Received February 24, 2019, accepted March 9, 2019, date of publication March 14, 2019, date of current version April 5, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2904749

Dynamic Sign Language Recognition Based on Video Sequence With BLSTM-3D Residual Networks

YANQIU LIAO[1], PENGWEN XIONG[1], WEIDONG MIN[2], (Member, IEEE), WEIQIONG MIN[3], AND JIAHAO LU[1]

[1]School of Information Engineering, Nanchang University, Nanchang 330031, China
[2]School of Software, Nanchang University, Nanchang 330047, China
[3]School of Tourism, Jiangxi Science & Technology Normal University, Nanchang 330038, China

Corresponding author: Weidong Min (minweidong@ncu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61762061and Grant 61663027, and in part by the Natural Science Foundation of Jiangxi Province, China, under Grant 20161ACB20004 and Grant 20181BAB211019.

ABSTRACT Sign language recognition aims to recognize meaningful movements of hand gestures and is a significant solution in intelligent communication between the deaf community and hearing societies. However, until now, the current dynamic sign language recognition methods still have some drawbacks with difficulties of recognizing complex hand gestures, low recognition accuracy for most dynamic sign language recognition, and potential problems in larger video sequence data training. In order to solve these issues, this paper presents a multimodal dynamic sign language recognition method based on a deep 3-dimensional residual ConvNet and bi-directional LSTM networks, which is named as BLSTM-3D residual network (B3D ResNet). This method consists of three main parts. First, the hand object is localized in the video frames in order to reduce the time and space complexity of network calculation. Then, the B3D ResNet automatically extracts the spatiotemporal features from the video sequences and establishes an intermediate score corresponding to each action in the video sequence after feature analysis. Finally, by classifying the video sequences, the dynamic sign language is accurately identified. The experiment is conducted on test datasets, including DEVISIGN_D dataset and SLR_Dataset. The results show that the proposed method can obtain state-of-the-art recognition accuracy (89.8% on the DEVISIGN_D dataset and 86.9% on SLR_Dataset). In addition, the B3D ResNet can effectively recognize complex hand gestures through larger video sequence data, and obtain high recognition accuracy for 500 vocabularies from Chinese hand sign language.

INDEX TERMS Dynamic sign language recognition, bi-directional LSTM, residual ConvNet, video sequence.

I. INTRODUCTION

Sign language [1] recognition is an effective technology in the communication between deaf communities and hearing societies, and it tends to be a wildly discussed topic with the increasing development of the research on interaction between human and machines. In recent years, automatic sign language recognition system creates a new way of human-computer-interaction by transferring sign gestures into text/speech, and this technology can be implemented by computer-aid technology, like deep learning. Nowadays, there are already many successful applications in this area, such assign language translation, sign language tutor, and special education [2]–[4]. These can help the deaf people have a fluent communication with the others. In the other hand, sign language is constructed by a series of action, which involves quick motions with similar characteristics. Hence the static sign language recognition is hard to deal with the complexity and large variations of vocabulary set in hand actions. Besides, it may even make a misunderstanding of some significant variations from signers. Therefore, the research on dynamic sign language recognition is a more efficient method to solve the related problems. Vision-based dynamic gesture recognition technology has the

The associate editor coordinating the review of this manuscript and approving it for publication was Zhanyu Ma.

characteristics of flexibility, scalability and low cost, which is currently popular in the research of gesture interaction technology. However, dynamic sign language recognition also has challenges in dealing with the complexity of the sign activities of finger motion from the large scale body background. Another difficulty is that extracting the most discriminating features from images or videos. Besides, how to choose an appropriate classifier is also a critical factor for producing accurate recognition results.

In order to help the deaf-mute have a normal communication in daily life, an increasing number of researchers are devoted into improving the issues described above and many achievements have been received on dynamic sign language recognition already. There are several approaches to solve the problem in the dynamic sign language recognition, which can be mainly categorized into two types: one is the method based on hand shapes and motion trajectory of hand gestures, and the other is the method based on video sequence of each sign language.

In the traditional dynamic sign language recognition, the features of hand shapes and motion trajectory are utilized mostly to recognize hand gestures. But these features cannot fulfill the requirements of the practical dynamic sign language recognition systems completely, because of the aforementioned challenging factors. With the rapid development of the deep learning theory [5], data-driven methods [6] have demonstrated amazing performances in object detection [7] and gesture recognition [8]. Different from the method based on hand shapes and motion trajectory of hand gestures, sign language recognition based on video sequences could fully utilize the temporal information which plays a key role in the gesture recognition process. Hands have the relative small size compared with the whole scene, so the effective spatial features of gestures may be overwhelmed in backgrounds. Therefore, learning spatiotemporal feature simultaneously will be more informative for dynamic sign language recognition.

Comparing with the two methods introduced above, the method based on video sequence for dynamic sign language recognition is prior for its performance and practicability. According to that, this paper will present a new modified method, BLSTM-3D Residual Network (B3D ResNet), to instead of the original method based on video sequence, which can get an optimal result inspired by [9] and [10]. The new model B3D ResNet for dynamic sign language recognition is proposed to extract spatiotemporal features and analyze feature sequence of each hand gestures. At the beginning, the B3D ResNet model is used to perform object location using a deep learning method, i.e. Faster R-CNN, to detect hand and segment the hand position from the background. Then, theB3D ResNet model is used to extract the features of the inputted video sequence and analyze the feature sequence. Through classifying the inputted video sequence, the hand gestures could be identified, and the dynamic sign language recognition could be effectively achieved.

In summary, the major contributions of this paper are:

a) A new end-to-end trainable neural network B3D ResNet is proposed for dynamic sign language recognition, which is capable of solving the complex hand gesture classification task with lower error rates and distinguishing the tiny difference between similar sign gesture from different signers by fine-tuning the B3D ResNet model.
b) Compared with the current methods, the proposed B3D ResNet is capable of training on larger dataset, which consist of 500 daily vocabularies rather than simple number and alphabet dataset.
c) With a preprocessing network, our method could achieve an excellent performance in classifying dynamic hand gestures, which is verified in the experiment.

The rest of the paper is organized as follows. Firstly, Section II will show the related work. Next Section III will show the method overview. Then the detailed framework and experiment results will be discussed in Section IV and Section V respectively. Finally, there is a conclusion in Section VI to summarize the entire research.

II. RELATED WORK

In this section, the related work of dynamic sign language recognition will be reviewed from two aspects: method based on hand shapes and motion trajectory and method based on video sequence.

