IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 19, 2019, accepted December 7, 2019, date of publication December 13, 2019,
date of current version December 23, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2959206

Multimodal Spatiotemporal Networks for Sign Language Recognition

SHUJUN ZHANG, WEIJIA MENG, HUI LI, AND XUEHONG CUI
College of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266061, China

Corresponding author: Shujun Zhang (lindazsj@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61702295 and Grant 61672305 and in part by the Key Research and Development Plan Project of Shandong Province, China, under Grant 2017GGX10127.

ABSTRACT Different from other human behaviors, sign language has the characteristics of limited local motion of upper limb and meticulous hand action. Some sign language gestures are ambiguous in RGB video due to the influence of lighting and background color, which affects the recognition accuracy. We propose a multimodal deep learning architecture for sign language recognition which effectively combines RGB-D input and two-stream spatiotemporal networks. Depth videos, as an effective compensation of RGB input, can supply additional distance information about the signer's hands. A novel sampling method called ARSS (Aligned Random Sampling in Segments) is put forward to select and align optimal RGB-D video frames, which improves the capacity utilization of multimodal data and reduces the redundancy. We get the hand ROI by joints information of RGB data for local focus in spatial stream. D-shift Net is proposed as depth motion feature extraction in temporal stream, which fully utilizes three dimensional motion information of the sign language. Both streams are fused by convolutional fusion layer to get complementary features. Our approach explored the multimodal information and enhanced the recognition precision. It obtains the state-the-of-art performance on the datasets of CSL (96.7%) and IsoGD (63.78%).

INDEX TERMS Sign language recognition, two-stream network, motion features, multimodal data.

I. INTRODUCTION

With the development of computer vision, research on single-person behavior recognition has made significant progress. However, it is still a challenging problem to locate small-scale and low-resolution sign language behavior recognition. Sign language recognition is a multidisciplinary research field involving pattern recognition, computer vision, natural language processing and linguistics. This paper mainly studies how to use the latest deep learning method with RGB-D multimodal input to overcome the above difficulties. Our research can be a good illumination for sign language recognition, small displacement behavior recognition, and intelligent systems.

Sign language recognition has always been an important research direction in the field of behavior recognition. In 2019, Microsoft Research brought together a diverse group of experts for an interdisciplinary workshop about sign language recognition, generation, and translation systems [1]. Traditional methods including kinds of pattern recognition and machine learning techniques. Young [2] proposed a Chinese sign language recognition system by exploring the temporal and spatial features of video sequences. 30 groups of the Chinese manual alphabet images were classified by SVM. Wang *et al.* [3] proposed a Chinese sign language similarity evaluation model considering visual, contour and trajectory features. It achieves the similarity estimation of visual features through comparing histogram. The Longest Common Subsequence is applied to the two feature strings. The algorithm calculates the contour similarity, and finally uses the multiple linear regression process to construct the similarity evaluation model. With the great success of deep learning technology in computer vision, the deep learning method has been proven to have a higher recognition accuracy than the traditional method. Pigou *et al.* [4] established an end-to-end deep neural network that combines temporal convolution and bidirectional cyclic neural networks. The network captures the temporal structure of sign language videos by adding time-dimensional convolutions with loop structure to improve frame-level gesture recognition in the video.

Most of the above-mentioned sign language recognition studies only consider the temporal or the spatial features

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

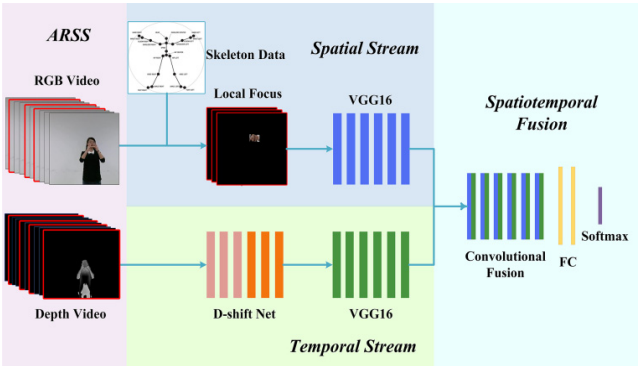

FIGURE 1. An overview of the proposed method. The proposed deep architecture is composed of four components: (a) ARSS, (b) Spatial stream with local focus, (c) Temporal stream with D-shift Net, (d) Spatiotemporal feature fusion and classification result.

of sign language, however, neurologists believe that both temporal features and spatial features play an important role in human cognition. The spatiotemporal two- stream network in deep learning divides the video features into time flow and spatial flow, conforming to the biological human visual perception. Its feasibility and efficiency have been verified on multiple behavior recognition standard data sets. Combining the spatiotemporal two-stream network method and multimodal data input, if the effect of the feature extraction can be enhanced, the performance index of the sign language recognition system will be greatly improved, and the degree of intelligence of the system will be further strengthened. It will be of great significance for intelligence and application.

Spatiotemporal two-stream networks generally use optical flow as input of time stream. The optical flow is the instantaneous velocity of moving objects on the observation imaging plane. This motion is reflected in the movement of the pixels. It uses the changes of the pixel in the frames sequence to find correlation between adjacent images. Optical flow method calculates motion features of objects between sequential frames. This feature requires continuous video input, and the huge calculation amount makes the speed of the whole network model significantly reduced. Therefore, continuous sampling is difficult to cover the entire video, which restricts the development of the optical flow two-stream network.

In this paper, a multimodal two-stream convolutional neural network is used to learn the sign language videos to form robust features and optimize the fusion mode to achieve the final sign language recognition. We propose a sign language recognition method based on multimodal two-stream neural network, as illustrated in Fig. 1.The main contributions are as follows: (1) We proposed a sampling method named Align Random Sampling within Segments (ARSS), which sample RGB data extraction spatial features, and sample aligned depth data extraction time features. (2) The D-shift Net is proposed as a depth motion feature extraction which adapts to the ARSS sampling method and makes full use of the temporal features of the depth data.

With the proposed networks, we combine temporal features and spatial features of sign language recognition. It shows advanced performance in behavior recognition even with small target displacement. This article will introduce the current work of sign language recognition in Section II. Section III will focus on the overall structure of the proposed method and Section IV will show the experimental results and analysis of the model. Finally, Section V will conclude the paper with discussions related to future work.

