Computer Systems Science & Engineering
DOI: 10.32604/csse.2023.035119
Article

Tech Science Press

Multimodal Spatiotemporal Feature Map for Dynamic Gesture Recognition

Xiaorui Zhang[1,2,3,*], Xianglong Zeng[1], Wei Sun[3,4], Yongjun Ren[1,2,3] and Tong Xu[5]

[1]Engineering Research Center of Digital Forensics, Ministry of Education, Jiangsu Engineering Center of Network Monitoring, School of Computer and Software, Nanjing University of Information Science & Technology, Nanjing, 210044, China
[2]Wuxi Research Institute, Nanjing University of Information Science & Technology, Wuxi, 214100, China
[3]Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology (CICAEET), Nanjing University of Information Science & Technology, Nanjing, 210044, China
[4]School of Automation, Nanjing University of Information Science & Technology, Nanjing, 210044, China
[5]University of Southern California, Los Angeles, California, USA
*Corresponding Author: Xiaorui Zhang. Email: zxr365@126.com

Received: 08 August 2022; Accepted: 26 October 2022

Abstract: Gesture recognition technology enables machines to read human gestures and has significant application prospects in the fields of human-computer interaction and sign language translation. Existing researches usually use convolutional neural networks to extract features directly from raw gesture data for gesture recognition, but the networks are affected by much interference information in the input data and thus fit to some unimportant features. In this paper, we proposed a novel method for encoding spatio-temporal information, which can enhance the key features required for gesture recognition, such as shape, structure, contour, position and hand motion of gestures, thereby improving the accuracy of gesture recognition. This encoding method can encode arbitrarily multiple frames of gesture data into a single frame of the spatio-temporal feature map and use the spatio-temporal feature map as the input to the neural network. This can guide the model to fit important features while avoiding the use of complex recurrent network structures to extract temporal features. In addition, we designed two sub-networks and trained the model using a sub-network pre-training strategy that trains the sub-networks first and then the entire network, so as to avoid the sub-networks focusing too much on the information of a single category feature and being overly influenced by each other's features. Experimental results on two public gesture datasets show that the proposed spatio-temporal information encoding method achieves advanced accuracy.

Keywords: Dynamic gesture recognition; spatio-temporal information encoding; multimodal input; pre-training; score fusion

1 Introduction

Gestures are another common form of communication besides spoken expressions, which can express semantic information as well as convey emotions and attitudes [1]. The purpose of gesture recognition is to recognize human gestures through computer technology, and it has applications in scenes such as emotion recognition, sign language translation, human-computer interaction, virtual reality, and healthcare [2–4], especially today when the new coronavirus is rampant [5], which provides a healthy and friendly way of contactless human-computer interaction. Gesture recognition can be divided into wearable-based methods and computer vision-based methods, depending on the device used. Compared with the former, the latter is less costly and more user-friendly [6].

Gesture recognition based on computer vision [7] can be further divided into static gesture recognition and dynamic gesture recognition. Dynamic gesture recognition relies on extracting deep features from key information such as hand shape, structure, and hand movement trajectory, which requires the method used to extract the key information from input data first, and then further extract deep features from them for gesture recognition. In recent years, the rise of deep learning has brought new solutions to the field of computer vision [8,9]. In the field of dynamic gesture recognition, convolutional layers combined with recurrent neural network structures are often used to extract spatio-temporal features of raw data for gesture recognition [10–13]. Although these network structures can extract features from gesture data to a certain extent, they are affected by many distracting information in the input data, such as background noise, hand color, finger length, etc., and tend to fit to some unimportant features. Single-modal input data contains limited features, so multi-modal fusion is often used to extract more effective features. However, direct training of multi-modal networks can lead to excessive influence of sub-networks on each other. We can summarize our contribution points as follows:

- A novel method for encoding spatio-temporal information is proposed, which can encode arbitrary frames of skeletal data and depth images into skeletal spatio-temporal feature maps and depth spatio-temporal feature maps. This method can integrate and enhance the effective features in the input data, while reducing the amount of interference information, which in turn guides the neural network model to fit the key features required for gesture recognition.
- A convolutional neural network with skeletal spatio-temporal feature maps and depth spatio-temporal feature maps as dual input modalities is proposed. According to the characteristics of skeletal data and depth images, two sub-networks are designed and the whole network based on these two sub-networks is trained by the sub-network pre-training strategy to improve the accuracy and reliability of gesture recognition.
- The effectiveness of the proposed method was verified on dataset SHREC’17 and dataset DHG-14/8. Its accuracy is superior to the method using recurrent neural network structure.

The rest of this paper is organized as follows. Section 2 outlines the work related to this paper. Section 3 describes the details of the proposed method. Section 4 provides an experiment of the effectiveness of the proposed method. Section 5 summarizes the work done in this paper and the direction of future work.

2 Related Works

This section details the data selection related to gesture recognition and research using deep learning techniques in the feature extraction and classification stages.

2.1 Data for Gesture Recognition

To be able to improve the accuracy of gesture recognition, we need to select appropriate data as input in the data collection stage [14]. Wearable devices-based gesture recognition will use various types of sensors to capture human motion signals. For example, [6] used the metacarpophalangeal (MCP) and proximal interphalangeal (PIP) joint angles of five fingers captured by a soft sensor-based embedded data glove as the input to the model. [15] developed a wearable surface electromyography (EMG) biosensing system with adaptive learning capability based on screen-printed conformal electrode arrays and implemented a neuro-inspired algorithm for real-time gesture classification. In [16], the starting point is to ensure the reliability and validity of the myoelectric sign recognition system by using linear discriminant analysis (LDA) and extreme learning machine (ELM) to reduce the redundant information of the surface EMG signal. Although wearable device-based methods can achieve good accuracy, they are not user-friendly due to affecting human motion.

Vision-based gesture recognition uses cameras to capture data, which allows users to express gestures more freely without being limited in movement by hardware, wiring, etc. [17] proposed a gesture feature extraction and recognition method based on image processing, which takes the original gesture image as the input of the model, and filters the noise in the image by denoising, binarizing, expanding and eroding it. Thereby, the gesture features are obtained. [18] constructed an automatic Arabic sign language recognition framework using RGB hand images and body skeletal data as inputs. [19] used Kinect cameras to capture both RGB and depth images, extracted the two sets of features separately using multiple networks and finally combined the two obtained features for classification. [20] also used the Kinect cameras to acquire the human skeleton, based on which an algorithm for the analysis and extraction of key points of the hand skeleton was proposed for gesture recognition. It is evident from the previous discussion that RGB images, skeletal data and depth images are commonly used data types for gesture recognition, while skeletal data and depth images have less background noise compared to RGB images.

