

Article

Sign Language Recognition Using Two-Stream Convolutional Neural Networks with Wi-Fi Signals

Chien-Cheng Lee ^{1,*}  and Zhongjian Gao ^{1,2}¹ Department of Electrical Engineering, Yuan Ze University, Taoyuan 320, Taiwan; gzj@fjssmu.edu.cn² School of Mechanical and Electrical Engineering, Sanming University, Sanming 365004, China

* Correspondence: cclee@saturn.yzu.edu.tw; Tel.: +886-3-4638800 (ext. 7323)

Received: 30 October 2020; Accepted: 15 December 2020; Published: 16 December 2020



Abstract: Sign language is an important way for deaf people to understand and communicate with others. Many researchers use Wi-Fi signals to recognize hand and finger gestures in a non-invasive manner. However, Wi-Fi signals usually contain signal interference, background noise, and mixed multipath noise. In this study, Wi-Fi Channel State Information (CSI) is preprocessed by singular value decomposition (SVD) to obtain the essential signals. Sign language includes the positional relationship of gestures in space and the changes of actions over time. We propose a novel dual-output two-stream convolutional neural network. It not only combines the spatial-stream network and the motion-stream network, but also effectively alleviates the backpropagation problem of the two-stream convolutional neural network (CNN) and improves its recognition accuracy. After the two stream networks are fused, an attention mechanism is applied to select the important features learned by the two-stream networks. Our method has been validated by the public dataset SignFi and adopted five-fold cross-validation. Experimental results show that SVD preprocessing can improve the performance of our dual-output two-stream network. For home, lab, and lab + home environment, the average recognition accuracy rates are 99.13%, 96.79%, and 97.08%, respectively. Compared with other methods, our method has good performance and better generalization capability.

Keywords: sign language recognition (SLR); two-stream CNN; SignFi; CSI; attention mechanism

1. Introduction

Sign language is an important way for deaf people to understand and communicate with each other. Communication barriers are often encountered between the deaf communities and people who do not know about sign language. Many researchers try to build a sign language recognition system to break these barriers [1]. Currently, sign language recognition systems are roughly divided into two categories: (i) device-based sign language recognition systems; (ii) device-free sign language recognition systems [2,3].

Wearable sensors are widely used in device-based sign language recognition systems. In 1983, Grimesws et al. invented a data glove for dynamic gesture recognition [4]. Shukor et al. used data gloves to obtain data on Malaysian sign language letters, numbers, and words [5]. Kanokod et al. recognized gestures through the time delay neural networks (TDNNs) algorithm, and the gesture data is obtained from data gloves based on pyrolytic graphite sheets (PGS) [6]. In general, the advantages of the wearable device-based sign language recognition method are that the input data is accurate and the recognition rate is high [2,4–6]. The disadvantages are also obvious. For example, wearable devices are often expensive and inconvenient to carry.

The device-free sign language recognition systems are usually inexpensive and not limited by the wearable device [2,7]. Several device-free sign language recognition systems use computer vision techniques with cameras. Koller conducted a survey on gesture recognition, focusing on the

RWTH-PHOENIX-Weather data set recorded by a stationary standard color camera [8]. For example, Cui et al. adopted a convolutional neural network (CNN) with stacked temporal fusion layers and a bi-directional recurrent neural network (RNN) to extract the spatiotemporal information of sign language videos [9]. Pu et al. proposed an alignment network with iterative optimization for video-based sign language recognition, including a three-dimensional (3D) ResNet and an encoder-decoder network [10]. In addition, depth information is also considered to improve the performance. Ohn-Bar et al. introduced a vision-based system consisting of RGB and depth descriptors to classify gestures. This method used RGB and depth images to modify the Histogram of Oriented Gradients (HOG) function to achieve higher classification accuracy [11]. Huang et al. proposed a 3D CNN which extracts automatically spatial-temporal features from raw video datasets collected with Microsoft Kinect sensor [12]. Aly et al. utilized an unsupervised principal component analysis network (PCANet) to extract the local features of sign language images captured from the Microsoft Kinect sensor. The extracted features were recognized by the support vector machine (SVM) [13]. It is worth noting that computer vision-based sign language recognition systems are not only susceptible to light conditions and obstacles, but also cause privacy issues [2,7].

With the widespread deployment of wireless networks, wireless related technologies have experienced very rapid growth. Gesture recognition based on commodity Wi-Fi sensing solutions has been widely studied [2,7]. Thus, a sign language recognition system that is non-intrusive and insensitive to lighting conditions has attracted the interest of many researchers. The gestures recognition system proposed by Melgarejo et al. used directional antenna technology to achieve fine-grained gesture recognition [14]. Shang and Wu used Wi-Fi signals to recognize different gestures and arm movements, called WiSign [15]. Wi-Finger, which can recognize nine-digit finger gestures from American Sign Language (ASL), was implemented on a commercial Wi-Fi infrastructure [16].

Most Wi-Fi sensing research uses Wi-Fi channel state information (CSI), which describes how the signal propagates from the transmitter to the receiver, and reveals combined effects such as scattering, fading, and power decay with distance. Zhou et al. collected the number of digital gesture information through CSI-based technology, and the recognition accuracy of these gestures can reach 96% through deep learning [17]. Ma et al. collected CSI traces for 276 sign gestures that are frequently used in daily life. The CSI dataset is called the SignFi dataset. The proposed method is called the SignFi method, which is a nine-layer CNN model for recognizing sign language gestures from laboratory, home, and mixed lab and home environments [1].

