Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.047861

ARTICLE

HgaNets: Fusion of Visual Data and Skeletal Heatmap for Human Gesture Action Recognition

Wuyan Liang[1] **and Xiaolong Xu**[2,*]

[1]School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing, 210023, China

[2]School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing, 210023, China

*Corresponding Author: Xiaolong Xu. Email: xuxl@njupt.edu.cn

Received: 20 November 2023 Accepted: 04 March 2024 Published: 25 April 2024

ABSTRACT

Recognition of human gesture actions is a challenging issue due to the complex patterns in both visual and skeletal features. Existing gesture action recognition (GAR) methods typically analyze visual and skeletal data, failing to meet the demands of various scenarios. Furthermore, multi-modal approaches lack the versatility to efficiently process both uniform and disparate input patterns. Thus, in this paper, an attention-enhanced pseudo-3D residual model is proposed to address the GAR problem, called HgaNets. This model comprises two independent components designed for modeling visual RGB (red, green and blue) images and 3D skeletal heatmaps, respectively. More specifically, each component consists of two main parts: 1) a multi-dimensional attention module for capturing important spatial, temporal and feature information in human gestures; 2) a spatiotemporal convolution module that utilizes pseudo-3D residual convolution to characterize spatiotemporal features of gestures. Then, the output weights of the two components are fused to generate the recognition results. Finally, we conducted experiments on four datasets to assess the efficiency of the proposed model. The results show that the accuracy on four datasets reaches 85.40%, 91.91%, 94.70%, and 95.30%, respectively, as well as the inference time is 0.54 s and the parameters is 2.74M. These findings highlight that the proposed model outperforms other existing approaches in terms of recognition accuracy.

KEYWORDS

Gesture action recognition; multi-dimensional attention; pseudo-3D; skeletal heatmap

1 Introduction

With the advancement of artificial intelligence, human gesture action recognition [1] (GAR) plays an important role in facilitating gesture communication between deaf and healthy people, and has widespread applications in the fields of human-computer interaction [2]. Relative to action recognition, the GAR is more complicated, which often utilize visual RGB (red, green and blue) images and skeletal data. Hence, there are two categories: Visual-based gesture action recognition (VGAR) and skeleton-based gesture action recognition (SGAR). In the VGAR task, the RGB images contain rich texture information [3]. And 2D Convolutional Neural Networks (CNN) are usually employed to

extract visual image features [4]. For instance, the gesture team [5] introduced a GAR system using 2D CNN to encode gesture features from the hands and upper body. Compared to 2D CNN, 3D CNN considers the temporal information and is used to extract spatiotemporal features from multiple video images [6]. Li et al. [7] proposed a 3D CNN model to learn and extract the spatiotemporal features from gesture videos.

In the SGAR task, unlike RGB images that becomes impractical in low-light conditions, the skeletal data portrays hand pose information, which is robust to context noise [8]. Graph Convolutional Networks (GCN) are used to process skeletal data. For example, Amorim et al. [9] introduced a spatiotemporal GCN to model the relationship between skeletal joints on the continuous multiple frames. Tunga et al. [10] proposed a GCN model to extract skeletal graph features and a transformer model to explore the temporal dependence between frames. However, the SGAR method is largely influenced by the accuracy of hand key points [11]. What's more, although existing VGAR and SGAR studies have demonstrated significant results [12], these studies highlight independent modalities, and ignore the overall applicability of GAR. This challenge stems from the inherent differences between skeletal and visual information. Skeletal data are represented in irregular graphs while visual data are organized on a regular grid. Huang et al. [13] proposed a multi-channel 3D CNN model to extract discriminative spatial-temporal features from color information, depth clues, and joint positions. However, the distinct properties of these modes impede the efficient exchange and integration of multimodal knowledge through a deep learning network. Therefore, many of reaches focus on different deep networks to encode visual and skeletal information. For instance, Kumar et al. [14] proposed a 5-stack convolution neural network (5S-CNN) to represent motion images and a Bi-directional Long Short-Term Memory Network (Bi-LSTM) with the hybrid 5S-CNN to encode skeletal data. Although this method achieved better recognition results, it cannot decrease the number of parameters and computational costs.

In this work, we aim to exploit a deep learning model that can process uniform and distinct input patterns to capture multi-modal features. Firstly, an attention-enhanced pseudo-3D residual model (HgaNets) is suggested to integrate the visual and skeletal data. Unlike characterizing a skeleton, which simply represents time-dependent 2D joint/bone locations, we introduce the heatmap stacks of joints to form 3D skeletal heatmaps, which is generated by OpenPose [15] technique. Furthermore, we have developed a multi-dimensional attention mechanism for the pseudo-3D residual module to better capture important spatial, temporal, and channel information. Finally, our proposed model can be trained in an end-to-end manner, providing the versatility for processing uniform and separate input data. This adaptation makes it well suited for processing a variety of multimodal information. This model allows us to use smaller networks (2.74M) and attains excellent accuracy on four datasets (85.40%, 91.91%, 94.70%, and 95.30%). Our main contributions are the following:

1) We present HgaNets, a lightweight attention-enhancing pseudo-3D residual model to capture multi-modal features for the GAR task.

2) We propose a multi-dimensional attention module to encode the important appearance, frame, and feature information. And we utilize the 3D skeletal heatmaps generated by the OpenPose to process the corresponding gesture joints.

3) The efficiency of the proposed approach is demonstrated through extensive experiments and outperforms others in terms of the recognition accuracy.

The remainder of this work is organized as follows: Related work is discussed in Section 2. Section 3 describes our proposed model in detail. Section 4 presents the experimental results, and the work is concluded in Section 5.

