electronics

MDPI

Article

EM-Sign: A Non-Contact Recognition Method Based on 24 GHz Doppler Radar for Continuous Signs and Dialogues

Linting Ye, Shengchang Lan *, Kang Zhang and Guiyuan Zhang

Department of Microwave Engineering, School of Electronic and Information Engineering, Harbin Institute of Technology, Harbin 150001, China; 20S005012@stu.hit.edu.cn (L.Y.); zhangkang@hit.edu.cn (K.Z.); 19S005011@stu.hit.edu.cn (G.Z.)

* Correspondence: lansc1015@hit.edu.cn

Received: 31 August 2020; Accepted: 22 September 2020; Published: 26 September 2020

Abstract: We studied continuous sign language recognition using Doppler radar sensors. Four signs in Chinese sign language and American sign language were captured and extracted by complex empirical mode decomposition (CEMD) to obtain spectrograms. Image sharpening was used to enhance the micro-Doppler signatures of the signs. To classify the different signs, we utilized an improved Yolov3-tiny network by replacing the framework with ResNet and fine-tuned the network in advance. This method can remove the epentheses from the training process. Experimental results revealed that the proposed method can surpass the state-of-the-art sign language recognition methods in continuous sign recognition with a precision of 0.924, a recall of 0.993, an F1-measure of 0.957 and a mean average precision (mAP) of 0.99. In addition, dialogue recognition in three daily conversation scenarios was performed and evaluated. The average word error rate (WER) was 0.235, 10% lower than in of other works. Our work provides an alternative form of sign language recognition and a new approach to simplify the training process and achieve a better continuous sign language recognition effect.

Keywords: continuous sign language recognition; Doppler radar; time-frequency analysis; complex empirical mode decomposition; Yolov3-tiny

1. Introduction

Human–computer interactions (HCIs) have developed prosperously in recent years, allowing various non-contact and flexible methods for different industrial and daily life backgrounds. For example, hand gestures can be used as inputs in many HCI scenarios and hand gesture recognition (HGR) is a key component widely applied in motion sensory games; intelligent driving; and assisting deaf communities and hearing societies [1,2]. Widely used in these disabled communities, sign language is considered a dynamic sequence of hand gestures that conveys meaningful semantic expressions from the brain and thus has received extensive attention in the latest HCI studies. The current sign language recognition available can be either sensor-based or vision-based. Sensor-based sign language recognition is achieved via sensors embedded in the data gloves [3] or mechanomyogram signals [4]. These sensor-based sign language recognition devices are relatively expensive and inconvenient to wear or carry. Vision-based sign language recognition utilizes images or videos of signs captured by cameras, and deep learning as the classifier to reach a fine recognition accuracy. In [5], alphabets were classified based on images and a deep neural network with a test accuracy of up to 70%. In [6], three specific words were recognized with videos based on a deep learning network with a final recognition accuracy of up to 91%. In [7], four dynamic words were distinguished by analyzing the video sequence with a combination of 3D residual convolutional

neural network and bi-directional long short-term memory (LSTM) networks. However, vision-based sign language recognition is highly affected by light and has security concerns about privacy leakages. Thus, exploration of new sign language recognition solutions is still in the spotlight of the HCI field to improve the status quo. Inspired by the rapid development in millimeter-wave radar sensors, electromagnetic (EM) waves are an emerging alternative in HCI. The Soli project by Alphabet utilized a customized 60 GHz FMCW radar chipset to manipulate a wearable device [8]. Zhang et al. [9] used continuous wave radar to collect data and a support vector machine (SVM) to classify four gestures, producing a classification accuracy of 88.56%. Kim et al. [10] used a 5.8 GHz Doppler radar combined with a deep convolutional neural network (DCNN) and achieved an accuracy of 87%. Therefore, EM wave is able to become the third medium which semantically represents sign language for deaf and dumb communities. Nevertheless, wielding EM waves to translate sign language is different from most state-of-the-art HGR problems in [11]. Hand gestures studied in the research based on radar sensors are commonly simple motions, such as directional movement or finger rotation. In contrast, sign language contains more versatile expressions of joint flexion and rotation [12]. Meanwhile, most HGR studies focus more on alphabets, numbers and single word recognition among the existing identification methods, and little attention is paid to the practical recognition of continuous signs or dialogues.

Therefore, this paper proposes to use EM waves to represent continuous sign language at the sentential level. Continuous signs consist of individual signs and the epentheses between adjacent signs, which are useless to the study. Without segmentation in terms of frames, we applied a complex empirical mode decomposition (CEMD) to extract non-stationary micro-Doppler signatures on the millimeter-wave radar echo spectrograms. Additionally, an improved yolov3 network was constructed to classify four signs in Chinese sign language (CSL) and American sign language (ASL) datasets. Accordingly, the rest of the paper is organized as follows: Section 2 shows the state-of-the-art of radar sensing and the its characteristics compared to other data modalities. Section 3 gives the methodology of CEMD-based time-frequency analysis. Section 4 presents an improved fine-tuned yolov3-tiny network to dynamically find the sign from the time-frequency spectrograms. After the experimental analysis in Sections 5 and 6, we draw the conclusions in Section 7.

2. State-Of-The-Art of Radar Sensing

2.1. State-Of-The-Art and Characteristic of Radar Sensing

In recent years, radar sensing has started to gain significant interest in many fields beyond defense and air-traffic control, opening new frontiers in radar [13]. With different modulations of transmitting waveforms, the capabilities of short-range radars, including Doppler radar, frequency modulated continuous wave (FMCW) radar sensors, stepped frequency continuous wave (SFCW) devices, and impulse response ultra-wideband (IR-UWB) radars, are also being investigated in automotive [14], indoor navigation [15] and biomedical and health care [16] industries. Featured for its simple structure, low cost and high integration in package, Doppler radar sensors in the millimeter wave band provide an appealing high sensitivity to moving targets and enable numerous applications, such as hand gesture, human activity and vital sign detections, which traditionally relied on visual images.

