Scholar Publishing Group
International Journal of Neural Network
https://doi.org/10.38007/NN.2021.020303
ISSN 2790-2692 Vol. 2, Issue 3: 19-26

A Sign Language Recognition Method Relying On Convolutional Recurrent Neural Network

Wei Dong*

Xizang Minzu University, Xizang, China

**corresponding author*

Keywords: Convolutional Neural Network, Recurrent Neural Network, Sign Language Recognition, CNN-LSTM Model

Abstract: As a communication method widely used among hearing-impaired people, sign language is a natural language that transmits information through the three-dimensional space of the "manual-visual" channel, and solves the communication problem of deaf people through human-computer interaction. The problem is a qualitative leap, and human-computer interaction effectively solves the communication problem between deaf people and ordinary people. Therefore, this paper relies on the convolutional recurrent neural network (CRNN) to study the sign language recognition (SLR) method. This paper first describes the two concepts of SLR and image processing, and then builds the CNN-LSTM model through the CRNN structure, network training and parameter settings, and finally analyzes the CNN-LSTM model. Model analysis shows that the CNN-LSTM model has high SLR accuracy and good recognition performance.

1. Introduction

Sign language is a body language specially used for the communication of deaf and dumb people. It expresses specific semantics by making specific movements and movement trajectories of the hands and arms. It also includes auxiliary factors such as facial expressions and the position of the hands relative to the body. A communication language for reading visually [1-2]. Sign language consists of a series of gestures, which is a typical time series data. At the same time, the changes of gestures involve spatial changes, so sign language data is essentially a spatiotemporal data [3]. Based on the rapid development of science and technology, intelligent gesture recognition machines have become an indispensable communication tool between deaf people and ordinary people. However, in the actual application environment, gestures will increase the difficulty of recognition due to different hand shapes, environments, and states, and people need intelligent devices that are more in line with life habits [4-5].

At present, many experts and scholars have conducted continuous and in-depth research on SLR

methods and convolutional neural networks, and have achieved fruitful results. For example, researchers such as Al-Shamayleh A S proposed a very compact in-place gated CRNN for end-to-end multi-channel speech enhancement, which utilizes in-place convolution for frequency pattern extraction and reconstruction, in-place The features effectively preserve spatial cues in each frequency bin, and utilize a novel spectral restoration method to effectively improve speech quality by predicting magnitude masks, mappings, and phases [6]. Researchers such as Saleh BM proposed a robust SLR method based on deep learning, which represented multimodal information through texture maps to describe hand position and motion, and used this information as the basis for two three-stream and two-stream CNN models. In order to learn to recognize robust features of dynamic signs, the experimental comparison revealed the superiority of the robust SLR method based on deep learning [7]. With the development of science and technology, more and more researchers conduct research on SLR methods based on CRNN.

This paper relies on convolutional recursive neural SLR methods for research and analysis. Therefore, the structure of this paper can be roughly divided into three parts: The first part is to explain the related concepts of SLR, mainly introducing two aspects, namely SLR. The definition related to image processing; the second part is to build the model, in this part, the CNN-LSTM model is built through the three aspects of CRNN structure, network training and parameter setting; the third part is to The analysis of the model, the third part includes the analysis of the accuracy of the model recognition and the analysis of the optimization algorithm.

2. Related Overview

2.1. Sign Language Recognition

Usually SLR is divided into two categories, isolated SLR (a single hand gesture) and continuous SLR (a sequence of hand gestures). Both gestures consist of manual elements and non-manual elements, the former including hand motion and hand shape, and the latter including facial expressions, head motions and body poses [8]. The input for isolated SLR is a sequence of sign language images, but the duration is short, containing only a single gesture, and only the upper body of the performer is recorded. The input data of continuous SLR is continuous sign sentences, including multiple gestures, so it is necessary to identify the boundaries between gestures during the recognition process [9]. Continuous SLR adopts the idea of first segmentation and then recognition. First, the time segmentation network is used to identify the gesture boundary frames in the continuous sign sentence, and the original input is divided into multiple isolated gesture fragments; then, the recognition network is used to analyze the gesture fragments. Feature extraction and classification, combining the recognized gesture label sequences to obtain continuous hand gesture sentences [10].