A. METHOD BASED ON HAND SHAPES AND MOTION TRAJECTORY

Dynamic sign language recognition method based on analyzing the features of hand shapes and motion trajectory with each hand gesture is a normal solution to the problems of dynamic sign language recognition. Primarily, work by some researchers is only based on analyzing the features of hand shapes. Kim *et al.* [11] explored the problem of finger-spelling recognition by deep neural network based on hand shape features. Literature [12] and [13] also focused on the sign language finger-spelling recognition. As for the method based on hand shapes, it could reflect the meaning of relatively simple hand gestures, like 31 alphabets and numbers, but this method was still restricted with the complex motion gestures for its limited consideration of the hand shape without the coherent hand motion. On the other hand, work by some researchers was only based on analyzing the motion trajectory of each hand gesture. Mohandes *et al.* [14] recognized hand gestures only based on hand motion trajectory with long short-term memory. With the help of sensor technology, such as leap motion controller [15], digital gloves data, surface electromyography accelerometer and gyroscope, [16]–[20] classified the hand gestures using the motion trajectory information obtained by sensors. However, these approaches are limited to some specific hand gestures, such as wave hand and gesticulate. Therefore, some works ware based on analyzing both the features of hand shapes and motion trajectory of each hand gesture. Many related researches ware accomplished successfully, for example, Kumar *et al.* [21] proposed a multimodal framework for isolated sign language recognition based on Kinect [22] sensor devices between single hand and double hand signs with a track model. Wang *et al.* [23] used a Sparse Observation description to indentify sign language based on hand postures and motions with RGB-D data. However, such sensor-based systems had the pitfalls in convenient practical situation and user comfort. Besides, these approaches would become quite complicated when the recognition object has lots of sign language vocabularies.

The dynamic sign language recognition based on analyzing the hand shapes features and motion trajectory as discussed above has some obvious shortcomings. On one hand, this method is effective for the sign language such as alphabet and numbers, but it is not suitable for the whole sign language vocabularies. On the other hand, this method is generally with the help of sensors. In this way, it may be inconvenient and uncomfortable for users. Besides, by combing hand shapes features and motion trajectory, the dynamic sign language recognition is also hard to obtain satisfying result because of the complex variable motion trajectory affected by interaction of hands and body joints.

B. METHOD BASED ON VIDEO SEQUENCE

More and more researchers have focused on the sign language recognition based on video sequence. For the method of dynamic sign language recognition based on video sequence, sensors are not necessary elements. The dynamic sign language recognition could be achieved by intelligent algorithm with analyzing video sequence features and classifying hand gestures. Cui *et al.* [24] designed a recurrent convolutional neural network for continuous sign language recognition fully based on video sequence. Huang *et al.* [25] proposed a framework of Hierarchical Attention Network with Latent Space for the generation of global-local video feature representations. Köpüklü *et al.* [26] added motion information into static images. But this approach was hard to have universal applicability, because some hand motion information could not be converted into static image. In [27], it presented a spotting-recognition framework for large-scale continuous gesture recognition. Camgoz *et al.* [28] used the sequence-to-sequence to learn the recognition problems of sign language. Kishore *et al.* [29] recognized continuous sign language by fuzzy classifying continuous sign language videos with the features combining tracking and shapes. Rao *et al.* [30] used neural network classifier for selfie continuous sign language recognition. This type of method mainly used deep learning technology to achieve the dynamic sign language recognition, such as the approaches of extracting the discriminative features of hand gestures with Convolutional Neural Networks (CNNs) in [31] and [32], the approaches of learning video sequence with Recurrent Neural Networks (RNNs) in [33], and the approaches of learning spatiotemporal sequence features combining CNNs and RNNs in [34] and [35]. These approaches performed well in dynamic sign language recognition based on video sequence compared with the method based on hand shapes and motion trajectory. So far, many researchers could be referred to the action recognition model or speech recognition model to recognize the hand gestures with video sequence, but majority ware not specially designed for the model of dynamic sign language recognition. The robustness and reliability ware poor in recognizing complex sign language which contains variable hand shapes and motion trajectory, as shown Fig.1 (a) and (b). Therefore, the method of complex sign language recognition had difficulty in tackling these complex and changeable information. While simple sign language, such as (c) and (d), only contained single hand shape and motion trajectory.

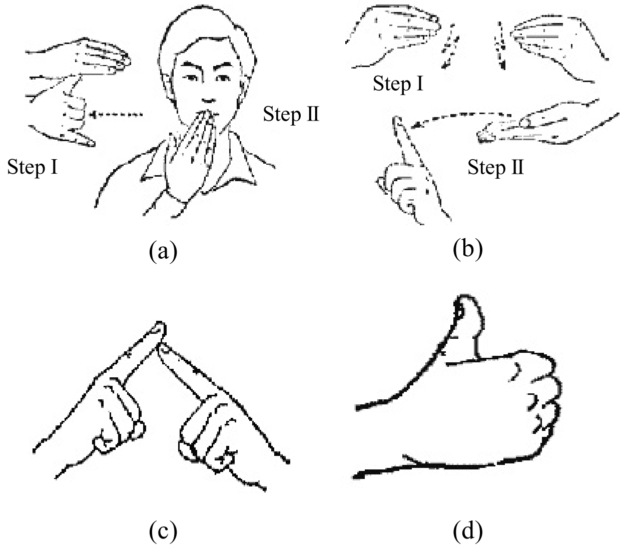


FIGURE 1. Four examples of sign language in DEVISIGN dataset. (a) Conceal. (b) Educate. (c) Person. (d) Good.

III. MATHOD OVERVIEW

Most of the current dynamic sign language recognition methods is based on the short-term spatiotemporal features and cannot recognize similar hand gestures, because the hand gesture is a long-term action. To solve this problem, a progressive approach to dynamic sign language recognition based on video sequence is proposed, whose framework is shown in Fig.2. This method is divided into three main parts. The first part is the object localization module based on Faster R-CNN, which is used to capture the information of hand position. The video frames are trained by Conv layers for feature extraction. The proposals are proposed by region proposal network. After that, proposals are mapped to the last layer of feature map. Then the fixed size of feature maps are generated by the ROI pooling layer. Through the classifier, the hand position could be detected correctly. With the accurate information of hand position, the hand region could be segmented from the background by coding algorithm. After that, the segmented video frames are inputted into the