II. RELATED WORK

With the development of deep learning, the method of extracting sign language features through neural networks has displayed excellent performance and gradually replaced traditional methods. Molchanov *et al.* [5] perform dynamic gesture detection and classification by cyclic 3D CNN, and joins time classification to train the network to predict category labels in undivided input streams. Pigou *et al.* [6] also follow the idea of using CNN to automatically extract features, which consisted of two CNNs, one for extracting hand features and the other for extracting upper body features. Both CNN are designed to containing of two convolutional layers sharing the same weights and fully connected layers. Liu *et al.* [7] use the trajectory of the four skeleton joint points as the network input and add the LSTM network for context information to the study of sign language recognition. To solve the single input problem of sign language recognition, Li *et al.* [8] propose new hand descriptors and LSTM-based time series modeling on these descriptors to achieve accurate recognition on 100 Chinese sign language words. Huang *et al.* [9] embed input data into the RNN network to concentrate on key frame in order to improve the recognition accuracy. Yang and Zhu [10], [11] propose a framework combining CNN with LSTM, and RGB and optical flow data as two inputs. The method is evaluated on the constructed small-scale sign language data set and met the real-time requirements of the small-scale sign language recognition system. Nasri et al. proposed two novel representations for the recognition of moving hand gestures, one is the contour-based similarity images (CBSIs) [12] and the other is spatio-temporal 3D surfaces [13]. Both of them can simultaneously divide the continuous gestures into disjointed gestures and recognize them.

A. RGB-D

With the application and promotion of depth cameras, RGB-D multi-mode input has been widely used in computer vision and object recognition. In addition to traditional feature extraction methods, RGB-D detection methods combined with deep learning convolutional neural network also develop fast. Multi-scale deep learning [14] adopts multimode convolutional neural network to integrate various modal data and learn representations of multiple spatial and temporal scales. The data modes integrated by the algorithm include the grayscale and depth video, as well as the joint pose information extracted from the depth map. They proposed a multi-scale neural model including combination of single-scale paths connected in parallel. Each path

learns a representation independently and performs gesture classification based on the input RGB-D video, joint pose descriptor, and its own time scale. The network separates the left and right hand for feature extraction, ignoring the positional interaction information of the two hands, and the number of networks is large, making the structure complex. Wu *et al.* [15] propose a semi-supervised hierarchical dynamic framework based on HMM for simultaneous gesture segmentation and recognition. Input observations are skeleton joint information, depth and RGB images, using a Gauss-Bernoulli deep belief network suitable for input forms to process bone dynamics, and applying convolutional neural networks to adjust and fuse batch depth and RGB images. Konstantinidis *et al.* [16] propose a sign language recognition RNN networks based on RGB, skeleton data, and facial expression features. The data fusion schemes are analyzed. Miao *et al.* [17] propose a multimodal gesture recognition method based on ResC3D network. One of the key ideas is to find a compact and effective video sequence representation. They use video enhancement techniques such as neural network and median filtering. Eliminate illumination variations and noise in the input video and sample key frames using a weighted frame consistency strategy. Lin *et al.* [18] propose a new model combining a masked Res-C3D network with skeletal data based on LSTM modeling, which process and segment RGB-D video data at the same time.

However, most of the above sign language recognition research is based on the traditional method, simple CNN or 3D-CNN. The framework which combines with RGB-D multimodal input and spatiotemporal two-stream network, is generally blank.

B. SPATIOTEMPORAL TWO-STREAM

Since spatiotemporal two-stream convolution neural network has proposed by Karen and Andrew [19], it has become a common method for behavior recognition. With the advantage of two kind of features and excellent recognition accuracy. The model integrates motion information by training another neural network on the optical flow. By using the appearance features and optical flow features, the accuracy of behavior recognition is significantly improved even by simply merging the probability scores. Many people have made a series of improvements on two-stream networks. Wang *et al.* [20]–[22] add the idea of segmentation and sparse sampling on the basis of two-stream network and proposed TSN network to fuse multiple segments and obtain more context information. The input of two-stream network uses warped optical flow fields to replace the original optical flow, which can eliminate the impact of camera movement. In addition, cross-form pre-training, regularization, data enhancement and other technologies are conducted in the training process to optimize. Zhu *et al.* [23] combine Flownet2.0 [24] with the two-stream network to extract the optical flow information, and took the optical flow features as the temporal ConvNets input. The optical flow information is generated online, which greatly saved the storage space. Sun *et al.* [25] propose optical flow guided feature to represent motion information. They utilize the Sobel operator and element-wise subtraction to calculate the spatial and temporal gradients respectively. Song *et al.* [26] propose Discriminative Motion Cue (DMC) to reduce noises in motion vectors and capture fine motion details. They train the DMC generator to approximate flow using a reconstruction loss and an adversarial loss, jointly with the downstream action classification task. Shou *et al.* [27] introduce a standard 3D CNN to mimic the motion stream by minimizing a feature-based loss compared to the flow stream. They show that the network reproduces the motion stream with high fidelity, and avoids flow computation at test time.

C. CONVERGENCE

In addition to improving the optical flow method, the final fusion method is also important. These classic two-stream networks usually adopt a post-fusion method. After the temporal stream and the spatial stream respectively obtain the recognition result, the recognition result is fused by the weight score. Feichtenhofer *et al.* [28] analyze various fusion methods for two-stream networks, and verify that the results of 3D convolution fusion between layers are better than simple post-fusion, and a convolutional layer fusion and post-fusion parallel use are proposed. This fusion method is for further experimentation on the effect of small data sets. Köpüklü *et al.* [29] propose the MMF model, which combines the optical flow and RGB features through the MLP in the FC6 layer, demonstrating that the feature fusion effect is better than the post-fusion. Crasto *et al.* [30] propose the TACNet, which use a transition-aware classifier in the fusion part to further distinguish transitional states by classifying action and transitional states simultaneously.

The above models based on neural network show good performance in sign language and behavior recognition. However, depth information is only used as a supplement to RGB spatial information and fused at the end of the network model. In fact, depth information reflects more important property of hand language, for example, the distance between the hand and the upper limb. Thus the use of RGB-D information is not sufficient in current research. In view of this, we propose frameworks for sign language recognition by combining RGB-D data with spatiotemporal two-stream network. Specifically, a more perfect spatiotemporal two-stream network model is obtained by combining the depth motion features with the spatial features after local focusing

III. MULTIMODAL SPATIOTEMPORAL NETWORKS

In this section, we will elaborate the proposed multimodal two-stream convolutional neural network. The schematic diagram is shown in Fig. 1. There are four core modules: multimodal input, local focus, D-shift Net, and convolutional fusion. (1) The ARSS method is proposed for optimal sampling and alignment of RGB and depth input, and a relatively complete key frame set of the video is obtained. (2) For RGB spatial stream, we use a local focus approach to obtain the

hand ROI, and try to avoid interference caused by background when extracting gesture features. (3) For depth temporal stream, D-shift Net is proposed in order to make the best use of motion features. (4) Two-stream features are fused at the convolutional layers to preserve the spatial and temporal information, in which the RGB and depth information are complemented effectively. The final recognition is implemented by Softmax.

A. ARSS

Video data is required to be sampled for input to the network. A good sampling method should cover the features of the whole video and significantly reduce the computation complexity. The classical continuous sampling method extracts randomly one concentrated video segment as input but often loses important information. In recent years, the ECO [31] and TRN [32] networks perform multiple sets of sampling on RGB videos, and transfer different sampling segments into multiple networks. They preserve key frames as much as possible, but significantly increase the amount of calculation.