2.2. Image Processing

After capturing the recognized gesture, the recognized area within the skin color is isolated. For better control over the identified regions, we mask them to enhance the features of the selected regions [11]. The main functions of mask processing are:

(1) Divide the image into an area of interest and an area to be processed. When performing mask processing, multiply the two areas to obtain a new area of interest. The image features in the area of interest remain unchanged, and the remaining areas are 0 [12].

(2) Mask processing can mask specific regions and does not participate in the parameter operation of image features [13].

(3) Extract features in the image that are similar to the mask [14].

(4) Extraction of special feature points [15].

After masking the extracted area, the foreground is separated to reduce the influence of the background skin color area on the recognized area in the foreground, and finally the corresponding gesture dataset that needs model training is obtained [16].

3. Model Building

3.1. CRNN Structure

In Chinese Sign Language, many words are represented by continuous hand movement transformation, so the continuous SLR method for Chinese Sign Language has more research value and application prospect. Vision-based continuous SLR methods aim to extract spatial and temporal features from continuous image inputs and utilize these features for recognition [17]. Convolutional recurrent neural has a strong ability to extract the spatial features of sign language images [18]. But convolutional recurrent neural cannot exploit the temporal features in consecutive image sequences. Therefore, this paper builds a CNN-LSTM model based on a recurrent neural network, and the model structure is shown in Figure 1. As can be seen from Figure 1, the medium that really connects the CNN and the LSTM network is the vector representation of the image features that are both the output of the CNN model and the input of the LSTM network.

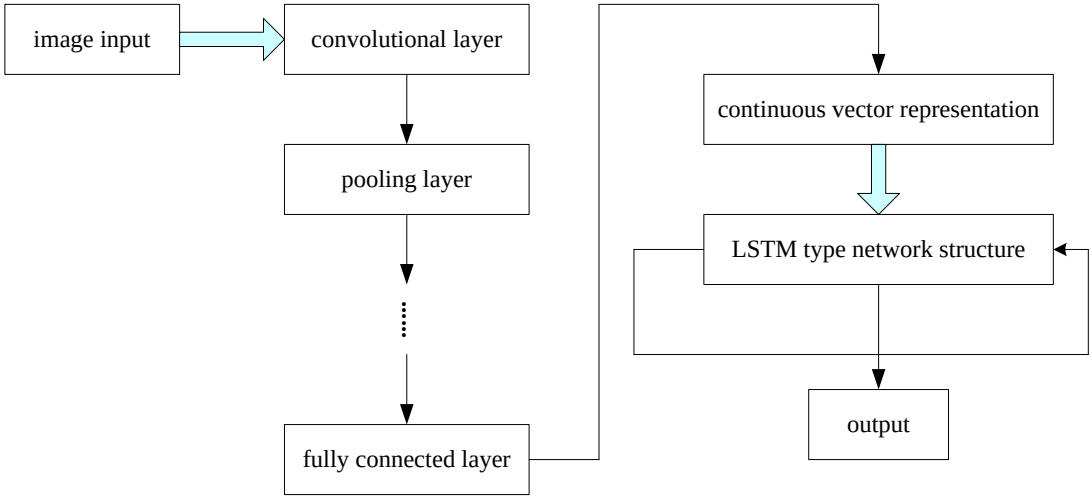

Figure 1. CNN-LSTM model structure diagram

3.2. Network Training

The following describes the process of training data for a convolutional neural network (CNN). The first part is to extract the representative features step by step, the convolution check the input image convolution algorithm, the final result is output by the activation function, and the recognition result value is obtained, that is, the forward propagation process. The output result value is compared with the expected value. If the binary difference is within the allowable range, the output result value can be classified as the final result. This is the best case for the experiment, but it often occurs rarely; The second stage is back-propagation. When the result value does not match the actual value, the training process starts to transmit from high-level to low-level features, calculates the error between neurons, obtains the error gradient, updates the weight parameters, and then calculates Each unit outputs, and then enters the second step of the forward propagation process, and repeats the training until the error is small or non-existent, and the training can be ended with a

fixed weight. The training process is shown in Figure 2.

