Research of a Sign Language Translation System Based on Deep Learning

Siming He
Ridley College, St. Catharines, Canada

Abstract—**Sign language is a visual language that uses hand gesture, change of hand shape and track information to express meaning, and is the main communication tool for people with hearing and language impairment. Sign language recognition can improve the problem that the number of people who need to use sign language is large but the popularity of sign language is poor, and provide a more convenient way of study, work and life for people with hearing and language impairment. Hand locating and sign language recognition methods can generally be divided into traditional methods and deep learning methods. In recent years, with the brilliant achievements of deep learning in the field of computer vision, it has been proved that the method based on deep learning has many advantages, such as rich feature extraction, strong modeling ability and intuitive training. Therefore, this paper studies hand locating and sign language recognition of common sign language based on neural network, and the main research contents include: 1. A hand locating network based on the Faster R-CNN is established to recognize the sign language video or the part of the hand in the picture, and the result of recognition is handed over to subsequent processing; 2. A 3D CNN feature extraction network and a sign-language recognition framework based on long and short time memory (LSTM) coding and decoding network are constructed for the sign language images of sequence sequences. The framework can improve the recognition accuracy by learning the context of sign language; 3. In order to solve the problem of RGB sign language image or video recognition in practical problems, this paper combines hand locating network, 3D CNN feature extraction network and LSTM encoding and decoding to build the recognition algorithm. Experimental results show that the recognition rate of this method is up to 99% in common vocabulary data set, which is better than other known methods.**

Keywords—Sign Language Recognition, Deep Learning, Convolutional Neural Network, Recurrent Neural Network, LSTM

I. Background Information

According to the World Health Organization, there are 328 million people worldwide suffer from impaired hearing loss, of whom 32 million are children. Sign language is the daily language of communication between deaf and dumb people, which is the most comfortable and natural way of communication between deaf and dumb people, and is also the main tool for special education schools to teach and convey ideas. Sign language is a natural language that conveys meaning through the shape, position, movement of hands and facial expressions. Similar to other natural languages, sign language has a standardized grammar and a complete vocabulary system. However, there are very few people with normal hearing who are proficient in sign language, and in many countries, the theoretical research on translation of sign language is still in its infancy.

Sign language recognition methods are easily affected by human movement, change of gesture scale, small gesture area, complex background, illumination and so on. And some sign language recognition methods must use gesture areas to input information. Therefore, robust hand locating is an important pretreatment step in sign language recognition. Compared with basic gestures, gestures in sign language are characterized by complex hand shape, blurred movement, low resolution of small target area, mutual occlusion of hands and faces, and overlapping of left and right hands.

In addition to the influence of complex background and light, a large number of sign language image sequences are needed in sign language recognition, and all these have brought great challenges to the accuracy and stability of hand locating in sign language recognition. Sign language is composed of continuous gestures, therefore, for sign language recognition, in addition to spatial domain, it also needs to capture motion information across multiple consecutive video frames. At the same time, how to build an efficient and suitable sign language recognition model has always been a hot research spot.

II. Research Status

The first problem of sign language research is data collection. The current method is that researchers record or shoot hand movements of sign language users through sensors or cameras, and carry out the corresponding feature extraction and model establishment of the data flow in the computer. Data acquisition can be done mainly with data gloves, video cameras, and new motion sensing devices.

In term of the hand locating, the current methods can be divided into traditional hand locating algorithm and hand locating algorithm based on deep learning. In the detection algorithm based on traditional methods, the researchers use prior knowledge to extract features, and combine the method of classifier to detect gestures. As the visual color feature of human hand, skin color feature can distinguish most objects from human hands. Some scholars also use Bayesian model to train skin color model with skin color pixel points as sample data, and use region growth to obtain relatively complete human hand region, making this method not easily disturbed by the shape change of human hand, scale change and rotation. With the successful application of Convolutional Neural Network (CNN) and target detection algorithm, a new research direction has been brought to hand locating. The Faster R-CNN model of Convolutional Neural Network is proposed, and this model uses the CNN feature for target recognition, which greatly improves the accuracy of hand locating.