2 Related Works

2.1 Visual-Based GAR

Early visual-based GAR often relies on manual feature [16] for representing gesture motion. For example, Lim et al. [17] proposed block-based histogram to explicitly encode optical flow and generate sign features. Zheng et al. [3] developed pyramid histograms of oriented gradient (M-PHOG), the additional motion information from three projected orthogonal planes was applied to generated 3D gesture motion map. Oliveira et al. [18] applied principal component analysis (PCA) to compress visual features, which helped to reduce the feature dimension and improve recognition accuracy. Furthermore, gesture recognition technologies based on deep learning can automatically encode gesture features, such as posture, shape, and speed of gestures. For instance, Al-Hammadi et al. [2] proposed a 3D CNN model as a discriminative spatiotemporal descriptor for hand gestures. To better represent spatio-temporal gesture features, Luqman et al. [19] used convolutional neural network and long short-term memory network to capture spatial features and learn temporal information, respectively.

2.2 Skeletal-Based GAR

Compared to the RGB images, human skeleton is a well-established modality, which is robust not only to contextual noise, but also to variations in viewpoint and appearance [20]. Currently, there are four kinds of deep learning frameworks to encode skeletal information, including Recurrent Neural Networks (RNN), CNN and GCN, Transformer networks. For example, Xiao et al. [8] proposed a recognition method with RNN, which naively treats the skeleton as vectors formed by the body joints coordinates. Kumar et al. [21] proposed CNN model to identify discriminative spatio-temporal features of each gesture, which was interpreted using joint angular displacement maps. Heidari et al. [22] proposed GCN topology with a temporal attention module to encode non-Euclidean skeleton structures. This method represents spatiotemporal graphs of body skeleton sequences. Zhang et al. [23] proposed a spatial-temporal Transformer network to model dependencies between skeletal joints using the Transformer self-attention operator. Compared with other methods, although Transformer approach can quickly obtain global skeletal information, it is weak in encoding human gestures from local features and short-term temporal information.

2.3 Multi-Modal GAR

Multi-modal GAR often use different frameworks to represent different gesture modes. For example, Pu et al. [24] applied a classical CNN model (LeNet) to extract skeletal trajectories and a 3D CNN to represent hand videos, and trained a classifier with support vector machine (SVM) to fuse these features. Wu et al. [25] proposed a deep belief network (DBN) Network and a 3D CNN network to respectively process skeletal dynamics and depth and RGB images, and used late fusion to fuse these representative features. Besides, two global and local spatial attention networks were used to focus on the hand/arm movements in the GAR task [26], which combined the predicted probabilities from those networks with the adaptive fusion strategies. Jing et al. [27] proposed a multi-channel and multi-modal framework with 3D CNN to incorporate multimodal (RGB, depth, motion, and skeleton joints) information. However, this approach directly encodes the skeleton using a 3D CNN-based network without considering skeletal information operated on an irregular graph and visual information represented on a regular grid, resulting in unsatisfactory performance. Instead, our approach focuses on different structures of visual and skeletal data, characterizing skeletal data as joint heatmaps in a 3D volume, retaining the irregular graphical information of the skeleton.

3 Methodology

Fig. 1 presents the overall framework of the proposed HgaNets model. It consists of two independent components that are designed to model the skeletal heatmaps and RGB videos, respectively. The two components share the same network structure and each of them consists of different pseudo-3D residual blocks (i.e., Res1, Res2, and Res3) and multi-dimensional attention blocks (MDT), a fully-connected layer and a softmax layer. There are three kinds of attention sub-maps (i.e., spatial map, temporal map and channel map) in each MDT block. In order to optimize training efficiency, we implemented a residual learning concept in each component. In the end, the output probabilities of the two components are further combined to obtain the final recognition result. The overall network structure is developed to describe the dynamic spatial-temporal-channel correlations of human gesture.

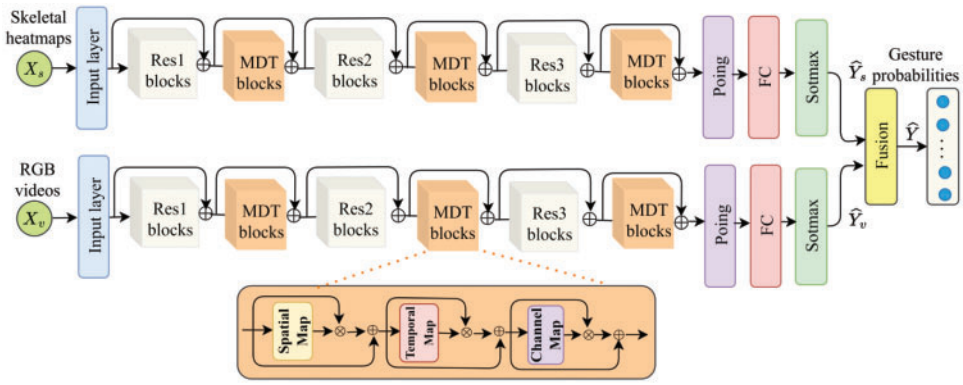

Figure 1: The framework of the HgaNets: Different pseudo-3D residual blocks: Res1, Res2, and Res3; MDT blocks: Multi-dimensional attention; Attention sub-maps: Spatial map, temporal map and channel map; FC: Fully-connected. $\oplus$ denotes the elementwise summation. $\otimes$ denotes the matrix multiplication

3.1 Skeletal Heatmaps

As illustrated in Fig. 1, we use the RGB videos and skeletal data as input, respectively. Specifically, we first define the visual RGB sample as $I \in \Re^{C\times T\times H\times W}$, where T is the number of frames, H and W are the height and width of the frame, respectively. Then, we reformulate 2D skeletal data as 3D skeletal heatmaps, as shown in Fig. 2. Formally, the size of a 3D skeletal heatmap is $K \times H \times W$, where K is the number of joints, H and W are the height and width of the frame, respectively. In this work, only the $K = 30$ upper skeletal joints (involving 8 joints from the body and 22 joints from the hands) are directly relevant for the GAR task. The skeletal joints can be estimated by OpenPose [15] technique and are often stored as coordinate-triplets (x_k, y_k, c_k), where (x_k, y_k) and c_k are respectively the coordinate and confidence score of the k_{th} joint. Finally, with the coordinate-triplets (x_k, y_k, c_k) of the skeletal joint, we obtain a joint heatmap J by composing K gaussian maps centered at every joint:

$$J_{kij} = e^{\frac{(i-x_k)^2+(j-y_k)^2}{2*\sigma^2}} * c_k \tag{1}$$

where σ controls the variance of gaussian maps. The input skeletal data is obtained by stacking all heatmaps along the temporal dimension, which thus has the size of $K \times T \times H \times W$.

