IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 28, 2020, accepted November 24, 2020, date of publication November 27, 2020,
date of current version December 11, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3041115

Attention-Based Sign Language Recognition Network Utilizing Keyframe Sampling and Skeletal Features

WEI PAN, XIONGQUAN ZHANG, AND ZHONGFU YE
Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei 230027, China
National Engineering Laboratory for Speech and Language Processing, Hefei 230027, China

Corresponding author: Zhongfu Ye (yezf@ustc.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61671418, and in part by the Advanced Research Fund of the University of Science and Technology of China.

ABSTRACT Sign language recognition(SLR) is a multidisciplinary research topic in pattern recognition and computer vision. Due to large amount of data from the continuous frames of sign language videos, selecting representative data to eliminate irrelevant information has always been a challenging problem in data preprocessing of sign language samples. In recent years, skeletal data emerged as a new type of data but received insufficient attention. Meanwhile, due to the increasing diversity of sign language features, making full use of them has also been an important research topic. In this paper, we improve keyframe-centered clips (KCC) sampling to get a new kind of sampling method called optimized keyframe-centered clips (OptimKCC) sampling to select key actions from sign language videos. Besides, we design a new kind of skeletal feature called Multi-Plane Vector Relation (MPVR) to describe the video samples. Finally, combined with the attention mechanism, we also use Attention-Based networks to distribute weights to the temporal features and the spatial features extracted from skeletal data. We implement comparison experiments on our own and the public sign language dataset under the Signer-Independent and the Signer-Dependent circumstances to show the advantages of our methods.

INDEX TERMS Sign language recognition, keyframe sampling, skeletal features, attention-based BLSTM.

I. INTRODUCTION

Sign language is an effective way for hearing impaired people to convey their ideas to others. Sign language recognition(SLR) provides good communication media between deaf-mute and ordinary people, which has important application value [1]. The research of SLR is mainly divided into isolated SLR and continuous SLR. The former aims at recognizing word by word, the other focuses on translating sentences from a sequence of actions. We will discuss the methodology of isolated SLR in this paper.

The traditional equipment for acquiring sign language data can be divided into data gloves and visual image systems [2]. The former use gesture motion sensors to get sequential trajectory data [3], [4]. This method can have a high identification rate but bring so much labor and financial consumption. In contrast, a visual image system uses cameras to collect information [5]. However, the identification rate is low, and real-time performance is poor, especially it is not able to collect large sign language datasets. After Microsoft launched Kinect-2.0, we can get the RGB, depth, and skeletal data simultaneously from a frame sampled from the video [6]. Using Kinect-2.0, we recorded a large vocabulary Chinese Sign Language(CSL) dataset with 200 sign language words. Each word contains 100 samples by 10 signers who repeated the same sign language word 10 times. The CSL dataset will be used for the experiments in our paper.

Sign language videos are composed of continuous frames sampled at a specific sampling rate by cameras. Due to the high sampling rate, the number of frames in a sign language video is large, which causes much data storage memory and redundant information between adjacent frames. Therefore, we need to select keyframes from the whole sign language video as the descriptor of it. In this way, the redundant information is eliminated, and data storage memory is greatly reduced, making the feature extraction

The associate editor coordinating the review of this manuscript and approving it for publication was Thomas Canhao Xu.

more convenient without negatively affecting recognition performance. In this paper, we design a method called optimized keyframe-centered clips(Optim KCC) sampling and received better results compared with the state-of-art method in [7].

In recent years, skeletal coordinate data emerged as a new type of data in SLR [8], [9]. It can eliminate the influence from the background and the illumination of the signing environment. and describe the three-dimensional spatial trajectories of finger joints. Besides, the storage memory of skeletal data is much less than that of RGB images and depth images. However, to our knowledge, the number of literatures researching on extracting skeletal features is limited. Thus, making full use of skeletal data, we design a kind of feature called Multi-Plane Vector Relation(MPVR) in this paper to get better descriptions of sign language videos.

With the development of the attention mechanism, researchers proposed Attention-Based networks for SLR, which distribute corresponding weights to different keyframes' features to achieve more efficient feature extraction. However, current research mainly focuses on the temporal attention mechanism. In this paper, we also consider the spatial attention mechanism. In the 3D skeletal data provided by Kinect-2.0, the XY plane represents the screen of the camera facing the signers, the YZ plane represents the ground, and the XZ plane represents the sidewall orthogonal to the previous two subplanes. The skeletal joint trajectories' projection onto the three subplanes has different importance in feature representation. Therefore, we design the spatial Attention-Based BLSTM referencing the Attention-Based network proposed in [10] to weight different subplanes' features.

The contribution of this paper can be summarized as follows:

- Based on the keyframe-centered clips(KCC) sampling proposed in [7], we improve it for better data preprocessing and feature descriptions.
- In this paper, a new kind of skeletal feature called Multi-Plane Vector Relation(MPVR) is proposed. We project each skeletal joint's 3D coordinate data to 3 subplanes to get 3 2D vectors. And then, we explore the vector relation in different subplanes, which is the principal component of the MPVR feature.
- Based on the Attention-Based network proposed in [10], we design a spatial Attention-Based BLSTM to distribute weights to corresponding subplanes' features in MPVR.
- According to the ideas proposed above, we implement the comparison experiments under the Signer-Independent and the Signer-Dependent circumstances distinguished by whether there are signers who appear in both training sets and test sets. The recognition accuracy under two cases can validate the adaptiveness of our networks to the practical Signer-Independent situation.

II. RELATED WORKS

In this section, we will review the work related to our research in this paper.

A. KEY FRAME SAMPLING

Huang *et al.* [7] proposed keyframe-centered clips(KCC) sampling, which aims at selecting a certain number of frames to describe the whole sign language videos. He got better recognition performance compared with other sampling methods. The keyframe extraction algorithm in this paper is based on [7].

B. FEATURE EXTRACTION

In the field of SLR, the initial research on feature extraction focused on extracting features such as HOG, LBP, optical flow, or SIFT [11] from RGB images and depth images using traditional image-processing algorithms [12]–[14]. With the development of deep learning, CNN [15]–[17] and RNN [18]–[21] can directly extract the temporal features or the spatial features from image data, which gradually made themselves become the mainstream research methods in SLR.

After Microsoft launched Kinect-2.0, skeletal data gradually received attention. In recent years, a few literatures began to research on extracting skeletal features and have made some progress. Kumar *et al.* [22] proposed joint distance and angular coded Color topographical descriptor(JTDT) and got 84.12% accuracy on the Indian sign language dataset. Rastgoo *et al.* [23] proposed the multi-view hand skeleton, which obtained skeletal coordinate information from multiple perspectives and achieved 99.6% accuracy on his own laboratory's dataset. The above works only stay in the use of rectangular coordinate data. In consideration of this, MPVR is designed by us, which uses polar coordinate data to describe vector relation in different subplanes.