On the problem of sign language interpretation, some

978-1-7281-4691-1/19/$31.00 ©2019 IEEE
DOI 10.1109/AIAM48774.2019.00083

scholars use HOG feature and three-dimensional track feature of human skeletal points to evaluate the number of sign language words in continuous sentences, and the recognition rate of 87.7% is achieved by using mathematical modeling method to recognize words one by one. With the development of neural network, Recurrent Neural Network (RNN) has also been widely used in sign language recognition.

III. Hand locating Based on Faster R-CNN

Sign language is mainly the movement of the hand, and has nothing to do with the complex background and movement of the human body. The interference of external factors will affect the result of sign language recognition. Therefore, the hand locating of sign language words is not only a crucial pretreatment step in sign language recognition, but also an essential step to extract gesture features and further gesture simulation.

In the image sequence of sign language, the gesture is accompanied by the interference of a large number of skin-like region, which has the characteristics of complex and variable hand shape, much special hand shape and fuzzy hand movement. Therefore, the traditional methods of skin color detection and template classification cannot locate and classify gestures well. As one of the most advanced target detection networks, Faster R-CNN integrates the RPN module and the Faster R-CNN measurement module into an end-to-end network, so as to obtain better performance in terms of speed and accuracy. Compared with the single-stage target detection algorithm such as YOLO, Faster R-CNN can better meet the requirements of accurate detection and location of gestures. The overall structure is shown in the Fig.1.

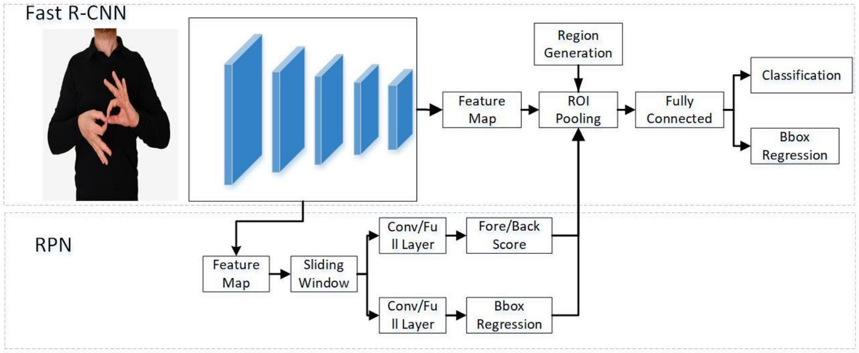

Figure 1. Structure chart of Faster R-CNN

Compared with Fast R-CNN, Faster R-CNN designs the RPN module. RPN is a convolutional neural network, which uses the shared CNN feature to generate candidate regions, and replaces the traditional selective search candidate region extraction algorithm, significantly improving the accuracy of candidate regions. At the same time, because the RPN module shares convolutional network with Fast R-CNN detection module, the detection speed of the Faster R-CNN module is also improved significantly.

We train the Faster R-CNN hand locating network by using a data set consisting of 40 common words and 10,000 sign language images, among which the optimizer uses Stochastic Batch Gradient Descent (SGD). The changing process of loss training is shown in Fig.2. It can be found from the figure that the loss value of network decreases gradually with the increase of training times, and at 55000 iterations, the network tends to converge without overfitting. In Table 1, it compares the accuracy of hand locating of Faster R-CNN and the other two methods.

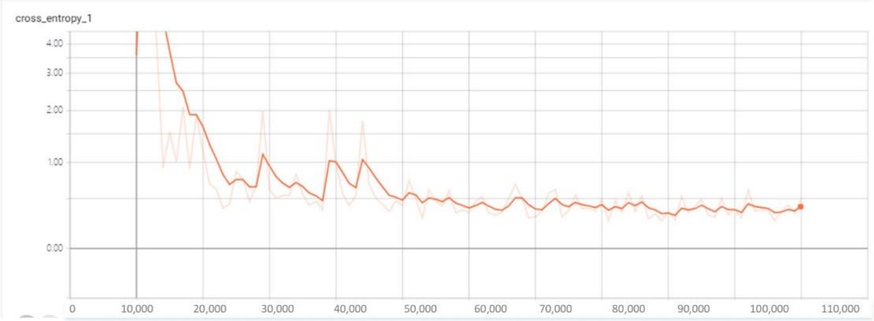

Figure 2. Trend chart of loss of network training

TABLE I. Detection results of each algorithm

Methors	mAP(%)	Right-Hand	Left-Hand	Both-Hand
YOLO	83.2	80.5	81.7	87.3
Fast R-CNN	89.0	86.1	88.4	92.5
Faster R-CNN	91.7	89.2	89.8	96.2