However, the challenges of sign language recognition based on Wi-Fi sensing come from three aspects. First, the CSI signal is mixed with background noise, signal interference, and multipath noise. Second, the diversity, complexity, and similarity of gestures make it difficult to distinguish. Third, sign language includes the positional relationship of gestures in space and the changes of actions over time. Ahmed et al. proposed a higher order statistics-based recognition (HOS-Re) model to extract higher order statistical features from SignFi dataset and select a robust feature subset as input to a multilevel support vector machine (SVM) classifier [3].

Our goal is to recognize gestures based on the SignFi dataset. We propose a novel gesture recognition method that combines singular value decomposition (SVD), dual-output two-stream network, and attention mechanism. SVD is used in the data preprocessing to reduce the noise to a certain extent. The proposed dual-output two-stream network, which combines a spatial-stream network and a motion-stream network, extracts spatial and temporal information, and classifies input patterns. The attention mechanism in deep learning is inspired by the mode of human attention thinking and has been widely used in natural language processing. In our work, the attention mechanism is utilized to select the important features from our dual-output two-stream network. In this study, we find that the additional auxiliary output can effectively alleviate the backpropagation problem of the two-stream network and improve its recognition accuracy.

In summary, the contributions of work are as follows:

1. This work shows how to process sign language data based on CSI traces through SVD. It not only makes sign language features more prominent, but also reduces noise and outliers to a certain extent. SVD helps to improve the recognition accuracy of the two-stream network, and has the characteristics of fast running, robustness, and generalization ability.
2. We explored a novel scheme, dual-output two-stream network. The two-stream network consists of a spatial-stream network and a motion-stream network. The input of the spatial stream network is a three-dimensional array (similar to an array of RGB images) composed of the amplitude and phase of each gesture. The array differences, which represent the amplitude and phase changes, are fed into the motion stream network. The convolutional features from the two streams are fused, and then an attention mechanism automatically selects the most descriptive features. The experimental results show that the dual output can effectively alleviate the back propagation problem of two-stream CNN and improve the accuracy.
3. The fine-tuning of an ImageNet pre-trained CNN model on CSI datasets has not yet been exploited. We explored CNN architectures with different model layers on CSI data.

2. Materials and Methods

2.1. Received Signal Strength Indicator and Channel State Information

The principle of wireless indoor behavior detection is to transmit the generated wireless signal through multipath transmission. Reflection and scattering will cause multiple superimposed signals to be received in an indoor environment. These signals are physically affected by human behavior in the transmission space and generate various environmental characteristic information. Therefore, the information extracted from multipath superimposed signals can be used to identify human behavior [18].

The most common data sources for the device free gesture recognition systems based on Wi-Fi signals are the Received Signal Strength Indicator (RSSI) and CSI [19]. RSSI is the most widely used signal indicator for wireless devices [20]. It describes the attenuation experienced during the propagation of a wireless signal. In the wireless sensor link, the RSSI of the wireless sensor unit will be changed with the movement of the person. In other words, the movement of the person can be detected based on the change of RSSI. Since the RSSI information is easy to capture, RSSI was used for hand gesture recognition in the early days [21]. The RSSI is a kind of coarse-grained information, mainly from the superimposition result of the receiver during signal transmission. It is affected by the multipath effect and environmental noise, so it has large fluctuations and poor stability [2].

RSSI only reflects the total amplitude of multipath overlap on the media access control (MAC) layer, while CSI is more fine-grained subcarrier information of the physical layer. For a multiple-input multiple-output (MIMO) wireless technology in combination with orthogonal frequency division multiplexing (OFDM) Wi-Fi system, CSI is mainly derived from the sub-carriers decoded by the OFDM [22]. It can effectively eliminate or reduce the interference caused by the multi-path effect. CSI contains amplitude and phase information under different sub-carriers, and each sub-carrier does not interfere with each other. Thus, CSI is more sensitive and reliable than RSSI. It has higher detection accuracy and sensitivity, so it can achieve more detailed motion detection [23].

A set of CSI data can be obtained from each received data packet of a wireless network card compatible with the IEEE 802.11n protocol standard. The amplitude and phase of a sub-carrier on the CSI data are shown in Equation (1):

$$H(f_k) = \|H(f_k)\|e^{j\angle H(f_k)} \quad (1)$$

where $H(f_k)$ is the CSI of the sub-carrier with a center frequency of f_k , $\|H(f_k)\|$ and $\angle H(f_k)$ are the amplitude and phase of the center frequency of f_k , respectively. They are the most important information in CSI data, and k represents the total number of sub-carriers.

2.2. Singular Value Decomposition

In linear algebra, singular value decomposition (SVD) is the factorization of real or complex matrices [24]. SVD is a decomposition method that can be applied to any matrix. There is always SVD for any matrix A , as shown in Equation (2):

$$A = U\Sigma V^T \quad (2)$$

Assuming that A is an $m \times n$ real or complex matrix, the obtained U is an $m \times m$ square matrix, and the orthogonal vector in U is called a left singular vector. Σ is an $m \times n$ rectangular diagonal matrix. Except for the diagonal elements, all elements of Σ are 0. The elements on the diagonal are called singular values. V^T is the transposed matrix of V , which is an $n \times n$ square matrix. The orthogonal vector is called the right singular value vector.

Generally speaking, the values on the Σ are in descending order [24]. The larger the value, the higher the importance of the dimension. We can choose the top singular values to approximate the matrix. This way not only extracts important features from the data, but also simplifies the data and eliminates noise and redundancy. The number of singular values depends on various factors, such as different datasets, recognition methods, and temporal and spatial characteristics.