2.2 Feature Extraction

In recent years, deep learning has made outstanding achievements in computer field, which has greatly promoted the development of computer vision field [21,22] has proposed a recurrent neural network structure that can simultaneously capture the spatial coherence among joints and the temporal evolution among skeletons, and can accurately predict human motion [23] uses deep convolution neural network to recognize people's emotions, so that people can objectively understand their real-time emotional state. More and more researchers choose to use deep learning as a technical direction for gesture recognition. [24] applied a transfer learning strategy to the convolutional neural network AlexNet to achieve faster feature learning for gesture recognition and used the Artificial bee colony optimization algorithm to optimize the hyperparameters. [25] proposed the first network with joint-aware features for both gesture recognition and 3D hand pose estimation, which can recognize and estimate both gestures and 3D hand pose using joint-aware features. In [26], to avoid manual segmentation of hand images with background noise, convolutional neural network based on a deep parallel structure is proposed, which can perform gesture recognition on images containing various confusing backgrounds. However, using neural network to process raw data directly, it is easy to fit unimportant features.

2.3 Multi-Mode Fusion

It was found that multimodal inputs tend to significantly improve the accuracy of gesture recognition [10,27] proposed an adaptive fusion method with multimodal weights considering the correlation between multimodalities. [28] decoupled hand gestures into two features, hand posture change and hand motion. And correspondingly, an end-to-end dual-stream network was proposed to model these two hand features, and finally performed gesture recognition by aggregating these two features. [29] proposed an end-to-end network, in which two pre-trained networks are fused using a score fusion technique. [30] proposed a multimodal input model for isolated sign language recognition. First, four categories of spatial features were obtained from each of RGB images and depth images, and the extracted features were used as the input of the convolutional neural network, which fused eight categories of spatial features. Then, temporal features were extracted by long short-term memory(LSTM). Finally, classification is performed to obtain accuracy comparable to state-of-the-art models. Unfortunately, they are all trained directly on the whole network model, ignoring the influence of the sub-networks on each other, which results in the sub-networks not fully extracting the features of individual modalities.

3 Proposed Method

In this section, the proposed method and model are described in detail. At the feature level, this paper proposes a spatio-temporal information encoding method that enhances the features required for gesture recognition in skeletal data and depth images; at the model level, a dual-stream input model consisting of two sub-networks with different structures to serve different input types is designed; at the training strategy level, a sub-network pre-training strategy is applied. Specific details are presented in each subsection below.

3.1 Overall Framework

As shown in Fig. 1, our proposed framework consists of four main components, which are the depth spatio-temporal feature map, skeletal spatio-temporal feature map, depth sub-network and skeletal sub-network. The first two of them are used as inputs of the latter two respectively, and the latter two are used to extract the features in the first two. Finally, the two sub-networks use score fusion and input into softmax function to obtain gesture recognition results.

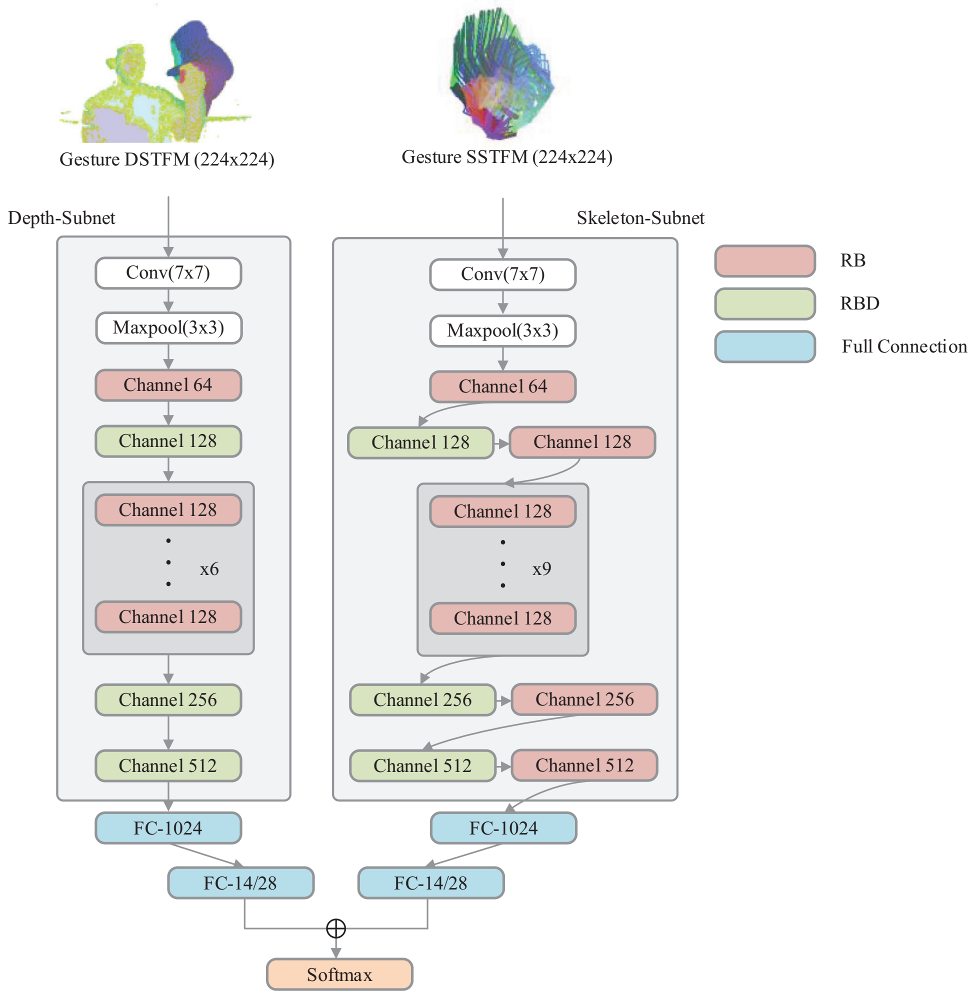


Figure 1: The overall structure of the model. The model is composed of depth subnet (left) and skeletal subnet (right). Their inputs are depth spatio-temporal feature map and skeletal spatio-temporal feature map respectively. The channel in the figure represents the number of output channels of convolution

3.2 Spatiotemporal Feature Map

Studies have shown that the shape, structure, contour, and motion trajectory of the hand are crucial features to distinguish different gestures, where the motion trajectory is mainly reflected in the spatial position of the hand at each frame. Using 2D convolution directly in neural networks to process raw gesture data is easy to fit to unimportant features, and it is difficult to process information in the temporal dimension. The spatio-temporal information encoding method proposed in this paper aims to enhance the aforementioned hand features in dynamic gestures and map them to a single-frame three-channel 2D spatio-temporal feature map. Thus, 2D convolutional layers can be fitted to the key features for gesture recognition, while avoiding the use of complex recurrent neural networks to extract temporal features.

3.2.1 Skeletal Spatiotemporal Feature Map

Skeletal data can clearly represent the spatial position of each joint of the hand in each frame, and there is not so much background noise that affects the recognition accuracy as in RGB images. We construct a skeletal spatio-temporal feature map by parsing and mapping 2D hand skeletal key point information and fusing skeletal, temporal and shape features in skeletal data.