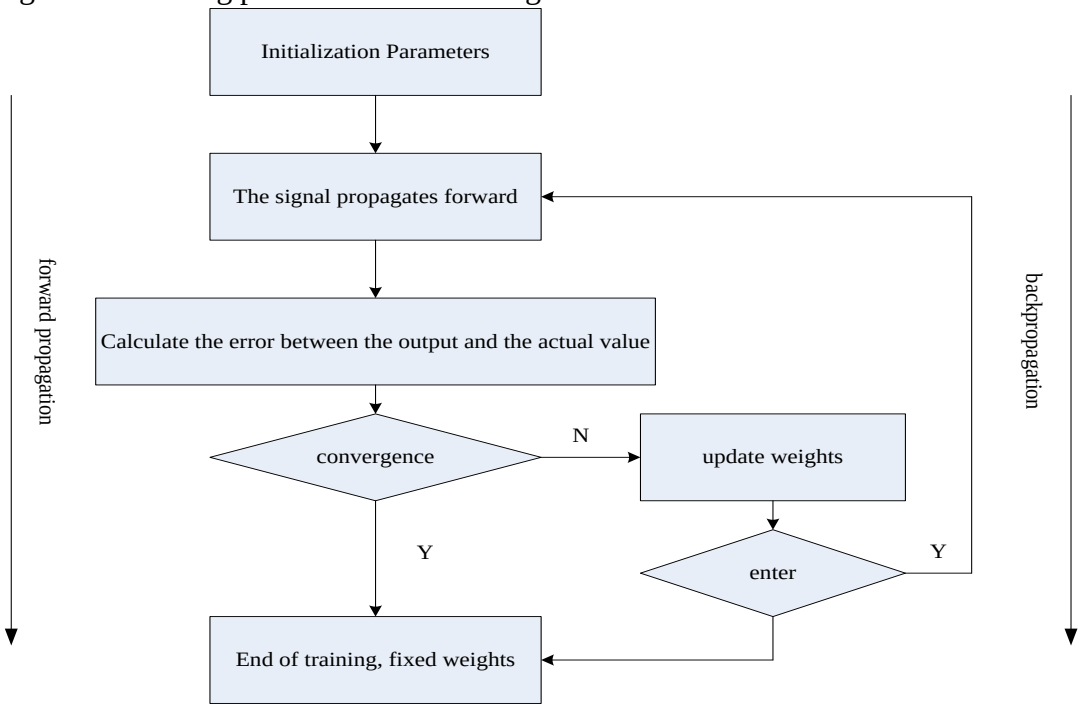

Figure 2. Flowchart of the training process

3.3. Parameter Setting

The CNN-LSTM fusion model is relatively complex. Considering the amount of training data, memory and time costs, it is extremely difficult to directly train the network for the entire fusion model. Therefore, a divide-and-conquer method is used to train the model. First, the model is trained and the parameter results of each layer are saved. Then, the entire fusion model is built, and the pre-trained parameters are directly loaded into the corresponding convolutional, pooling, and fully connected layers. Finally, the upper body images in the sign language videos are continuously captured and preprocessed uniformly. Each sign language vocabulary corresponds to multiple videos, and each video corresponds to a set of preprocessed pictures. Therefore, the training data set can be regarded as a collection of multiple sets of pictures with multiple words. These pictures are scaled in groups, and the pictures are ordered into a dataset for training LSTM-type networks. When training the CNN-LSTM fusion model, the entire data set is directly input into the network structure, and the parameters of each layer are fixed, and only the parameters of the LSTM network part are trained. This training method improves the parameter training speed of the CNN-LSTM fusion model. In this paper, when training the LSTM network structure, the initial weights are randomly assigned, which obey the natural distribution in the range of 0.01. The network algorithm is optimized according to the following formula:

$$LSTM\left|f\right|_d = \sqrt{E\left|f^2\right|_d + \alpha} \tag{1}$$

$$\Delta w_d = -\frac{\beta}{LSTM\left|f\right|_d} * f_d \tag{2}$$

where fd is the current gradient, α is a small number to keep the denominator non-zero, β is the initial learning rate, and E is the expectation.

4. Model Analysis

4.1. Model Recognition Accuracy Analysis

In order to compare the advantages of the CNN-LSTM network model in dynamic recognition, the CNN-LSTM model was compared with the HMM model and the CNN-RBM model, and the recognition rates of the three models for dynamic continuous sign language were compared through specific data, and a comparative experiment was conducted. The experimental results are as follows shown in Table 1.