A sign language behavior usually consists a series of several basic actions whose gestures and time coherence can represent the characteristics of the semantics. Therefore, the original behavior video can be divided into multiple segments. Meanwhile, there exists many redundant frames in each segment. We can reasonably use one frame to replace its adjacent frames while keeping the behavior's continuity and completeness. For this reason, we present an approach of equal interval segmentation and random sampling in segments to cover key information and effectively remove redundancy. At the same time, the multimodal input requires not only temporal information reservation between frames, but also the time and space alignment of the two-stream images to obtain meaningful fusion features. Integrating the above ideas, we propose the ARSS algorithm to corresponding sample the depth and RGB data of the same sign language action of the same person. The ARSS algorithm includes two procedures:

First, aligning the spatial position. Due to the different resolutions of the depth camera and the RGB camera, the image sizes of the two video frames are different. RGB and depth images are calibrated by unifying the positions of the corresponding joint points. When the coordinates of the same joint point on the two images are the same, we crop the non-overlapping parts of two images to make them uniform in size.

Second, aligning the temporal position. RGB data is extracted for each RGB video by equal interval segmentation and intra-segment random sampling. The corresponding depth frames are selected from the depth video. We fill the last frame of the depth video into the depth set in order to extract the depth motion feature using D-shift Net. The RGB frame set and the depth frame set are respectively formed. An example of a specific process is shown in Fig. 2.

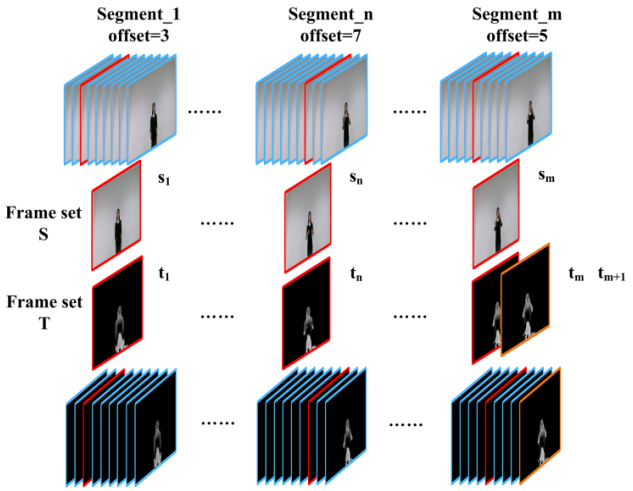

FIGURE 2. ARSS diagram. Input videos are divided into M segments, and random but aligned frames in each video are selected. (1) The RGB video V is equally divided into M segments, and we sample one frame of RGB data randomly from each segment to form an RGB frame set $S = \{s_1, s_2, \ldots s_M\}$ (2) The depth video V′ is divided into M video segments, each frame of which is selected according to the same RGB frame, and the last frame of the video is filled to generate a depth frame set $T = \{t_1, t_2, \ldots t_{M+1}\}$.

B. LOCAL FOCUS

In the field of sign language recognition, the hand detection and location has always been a major problem. In most current dataset and application, the operator usually stands still, only the upper limbs and two hands make movements, Hand is the most flexible limb of the human body and it can produce very detailed actions. The direction of the palm, the gesture of the five fingers and the distance of the hand from the upper limb, all have an influence on sign language semantics. Therefore, the hand action has evident locality and independence from other parts of human body and the background. Aiming at this point, we put forward a local focus approach to get the hand ROI for more direct and precise recognition. Some studies use deep learning networks to generate this local information [33], [34], or use attention module to focus on features from relevant spatial parts as LSTA networks [35]. Differently, we conduct numerical computation to get hand ROI for less complexity and higher efficiency, as illustrated in FIG.3.

The Kinect camera captures 25 joints information. Among them, the left and right hand joints are exactly the centers of the hand ROI. And the size of the ROI is calculated through the relationship between the wrist and elbow joints. Obviously, hand is shorter than forearm even if it is fully open, and when there are other gestures, such as grasping or indicating, the hand size is smaller. In order to locate the complete hand region, we use the length of the human forearm H, as the side length of the hand ROI. According to the joints data, H can be calculated by the Euclidean distance formula:

$$\mathrm{H} = \sqrt{(P_{10x} - P_{9x})^2 + (P_{10y} - P_{9y})^2} \tag{1}$$

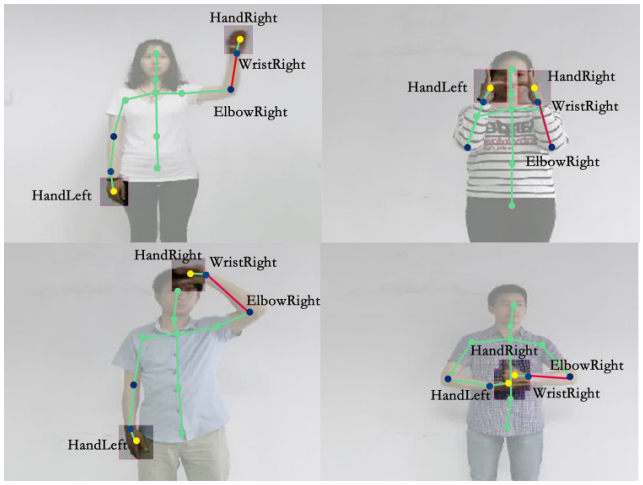

FIGURE 3. We calculate the Euclidean distance H between the joints of the WristRight and the ElbowRight. The size of ROI is equal to 0.9∗H. The center of ROI is the joints of HandLeft and HandRight.

where (P_{9x}, P_{9y}) are the coordinates of the right elbow joint, and (P_{10x}, P_{10y}) are the coordinates of the right wrist joint, since the number of right elbow joint and right wrist joint in Kinect are respectively 9 and 10.

In rare cases, when the arm is perpendicular to the camera, the distance between the wrist and the elbow joint is small, and the length of the arm is not accurately calculated. Therefore, we set a length threshold H_{min} by data analysis to guarantee the robustness and stability of our algorithm.

Unlike the global-local mode [14], [36], this paper keeps the left and right hand ROI in the same RGB image without dividing into two inputs to preserve their relative position and interaction.

The RGB frame set S is locally focused to S', and is transferred to the spatial stream VGG16 for feature extraction, therefore, the VGG16 pays more attention to the key spatial information of the sign language behavior.

C. D-SHIFT NET

The ARSS sampling method obtains a set of intermittent frames, covering the entire sign language action. The random method in the segment causes discontinuity between frames, which cannot meet the fundamental assumption of optical flow which is small displacement of adjacent frames. At the same time, in order to capture the property of the sign language hidden in depth information, this paper uses depth data to emphasize the motion in the direction perpendicular to the camera lens, which is very meaningful for the Chinese sign language with small motion range and a large number of relative front and rear displacements. As the depth changes, the pixel value also changes accordingly, which is contrary to the fundamental assumption that the brightness of the object in optical flow is constant. Therefore, the classical optical flow algorithm [23], [37]–[39] is not applicable to the multimodal input mode of this paper. To this end, we modify and improve the classic FlowNet2.0 algorithm and present

TABLE 1. Architecture of d-shift net.