In sign language recognition, both visual and radar images for a sign are highly related to the pose of the hand and how it temporally evolves. However, the information that they can provide is extremely different. There are many released visual image datasets to support hand pose estimation and further promote the sign language recognition effectively [7,17–19]. These datasets consist of hundreds of thousands of RGB 2D or 3D images and videos to describe the hand in different poses with different backgrounds and illuminations. Additionally, images in the dataset can be arbitrarily rotated, flipped and cropped to enhance the robustness of the recognition. However, those functions cannot be used in radar sensing. Due to the frequency band, modulation bandwidth and modulated waveforms of different radars, data consistence can not be easily guaranteed. Unlike video, radar measurements are not inherently images, but actually form a time-stream of complex I/Q data from which line-of-sight

distance and radial velocity may be computed [20]. In this work, the radar echo signals require further processing to become visualized in the modality of micro-Doppler signatures of motion. This modality shows the correspondence of time and Doppler rather than hand poses. The descriptions of this correspondence do not rely on the light illumination and are barely affected by the background, as only moving targets can be shown in the spectrograms after time-frequency analysis. Although not supported by large standardized datasets, radar sensing is favored as a new approach in sign language recognition.

2.2. State-Of-The-Art of 24 Ghz Radar Integrated Implementation

Note that 24 GHz radar is widely used in automotive anti-collision and vital sign recognition. Due to the frequency band which they work in, 24 GHz radar sensors can reduce the size of the antennas. With the development of microwave integrated circuits (MIC), the radar components can be integrated on a planar structure with the antennas on one side and the baseband circuits on the other side. At present, the majority of companies have produced their own radar chip products for different needs. Single-chip radars, such as the front-end transceiver chipset for the radar system from Freescale and Infienon, provide highly integrated technology solutions in many scenarios. Another type of low-cost radar is composed of transistors and microstrip circuits, such as the radar sensor from RFbeam used in our work. Moreover, many integrated circuit companies, such as Texas Instruments and Analog Devices, have extended their product lines to the millimeter wave band. They have released many new independent RF modules on the shelf, such as mixers, amplifiers and local oscillators. The all-round and open framework makes radar design more flexible.

3. Sign Language Representation Based on Doppler Radar

3.1. Micro-Doppler Signatures of Sign Language

A Doppler radar can transmit EM wave at a wavelength of λ and the signal can be mathematically expressed as

$$s(t) = cos(2\pi ft + \phi(t)) \tag{1}$$

where f and $\phi(t)$ represent the carrier frequency and phase. The echo signal can be expressed as

$$r(t) = cos(2\pi ft - \frac{4\pi d_0}{\lambda} - \frac{4\pi x(t)}{\lambda}) \tag{2}$$

where d_0 is the initial distance between object and antenna and $x(t)$ is the distance shift caused by the relative movement. According to Doppler effect, the velocity of the moving object can be measured. The center frequency shift, that is, Doppler frequency, can be expressed as

$$f_d = \frac{2v}{\lambda} cos\theta \tag{3}$$

where θ is the incident angle to radar. Therefore, a sign can be represented by a series of micro-Doppler signatures of echo signals.

3.2. Continuous Sign Representation at the Sentential Level

A daily conversation based on sign language can be visualized by processing the echo signals to find the spectrograms, or micro-Doppler signatures of the hand motion. Figure 1 shows a daily conversation scenario captured by 24 GHz Doppler radar sensor with the sign dialogue segments "How are you?"; "Fine, Thank you"; and "Do you understand? Yes." The top layer in Figure 1 shows the visual hand motions, the middle layer shows the translated spectrograms related to the individual words after time-frequency analysis and the bottom layer shows the translated spectrograms after time-frequency analysis related to the completed sentence. In Figure 1, a movement transition exists

between two consecutive signs, namely, an epenthesis, which results from moving the hands from the end of one sign to the beginning of the next sign. The epentheses do not correspond to any semantic content but can involve changes in hand gestures, sometimes resulting in a crossover between two consecutive sign sentences, even longer than real signs. Consequently, to use EM waves to recognize signs, the solutions to two key problems are needed: (1) describe the micro-Doppler signature in a fine-grained way; (2) identify and remove the epentheses prior to the recognition.

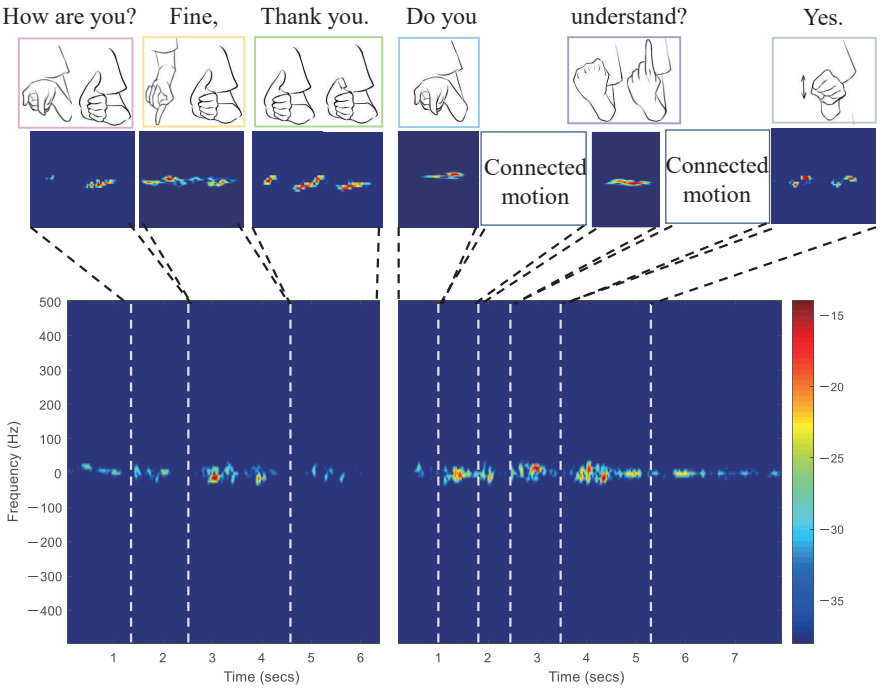


Figure 1. Time-frequency spectrograms to characterize dialogues. The demonstrated dialogue: "How are you? Fine, Thank you"; and "Do you understand? Yes."

