Behavioral Recognition of Skeletal Data Based on Targeted Dual Fusion Strategy

Xiao Yun[1], Chenglong Xu[1], Kevin Riou[2], Kaiwen Dong[1]*, Yanjing Sun[1], Song Li[1], Kevin Subrin[2], Patrick Le Callet[2]

[1]School of Information and Control Engineering, China University of Mining and Technology, Xuzhou, 221116, China
[2]Nantes Université, Ecole Centrale Nantes, CNRS, LS2N, UMR 6004, Nantes, France
{xyun, clongxu, dongkaiwen, yjsun, lisong}@cumt.edu.cn, {kevin.riou, kevin.subrin, patrick.lecallet}@univ-nantes.fr

Abstract

The deployment of multi-stream fusion strategy on behavioral recognition from skeletal data can extract complementary features from different information streams and improve the recognition accuracy, but suffers from high model complexity and a large number of parameters. Besides, existing multi-stream methods using a fixed adjacency matrix homogenizes the model's discrimination process across diverse actions, causing reduction of the actual lift for the multi-stream model. Finally, attention mechanisms are commonly applied to the multi-dimensional features, including spatial, temporal and channel dimensions. But their attention scores are typically fused in a concatenated manner, leading to the ignorance of the interrelation between joints in complex actions. To alleviate these issues, the Front-Rear dual Fusion Graph Convolutional Network (FRF-GCN) is proposed to provide a lightweight model based on skeletal data. Targeted adjacency matrices are also designed for different front fusion streams, allowing the model to focus on actions of varying magnitudes. Simultaneously, the mechanism of Spatial-Temporal-Channel Parallel Attention (STC-P), which processes attention in parallel and places greater emphasis on useful information, is proposed to further improve model's performance. FRF-GCN demonstrates significant competitiveness compared to the current state-of-the-art methods on the NTU RGB+D, NTU RGB+D 120 and Kinetics-Skeleton 400 datasets. Our code is available at: https://github.com/sunbeam-kkt/FRF-GCN-master.

Introduction

Human action recognition (HAR) aims to determine the current behavior category of a person based on a series of well-trained recognition models, and it has wide range of applications, such as human-computer interaction (Liu and Wang 2020), video surveillance (Xin et al. 2023), and autonomous driving (Saleh et al. 2022). In particular, after the introduction of skeleton-based HAR by ST-GCN (Yan, Xiong, and Lin 2018), algorithms based on GCN for skeleton-based behavior recognition have emerged rapidly. Recent approaches on GCN based behavior recognition mostly focused on 3 axis.

The first is how the adjacency matrix in GCN can better learn the relationship between joints to enhance the learning ability of the model. ST-GCN adopts the natural connections between human joints as the topological relations to be learned, but is unable to establish new generative relations as unnatural connections. 2S-AGCN (Shi et al. 2019b) proposes an adaptive adjacency matrix to solve this problem in order to learn more information. Further, MS-AAGCN (Shi et al. 2020) adds learnable coefficients to the adjacency matrix to make it more flexible. It has also been argued that more relations between joints do not always help action learning, for example, STSF-GCN (Fang et al. 2022) has obtained relatively good performance using fewer relations on a GCN model with joints as data input.

The second is different attention mechanisms are adopted to better capture the spatio-temporal as well as the channel attention scores. Numerous studies have explored the influence of attentional mechanisms on action judgments, which is very reasonable from a biological point of view. All attention mechanisms are roughly divided into three categories: temporal (Liu et al. 2020), spatial (Song et al. 2020), and channel (Chen et al. 2021c). Many scholars have addressed one or two of these, but more advanced models incorporate all three, such as DC-GCN (Zhou et al. 2023), AM-GCN (Sun et al. 2022).

The last is how multi-stream information can be better fused. Different types of data often contain distinct feature information, and the fusion of multiple streams of information generally achieves better results than using a single stream. Currently, most behavior recognition models based on skeleton data employ architectures that fuse multiple streams of information (Hu et al. 2022; Qin et al. 2022; Wu, Zhang, and Zou 2023; Zhang et al. 2023). The data fusion schemes used can be broadly classified into two categories: 1) Rear fusion of data (Chen et al. 2021d; Liu et al. 2022a; Xiong et al. 2022), where each stream of information is processed by the same model to obtain behavior category scores. These scores are then combined through weighted fusion to produce a final behavior classification result. 2) Front fusion of data (Song et al. 2020, 2022) where the input data is fused prior to being processed by the model, followed by feature extraction. We neutralized these two schemes in FRF-GCN and obtained better fusion results.