Layer	Kernel	Str.	Ch I/O	In Res	Out Res	Input
Conv1	3× 3	1	33/64	224 × 224	224 × 224	Frame
Conv1_1	3× 3	1	64/64	224 × 224	224 × 224	Conv1
Conv2	3× 3	2	64/128	224 × 224	112 × 112	Conv1_1
Conv2_1	3× 3	1	128/128	112 × 112	112 × 112	Conv2
Conv3	3× 3	2	128/256	112 × 112	56 × 56	Conv2_1
Conv3_1	3× 3	1	256/256	56 × 56	56 × 56	Conv3
Conv4	3× 3	2	256/512	56 × 56	28 × 28	Conv3_1
Conv4_1	3× 3	1	512/512	28 × 28	28 × 28	Conv4
Conv5	3× 3	2	512/512	28 × 28	14 × 14	Conv4_1
Conv5_1	3× 3	1	512/512	14 × 14	14 × 14	Conv5
Conv6	3× 3	2	512/1024	14 × 14	7 × 7	Conv5_1
Conv6_1	3× 3	1	1024/1024	7 × 7	7 × 7	Conv6
Flow6(loss6)	3× 3	1	1024/20	7 × 7	7 × 7	Conv6_1
Deconv5	4× 4	2	1024/512	7 × 7	14 × 14	Conv6_1
Xconv5	3× 3	1	1044/512	14 × 14	14 × 14	Deconv5+Flow6+Conv5_1
Flow5 (loss5)	3× 3	1	512◁20	14 × 14	14 × 14	Xconv5
Deconv4	4× 4	2	512/256	14 × 14	28 × 28	Xconv5
Xconv4	3× 3	1	788/256	28 × 28	28 × 28	Deconv4+Flow5+Conv4_1
Flow4 (loss4)	3× 3	1	256/20	28 × 28	28× 28	Xconv4
Deconv3	4× 4	2	256/128	28 × 28	56 × 56	Xconv4
Xconv3	3× 3	1	404/128	56 × 56	56 × 56	Deconv3+Flow4+Conv3_1
Flow3 (loss3)	3× 3	1	128/20	56× 56	56 × 56	Xconv3
Deconv2	4× 4	2	128/64	56 × 56	112× 112	Xconv3
Xconv2	3× 3	1	212/64	112× 112	112× 112	Deconv2+Flow3+Conv2_1
Flow2 (loss2)	3× 3	1	64/20	112× 112	112× 112	Xconv2

a kind of depth displacement network called D-shift Net. It adapts to the ARSS mechanism and captures the motion temporal information of the depth data, which better reflects the feature changes of the sign language in 3D space.

The specific improvements include: (1) First, deleting the first convolutional layer with a large receptive field, and reducing the step size of the second convolutional layer to one. (2) Second, we made the beginning of the network deeper by exchanging the 7×7 and 5×5 kernels with multiple 3 × 3 kernels. Detailed parameters are shown in TABEL I. (3) Third, a convolution layer is inserted between each deconvolutional layer of the expanded convolution portion to obtain a smoother depth displacement characteristic.

Further, the original network is modified to be an unsupervised model as Fig. 4. We calculate the moving matrix in the horizontal direction (X direction) and the vertical direction (Y direction) of the adjacent frames in depth frame set T and denote it as depth motion feature D. Each value in the matrix represents the distance in the X direction or Y direction between the pixel $t_{p+1}(i, j)$ and the pixel $t_p(i', j')$ with the same value, denoted as $D^x_{i,j}$ and $D^y_{i,j}$. Then the frame t_p is reconstructed by frame t_{p+1} and D. The reconstructed frame is recorded as t'_p. The formula of reconstruction is expressed as:

$$t'_p = t_{p+1}\left(i + D^X_{i,j}, j + D^Y_{i,j}\right) \tag{2}$$

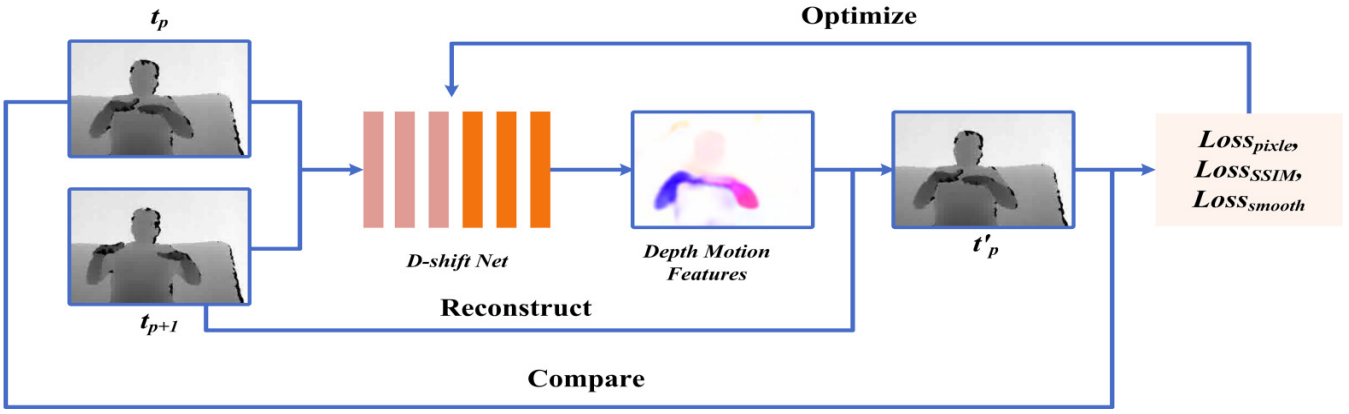

FIGURE 4. The unsupervised model is divided to 5 step. (1)Input frames t_p and t_{p+1} into the D-shift Net. (2) Extract the depth motion feature D between two frames. (3) Reconstruct the t_p by t_{p+1} and D. (4) Compare t_p and t'_p and calculate three kinds of loss. (5) Optimize D-shift Net.

Three kinds of target functions is used to minimizing the difference between $t_p\,(i,j)$ and $t'_p(i,j)$ and improve the quality of the feature D. The specific calculation method is as follows.

1) PIXELS ERROR

We subtract per-pixel value to represent the pixel-level difference of $t_p\,(i,j)$ and $t'_p(i,j)$. The loss function takes the form:

$$L_{pixel} = \frac{1}{N}\sum_{j=1}^{m}\sum_{i=1}^{n}\rho\left(t_p\,(i,j) - t'_p(i,j)\right) \tag{3}$$

where N is the total number of pixels of a frame image, n and m are the height and width of the current frame, so $N = n \times m$; ρ is Charbonnier error.