As can be seen from Table I, compared with the reference method YOLO, the Faster R-CNN increases the mAP detection results from 83.2% to 91.7%. The detection accuracy of Faster R-CNN increases from 89.0% to 91.7% compared with Fast-RCNN. The experiment shows that the Faster R-CNN is more suitable for hand locating in sign language, which can better detect and locate gestures effectively under the interference of skin color, background, movement blur and hand, face occlusion and so on. The hand locating by Faster R-CNN is shown in the Fig.3

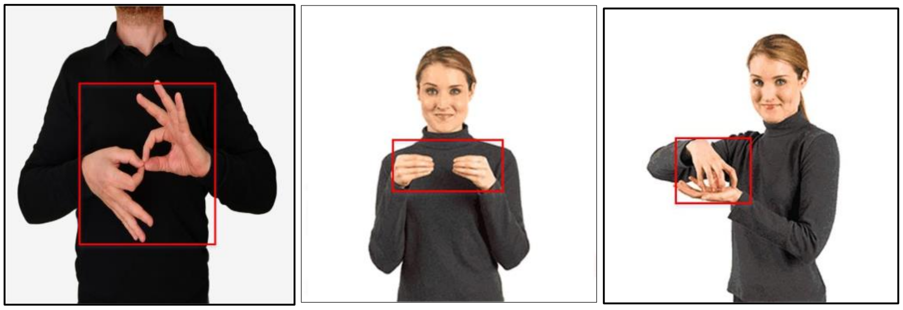

Figure 3. Detection Result of Faster R-CNN

IV. 3D CNN Network Structure

In order to solve the feature extraction and classification of sign language, this paper proposes a new 3D CNN structure, as shown in Figure 1. The samples pass through 4 Conv Blocks in sequence, and the input of each Conv Block is sampled to the same size as the output through the operation of 3D Average Pooling. Then they are concatenated with the feature map extracted by Conv Block to form new features, and input into the next Conv Block. After the output of Convblock5, the feature graph of 1×1×1 is obtained by using a Global Max Pooling, which passes through Dense Block, and get the probability of dichotomy by using Sigmoid function.

A. Conv Block Structure

The sequence of Conv Block consists of 1 BN layer, activation function, 3D convolutional layer and 3D Max Pooling layer. The first four Conv Blocks use CReLU as the activation function, and the 5 Conv Block CReLUs replace Leaky ReLU, and do the operation of Pooling. Batch Normalization is usually located in front of the activation function. Through the normalization operation of each dimension of the data in Batch, it prevents saturated nonlinear function from gradient dispersion due to too large or too small input. At the same time, the normalization operation can prevent a large learning rate from causing a gradient explosion of back propagation, reduce the requirement of parameter initialization, and accelerate CNN training. The activation function is the CReLU function, that is:

$$\mathrm{CReLU}(x) = [\mathrm{ReLU}(x), \mathrm{ReLU}(-x)] \qquad (1)$$

It doubles the input dimension, removes the convolution kernel of Pair-Grouping Phenomenon, reduces redundancy, and improves parameter utilization ratio. In the fifth Conv Block, it doesn't do the operation of 3D Max Pooling. The activation function is the Leaky ReLU, and takes the parameter $a = 0.01$, that is:

$$\text{Leaky ReLU}(x) = \max(x\,, ax) \qquad (2)$$

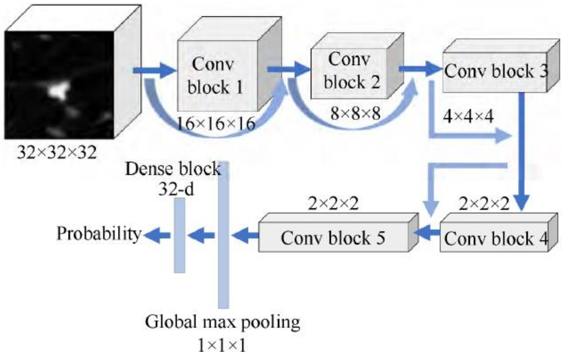

Figure 4. 3D CNN Network Structure

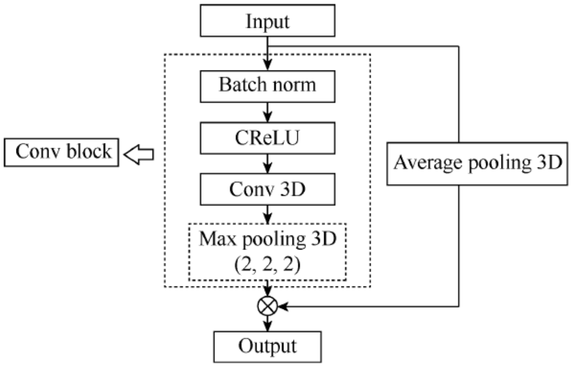

Figure 5. Structural Chart of Conv Block

B. Dense Block Structure

Conv Block extracts features from images, while Dense Block acts as a classifier. The structure of Dense Block is shown in Fig.6, which consists of 2 Fully Connected Layer, 1 Batch norm layer and 1 Leaky ReLU activation function. Finally, it generates the confidence of binary classification by Sigmoid function.

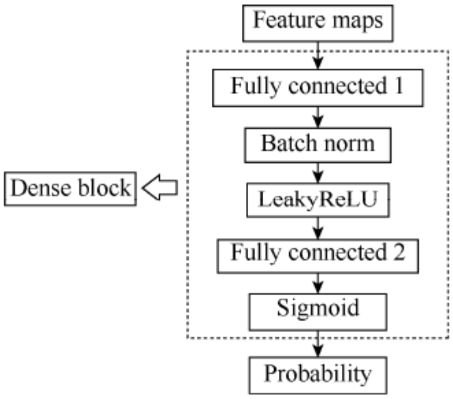

Figure 6. Structural Chart of Dense Block

V. Sign Language Vocabulary Recognition Based on LSTM Encoding and Decoding

LSTM encoding and decoding network applies the sequence-to-sequence model to sign language recognition, which enables the sign language recognition model to handle input frames of variable length, and also learn and use the temporal structure information of the image sequences of sign language recognition.

The long and short time memory neural network (LSTM) used in this project is a chain-type network proposed by Hochreiter et al based on time series. Gers et al. added forget gate on the basis of Hochreiter et al. The LSTM model used in this paper is the most basic LSTM model, with 34 hidden layers. The calculation process mainly involved in LSTM model is shown in the following formulas:

$$\begin{aligned} \mathrm{i}_t &= \sigma(W_i * (h_{t-1} + x_t) + b_i) \\ \mathrm{f}_t &= \sigma(W_f * (h_{t-1} + x_t) + b_f) \\ \mathrm{o}_t &= \sigma(W_o * (h_{t-1} + x_t) + b_o) \\ \tilde{\tilde{\mathrm{C}}}_t &= \tanh(W_c * (h_{t-1} + x_t) + b_c) \\ C_t &= f_t * \tilde{\tilde{\mathrm{C}}}_t + \mathrm{i}_t * \tilde{\tilde{\mathrm{C}}}_t \\ h_t &= o_t * \tanh(C_t) \end{aligned} \tag{3}$$

The first three are the formulas of input gate, forget gate and output gate respectively, and the fourth and fifth formulas update the node state. The final formula calculates the final output of the memory unit. LSTM model processes the information of sequence data based on the linear form of superposition, avoiding the problem of gradient disappearance, and also learns information with a longer learning cycle, which is very suitable for this project. Fig.7 is the structural chart of LSTM model.