The 2D hand skeletal key points record the 2D coordinates of multiple key points of the hand in 2D space, where n is used to denote the number of frames of skeletal data. For the sake of description, the connections of skeletal key points in the skeletal channel are referred as skeletal connections and the connections of the skeletal key points in the shape channel are referred as shape connections.

First, connect the key points of the bones according to the connection method of the hand bones. While connecting the hand key points into lines, each skeletal connection is assigned a different value using Eq. (1), thereby distinguishing different bones. By this method the skeletal channel is obtained for each frame.

$$F_o^s = (o+1)/s * 150 + 50 \tag{1}$$

where $\mathrm{F}_\mathrm{o}^\mathrm{s}$ denotes the value of the No. o skeletal connection and s denotes the total number of skeletal connections.

The shape channel mainly presents the shape of the hand in each frame, which connects the fingertips of each finger and wrist joints to form a closed geometry. Similarly, the shape channel for each frame is constructed by assigning different values to each shape connection using Eq. (2).

$$F_o^h = (o+1)/h * 150 + 50 \tag{2}$$

where $\mathrm{F}_\mathrm{o}^\mathrm{h}$ denotes the value of the No. o shape connection and s denotes the total number of skeletal connections.

Finally, the temporal channel of frame i is constructed by taking i/n as the temporal weight and multiplying this value by 255 as the value of all skeletal connections and shape connections in that frame.

Finally, the skeletal channel, temporal channel and shape channel of each frame are superimposed in time order. Taking frame i of the skeletal channel as an example. Firstly, frame i is binarized with 240 thresholds to get the mask of skeletal channel of frame i. The superimposed result of the previous frame is bitwise summed with mask, and frame i is bitwise summed with inverse of mask. Then Pixel-by-pixel summation of the two results is performed to obtain the superposition result of the current frame. This operation is repeated for frames 1 to n to obtain the superimposed skeletal channel, and for frame 1, the superimposed result of the previous frame is set as a single-channel image with each pixel value of 255. The three single-channel images obtained from the final superposition calculation are stacked to form the skeletal spatio-temporal feature map, as in Fig. 2.

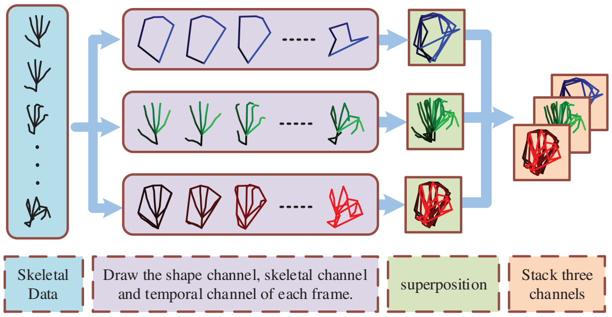

Skeletal Data
Draw the shape channel, skeletal channel and temporal channel of each frame.
superposition
Stack three channels

Figure 2: Skeleton spatiotemporal feature map construct process

3.2.2 Depth Spatiotemporal Feature Map

Depth images can well describe the area and shape of body parts, but does not specify the location of each part. Combining depth images with skeletal data can effectively improve the recognition accuracy of the model. We also enhance the location of different regions in the depth image so that the neural network can effectively detect the information of different regions. Based on depth information, temporal information and contour information, we further propose a deep spatio-temporal feature map.

The depth image itself is a single-channel image, and the contour extraction of the depth image is performed using the Canny algorithm [31], which can extract the image contour well by Gaussian filtering, normalization, calculation of gradient magnitude and direction, non-extreme suppression, and double-threshold judgment operations. We use the extracted contours as the contour channel for each frame.

Since the depth image itself has a small value, we multiply each pixel of the depth image by 100 to enhance the distinction between different regions and use it as a depth channel. Like the temporal channel of the skeletal spatio-temporal feature map, i/n is used as the temporal weight of the time channel in frame i. The difference is that the depth image is binarized here, and $i/n \times 255$ is assigned to the human body region in the depth image, while the non-human body region is assigned 0. The depth channel, temporal channel, and contour channel of each frame are stacked to form the depth spatio-temporal feature map.

3.3 Subnet Structure

The model proposed in this paper is shown in Fig. 1. The whole model consists of two sub-networks: the depth sub-network and the skeletal sub-network. The inspiration of the two sub-networks comes from Resnet-18 [32]. The whole network is mainly composed of two modules, as shown in Fig. 3 below. The residual-block (RB) is composed of two 3×3 convolutions connected in series, and a shortcut indicates that the input is added with the results of the two convolutions before output. The difference between the residual-block with downsampling (RBD) and the residual-block is that the first 3×3 convolution stride of the former has changed from 1 to 2, and there is also an additional 1×1 convolution with a stride of 2 for downsampling on the shortcut. The input size of the two subnets is 224×224. The first two layers are a convolution layer with a size of 7×7, stride of 2 and a padding of 3. A pooling layer size is 3×3, stride of 2 and a padding of 1.

The main difference between these two subnets is the different number of RB and RBD used, RB and RBD are represented by two different colored rectangles on Fig. 1, where the channels depicted in the figure indicates the number of output channels of the convolutional layer, i.e., the number of convolutional kernels. In addition, Gaussian Error Linear Unit(GELU) and Rectified Linear Unit(RELU) are commonly used activation functions in neural networks. They have different sensitivity to different data, depth subnet uses GELU as the activation function after each convolution, while skeletal subnet uses RELU as the activation function after each convolution.

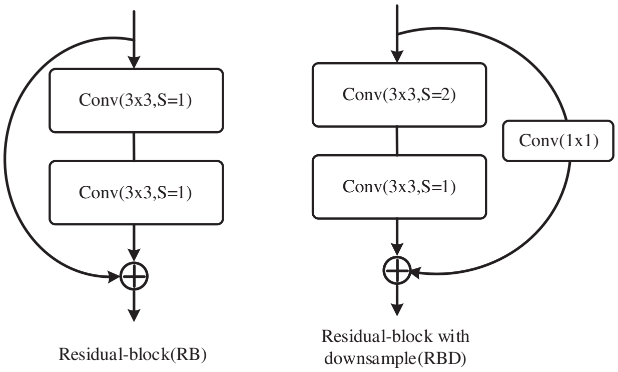


Figure 3: Residual block (RB) and residual block with downsampling(RBD) are the basic blocks of depth subnet and skeletal subnet, where s represents stride

3.4 Pre-Training Strategy

Because of the different inputs of the two sub-networks, they need to pay attention to different features. When these two sub-networks are trained separately, they do not influence each other and focus on their own parts, but when they are combined into a complete trained model, they will have a certain influence on each other's parameters. That is to say, the weights obtained by training the two models separately and obtained by directly training the complete model are different. In a word, the sub-networks trained separately are more focused on their own fields, and more complementary information will be considered between sub-networks when training the complete model directly.