2) STRUCTURAL SIMILARITY ERROR

L_{SSIM} represents the structural similarity index of two frame. It is a fully referenced image quality evaluation index, which measures image similarity from brightness, contrast and structure.

$$L_{SSIM} = \frac{1}{N}\sum_{j=1}^{m}\sum_{i=1}^{n}\left(1 - SSIM\left(t_p\,(i,j)\,,t'_p\,(i,j)\right)\right) \tag{4}$$

SSIM is specifically expressed as:

$$\mathrm{SSIM}\,(x,y) = \frac{\left(2\mu_x\mu_y + c_1\right)\left(\sigma_{xy} + c_2\right)}{\left(\mu_x^2 + \mu_y^2 + c_1\right)\left(\sigma_x^2 + \sigma_y^2 + c_2\right)} \tag{5}$$

μ_x, μ_y is the average value of pixel (x, y). σ_x, σ_y is the standard deviation of x, y. $\sigma_{x,y}$ is covariance. c_1, c_2 is a constant to avoid a zero denominator.

3) APERTURE ERROR

Calculating the aperture error: Similar to the optical flow, the depth motion feature also has an observation window problem, that is, the aperture problem, and the aperture error is calculated by the objective function L_{smooth}:

$$L_{smooth} = \rho\left(\nabla D_i^x\right) + \rho\left(\nabla D_j^x\right) + \rho\left(\nabla D_i^y\right) + \rho\left(\nabla D_j^y\right) \tag{6}$$

wherein, ∇D_i^x and ∇D_j^x represent the gradient of the horizontal depth flow in the horizontal and vertical directions, ∇D_i^y and ∇D_j^y represent the gradient of the vertical depth flow in the horizontal and vertical directions, ρ is the Charbonnier error.

4) TOTAL ERROR L

$$\mathrm{L} = \lambda_1 \cdot L_{pixel} + \lambda_2 \cdot L_{SSIM} + \lambda_3 \cdot L_{smooth} \tag{7}$$

where $\lambda_1, \lambda_2, \lambda_3$ are the weighting factors and $\lambda_1 + \lambda_2 + \lambda_3 = 1$.

Backpropagation is performed with L as the objective function of the D-shift Net model. The training process is stopped when iterating to L convergence. The trained model is used to extract the depth motion feature map between each pair of adjacent frames to form 10 feature map as a set T'.

The D-shift Net can extract the motion features contained in the depth information, and meets the condition of interval frames obtained by the ARSS, so that the motion features can cover the entire sign language video. This procedure needs small storage space and has high calculation speed with more than 120 frames per second.

D. TWO-STREAM NEURAL NETWORK FUSION

The classic two-stream framework consist of two networks which merge to get the final recognition result. The most common way is to combine the scores of the two streams classifier results in a certain proportion. This method is easy to implement, and does not need to consider the dimension alignment problem of the two streams. However, there are not interaction between two-stream features, so the RGB and D information, temporal and spatial information cannot be complemented with each other by joint training.

In this paper, we use convolutional fusion to combine a two-stream VGG16 network, which the locally focused RGB frame set S' and a depth motion feature set T' are imported to.

Before the fifth set of convolutions, the structure is the same as the classic VGG16 network. The convolutional layer

uses 3×3 convolution kernels. The number of channels is 64, 64, 128, 128, 256, 256, 256, 512, 512, 512, and the pooling mode is 2×2 MAX Pooling. We insert the fusion structure before the fifth set of convolutions. The spatiotemporal two-stream features at this time are respectively as x_a and x_b, and the convolutional fusion model can be expressed as

$$y^{conv} = f^{conv}(x_a, x_b) \tag{8}$$

In performing convolutional fusion, the two feature maps x_a, x_b are first stacked together, and then the channel is convoluted using a $1 \times 1 \times 2$D convolution kernel f^{conv}. Here, the convolution kernel f is used to reduce the dimension twice, and the weight combination of the two feature maps x_a and x_b can be modeled at the same spatial position. f^{conv} learns to minimize entropy loss function of the correspondence between two feature maps when used as a filter kernel in a network.

The fifth set of convolutions is performed on the merged feature y, the number of channels is 512, 512, 512, and the pooling mode is 2×2 Max Pooling. Following the two 4096-dimensional fully connected layers, the neurons are discarded at a dropout rate of 0.9, 0.8, respectively. The final result is classified using Softmax as a classifier.

In training stage, firstly, the temporal network and the spatial networks are trained separately, and then the trained model is frozen to train the fusion structure. When the loss no longer drops, we unfreeze two stream structure to retrain until convergence.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, the proposed method will be evaluated systematically on two public datasets: Chinese Sign Language dataset (CSL) [40] and the ChaLearn LAP large-scale isolated gesture dataset (IsoGD) [14]. First, the two datasets will be described briefly. And then, the training processes will be described in detail. Finally, the evaluation results will be reported respectively.

A. DATASETS

1) CSL

To solve the Chinese sign language problem, experiments were conducted using the CSL data set. The CSL data set is proposed by University of Science and Technology of China for large-scale Chinese sign language recognition. It is recorded by Kinect2.0 and provides three kinds of data: RGB video, depth video and joints information. A vocabulary of 500 Chinese sign language words are included, with 50 people participating in the recording to form a total of 125,000 video instances.

2) ISOGD

At the same time, in order to test the versatility of our method, this paper tests on IsoGD which is proposed by "2016 Looking at People CVPR Challenge" and commonly used for international research on behavior recognition algorithms. The focus of the challenges is "large-scale" learning and "user independent" gesture recognition from RGB or RGB-D videos. The dataset contains a total of 47,933 instances, which consists of 249 categories of gestures recorded by 21 people.

B. EXPERIMENTAL DETAILS

1) ENVIRONMENT CONFIGURATION

The experiments are tested on Ubuntu14.04 system. The server is equipped with NVIDIA GTX 1080Ti and the CPU is Intel Xeon E5-2620 v4. The deep learning framework is Caffe.

