data-intensive skeleton and bending sensor data for fusion. The network structure is shown in Figure 5. We use Keras to construct the code, and the sampling frequency is 60 Hz. The input consists of three seconds of data, with each frame containing 88 data points, resulting in a shape of [88, 180]. The 1D-CNN has one layer with 32 filters, a kernel size of 3, and 'valid' padding, followed by one layer of BiLSTM. The learning rate is set to 0.001.

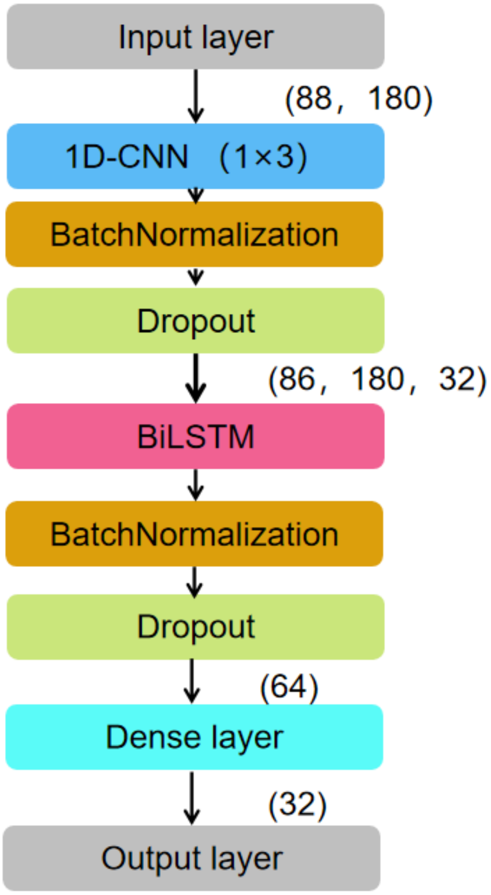

**Figure 5.** The 1D-CNN–BiLSTM model structure diagram.

## 4. Implementation

### *4.1. Outline*

The model of this sign language recognition system is shown in Figure 3. First, we construct a data collection system that includes data gloves and cameras to collect bending data. Then, we create a dataset. This dataset contains video data and finger-bending data from sign language performances. Next, the hand skeleton is predicted from the sign language video. The hand skeleton is estimated using MediaPipe. Finally, the sensor data and skeletal data are fused, and the model is trained using CNN + BiLSTM. The model for gesture estimation is formed.

### *4.2. Bending Sensor Glove Design*

This part describes the design of the original glove, the sensors, the sensor controllers, and the sensor data structures. Figures 6 and 7 shows the actual 2-axis bending sensor glove. The fingertips are securely attached while the rest is loosely secured to ensure that the sensor does not come loose. As a result, parts for fixing the sensors were created using a 3D printer. The fingertip part is designed so that the sensor can be inserted and fixed. Additionally, if the sensor is fixed at every part, it would restrict finger movement, thereby making it difficult to express sign language. Therefore, parts other than the fingertips are not fixed. Furthermore, during actual use, white gloves are worn to conceal the sensors. This prevents MediaPipe from failing to recognize the sensor glove as a hand. Then Raspberry Pi Pico is used as a controller to control the sensor. Note that the sensor gloves produce different values depending on the individual using them, even when the same hand pose is applied.