In this paper, we first train the two sub-networks separately, so that the two sub-networks can fully extract the features of their respective input modalities. The complete model is further trained with the respective weights and biases of the two subnetworks, so that these two subnetworks can consider each other's information to some extent, while fully extracting the features of their respective input modalities, for example, allowing the skeletal subnetwork to analyze the skeletal structure with reference to the contour information. Such a training method not only allows the sub-network not to focus too much on itself and cut off from the other sub-network, but also does not let the two sub-networks have too much influence on each other and lose effective information, so as to obtain a model with higher accuracy.

4 Experiments and Results

In this section, the proposed method is evaluated. Firstly, the datasets used for the experiments are presented. Then, the method is compared with current state-of-art methods on two datasets. Finally, various modules of the proposed method are validated, including different coding schemes, two sub-networks, and the application of pre-training strategy.

4.1 Training Details

To alleviate the overfitting problem, we used a random cropping method to implement data enhancement on the spatio-temporal feature maps. The proposed model was constructed using PyTorch framework [33] with both depth spatio-temporal feature map and skeletal spatio-temporal feature map as inputs to the model, both with a resolution of 224 × 224. For the whole model, we used Stochastic Gradient Descent (SGD) as the optimizer for an initial learning rate of 0.005, a momentum of 0.9, a weight decay of

0.00001, epoch of 100, and small batch of 32. Cross-entropy loss function can help model learn information between classes, so the network is trained to minimize the cross-entropy loss between the predicted and actual labels during the training process.

4.2 Comparison with State-of-the-Art Methods

We compared the proposed method and model on the SHREC'17 dataset and the DHG-14/28 dataset, respectively, with some models that have recently achieved advanced results. These two datasets were chosen because they are publicly available and provide skeletal data and depth data for easy evaluation of the proposed models and methods, and they are also two challenging datasets.

4.2.1 SHREC'17

SHREC'17 dataset [34] is a dataset containing 14 gestures, which can be divided into coarse-grained gestures and fine-grained gestures according to gesture characteristics, as shown in Table 1 Among them, each gesture is performed by a single finger and the whole hand, so it can be further divided into 28 gestures. Each gesture was performed 1 to 10 times by 28 participants in two ways, and then 2800 frame sequence data were obtained. Each frame included skeletal data and depth images, in which the skeletal data contained the 2D and 3D spatial coordinates of 22 joints in each frame. The depth image size was 640 × 480 pixels, and the length of each sample ranged from 20 to 50 frames. In this paper, only 2D skeletal data and depth images are used.

Table 1: List of the 14 gestures in SHREC'17 dataset

Class	Name of the gesture	Type of the gesture
1	Grab	Fine
2	Tap	Coarse
3	Expand	Fine
4	Pinch	Fine
5	Rotation CW	Fine
6	Rotation CCW	Fine
7	Swipe right	Coarse
8	Swipe left	Coarse
9	Swipe up	Coarse
10	Swipe down	Coarse
11	Swipe X	Coarse
12	Swipe V	Coarse
13	Swipe +	Coarse
14	Shake	Coarse

For all experiments conducted on the SHREC'17 dataset, we followed the data splitting scheme given in the dataset: 70% of the samples as the training set and the rest as the test set, i.e., 1960 samples as the training set and 840 samples as the test set, which is also the same data splitting scheme used in those studies [31,35–39] that we want to compare.

The results of the comparison experiments on SHREC'17 are shown in Table 2. Compared with methods in the table, the method proposed in this paper outperforms all the methods on 14 gestures, where [39] uses a bi-directional long and short-term memory network (Bi-LSTM) to extract temporal features. It can be seen that our method has the same effect as that using Bi-LSTM, although the accuracy is slightly lower than that method on 28 gestures. This indicates that our proposed method and model are indeed capable of extracting features in the temporal dimension without using recurrent neural networks and their variants, but the deep feature extraction capability leaves something to be desired.

Table 2: Comparison of top-1 recognition rates (%) with state-of-the-art methods on SHREC'17 dataset

Method	Modality	14 Gesture	28 Gesture
Key frames [31]	Depth sequence	82.9	71.9
SoCJ + HoHD + HoWR [35]	3D-Skeleton	88.24	81.90
MFA-Net [36]	3D-Skeleton	91.31	86.55
M-sihrhs [37]	3D-Skeleton	92.14	85.69
ST-GCN [38]	3D-Skeleton	92.7	87.7
gVar-FL-fusion [39]	Depth sequence + 2D-skeleton	93.33	**90.24**
Ours	Depth sequence + 2D-skeleton	**93.45**	88.57

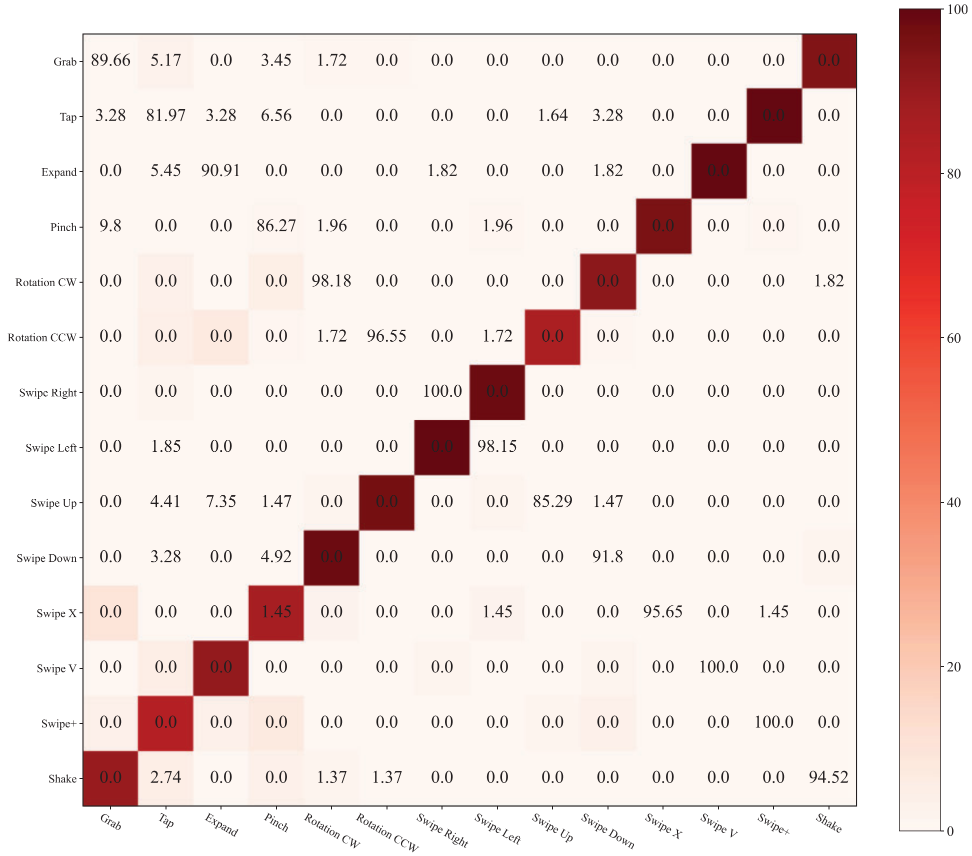
A closer look reveals that the use of multimodal data is more effective than single class data, and skeletal data is more effective than depth data. This may be because depth data is more ambiguous and more noisy than skeletal data, and the combination of the two can provide complementary information to some extent. In addition, the recognition accuracy of 28 class gestures is always lower than that of 14 class gestures, because 28 class gestures require higher details of features. Figs. 4 and 5 show the confusion matrices of 14 gestures and 28 gestures on the SHREC'17 dataset, respectively, and overall satisfactory results are achieved. However, the recognition of some similar gestures could be improved, for example, the gesture Pinch could be easily mistaken for Grab.