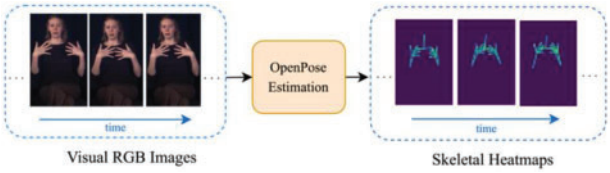

Figure 2: Heatmaps of skeleton with 30 joints are extracted by OpenPose [15] technique

3.2 Pseudo-3D Residual Model

The Pseudo-3D Residual Network (Pseudo-res3D) [28] is an improvement of the Residual Unit [6] for learning spatiotemporal features in videos. Specifically, the 3D convolution ($\mathbf{3} \times \mathbf{3} \times \mathbf{3}$) in the Residual Unit block is replaced by a combination of 2D convolution ($\mathbf{1} \times \mathbf{3} \times \mathbf{3}$) and 1D convolution ($\mathbf{3} \times \mathbf{1} \times \mathbf{1}$) in the Pseudo-res3D Block, as shown in Fig. 3. The former $\mathbf{1} \times \mathbf{3} \times \mathbf{3}$ convolution is used to obtain the characteristics of space dimension; The latter $\mathbf{3} \times \mathbf{1} \times \mathbf{1}$ convolution effectively reduces parameter calculation. Therefore, in this work, based on a 50-layer Residual Network (50-layer Resnet) [6], we introduced a deep pseudo-3D residual learning model with three different pseudo-3D residual blocks, i.e., Res1, Res2, and Res3, as described in Fig. 1.

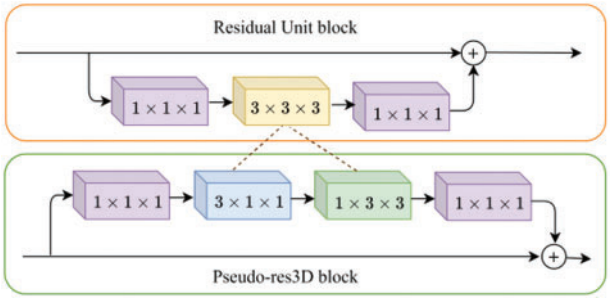

Figure 3: Illustration of the residual unit block [6] (top) and the Pseudo-res3D block [28] (bottom)

Furthermore, increasing the depth of the pseudo-3D residual model can enhance the ability to extract spatiotemporal features. For the pseudo-3D residual model with three different blocks, there are a total of three depth coefficients, namely, $\boldsymbol{L1}$, $\boldsymbol{L2}$ and $\boldsymbol{L3}$, as shown in Table 1. Here, following [6], to adapt 50-layer Residual Network for human gesture recognition, we remove the original first stage in the network and adjust the number of depth coefficients used for each block to 4, 6 and 3, i.e., $\boldsymbol{L1} = \mathbf{4}$, $\boldsymbol{L2} = \mathbf{6}$ and $\boldsymbol{L3} = \mathbf{3}$.

3.3 Multi-Dimensional Attention

We introduce multi-dimensional attention (MDT) to capture dynamic spatial, temporal and feature information in gesture action, as shown in Fig. 1. The MDT blocks consist of three attention maps: Spatial, temporal, and channel.

Spatial Map: The shape of gesture hand has rich dynamics in spatial dimension, and thus, we introduce a spatial attention [29] to capture the dynamic shape changes. Take the spatial map in the input visual data as an example:

$$\mathrm{M}_s = \sigma\left(g_s\left(AvgPool_t\left(X_v^r\right)\right)\right) \tag{2}$$

where $\boldsymbol{X}_v^r \in \Re^{C_r \times T_r \times H_r \times W_r}$ represents the visual data. Average pooling ($\boldsymbol{AvgPool}_t$) denotes the temporal average pooling across all frames, and a 2D convolution g_s is used to capture dynamic information along spatial dimension. The Sigmoid $\boldsymbol{\sigma}$ is used as the activation function to generate a spatial attention map $\mathbf{M}_s \in \Re^{1\times 1\times H_r \times W_r}$.

Table 1: The pseudo-3D residual block in the proposed model

Module	Block	Output size
Res1	$\begin{bmatrix} 1\times 1\times 1, \\ 1\times 3\times 3, \\ 1\times 1\times 1 \end{bmatrix} \times L1$	$T \times N \times N$
Res2	$\begin{bmatrix} 3\times 1\times 1, \\ 1\times 3\times 3, \\ 1\times 1\times 1 \end{bmatrix} \times L2$	$T \times (N/2) \times (N/2)$
Res3	$\begin{bmatrix} 3\times 1\times 1, \\ 1\times 3\times 3, \\ 1\times 1\times 1 \end{bmatrix} \times L3$	$T \times (N/2) \times (N/2)$

Temporal Map: The value information of human gesture is also present in different frames. Based on spatial map, we design a temporal map to assign different importance to gesture data:

$$\mathrm{M}_t = \sigma\left(g_t\left(AvgPool_s\left(X_v^r\right)\right)\right) \tag{3}$$

where average pooling ($\boldsymbol{AvgPool}_s$) is the average pooling along the spatial dimension, and a 1D convolution g_t is employed to capture temporal information. The sigmoid activation function σ is used to normalize the attention scores $\mathbf{M}_t \in \Re^{1\times T_r \times 1\times 1}$.