2.3. SignFi Dataset

The SignFi dataset contains the CSI data, which were extracted by the 802.11n CSI-Tool on the Intel WiFi Link 5300 device with three antennas. The dataset was collected through a transmitter with three external antennas and a receiver with one internal antenna. Figure 1 shows the measurement scenes of the lab and home environments. The 802.11n CSI-Tool provides CSI values of 30 sub-carriers, which were sampled approximately every 5 milliseconds. The duration of each gesture is about 1 s, so there were 200 CSI samples for each gesture. The CSI data was stored as a 3D matrix of complex values representing amplitude and phase information. The size of the 3D matrix is $200 \times 30 \times 3$. The 3D amplitude and phase matrices are similar to digital images with spatial resolution of $H \times W$ and C color channels. Thus, the CSI data can be regarded as images. The three color channels correspond to the three antenna signals.

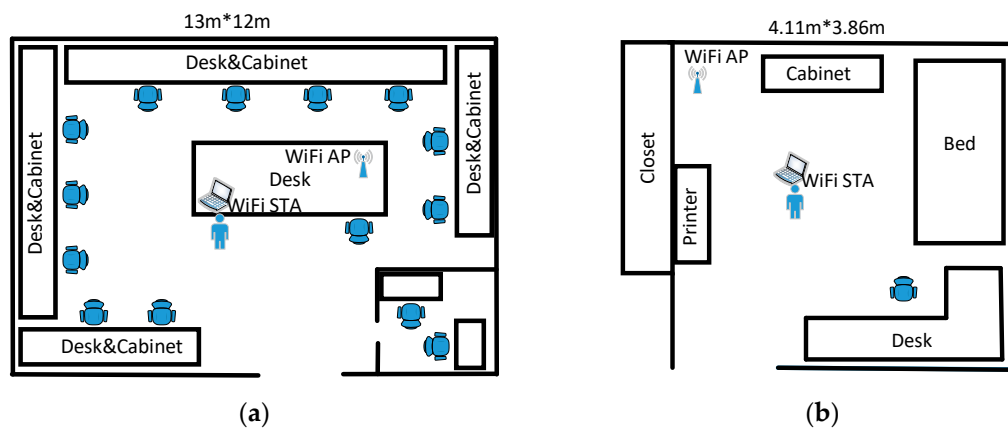


Figure 1. Measurement scenes of the lab and home environments. (a) Lab environment. (b) Home environment.

The SignFi dataset consisted of two parts. The first part included 276 gestures, a total of 8280 instances from the same user. Among them, 5520 instances and 2760 instances were obtained in the laboratory and home. Each gesture had 20 and 10 instances in the laboratory and home, respectively. The second part included 150 gestures with 7500 instances collected from five users in the laboratory, 50 instances of each gesture and 10 instances of each user. The dataset was further divided into four groups to train and

evaluate our method including Home276, Lab276, Lab+Home276, and Lab150 groups. The number of gestures were 276 and 150. Table 1 shows the statistics of the SignFi dataset.

Table 1. Statistics of the SignFi dataset.

Number of Users	Data Groups	Number of Gesture Categories	Number of Gesture Instances	Number of Instances of Each Gesture per User
1	Home276	276	2760	10
1	Lab276	276	5520	20
1	Lab + Home276	276	8280	20 + 10
5	Lab150	150	7500	10

2.4. Data Preprocessing

The amplitude and phase can be obtained from the 3D matrix of the raw CSI. Their size is $200 \times 30 \times 3$. The amplitude and phase of an antenna can be obtained from Equations (3) and (4):

$$\|H(f_k)\| = \sqrt{(Re)^2 + (Im)^2} \quad (3)$$

$$\angle H(f_k) = \text{actan}(Im/Re) \quad (4)$$

Note that we directly get the angular degree value of the phase without unwrapping the phase to eliminate the phase shift like SignFi method [1]. We combined the amplitude and phase of each gesture and reshaped it into a combination matrix with a size of $200 \times 60 \times 3$ as the input data of the spatial-stream network. The input of the motion-stream network was the difference (size of $199 \times 60 \times 3$) of the above combination matrix, which was a concatenation of amplitude difference and phase difference. The difference comes from two consecutive instances and describes the changes in amplitude and phase. It indicates the change of the gesture corresponding to the salient area of movement. Then, two types of modality data, namely combination matrix and their difference matrix, were preprocessed by SVD to remove redundant and irrelevant noise.

Figure 2 shows the combination matrix and difference matrix before and after SVD preprocessing of sign language “GO” in the home and laboratory environments. Each picture in Figure 2 represents a 3D matrix. Figure 2a,b,e,f are the combination matrices with a size of $200 \times 60 \times 3$, and Figure 2c,d,g,h are the difference matrices of a size of $199 \times 60 \times 3$. The Y axis represents the first dimension, and the X axis represents the second dimension. The RGB color is the third dimension representing three antenna signals. On the X axis, the first half (0–29) is the amplitude information and the second half (30–59) is the phase information.

From Figure 2, we observed: (1) Combination matrices are more colorful than difference matrices. The color channels correspond to the three antenna signals. The richer the color, the greater the diversity of the signal, and the more information it contains. Thus, the combination matrices contain more information than the difference matrices. (2) The same user performs the same gesture, and the difference between the home and laboratory environment results in different CSI data, especially in the combination matrices, as shown in Figure 2a,b,e,f. In other words, the amplitude and phase of CSI are easily affected by the environment. However, the difference matrices are less affected by the environment. (3) We perform SVD preprocessing on the amplitude and phase of CSI data, respectively. In order to strike a balance between data feature integrity and noise elimination, SVD only selects the top 20 of the 30 singular value rankings, namely SVD_20. From Figure 2c,d,g,h, we can know that the matrix signal becomes smoother after performing SVD.