C. NETWORK LEARNING

The types of SLR networks are closely related to the development of computer vision and pattern recognition. HMM is one of the classical models [24], [25]. HMM can model continuous frames in the time domain and extract temporal features. Based on traditional HMM, Zhang *et al.* [26] and Guo *et al.* [27] proposed adaptive HMM. Pu *et al.* [28] applied HMM to trajectory modeling. In addition, SVM [12], [18], [29], CRF [30], [31], and some of their variants have also been used in SLR.

With the development of deep learning, SLR gradually relies on neural networks [19]. Zamora-Mora and Chacn-Rivas [32] used CNN for real-time hand detection as the tool of SLR. Al-Hammadi *et al.* [33] used 3DCNN to extract temporal and spatial features simultaneously and got better recognition performance. Besides, RNN has also been widely welcomed. For example, Xiao *et al.* [34] used LSTM to realize multimedia fusion for Chinese SLR. Li *et al.* [35]

proposed an encoder-decoder model using LSTM to model different features of hand shapes.

In recent years, feature fusion gradually receives attention. To make full use of different types of features, Su and Zhu [36] combined the CNN and LSTM to form the fusion network, H. Zhou and W. Zhou also designed the spatial-temporal Multi-Cue network [37] for fully exploring the features from different cues.

Due to the advantages of the attention mechanism, researchers also began to transfer it to SLR. For example, Huang *et al.* [10] used Attention-Based 3DCNN to distribute different weights to different frames in video sequences. According to this idea, we not only use a temporal Attention-Based BLSTM to weight keyframes' representation but also add a spatial Attention-Based BLSTM to weight the sub-planes' features. Then we fuse them to obtain the fusion network.

III. OUR METHOD

A. KEY FRAME SAMPLING

1) SAMPLING METHOD

Before feature extraction, we need to select keyframes from sign language videos. Since each sign language video exists in the format of continuous frames, keyframe sampling is downsampling all the frames of the video to select some representative frames as the descriptor of the whole video. Currently, the standard method of keyframe sampling is uniform sampling, which means that if we select N keyframes from the sign language video with L frames, the index of the ith keyframe K_i is:

$$K_i = [\frac{iL}{N+1}](1 \leq i \leq N) \tag{1}$$

This method does not consider the importance of different frames. Therefore, we refer to KCC sampling in [7] and propose OptimKCC sampling to extract the key actions.

2) KCC SAMPLING

Firstly, we take the first frame as the referenced frame, and we search the keyframe from the subsequent n(hyper parameter) frames. We denote $D_i(1 \leq i \leq n)$ as the Euclidean distance between the pixels of the ith frame and the referenced frame. Long distance means low similarity.

Secondly, we sort the sequence $\{D_i\}_{i=1}^{n}$ to get a new sequence $\{D_{s_i}\}_{i=1}^{n}$ with decreasing similarity ($D_{s_1} \leq D_{s_2} \leq \ldots \leq D_{s_n}$), in which $\{s_1, s_2, \ldots, s_n\} = \{1, 2, \ldots, n\}$. Then, we classify n frames into two categories by threshold segmentation. One is similar to the referenced frame; the other is dissimilar to it. We assume the first k frames corresponding with $D_{s_i}(1 \leq i \leq k)$ as the similar frames. Then, we design the criterion function as:

$$\mathcal{C}(k) = \frac{(m_1 - m_2)^2}{{\sigma_1}^2 + \sigma_2^2} \tag{2}$$

m_1 and m_2 represent the means of the first k and the subsequent $(n-k)$ similarity values. σ_1 and σ_2 represent the standard deviation of the first k and the subsequent $(n-k)$ similarity values.

According to the principle of optimal classification, the result should make the largest mean square error(MSE) between classes and the least MSE within every class, which means the optimal solution k^* should satisfy:

$$\mathcal{C}(k^*) \geq \mathcal{C}(k)(1 \leq k \leq n-1) \tag{3}$$

After finding k^* according to (3), we select the frame which appears earliest in the video from the $(n-k^*)$ dissimilar frames as the keyframe. And we set it as the next referenced frame to find the next keyframe in the same way, until the number of remaining frames is less than n.

3) OPTIMIZATION

According to KCC, for each sample, we gradually change n to select keyframes with the fixed number of N. Take the video sample with L frames as an example, we use $X = (x_1, x_2, \ldots, x_L)$ to denote the frame sequence, and $Y = (y_1, y_2, \ldots, y_N)$ to denote the keyframe sequence selected from X using KCC sampling. We assume that:

$$y_i = x_{s_i}(1 \leq i \leq N, 1 \leq s_i \leq L) \tag{4}$$

Because $L > N$, we refer to DTW distance as the measurement of similarity between X and Y. Firstly, we construct a matrix $\mathcal{M} \in R^{L\times N}$, in which

$$\mathcal{M}_{ij} = D(x_i, y_j)(1 \leq i \leq L, 1 \leq j \leq N) \tag{5}$$

$D(x_i, y_j) = ||x_i - y_j||_2$ represents the Euclidean distance between the pixels of x_i and y_j. Long distance means low similarity.

We use a path $\mathcal{P}$ in matrix $\mathcal{M}$ which starts from the coordinate $(1, 1)$ and ends at (L, N) to match the sequence X and Y. For each point (i, j), the next point along the path can only be one of the following points:

$$(i+1, j), (i, j+1), (i+1, j+1)(1 \leq i \leq L, 1 \leq j \leq N) \tag{6}$$

Each point along the path can be regarded as the matched point between the two sequences. The summation of all the elements along the path, which is shown in eq (7), is defined as the accumulative distance between X and Y:

$$\gamma = \sum_{(i,j)\in\mathcal{P}} \mathcal{M}_{ij} \tag{7}$$

Our objective is to find a path $\mathcal{P}^*$ generating the least accumulative distance, which is defined as the DTW distance:

$$\mathcal{P}^* = \arg\min_{\mathcal{P}} \gamma \tag{8}$$

The DTW distance can be calculated by DTW algorithm, in which $\gamma(1, 1) = \mathcal{M}_{11}$ and $\gamma(L, N)$ is the final result:

$$\gamma(i,j)=\mathcal{M}_{ij} + min\{\gamma(i-1, j), \gamma(i, j-1), \gamma(i-1, j-1)\} \tag{9}$$

We attempt to optimize the result of KCC sampling by using the conception of DTW distance. We set Y as the initial sequence and gradually approach the optimal result using the greedy algorithm. The flowchart of the algorithm is shown as follow:

```
Initialization: $s_0 = 1, s_{N+1} = L, j^* = s_1$
for $1 \leq i \leq N$:
    for $s_{i-1} \leq j \leq s_{i+1}$:(search one by one)
        $y_i = x_j$ (change $y_i$ to get a new sequence)
        if $\gamma(X, Y) < \gamma_{min}$ :
            $\gamma_{min} = \gamma(X, Y)$
            $j^* = j$
    $s_i = j^*$
    $y_i = x_{j^*}$(renew $y_i$)
```

We use Y^* to denote the final keyframe sequence got from the above algorithm, which will be used for data processing and feature extraction.