Channel Map: There are discriminative features in the channel dimension from the human gesture. Therefore, we also calculate a channel attention map:

$$\mathrm{M}_c = \sigma\left(g_{c2}\left(ReLu\left(g_{c1}\left(AvgPool_{st}\left(X_v^r\right)\right)\right)\right)\right) \tag{4}$$

where average pooling ($AvgPool_{st}$) is used to average the gesture data along both the spatial and temporal dimensions. Two linear functions, i.e., g_{c1} and g_{c2} are used to represent the discriminative features along the channel dimension. The resulting channel map is $\mathrm{M}_c \in \Re^{C\times 1\times 1\times 1}$ and the activation function is rectified linear unit ($ReLu$).

Structure of attention sub-maps: The three sub-maps in MDT blocks described above can be arranged in different structures: Parallel or sequential structures with different orders. Finally, we validated that sequential structures are generally better than parallel structure, where sequential order is the spatial, temporal, and channel map, as shown in Fig. 1.

In conclusion, the pseudo-3D residual model can well capture the discriminatingly spatiotemporal features for GAR. The multidimensional attention module is added to further pay attention to the spatial-temporal-channel information of gesture data. Finally, a fully connected layer is added to ensure that the output feature of each component has the same dimension and shape. And then, these output features are converted into probabilities belonging to specific gesture categories by the softmax layer.

3.4 Multi-Modal Fusion

In this section, we will discuss how to integrate the output from the two modal components. Take recognizing 3D skeletal heatmaps and RGB images as two modal examples. It is obvious that different sources of gesture show unique information, so that the output of the skeletal and visual components is more crucial. Consequently, the effect weights of the two modal components are different when the output of the two components is fused, and they should be learned from the dynamic gesture data. There are different manners to integrate these two components: Early fusion or late fusion. Finally, we found that the late fusion is better, as shown in Section.4.3.2. Therefore, the final recognition result after late fusion is:

$$\hat{Y} = \alpha * \widehat{Y}_s + (1 - \alpha) * \widehat{Y}_v \tag{5}$$

where the $\widehat{Y}_s$ and $\widehat{Y}_v$ are provided by the two components described in Fig. 1, while $\boldsymbol{\alpha}$ represents the coefficient to control the contributions of each component. Its value is optimized through cross-validation, with $\boldsymbol{\alpha}$ set as 0.5.

Algorithm 1 is the pseudo-code of the proposed model, which describes how the classification model recognizes gesture action.

Algorithm 1: Classifier model for gesture action recognition

Data: Input

D_1: $X_v \in \Re^{C\times T\times H\times W}$ represents the input of visual RGB images, where T frames in temporal dimension and $H \times W$ is the resolution of the image in the spatial dimension, C is the number of RGB channels.

D_2: $X_s \in \Re^{K\times T\times H\times w}$ represents the input of 3D skeletal heatmaps, where T frames in the temporal dimension and $H \times W$ is the resolution of the heatmap in the spatial dimension, K is the number of human joints.

Label: Y : $\mathrm{Y} \in \Re^{1\times C}$ represents the label of each type, C denotes the number of classes.

1. Obtain the visual features by the proposed model using the D_1,
2. Obtain the probabilities $\hat{Y}_v$ from the visual features by softmax layer.
3. Obtain the skeletal features by the proposed mode using the D_2,
4. Obtain the probabilities $\hat{Y}_s$ from the skeletal features by softmax layer.
5. Fuse the outputs of the two components by feeding their respective probabilities to Eq. (5).
6. Optimize the proposed model using label Y by standard gradient descent back-propagation (SGD).

Result: Output the prediction $\hat{Y}$

4 Experiments

4.1 Datasets

The proposed model experiments on two public datasets: DEVISIGN-D [30] and ASLLVD [31]. The DEVISIGN-D is collected by the visual information processing and learning of the Chinese academy of sciences, which covers 500 daily Chinese gesture vocabularies and contains 6000 video samples. The ASLLVD is collected by Boston University, containing 10k samples covering 2745 gestures, with each gesture containing between 1 and 18 samples articulated by different individuals. In the ASLLVD dataset, some gestures differ primarily in hand pose, but the overall motions of the arm may be very similar, as shown in Fig. 4. Following [32], it is suggested to split the DEVISIGN-D and ASLLVD datasets into four benchmarks: (1) DSLI, which contains 500 classes and is divided into a training set (450 videos), a test set (750 videos) and a validation set (750). (2) DSLII, which contains

50 classes and includes 620 video samples randomly split into an 80% training set and a 20% test set. (3) ASLI and (4) ASLII, which contain 20 signs and 1080 samples selected from the ASLLVD (80% training data and 20% test data).

Figure 4: Examples of gesture action in the ASLI dataset

4.2 Parameter Settings

All experiments are performed on the PyTorch deep learning framework with the RTX 3090 GPUs. The batch size is 32 and the cross-entropy loss function is applied. while the optimization strategy uses the stochastic gradient descent (SGD) with Nesterov momentum (0.9). For the DSLI dataset, we set the learning rate to 0.01 and divided it by 10 at the 25_th epoch, 50_th epoch, and 75_th epoch. For the DSLII, ASLI, and ASLII datasets, we set the learning rate to 0.01, and divided it by 10 at the 50_th epoch, 100_th epoch, and 150_th epoch.