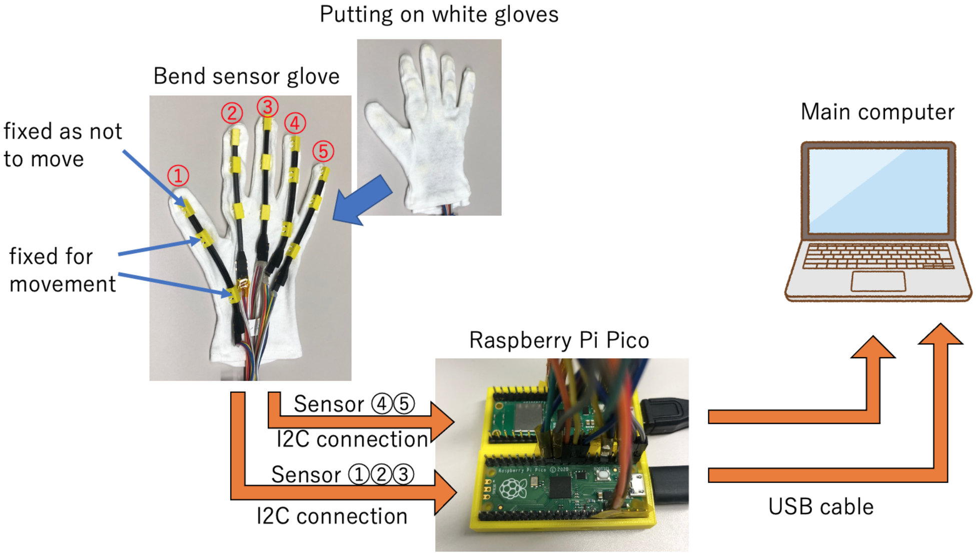

**Figure 6.** Sensor glove design. The five bending sensors (1–5) transmit data to the PC through two Raspberry Pi Pico.

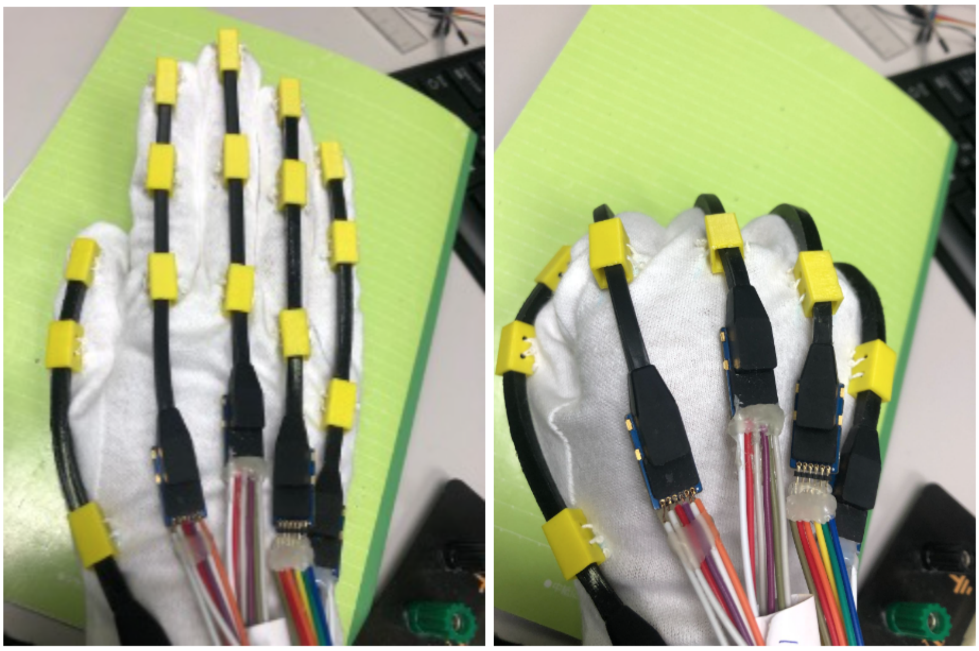

**Figure 7.** Sensor glove.

#### The 2-Axis Bending Sensor

The sensor used is a 2-axis bending sensor, as shown in Figure 8, developed by Bend Labs. Compared to conventional sensors, this 2-axis bending sensor measures angular displacement with greater accuracy. The sensor output is the angular displacement as computed from the vectors defined by the ends of the sensor (v1 and v2) [30].

**Figure 8.** The 2-axis bending sensor from Bend Labs.

### *4.3. Sign Language Dataset*

First, we create a dataset consisting of sign language videos to generate skeleton data. The dataset comprises original data contributed by laboratory members. The dataset includes 32 words from Japanese sign language. The Japanese language is represented by 46 letters. These letters include vowels (a, i, u, e, o) and consonants (k, s, t, n, h, m, y, r, w). The list of letters used in this research is presented. In Japanese, there are letters with only vowels, combinations of vowels and consonants, and special characters, like 'nn'. The table shows consonants in columns and vowels in rows. The first column from the right is for vowels only ('/' indicates no consonants) and "nn" appears at the end of the column for the consonant 'n'.

### *4.4. Image Data Collection*

The dataset has videos of four people for each word shot at 60 fps with a green screen background. The sensor glove is worn on the right hand. Sign language words are basically fixed, such as clenching a fist or raising only the index finger, without moving the hand. However, some sign language words are expressed by moving the hand. For example, "ri", "no", "nn", and "mo" are in the word list. "mo" is a finger movement only, but "ri", "no", and "nn" are expressed by moving the wrist.

#### 4.4.1. Key Point Estimation

We predict skeletal data from videos of sign language, where the participants are wearing sensor gloves. MediaPipe estimates 21 keypoints and makes them skeleton data. Keypoint coordinates are 3D (x, y, z) and 60 frames are acquired per second.