3.3. CEMD-Based Time-Frequency Analysis

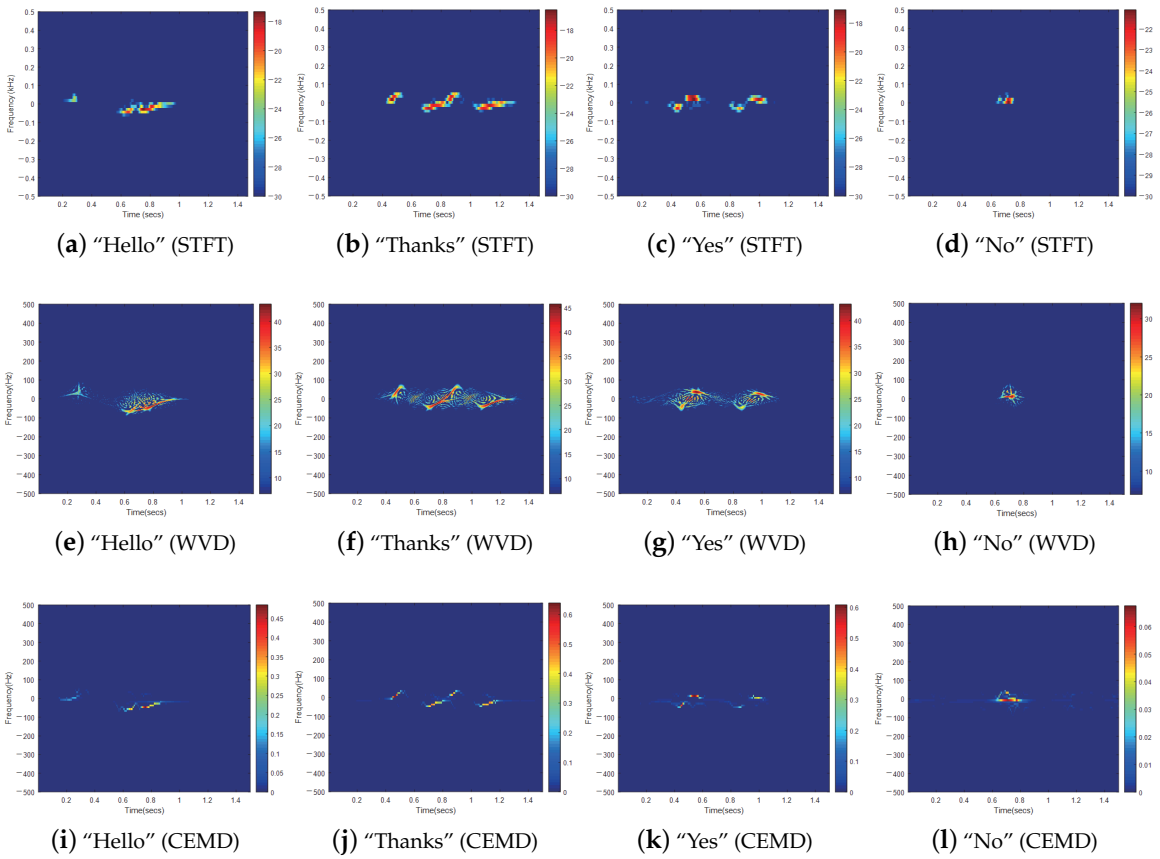
In order to find the intrinsic connection between real signs and frequency spectrograms, time-frequency analysis is frequently used as an effective tool in radar signal processing. Short-time Fourier transform (STFT) and Winger–Ville distribution (WVD) are popular time-frequency analysis methods. In STFT, the time and frequency resolutions are limited due to the fixed window length. Thus the spectrogram obtained by STFT has many lumps caused by low resolution of time-frequency analysis. WVD is a non-linear transformation that leads to the cross-terms in calculating multi-component signals. Although the spectrogram gives the best time-frequency resolution, the cross-terms inevitably affect the recognition of signal terms. Compared to the methods above, with a high time-frequency resolution and shorter processing time, CEMD is utilized to pre-process the signals by many time-frequency analysis methods [21]. As the extension of EMD [22] in the complex domain, CEMD can adaptively decompose the input signal according to its own characteristics. Hence, CEMD for processing in this study could clearly represent the time and frequency distribution. The comparison between STFT, WVD and CEMD is shown in Figure 2. The cell intensity of the spectrogram represents power spectrum density (PSD), expressed by the color intensity. For example, the red region on the spectrogram corresponds to the strong PSD of the radar echo signals. Different signs have different PSD distributions. For a single sign, the peak of the PSD can be always found. However, for a stream of signs, such as a sign sentence with an unlimited length for words or phrases, some signs are invisible because the peak PSD of each sign may be relatively low compared to that of other signs, as shown in Figure 3a.

To avoid this problem, PSD sharpening can be used to enhance the edge of the spectrogram and PSD transition. The original spectrogram is $m(t, f)$ and the sharpened spectrogram $g(t, f)$ can be obtained by

$$\nabla^2 m(t, f) = m(t, f+1) + m(t+1, f) + m(t-1, f) + m(t, f-1) - 4m(t, f) \tag{4}$$

$$\nabla^2 g(t, f) = m(t, f) + \alpha \nabla^2 m(t, f) \tag{5}$$

where α is the enhancement factor and the operator ∇^2 is Laplacian operator. Background noise can be filtered out on the spectrogram by threshold detection described by Equation (5). In this case, a time-frequency spectrogram is transformed more visually clearly from Figure 3a to Figure 3b. Figure 3b also reveals that the sharpened spectrogram makes it easy to make the labels. The categories of the signs have been marked with red boxes.

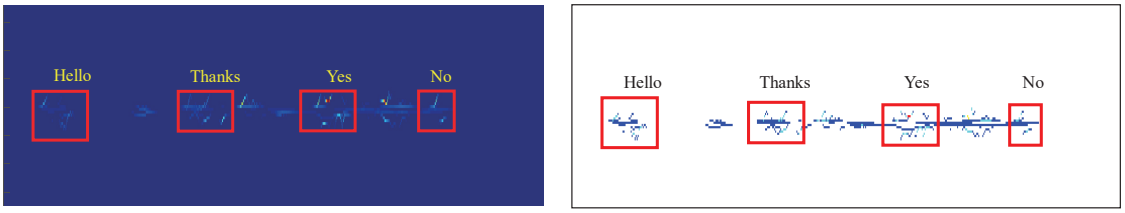


(**a**) "Hello" (STFT) (**b**) "Thanks" (STFT) (**c**) "Yes" (STFT) (**d**) "No" (STFT)

(**e**) "Hello" (WVD) (**f**) "Thanks" (WVD) (**g**) "Yes" (WVD) (**h**) "No" (WVD)

(**i**) "Hello" (CEMD) (**j**) "Thanks" (CEMD) (**k**) "Yes" (CEMD) (**l**) "No" (CEMD)

Figure 2. The spectrograms obtained by three different time-frequency analysis methods.