Because that the sequence Y^* is based on Y, it preserves the characteristic that it considers the different weights of different frames. Besides, Y^* shows more similarity between X compared with Y. So, we can conclude that Y^* can better capture the visual tempo of the video and fully describe the sign language video.

B. MPVR(MULTI PLANE VECTOR RELATION) FEATURE

1) SKELETAL DATA

Kinect-2.0 can capture the 3D coordinate data of 25 skeletal joints. We take the spine joint, which keeps still during almost the whole process of sign language demonstration, as the new coordinate origin to normalize the 3D skeletal coordinate data to eliminate the influence from the heights and the body shapes of signers. In the new 3D coordinate space, the lines connecting the spine joint with other joints can be viewed as 3D vectors.

We select 10 joints closely related to sign language demonstration: thumb, wrist, elbow, index fingertip, and palm center on the left and right sides to get 10 3D vectors. The extraction of MPVR is based on these 3D vectors.

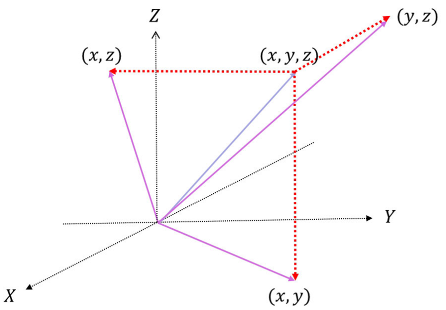


FIGURE 1. Projection. 3D vector (x, y, z) is projected to 3 orthogonal subplanes to form 3 2D vectors (x, y), (x, z), and (y, z).

2) MPVR FEATURE EXTRACTION

Multi-Plane(MP): The meaning of multi-plane is that we project the 3D vector (x, y, z) onto the three orthogonal 2D planes, which are the screen of the camera, the ground, and the sidewall, to obtain three 2D vectors (as shown in Fig. 1). With the same operation on each joint's coordinate, we can get 10 vectors in each plane.

Vector Relation(VR): Take the XY plane as an example, we use $\mathcal{V}_i (1 \leq i \leq 10)$ to represent the 10 vectors in this subplane. For 2 vectors $\mathcal{V}_i(x_i, y_i)$ and $\mathcal{V}_j(x_j, y_j)$, we can use transformation formula to get the polar coordinate (P_i, Θ_i) and (P_j, Θ_j) $(0 \leq \Theta_i, \Theta_j < 2\pi)$ from the rectangular coordinate. We use Θ_{ij} to represent the counterclockwise rotation angle from $\mathcal{V}_i$ to $\mathcal{V}_j$ (As shown in Fig. 2).

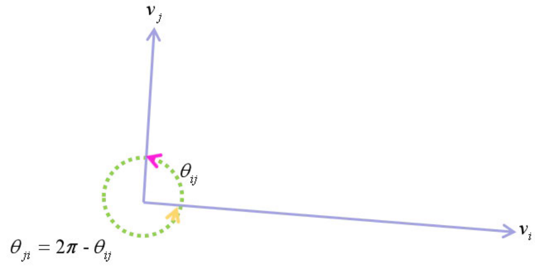


FIGURE 2. Counterclockwise Rotation Angle.Vector $\mathcal{V}_i$ rotates Θ_{ij} counterclockwise to have the same direction as vector $\mathcal{V}_j$. Thus $\mathcal{V}_j$ rotates $(2\pi - \Theta_{ij})$ counterclockwise to have the same direction as $\mathcal{V}_i$.

According to the definition of the counterclockwise rotation angle, we can get that:

$$\Theta_{ij} = \Theta_j - \Theta_i (\Theta_i \leq \Theta_j) \tag{10a}$$

$$\Theta_{ij} = 2\pi + \Theta_j - \Theta_i (\Theta_i > \Theta_j) \tag{10b}$$

$$\Theta_{ij} + \Theta_{ji} = 2\pi \tag{10c}$$

MPVR feature extraction: For one of the three subplanes, we use the vector $\mathcal{M}' \in R^{10}$ to represent $\mathcal{P}_i (1 \leq i \leq 10)$. Meanwhile, we use the matrix $\mathcal{M}'' \in R^{10 \times 10}$ to represent the argument of each vector and the counterclockwise rotation angle between every two vectors:

$$\mathcal{M}'_i = \mathcal{P}_i (1 \leq i \leq 10) \tag{11a}$$

$$\mathcal{M}''_{ij} = \Theta_{ij} (i \neq j, 1 \leq i \leq 10, 1 \leq j \leq 10) \tag{11b}$$

$$\mathcal{M}''_{ii} = \Theta_i (1 \leq i \leq 10) \tag{11c}$$

In this way, we get the matrix $\mathcal{M} = concat(\mathcal{M}', \mathcal{M}'') \in R^{10 \times 11}$ as the vector relation feature of one subplane. Assume that the vector relation features of the three subplanes are $\mathcal{M}_{xy}$, $\mathcal{M}_{xz}$, and $\mathcal{M}_{yz}$. We stack them to form the 3D tensor $stack(\mathcal{M}_{xy}, \mathcal{M}_{xz}, \mathcal{M}_{yz}) \in R^{3 \times 10 \times 11}$ as the MPVR of one keyframe. Assume that N keyframes are selected, we stack their MPVRs to form the tensor with size $N \times 3 \times 10 \times 11$ as the MPVR of the whole sign language video.

The main advantages of MPVR lie in:

- **Scale invariance**: Due to the normalization of the skeletal data, the length of the skeletal joints' vectors can be robust to the diversity of the signers' heights and body shapes.
- **Equivalent reconstructability**: According to the given feature matrices $\mathcal{M}_{xy}$, $\mathcal{M}_{xz}$, and $\mathcal{M}_{yz}$, we can reconstruct the original spatial distribution of skeletal joints.

- **Rotation invariance**: In the process of data acquisition, due to the shake of the camera, the spatial coordinate will change suddenly, resulting in discontinuity and instability of the rectangular coordinate data. However, the length of the skeletal vectors and the counterclockwise rotation angles between them do not change with translation and rotation of the plane facing the signers. Therefore, the features can eliminate the error caused by the camera shaking.
- **Multidirectional**: We take the skeletal trajectories' projection onto the three orthogonal subplanes into consideration, which fully explores the trajectory features during the sign language demonstration.

C. NETWORK

1) ATTENTION-BASED BLSTM

After obtaining the features from the skeletal data, we need to feed them into networks for training. Currently, BLSTM is widely used for extracting features from the sequential data [34], [36]. However, this network is not sensitive to the fact that different keyframes have different importance. Besides, the corresponding weights of the three orthogonal subplanes in describing the sign language video have also not been considered. To solve the problem, we adopt the Attention-Based BLSTM proposed in [10] to weight the features of the keyframes and the subplanes in MPVR. The general structure of the Attention-Based network is shown in Fig. 3.