#### 4.4.2. Calculating Joint Angle

Finger angles are calculated from the skeleton data obtained with MediaPipe. This is useful for data argumentation of the dataset. There is one finger angle for each joint, and angles are calculated by the inner product. For example, the two adjacent segments of the finger are $\vec{a}$ and $\vec{b}$.

$$\vec{a} = (x_j - x_i, y_j - y_i, z_j - z_i) = (a_1, a_2, a_3), \vec{b} = (x_j - x_i, y_j - y_i, z_j - z_i) = (b_1, b_2, b_3)$$

If $\cos\theta$ is the interior angle of $\vec{a} = (a_1, a_2, a_3), \vec{b} = (b_1, b_2, b_3)$, it is calculated by

$$\cos\theta = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}||\vec{b}|} = \frac{a_1 b_1 + a_2 b_2 + a_3 b_3}{\sqrt{a_1^2 + a_2^2 + a_3^3}\sqrt{b_1^2 + b_2^2 + b_3^2}}$$

### *4.5. Collecting Sensor Data*

This section describes the original bending sensor glove and finger angle data collection. We designed and created an original bending sensor glove specifically for collecting finger angles. The glove is worn on the right hand. The data collected from the glove, including time stamps and the 2-axis angles of the five fingers, is saved as a text file on the main computer. Additionally, a video of the sign language performance is recorded simultaneously with the collection of bending sensor data. The finger angles acquired alongside the bending sensor data, together with the simultaneously captured images, support the process of image recognition.

### *4.6. Data Fusion*

Skeleton data are acquired using MediaPipe. Finger joint angles are then calculated from this skeleton data, and subsequently, sensor data is fused with it. As shown in Figure 9. The skeleton data comprise 63 points (21 keypoints × 3 dimensions), The finger joint angle data consist of 15 points (5 fingers × 3 joint angles), and the sensor data include 10 points (5 fingers × 2-axis).

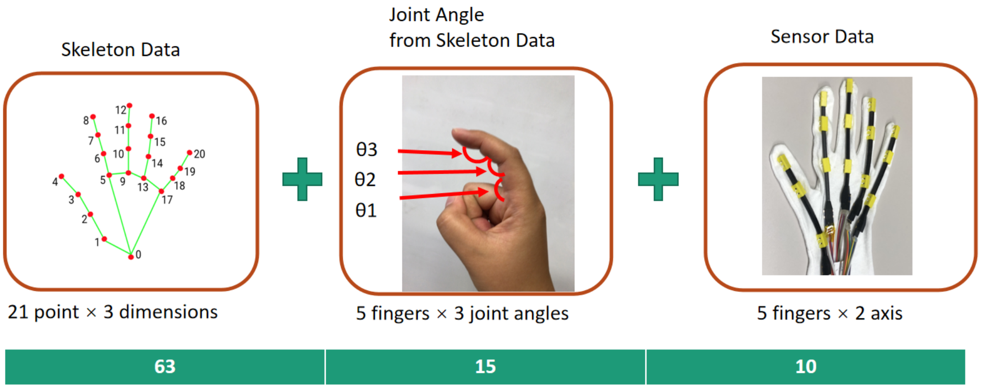

**Figure 9.** Data fusion. (The red arrows indicate the MCP, PIP, and DIP positions of the index finger, and $\theta 1$, $\theta 2$, and $\theta 3$ represent the joint angles.)

## 5. Experiment and Evaluation

### *5.1. Experiment Purpose*

In this experiment, we aimed to evaluate the sign language recognition performance of the fusion system comprising bending sensor gloves and computer vision. The experimental evaluation and discussion will involve comparing the results of sign language recognition using only skeleton data with those using all fused data.

### *5.2. Experiment Setting*

We prepared a bending sensor glove and a camera to collect data. The camera uses GoPro Hero10. Each action was collected for three seconds at one time. The camera, a GoPro Hero10, captures high-resolution images (1080, 1920) at 60 fps. Its compact size allows for fast and efficient recording. Additionally, a green screen was used to maintain uniform background colors. Participants wore sensor gloves to collect both sensor data and video data simultaneously. We collected sign language data from seven participants (three females and four males). The bending sensor data glove we built has a sampling rate of 60 Hz. Thirty-two sign language words were executed, as shown in Figure 10. Each action was repeated 10 times per person, collecting 3 s each time. The sampling rate of the system is 60 Hz. The dataset has 403,200 frames of data collected. In total, approximately 6720 s (112 min) of data were collected.