(**a**) Original continuous spectrogram (**b**) Sharpened continuous spectrogram

Figure 3. The spectrograms of continuous signs.

Based on the Doppler radar model in the section above, the proposed CEMD-based time-frequency analysis algorithm is described in Table 1.

Table 1. CEMD-based time-frequency analysis algorithm.

Input: original radar time series $r(t)$.
Output: spectrogram obtained by CEMD.
(1) Pass the Fourier transform of the signal through an ideal bandpass filter to get the positive and negative frequency components, $R_+(e^{j\omega})$ and $R_-(e^{j\omega})$. (2) Inverse Fourier transform to $R_+(e^{j\omega})$ and $R_-(e^{j\omega})$ and take the real part. The real sequence $r_+(n)$ and $r_-(n)$ are obtained. (3) Perform standard EMD to $r_+(n)$ and $r_-(n)$. The order of the intrinsic mode functions (IMF) is N. (4) Iteration from 1 to N. Do the following operations. (a) Hilbert transform to $imf_i(t)$ and obtain the matrix $m_i(t, f)$. (b) Add the newly obtained matrix to the previous matrix. (5)End of iteration (6) The final matrix $m(t, f)$ is the spectrogram of the gesture data. The horizontal axis of the spectrogram is time and the vertical axis is frequency. (7) Follow Equations (4) and (5) to sharpen $m(t, f)$ for $g(t, f)$.

4. Yolov3-Based Sign Language Detection

As described above, there is no practical point in training and identifying the epentheses in sign language. However, they can also be studied as objects unnecessarily. To distinguish the epentheses from a sequence of signs, the technique of framing was used in the previous work [23] via splitting the sequence into different lengths. However, the fault tolerance rate is unsatisfactory because the sequence is empirically split. For different signs, the segmented sequences may not be compatible. Consequently, we propose a Yolo model-based sign detection network as the classifier. The Yolo models firstly proposed in 2016 are well-known in object detection [24]. They are featured in finding the interesting information directly and ignoring the interference from video streams, and becoming ideal candidates for real-time conditions and on-device deployment environments. This technique predicts the probability of multiple sign objects in different sign categories along with their temporal evolution over the stream of time-frequency spectrograms. Each sign object category can be classified based on its time-frequency analysis features, including the colored points and regions, as shown in Figure 2. When the features are fed to the network, it can predict which categories of object the features belong to and the occupied region with bouding box.

4.1. Network Input

The input of Yolo is objects of different sizes and orientations in the data flow, often overlapping with other objects. Unlike the previous objects, in this paper, the objects in our data flow are approximately equal in size and have the same orientation. Meanwhile, the objects are clustered near the axis where $f = 0$ and arranged distanced in the order of the time axis. Due to these characteristics, the meanings of the corresponding signs can be marked in order, making understanding the meaning of the entire sequence easy.

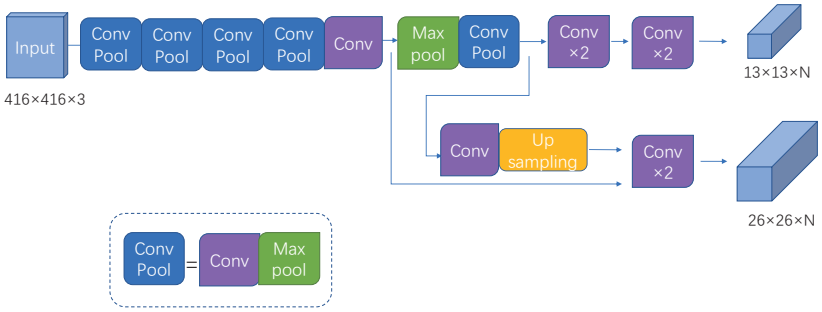
4.2. Yolov3 and Yolov3-Tiny

Yolov3 [25] predicts the bounding box of the object directly from the whole image, which turns the detection issue into logistic regression. Yolov3 is an end-to-end convolutional neural network and uses Darknet-53 as the network framework. Instead of pooling layers, convolutional layers with a stride size of 2 and a kernel of 3×3 are used to reduce the loss of feature extracted. The fully connected layers are also replaced by convolutional layers. Yolov3 divides the input image into $S \times S$ grid cells where S has three different scales. A certain grid cell predicts the object if the center coordinates of the object fall within this grid. Each grid cell predicts three bounding boxes according to three anchor boxes and each bounding box predicts four coordinate offsets, including coordinate position and

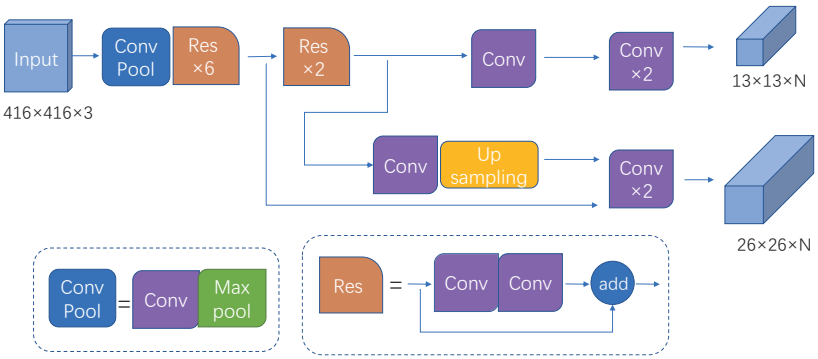
length; and width of the box, confidence score and number of categories. Thus the size of the final tensor is $S \times S \times 3 \times (5 + N)$, where N is the number of categories.

In Yolov3, the loss function is divided into three parts, including Giou loss, objectness loss and classification loss. Giou is caused by the offset of the bounding box to the ground truth. Objectness loss is the probability of whether the bounding box contains the object. Classification loss is caused by the offset of the category prediction.

Yolov3-tiny is a simplified version of Yolov3 and inherits the basic method of Yolov3. Yolov3-tiny only consists of 13 convolutional layers and divides the image into two scales, as shown in Figure 4a. Although the accuracy of Yolov3-tiny is reduced to a certain extent, the calculation speed is greatly improved with fewer network layers.



(a) Original Yolov3-tiny network



(b) Improved Yolov3-tiny network

Figure 4. A Yolov3-based dynamic sign language detection network.