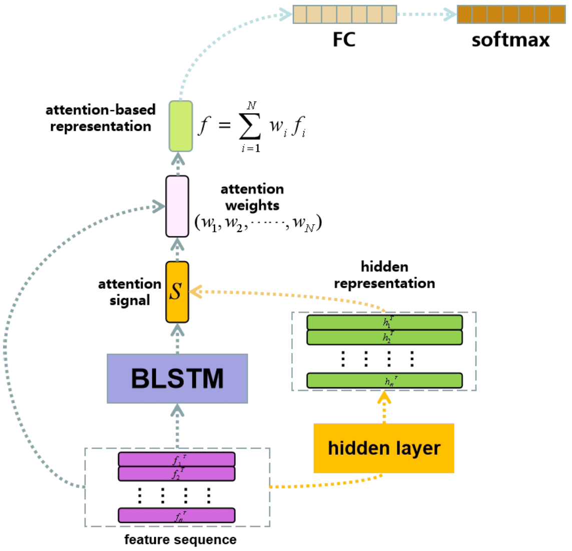


FIGURE 3. Attention-Based LSTM. Feature sequence $\mathcal{F} \in R^{N\times L}$. Attention signal $\mathbf{s} \in R^{256}$. Hidden layer outputs the hidden representation $\mathbf{H} \in R^{N\times 256}$. Weight vector $\mathcal{W} \in R^{N}$.

As shown in Fig. 3, $\mathcal{F} = (\mathbf{f_1}, \mathbf{f_2}, \ldots, \mathbf{f_N})^{\mathbf{T}} \in R^{N\times L}$ represents feature sequence. N means the number of the feature vectors, and L means the length of each feature vector. We set the number of hidden units in the BLSTM to be 128 and feed $\mathcal{F}$ into the BLSTM. The hidden neurons in the hidden layer are composed of Multi-Layer Perceptrons(MLP).

In this way, we get the attention signal from the BLSTM:

$$\mathbf{s} = \mathbf{BLSTM}(\mathcal{F}) \in R^{256} \tag{12}$$

Hidden layer $\mathcal{H}$ outputs the hidden representation $\mathbf{H}$:

$$(\mathbf{h_1}, \mathbf{h_2}, \ldots, \mathbf{h_N})^{\mathbf{T}} = \mathbf{H} = \mathcal{H}(\mathcal{F}) \in R^{N\times 256} \tag{13}$$

where $\mathcal{H} = \mathcal{A}(\mathcal{F}\mathbf{C})$. $\mathcal{F}\mathbf{C}$ means fully-connected layer and $\mathcal{A}$ means activation function.

Then we calculate the weight vector $(w_1, w_2, \ldots, w_N) = \mathcal{W} \in R^N$:

$$w_i = \frac{\mathbf{e}^{\mathbf{h_i^T s}}}{\sum \mathbf{e}^{\mathbf{h_i^T s}}} (1 \le i \le N) \tag{14}$$

Finally, we weight different feature vectors by the weight vector $\mathcal{W}$ to get the final feature vector:

$$\mathbf{f} = \sum w_i \mathbf{f_i} \in R^L \tag{15}$$

We use $\mathcal{B}$ to denote the whole Attention-Based BLSTM, thus we get:

$$\mathbf{f} = \mathcal{B}(\mathbf{f}) \in R^L \tag{16}$$

D. TEMPORAL-SPATIAL ATTENTION-BASED BLSTM

Assume that we choose $N = 8$ as the number of the keyframes in our experiments and denote the Temporal Attention-Based BLSTM and the Spatial Attention-Based BLSTM as $\mathcal{B}_T$ and $\mathcal{B}_S$. Then we feed the feature sequence $\mathbf{f}$ into them. The structure of the fused network is shown in Fig. 4.

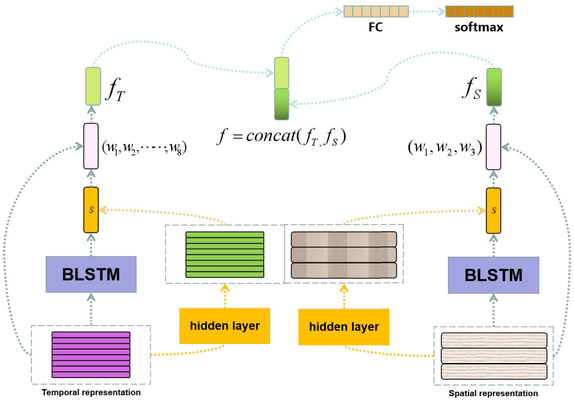


FIGURE 4. Fused Attention-Based BLSTM. The feature sequence is fed into $\mathcal{B}_T$ and $\mathcal{B}_S$ with $\mathbf{f_T}$ and $\mathbf{f_s}$ as the outputs. Then, we concatenate them and feed them into the fully-connected layer and the softmax layer to get the probability distribution vector **p**.

As for $\mathcal{B}_T$, we set the size of $\mathbf{f}$ to be $8 \times 3 \times 10 \times 11$. Thus, we get the temporal feature vector:

$$\mathbf{f_T} = \mathcal{B}_\mathbf{T}(\mathbf{f}) \in R^{330} \tag{17}$$

Similarly, as for $\mathcal{B}_S$, we set the size of $\mathbf{f}$ to be $3 \times 8 \times 10 \times 11$ and get the spatial feature vector:

$$\mathbf{f_S} = \mathcal{B}_\mathbf{S}(\mathbf{f}) \in R^{880} \tag{18}$$

To make full use of the temporal and spatial characteristics, we concatenate them to form the fusion feature:

$$\mathbf{F} = concat(\mathbf{f_T}, \mathbf{f_s}) \in R^{1210} \tag{19}$$

Finally, we feed $\mathbf{F}$ into the fully-connected layer and the softmax layer to get the probability distribution vector, where C is the number of the classes.:

$$\mathbf{p} = \mathbf{softmax}(\mathcal{F}\mathbf{C}(\mathbf{F})) \in R^C \tag{20}$$

E. LOSS FUNCTION

We use Cross Entropy as the loss function. For a probability distribution vector $\mathbf{p} = (p_1, p_2, \ldots, p_C)$, if the ground truth label is $i(1 \leq i \leq C)$, the loss function is:

$$\mathcal{L}oss = -ln(p_i) \tag{21}$$

IV. EXPERIMENT

A. IMPLEMENT DETAILS

1) DATASET

Our experiments were implemented on the DEVISIGN sign language dataset released by the the Chinese Academy of Sciences and Chinese Sign Language(CSL) dataset recorded by us.