Table 1. Recognition algorithm accuracy numerical table

	80% Training Data Training Recognition Rate(%)	80% Training Data Test Recognition Rate(%)
HMM	91.45	92.64
CNN-RBM	94.78	92.17
CNN-LSTM	98.59	97.63

It can be seen from Table 1 that under the same conditions in the test environment, the training recognition rates of the SLR methods of the HMM, CNN-RMB and CNN-LSTM models are 91.45%, 94.78%, and 98.59%, respectively, and the test recognition rates are 92.64% %, 92.17%, 97.63%, through the comparative analysis of the data, it is found that the accuracy of CNN-LSTM SLR is significantly higher than that of HMM and CNN-RBM, so the CNN-LSTM model has higher SLR. Accuracy and good robustness.

4.2. Analysis of Optimization Algorithms

In the experiment, four algorithms were selected for testing. The four algorithms were 3DCNN, LSTM, BiLSTM and RN-BiLMTS algorithms. In order to obtain the test results quickly, we selected the CNN structure of 128-dimensional vector for testing, and carried out 50 experiments on the four algorithms. Iteratively, by observing the convergence speed of the loss function, the optimization algorithm that is most suitable for the training of the CNN-LSTM network model of this network is selected.

The LSTM network is the best way to solve sequence analysis problems, selectively forgetting unimportant data during training. The BiLSTM network can encode information from the back to the front, and the forward LSTM and the backward LSTM are combined into a BiLSTM. The RN-BiLMTS algorithm is a combination model of ResNet and BiLSTM network. The algorithm combines the underlying features in the video with the extracted deep features, which can accurately and effectively distinguish different gestures. Figure 3 shows the recognition accuracy obtained experimentally on the self-built data set, and Figure 4 lists the comparison curve of the loss rate between the RN-BiLMTS algorithm and the first three proposed algorithms, which intuitively shows the change trend.