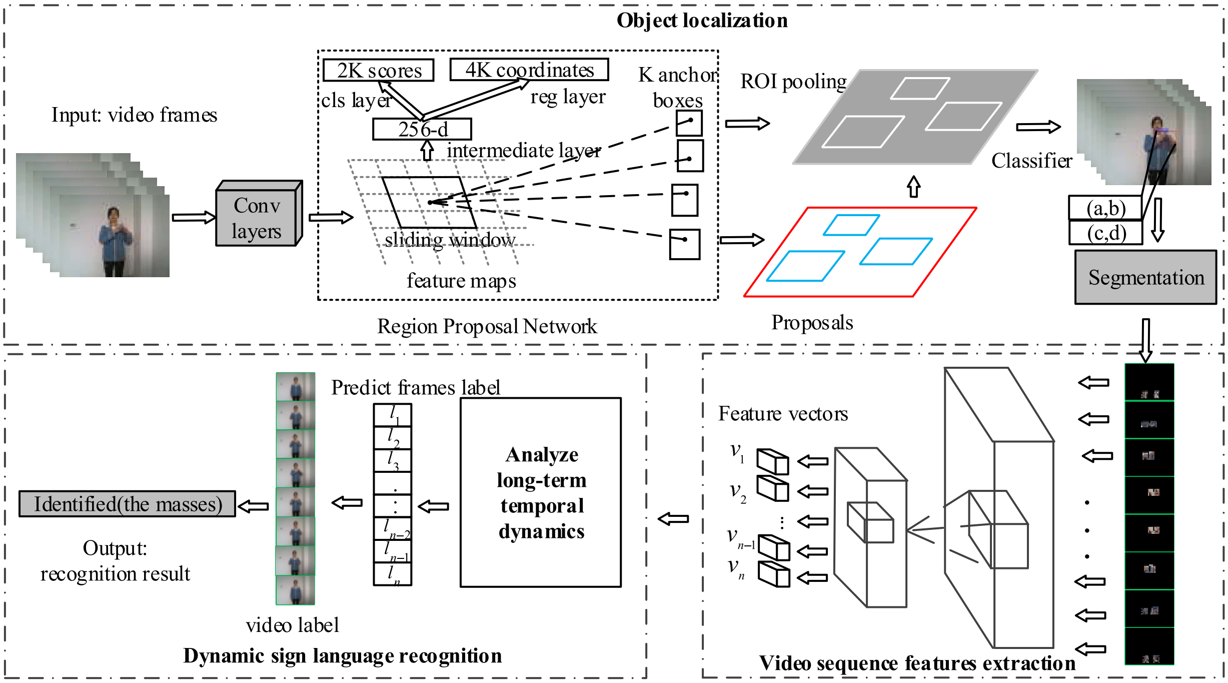


FIGURE 2. Overview of our proposed method for dynamic sign language recognition.

next part for long-term spatiotemporal features extraction. The second part is the video sequence features extraction module, which performs the task of long-term spatiotemporal features extraction with inputting segmented video frames. In this step, the networks are trained on chunks of full-time video, so the spatial context of the action performed can be preserved favorably. The feature vectors could be obtained by training full-time segmented videos on B3D ResNet model. Each video feature vector will be provided to the third part for analyzing dynamic information of sign language. Eventually, to create a joint representation for these independent streams, the third part is dynamic sign language recognition module, which can analyze long-term temporal dynamics and predict the hand gesture label. Through analyzing each video feature vectors, the frames label could be predicted. Thus, the video sequence label could be predicted. According to the top label prediction scores, this label will be regarded as the label of video sequence and be outputted as the recognition result. Therefore, the dynamic sign language could be recognized effectively.

IV. DYNAMIC SIGN LANGUAGE RECOGNITION WITH B3D RESNET

A. OBJECT LOCATION USING FASTER R-CNN

Hand detection is crucial for temporal segmentation and the subsequent recognition module. In order to obtain the accurate information of hand location in the frame images, it's essential to choose excellent object detection algorithm. Compared with SSD [37], YOLO [38] and other methods [39]–[41], Faster R-CNN [42] has higher accuracy and stronger robustness for smaller object detection. As shown in Fig.3, the object detection method of SSD and YOLO mainly consists of two modules, including detection generator and feature extractor. In addition, the object detection process is completed in a neural network. But for smaller targets, there's litter information left after multiple levels of convolution, which makes these models perform poorly in detecting small object targets. In this paper, the hand region is a smaller proportion in the whole video sequence. Therefore, the Faster R-CNN is more appropriate and chosen to detect the hand. The object detection method Faster R-CNN mainly consists of three modules, including proposal generator,

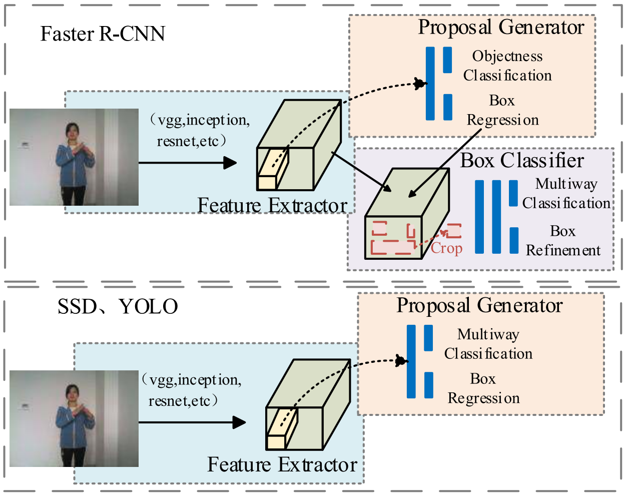


FIGURE 3. The difference of faster R-CNN, SDD and YOLO.

feature extractor and box classifier, which are finally unified into a depth network framework. All the calculations are completely run in the GPU, greatly improving the speed.

As shown in the object localization module from Fig. 2, when the video frames are inputted into the networks, the image features can be extracted by a fully convolutional network separately. These extracted feature maps are concatenated and considered as the final feature representation. Then region proposal network slides a small network over the*n-by-n* convlutional feature map, after that, a small sliding window is chosen as the input of the small network. Consequently, each sliding window is decreased to 256-dimension features contained in two sliding full-connected layers, i.e. a regression layer (*reg layer*) and a classification layer (*cls layer*). Each sliding window is predicted as multiple region proposals to find the number of maximum possible proposals for each location (denoted as k). Therefore, *4k* outputs encode the coordinates of k boxes in *reg layer*, and the *2k* outputs estimate probability of object for each proposal in *cls layer*. Generally, k proposals are relative to k reference boxes which are defined as anchors. An anchor is centered as the sliding window and connected with a scale aspect ratio as shown in the Region Proposal Network in Fig.2. RPN is used to generate high quality regions of interest (ROI). After that, the classification and bounding box regression will be done for each ROI. Finally, the non-maximum suppression is performed independently for each class. After accurate hand position is obtained, the hand can be segmented from background by the coordinate points of accurate hand position in the video frames.

TABLE 1. The result of hand detection.

Object	Real result	Prediction result		Recall	Precision	Accuracy
		Positive	Negative			
Hand	True	4443	101	99.84%	96.54%	96.48%
	False	159	7			

As shown in the Table 1, the Faster R-CNN has high accuracy for hand detection. This result is reflected in the followed parameters: $recall = \frac{TP}{TP+FN}$, $precision = \frac{TP}{TP+FP}$ and $accuracy = \frac{TP+TN}{TP+TN+FP+FN}$. Hence, using Faster R-CNN to detect hand, the precise information of position could be obtained for the segmentation from background in the frame images.

B. OUR METHOD-B3D RESNET MODEL

In this paper, the B3D ResNet model is proposed to recognize dynamic sign language based on video sequence. In details, this model can accomplish video sequence features extraction and learn long-term spatiotemporal features. For the dynamic sign language recognition, different dynamic sign language gestures generally correspond to different videos with different labels. Therefore the hand gesture could be recognized by classifying the labels. Through extracting spatiotemporal feature of different videos and classifying videos with different characteristics, this model can perfectly realize the various dynamic sign language including complex hand gestures. To improve the recognition accuracy of dynamic sign language, the feature sequences are further analyzed by the bi-directional long short-term memory (Bidirectional-LSTM) [43], [44] units. After that, the final features are classified by soft-max.

Besides, the dataset of dynamic sign language gestures based on video sequences is too large for the short training network. While the networks with residual connection are easier to optimize, concurrently, it could gain more accurate recognition accuracy when the network structure is deeper. So this structure is specially designed and covered in our B3D ResNet model. The B3D ResNet model is described as follows.