4.3. Improved Yolov3-Tiny Network

In order to further increase the learning depth of the network, an improved Yolov3-tiny network is shown in Figure 4b. We replaced the framework with ResNet [26]. The residual network can solve the problem of gradient disappearance and gradient explosion when the number of network layers increases. The input size of the network is unified to $416 \times 416 \times 3$. The entire network includes eight residual blocks and five downsamplings. The kernel size is 3×3 in each residual block. Two different sizes, 26×26 and 13×13, are output through route layers after the fourth and fifth downsampling. The route layer stitches tensors of different layers together to realize the connection from the shallow network to the deep network. The final convolutional layer is $13 \times 13 \times 27$.

5. Experiments

5.1. Doppler Radar Sensor Prototype System

In order to validate the effectiveness and efficiency of the proposed method in sign language recognition, a prototype radar system was constructed as shown in Figure 5. In this prototype system,

a 24 GHz shelf radar sensor from RFbeam company, K-LC7, was selected to provide one transmitting antenna and two receiving antennas with a gain of 8.6 dBi. A 5 V DC power supplier powered the prototype system. Due to no amplifier being embedded inside the radar sensor, the output signals directly from the radar were weak. To enhance the signal-to-noise ratio, a customized operational amplifier module was designed and manufactured. The operational amplifier was LMV722 which had a unity-gain bandwidth of 10 MHz, a slew rate of 5 V/ms and dual ports. The amplifier module was designed based on the principle of voltage parallel negative feedback. As shown in Figure 5, a reference voltage was provided for the positive input. C1 and R2 formed a high-pass filter and C3 and R3 formed a low-pass filter. The cut-off frequency of the high-pass filter was 3.18 Hz and the cut-off frequency of the low-pass filter was 3.09 kHz. Meanwhile, the PCB was designed to ensure the radar system can be directly plugged into the input channel of analogue-to-digital (A/D) converter which made the sign collection more convenient. The gain of the designed amplifier module was 52 dB. The intermediate frequency (IF) I/Q output signals were amplified by the amplifier, and then transferred to the National Instruments USB6211 A/D converter to a desktop computer. Combined with the DAQ assistant in LabView, radar echo signals reflected by signs can be collected and processed.

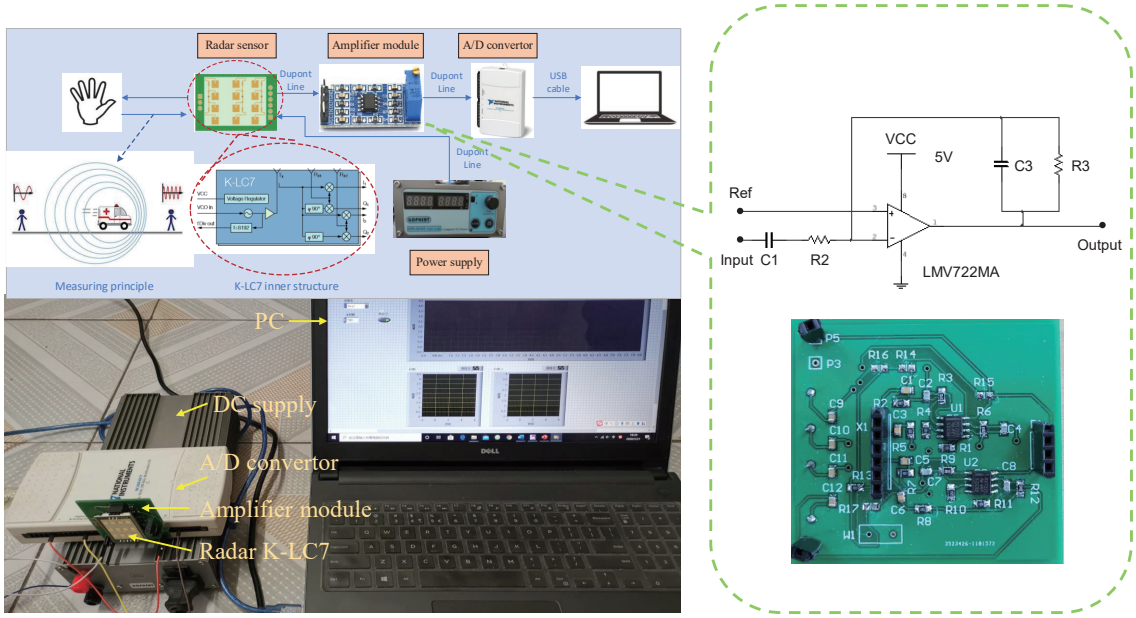


Figure 5. The entire structure of the prototype radar system.

5.2. Dataset

The sign language measured was signed by a single hand, as shown in Figure 6. The signs included "Hello" and "Thanks" in CSL; and "Yes" and "No" in ASL. The first line of Figure 6 has the initial poses of the signs and the second line shows the final poses of the signs. "Yes" and "Thanks" needed to be performed twice for meaning expression. For the convenience of the measurement process, the prototype system was placed at the edge of the table. The signer gestured straightly towards the radar at the same distance of 10 cm. Based on the low pass filter of the operational amplifier, the velocity of the signs needed to be greater than 2 cm/s. The sampling frequency was 1 kHz. Three different datasets were collected, including single signs for residual block fine-tuning, continuous signs for training and dialogues for evaluating the entire network. For the single signs, the sampling time was 1.5 s. For continuous signs, the signers gestured the four signs continuously in five different orders and paused during switching the gestures. Each set of signs was signed for 8 s and repeated 40 times. In this way, the size of the raw dataset was 800. In addition, we added noise, which was white Gaussian noise with a signal to noise ratio of 22 dB, to the original signals in order

to expand the dataset to 1600. To further investigate the dynamic sign recognition, we also collected several sets of sign dialogues combining these four gestures studied with other words unstudied. There were 150 sets of dialogues collected and each sign was repeated 50 times.

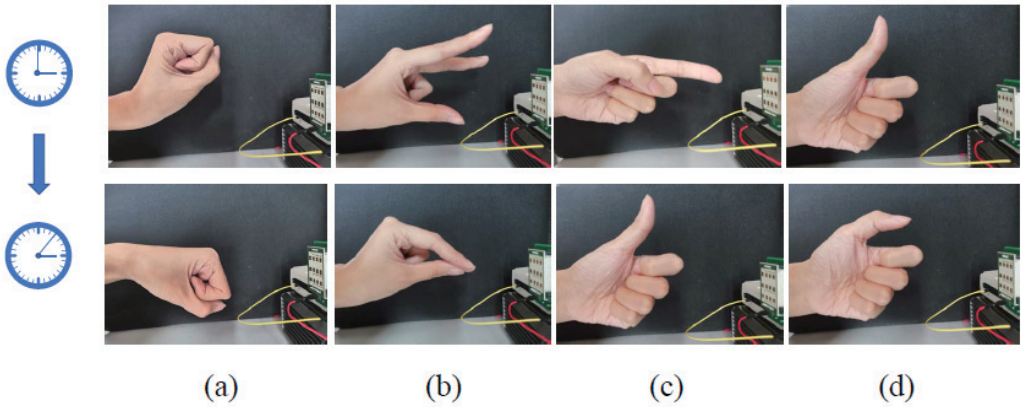


Figure 6. Four signs measured: (**a**) "Yes" in ASL, (**b**) "No" in ASL, (**c**) "Hello" in CSL, (**d**) "Thanks" in CSL.