4.2.2 DHG-14/28

To validate the generalization performance of the model, we also validated the proposed method and model on DHG-14/28 dataset [35]. DHG-14/28 dataset is like SHREC'17 dataset, and it is also a dataset containing 14 gestures. Each gesture was performed five times by 20 participants in two ways, and a total of 2,800 frame sequences were obtained. The data type and collection method are the same as those of SHREC'17 Dataset.

The experiment on DHG-14/28 dataset adopts the same data splitting scheme as those methods to be compared [13,36,40,41]. Each time, the data of 19 participants is used as the training set, and the data of the remaining one participant is used as the test set. By setting different participants as the test set, 20 experiments are conducted, and the average value is used result.

The results of the comparison experiments on the DHG-14/28 dataset are shown in Table 3 The recognition accuracy of the proposed method in this paper is 87.14% for 14 gestures and 84.36% for 28 gestures, and it can also be seen that our method outperforms most dynamic gesture recognition methods using LSTM. To further observe the cross-validation results, we visualized the cross-validation of 14 gesture and 28 gesture, as shown in Figs. 6 and 7. It can be seen from the figure that good recognition results can be achieved for most subjects. But whether it is 14 gestures or 28 gestures, the gesture of No.2 subject is always the most difficult to recognize, followed by No.6. It may be necessary to improve the generalization performance of the model to alleviate this kind of problem of insufficient recognition accuracy caused by the large deviation of the speaker's gesture from the standard. At the same time, the cross-validation results also verify the conclusion in Section 4.4.2, that is, gesture recognition by two sub-networks alone can also achieve good accuracy, but it is not as good as that by combining the two sub-networks. Here, the accuracy of the depth sub-network is slightly better than that of the skeletal sub-network, while the performance of the skeletal sub-network is a little better in the experiment in Section 4.4.2, which shows that the two sub-networks have different performances on different datasets. These two can make up for each other's shortcomings to a certain extent.



	Grab	Tap	Expand	Pinch	Rotation CW	Rotation CCW	Swipe Right	Swipe Left	Swipe Up	Swipe Down	Swipe X	Swipe V	Swipe+	Shake
Grab	89.66	5.17	0.0	3.45	1.72	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Tap	3.28	81.97	3.28	6.56	0.0	0.0	0.0	0.0	1.64	3.28	0.0	0.0	0.0	0.0
Expand	0.0	5.45	90.91	0.0	0.0	0.0	1.82	0.0	0.0	1.82	0.0	0.0	0.0	0.0
Pinch	9.8	0.0	0.0	86.27	1.96	0.0	0.0	1.96	0.0	0.0	0.0	0.0	0.0	0.0
Rotation CW	0.0	0.0	0.0	0.0	98.18	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.82
Rotation CCW	0.0	0.0	0.0	0.0	1.72	96.55	0.0	1.72	0.0	0.0	0.0	0.0	0.0	0.0
Swipe Right	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Swipe Left	0.0	1.85	0.0	0.0	0.0	0.0	0.0	98.15	0.0	0.0	0.0	0.0	0.0	0.0
Swipe Up	0.0	4.41	7.35	1.47	0.0	0.0	0.0	0.0	85.29	1.47	0.0	0.0	0.0	0.0
Swipe Down	0.0	3.28	0.0	4.92	0.0	0.0	0.0	0.0	0.0	91.8	0.0	0.0	0.0	0.0
Swipe X	0.0	0.0	0.0	1.45	0.0	0.0	0.0	1.45	0.0	0.0	95.65	0.0	1.45	0.0
Swipe V	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0
Swipe+	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0
Shake	0.0	2.74	0.0	0.0	1.37	1.37	0.0	0.0	0.0	0.0	0.0	0.0	0.0	94.52

Figure 4: The 14 gesture confusion matrix of the proposed approach for SHREC'17

4.3 *Ablation Studies*

To verify the effectiveness of proposed method and model, we verify the different coding schemes of spatio-temporal feature map, two sub-networks of proposed model and pre-training strategy on SHREC'17 dataset. The evaluation criteria are top-1 accuracy and top-5 accuracy. Top-1 accuracy refers to the probability that the item with the largest predicted value is the correct class, and top-5 accuracy refers to the probability that the top five items with the largest predicted value contain the correct class.

4.3.1 Different Coding Schemes

Firstly, the skeletal, temporal and shape features of the skeletal spatio-temporal feature map are input into the skeletal sub-network in different combinations to verify the effectiveness of each coding scheme of the skeletal spatio-temporal feature map. The experimental results are shown in Table 4.