1) ARCHITECTURE OF B3D RESNET MODEL

The B3D ResNet model is proposed to capture the spatiotemporal features information for a video representation and analyze long-term temporal dynamic feature sequences. Fig. 4 shows the detailed structure of B3D ResNet, which mainly includes 17 convolutional layers, two Bidirectional-LSTM layers, one fully connected layer, and one soft-max layer. After the hand position has been localized in the input frames, the video sequences are transmitted into the B3D ResNet model for obtaining spatiotemporal features of the video sequence. With these processed short-term spatiotemporal features sequences, two Bidirectional-LSTM layers can analyze long-term temporal dynamic feature sequence. Consequently, an intermediate score is obtained corresponding to each action. Finally, the soft-max layer can classify the video sequence label and then recognize dynamic sign language gestures.

In the B3D ResNet for dynamic sign language recognition model, there are eight frames of size 112 × 112, centered on the current frame and used as inputs to the B3D ResNet model via three channels with 3D tensors of $L \times H \times W$, where L, H, and W are temporal length, height, and width respectively. Then, 3D convolutions are applied with a kernel size of $7 \times 7 \times 3$ (7×7 in the spatial dimension and 3 in the temporal dimension) on each of the three channels separately. The $2 \times 2 \times 1$ down-sampling is applied on each of the feature maps in the convolution layer, which leads to the same number of features maps with a reduced spatial resolution. The next convolution layers C2_x are obtained by applying 3D convolution with a kernel size of $3 \times 3 \times 3$ on each of three channels. The shortcut connections are inserted to the networks which could be turned into counterpart residual version. The identity shortcuts can be directly used when the input and output are with the same dimensions. When the shortcuts go across feature maps of two sizes, they are performed with a stride of 2. The next layers C3_x, C4_x and C5_x have the same application. All of these layers are obtained by

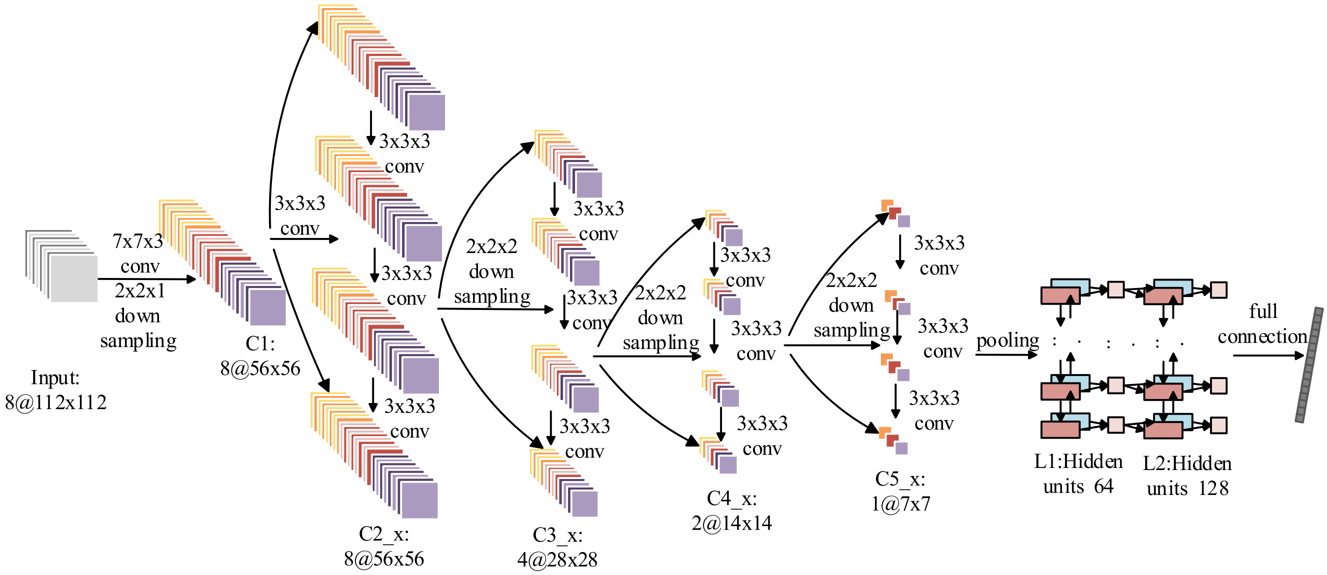


FIGURE 4. The B3D ResNet for dynamic sign language recognition.

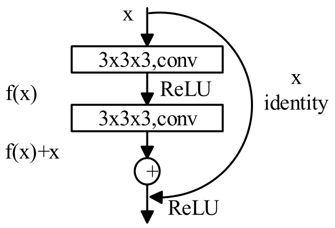


FIGURE 5. Building block of 3D residual unit.

applying 3D convolution with a kernel size of 3 × 3 × 3 on each of the three channels. The 2 × 2 × 2 down-sampling on each feature maps, which leads to the same number of feature maps with a reduced spatial resolution. After that, inserting shortcut connections between every two layers turns the network into its counterpart residual version with a stride of 2. Next the multiple layers of convolution and down-sampling, the eight input frames have been converted into a 1024D feature vector capturing the motion information in the input frames. Therefore, the B3D ResNet is provided with full-length video (arranged as a temporal sequence of 16-frame chunks), and the feature vectors are then fed into an LSTM network running in two directions. Each directional LSTM's hidden states layer, the fully connected layer and the soft-max layer are combined to obtain an intermediate score corresponding to each action. Finally, the scores for the two LSTMs are averaged to get action-specific scores.

2) SPATIOTEMPORAL FEATURE EXTRACTION OFB3D RESNET MODEL

The first step of the B3D ResNet model is extracting the features of input video sequence. As for image recognition problems, it is effective when utilizing the 3D convolution [45], [46] for capturing spatial and temporal dimension from videos. By the construction of 3D convolution, the feature maps in the convolution layer are connected to multiple contiguous frames in the previous layer, and then motion information is captured. The design principle of 3D CNNs is carried out with the 3D convolution kernel which can extract one type of features from the frame cube. At each feature map of any single layer, the value at position (a, b, c) is given by (1).

$$v^{abc} = \tanh\left(t\sum_{h=0}^{H-1}\sum_{w=0}^{W-1}\sum_{d=0}^{D-1} xy^{(a+h)(b+w)(c+d)} + z\right) \tag{1}$$

where $\tanh(\cdot)$ is the hyperbolic tangent function, and the index t and the value x are connection parameters of the current feature maps, H, W and D are the height, width and temporal dimension the 3D kernel, and z is the bias for the feature map.

However, deeper 3D CNNs are more effective and can be expected to facilitate progress in computer vision for videos [47]. In addition, adding residual connection [48] to 3D CNNs could simplify the training of deeper networks. Instead of directly learning unreferenced non-linear functions, our method utilizes the additive residual function with reference to the unit inputs, which is realized through a shortcut connection and helps to learn spatiotemporal features. Formally, denoting the desired underlying mapping as $h(x)$, we let the stacked nonlinear layers fit another mapping of $f(x) = h(x) - x$. The original mapping is recast into (2).

$$y = f(x, w) + x \tag{2}$$

where x is the value output from the previous layer of the neuron, w is the weight of the neuron; and y is the output value from the activation function within a neuron. This structure in the neural network is shown in Fig. 5. To develop 2D Residual unit into 3D architectures for encoding spatiotemporal video information, the basic Residual unit is modified following the principle of 3D convolution as introduced above, as shown in Fig. 5. The 3D convolutions are applied with a kernel size of 3×3×3 (3×3 in the spatial dimension and 3 in the temporal dimension) on each of the three channels separately.