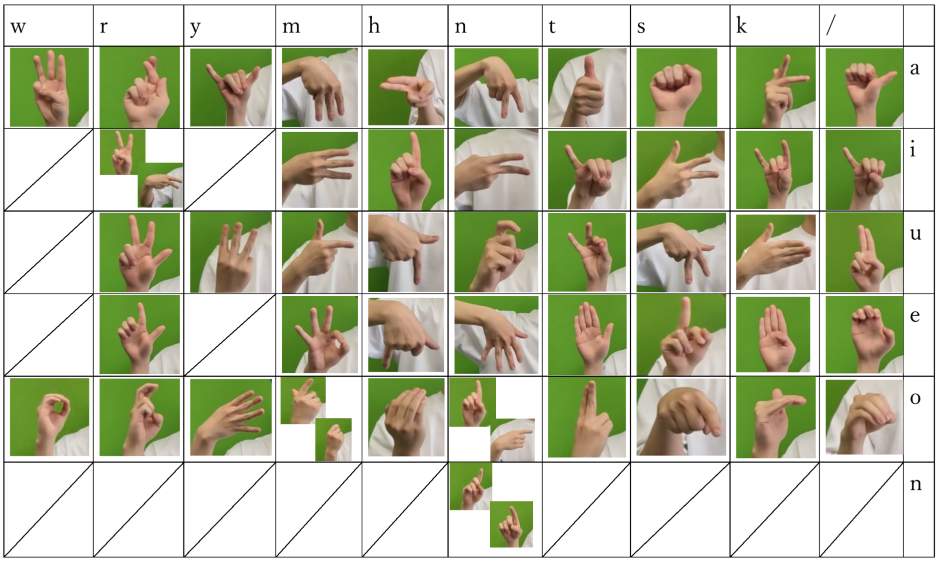

**Figure 10.** Japanese sign language letter list.

To demonstrate the sensor's effectiveness, occlusion was introduced in the sign language videos, which were then processed and recognized by MediaPipe. In this occlusion scenario, a paper strip measuring 15 cm in length and 3 cm in width was used to cover the little finger, and data were collected under this specific type of occlusion. This specific scenario is illustrated in Figure 11.

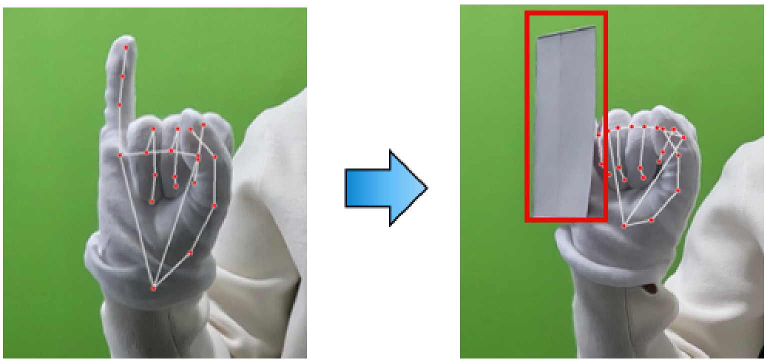

**Figure 11.** Occlusion data. (The occlusion way is as shown in the red box by covering part of the hand with a paper strip.)

### 5.3. Experiment Process

Initially, the subject performed stationary movements while maintaining a flat hand, which provided calibration data for the glove. The gloves and the camera were turned on simultaneously to obtain synchronized data. Subjects were guided through each gesture.

### 5.4. Experiment Results

The model was trained with k-Fold cross-validation. When training with a small dataset, the training accuracy could be misleadingly high. If this is the case, the accuracy in training may be high, but the accuracy in testing may be lowered, resulting in overfitting. To prevent this, a technique known as k-Fold cross-validation is employed. In k-Fold cross-validation, the data is divided into k segments, with some segments used for validation and others for training. Since all the divided data are used once for validation data, training is performed k times. The result is calculated as the average of the k training accuracies. The cross-entropy method is used for calculating the loss function. If the probability distributions of p and q are approximate, the cross-entropy loss is smaller. In other words, the closer the learning accuracy approaches 1, the closer the result approaches 0. Results for the skeleton data only are shown below. Additionally, the training and evaluation data are split at a ratio of 4:1. The training data is used for the training process, evaluation data for evaluation during training, and test data for the final model evaluation.

The training curve is shown in Figures 12 and 13. The blue line represents the accuracy of training, the orange line represents the accuracy of validation, and the green line represents the accuracy of validation when using test data. The confusion matrix using only skeleton data or using fused data is shown in Figures 14 and 15. In the skeleton-only validation, cross-validation was performed five times, resulting in an average training accuracy of 85.9% and an accuracy of 73.5% when using test data. In the fusion data validation, cross-validation was also performed five times. The average training accuracy was 99.2%, while the accuracy was 96.5% for training and 84.13% for test data.