DEVISIGN dataset includes 500 sign language words and used Kinect-1.0 to capture RGB, depth, and skeletal data. The 500 words cover signs with fundamental postures to complex postures variations. The data covers 8 different signers. 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).

CSL dataset contains 200 sign language words, which are collected by Kinect-2.0. All the 200 words are from the Chinese Sign Language Textbooks. Each word in CSL dataset contains 100 video samples obtained by 10 signers who repeated the same sign language word 10 times. CSL dataset can provide more detailed skeletal information than DEVISIGN dataset because of the superiority of Kinect-2.0 over Kinect-1.0.

2) CONTENT OF THE EXPERIMENTS

We did self-comparison experiments on CSL to validate the effect of keyframe sampling and the attention mechanism. After that, we realized different methods proposed in other literature on DEVISIGN and CSL dataset to validate the advantages of our methods.

Besides, we did experiments under two cases: Signer-Independent and Signer-Dependent. The former means that the signers in the training set are completely different from those in the test set. The latter means that there are some signers appear in both datasets.

Obviously, experiments under the Signer-Independent circumstance is more challenging but has more practical application value. By comparing the recognition results under the two cases, the networks' robustness to the Signer-Independent circumstance can be observed. Now, many works researching SLR tend to include the two cases in the experiments [14], [18].

The experiments were conducted on GPU 1080Ti with the stochastic gradient descent(SGD) optimizer and the CrossEntropyLoss criteria. We set batch size = 8, learning rate = 0.01, learning decay = 0.99, momentum = 0.9. 80% of the samples were used for training, 5% for validation, and the remaining 15% for testing.

B. EXPERIMENTAL RESULTS

1) EXPERIMENTS ON THE NUMBER OF KEYFRAMES

Firstly, we changed the number of keyframes N and conducted experiments on CSL dataset. The results are shown in Fig. 5-9.

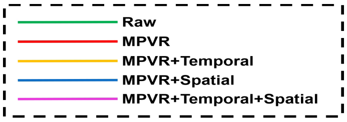


FIGURE 5. Illustration.The lines with different colors mean different methods listed in Table 1.

TABLE 1. Illustration for different methods.

METHOD	
BLSTM with original coordinate data	Raw
BLSTM with MPVR	MPVR
Temporal Attention-Based BLSTM with MPVR	MPVR+temporal
Spatial Attention-Based BLSTM with MPVR	MPVR+spatial
Fused Attention-Based BLSTM with MPVR	MPVR+temporal+spatial

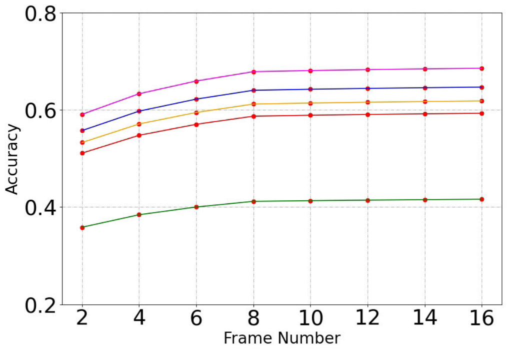


FIGURE 6. Illustration. The curves of the accuracy concerning the number of keyframes *N* under the case of Signer-Independent using KCC sampling.

With the number of keyframes N increasing from 2 to 16, the recognition accuracy also gradually increases. However, after N equals 8, the rising speed drops sharply, and the

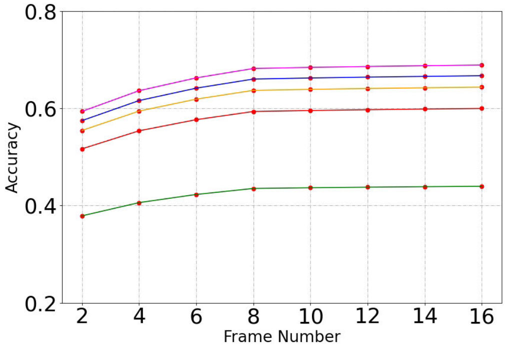


FIGURE 7. Illustration.The curves of the accuracy concerning the number of keyframes *N* under the case of Signer-Indpendent using OptimKCC sampling.

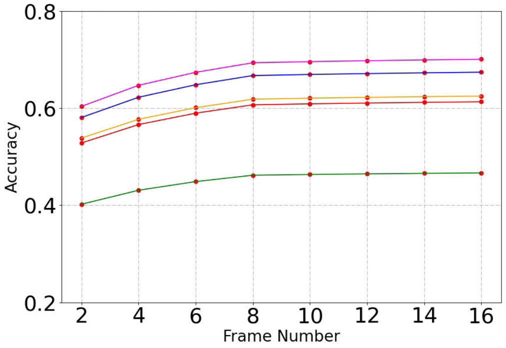


FIGURE 8. Illustration.The curves of the accuracy concerning the number of keyframes *N* under the case of Signer-Dependent using KCC sampling.

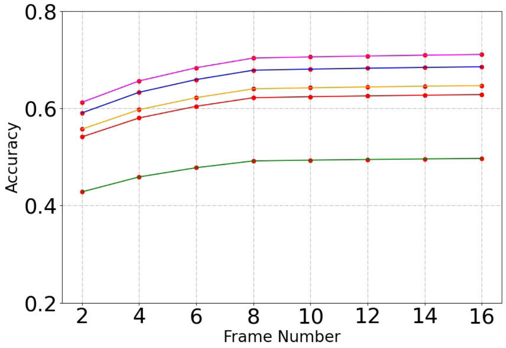


FIGURE 9. Illustration.The curves of the accuracy concerning the number of keyframes *N* under the case of Signer-Dependent using OptimKCC sampling.

accuracy reaches saturation. So, we can conclude that when $N = 8$, the keyframes can fully describe sign language videos. Considering that the number of frames of most sign language videos varies from 80 to 120, keyframe sampling can significantly reduce the data storage memory without bringing significantly negative influence on recognition performance. Subsequent comparison experiments are based on the results when $N = 8$.

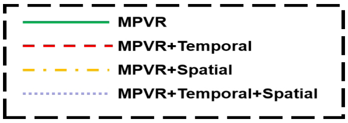


FIGURE 10. Illustration. Different types of lines represent different methods.

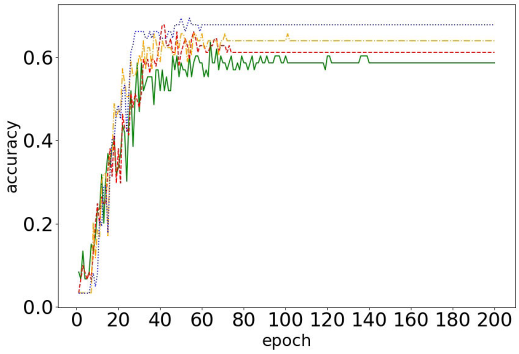


FIGURE 11. Illustration.The recognition accuracy during training under the case of Signer-Independent using KCC sampling.