5.3. Classification with Improved Yolov3-Tiny Network

Before the classification, we used Window LabelImg to label objection and then trained the data with the improved Yolov3-tiny network. The total continuous dataset was divided into a training set and a testing set with the ratio of 9:1. In order to reduce the training time and avoid over-fitting, we used fine-tuning in the training process [27,28]. Since the improved Yolov3-tiny network was modified based on the ResNet18 with the same downsampling layers with ResNet18, we trained the ResNet18 with the single sign data in advance. Then we fine-tuned the network by loading the pretraining weights of the eight residual blocks and trained the other layers with the continuous dataset. The training parameters are shown in Table 2. The sizes of the initial anchor boxes can be obtained by k-means clustering algorithm—respectively, (38, 33), (24, 28), (33, 33), (25, 31), (23, 30), (21, 28).

Table 2. Training parameter settings.

Parameter	Value
Epoch	300
Batch size	16
Initial learning rate	0.005
Weight decay	0.0005
IOU threshold	0.5
Confidence threshold	0.3
Optimizer	Adam

6. Results and Analysis

6.1. Continuous Sign Recognition

After training, we used the testing set to evaluate the results of the training. The training loss of the dataset of 1600 is shown in Figure 7. The network marked the predicted category and confidence score of signs and framed each sign predicted, as shown in Figure 8. The quantitative effect of the network could be evaluated by precision, recall and mean average precision (mAP). Precision is how many signs in the set are accurately predicted. Recall is how many signs predicted as corresponding categories are accurate. The mAP is the accuracy of the total validation set. The results of 800 and 1600 dataset sizes are shown in

Table 3. It can be seen from the results that the network could achieve a better result with the increasing dataset size. When the dataset size is not large enough, data augmentation by adding noise to the original signal could improve the training effect to a certain extent.

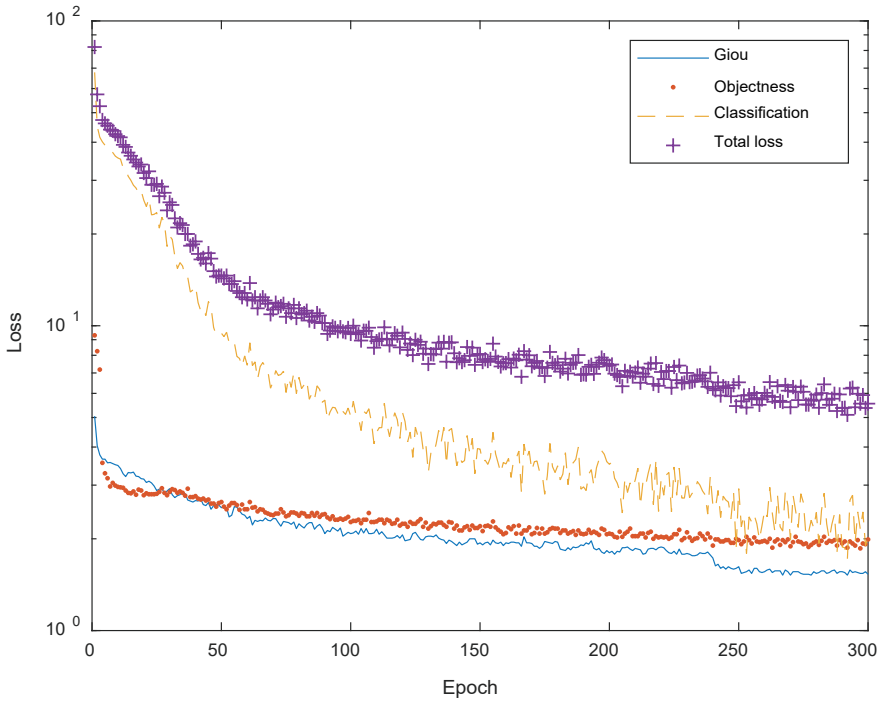


Figure 7. Training loss.

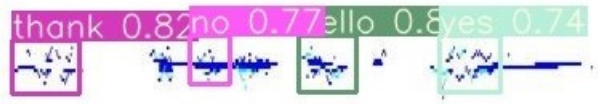


Figure 8. The predicted categories and confidence scores of continuous signs.

Table 3. The recognition results of different dataset sizes.

Dataset Size	Precision	Recall	mAP	F1-Measure
800	0.812	0.958	0.932	0.879
1600	0.924	0.993	0.99	0.957

The mAP compared to other methods is also shown in Table 4. According to the results, our method reaches an mAP up to 99%, which is higher than most of the previous works.

Table 4. Mean average precision (mAP) comparison with other methods having different data acquisitions and classifications.

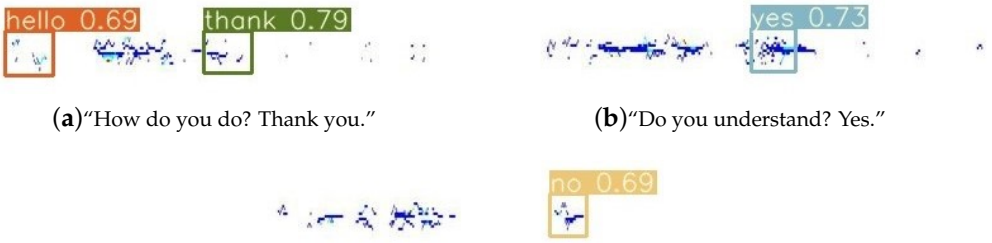
Data Acquisition	Classification	mAP
Data glove [29]	Orientation and motion	94%
Video [30]	SVM	96%
Video [31]	CNN	92.88%
Video [32]	Inception+RCNN	93%
Image [33]	Faster-RCNN+3D CNN+LSTM	99%
Radar(our method)	Improved yolov3-tiny	**99%**

6.2. Dialogue Recognition

Word error rate (WER) was used to evaluate the recognition effect of the proposed network. In this paper, WER is defined as

$$WER = \frac{insertions + substitutions}{total\ words} \tag{6}$$

We used the dialogue dataset to evaluate the trained Yolov3-tiny network. For the dialogue recognition, the WERs of our work and other works are shown in Table 5. The predicted categories and confidence scores are shown in Figure 9. In the actual application scenario, the WER of single sentence dialogue was below 0.28, which is 10% lower than that of other methods, even though many untrained signs were involved.