For sign language recognition in videos, the B3D ResNet model could automatically extract spatiotemporal features from the input video sequence by applying residual connections to the 3D CNNs. Along with these spatial features, it is desirable to capture the motion information encoded in multiple contiguous frames. Besides, it is easy for the B3D ResNet model to train on larger video sequence data set.

3) LEARN SPATIOTEMPORAL SEQUENCE FEATURE OF B3D RESNET MODEL

With the above steps, the short-term spatiotemporal features for specific length are extracted. As for an entire video including several spatiotemporal features, the long-term spatiotemporal features of the video also need to be obtained. Hence, the B3D ResNet model utilizes Bi-directional LSTM unit, which contains six shared weights $w_1 - w_6$ and integrates information from the future as well as the past to make a prediction for each chunk in the video sequence, illustrated in Fig.6. In the Bi-directional LSTM unit, the forward layer and the backward layer are connected to the output layer. Moreover, LSTMs are an important part of deep learning models to analyze long-term temporal dynamics for human gesture recognition [49]–[51]. Conceptually, the memory cell stores the past contexts, the input and output gates allow the cell to store contexts for a long period of time. Meanwhile, the memory in the cell can be cleared by the forget gate. The special design of LSTM allows it to capture long-range dependencies, which often occurs in image-based sequences. Therefore, they are expected to have a prior performance at predicting the temporal boundaries of an action compared to a unidirectional LSTM network.

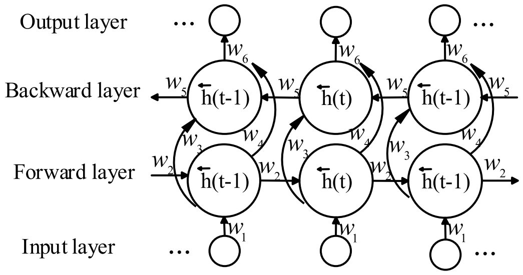


FIGURE 6. **The structure of deep bidirectional LSTM cell used in this paper.**

Formally, the following formulas are given, including an input sequence $x = \{x_1, x_2, \ldots, x_t\}$, the cell states $c = \{c_1, c_2, \ldots, c_t\}$ and the hidden states $h = \{h_1, h_2, \ldots, h_t\}$. The gates i_t, f_t, o_t and c_t are the input gate, forget gate,output gate, and memory cell activation vectors respectively. The equations for a Bi-directional LSTM cell are as follows:

$$i_t = \sigma \left(w_{xi} x_t + w_{hi} h_{t-1} + b_i \right) \tag{3}$$

$$f_t = \sigma \left(w_{xf} x_t + w_{hf} h_{t-1} + b_f \right) \tag{4}$$

$$o_t = \sigma \left(w_{xo} x_t + w_{ho} h_{t-1} + b_o \right) \tag{5}$$

$$g_t = \tanh \left(w_{xc} x_t + w_{hc} h_{t-1} + b_c \right) \tag{6}$$

$$c_t = f_t c_{t-1} + i_t g_t \tag{7}$$

$$h_t = o_t \tanh \left(c_t \right) \tag{8}$$

where sigmoid σ is a function, and tanh is the hyperbolic tangent function. The forget gate f_t decides when information should be cleared from the memory cell c_t. The input gate i_t decides when new formation should be incorporated into the memory. The tanh layer g_t generates a candidate set of values which will be added to the memory cell if the input gate allows it. Referred to (7), based on the output of the forget gate f_t, input gate i_t and the new candidate values g_t, the memory cell c_t is updated. In (8), the output gate o_t controls the status and memory information of the hidden state. Finally, the hidden state is represented as a product between a function of the memory cell state and the output gate.

Therefore, the B3D ResNet model could obtain the full features of the input video. For dynamic sign language recognition, the B3D ResNet model has a strong capability of capturing contextual information within a sequence. Using contextual cues for image-based sequence recognition is more stable and helpful than treating each symbol independently. At the same time, error differentials could be back-propagated to its input. Besides, the B3D ResNet model is able to operate on sequences with arbitrary lengths, traversing from starts to ends.

V. EXPERIMENTS

A. DATASETS

In order to demonstrate that the proposed network can effectively recognize dynamic sign language, two testing datasets including DEVISIGN-D dataset [52] and SLR_Dataset are chosen for the contrast test. Fig.7 shows some examples of the two datasets. More details regarding the two datasets are given below.

1) DEVISIGN-D DATASET

DEVISIGN-D is a chinese sign language dataset, which provides the worldwide researchers of sign language recognition community with a large vocabulary Chinese sign language dataset for training and evaluating their algorithms. It composed of 500 daily vocabularies. the data covers 8 different signers. Among them, the vocabularies are recorded twice for 4 signers (2 males and 2 females) and once for the other 4 signers (2 males and 2 females). It totally includes 6000 videos.

2) SIGN LANGUAGE RECOGNITION DATASET (SLR_DATASET)

The SLR_Dataset is collected by Huang *et al.* [25] and released on their project web page (http://mccipc.ustc.edu.cn/mediawiki/index.php/SLR_Dataset).A Microsoft Kinect

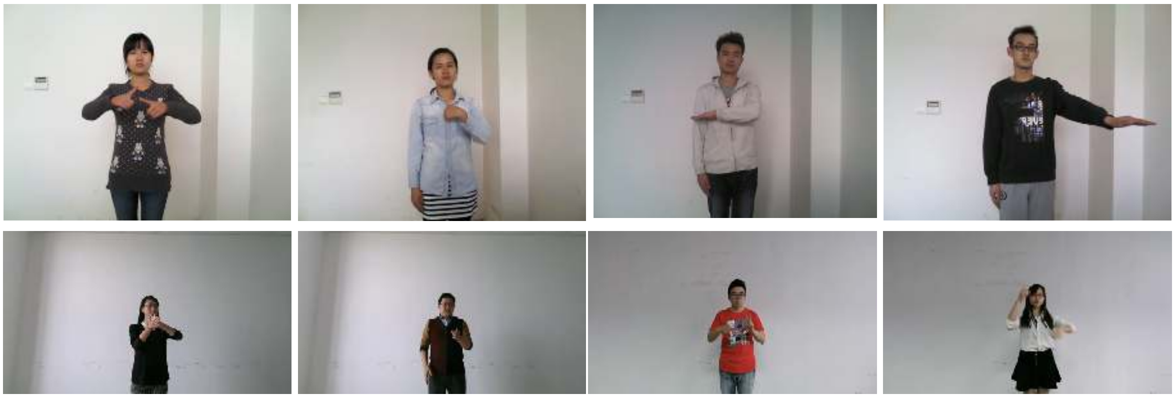


FIGURE 7. Some examples of the DEVISIGN-D dataset (the first line) and SLR_Dataset (the second line).

TABLE 2. Experimental configurations.

Parameter Setting	
basic learning rate	0.1
learning rate policy	step
gamma	0.1
momentum	0.9
weight decay	0.00005
batch size	2
Platform and Device	
Platform	Caffe
GPU	Quadro P4000

camera is used for recording video, and provides RGB, depth and body joints modalities in all videos. In this paper, only the RGB modality is used. The SLR_Dataset contains 25K labeled video instances, with100+ hours of total video footage by 50 signers and every video instance is annotated by a professional Chinese sign language teacher.