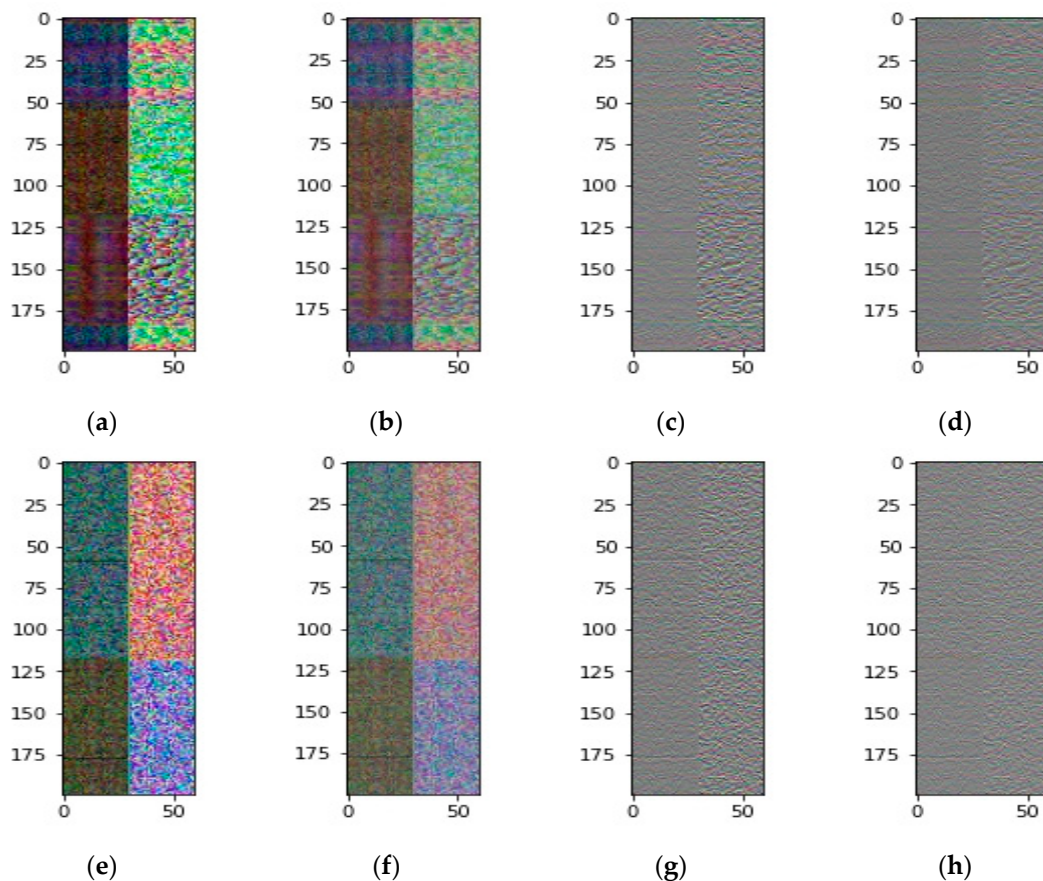


Figure 2. Channel State Information (CSI) matrices before and after singular value decomposition (SVD) preprocessing of sign language “GO” in home and laboratory environments. (a,b) are the combination matrices before and after SVD preprocessing in the home. (c,d) are the difference matrices before and after SVD preprocessing in the home. (e,f) are the combination matrices before and after SVD preprocessing in the lab. (g,h) are the difference matrices before and after SVD preprocessing in the lab.

2.5. Dual-Output Two-Stream Convolutional Neural Network

Convolutional Neural Network (CNN) is the most successful neural network in the field of deep learning [25]. The network avoids the complicated processing of images and can directly input the original images to achieve end-to-end results. CNN is derived from Hubel and Wiesel’s study of the cat brain visual system in 1962 [26]. In 1998, Yann Lecun proposed the LeNet-5 network to solve the visual task of handwritten digit recognition [27]. In the 2012 ImageNet image recognition competition, Hinton used AlexNet to greatly improve the accuracy of image recognition and subvert the field of image recognition [28]. This made CNN attract much attention and has become a research hotspot. In order to improve the performance of CNN, several improved CNNs have been proposed, such as ZFNet [29], VGGNet [30], GoogleNet [31], ResNet [32], DenseNet [33], and ResNeXt [34]. These networks focus on three important aspects: depth, width, and cardinality. At the same time, the CNN network structure has been developed in terms of attention mechanism, efficiency, and automation. The most famous are SENet [35], CBAM [36], SqueezeNet, MobileNet [37], NASNet [38], and EfficientNet [39].

In video behavior recognition, Simonyan et al. proposed a two-stream CNN structure for RGB input and optical flow input [40]. They used two identical CNN structures for training and merged through a post-fusion method. This is an effective way in the field of behavior recognition when the training dataset is limited. A lot of research has been conducted based on this architecture [41–43]. For example, Wang et al. proposed a time segmentation network (TSN), which divides the input video into several segments and sparsely samples two-stream features from these segments [41]. Feichtenhofer et al. extended the two-stream CNN and proposed a spatio-temporal CNN [42].

Figure 3 shows the architecture of our proposed sign language recognition method, which combined SVD, a dual-output two-stream network, and an attention mechanism. The SignFi dataset was collected by CSI measurement. The raw CSI data are a 3D matrix that can be regarded as an image. Thus, computer vision techniques and CNN models can be used to process the CSI data.