(**a**)"How do you do? Thank you." (**b**)"Do you understand? Yes."

(**c**)"Do you understand? No."

Figure 9. The predicted categories and confidence scores of dialogues.

Table 5. The recognition results of three dialogues and word error rate (WER) comparison with other works.

Dialogue	Insertion	Substitution	WER
How do you do? Thank you.	0.1	0.18	0.28
Do you understand? Yes.	0.06	0.16	0.22
Do you understand? No.	0.1	0.06	0.16
Total	0.09	0.145	**0.235**
Camgoz, et al. [34]	-	-	0.407
Zhang, et al. [19]	-	-	0.383
Pu, et al. [35]	-	-	0.367
Wei, et al. [36]	-	-	0.268

6.3. Discussion on Reuse of 24 Ghz Automotive Radar in Sign Language Recognition

As described in Section 2, 24 GHz radar sensors can also be utilized as automotive radar for blind zone detection, lane change assistance and parking aid. With strong ability in target detection, the automotive radar is also theoretically capable of extending its functional boundary from traditional vehicle navigation to sign language recognition. However, there are still differences between automobile navigation and sign language recognition in the actual application scenario. Automotive radar has a longer detection distance and has certain requirements for the beam patterns and polarization. Based on these characteristics, the microstrip antenna array has been proposed within the automobile radar design—the radar sensor is always very large in size, with high integration of circuits and components and a hefty price. However, the very different requirements for sign language recognition compared with automobile radars are shorter distance, wider field of view and lower polarization dependence. Those differences can be explained by the fact that the scenario for sign language recognition mostly involves consumer electronics such as laptops, tablets, mobile phones and other wearable electronics. Therefore, further optimizations such as modification of antenna gain, amplifier gain and array arrangement are required for the direct use of automotive radar for sign language recognition.

7. Conclusions

In this paper, we investigated the feasibility of continuous sign language recognition based on a Doppler radar sensor. The continuous signs consist of four signs in CSL and ASL, including "Hello"; "Thanks"; "Yes"; and "No." The signs are captured by a 24 GHz radar system and the micro-Doppler signatures are extracted by CEMD in time-frequency analysis. To distinguish the interesting signs from the epentheses, a fine-tuned improved Yolov3-tiny network is utilized based on pretrained Resnet18. A prototype system was constructed for the evaluation of the proposed method at both the continuous sign level and the sentential sign level. The experiment results show that the precision, recall, mAP and F1-measure of the continuous sign recognition are 0.924, 0.993, 0.99 and 0.957. The average WER of the dialogue recognition is up to 0.235. Compared with sign language recognition based on video and inertial sensors, sign language recognition represented by EM waves reaches a higher accuracy and lower WER. In conclusion, the proposed method provides a new potential for non-contact sign language recognition.

Author Contributions: S.L. and K.Z. conceived the concept; L.Y. designed the experiment; L.Y. built the experimental system with G.Z. and performed the whole experiment, including data collection, sign feature extraction and network training and testing; K.Z. evaluated the training process. L.Y and S.L. contributed to writing this paper. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the China National Natural Science Foundation Project under grant 61601141 and the Fundamental Research Funds for the Central Universities (grant number HIT.NSRIF.201817).

Acknowledgments: The authors acknowledge the participants in the experiments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhang, X.; Wu, Q.; Zhao, D. Dynamic Hand Gesture Recognition Using FMCW Radar Sensor for Driving Assistance. In Proceedings of the 2018 10th International Conference on Wireless Communications and Signal Processing (WCSP), Hangzhou, China, 18–20 October 2018; pp. 1–6.
2. Yang, S.; Premaratne, P.; Vial, P. Hand gesture recognition: An overview. In Proceedings of the 2013 5th IEEE International Conference on Broadband Network Multimedia Technology, Guilin, China, 17–19 November 2013; pp. 63–69.
3. Abraham, E.; Nayak, A.; Iqbal, A. Real-Time Translation of Indian Sign Language using LSTM. In Proceedings of the 2019 Global Conference for Advancement in Technology (GCAT), Bangaluru, India, 18–20 October 2019; pp. 1–5.
4. Feng, W.; Xia, C.; Zhang, Y.; Yu, J.; Jiang, W. Research on Chinese Sign Language Recognition Methods Based on Mechanomyogram Signals Analysis. In Proceedings of the 2019 IEEE 4th International Conference on Signal and Image Processing (ICSIP), Wuxi, China, 19–21 July 2019; pp. 46–50.
5. Krishnan, P.T.; Balasubramanian, P. Detection of Alphabets for Machine Translation of Sign Language Using Deep Neural Net. In Proceedings of the 2019 International Conference on Data Science and Communication (IconDSC), Bangalore, India, 1–2 March 2019; pp. 1–3.
6. Ku, Y.; Chen, M.; King, C. A Virtual Sign Language Translator on Smartphones. In Proceedings of the 2019 Seventh International Symposium on Computing and Networking Workshops (CANDARW), Nagasaki, Japan, 26–29 November 2019; pp. 445–449.
7. Liao, Y.; Xiong, P.; Min, W.; Min, W.; Lu, J. Dynamic Sign Language Recognition Based on Video Sequence With BLSTM-3D Residual Networks. *IEEE Access* **2019**, *7*, 38044–38054. [CrossRef]
8. Lien, J.; Gillian, N.; Karagozler, M.E.; Amihood, P.; Schwesig, C.; Olson, E.; Raja, H.; Poupyrev, I. Soli: Ubiquitous gesture sensing with millimeter wave radar. *ACM Trans. Graph. (TOG)* **2016**, *35*, 1–19. [CrossRef]
9. Zhang, S.; Li, G.; Ritchie, M.; Fioranelli, F.; Griffiths, H. Dynamic hand gesture classification based on radar micro-Doppler signatures. In Proceedings of the 2016 CIE International Conference on Radar (RADAR), Guangzhou, China, 10–13 October 2016; pp. 1–4.