B. NETWORK TRAINING

The B3D ResNet model is implemented based on the deep learning platform Caffe [53], the GPU is Quadro P4000 with 8 GB of memory. When training the model, the batch size is set to two. The layers have a basic learning rate of 0.1 and a momentum of 0.9. The learning rate policy is the way of step by 10k size, and the decay parameter of 0.00005 is used. In order to exhibit the experimental configurations more clearly, more experimental configurations could be seen in table 2. Our datasets are not very large, so we take the following effective strategies to avoid over-fitting. One well known method is called data augmentation, in which video frames are randomly cropped to 112 × 112. Another strategy is batch normalization [54], which aims to reduce internal covariate shift and is applied in all convolution layers to dramatically accelerate the training process of deep neural nets. Finally, the learning rate is set to 0.1 and divided by 10 after every 10k of iterations. Training process is well done after 90k iterations.

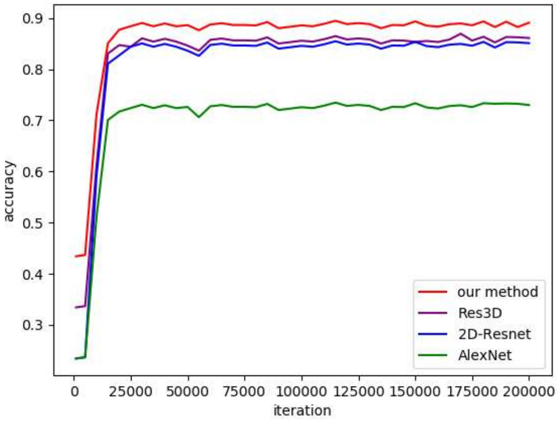


FIGURE 8. The comparison of gesture recognition results.

After the experimental parameters are set, the B3D ResNet model is training for the dynamic sign language recognition, which mainly can extract spatiotemporal features from the input videos for analyzing long-term temporal dynamics to predict the hand gesture label. To evaluate the performance of the B3D ResNet model for dynamic sign language recognition, the recognition accuracy is employed as the criterion. Here we compared the proposed method, the B3D ResNet model, with some traditional models for sequence action recognition, such as Res3D [9], 2D-Resnet [48], and AleXnet [31], based on DEVISIGN-D dataset. The comparison of dynamic sign language recognition results are shown in Fig. 8. When training these models until about 20k iteration, recognition accuracy begins to reach the maximum. The results shows that the accuracy of Res3D, 2D-Resnet and AleXnet are 86.6%, 85%, and 73.8% respectively, while the accuracy of our method is 89.9%, which outperforms the other method by 3.3% at least. Hence, the experiment demonstrates that the B3D ResNet model could effectively achieve the task of dynamic sign language recognition.

C. EVALUATION ONOBJECT LOCALIZATION

Localizing the hand is indispensable, because the hand always occupies only a limited space and leaves a large

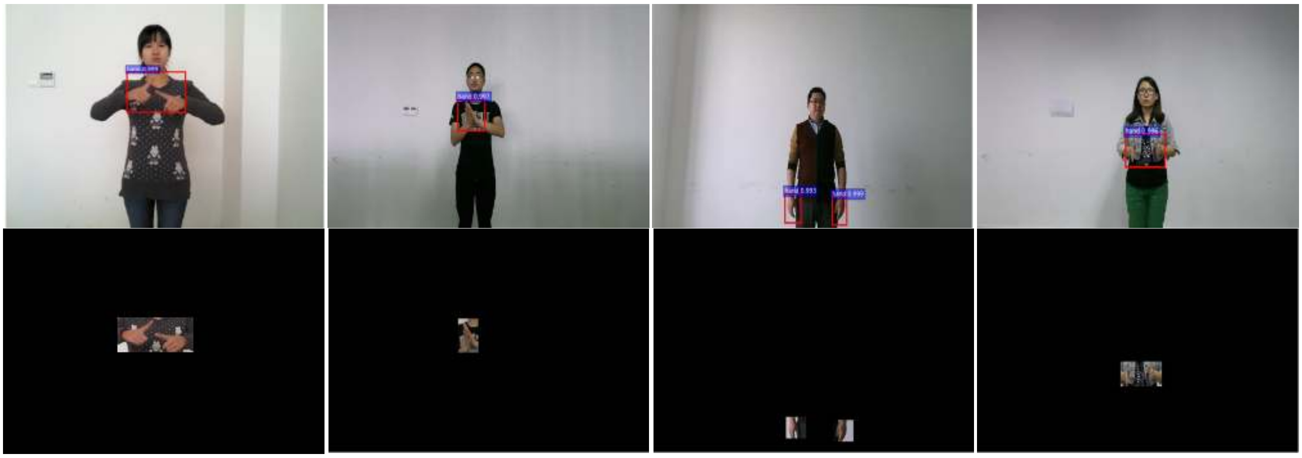


FIGURE 9. Examples of hand detection and segmentation results, there are two rows in total, which represent detection and segmentation results. Red rectangles indicate the hand position by Faster R-CNN at high detection accuracy.

TABLE 3. Results on object localization.

Training strategy	Data set	Accuracy
Strategy 1	DEVISIGN-D	89.8%
	SLR_Dataset	86.9%
Strategy 2	DEVISIGN-D	43.7%
	SLR_Dataset	50.2%

background area within the image. Therefore, with object localization step, it would reduce the calculated amount of the B3D ResNet model to improve the recognition accuracy. In the object localization module, the Faster R-CNN model is used to detect the object of hand. After that the information of hand position is outputted as a result. Next the hand could be segmented from the background in the input frames with these accurate hand positions. In order to exhibit the object localization module, Fig. 9 shows some examples of the experimental results. In order to verify this approach, the object localization module is evaluated on DEVISIGN-D dataset and SLR_Dataset with two different training strategies:

Strategy 1 with object localization module: At the beginning, the video frames are inputted into the object localization module. After the hand is segmented from the background, the preprocessing video sequences will be trained based on the B3D ResNet model.

Strategy 2 without object localization module: The video frames are directly trained based on the B3D ResNet model. As illustrated in Table 3. The results show that Strategy 1 with object localization module has better recognition accuracy than Strategy 2 without object localization module.

D. EVALUATION ON DEVISIGN-D AND SLR_DATASET

The performances of the B3D ResNet model training on the DEVISIGN-D dataset and SLR_Dataset are shown in Table 4. It can be seen that our model outperforms the previous published method for dynamic sign language recognition.

TABLE 4. Comparison with other methods on DEVISIGN-D and SLR_Dataset.