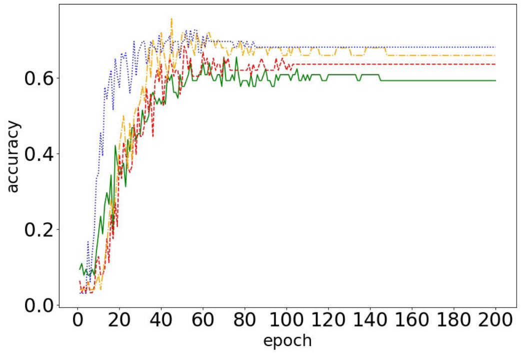


FIGURE 12. Illustration.The recognition accuracy during training under the case of Signer-Independent using optimKCC sampling.

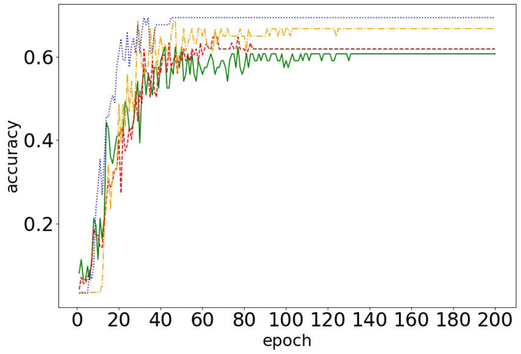


FIGURE 13. Illustration.The recognition accuracy during training under the case of Signer-Dependent using KCC sampling.

2) EXPERIMENTS ON THE ATTENTION MECHANISM

Under the above parameter settings, we implemented experiments with different methods and recorded accuracy and loss during the training process. The accuracy-epoch curves and loss-epoch curves under the Signer-Independent and the Signer-Dependent cases are shown in Fig. 10-18.

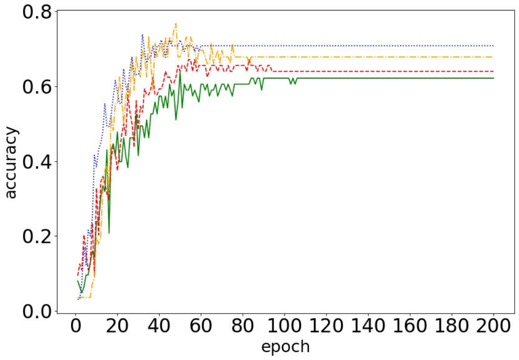


FIGURE 14. Illustration.The recognition accuracy during training under the case of Signer-Dependent using optimKCC sampling.

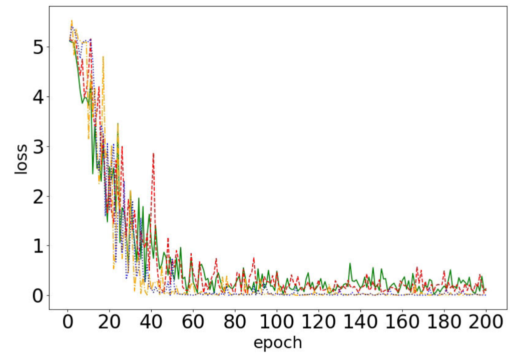


FIGURE 15. Illustration.The loss during training under the case of Signer-Independent using KCC sampling.

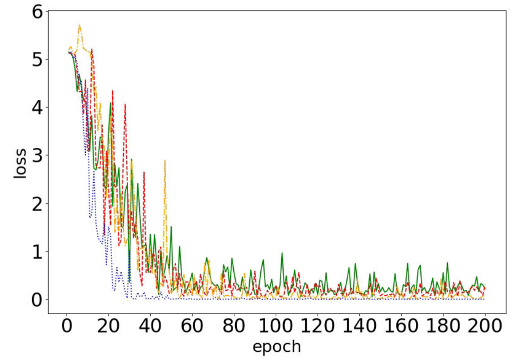


FIGURE 16. Illustration.The loss during training under the case of Signer-Independent using optimKCC sampling.

3) COMPARATIVE EXPERIMENTS WITH OTHER METHODS

To validate the advantages of the Fused Attention-Based BLSTM with MPVR, we did experiments on DEVISIGN and CSL dataset with our methods and some state-of-art methods researching extracting skeletal features for SLR. The experimental results are shown in Tables 2 and 3.

C. ANALYSIS OF EXPERIMENTAL RESULTS

From the experimental results, we can observe that:

- Compared with KCC sampling in [7], OptimKCC sampling could slightly improve the recognition accuracy. OptimKCC sampling preserves the characteristic that it considers the different weights of different frames. Besides, its results show more similarity between the original samples compared with the results of KCC sampling. The experiments indeed confirm that optimKCC sampling could better capture the representation of the sign language videos.
- Compared with original coordinate data, the new feature MPVR can significantly enhance recognition accuracy. Besides, the networks with the attention mechanism have better performance, and the Spatial Attention-Based BLSTM performs even better than the Temporal Attention-Based BLSTM, from which we can conclude that considering weight distribution of different subplanes' features can better describe the sign language videos.
- As expected, the recognition accuracy under the Signer-Dependent circumstance is higher than that under the Signer-Independent circumstance. Nevertheless, we can find that there is not much gap between two cases, which indicates that our networks can avoid the overfitting phenomenon on the training set under the Signer-Independent case, and shows that our networks can be robust to such circumstance and demonstrate practical application value.

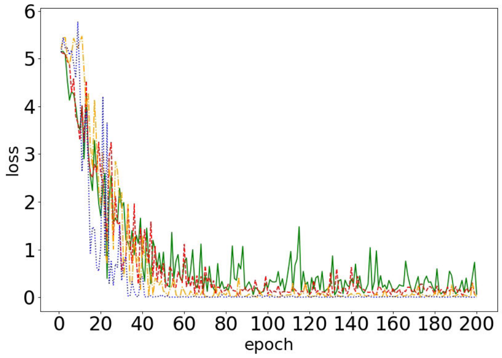


FIGURE 17. Illustration.The loss during training under the case of Signer-Dependent using KCC sampling.

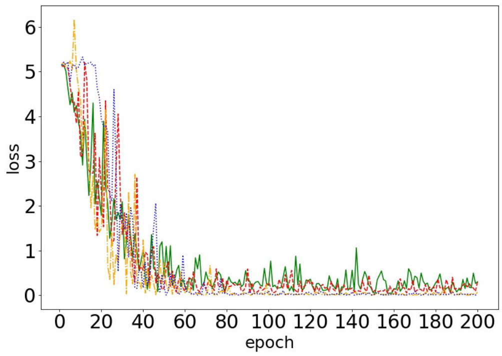


FIGURE 18. Illustration.The loss during training under the case of Signer-Dependent using optimKCC sampling.

TABLE 2. The experimental results on CSL dataset.