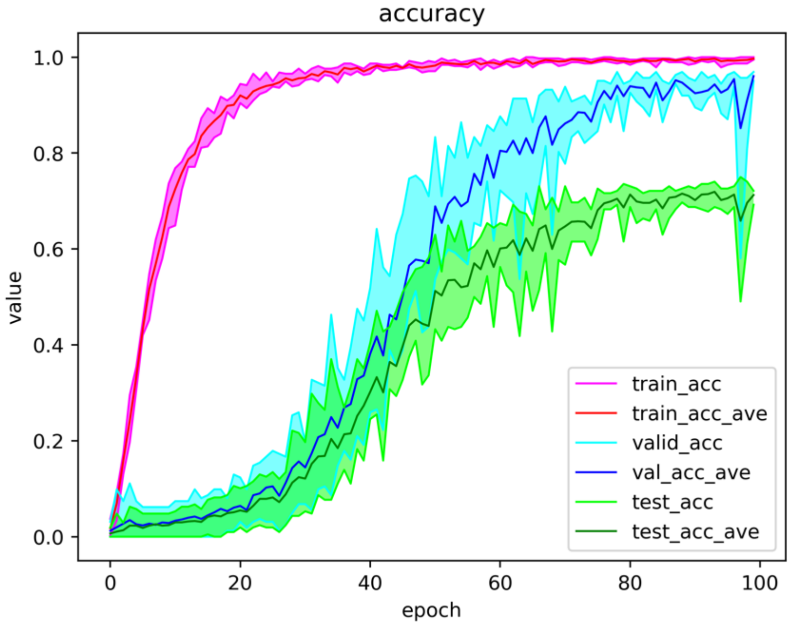

**Figure 12.** Accuracy curve of only skeleton data.

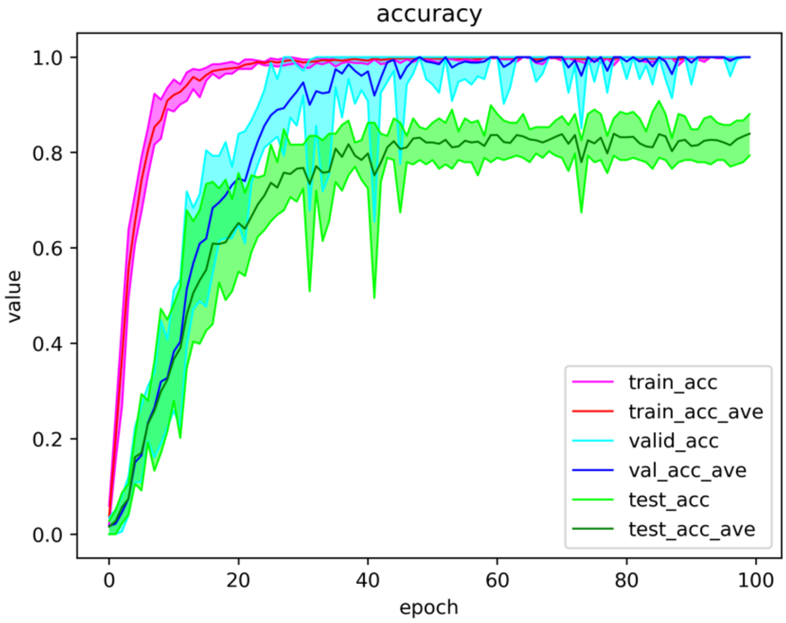

**Figure 13.** Accuracy curve of fusion data.

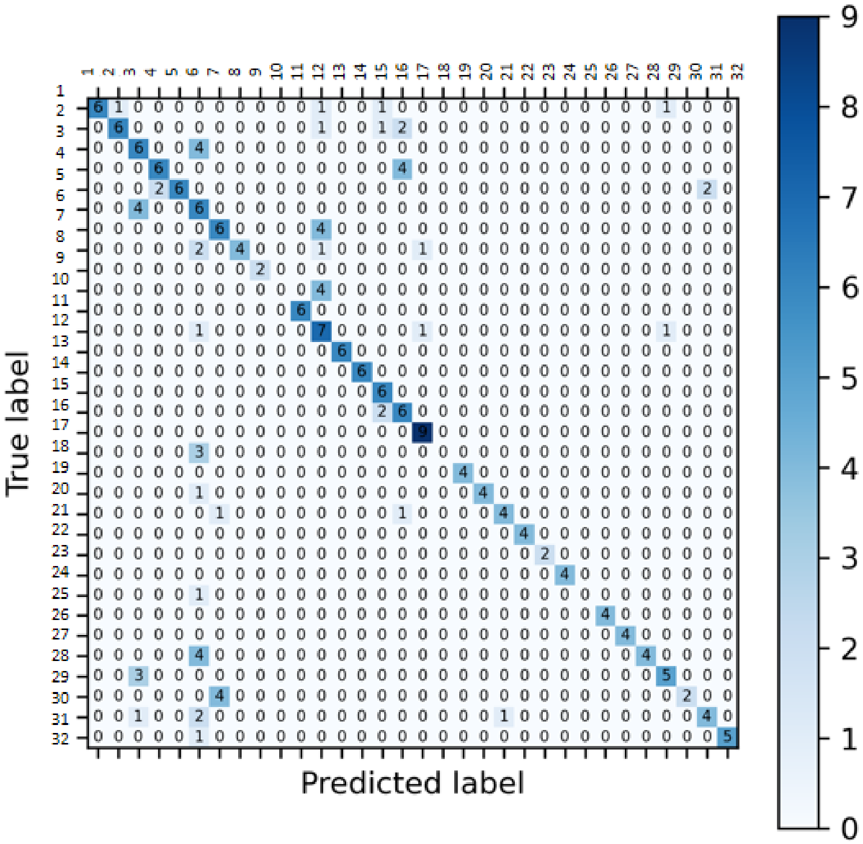

**Figure 14.** Confusion Matrix: Only Skeleton Data.

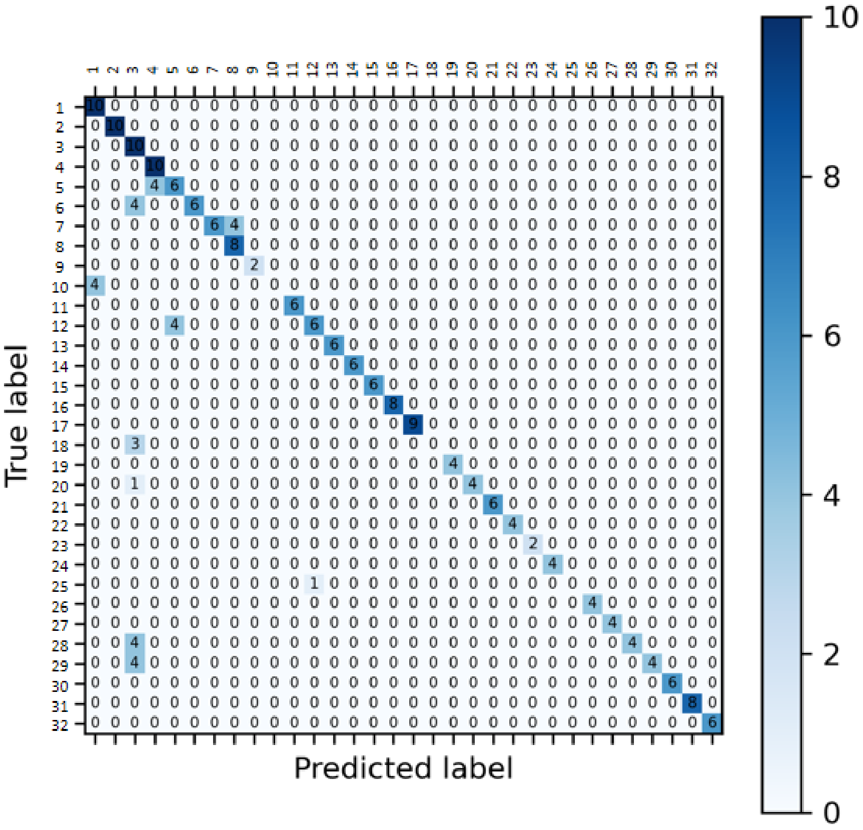

**Figure 15.** Confusion Matrix: Fusion Data.

### *5.5. Discussion*

The overall recognition rate of the fused system is improved compared to using only skeleton data. Furthermore, the fused system requires fewer epochs to achieve a stable recognition rate and exhibits lower overfitting. However, there are some areas that require improvement. First, some sign language movements are indistinguishable using only bending sensors. This is because the values recorded by the bending sensors are exactly the same, leading to conflicts in recognition judgments. Additionally, when recognizing partially similar sign languages, the similarity in added sensor data values results in lower recognition accuracy for certain actions. When the recognition effect using only skeleton data is poor, the sensor has a complementary effect.

Our primary comparison involved the use of features obtained solely by the camera versus a combination of information from both the bending sensor and the camera. The result of fusing the two pieces of information is better. First of all, this shows that the information from the bending sensor and the camera are complementary. Second, the fusion algorithm we use can effectively use complementary information to improve the recognition rate.