Methods	Accuracy	
	DEVISIGN-D	SLR_Dataset
BLSTM-NN [21]	60.3%	56.6%
HMM-DTC [23]	74.4%	65.2%
DNN [11]	70.8%	65.8%
C3D [19]	78.3%	73.5%
B3D ResNet (ours)	89.8%	86.9%

To be emphasized, all the compared models are trained on the same datasets. As can be seen from the above table data, the SLR_Dataset is very challenging, due to the intricacy of its collection. It is gratifying that our method outperforms these methods by a large margin. Concretely, in the DEVISIGN-D dataset and SLR_Dataset, the results of our method are 89.8% and 86.9% separately, which are separately 29.5% and 30.3% higher than BLSTM-NN [21], 25.4% and 21.7% higher than HMM-DTC [23], 19% and 21.1% higher than DNN [11], and 11.5% and 13.4% higher than C3D [19]. The comparison results show that the proposed method obtains the state-of-the-art recognition accuracy of dynamic sign language on both of the two testing datasets.

VI. CONCLUSION

In this paper, we propose a new model for the task of dynamic sign language recognition based on 3D Residual ConvNet and Bi-directional LSTM networks. By analyzing the video sequence, the proposed model could effectively recognize different hand gestures with extracting video spatiotemporal features and analyzing features sequence. With our model, it could get a good performance on complex or similar sign language recognition. In the experimental result on DEVISIGN_D dataset and SLR_dataset, different sign language could be accurately and effectively distinguished, even some quite similar pairs of hand gestures. Moreover, the evaluation results demonstrate that analyzing spatiotemporal features in our proposed method is more effective for dynamic sign language recognition. The experimental results demonstrate that our proposed method improves accuracy and overall performance in dynamic sign language recognition.

In future work, the sign sentence recognition for real world video in complex environment will be considered for further research, and the evaluation of experiment should be considered with more characteristics, like robustness, sensitivity and so on.

VII. ACKNOWLEDGMENT

(Yanqiu Liao and Pengwen Xiong contributed equally to this work.) The authors thank VIPL (Visual Information Processing and Learning, ICT) for releasing the DEVISIGN database.

REFERENCES

[1] U. Zeshan, "Sign languages of the world," in *Encyclopedia Language & Linguistics*, 2nd ed. K. Brown, Ed. Amsterdam, The Netherlands: Elsevier, 2006, pp. 358–365.

[2] S. C. W. Ong and S. Ranganath, "Automatic sign language analysis: A survey and the future beyond lexical meaning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 6, pp. 873–891, Jun. 2005.

[3] Z. Zafrulla, H. Brashear, P. Yin, P. Presti, T. Starner, and H. Hamilton, "American sign language phrase verification in an educational game for deaf children," in *Proc. IEEE 20th Int. Conf. Pattern Recognit. (ICPR)*, Istanbul, Turkey, Aug. 2010, pp. 3846–3849.

[4] Z. Ma, H. Yu, W. Chen, and J. Guo, "Short utterance based speech language identification in intelligent vehicles with time-scale modifications and deep bottleneck features," *IEEE Trans. Veh. Technol.*, vol. 68, no. 1, pp. 121–128, Jan. 2019.

[5] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[6] J. Han *et al.*, "Representing and retrieving video shots in human-centric brain imaging space," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2723–2736, Jul. 2013. doi: 10.1109/TIP.2013.2256919.

[7] D. Zhang, D. Meng, and J. Han, "Co-saliency detection via a self-paced multiple-instance learning framework," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 5, pp. 865–878, May 2017. doi: 10.1109/TPAMI.2016.2567393.

[8] M. Asadi-Aghbolaghi *et al.*, "A survey on deep learning based approaches for action and gesture recognition in image sequences," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Washington, DC, USA, May 2017, pp. 476–483.

[9] D. Tran, J. Ray, Z. Shou, S.-F. Chang, and M. Paluri. (2017). "Convnet architecture search for spatiotemporal feature learning." [Online]. Available: https://arxiv.org/abs/1708.05038

[10] B. Singh, T. K. Marks, M. Jones, O. Tuzel, and M. Shao, "A multi-stream bi-directional recurrent neural network for fine-grained action detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 1961–1970.

[11] T. Kim *et al.*, "Lexicon-free fingerspelling recognition from video: Data, models, and signer adaptation," *Comput. Speech Lang.*, vol. 46, pp. 209–232, Nov. 2017.

[12] B. Kang, S. Tripathi, and T. Q. Nguyen, "Real-time sign language fingerspelling recognition using convolutional neural networks from depth map," in *Proc. 3rd IAPR Asian Conf. Pattern Recognit. (ACPR)*, Kuala Lumpur, Malaysia, Nov. 2015, pp. 136–140.

[13] S. Aly, B. Osman, W. Aly, and M. Saber, "Arabic sign language fingerspelling recognition from depth and intensity images," in *Proc. 12th Int. Conf. Comput. Engineer. (ICENCO)*, Cairo, Egypt, Dec. 2016, pp. 99–104.

[14] M. Mohandes, M. Deriche, and J. Liu, "Image-based and sensor-based approaches to arabic sign language recognition," *IEEE Trans. Human-Mach. Syst.*, vol. 44, no. 4, pp. 551–557, Aug. 2014.

[15] C.-H. Chuan, E. Regina, and C. Guardino, "American sign language recognition using leap motion sensor," in *Proc. 13th Int. Conf. Mach. Learn. Appl. (ICMLA)*, Detroit, MI, USA, Dec. 2014, pp. 541–544.

[16] T. Sonawane, R. Lavhate, P. Pandav, and D. Rathod, "Sign language recognition using leap motion controller," *Internation J. Advance Res. Innov. Ideas Edu.*, vol. 3, no. 2, pp. 1878–1883, Mar. 2017.

[17] K. Li, Z. Zhou, and C.-H. Lee, "Sign transition modeling and a scalable solution to continuous sign language recognition for real-world applications," *ACM Trans. Accessible Comput.*, vol. 8, no. 2, Jan. 2016, Art. no. 7.

[18] X. Yang, X. Chen, X. Cao, S. Wei, and X. Zhang, "Chinese sign language recognition based on an optimized tree-structure framework," *IEEE J. Biomed. Health Inform.*, vol. 21, no. 4, pp. 994–1004, Jul. 2017.

[19] T. Liu, W. Zhou, and H. Li, "Sign language recognition with long short-term memory," in *Proc. IEEE Int. Conf. Image Process.(ICIP)*, Phoenix, AZ, USA, Sep. 2016, pp. 2871–2875.

[20] Z. Ma, Y. Lai, W. B. Kleijn, Y.-Z. Song, L. Wang, and J. Guo, "Variational Bayesian learning for Dirichlet process mixture of inverted Dirichlet distributions in non-Gaussian image feature modeling," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 2, pp. 449–463, Feb. 2019.

[21] P. Kumar, H. Gauba, P. P. Roy, and D. P. Dogra, "A multimodal framework for sensor based sign language recognition," *Neurocomputing*, vol. 259, pp. 21–38, Oct. 2017.

[22] K. Lai, J. Konrad, and P. Ishwar, "A gesture-driven computer interface using Kinect," in *Proc. IEEE Southwest Symp. Image Anal. Interpretation (SSIAI)*, Santa Fe, NM, USA, Apr. 2012, pp. 185–188.

[23] H. Wang, X. Chai, and X. Chen, "Sparse observation (so) alignment for sign language recognition," *Neurocomputing*, vol. 175, no. 29, pp. 674–685, Jan. 2016.

[24] R. Cui, H. Liu, and C. Zhang, "Recurrent convolutional neural networks for continuous sign language recognition by staged optimization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 7361–7369.