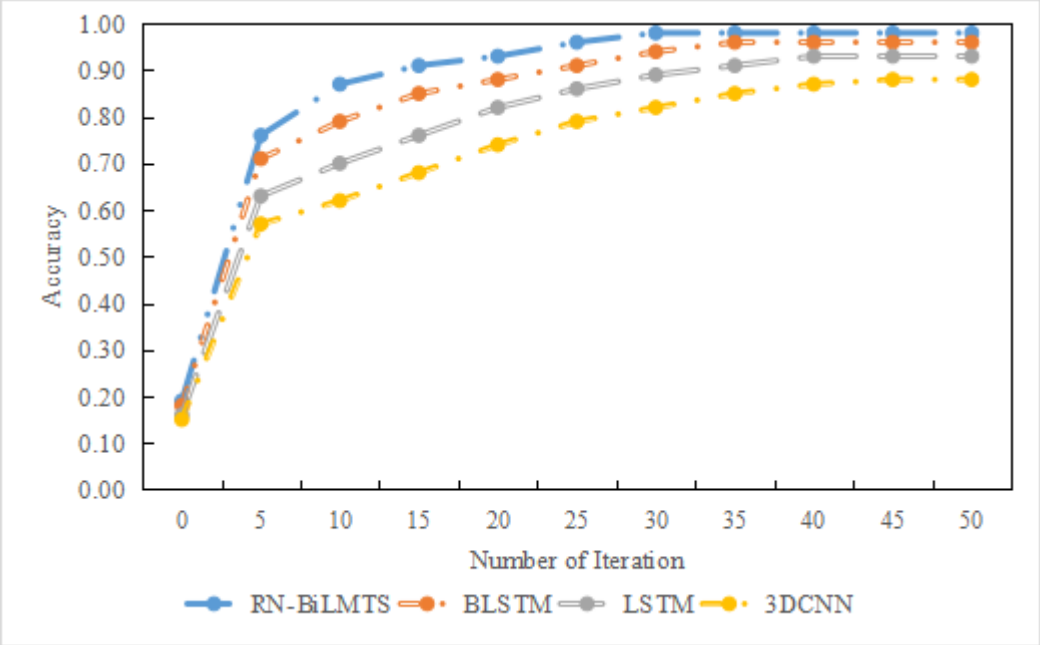

Figure 3. Accuracy curves of different algorithms

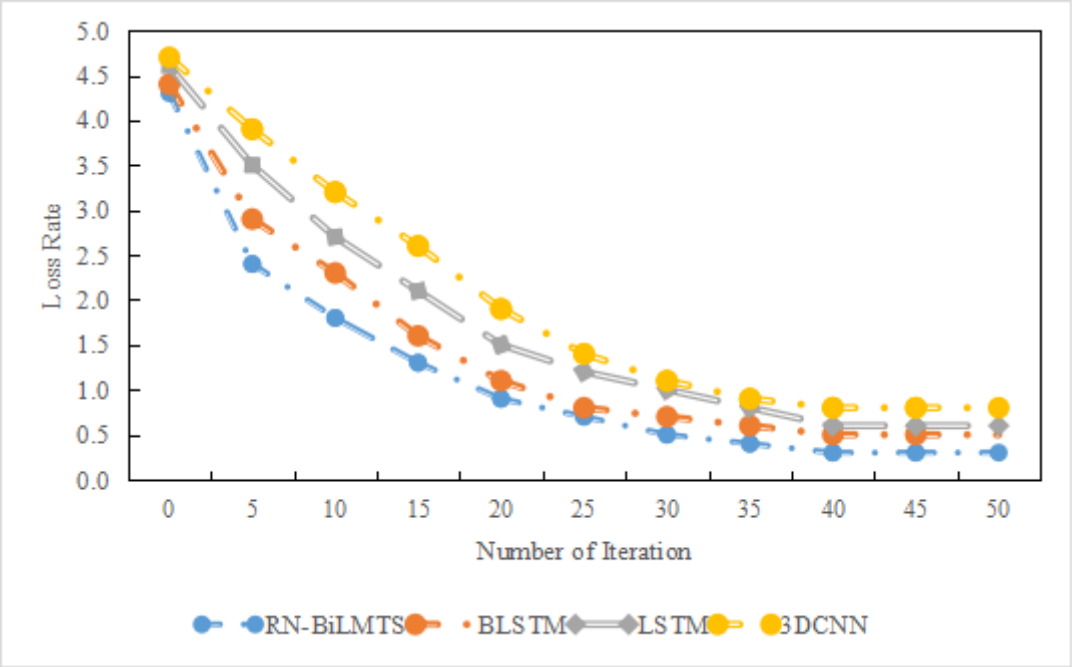

Figure 4. Loss function graph

From the analysis of the accuracy rate curve in Figure 3, it can be seen that the accuracy rate of RN-BiLMTS is the highest, indicating that the algorithm has the best recognition effect and good performance of the algorithm. It can be seen from Figure 4 that RN-BiLMTS has the fastest training speed, the fastest convergence speed, high accuracy, and high fit between samples and the loss function. The accuracy rate reached the highest after iterating to 35 rounds, and the highest accuracy rate was as high as 98%, so the accuracy curve no longer changed significantly. In Figure 4, the RN-BiLMTS loss function curve was iterated to 40 rounds. The number of times increases, the network becomes more and more mature, and the generalization ability gradually improves.

Comprehensive analysis can be obtained through the analysis of the data set. Compared with the other three algorithms, the RN-BiLMTS optimization algorithm has higher SLR accuracy., the recognition effect is the best.

5. Conclusion

The improvement of the SLR method is beneficial to the communication and communication between the deaf and the ordinary people, so this paper relies on the CRNN to study the SLR method. This paper studies the SLR method based on CRNN, and draws the following conclusions. Through the comparative analysis of CNN-LSTM and HMM method and CNN-RBM method, it is found that the CNN-LSTM model has high accuracy and good robustness in SLR. Through the comparative experimental analysis of the four algorithms of 3DCNN, LSTM, BiLSTM and RN-BiLMTS, it can be seen that the RN-BiLMTS optimization algorithm has high SLR accuracy and the best recognition effect. In this paper, the method of SLR is studied and analyzed, but there are still many areas for improvement. Research on SLR based on CRNN is a good research direction.

Funding

This article is not supported by any foundation.

Data Availability

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Conflict of Interest

The author states that this article has no conflict of interest.