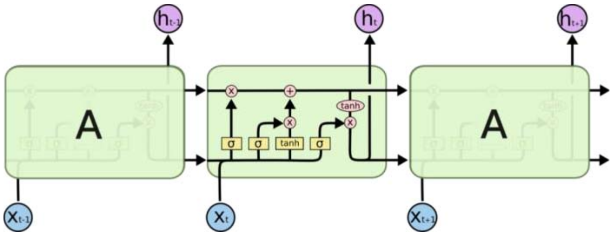

Figure 7. Structural Chart of LSTM Model

VI. Sign Language Recognition System Based on 3D CNN and LSTM Encoding and Decoding

In sign language recognition, the changes of hand shape and gesture shape have a crucial influence on the recognition results. However, the gesture area in sign language video is small, and the whole video frame contains the interference of external factors such as complex background area, the wear of sign language speakers and non-specific sign language speakers. At the same time, for sign language recognition based on RGB image sequence, in addition to the spatial domain, the capturing of movement information across multiple consecutive video frames is also critical. Therefore, it is necessary to study how to extract the features from the sign language image sequences, so as to obtain the distinctive sign language feature expression.

This paper uses 3D CNN which can extract video representation rich in sports information, and also uses double-way 3D CNN which can input the whole video frame and the acquired gesture image sequence as two channels, so as to extract features. Then it inputs the features into the LSTM encoding and decoding network, and builds a sign language model to realize the sequence-to-sequence sign language recognition. The overall process is shown in the Fig.8.

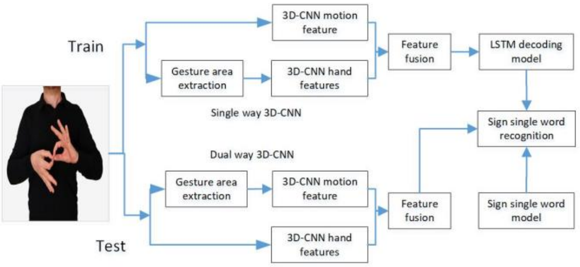

Figure 8. Double-way 3D CNN and LSTM Encoding and Decoding Network Structure

The overall process framework consists of three parts: the detection and tracking method used for gesture area acquisition, the feature extraction method of 3D CNN used for feature extraction of sign language, and the LSTM encoding and decoding network used for sign language word unit recognition.

It uses 3D CNN for feature extraction of image fragment sequences, and select Pool5 layer feature for integration, so as to combine the global information $\mathrm{M}_{\mathrm{global}}$ from the upper road and the local information $\mathrm{M}_{\mathrm{local}}$ from the lower road. The formula of feature integration is as follows:

$$\mathrm{C}_{global} = F_{global}(X_i) \tag{5}$$

$$\mathrm{C}_{local} = F_{local}(X_h) \tag{6}$$

$$\mathrm{C} = \mathrm{G}(\mathrm{C}_{global}, \mathrm{C}_{local}) \tag{7}$$

Among them, X_{i} represents the complete sequence of images, X_{h} is the acquired sequence of gesture images, and through the transformation functions $\mathrm{F}_{\mathrm{global}}$ and $\mathrm{F}_{\mathrm{local}}$, it uses the trained $\mathrm{M}_{\mathrm{global}}$ and $\mathrm{M}_{\mathrm{local}}$ models to spread forward, so as to get the features $\mathrm{C}_{\mathrm{global}}$ and $\mathrm{C}_{\mathrm{local}}$ of Pool5. Finally, carry out the feature integration operation G, to get the global situation after integration - local feature C, the formula is as follows:

$$\mathrm{C(p)} = \mathrm{concate}(\mathrm{C}_{global}(\mathrm{p}), \mathrm{C}_{local}(\mathrm{p})) \qquad (8)$$

The stitching function is represented by Concate, and for fragment p, the global-local feature $\mathrm{C(p)}$ can be finally obtained after integration. Finally, this feature is input into LSTM encoding and decoding network for sign language semantic coding and sign language word unit decoding.