4.3 Ablation Study

4.3.1 Empirical the MDT Module

We first studied the effect of the multi-dimensional attention (MDT) blocks by comparing the pseudo-3D residual model on the DSLI dataset. The results are shown in Table 2. There are three sub-maps in the MDT module presented in Section 3.3, i.e., spatial map, temporal map, and channel map, denoted as Res3D-S, Res3D-T, and Res3D-C, respectively. As expected, the three sub-maps produce better the accuracy of validation and test than the basic pseudo-3D model. Then we validated the

performance of both the parallel and sequential structures for each of the sub-maps, shown as Res3D-PA and Res3D-SA. It shows that the sequential structure for each of the sub-maps is slightly better. Consequently, we introduced the sequential attention structure into the pseudo-3D residual model.

Table 2: The effect of MDT module on the DSLI dataset

Methods	Validation (%)	Test (%)
Pseudo-3D	83.71	82.81
Res3D-S	85.32	84.31
Res3D-T	85.79	85.09
Res3D-C	85.71	84.67
Res3D-PA	85.83	84.90
Res3D-SA	86.27	85.40

4.3.2 Effectiveness of the Multi-Modal Fusion

In this section, we evaluated the contribution of using skeletal heatmaps and RGB images on the DSLI dataset, and the findings are presented in Table 3. The results demonstrate a significant improvement in the skeletal-visual fusion method. One possible explanation is that skeletal information is more robust, because it benefits from training using large and highly varied data. Instead, RGB images rely only on the raw data and are learned solely from the training set, which may lead to overfitting. As expected, combining the skeletal and visual information produced a significant improvement. We then tested the performance of the two components during the early input stage and the later stage, shown as early fusion and late fusion, respectively. This suggests that integrating these two components in later stages brings improvement. Finally, we applied the later fusion approach to the proposed model and obtained the better results.

Table 3: The recognition accuracy for the different modalities on the DSLI dataset

Module	Validation (%)	Test (%)
Skeletal heatmaps	81.69	82.70
RGB images	72.9	76.51
Early fusion	85.20	84.23
Later fusion	86.27	85.40

Additionally, we studied the complementarity between skeletal and visual data by drawing confusion matrices on ASLI dataset, as shown Figs. 5 and 6. Fig. 5 shows the confusion matrices between the skeletal heatmaps (left) and the RGB images (right). We can observe that the color of the "afraid" gesture in visual confusion matrices is more yellow than in skeletal, which shows that RGB images help a lot for the "afraid" class with fear emotion. And the color of the "Eat" and "Drink" gestures in skeletal confusion matrices are more yellow than in visual, which shows that the skeleton significantly different between the "Eat" and "Drink" samples. As shown in Fig. 4, the "Eat" and "Drink" gestures have the same emotion, but their gestures are different. Thus, this model can distinguish between them based on whether the hand is raised above the mouth by skeletal component.

Likewise, this model also can distinguish similar gestures based on their different emotion by visual information component. As expected, we fuse the skeletal and visual data, which brings confusion matrices improvement, as shown in Fig. 6.

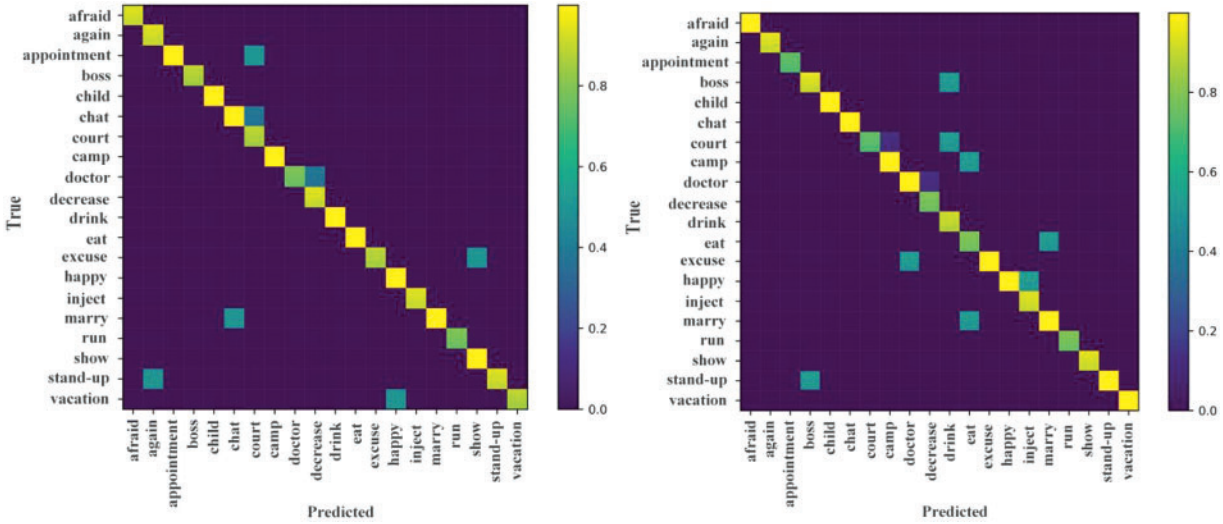

Figure 5: Confusion matrices on the ASLI dataset. Skeletal heatmaps (left); RGB images (right)