2) PARAMETERS

INPUT In the multimodal input, the spatially aligned image size is 512×396, and the unified zoom is 224×224 when entering the network. According to ARSS method, the RGB video and its corresponding depth video of the same behavior are divided into $M = 10$ segments, and one frame is randomly sampled in each segment to form the frame sets S, T. For the depth frame set T, the last frame of the depth video is added to align RGB data. The batch_size of each stream input is 8.

a: LOCAL FOCUS

During the local focus stage, the length of the human forearm H is calculated by the human elbow P_9-*ELBOW_RIGHT* and the wrist joints P_{10}-*WRIST_RIGHT* coordinates. The length threshold H_{min} is set to be 24.

b: HYPER-PARAMETERS

The generalized Charbonnier parameter α is 0.4 in the L_{pixel} objective function and 0.3 in the L_{smooth} objective function. The dropout rates after the two FC layers are respectively 0.9 and 0.8.

c: INITIALIZATION

We use different ways to initialize the two stream networks. The spatial network VGG16 is initialized by Mrsa, and the D-shift Net and temporal network VGG16 in this paper are finetuned with the pre-training model trained on UCF101 to enhance the convergence rate.

d: TRAINING

The two-stream models are first trained separately, then the feature extraction layers are frozen in order to train the fusion structure. We use Adam as the optimization algorithm. The momentum is 0.9 and the weight_decay is 0.0005. A total of 25,000 iterations are used averagely for training, and the adaptive learning rate is adopted. Among them, the basic learning rate of the D-shift Net is 10^{-5}, and that of the two-stream feature extraction network VGG16 and fusion structure is 0.01. We set lr_policy for "multistep", where gamma is 0.1. Stepvalues are 5000, 9500, 14000 and 20000.

e: TESTING

5-fold cross-validation is adopted to divide the original data set and training and testing are carried out for each partition,

TABLE 2. Comparison of recognition accuracy of different model on the validation subset of CSL.

Model	Modality	Accuracy (%)
VGG16	RGB	60.7
D-shift+VGG16	RGB	92.5
D-shift+VGG16	Depth	95.8
FlowNet2.0+VGG16	RGB	85.3
Multimodal Fusion	**RGB-D**	**96.7**

TABLE 3. Comparison of proposed method and other methods on the validation subset of CSL.

Model	Modality	Accuracy (%)
STIP-SVM-FV-SVM[41]	RGB	61.8
iDTs-SVM-FV-SVM[42]	RGB	68.5
GMM-HMM[43]	RGB-D + Skeleton	56.3
C3D[44]	RGB-D	78.1
Huang et al.[45]	RGB-D + Skeleton	88.7
Multimodal Fusion	**RGB-D** + Skeleton	**96.7**

and the final score is obtained by averaging the result of 5 times. The ratio of training data set to testing data set is 3:1 in each partition.

C. EVALUATION ON CSL

As a motion feature extraction network, D-shift can use not only depth data as input, but also RGB data as input to extract features. On CSL dataset, the network with depth data as input shows better performance because the sign language behavior has many displacements in the depth direction. For RGB data, these shifts are not obvious enough. Therefore, if there is fine depth data, we can consider using depth data for motion capture to obtain better motion features for classification.

We compared the RGB+D-shift+VGG16 model with the optical flow algorithm RGB+FlowNet2.0+VGG16 model. The input sampling method of the FlowNet2.0 model is divided into 3 segments on average, and each segment takes 5 frames in succession. As shown in Table 2, the proposed model on CSL is significantly better than the FlowNet2.0+VGG16 model. In the long sign language video, the continuous sampling method required by the optical flow is difficult to cover the entire video, and the temporal characteristics of the sign language cannot be well learned. And the randomness of the increased speed of our sampling method helps to adapting to the speed difference of different sign language operators.

TABLE 3 displays the comparison results with the previously published methods on the validation set of CSL. STIPs [41] are commonly used spatiotemporal features, generated by detecting the 3D Harris angles in the video and calculating the HOG and HOF features around the detection points. iDTs [42] are also currently good manual annotation features composed of trajectory, HOG, HOF and MBH. They are based on optical flow tracking and low horizontal gradient histogram.

After extracting the STIPs and iDTs features from the video, these features are encoded into Fisher Vector using the DFT fisher toolbox, and finally the SVM is used to classify the encoded features. GMM-HMM [43] is a traditional method in time series pattern recognition, which can better construct and classify time series features in sign language video. The above-mentioned methods are traditional feature extractors used on CSL, obviously, their accuracy lags far behind the deep learning methods. C3D [44] is a common neural network for activity recognition. It introduces the attention mechanism named the Attention-pooling method to classify the features. Huang *et al.* [45] used 3D convolutional neural networks and convolutional Long-Short-Term-Memory (LSTM) networks. They believe that learning spatiotemporal features simultaneously is more suitable than learning spatial and temporal features consecutively or separately for gesture recognition. To learn spatiotemporal features synchronously, we use convolutional layer to combine RGB and depth, as well as temporal and spatial information. As a result, the recognition accuracy is improved

D. EVALUATION ON ISOGD

There is no skeleton information on IsoGD dataset. To achieve the part of local focus, we use OpenPose [49], [50] to get wrist and elbow joints. OpenPose is a kind of skeleton extraction technique which combines Part Confidence Maps and Part Affinity Fields to get person's skeleton by greedy algorithm and bipartite matching. OpenPose method can output: (1) 25 key points of body which include the same wrist and elbow joints our Local Focus part used; (2) 2x21 hand key points including all hand joint and endpoint. However, this method does not extract the same *HandLeft* and *HandRight* points like Kinect, so we use metacarpophalangeal joint of middle finger as the center of hand ROI, as shown with the yellow point in Fig.5. Furthermore, when the method cannot locate the metacarpophalangeal joint, the point on extension line of wrist and elbow joints, whose distance from wrist joint is H/2, is regarded as the center. The result of Local Focus on IsoGD is shown in Fig. 5.

On IsoGD, as shown in Table 4, the FlowNet2.0 extraction feature model is better than the D-shift model, which is closely related to the quality of the depth video in IsoGD dataset. The depth image of IsoGD is acquired by Kinect 1.0, which is noisy and insensitive to relatively small hand movements. Therefore, for D-shift Net, low-noise and fine-quality depth data is very important.

Compared to ResC3D [17], [48], our model has significant room for improvement. So far, we have only used a simple VGG16 network in the dual stream network section. Such networks tend to lose some temporal information when extracting high-level features, especially in spatial networks. In the spatial network structure, we used stacked RGB inputs. This type of input retains only a small amount of temporal information which is evenly distributed. Further work can

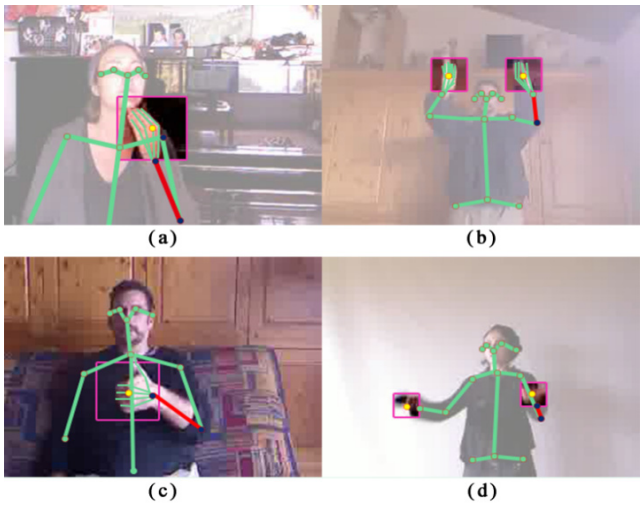

FIGURE 5. Skeleton extraction method named OpenPose is used on IsoGD. Skeleton data can be extracted in most of gesture such as (a) (b) (c). (d) shows the result that if hands are difficult to be recognized, the point in extension cord of wrist and elbow joints works well.