The input end of the algorithm model employs feature splicing. It uses 1D-CNN to analyze the relationship between features, followed by BiLSTM, which analyzes the temporal relationship of features. The specific parameters of the model are obtained through the backpropagation algorithm and data-driven training.

In our study on system robustness, we identified that the primary sources of errors were attributable to variations in environmental conditions and individual differences among users. We observed that environmental factors, particularly background elements, could potentially skew the results. To address this, our implementation leverages MediaPipe for keypoint extraction from images. MediaPipe's effectiveness in background suppression is a result of its training on extensive datasets, allowing for more accurate keypoint detection irrespective of varied backgrounds. Furthermore, data variability due to individual differences poses a challenge to the robustness of our system. We will address these shortcomings by expanding the dataset to include broader data reflecting different individuals.

## 6. Conclusions

In this research, we aimed to improve sign language recognition with occlusion accuracy by combining CNN + BiLSTM and also combining bending sensor data with skeleton data. The combination of the CNN + BiLSTM method with sensor data enabled better finger character recognition than using either the CNN or BiLSTM method alone. However, there were limitations in acquiring spatial information, such as blind spot problems. Therefore, we used a 2-axis bending sensor to assist with spatial information. The performance evaluation of the original 2-axis bending glove further strengthened the spatial information of sign language. By using sensor data, we were able to improve sign language recognition accuracy in the presence of occlusion compared to skeleton data alone.