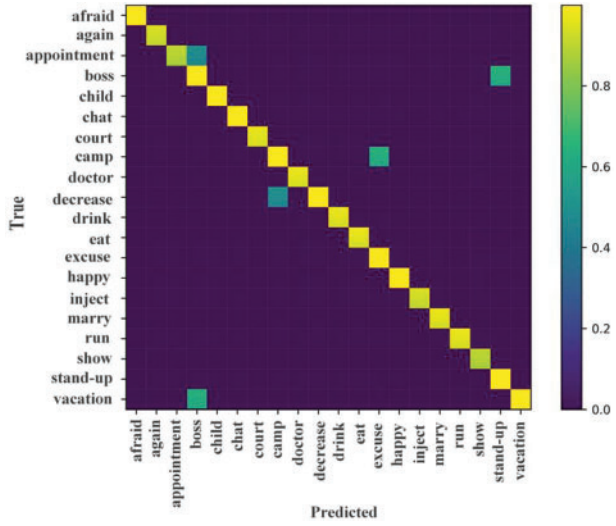

Figure 6: Confusion matrices for the skeletal-visual data on the ASLI dataset

4.4 Comparison with the State-of-the-Arts

We compared the final model with nine baselines on the ASLI and ASLII datasets. The results are shown in Table 4. The baselines used for comparisons include the methods [9,16–18,29,32–35]. The proposed model (Ours) achieves the state-of-the-art performance with a large margin on the datasets. Specifically, when compared to the baseline of [17] and [32], the accuracy of the proposed model on the ASLI dataset achieves an improvement of 9.7% and 6.82% and on the ASLII dataset improves by 10.3% and 7.32%, respectively.

Table 4: Comparison of results with state-of-the-art related works

Module	Accuracy (%)
MHI [33]	10
MEI [16]	25
PCA [18]	45
HOF [34]	70
BHOF [17]	85
ST-GCN [9]	56.82
AGCN [35]	80.96
AAGCN [29]	85.79
Aegles [32] (ASLI)	87.88
Aegles [32] (ASLII)	87.98
Ours (ASLI)	94.70
Ours (ASLII)	95.30

4.5 Performance for Recognition Accuracy

We analyzed the recognition accuracy of the proposed model on the DSLII and ASLII datasets, as shown in Figs. 7 and 8, respectively. The accuracy on Figs. 7 and 8 is 91.91% and 95.3%, respectively. Fig. 9 shows the recognition confusion matrix for the ASLII (left) and DSLII (right) datasets, respectively. As shown in the figure, the proposed model achieves high accuracy and demonstrates excellent recognition performance on different datasets.

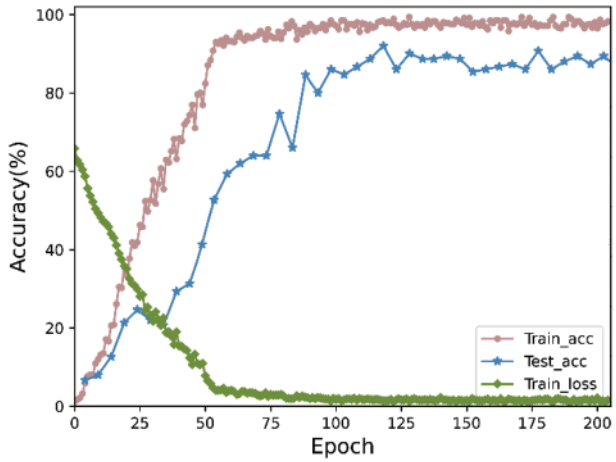

Figure 7: Train and test accuracies on the DSLII dataset

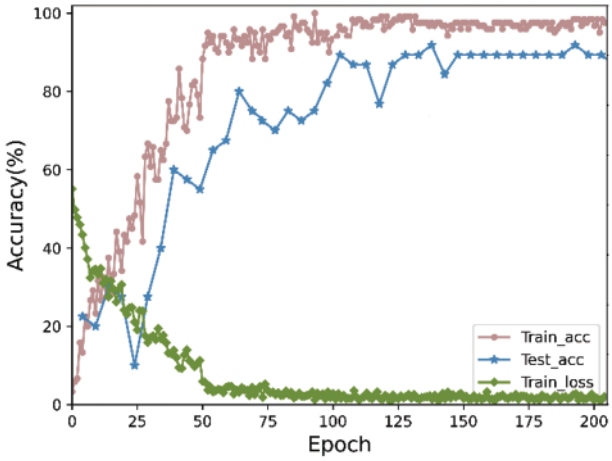

Figure 8: Train and test accuracies on the ASLII dataset

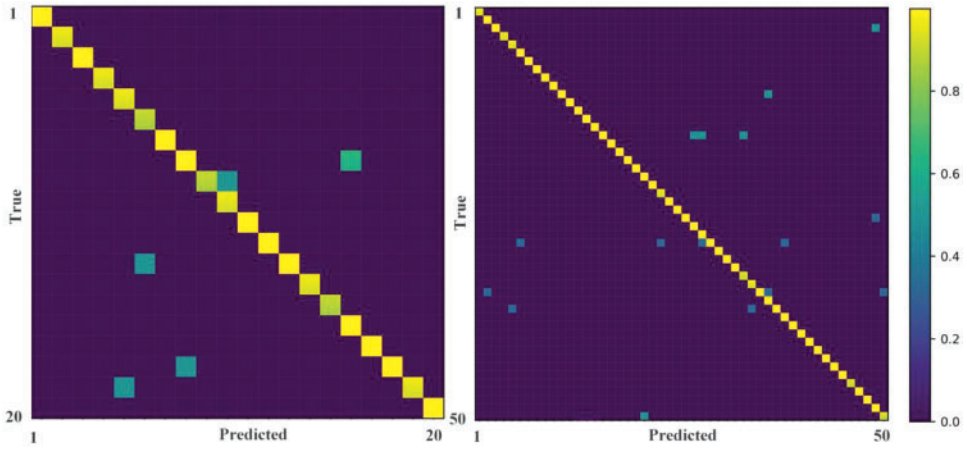

Figure 9: Confusion matrices for testing performance on the ASLII (left) and DSLII (right) datasets