TABLE 4. Comparison of proposed method and other methods on the validation subset of IsoGD.

Model	Modality	Accuracy (%)
C3D+LSTM[46]	RGB-D	51.02
Zhang et al. [47]	RGB-D	58.65
Wang et al. [36]	RGB-D	60.81
ResC3D [17]	RGB-D	64.40
MFFs [29]	RGB	57.40
Lin et al. [48]	RGB-D	64.37
D-shift+VGG16	Depth	56.54
FlowNet2.0+VGG16	RGB	57.62
Multimodal Fusion	**RGB-D**	**63.78**

be combined with C3D networks to retain more temporal information and use a deeper network framework.

E. DISCUSSION

We will compare the effects of the proposed depth motion features and the classical RGB optical flow features in this section. When the depth image quality is satisfactory, the depth motion feature shows excellent performance, especially for the behavior series with apparent displacements in distance to the camera. We show the example of qualitative results on IsoGD in Fig. 6.

Fig. 6(a) shows that the depth motion feature obtains a clearer and more complete motion trajectory than the RGB optical flow, making the feature more vivid, when the motion has obvious depth displacement and the horizontal displacement is small.

Fig. 6(b) displays that when the moving parts are similar in color, the optical flow features are obvious poor. At the same time, the RGB optical stream cannot locate the pixels through the color block, but the depth motion network can extract the

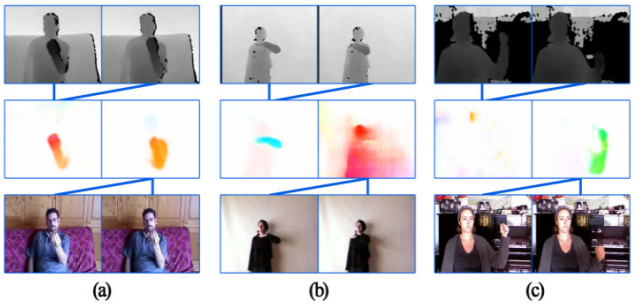

FIGURE 6. (a) demonstrates the validity and superiority of depth motion features at the depth motion level. (b) shows the effect of color on depth motion features and optical flow features. (c) displays the influence of background noise on motion feature extraction quality from depth data.

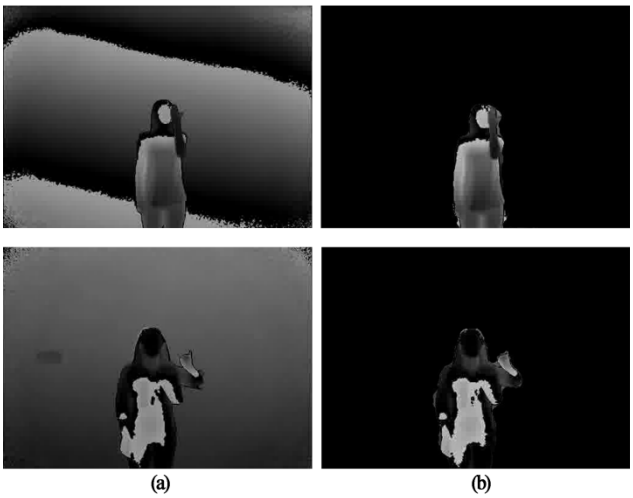

FIGURE 7. (a) CSL original image (b) contour extracted image.

motion features through the change of the depth information. In addition, depth information ignores the effects of light and shadow changes, which greatly avoids the interference of the environment and captures the gesture features more accurate.

Fig. 6(c) shows the fatal effect of depth image quality on feature extraction. The depth data of IsoGD has some depth missing area, which is represented by a depth value of zero. The occlusion problem and the surface material of the object will affect the acquisition of the Kinect depth image. These noises affect the quality of the depth image seriously, resulting in inaccurate depth motion feature extraction. If the background noise is filtered in pre-process, the quality of depth motion features will be significantly improved.

We segmented foreground character on CSL by the contour extraction algorithm Morphological GAC [51], as illustrated in Fig. 7. Contour extraction algorithm can effectively remove the interference of background noise, but it performs poorly on the IsoGD dataset with complex background. Choosing the right pre-processing method to optimize the data set usually yields better results on IsoGD.

V. CONCLUSION

In this paper, the framework is proposed for sign language recognition, which combines both RGB-D input and

two-stream spatiotemporal networks. For aligned multimodal input, the ARSS approach covers key information and effectively removes redundancy. Local focus of the hand optimize the input of spatial network. And D-shift Net generates depth motion features to explore depth information effectively. A convolutional fusion is subsequently conducted to fuse two-stream features and better recognition results. Our future work could involve optimizing the image quality of depth video for more effective motion features extraction and uniting both depth motion features and RGB optical flow, as well as improving the recognition speed without reducing precision.

REFERENCES

[1] D. Bragg, O. Koller, M. Bellard, L. Berke, P. Boudrealt, A. Braffort, N. Caselli, M. Huenerfauth, H. Kacorri, T. Verhoef, and C. Vogler, "Sign language recognition, generation, and translation: An interdisciplinary perspective," 2019, *arXiv:1908.08597*. [Online]. Available: https://arxiv.org/abs/1908.08597

[2] G. O. Young, "Synthetic structure of industrial plastics," in *Plastics*, vol. 3, J. Peters, Ed., 2nd ed. New York, NY, USA: McGraw-Hill, 1964, pp. 15–64.

[3] L.-C. Wang, R. Wang, D.-H. Kong, and B.-C. Yin, "Similarity assessment model for Chinese sign language videos," *IEEE Trans. Multimedia*, vol. 16, no. 3, pp. 751–761, Apr. 2014.

[4] L. Pigou, A. van den Oord, S. Dieleman, M. Van Herreweghe, and J. Dambre, "Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video," 2015, *arXiv:1506.01911*. [Online]. Available: https://arxiv.org/abs/1506.01911

[5] Y. X. Molchanov, S. Gupta, K. Kim, S. Tyree, and J. Kautz, "Online detection and classification of dynamic hand gestures with recurrent 3D convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 4207–4215.

[6] L. Pigou, S. Dieleman, P. J. Kindermans, and B. Schrauwen, "Sign language recognition using convolutional neural networks," in *Proc. Workshop Eur. Conf. Comput. Vis.*, Zurich, Switzerland, 2014, pp. 572–578.

[7] 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.

[8] 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.* Cham, Switzerland: Springer, 2017, pp. 719–728.

[9] 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.

[10] S. Yang and Q. Zhu, "Continuous Chinese sign language recognition with CNN-LSTM," in *Proc. 9th Int. Conf. Digit. Image, Int. Soc. Opt. Photon.*, vol. 10420, 2017, Art. no. 104200F.