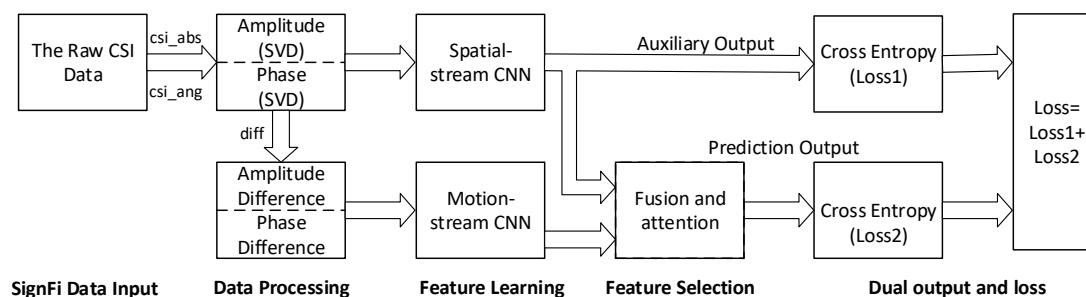


Figure 3. Architecture of the proposed sign language recognition method.

The amplitude and phase information contain noise and a certain phase offset. In our method, SVD was first used to remove redundant and irrelevant noise in amplitude and phase. Then, they were concatenated and converted to a 3D matrix, which is similar to an array of RGB images. After the SVD processing, the resulting matrix was fed into the spatial-stream CNN, which is the top layer of our dual-output two-stream network. Sign language includes not only the positional relationship of gestures in space, but also the changes of actions over time. We introduced the amplitude difference and phase difference information, which represented changes in amplitude and phase respectively. The difference matrix was input to the motion-stream CNN, which is the bottom layer of our dual-output two-stream network.

The proposed dual-output two-stream network is shown in Figure 4. In total, two types of modality data, combination matrix and difference matrix, were input into the network. In this study, the ResNet model was used for two stream CNNs. The convolutional layers in CNNs extracted multiple levels of features. When two streams are fused by concatenation, the attention mechanism (CBAM) [36] module will automatically select the most descriptive features learned by the two stream networks. Then, batch normalization (BN) is used to prevent overfitting. The ensemble prediction was the final output, as shown in the bottom layer of Figure 4. The two cross-entropy losses were combined to optimize the learning process. The dual-output and two cross-entropy losses in this structure mainly borrowed the ideas of GoogleNet architecture. The additional classification network mainly provided gradient training for the previous convolution. When the network deepens, the gradient cannot be effectively transmitted from back to front, and network parameters cannot be updated. Such a branch can alleviate the gradient propagation problem.

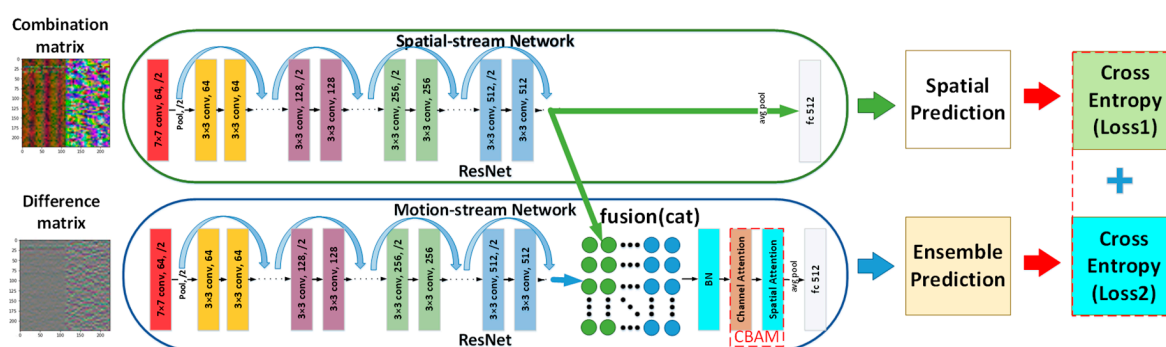


Figure 4. The framework of our dual-output two-stream network.

Most CNNs provide the pre-trained models based on the ImageNet dataset. For the pre-trained models, the CSI data is unknown. The transfer learning allowed us to use a small amount of newly

labeled data to build a high-quality classification model for the new data. Therefore, we used the transfer learning to fine-tune the pre-trained CNN models to speed up the training and improve the accuracy. In transfer learning, we freeze the first five layers of the pre-trained model and train the remaining layers. In this way, we retain the generic features learned from the ImageNet data set, and also learn domain knowledge from the CSI data.

3. Experimental Results

3.1. Network Training and Test Settings

We conducted experiments on sign language recognition tasks and performed all experiments on a PC with Intel(R) Xeon(R) CPU E5-2630 v3 @ 2.40 GHz CPU and GeForce GTX TitanX GPU with 12 GB of memory. We used momentum stochastic gradient descent (SGDM) with a momentum of 0.9 and an initial learning rate of 0.0001 to train the network to update the weights and biases. The activation function is a rectified linear unit (ReLU). The batch size is set to 16, and the training period is 500. The width and height of the input matrix data are resized to 224×224 . The network weights are initialized using ImageNet's pre-trained model. Our experiment followed the paper of SignFi method training and evaluation scheme, using non-repetitive five-fold cross-validation. The training sample is 80% and the test sample is 20%, which is consistent with the references [1,3]. In order to preserve the percentage of test samples for each category, stratified K -fold was applied. The dual-output two-stream network contains a lot of batch normalization, so it is necessary to shuffle the training data after each training cycle.