4.6 Performance for Parameter and Time Cost

In order to evaluate the runtime performance of the proposed model, we analyzed the parameters and runtime of the model with other baseline models on the DSLI dataset, as shown in Table 5. We set the size of the input RGB images to $\mathbf{1} \times \mathbf{3} \times \mathbf{300} \times \mathbf{56} \times \mathbf{56}$, denoting 1 batch, 300 frames 5, 3 channels, 56 height and 56 width, and the size of the input 3D skeletal heatmaps to $\mathbf{1} \times \mathbf{28} \times \mathbf{300} \times \mathbf{56} \times \mathbf{56}$, denoting 1 batch, 300 frames, 28 joints, 56 height and 56 width. Runtime represents the time required to run 100 models using RTX 3090 GPUs. Our model requires only 2.7M parameters, which is much smaller than other methods. Furthermore, the inference time of our method is 0.39 s, while the baseline from [32] is 0.33 s. The extra computation mainly comes from the multimodal fusion calculations, but it is still a small fraction of the [32] baseline. Moreover, increasing the number of modes brings additional computational costs but not extra parameters. These evaluation results demonstrate that our proposed method can run in real time and is helpful for GAR.

Table 5: Runtime performance with existing methods on the DSLI dataset

Method	#Parameters (M)	Times (s)
ST-GCN [9]	3.12	0.09
AGCN [35]	3.47	0.26
AAGCN [29]	3.78	0.31
Aegles [32]	4.60	0.33
Ours (skeleton)	2.74	0.39
Ours (RGB)	2.74	0.39
Ours (skeleton + RGB)	2.74	0.54

5 Conclusion

In this paper, we propose an attention-enhanced pseudo-3D residual model (HgaNets) for the GAR. This model consists of two independent components with the same pseudo-3D residual convolution that encode the RGB videos and 3D skeletal heatmaps in an end-to-end manner. To further enhance the proposed model, we introduce a multi-dimensional attention module in the pseudo-3D residual model that focuses on discriminating spatial, temporal and feature information. Finally, the outputs of the two components are fused to generate the recognition results. Experiments are conducted on the four datasets: DSLI, DSLII, ASLI, and ASLII, and the accuracy of the proposed model is 85.40%, 91.91%, 94.70%, and 95.30%, respectively. Moreover, our model has a size of only 2.74M on NVIDIA GeForce GTX 1070 GPU and an inference time of 0.54 s, demonstrating the ability to run in real-time for the GAR tasks. Future research will focus on deep learning models to process multimodal embedding and cross-modal interaction between visual and textual semantics for the GAR task.

Acknowledgement: We are thankful to the reviewers for helping us improve this paper's quality.

Funding Statement: This study was supported by the National Natural Science Foundation of China under Grant No. 62072255.

Author Contributions: Study conception and design: Wuyan Liang, Xiaolong Xu; data collection: Wuyan Liang; analysis and interpretation of results: Wuyan Liang, Xiaolong Xu; draft manuscript preparation: Wuyan Liang, Xiaolong Xu. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The ASLLVD and DEVISIGN-D datasets considered in this study are publicly available, which can be accessed from https://www.bu.edu/asllrp/av/dai-asllvd.html and http://vipl.ict.ac.cn/homepage/ksl/data.html, respectively.

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

References

[1] Y. Liu, Y. Fan, and Z. Wu, "Ultrasound-based 3-D gesture recognition: Signal optimization, trajectory, and feature classification," *IEEE Trans. Instrum. Meas.*, vol. 72, no. 4, pp. 1–12, 2023. doi: 10.1109/TIM.2023.3235438.

[2] M. Al-Hammadi, G. Muhammad, and M. A. Mekhtiche, "Hand gesture recognition for sign language using 3DCNN," *IEEE Access*, vol. 8, pp. 79491–79509, 2020.

[3] L. Zheng and B. Liang, "Sign language recognition using depth images," in *2016 14th Int. Conf. Control, Autom., Robot. Vision (ICARCV)*, Phuket, Thailand, 2016, pp. 1–6.

[4] G. Park, V. K. Chandrasegar, and J. Koh, "Accuracy enhancement of hand gesture recognition using CNN," *IEEE Access*, vol. 11, pp. 26496–26501, 2023. doi: 10.1109/ACCESS.2023.3254537.

[5] L. Pigou and S. Dieleman, "Sign language recognition using convolutional neural networks," in *European Conf. Comput. Vision*, Zurich, Switzerland, 2014, pp. 572–578.

[6] K. He, X. Zhang, and J. Sun, "Deep residual learning for image recognition," in *Proc. 2016 IEEE Conf. Comput. Vision Pattern Recogn.*, Las Vegas, NV, USA, 2016, pp. 770–778.

[7] Y. Li, Q. Miao, and K. Tian, "Large-scale gesture recognition with a fusion of RGB-D data based on the C3D model," in *2016 23rd Int. Conf. Pattern Recogn.*, Cancun, Mexico, 2016, pp. 25–30.

[8] Q. Xiao, M. Qin, and Y. Yin, "Skeleton-based Chinese sign language recognition and generation for bidirectional communication between deaf and hearing people," *Neural Netw.*, vol. 125, no. 1, pp. 41–55, 2020. doi: 10.1016/j.neunet.2020.01.030.

[9] C. C. Amorim, D. Macêdo, and C. Zanchettin, "Spatial-temporal graph convolutional networks for sign language recognition," in *Int. Conf. Artif. Neural Netw.*, Munich, Germany, 2019, pp. 646–657.

[10] A. Tunga, S. V. Nuthalapati, and J. Wachs, "Pose-based sign language recognition using GCN and BERT," in *Proc. IEEE Winter Conf. Appl. Comput. Vision Works.*, Waikola, HI, USA, 2021, pp. 31–40.