References

[1] Lee M, Lee S, Lee J, et al. Predicting Maximum Wind Speed of Typhoons based on Convolutional Recurrent Neural Network via COMS Satellite Data. Journal of Korean Institute of Industrial Engineers. (2019) 45(4):349-360. https://doi.org/10.7232/JKIIE.2019.45.4.349

[2] Park M, Kim H, Park S. End-to-End Autonomous Driving Based on Convolutional Recurrent Neural Network. Journal of Korean institute of intelligent systems. (2019) 29(4):297-301. https://doi.org/10.5391/JKIIS.2019.29.4.297

[3] Kumar E K, Kishore P, Kumar M, et al. Three-Dimensional Sign Language Recognition With Angular Velocity Maps and Connived Feature ResNet. IEEE Signal Processing Letters. (2018) 25(12):1860-1864. https://doi.org/10.1109/LSP.2018.2877891

[4] Savant R, Ajay A. Indian Sign Language Recognition System for Deaf and Dumb Using Image Processing and Fingerspelling: a Technical Review. National Journal of System and Information Technology. (2018) 11(1):23-34.

[5] Gndz C, Polat H. Turkish Sign Language Recognition Based on Multistream Data Fusion. Turkish Journal of Electrical Engineering and Computer Sciences. (2021) 29(2):1171-1186. https://doi.org/10.3906/elk-2005-156

[6] Al-Shamayleh A S, Ahmad R, Jomhari N, et al. Automatic Arabic Sign Language Recognition: A Review, Taxonomy, Open Challenges, Research Roadmap And Future Directions. Malaysian Journal of Computer Science. (2020) 33(4):306-343. https://doi.org/10.22452/mjcs.vol33no4.5

[7] Saleh B M, Al-Beshr R I, Tariq M U. D-Talk: Sign Language Recognition System for People with Disability using Machine Learning and Image Processing. International Journal of Advanced Trends in Computer Science and Engineering. (2020) 9(4):4374-4382. https://doi.org/10.30534/ijatcse/2020/29942020

[8] Tabassum T, Mahmud I, Uddin M P, et al. Enhancement of Single-Handed Bengali Sign Language Recognition Based on Hog Features. Journal of Theoretical and Applied Information Technology. (2020) 98(5):743-756.

[9] Kasmin F. A Comparative Analysis of Filters towards Sign Language Recognition. International Journal of Advanced Trends in Computer Science and Engineering. (2020) 9(4):4772-4782. https://doi.org/10.30534/ijatcse/2020/84942020

[10] Kaushik N, Rahul V. A Survey of Approaches for Sign Language Recognition System. International Journal of Psychosocial Rehabilitation. (2020) 24(1):1775-1783. https://doi.org/10.37200/IJPR/V24I1/PR200278

[11] Mallick T, Balaprakash P, Rask E, et al. Graph-Partitioning-Based Diffusion Convolutional Recurrent Neural Network for Large-Scale Traffic Forecasting: Transportation Research Record. (2020) 2674(9):473-488. https://doi.org/10.1177/0361198120930010

[12] Hoshi I, Shimobaba T, Kakue T, et al. Single-Pixel Imaging Using a Recurrent Neural Network Combined with Convolutional Layers. Optics express. (2020) 28(23):34069-34078. https://doi.org/10.1364/OE.410191

[13] Rokade P, Sali N, Shinde D, et al. Indian Sign Language Recognition System in Marathi Language Text. International Journal of Computer Sciences and Engineering. (2019) 7(5):881-885. https://doi.org/10.26438/ijcse/v7i5.881885

[14] Elakkiya R, Vanitha V. Interactive Real Time Fuzzy Class Level Gesture Similarity Measure based Sign Language Recognition Using Artificial Neural Networks. Journal of Intelligent & Fuzzy Systems. (2019) 37(5):6855-6864. https://doi.org/10.3233/JIFS-190707

[15] Ibrahim N B, Zayed H H, Selim M M. Advances, Challenges, and Opportunities in Continuous Sign Language Recognition. Journal of Engineering and Applied Sciences. (2019) 15(5):1205-1227. https://doi.org/10.36478/jeasci.2020.1205.1227

[16] Kotari M. Sign Language Recognition using Image-based Hand Gesture. Global Journal of Engineering Science and Research Management. (2019) 2(5):237-240.

[17] ME Martínez-Gutiérrez, JR Rojano-Cáceres, E Benítez-Guerrero, et al. Data Acquisition Software for Sign Language Recognition. Research in Computing Science. (2019) 148(3):205-211. https://doi.org/10.13053/rcs-148-3-17

[18] Ravi S, Maloji S, Polurie V, et al. Sign language recognition with multi feature fusion and ANN classifier. Turkish Journal of Electrical Engineering and Computer Sciences. (2018) 26(6):2872-2886. https://doi.org/10.3906/elk-1711-139