However, there are still some problems in the system that need to be solved urgently. The current handling of occlusion data is relatively simplistic. There is a lack of consideration for varying conditions such as shadows and overexposure. This may lead to a decrease in accuracy in practical applications. Additionally, the bending sensor alone cannot distinguish between certain sign language movements that have identical finger bends, thereby limiting its effectiveness in providing discriminative information. When the camera is blocked, there are limitations to using only bending sensors for supplementary information. Therefore, it is necessary to integrate inertial sensors to capture the hand's posture and the upper limb's skeleton.

Future work can be made to provide the system with complementary hand movement measurement data of more different modalities and try to improve the data fusion method. The recognition rate using only bending sensors is very low. Integrating IMU sensors into the data gloves presents an opportunity to further enhance the fusion system's sign language recognition rate in occlusion situations. Furthermore, the credibility or reliability of the bending sensor and camera fluctuates with environmental changes. Introducing Trust-Level Fusion to combine the data could potentially improve system stability.

**Author Contributions:** Conceptualization, C.L. and M.K.; methodology, C.L. and M.K.; software, M.K. and C.L.; validation, C.L. and M.K.; formal analysis, C.L.; investigation, C.L.; resources, L.J.; data curation, M.K.; writing—original draft preparation, M.K.; writing—review and editing, C.L.; visualization, M.K. and C.L.; supervision, L.J.; project administration, L.J.; funding acquisition, L.J. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by JSPS KAKENHI Grant Number 22K12114, the JKA Foundation, and NEDO Intensive Support for Young Promising Researchers Number JPNP20004.

**Data Availability Statement:** Data are contained within the article.

**Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

## References

1. World Health Organization. World Report on Hearing. 2021. Available online: https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss (accessed on 18 September 2023).
2. Adeyanju, I.; Bello, O.; Adegboye, M. Machine learning methods for sign language recognition: A critical review and analysis. *Intell. Syst. Appl.* **2021**, *12*, 200056. [CrossRef]
3. Joksimoski, B.; Zdravevski, E.; Lameski, P.; Pires, I.M.; Melero, F.J.; Martinez, T.P.; Garcia, N.M.; Mihajlov, M.; Chorbev, I.; Trajkovik, V. Technological Solutions for Sign Language Recognition: A Scoping Review of Research Trends, Challenges, and Opportunities. *IEEE Access* **2022**, *10*, 40979–40998. [CrossRef]
4. Amin, M.S.; Rizvi, S.T.H.; Hossain, M.M. A Comparative Review on Applications of Different Sensors for Sign Language Recognition. *J. Imaging* **2022**, *8*, 98. [CrossRef] [PubMed]
5. Al-Qurishi, M.; Khalid, T.; Souissi, R. Deep Learning for Sign Language Recognition: Current Techniques, Benchmarks, and Open Issues. *IEEE Access* **2021**, *9*, 126917–126951. [CrossRef]
6. Lu, C.; Dai, Z.; Jing, L. Measurement of Hand Joint Angle Using Inertial-Based Motion Capture System. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 1–11. [CrossRef]
7. Faisal, M.; Abir, F.F.; Ahmed, M.; Ahad, M.A.R. Exploiting domain transformation and deep learning for hand gesture recognition using a low-cost dataglove. *Sci. Rep.* **2022**, *12*, 21446. [CrossRef] [PubMed]
8. Lu, C.; Amino, S.; Jing, L. Data Glove with Bending Sensor and Inertial Sensor Based on Weighted DTW Fusion for Sign Language Recognition. *Electronics* **2023**, *12*, 613. [CrossRef]
9. Zakariah, M.; Alotaibi, Y.A.; Koundal, D.; Guo, Y.; Elahi, M.M. Sign Language Recognition for Arabic Alphabets Using Transfer Learning Technique. *Comput. Intell. Neurosci.* **2022**, *2022*, 4567989. [CrossRef] [PubMed]
10. Mukai, N.; Yagi, S.; Chang, Y. Japanese Sign Language Recognition based on a Video accompanied by the Finger Images. In Proceedings of the 2021 Nicograph International (NicoInt), Tokyo, Japan, 9–10 July 2021; pp. 23–26.
11. Samaan, G.H.; Wadie, A.R.; Attia, A.K.; Asaad, A.M.; Kamel, A.E.; Slim, S.O.; Abdallah, M.S.; Cho, Y.I. MediaPipe's Landmarks with RNN for Dynamic Sign Language Recognition. *Electronics* **2022**, *11*, 3228. [CrossRef]
12. Purkait, P.; Zach, C.; Reid, I.D. Seeing Behind Things: Extending Semantic Segmentation to Occluded Regions. In Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 3–8 November 2019; pp. 1998–2005.
13. Zhang, Z. Microsoft Kinect Sensor and Its Effect. *IEEE Multim.* **2012**, *19*, 4–10. [CrossRef]
14. Guna, J.; Jakus, G.; Pogacnik, M.; Tomažič, S.; Sodnik, J. An Analysis of the Precision and Reliability of the Leap Motion Sensor and Its Suitability for Static and Dynamic Tracking. *Sensors* **2014**, *14*, 3702–3720. [CrossRef] [PubMed]
15. Kothadiya, D.; Bhatt, C.; Sapariya, K.; Patel, K.; Gil-González, A.B.; Corchado, J.M. Deepsign: Sign Language Detection and Recognition Using Deep Learning. *Electronics* **2022**, *11*, 1780. [CrossRef]
16. Singh, A.; Wadhawan, A.; Rakhra, M.; Mittal, U.; Ahdal, A.A.; Jha, S.K. Indian Sign Language Recognition System for Dynamic Signs. In Proceedings of the 2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 13–14 October 2022; pp. 1–6. [CrossRef]
17. Avola, D.; Bernardi, M.; Cinque, L.; Foresti, G.L.; Massaroni, C. Exploiting Recurrent Neural Networks and Leap Motion Controller for the Recognition of Sign Language and Semaphoric Hand Gestures. *IEEE Trans. Multimed.* **2018**, *21*, 234–245. [CrossRef]
18. Ge, L.; Cai, Y.; Weng, J.; Yuan, J. Hand PointNet: 3D Hand Pose Estimation Using Point Sets. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8417–8426. [CrossRef]