[25] J. Huang, W. Zhou, Q. Zhang, H. Li, and W. Li, "Video-based sign language recognition without temporal segmentation," in *Proc. 32nd AAAI Conf. Artif. Intell. (AAAI-18)*, New Orleans, LA, USA, Feb. 2018, pp. 1–8. [Online]. Available: https://arxiv.org/abs/1801.10111

[26] O. Köpüklü, N. Köse, and G. Rigoll, "Motion fused frames: Data level fusion strategy for hand gesture recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Salt Lake City, UT, USA, 2018, pp. 2184–21848, doi: 10.1109/CVPRW.2018.00284.

[27] Z. Liu, X. Chai, Z. Liu, and X. Chen, "Continuous gesture recognition with hand-oriented spatiotemporal feature," in *Proc. IEEE Conf. Comput. Vis. Workshops*, Venice, Italy, Oct. 2017, pp. 3056–3064.

[28] N. C. Camgoz, S. Hadfield, O. Koller, and R. Bowden, "SubUNets: End-to-end hand shape and continuous sign language recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 3075–3084.

[29] P. V. V. Kishore, D. A. Kumar, E. N. D. Goutham, and M. Manikanta, "Continuous sign language recognition from tracking and shape features using fuzzy inference engine," in *Proc. Int. Conf. Wireless Commun., Signal Process. Netw. (WiSPNET)*, Chennai, India, Mar. 2016, pp. 2165–2170.

[30] G. A. Rao, P. V. V. Kishore, A. S. C. S. Sastry, D. A. Kumar, and K. Kumar, "Selfie continuous sign language recognition with neural network classifier," in *Proc. 2nd Int. Conf. Micro-Electron., Electromagn. Telecommun.* Singapore: Springer, 2018, pp. 31–40.

[31] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. 25th Int. Conf. Neural Inform. Process. Syst.*, Lake Tahoe, NV, USA, Dec. 2012, pp. 1097–1105.

[32] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 677–691, Apr. 2017.

[33] N. Srivastava, E. Mansimov, and R. Salakhudinov, "Unsupervised learning of video representations using lstms," in *Proc. 32th Int. Conf. Mach. Learn.*, Lille, France, Jul.2015, pp. 843–852.

[34] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Sequential deep learning for human action recognition," in *Proc. Int. Workshop Hum. Behav. Understand.*, Berlin, Germany, Mar. 2011, pp. 29–39.

[35] J. Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. G. Monga, "Beyond short snippets: Deep networks for video classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 4694–4702.

[36] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1110–1118.

[37] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, Oct. 2016, pp. 21–37.

[38] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 779–788, doi: 10.1109/CVPR.2016.91.

[39] J. Han, D. Zhang, X. Hu, L. Guo, J. Ren, and F. Wu, "Background prior-based salient object detection via deep reconstruction residual," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 8, pp. 1309–1321, Aug. 2015. doi: 10.1109/TCSVT.2014.2381471.

[40] Z. Ma, J.-H. Xue, A. Leijon, Z.-H. Tan, Z. Yang, and J. Guo, "Decorrelation of neutral vector variables: Theory and applications," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 1, pp. 129–143, Jan. 2018.

[41] T. Zhang *et al.*, "Predicting functional cortical ROIs via DTI-derived fiber shape models," *Cerebral Cortex*, vol. 22, no. 4, pp. 854–864, Apr. 2011.

[42] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[43] A. Graves, S. Fernández, and J. Schmidhuber, "Bidirectional LSTM networks for improved phoneme classification and recognition," in *Proc. Int. Conf. Artif. Neural Netw.*, Warsaw, Poland. Berlin, Germany: Springer, Sep. 2005, pp. 799–804.

[44] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[45] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[46] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 6546–6555.

[47] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)]*, Santiago, Chile, Dec. 2015, pp. 4489–4497.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[49] Z. Ma, A. E. Teschendorff, A. Leijon, Y. Qiao, H. Zhang, and J. Guo, "Variational Bayesian matrix factorization for bounded support data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 4, pp. 876–889, Apr. 2015.

[50] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, Mar. 1994.

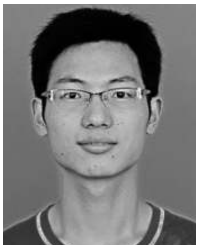
[51] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, "Learning precise timing with LSTM recurrent networks," *J. Mach. Learn. Res.*, vol. 3, pp. 115–143, Mar. 2003.

[52] H. Wang, X. Chai, X. Hong, G. Zhao, and X. Chen, "Isolated sign language recognition with grassmann covariance matrices," *ACM Trans. Accessible Comput.*, vol. 8, no. 4, May 2016, Art. no. 4.

[53] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22th ACM Int. Conf. Multimedia*, Orlando, FL, USA, Nov. 2014, pp. 675–678.

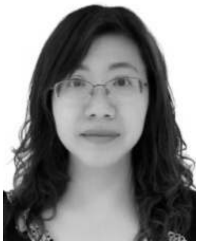
[54] S. Ioffe and C. Szegedy. (2015). "Batch normalization: Accelerating deep network training by reducing internal covariate shift." [Online]. Available: https://arxiv.org/abs/1502.03167

YANQIU LIAO received the B.E. degree in communication engineering from Nanchang University, China, in 2016, where she is currently pursuing the master's degree in behavior detection and computer vision.

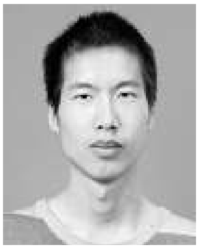


PENGWEN XIONG received the B.S. degree from the North University of China, in 2009, and the Ph.D. degree in instrument science and technology from Southeast University, China, in 2015. He visited the Laboratory for Computational Sensing and Robotics, Johns Hopkins University, from 2013 to 2014. He is currently an Associate Professor with the School of Information Engineering, Nanchang University. His research interests include artificial intelligence and robotic sensing and controlling.

WEIDONG MIN (M'12) received the B.E., M.E., and Ph.D. degrees in computer application from Tsinghua University, China, in 1989, 1991, and 1995, respectively, where he was an Assistant Professor, from 1994 to 1995. From 1995 to 1997, he was a Postdoctoral Researcher with the University of Alberta, Canada. From 1998 to 2014, he was a Senior Researcher and a Senior Project Manager with Corel and other companies in Canada. From 2011 to 2014, he cooperated with the School of Computer Science and Software Engineering, Tianjin Polytechnic University, China. Since 2015, he has been a Professor with Nanchang University, China, where he is currently a Professor and the Dean of the School of Software. He is also an Executive Director of the China Society of Image and Graphics. His current research interests include image and video processing, artificial intelligence, big data, distributed systems, and smart city information technology.



WEIQIONG MIN received the B.E. degree in international culture and the M.E. degree in international trade from Kyushu University, Japan, in 2002 and 2006, respectively. She is currently an Assistant Professor with the School of Tourism, Jiangxi Science and Technology Normal University, China. Her research interests include image and video processing, and hi-tech applications in tourism.



JIAHAO LU received the B.E. degree in civil engineering from the Zhejiang University of Science and Technology, China, in 2017. He is currently pursuing the master's degree in object detection with Nanchang University, China.