3.2. SignFi Dataset Evaluation

We quantify the performance of dual-output two-stream network on the benchmark SignFi dataset. The first evaluation is to show the effect of the attention mechanism. We use ResNet50 in our model with and without attention mechanism, respectively. The combination modality of SVD preprocessing on all data groups is used as input. The evaluation results are shown in Table 2. We can observe that the attention mechanism can indeed improve accuracy.

Table 2. Evaluation of attention mechanism.

Data Groups	Accuracy without Attention	Accuracy with Attention
Home276	99.06%	99.13%
Lab276	96.52%	96.79%
Lab + Home276	97.04%	97.08%
Lab150	95.81%	95.88%

The next evaluation is to test the Home276 data group. We use ResNet18 and ResNet50 models in our two stream CNNs to understand which is more suitable for CSI data in sign language recognition tasks. The results in Table 3 show that, compared with other methods and single stream CNNs, the dual-output dual-stream network with ResNet50 and SVD obtains competitive results. In Table 3 we can also know that SVD preprocessing and difference matrices also improves the accuracy. The evaluation of Lab276 data group is shown in Table 4. SVD still improves the accuracy of our dual-output two-stream network. However, the best recognition accuracy (96.79%) of our method is lower than SignFi method (98.91%) and HOS-Re method (98.26%). The laboratory environment is likely to be more complicated than the home environment. The SignFi method uses the unwrapping transform to preprocess the CSI phase data, while the HOS-Re method extracts the third-order cumulant feature that can reduce signal noise.

Table 3. Comparisons of dual-output two-stream network with other methods on Home276 group.

Network Model	Modality	Preprocessing	Accuracy
SignFi [1] (9-layer CNN)	combination	no	93.98%
SignFi [1] (9-layer CNN)	combination	Phase Shift	98.91%
HOS-Re [3] (SVM)	combination	no	98.26%
Single-stream ResNet18	combination	no	97.90%
Single-stream ResNet18	combination	SVD_20	97.83%
Single-stream ResNet50	combination	no	99.02%
Single-stream ResNet50	combination	SVD_20	98.87%
Dual-output two-stream with ResNet18	combination + difference	no	97.75%
Dual-output two-stream with ResNet18	combination + difference	SVD_20	98.55%
Dual-output two-stream with ResNet50	combination + difference	no	98.88%
Dual-output two-stream with ResNet50	combination + difference	SVD_20	99.13%

SVD_20 represents the top 20 singular values of amplitude and phase with SVD. Combination modality is combination of amplitude and phase. Difference modality means the difference of combination matrix.

Table 4. Comparisons of dual-output two-stream network with other methods on Lab276 group.

Network Model	Modality	Preprocessing	Accuracy
SignFi [1] (9-layer CNN)	combination	no	95.72%
SignFi [1] (9-layer CNN)	combination	Phase Shift	98.01%
HOS-Re [3] (SVM)	combination	no	97.84%
Single-stream ResNet18	combination	no	95.99%
Single-stream ResNet18	combination	SVD_20	95.54%
Single-stream ResNet50	combination	no	96.47%
Single-stream ResNet50	combination	SVD_20	96.30%
Dual-output two-stream with ResNet18	combination + difference	no	96.32%
Dual-output two-stream with ResNet18	combination + difference	SVD_20	96.54%
Dual-output two-stream with ResNet50	combination + difference	no	96.63%
Dual-output two-stream with ResNet50	combination + difference	SVD_20	96.79%

SVD_20 represents the top 20 singular values of amplitude and phase with SVD. Combination modality is combination of amplitude and phase. Difference modality means the difference of combination matrix.

In order to verify the generalization of our proposed network, we also mixed the Home276 group and Lab276 group together. Examples of the mixed group are randomly divided into training data and test data at a ratio of 8:2. The accuracies reported in Table 5 clearly show that our proposed method with ResNet50 (97.08%) is superior to other methods (94.81%, 96.34%) even without SVD preprocessing. SVD still improves performance a bit.

Table 6 shows the comparison results of Lab150 data group. In our proposed method, the accuracy of using SVD preprocessing is also higher than that of not using SVD preprocessing. However, our best result has an accuracy of 95.88%, which is lower than the HOS-Re method (96.23%), but higher than the SignFi method (86.66%). The accuracy difference between our method and the HOS-Re method is less than 1%. Therefore, our method has similar performance to HOS-Re and is significantly better than SignFi method in this evaluation.

Table 5. Comparisons of dual-output two-stream network with other methods on Lab+Home276 group.

Network Model	Modality	Preprocessing	Accuracy
SignFi [1] (9-layer CNN)	combination	no	92.21%
SignFi [1] (9-layer CNN)	combination	Phase Shift	94.81%
HOS-Re [3] (SVM)	combination	no	96.34%
Single-stream ResNet18	combination	no	95.97%
Single-stream ResNet18	combination	SVD_20	95.95%
Single-stream ResNet50	combination	no	96.67%
Single-stream ResNet50	combination	SVD_20	96.60%
Dual-output two-stream with ResNet18	combination + difference	no	96.3%
Dual-output two-stream with ResNet18	combination + difference	SVD_20	96.75%
Dual-output two-stream with ResNet50	combination + difference	no	96.78%
Dual-output two-stream with ResNet50	combination + difference	SVD_20	97.08%

SVD_20 represents the top 20 singular values of amplitude and phase with SVD. Combination modality is combination of amplitude and phase. Difference modality means the difference of combination matrix.

Table 6. Comparisons of dual-output two-stream network with other methods on Lab150 group.