10. Kim, Y.; Toomajian, B. Application of Doppler radar for the recognition of hand gestures using optimized deep convolutional neural networks. In Proceedings of the 2017 11th European Conference on Antennas and Propagation (EUCAP), Paris, France, 19–24 March 2017; pp. 1258–1260.
11. Kulhandjian, H.; Sharma, P.; Kulhandjian, M.; D'Amours, C. Sign Language Gesture Recognition Using Doppler Radar and Deep Learning. In Proceedings of the 2019 IEEE Globecom Workshops (GC Wkshps), Waikoloa, HI, USA, 9–13 December 2019; pp. 1–6.
12. Hou, J.; Aoki, Y. A real-time interactive non-verbal communication system through semantic feature extraction. In Proceedings of the IEEE International Conference on Multimedia and Expo, Lausanne, Switzerland, 26–29 August 2002; Volume 2, pp. 425–428.
13. Fioranelli, F.; Le Kernec, J.; Shah, S.A. Radar for Health Care: Recognizing Human Activities and Monitoring Vital Signs. *IEEE Potentials* **2019**, *38* 16–23. [CrossRef]
14. Choi, J.; Va, V.; Gonzalez-Prelcic, N.; Daniels, R.; Bhat, C.R.; Heath, R.W. Millimeter-Wave Vehicular Communication to Support Massive Automotive Sensing. *IEEE Commun. Mag.* **2016**, *54* 160–167. [CrossRef]
15. Lan, S.; Yang, C.; Liu, B.; Qiu, J.; Denisov, A. Indoor real-time multiple moving targets detection and tracking using UWB antenna arrays. In Proceedings of the 2015 International Symposium on Antennas and Propagation (ISAP), Hobart, Australia, 9–12 November 2015; pp. 1–4.
16. Shah, S.A.; Fioranelli, F. RF Sensing Technologies for Assisted Daily Living in Healthcare: A Comprehensive Review. *IEEE Aerosp. Electron. Syst. Mag.* **2019**, *34*, 26–44. [CrossRef]
17. Liu, J.; Ding, H.; Shahroudy, A.; Duan, L.Y.; Jiang, X.; Wang, G.; Kot, A.C. Feature Boosting Network For 3D Pose Estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 494–501. [CrossRef] [PubMed]
18. Yang, S.; Liu, J.; Lu, S.; Er, M.H.; Kot, A.C. Collaborative learning of gesture recognition and 3D hand pose estimation with multi-order feature analysis. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020.
19. Zhang, Z.; Pu, J.; Zhuang, L.; Zhou, W.; Li, H. Continuous Sign Language Recognition via Reinforcement Learning. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 285–289.
20. Gurbuz, S.Z.; Gurbuz, A.C.; Malaia, E.A.; Griffin, D.J.; Crawford, C.; Rahman, M.M.; Aksu, R.; Kurtoglu, E.; Mdrafi, R.; Anbuselvam, A.; et al. A Linguistic Perspective on Radar Micro-Doppler Analysis of American Sign Language. In Proceedings of the 2020 IEEE International Radar Conference (RADAR), Washington, DC, USA, 28–30 April 2020; pp. 232–237.
21. Tanaka, T.; Mandic, D.P. Complex Empirical Mode Decomposition. *IEEE Signal Process* **2007**, *14*, 101–104. [CrossRef]
22. Huang, N.; Shen, Z.; Long, S.; Wu, M.; Shih, H.; Zheng, Q.; Yen, N.; Tung, C.; Liu, H. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc. A-Math. Phys. Eng. Sci.* **1998**, *454*, 903–995. [CrossRef]
23. Suh, J.S.; Ryu, S.; Han, B.; Choi, J.; Kim, J.; Hong, S. 24 GHz FMCW Radar System for Real-Time Hand Gesture Recognition Using LSTM. In Proceedings of the 2018 Asia-Pacific Microwave Conference (APMC), Kyoto, Japan, 6–9 November 2018; pp. 860–862.
24. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
25. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
26. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
27. Izidio, D.M.F.; Ferreira, A.P.A.; Barros, E.N.S. An Embedded Automatic License Plate Recognition System Using Deep Learning. In Proceedings of the 2018 VIII Brazilian Symposium on Computing Systems Engineering (SBESC), Salvador, Brazil, 6–9 November 2018; pp. 38–45.
28. Carolis, B.D.; Ladogana, F.; Macchiarulo, N. YOLO TrashNet: Garbage Detection in Video Streams. In Proceedings of the 2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS), Bari, Italy, 27–29 May 2020; pp. 1–7.

29. Kau, L.; Su, W.; Yu, P.; Wei, S. A real-time portable sign language translation system. In Proceedings of the 2015 IEEE 58th International Midwest Symposium on Circuits and Systems (MWSCAS), Fort Collins, CO, USA, 2–5 August 2015; pp. 1–4.
30. Kamrul Hasan, S.M.; Ahmad, M. A new approach of sign language recognition system for bilingual users. In Proceedings of the 2015 International Conference on Electrical & Electronic Engineering (ICEEE), Rajshahi, Bangladesh, 4–6 November 2015; pp. 33–36.
31. Rao, G.A.; Syamala, K.; Kishore, P.V.V.; Sastry, A.S.C.S. Deep convolutional neural networks for sign language recognition. In Proceedings of the 2018 Conference on Signal Processing Additionally, Communication Engineering Systems (SPACES), Vijayawada, India, 4–5 January 2018; pp. 194–197.
32. Bantupalli, K.; Xie, Y. American Sign Language Recognition using Deep Learning and Computer Vision. In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 10–13 December 2018; pp. 4896–4899.
33. He, S. Research of a Sign Language Translation System Based on Deep Learning. In Proceedings of the 2019 International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM), Dublin, Ireland, 16–18 October 2019; pp. 392–396.
34. Camgoz, N.C.; Hadfield, S.; Koller, O.; Bowden, R. SubUNets: End-to-End Hand Shape and Continuous Sign Language Recognition. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 3075–3084.
35. Pu, J.; Zhou, W.; Li, H. Iterative Alignment Network for Continuous Sign Language Recognition. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 4160–4169.
36. Wei, C.; Zhao, J.; Zhou, W.; Li, H. Semantic Boundary Detection with Reinforcement Learning for Continuous Sign Language Recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2020**. [CrossRef]