CSL dataset	Signer-Independent		Signer-Dependent	
METHOD	KCC [7]	optimKCC	KCC [7]	optimKCC
Raw	41.16%	44.06%	46.17%	49.93%
MPVR	58.67%	59.73%	60.67%	62.52%
MPVR+temporal [10]	61.17%	64.16%	61.83%	64.31%
MPVR+spatial	64.00%	66.52%	66.70%	67.96%
MPVR+temporal+spatial	67.83%	68.32%	69.33%	70.24%
multi-view skeleton [23]	67.72%	67.94%	69.36%	70.11%
shapecontext+HMM [26]	64.78%	65.89%	66.23%	67.52%
trjectory+HMM [28]	62.58%	63.32%	64.17%	64.55%
iDTs [38]	61.22%	61.78%	63.08%	63.57%

TABLE 3. The experimental results on DEVISIGN dataset.

DEVISIGN dataset	Signer-Independent		Signer-Dependent	
METHOD	KCC [7]	optimKCC	KCC [7]	optimKCC
Raw	34.89%	37.08%	39.38%	41.82%
MPVR	49.93%	50.66%	51.68%	52.91%
MPVR+temporal [10]	52.03%	54.23%	52.78%	54.33%
MPVR+spatial	54.61%	56.19%	57.03%	57.86%
MPVR+temporal+spatial	57.71%	58.00%	58.79%	60.31%
multi-view skeleton [23]	57.41%	57.85%	59.11%	59.87%
shapecontext+HMM [26]	55.14%	55.83%	56.41%	57.14%
trajectory+HMM [28]	53.26%	53.97%	54.61%	55.03%
iDTs [38]	52.67%	52.67%	53.21%	53.38%

- As shown in Fig. 10-18, when combined with the attention mechanism, the performance of the networks is significantly improved in several aspects, including higher accuracy, faster convergence speed, and lower loss function with slighter fluctuation amplitude. The Fused Temporal-Spatial Attention-Based BLSTM has the highest recognition accuracy, the fastest convergence speed, and the lowest loss function, which means the optimal performance.
- As shown in Table 2. and Table 3., our methods performed better than other state-of-art and classical methods using skeletal features for SLR. It shows that MPVR and the attention mechanism can fully explore the temporal and the spatial features of sign language videos and consider the importance of different kinds of features and different components of a feature, which means the better ability to capture the representation of the whole sign language videos.

V. CONCLUSION

In this paper, we proposed a kind of Attention-Based network utilizing the OptimKCC sampling and the MPVR skeletal feature to improve the accuracy of SLR. First of all, we designed OptimKCC sampling based on [7] to get the keyframes from sign language videos. Secondly, we projected the skeletal joints' coordinate data to 3 orthogonal subplanes to get several 2D vectors and extracted vector relation from different subplanes as the MPVR skeletal feature.

Afterward, based on the attention mechanism, we adopted a temporal Attention-Based BLSTM and a spatial Attention-Based BLSTM for distributing weights to the features of different keyframes and different subplanes in MPVR.

Under different cases, we conducted experiments on our laboratory's CSL dataset and the public DEVISIGN dataset. The experimental results showed the advantages of our methods.

REFERENCES

[1] A. A. Kindiroglu, O. Ozdemir, and L. Akarun, "Temporal accumulative features for sign language recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Seoul, South Korea, Oct. 2019, pp. 1288–1297, doi: 10.1109/ICCVW.2019.00164.

[2] Y. Han, "A low-cost visual motion data glove as an input device to interpret human hand gestures," *IEEE Trans. Consum. Electron.*, vol. 56, no. 2, pp. 501–509, May 2010, doi: 10.1109/TCE.2010.5505962.

[3] C. Oz and M. C. Leu, "American sign language word recognition with a sensory glove using artificial neural networks," *Eng. Appl. Artif. Intell.*, vol. 24, no. 7, pp. 1204–1213, Oct. 2011.

[4] I. Hussain, A. K. Talukdar, and K. K. Sarma, "Hand gesture recognition system with real-time palm tracking," in *Proc. India Conf.*, 2015, pp. 1–6.

[5] A. Tuntakurn, S. S. Thongvigitmanee, V. Sa-Ing, S. Hasegawa, and S. S. Makhanov, "Natural interactive 3D medical image viewer based on finger and arm gestures," in *Proc. 6th Biomed. Eng. Int. Conf.*, Oct. 2013, pp. 1–5.

[6] K. K. Biswas and S. K. Basu, "Gesture recognition using Microsoft Kinect," in *Proc. Int. Conf. Automat.*, Dec. 2012, pp. 100–103.

[7] S. Huang, C. Mao, J. Tao, and Z. Ye, "A novel chinese sign language recognition method based on keyframe-centered clips," *IEEE Signal Process. Lett.*, vol. 25, no. 3, pp. 442–446, Mar. 2018, doi: 10.1109/LSP.2018.2797228.

[8] Q. Xiao, M. Qin, and Y. Yin, "Skeleton-based chinese sign language recognition and generation for bidirectional communication between deaf and hearing people," *Neural Netw.*, vol. 125, pp. 41–55, May 2020.

[9] C. Li, Y. Hou, P. Wang, and W. Li, "Joint distance maps based action recognition with convolutional neural networks," *IEEE Signal Process. Lett.*, vol. 24, no. 5, pp. 624–628, May 2017, doi: 10.1109/LSP.2017.2678539.

[10] J. Huang, W. Zhou, H. Li, and W. Li, "Attention-based 3D-CNNs for large-vocabulary sign language recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 9, pp. 2822–2832, Sep. 2019, doi: 10.1109/TCSVT.2018.2870740.

[11] G. Joshi, S. Singh, and R. Vig, "Taguchi-TOPSIS based HOG parameter selection for complex background sign language recognition," *J. Vis. Commun. Image Represent.*, vol. 71, Aug. 2020, Art. no. 102834.

[12] K.-S. Kim and H.-I. Choi, "Sign language recognition system using SVM and depth camera," *J. Korea Soc. Comput. Inf.*, vol. 19, pp. 63–72, 2014, doi: 10.9708/jksci.2014.19.11.063.

[13] S. Reshna and M. Jayaraju, "Spotting and recognition of hand gesture for indian sign language recognition system with skin segmentation and SVM," in *Proc. Int. Conf. Wireless Commun., Signal Process. Netw. (WiSPNET)*, Mar. 2017, pp. 386–390.

[14] W. Aly, S. Aly, and S. Almotairi, "User-independent American sign language alphabet recognition based on depth image and PCANet features," *IEEE Access*, vol. 7, pp. 123138–123150, 2019, doi: 10.1109/ACCESS.2019.2938829.

[15] K. Lim, A. W. Tan, C. Lee, and S. C. Tan, "Isolated sign language recognition using convolutional neural network hand modelling and hand energy image," *Multim. Tools Appl.*, vol. 78, no. 14, p. 19 917–19 944,2019, doi: 10.1007/s11042-019-7263-7.