Network Model	Modality	Preprocessing	Accuracy
SignFi [1] (9-layer CNN)	combination	no	-
SignFi [1] (9-layer CNN)	combination	Phase Shift	86.66%
HOS-Re [3] (SVM)	combination	no	96.23%
Single-stream ResNet18	combination	no	93.60%
Single-stream ResNet18	combination	SVD_20	94.20%
Single-stream ResNet50	combination	no	95.24%
Single-stream ResNet50	combination	SVD_20	95.53%
Dual-output two-stream with ResNet18	combination + difference	no	91.18%
Dual-output two-stream with ResNet18	combination + difference	SVD_20	94.67%
Dual-output two-stream with ResNet50	combination + difference	no	95.75%
Dual-output two-stream with ResNet50	combination + difference	SVD_20	95.88%

SVD_20 represents the top 20 singular values of amplitude and phase with SVD. Combination modality is combination of amplitude and phase. Difference modality means the difference of combination matrix.

According to Tables 3–6, we can conclude that using SVD preprocessing can indeed improve the performance of our dual-output two-stream network. As the CNN model in our dual-output dual-stream network, ResNet50 is more suitable for CSI data in sign language recognition tasks than ResNet18. Although the accuracy of our proposed method is lower than other methods in the Lab276 group and Lab150 group, the best results can be obtained in a mixed environment. This means that our method has better generalization capability than other methods.

4. Discussion

Deep learning models generally have achieved greater success due to the availability of massive datasets and extended model depth and parameterization. However, in practice, factors such as memory and computation time during training and testing are important factors to consider when choosing a model from a large number of models. In addition, the success of deep learning also depends on the training data and the model generalization, which is very important for deploying models in practical use because it is difficult to collect training data and train individual models for all

different environments. In other words, the generalization capability is more important for practical use. According to the evaluation results shown in Table 5, our method has better practicability than other methods.

The diversity of input data is very helpful in CNN-based methods. CNN-based methods extract features through training. Input diversity means that CNN can extract more types of features. This can avoid overfitting during network training. Therefore, the proposed dual-output two-stream network uses two modalities of input data and achieves good performance. Moreover, input data containing redundant and irrelevant noise must be preprocessed. This can be proved in the above experiments. Tables 3–6 show that SVD preprocessing can improve the performance of our dual-output two-stream network.

In this study, the experimental results also show that deep learning is not always successful. It can be seen from Table 6 that the HOS-Re method obtains the best result. This method is a traditional machine learning method. It relies on manual feature engineering to calculate a large number of features and use SVM as a classifier. The method is different from CNN-based methods such as SignFi method and our method, which can automatically extract features through training. Through this evaluation, we can know that as long as good functions can be found, traditional machine learning based on feature engineering is still worthy of attention.

5. Conclusions

The sign language includes the positional relationship of gestures in space and the changes of actions over time. In this study, we proposed a dual-output two-stream network and provided two types of modality input data: the combined matrix of amplitude and phase and its difference matrix. SVD completed the preprocessing of the input data. Then, the attention mechanism was used to select the features learned by the network. We evaluated our proposed network on a public dataset SignFi. Experimental results showed that SVD preprocessing improves the performance of our dual-output two-stream network. Compared with other methods, our method has good performance and better generalization capability.

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

Funding: This research was funded by Ministry of Science and Technology, Taiwan, grant number MOST 109-2221-E-155-054.

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

References

1. Ma, Y.; Zhou, G.; Wang, S.; Zhao, H.; Jung, W. SignFi: Sign language recognition using WiFi. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, March 2018*; Association for Computing Machinery: New York, NY, USA, 2018; Volume 2, p. 23.
2. Ahmed, H.F.T.; Ahmad, H.; Aravind, C. Device free human gesture recognition using Wi-Fi CSI: A survey. *Eng. Appl. Artif. Intell.* **2020**, *87*, 103281. [[CrossRef](#)]
3. Farhana Thariq Ahmed, H.; Ahmad, H.; Phang, S.K.; Vaithilingam, C.A.; Harkat, H.; Narasingamurthi, K. Higher Order Feature Extraction and Selection for Robust Human Gesture Recognition using CSI of COTS Wi-Fi Devices. *Sensors* **2019**, *19*, 2959. [[CrossRef](#)] [[PubMed](#)]
4. Grimes, G.J. Digital Data Entry Glove Interface Device. US Patent 4,414,537, 8 November 1983.
5. Shukor, A.Z.; Miskon, M.F.; Jamaluddin, M.H.; bin Ali, F.; Asyraf, M.F.; bin Bahar, M.B. A new data glove approach for Malaysian sign language detection. *Procedia Comput. Sci.* **2015**, *76*, 60–67. [[CrossRef](#)]
6. Kanokoda, T.; Kushitani, Y.; Shimada, M.; Shirakashi, J.-I. Gesture prediction using wearable sensing systems with neural networks for temporal data analysis. *Sensors* **2019**, *19*, 710. [[CrossRef](#)]