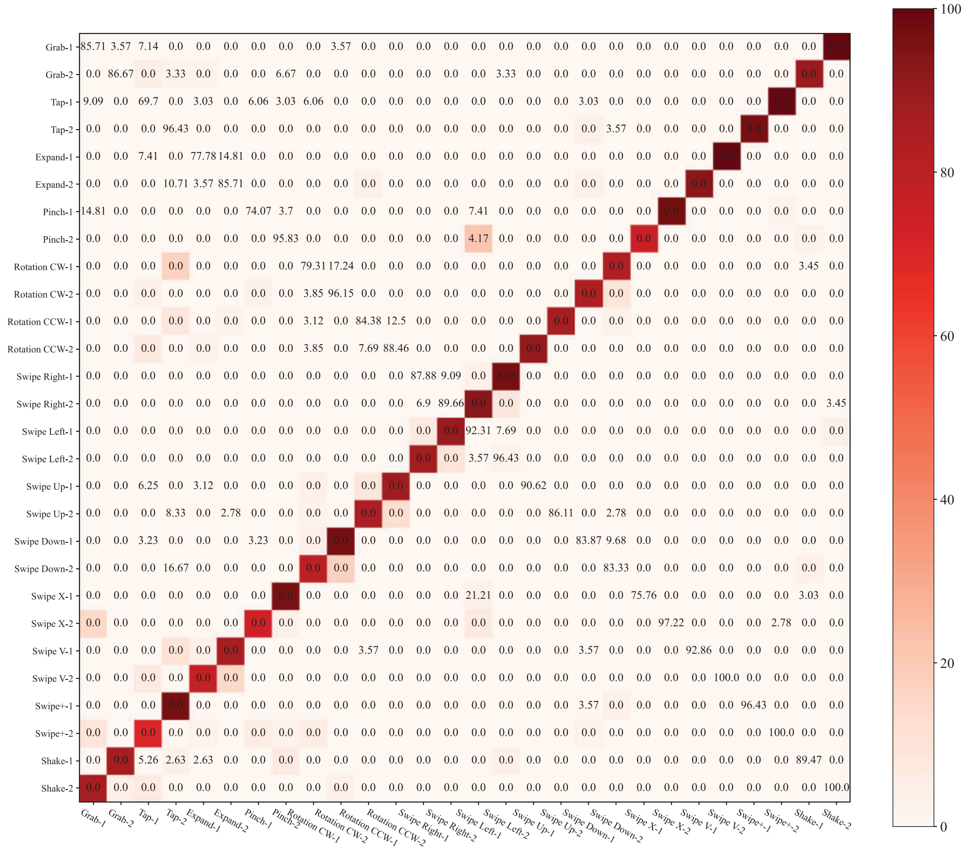


Figure 5: The 28 gesture confusion matrix of the proposed approach for SHREC'17

Table 3: Comparison of recognition rates (%) with state-of-the-art methods on DHG-14/28 dataset

Method	14 Gesture	28 Gesture
RNN + Motion feature [40]	84.68	80.32
3D CNN + LSTM [13]	85.6	81.1
MFA-Net [36]	85.75	81.04
Smedt et al. [41]	86.86	84.22
Ours	**87.14**	**84.36**

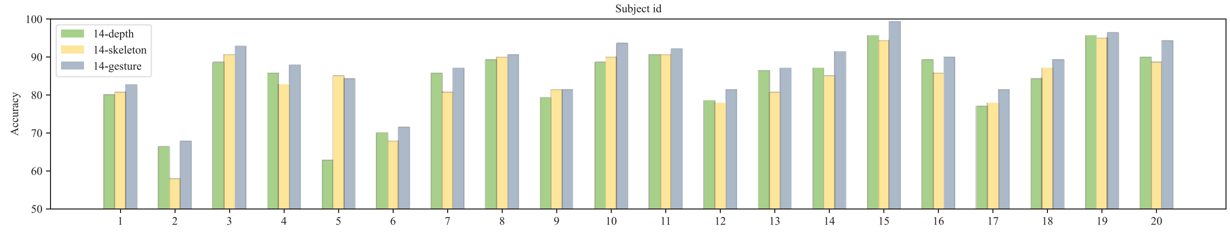


Figure 6: Cross-validation results of 14 gesture on DHG-14/28 dataset. '14-depth' means the recognition result of the depth subnet, '14-skeleton' means the recognition result of the skeletal subnet, and '14-gesture' means the recognition result of the complete network

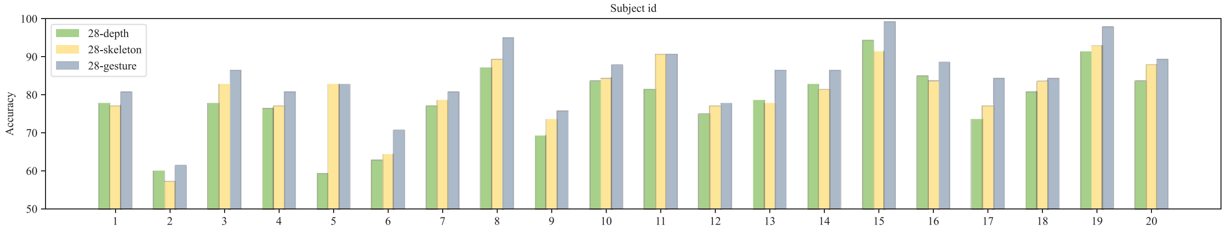


Figure 7: Cross validation results of 28 gestures on DHG-14/28 dataset. '28-depth' means the recognition result of the depth subnet, '28-skeleton' means the recognition result of the skeletal subnet, and '28-gesture' means the recognition result of the complete network

Table 4: Comparison of the different skeleton encoding schemes on the SHREC'17 dataset in terms of recognition accuracy

Channels	14-top1 (%)	14-top5 (%)	28-top1 (%)	28-top5 (%)
skeletal	81.90	98.45	76.90	95.23
temporal	86.07	97.62	81.43	95.95
shape	79.52	97.38	75.00	96.07
skeletal + temporal	89.00	98.57	83.21	97.74
shape + temporal	88.10	99.05	84.14	**98.69**
skeletal + shape	82.50	97.98	78.93	96.55
skeletal + temporal + shape	**89.17**	**99.29**	**84.52**	98.33

As can be seen from Table 4 we have conducted seven groups of experiments to verify the effect of the coding scheme of the skeletal spatio-temporal feature map, and each group of experiments has tested 14 gestures and 28 gestures respectively. Each coding scheme of the skeletal spatio-temporal feature map has achieved good results, and the temporal channel is the best among the three channels. When skeletal, temporal and shape are combined, the overall effect is the best in all groups. The above experiments show that our proposed skeletal spatio-temporal feature map is effective for gesture recognition.

Similarly, we use the depth sub-network to verify the different combinations of depth, temporal and contour features of the depth spatio-temporal feature map. The experimental results are shown in Table 5.

Table 5: Comparison of the different depth encoding schemes on the SHREC'17 dataset in terms of recognition accuracy

Channels	14-top1 (%)	14-top5 (%)	28-top1 (%)	28-top5 (%)
depth	70.71	95.48	60.83	88.69
temporal	83.45	98.13	72.62	96.79
contour	71.19	93.69	62.14	89.52
depth + temporal	85.24	97.68	74.76	97.14
contour + temporal	85.12	98.05	74.40	96.43
depth + contour	75.71	95.71	69.64	93.93
depth + temporal + contour	85.95	98.69	74.76	95.24

It can be seen from the table that, like the skeletal spatio-temporal map, the temporal channel is more effective than the other two channels, almost reaching the effect when the three channels are combined. The combination of temporal channel and depth channel is similar to that of temporal channel and contour channel. The best effect still appears when the three channels are combined, which shows that each channel of the depth spatio-temporal feature map can effectively improve the accuracy of gesture recognition.

4.3.2 Subnet

We use different parameters to train the two sub-networks. For the skeletal sub-network, Adam is used as an optimizer to train quickly with a learning rate of 0.0001. For the depth sub-network, SGD optimizer is used to train with a learning rate of 0.05, a momentum of 0.9 and a weight of 0.00001, with both batch size of 32 and epoch of 200.