VII. Experimental Results and Analysis

We divide a total of 40 common words and 10,000 sign language images into training set and verification set according to the ratio of 8:2, and the parameters used are shown in the following Table II:

TABLE II. Parameter design of network

Parameter	Image_Global	Fusion_Feature	Embed_Words	Words
	RGB 18*18	1024dims	48dims	40dims
Parameter	Image_feature	decoder_LSTM	encoder_steps	hand_feature
	512dims	1000	30	512dims
Parameter	Optimizer	encoder_LSTM	decoder_steps	Batch_size
	Adam	1000	5	100

In the training, the single-way gesture feature is used as input, and the double-way global-local feature is used as input respectively. The number of training rounds is 200, and the variation trend of loss during the respective training is shown in Fig.8. According to the figure, the input network of single-way gesture feature tends to converge at 150 turns, and the input network of double-way global-local feature tends to converge when the number of turns is about 100. The recognition results are shown in Table III.

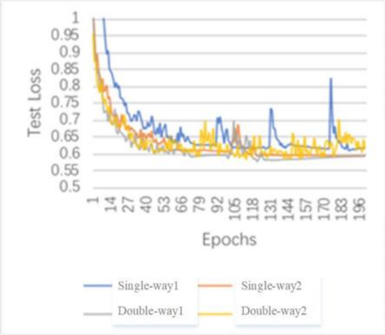

Figure 9. Variation Trend of Loss During Training

TABLE III. The recognition rate of each method

Method	Feature	Accuracy Rate
LSTM_fc	Trajectory feature	91.6%
3D-CNN	Gesture picture+3D Feature	91.5%
3D ResNet-18+SVM_local	Gesture picture+3D Feature	96.9%
3D ResNet-18+SVM_fusion	Global-local+3D Featur	98.3%
Our method_fusion	Global-local+3D Featur	99.0%

The recognition accuracy of other methods is shown in Table 3. According to the data, the recognition results of the method used in this paper reach 99.0%, which is higher than the above experimental data. From this we can see that the global-local feature realizes the complementarity of global and local features, and improves the experimental results. At the same time, compared to other methods, this method only uses RGB data stream, which is helpful for the application of sign language recognition in daily life.

VIII. Conclusion and Prospect

A. Conclusion

With the continuous breakthrough of neural network in artificial intelligence, computer vision and other related fields, neural network has brought dynamic new methods to the study of sign language recognition based on vision. Starting from the task of common sign language word recognition, and focusing on the topic of sign language locating and sign language recognition based on neural network, this paper discusses the design of hand locating algorithm based on deep learning, the feature extraction based on 3D CNN and the recognition algorithm based on recurrent neural network LSTM, and achieves better recognition results than other methods on common vocabulary data sets.

B. Prospect

Although this paper has achieved high accuracy, the data set is limited in scope and does not include all the sign language words. As a research direction with broad application and development space, sign language recognition still has much room for improvement. At the same time, most methods of sign language recognition now only consider the accuracy of the algorithm. However, for the application of sign language recognition in real scene, real-time performance is another important index. Therefore, it is also a direction worthy of breakthrough about how to improve the speed of hand locating and recognition of sign language words.

References

[1] Zhang,Ce,etal. “A hybrid MLP-CNN classifier for very fine resolution remotely sensed image classification.” ISPRS Journal of Photogrammetry and Remote Sensing140 (2018):133-144.

[2] Oyedotun,Oyebade K., and Adnan Khashman. “Deep learning in vi-sion-based static hand gesture recognition.”Neural Computing and Ap-plications 28.12(2017):3941-3951.

[3] Neverova,Natalia,etal. “Multi -scale deep learning for gesture detec-tion and localization.” Workshop at the European conference on computer vision. Springer,Cham,2014.2010,06(21):6045-6046.

[4] Huang J,Zhou W,Zhang Q,etal. Video -based Sign Language Recognition without Temporal Segmentation[J].2018.

[5] LeCun,Yann,Yoshua Bengio,and Geoffrey Hinton. “Deep learning.” nature 521.7553 (2015):436.

[6] SIMONYAN K,ZISSERMAN A. Very deep convolutional networks for large - scale image recognition[EB / OL].details /61625387.

[7] HE K,ZHANG X,REN S,et al. Deep residual learning for image recognition[J].IEEE conference on computer vision and pattern recognition(CVPR) ,2016: 770 - 778.

[8] REN S,HE K,GIRSHICK R,et al. Faster r -cnn: To-wards real - time object detection with region proposal net-works[J]. Advances in neural information processing sys-tems,2015: 91 - 99.