7. Ma, Y.; Zhou, G.; Wang, S. WiFi sensing with channel state information: A survey. *ACM Comput. Surv. (CSUR)* **2019**, *52*, 1–36. [[CrossRef](#)]
8. Koller, O. Quantitative survey of the state of the art in sign language recognition. *arXiv* **2020**, arXiv:2008.09918.
9. Cui, R.; Liu, H.; Zhang, C. A deep neural framework for continuous sign language recognition by iterative training. *IEEE Trans. Multimed.* **2019**, *21*, 1880–1891. [[CrossRef](#)]
10. Pu, J.; Zhou, W.; Li, H. Iterative alignment network for continuous sign language recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 4165–4174.
11. Ohn-Bar, E.; Trivedi, M.M. Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations. *IEEE Trans. Intell. Transp. Syst.* **2014**, *15*, 2368–2377. [[CrossRef](#)]
12. Huang, J.; Zhou, W.; Li, H.; Li, W. Sign language recognition using 3d convolutional neural networks. In Proceedings of the 2015 IEEE International Conference on Multimedia and Expo (ICME), Turin, Italy, 29 July 2015; pp. 1–6.
13. Aly, W.; Aly, S.; Almotairi, S. User-independent American sign language alphabet recognition based on depth image and PCANet features. *IEEE Access* **2019**, *7*, 123138–123150. [[CrossRef](#)]
14. Melgarejo, P.; Zhang, X.; Ramanathan, P.; Chu, D. Leveraging directional antenna capabilities for fine-grained gesture recognition. In Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, Seattle, WA, USA, 13–17 September 2014; pp. 541–551.
15. Shang, J.; Wu, J. A robust sign language recognition system with multiple Wi-Fi devices. In Proceedings of the Workshop on Mobility in the Evolving Internet Architecture, Los Angeles, CA, USA, 25 August 2017; pp. 19–24.
16. Li, H.; Yang, W.; Wang, J.; Xu, Y.; Huang, L. WiFinger: Talk to your smart devices with finger-grained gesture. In Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, Heidelberg, Germany, 12–16 September 2016; pp. 250–261.
17. Zhou, Q.; Xing, J.; Li, J.; Yang, Q. A device-free number gesture recognition approach based on deep learning. In Proceedings of the 2016 12th International Conference on Computational Intelligence and Security (CIS), Wuxi, China, 16–19 December 2016; pp. 57–63.
18. Kosba, A.E.; Saeed, A.; Youssef, M. Robust WLAN device-free passive motion detection. In Proceedings of the 2012 IEEE Wireless Communications and Networking Conference (WCNC), Paris, France, 1–4 April 2012; pp. 3284–3289.
19. Yang, Z.; Zhou, Z.; Liu, Y. From RSSI to CSI: Indoor localization via channel response. *ACM Comput. Surv. (CSUR)* **2013**, *46*, 1–32. [[CrossRef](#)]
20. Zhou, Z.; Wu, C.; Yang, Z.; Liu, Y. Sensorless sensing with WiFi. *Tsinghua Sci. Technol.* **2015**, *20*, 1–6. [[CrossRef](#)]
21. Zheng, W.; Zhang, D. HandButton: Gesture recognition of transceiver-free object by using wireless networks. In Proceedings of the IEEE International Conference on Communications, London, UK, 8–12 June 2015; pp. 6640–6645.
22. Choi, J.-S.; Lee, W.-H.; Lee, J.-H.; Lee, J.-H.; Kim, S.-C. Deep learning based NLOS identification with commodity WLAN devices. *IEEE Trans. Veh. Technol.* **2017**, *67*, 3295–3303. [[CrossRef](#)]
23. Kim, S.-C. Device-free activity recognition using CSI & big data analysis: A survey. In Proceedings of the 2017 Ninth International Conference on Ubiquitous and Future Networks, Milan, Italy, 4–7 July 2017; pp. 539–541.
24. Kalman, D. A singularly valuable decomposition: The SVD of a matrix. *Coll. Math. J.* **1996**, *27*, 2–23. [[CrossRef](#)]
25. Soffer, S.; Ben-Cohen, A.; Shimon, O.; Amitai, M.M.; Greenspan, H.; Klang, E. Convolutional neural networks for radiologic images: A radiologist’s guide. *Radiology* **2019**, *290*, 590–606. [[CrossRef](#)] [[PubMed](#)]
26. Hubel, D.H.; Wiesel, T.N. Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex. *J. Physiol.* **1962**, *160*, 106. [[CrossRef](#)] [[PubMed](#)]
27. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [[CrossRef](#)]
28. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–8 December 2012; pp. 1097–1105.
29. Zeiler, M.D.; Fergus, R. Visualizing and understanding convolutional networks. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 818–833.

30. Mateen, M.; Wen, J.; Song, S.; Huang, Z. Fundus image classification using VGG-19 architecture with PCA and SVD. *Symmetry* **2019**, *11*, 1. [[CrossRef](#)]
31. Zhong, Z.; Jin, L.; Xie, Z. High performance offline handwritten chinese character recognition using GoogLeNet and directional feature maps. In Proceedings of the 2015 13th International Conference on Document Analysis and Recognition (ICDAR), Tunis, Tunisia, 23–26 August 2015; pp. 846–850.
32. Marsden, M.; McGuinness, K.; Little, S.; O'Connor, N.E. Resnetcrowd: A residual deep learning architecture for crowd counting, violent behaviour detection and crowd density level classification. In Proceedings of the 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Lecce, Italy, 29 August–1 September 2017; pp. 1–7.
33. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
34. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1492–1500.
35. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
36. Woo, S.; Park, J.; Lee, J.-Y.; So Kweon, I. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
37. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Mobilenets, H.A. Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
38. Zoph, B.; Vasudevan, V.; Shlens, J.; Le, Q.V. Learning transferable architectures for scalable image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8697–8710.
39. Tan, M.; Le, Q.V. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv* **2019**, arXiv:1905.11946.
40. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 568–576.
41. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Van Gool, L. Temporal segment networks: Towards good practices for deep action recognition. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 20–36.
42. Feichtenhofer, C.; Pinz, A.; Wildes, R.P. Spatiotemporal multiplier networks for video action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4768–4777.
43. Feichtenhofer, C.; Pinz, A.; Zisserman, A. Convolutional two-stream network fusion for video action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1933–1941.

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).