[16] E. K. Kumar, P. V. V. Kishore, A. S. C. S. Sastry, M. T. K. Kumar, and D. A. Kumar, "Training CNNs for 3-D sign language recognition with color texture coded joint angular displacement maps," *IEEE Signal Process. Lett.*, vol. 25, no. 5, pp. 645–649, May 2018.

[17] S. S. Shanta, S. T. Anwar, and M. R. Kabir, "Bangla sign language detection using SIFT and CNN," in *Proc. Int. Conf. Comput., Commun. Netw. Technol.*, Jul. 2018, pp. 1–6.

[18] S. Aly and W. Aly, "DeepArSLR: A novel signer-independent deep learning framework for isolated arabic sign language gestures recognition," *IEEE Access*, vol. 8, pp. 83199–83212, 2020, doi: 10.1109/ACCESS.2020.2990699.

[19] A. Kiani Sarkaleh, F. Poorahangaryan, B. Zanj, and A. Karami, "A neural network based system for persian sign language recognition," in *Proc. IEEE Int. Conf. Signal Image Process. Appl.*, Kuala Lumpur, Malaysia, 2009, pp. 145–149, doi: 10.1109/ICSIPA.2009.5478627.

[20] S. C J and L. A, "Signet: A deep learning based indian sign language recognition system," in *Proc. Int. Conf. Commun. Signal Process. (ICCSP)*, Apr. 2019, pp. 596–600.

[21] S. He, "Research of a sign language translation system based on deep learning," in *Proc. Int. Conf. Artif. Intell. Adv. Manuf. (AIAM)*, Oct. 2019, pp. 392–396.

[22] E. K. Kumar, P. V. V. Kishore, M. T. Kiran Kumar, and D. A. Kumar, "3D sign language recognition with joint distance and angular coded color topographical descriptor on a 2 stream CNN," *Neurocomputing*, vol. 372, pp. 40–54, Jan. 2020.

[23] R. Rastgoo, K. Kiani, and S. Escalera, "Hand sign language recognition using multi-view hand skeleton," *Expert Syst. Appl.*, vol. 150, Jul. 2020, Art. no. 113336, doi: 10.1016/j.eswa.2020.113336.

[24] W. Gao, G. Fang, D. Zhao, and Y. Chen, "A chinese sign language recognition system based on SOFM/SRN/HMM," *Pattern Recognit.*, vol. 37, no. 12, pp. 2389–2402, Dec. 2004.

[25] W. Gao, J. Ma, J. Wu, and C. Wang, "Sign language recognition based on HMM/ANN/DP," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 14, no. 5, pp. 587–602, Aug. 2000.

[26] J. Zhang, W. Zhou, C. Xie, J. Pu, and H. Li, "Chinese sign language recognition with adaptive HMM," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2016, pp. 1–6, doi: 10.1109/ICME.2016.7552950.

[27] D. Guo, W. Zhou, M. Wang, and H. Li, "Sign language recognition based on adaptive HMMS with data augmentation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 2876–2880, doi: 10.1109/ICIP.2016.7532885.

[28] J. Pu, W. Zhou, J. Zhang, and H. Li, "Sign language recognition based on trajectory modeling with HMMs," in *MultiMedia Modeling* (Lecture Notes in Computer Science), vol. 9516, Q. Tian, N. Sebe, G. J. Qi, B. Huet, R. Hong, and X. Liu, Eds. Cham, Switzerland: Springer, 2016, doi: 10.1007/978-3-319-27671-7_58.

[29] P. Q. Thang, N. T. Thuy, and H. T. Lam, "The SVM, SimpSVM and RVM on sign language recognition problem," in *Proc. 7th Int. Conf. Inf. Sci. Technol. (ICIST)*, Apr. 2017.

[30] H. D. Yang and S. W. Lee, "Garbage model formulation for sign language spotting with conditional random fields(internationa session 7)," *Tech. Rep. Prmu*, vol. 107, no. 281, pp. 179–185, 2007.

[31] R. Su, X. Chen, S. Cao, and X. Zhang, "Random forest-based recognition of isolated sign language subwords using data from accelerometers and surface electromyographic sensors," *Sensors*, vol. 16, no. 1, p. 100, Jan. 2016.

[32] J. Zamora-Mora and M. Chacn-Rivas, "Real-time hand detection using convolutional neural networks for costarican sign language recognition," in *Int. Conf. Inclusive Technol. Educ.*, Oct. 2020, pp. 180–186.

[33] M. Al-Hammadi, G. Muhammad, W. Abdul, M. Alsulaiman, M. A. Bencherif, and M. A. Mekhtiche, "Hand gesture recognition for sign language using 3DCNN," *IEEE Access*, vol. 8, pp. 79491–79509, 2020, doi: 10.1109/ACCESS.2020.2990434.

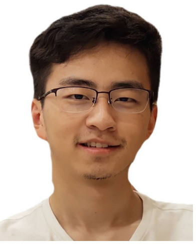
[34] Q. Xiao, M. Qin, P. Guo, and Y. Zhao, "Multimodal fusion based on LSTM and a couple conditional hidden Markov model for chinese sign language recognition," *IEEE Access*, vol. 7, pp. 112258–112268, 2019, doi: 10.1109/ACCESS.2019.2925654.

[35] X. Li, C. Mao, S. Huang, and Z. Ye, "Chinese sign language recognition based on SHS descriptor and encoder-decoder LSTM model," in *Proc. Chin. Conf. Biometric Recognit.*, 2017, pp. 716–728.

[36] S. Yang and Q. Zhu, "Continuous chinese sign language recognition with CNN-LSTM," in *Proc. 9th Int. Conf. Digit. Image Process. (ICDIP)*, Jul. 2017, Art. no. 104200F.

[37] H. Zhou, W. Zhou, Y. Zhou, and H. Li, "Spatial-temporal multi-cue network for continuous sign language recognition," *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 13–16.

[38] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *Int. J. Comput. Vis.*, vol. 103, no. 1, pp. 60–79, May 2013.



WEI PAN received the B.E. degree in communication engineering from the University of Science and Technology of China, Hefei, China, in 2019, where he is currently pursuing the degree. His research interest includes sign language recognition.

XIONGQUAN ZHANG received the B.E. degree in communication engineering from the Hefei University of Technology, Hefei, China, in 2018. He is currently pursuing the degree with the University of Science and Technology of China. His research interest includes hand pose estimation.

ZHONGFU YE received the B.Eng. and M.S. degrees in electronic and information engineering from the Hefei University of Technology, Hefei, China, in 1982 and 1986, respectively, and the Ph.D. degree from the University of Science and Technology of China, Hefei, in 1995. He is currently a Professor with the University of Science and Technology of China. His current research interests include statistical and array signal processing, speech processing, sign language recognition, and hand pose estimation.