[11] S. Yang and Q. Zhu, "Video-based Chinese sign language recognition using convolutional neural network," in *Proc. IEEE 9th Int. Conf. Commun. Softw. Netw. (ICCSN)*, Guangzhou, China, May 2017, pp. 929–934.

[12] N. Neverova, C. Wolf, T. W. Graham, and F. Nebout, "Multi-scale deep learning for gesture detection and localization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 474–490.

[13] S. Nasri, A. Behrad, and F. Razzazi, "A novel approach for dynamic hand gesture recognition using contour-based similarity images," *Int. J. Comput. Math.*, vol. 92, no. 4, pp. 662–685, 2015.

[14] S. Nasri, A. Behrad, and F. Razzazi, "Spatio-temporal 3D surface matching for hand gesture recognition using ICP algorithm," *Signal, Image Video Process.*, vol. 9, no. 5, pp. 1205–1220, 2015.

[15] D. Wu, L. Pigou, P.-J. Kindermans, N. D.-H. Le, L. Shao, J. Dambre, and J.-M. Odobez, "Deep dynamic neural networks for multimodal gesture segmentation and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1583–1597, Aug. 2016.

[16] D. Konstantinidis, K. Dimitropoulos, and P. Daras, "A deep learning approach for analyzing video and skeletal features in sign language recognition," in *Proc. IEEE Int. Conf. Imag. Syst. Techn. (IST)*, Krakow, Poland, Oct. 2018, pp. 1–6.

[17] Q. Miao, Y. Li, W. Ouyang, Z. Ma, X. Xu, W. Shi, X. Cao, Z. Liu, X. Chai, and Z. Liu, "Multimodal gesture recognition based on the Resc3D network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Oct. 2017, pp. 3047–3055.

[18] C. Lin, J. Wan, Y. Liang, and S. Z. Li, "Large-scale isolated gesture recognition using a refined fused model based on masked Res-C3D network and skeleton LSTM," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Xi'an, China, May 2018, pp. 52–58.

[19] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 1, 2014, pp. 568–576.

[20] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, and D. Lin, "Temporal segment networks: Towards good practices for deep action recognition," 2016, *arXiv:1608.00859*. [Online]. Available: https://arxiv.org/abs/1608.00859

[21] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks for action recognition in videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2740–2755, Nov. 2019.

[22] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao, "Towards good practices for very deep two-stream ConvNets," 2015, *arXiv:1507.02159*. [Online]. Available: https://arxiv.org/abs/1507.02159

[23] Y. Zhu, Z. Lan, S. Newsam, and A. Hauptmann, "Hidden two-stream convolutional networks for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Apr. 2017, pp. 363–378.

[24] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "FlowNet 2.0: Evolution of optical flow estimation with deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 1647–1655.

[25] S. Sun, Z. Kuang, L. Sheng, W. Ouyang, and W. Zhang, "Optical flow guided feature: A fast and robust motion representation for video action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 1390–1399.

[26] L. Song, S. Zhang, G. Yu, and H. Sun, "TACNet: Transition-aware context network for spatio-temporal action detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 11987–11995.

[27] Z. Shou, Z. Yan, K. Yannis, S. Laura, and S. F. Chang, "DMC-Net: Generating discriminative motion cues for fast compressed video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 1268–1277.

[28] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 1933–1941.

[29] 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, Jun. 2018, pp. 2103–2111.

[30] N. Crasto, P. Weinzaepfel, K. Alahari, and C. Schmid, "MARS: Motion-augmented RGB stream for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 7882–7891.

[31] M. Zolfaghari, K. Singh, and T. Brox, "ECO: Efficient convolutional network for online video understanding," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 713–730.

[32] B. Zhou, A. Andonian, A. Oliva, and A. Torralba, "Temporal relational reasoning in videos," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 803–818.

[33] 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.

[34] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.

[35] S. Swathikiran, E. Sergio, and L. Oswald, "LSTA: Long short-term attention for egocentric action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 9954–9963.

[36] H. Wang, P. Wang, Z. Song, and W. Li, "Large-scale multimodal gesture recognition using heterogeneous networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Oct. 2017, pp. 3129–3137.

[37] J. S. Pérez, E. Meinhardt-Llopis, and G. Facciolo, "TV-L1 optical flow estimation," *Image Process. On Line*, vol. 3, pp. 137–150, Jul. 2013.

[38] R. Gao, B. Xiong, and K. Grauman, "Im2Flow: Motion hallucination from static images for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 5937–5947.
[39] D. Sun, X. Yang, M. Liu, and J. Kautz, "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 8934–8943.
[40] 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.*, New Orleans, LA, USA, 2018, pp. 2257–2264.
[41] I. Laptev and T. Lindeberg, "Space-time interest points," in *Proc. 9th IEEE Int. Conf. Comput. Vis. (ICCV)*, Nice, France, Oct. 2003, pp. 432–439.
[42] H. Wang, A. Kläser, C. Schmid, and C. Liu, "Action recognition by dense trajectories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Providence, RI, USA, Jun. 2011, pp. 3169–3176.
[43] A. Tang, K. Lu, Y. Wang, J. Huang, and H. Li, "A real-time hand posture recognition system using deep neural networks," *ACM Trans. Intell. Syst. Technol.*, vol. 6, no. 2, p. 21, May 2015.
[44] 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.
[45] 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.
[46] G. Zhu, "Multimodal gesture recognition using 3-D convolution and convolutional LSTM," *IEEE Access*, vol. 5, pp. 4517–4524, 2017.
[47] L. Zhang, G. Zhu, P. Shen, J. Song, S. A. Shah, and M. Bennamoun, "Learning spatiotemporal features using 3DCNN and convolutional LSTM for gesture recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Oct. 2017, pp. 3120–3128.
[48] C. Lin, X. Lin, Y. Xie, and Y. Liang, "Abnormal gesture recognition based on multi-model fusion strategy," *Mach. Vis. Appl.*, vol. 30, no. 5, pp. 889–900, 2019.
[49] Z. Cao, T. Simon, S. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 7291–7299.
[50] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 4645–4653.
[51] P. Marquez-Neila, L. Baumela, and L. Alvarez, "A morphological approach to curvature-based evolution of curves and surfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 2–17, Jan. 2014.

SHUJUN ZHANG is currently an Associate Professor with the College of Information Science and Technology, Qingdao University of Science and Technology, China. Her research directions include computer vision and virtual reality.

WEIJIA MENG is currently pursuing the master's degree with the College of Information Science and Technology, Qingdao University of Science and Technology, China. She is majoring in computer vision.

HUI LI is currently an Associate Professor with the College of Information Science and Technology, Qingdao University of Science and Technology, China. His research direction includes computer vision.

XUEHONG CUI is currently a Lecturer with the College of Information Science and Technology, Qingdao University of Science and Technology, China. Her research direction includes computer vision.