19. Santavas, N.; Kansizoglou, I.; Bampis, L.; Karakasis, E.G.; Gasteratos, A. Attention! A Lightweight 2D Hand Pose Estimation Approach. *IEEE Sens. J.* **2020**, *21*, 11488–11496. [CrossRef]
20. Ge, L.; Ren, Z.; Li, Y.; Xue, Z.; Wang, Y.; Cai, J.; Yuan, J. 3D Hand Shape and Pose Estimation from a Single RGB Image. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 10825–10834.
21. O'Shea, K.; Nash, R. An Introduction to Convolutional Neural Networks. *arXiv* **2015**, arXiv:1511.08458.
22. Zhang, S.; Zheng, D.; Hu, X.; Yang, M. Bidirectional long short-term memory networks for relation classification. In Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation, Shanghai, China, 30 October–1 November 2015; pp. 73–78.
23. Chiu, J.P.C.; Nichols, E. Named Entity Recognition with Bidirectional LSTM-CNNs. *Trans. Assoc. Comput. Linguist.* **2015**, *4*, 357–370. [CrossRef]
24. Kavianpour, P.; Kavianpour, M.; Jahani, E.; Ramezani, A. A CNN-BiLSTM Model with Attention Mechanism for Earthquake Prediction. *arXiv* **2021**, arXiv:2112.13444.
25. Lugaresi, C.; Tang, J.; Nash, H.; McClanahan, C.; Uboweja, E.; Hays, M.; Zhang, F.; Chang, C.L.; Yong, M.G.; Lee, J.; et al. MediaPipe: A Framework for Building Perception Pipelines. *arXiv* **2019**, arXiv:1906.08172.
26. Chu, X.; Liu, J.; Shimamoto, S. A Sensor-Based Hand Gesture Recognition System for Japanese Sign Language. In Proceedings of the 2021 IEEE 3rd Global Conference on Life Sciences and Technologies (LifeTech), Nara, Japan, 9–11 March 2021; pp. 311–312.
27. Faisal, M.A.A.; Abir, F.F.; Ahmed, M.U. Sensor Dataglove for Real-time Static and Dynamic Hand Gesture Recognition. In Proceedings of the 2021 Joint 10th International Conference on Informatics, Electronics & Vision (ICIEV) and 2021 5th International Conference on Imaging, Vision & Pattern Recognition (icIVPR), Kitakyushu, Japan, 16–20 August 2021; pp. 1–7.
28. Rao, G.M.; Sowmya, C.; Mamatha, D.; Sujasri, P.A.; Anitha, S.; Alivela, R. Sign Language Recognition using LSTM and Media Pipe. In Proceedings of the 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 17–19 May 2023; pp. 1086–1091. [CrossRef]
29. Zhang, F.; Bazarevsky, V.; Vakunov, A.; Tkachenka, A.; Sung, G.; Chang, C.L.; Grundmann, M. MediaPipe Hands: On-device Real-time Hand Tracking. *arXiv* **2020**, arXiv:2006.10214.
30. Soft Angular Displacement Sensor Theory Manual. 2018. Available online: https://www.nitto.com/us/en/others/nbt/assets/pdf/ad_theory_guide.pdf (accessed on 18 September 2023).

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.