The validation results of two sub-networks are shown in Table 6 From the table, it is obvious that a single sub-network can also achieve good accuracy, especially the skeletal sub-network, but it is not as good as the effect of using two sub-networks at the same time. If the two sub-networks are combined, the accuracy is obviously improved.

Table 6: Comparison of the different net on the SHREC'17 Dataset in terms of recognition accuracy

Net	14-top1 (%)	14-top5 (%)	28-top1 (%)	28-top5 (%)
skeletal subnet	89.17	99.29	84.52	98.33
depth subnet	85.95	98.69	74.76	95.24
skeletal subet + depth subnet	93.45	99.64	88.57	99.17

4.3.3 Pre-Training Strategy

We apply the sub-network pre-training strategy to the complete model. First, we train the two sub-networks separately, and save their weights with the best effect. Then, we load their weights to train the complete network. The experimental results are shown in Table 7. From the table, it is obvious that the accuracy of top1 without pre-training strategy is 92.02%, and that of top1 with pre-training strategy is 93.45%. The application of pre-training strategy improves the accuracy of our model by 1.5%, while it improves the accuracy of 28 gesture by nearly 3%.

Table 7: Comparison of the different training strategy on the SHREC'17 Dataset in terms of recognition accuracy

Strategy	14-top1(%)	14-top5 (%)	28-top1 (%)	28-top5 (%)
without pre-training strategy	92.02	99.17	85.95	98.10
with pre-training strategy	93.45	99.64	88.57	99.17

5 Conclusion

In this paper, our main work is to propose a new spatio-temporal information coding method and a multi-modal neural network model. The proposed spatio-temporal information coding method can map the skeletal data and depth images of any frame into a single-frame three-channel spatio-temporal feature map to strengthen the spatio-temporal features needed for gesture recognition, thus avoiding the network fitting to unimportant features. The proposed neural network model consists of two sub-networks, which have different structures to fully extract the features of the two types of inputs. At the same time, we apply a sub-network pre-training strategy when training the network, that is, training the sub-network first and then training the complete network. Experiments on datasets SHREC'17 and DHG-14/28 show that the proposed spatio-temporal feature map, two sub-networks and pre-training strategy of sub-networks are all effective, and the complete model achieves the same accuracy as the model using recurrent neural network.

However, from the experimental results, it can also be seen that the effect of the depth spatio-temporal feature map on the SHREC'17 dataset is not as good as that of the skeletal spatio-temporal feature map, which may be mainly due to the fact that the thresholds of the depth maps of different subject in the dataset are not exactly the same and there is a small amount of noise, resulting in over-fitting of the network. Therefore, we may look for a better depth spatio-temporal coding scheme in the future. We may need to further improve the generalization performance of the model for the problem of the poor recognition effect of some subject in the dataset, such as pre-training on larger datasets, and then applying the transfer learning strategy on the target datasets. The proposed method has a certain limitation on the number of video frames, if the number of frames is too high, the feature map will be too chaotic, which may require more research on frame sampling in the future. In addition, we can also study better model structure, such as reducing features by 1×1 convolution and extracting features by convolution of different receptive fields, to further reduce the computation and improve the accuracy.

Acknowledgement: We are grateful to Nanjing University of Information Science and Technology for providing research environment and computing equipment.

Funding Statement: This work was supported, in part, by the National Nature Science Foundation of China under grant numbers 62272236; in part, by the Natural Science Foundation of Jiangsu Province under grant numbers BK20201136, BK20191401; in part, by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD) fund.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] G. Devineau, W. Xi, F. Moutarde and J. Yang, "Deep learning for hand gesture recognition on skeletal data," in *Proc. of the 13th IEEE Int. Conf. on Automatic Face & Gesture Recognition (FG 2018)*, Xi'an, China, pp. 106–113, 2018.

[2] X. R. Zhang, W. Z. Zhang, W. Sun and A. G. Song, "A new soft tissue deformation model based on runge-kutta: Application in lung," *Computers in Biology and Medicine*, vol. 148, pp. 105811–105822, 2022.

[3] M. Oudah, A. Al-Naji and J. Chahl, "Hand gesture recognition based on computer vision: A review of techniques," *Journal of Imaging*, vol. 6, no. 8, pp. 1–29, 2020.

[4] B. K. Chakraborty, D. Sarma, M. K. Bhuyan and K. F. MacDorman, "Review of constraints on vision-based gesture recognition for human-computer interaction," *IET Computer Vision*, vol. 12, no. 1, pp. 3–15, 2018.

[5] X. R. Zhang, J. Zhou, W. Sun and S. Jha, "A lightweight CNN based on transfer learning for COVID-19 diagnosis," *Computers, Materials & Continua*, vol. 72, no. 1, pp. 1123–1137, 2022.

[6] M. Lee and J. Bae, "Deep learning based real-time recognition of dynamic finger gestures using a data glove," *IEEE Access*, vol. 8, pp. 219923–219933, 2020.

[7] X. R. Zhang, W. Z. Zhang, W. Sun, H. L. Wu, A. G. Song *et al.,* "A Real-time cutting model based on finite element and order reduction," *Computer Systems Science and Engineering*, vol. 43, no. 1, pp. 1–15, 2022.

[8] M. Asadi-Aghbolaghi, A. Clapes, M. Bellantonio, H. J. Escalante, V. Ponce-Lopez *et al.,* "A survey on deep learning based approaches for action and gesture recognition in image sequences," in *Proc. of the 12th IEEE Int. Conf. on Automatic Face and Gesture Recognition*, Washington, DC, USA, pp. 476–483, 2017.

[9] W. Sun, G. Z. Dai, X. R. Zhang, X. Z. He and X. Chen, "TBE-Net: A three-branch embedding network with part-aware ability and feature complementary learning for vehicle re-identification," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 9, pp. 14557–14569, 2022.

[10] D. H. Fan, H. J. Lu, S. G. Xu and S. Cao, "Multi-task and multi-modal learning for RGB dynamic gesture recognition," *IEEE Sensors Journal*, vol. 21, no. 23, pp. 27026–27036, 2021.

[11] S. A. Alameen and A. M. Alhothali, "A lightweight driver drowsiness detection system using 3DCNN with lSTM," *Computer Systems Science and Engineering*, vol. 44, no. 1, pp. 895–912, 2022.

[12] S. Ameur, A. Ben Khalifa and M. S. Bouhlel, "A novel hybrid bidirectional unidirectional LSTM network for dynamic hand gesture recognition with leap motion," *Entertainment Computing*, vol. 35, pp. 100373, 2020.

[13] J. C. Núñez, R. Cabido, J. J. Pantrigo, A. S. Montemayor and J. F. Vélez, "Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition," *Pattern Recognition*, vol. 76, pp. 80–94, 2018.