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

[12] N. C. Camgoz and O. Koller, "Multi-channel transformers for multi-articulatory sign language translation," in *European Conf. Comput. Vision*, Glasgow, UK, 2020, pp. 301–319.

[13] J. Huang, W. Zhou, and H. Li, "Sign language recognition using 3D convolutional neural networks," in *2015 IEEE Int. Conf. Multimed. Expo*, Turin, Italy, 2015, pp. 1–6.

[14] R. Kumar and S. Kumar, "Multi-view multi-modal approach based on 5S-CNN and BiLSTM using skeleton, depth and RGB data for human activity recognition," *Wirel. Person. Commun.*, vol. 130, no. 2, pp. 1141–1159, 2023. doi: 10.1007/s11277-023-10324-4.

[15] Z. Gao and G. Hidalgo, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, 2021. doi: 10.1109/TPAMI.2019.2929257.

[16] M. Muneeb, H. Rustam, and A. Jalal, "Automate appliances via gestures recognition for elderly living assistance," in *2023 4th Int. Conf. Adv. Comput. Sci. (ICACS)*, Lahore, Pakistan, IEEE, 2023, pp. 1–6.

[17] K. M. Lim and A. W. C. Tan, "Block-based histogram of optical flow for isolated sign language recognition," *J. Vis. Commun. Image Rep.*, vol. 40, no. 1, pp. 538–545, 2016. doi: 10.1016/j.jvcir.2016.07.020.

[18] M. Oliveira, A. Sutherland, and M. Farouk, "Two-stage PCA with interpolated data for hand shape recognition in sign language," in *2016 IEEE Appl. Imagery Pattern Recogn. Works. (AIPR)*, Washington DC, USA, 2016, pp. 1–4.

[19] H. Luqman and E. Elalfy, "Utilizing motion and spatial features for sign language gesture recognition using cascaded CNN and LSTM models," *Turkish J. Electr. Eng. Comput. Sci.*, vol. 30, no. 7, pp. 2508–2525, 2022. doi: 10.55730/1300-0632.3952.

[20] X. You, Q. Gao, and H. Gao, "A feature fusion network for skeleton-based gesture recognition," in *Intell. Robot. Appl.: 16th Int. Conf.*, Hangzhou, China, 2023, pp. 67–78.

[21] E. K. Kumar, P. V. V. Kishore, and D. A. Kumar, "Training CNNs for 3-D sign language recognition with color texture coded joint angular displacement maps," *IEEE Signal Process Lett.*, vol. 25, no. 5, pp. 645–649, 2018. doi: 10.1109/LSP.2018.2817179.

[22] N. Heidari and A. Iosifidis, "Temporal attention-augmented graph convolutional network for efficient skeleton-based human action recognition," in *25th Int. Conf. Pattern Recogn.*, Milan, Italy, 2021, pp. 7907–7914.

[23] Q. Zhang, T. Wang, and M. Zhang, "Spatial temporal transformer network for skeleton-based action recognition," in *2021 China Autom. Congress (CAC)*, Beijing, China, 2021, pp. 7029–7034.

[24] J. Pu, W. Zhou, and H. Li, "Sign language recognition with multi-modal features," in *Pacific Rim Conf. Multimed.*, Xi'an, China, 2016, pp. 252–261.

[25] D. Wu, L. Pigou, and P. Kindermans, "Deep dynamic neural networks for multi-modal gesture segmentation and recognition," *IEEE Trans. Pattern Anal.*, vol. 38, no. 8, pp. 1583–1597, 2016. doi: 10.1109/TPAMI.2016.2537340.

[26] Q. Yuan, J. Wan, and C. Lin, "Global and local spatial-attention network for isolated gesture recognition," in *Chinese Conf. Biometr. Recogn.*, Zhuzhou, China, 2019, pp. 84–93.

[27] L. Jing, E. Vahdani, and M. Huenerfauth, "Recognizing american sign language manual signs from RGB-D videos," arXiv preprint arXiv:1906.02851, 2019.

[28] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3D residual networks," in *Proc. IEEE Int. Conf. Comput. Vision*, Venice, Italy, 2017, pp. 5534–5542.

[29] L. Shi, Y. Zhang, and J. Cheng, "Skeleton-based action recognition with multi-stream adaptive graph convolutional networks," *IEEE Trans. Image Process.*, vol. 29, pp. 9532–9545, 2020. doi: 10.1109/TIP.2020.3028207.

[30] X. Chai, H. Wang, and X. Chen, "The devisign large vocabulary of chinese sign language database and baseline evaluations," *Technical Report VIPL-TR-14-SLR-001*, Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, 2014.

[31] A. T. Neidle and S. Sclaroff, "Challenges in development of the American sign language lexicon video dataset (ASLLVD) corpus," in *5th Works. Rep. Process. Sign Lang.: Interact. between Corpus Lexicon, LREC*, 2012.

[32] W. Liang, X. Xu, and F. Xiao, "Human gesture recognition of dynamic skeleton using graph convolutional networks," *J. Electron. Image*, vol. 32, no. 2, pp. 021402, 2022. doi: 10.1117/1.JEI.32.2.021402.

[33] V. Athitsos, C. Neidle, S. Sclaroff, and J. Nash, "The American sign language lexicon video dataset," in *IEEE Comput. Soc. Conf. Comput. Vision Pattern Recogn. Works.*, Anchorage, AK, USA, 2008, pp. 1–8.

[34] R. V. Babu and K. R. Ramakrishnan, "Recognition of human actions using motion history information extracted from the compressed video," *Image Vision Comput.*, vol. 22, no. 8, pp. 597–607, 2004. doi: 10.1016/j.imavis.2003.11.004.

[35] L. Shi, Y. Zhang, and J. Cheng, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *IEEE Conf. Comput. Vision Pattern Recogn.*, Long Beach, CA, USA, 2019, pp. 12018–12027.