[14] M. Yasen and S. Jusoh, "A systematic review on hand gesture recognition techniques, challenges and applications," *PeerJ Computer Science*, vol. 2019, no. 9, pp. 1–30, 2019.

[15] A. Moin, A. Zhou, A. Rahimi, A. Menon, S. Benatti *et al.,* "A wearable biosensing system with in-sensor adaptive machine learning for hand gesture recognition," *Nature Electronics*, vol. 4, no. 1, pp. 56–63, 2020.

[16] J. X. Qi, G. Z. Jiang, G. F. Li, Y. Sun and B. Tao, "Intelligent human-computer interaction based on surface EMG gesture recognition," *IEEE Access*, vol. 7, pp. 61378–61387, 2019.

[17] Y. N. Wang, Y. M. Yang and P. Y. Zhang, "Gesture feature extraction and recognition based on image processing," *Traitement du Signal*, vol. 37, no. 5, pp. 873–880, 2020.

[18] M. A. Bencherif, M. Algabri, M. A. Mekhtiche, M. Faisal, M. Alsulaiman *et al.,* "Arabic sign language recognition system using 2D hands and body skeleton data," *IEEE Access*, vol. 9, pp. 59612–59627, 2021.

[19] W. T. Cheng, Y. Sun, G. F. Li, G. Z. Jiang and H. H. Liu, "Jointly network: A network based on CNN and RBM for gesture recognition," *Neural Computing and Applications*, vol. 31, pp. 309–323, 2019.

[20] D. Jiang, G. F. Li, Y. Sun, J. Y. Kong and B. Tao, "Gesture recognition based on skeletonization algorithm and CNN with ASL database," *Multimedia Tools and Applications*, vol. 78, no. 21, pp. 29953–29970, 2018.

[21] W. Sun, L. Dai, X. R. Zhang, P. S. Chang and X. Z. He, "RSOD: Real-time small object detection algorithm in UAV-based traffic monitoring," *Applied Intelligence*, vol. 52, no. 8, pp. 8448–8463, 2022.

[22] X. B. Shu, L. Y. Zhang, G. J. Qi, W. Liu and J. H. Tang, "Spatiotemporal co-attention recurrent neural networks for human-skeleton motion prediction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 3300–3315, 2021.

[23] D. Arnab. and D. Kousik, "Emotion recognition using deep learning in pandemic with real-time email alert," in *Proc. of Third Int. Conf. on Communication, Computing and Electronics Systems*, Singapore, Springer, vol. 844, pp. 175–190, 2022.

[24] T. Ozcan and A. Basturk, "Transfer learning-based convolutional neural networks with heuristic optimization for hand gesture recognition," *Neural Computing and Applications*, vol. 31, no. 12, pp. 8955–8970, 2019.

[25] S. Y. Yang, J. Liu, S. J. Lu, M. H. Er and A. C. Kot, "Collaborative learning of gesture recognition and 3D hand pose estimation with multi-order feature analysis," in *Proc. European Conf. on Computer Vision*, Cham, Springer, vol. 12348, pp. 769–786, 2020.

[26] V. Adithya and R. Rajesh, "A deep convolutional neural network approach for static hand gesture recognition," *Procedia Computer Science*, vol. 171, pp. 2353–2361, 2020.

[27] H. J. Duan, Y. Sun, W. T. Cheng, D. Jiang, J. T. Yun *et al.,* "Gesture recognition based on multi-modal feature weight," *Concurrency and Computation: Practice and Experience*, vol. 33, no. 5, pp. 1–17, 2020.

[28] J. B. Liu, Y. C. Liu, Y. Wang, V. Prinet, S. M. Xiang *et al.,* "Decoupled representation learning for skeleton-based gesture recognition," in *Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp. 5750–5759, 2020.

[29] S. J. Prakash, A. J. Prakash, P. Pławiak and S. Samantray, "Real-time hand gesture recognition using fine-tuned convolutional neural network," *Sensors*, vol. 22, no. 3, pp. 1–14, 2022.

[30] R. Rastgoo, K. Kiani and S. Escalera, “Hand pose aware multimodal isolated sign language recognition,” *Multimedia Tools and Applications*, vol. 80, no. 1, pp. 127–163, 2021.

[31] J. Canny, “A computational approach to edge detection john,” *Readings in Computer Vision*, vol. 8, no. 6, pp. 184–203, 1987.

[32] K. M. He, X. Y. Zhang, S. Q. Ren and J. Sun, “Deep residual learning for image recognition,” in *Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 770–778, 2016.

[33] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury *et al.,* “PyTorch: An imperative style, high-performance deep learning library,” *Advances in Neural Information Processing Systems*, vol. 32, pp. 8026–8037, 2019.

[34] D. S. Quentin, W. Hazem, P. V. Jean, G. Joris, L. S. Bertrand *et al.,* “SHREC’17 track: 3D hand gesture recognition using a depth and skeletal dataset,” in *Proc. of the 10th Eurographics Workshop on 3D Object Retrieval*, Lyon, France, pp. 23–24, 2017.

[35] Q. De Smedt, H. Wannous and J. P. Vandeborre, “Skeleton-based dynamic hand gesture recognition,” in *Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Las Vegas, NV, USA, pp. 1206–1214, 2016.

[36] X. H. Chen, G. J. Wang, H. K. Guo, C. Zhang, H. Wang *et al.,* “MFA-Net: Motion feature augmented network for dynamic hand gesture recognition from skeletal data,” *Sensors (Switzerland)*, vol. 19, no. 2, pp. 239, 2019.

[37] Y. H. Zhang, Z. Y. Zhao, W. Li and L. X. Duan, “Multi-scale enhanced active learning for skeleton-based action recognition,” in *Proc. of the 2021 IEEE Int. Conf. on Multimedia and Expo (ICME)*, Shenzhen, China, pp. 1–6, 2021.

[38] S. J. Yan, Y. J. Xiong and D. H. Lin, “Spatial temporal graph convolutional networks for skeleton-based action recognition,” in *Proc. of the 32th AAAI Conf. on Artificial Intelligence*, New Orleans, LA, USA, vol. 32, 2018.

[39] H. Mahmud, M. M. Morshed and M. K. Hasan, “A deep learning-based multimodal depth-aware dynamic hand gesture recognition system,” *arXiv:2107.02543*, 2021.

[40] X. Chen, H. Guo, G. Wang and L. Zhang, “Motion feature augmented recurrent neural network for skeleton-based dynamic hand gesture recognition,” in *Proc. of the 2017 IEEE Int. Conf. on Image Processing (ICIP)*, Beijing, China, pp. 2881–2885, 2017.

[41] Q. De Smedt, H. Wannous and J. -P. Vandeborre, “Heterogeneous hand gesture recognition using 3D dynamic skeletal data,” *Computer Vision and Image Understanding*, vol. 181, pp. 60–72, 2019.