We noticed 3 main limitations in these recent develop-

*Corresponding author

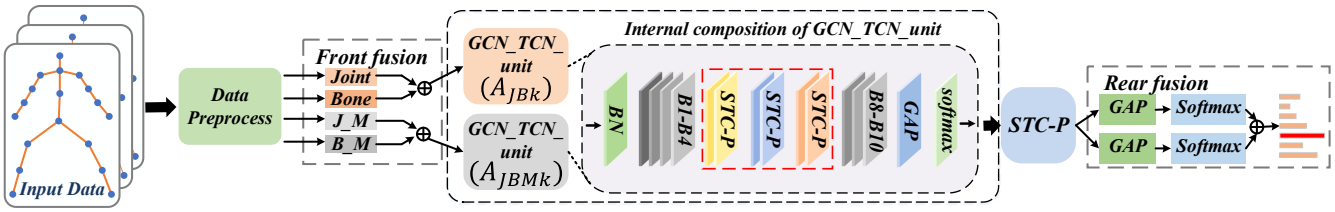

Figure 1: Overall flow chart of FRF-GCN, the J_M and B_M are the joint motion and bone motion, respectively, where $\oplus$ denotes information fusion (including forward fusion and backward fusion), GCN_TCN_unit is the spatial temporal joint unit, the upper and lower branches use a targeted adjacency matrix A_{JB} and A_{JBM}, STC-P is the spatial temporal channel parallel attention mechanism, GAP and Softmax are the global average pooling and classifier, respectively.

ments: (1) Different adjacency matrices are not selected according to the different characteristics of multi-stream information, which may overlook valuable information. (2) Most of the existing attention mechanisms use a serial processing approach, learning spatial, temporal and channel features sequentially or in a different order. The attention learned at the back-end may be influenced by the front-end, while some interconnections may be lost. Moreover, we know from our experiments that joint spatio-temporal attention and channel attention do not seem to enjoy equal status for behavioral classification tasks. (3) Incorporating novel modalities increases the complexity of the architecture, and while some efficient solutions have been proposed for multi-stream GCN based behavior recognition, there is still a significant gap with lightweight solutions.

To alleviate these limitations, we propose the FRF-GCN, which incorporates the following 3 key novelties:

1. Instead of using a fixed adjacency matrix, we propose targeted adjacency matrices for the fusion of two different information sources. This approach enhances the efficiency of GCN computations and minimizes redundant calculations.

2. To simultaneously capture attention scores in the temporal, spatial, and channel dimensions of skeleton data, we introduce the STC-P attention mechanism. By incorporating it into FRF-GCN, we preserve the interdependencies between temporal, spatial, and channel attention, resulting in more effective extraction of skeleton information.

3. We propose a lightweight architecture for fusing multiple streams of skeleton information through bidirectional fusion. This approach reduces the model's parameterization while maintaining its performance.

FRF-GCN achieves a balance between parameter size and performance on the NTU60, NTU120 and Kinetics-Skeleton datasets, demonstrating competitiveness against current state-of-the-art models.

Related Works

Multi-Stream Information Data Fusion

In recent years, many studies have demonstrated that multi-stream information fusion can lead to better performance, such as MS-AAGCN (Shi et al. 2020), MST-GCN (Chen et al. 2021d) and 4S-ACE-Ens (Qin et al. 2022). The number of information streams ranges from dual streams (Shi et al. 2019b; Wu, Wu, and Kittler 2021), to triple streams (Song et al. 2022), quadruple streams (Chen et al. 2021a; Yang et al. 2021), and even more (Chen et al. 2021b). The types of information are also diverse, including not only joint bones and their motion information, but also angular information and so on. They are fused at the end or at the beginning of the model to obtain better results after fusion. Based on the results done in previous studies, it is demonstrated that front fusion leads to a significant reduction in the number of parameters but a decrease in performance, while rear fusion leads to a significant increase in performance but an exponential increase in the number of parameters.

In contrast, our FRF-GCN model utilizes a dual data fusion scheme, combining both forward and post fusion approaches, along with targeted adjacency matrices. This approach effectively reduces the parameter count while minimizing performance loss. By leveraging both forward and post fusion techniques, our model achieves a balance between parameter reduction and preserving performance.

Attention Mechanisms in Behavior Recognition

Attention plays an extremely important role in behavioral recognition models, and there are two main categories of its forms of action: local attention and global attention.

In previous studies, local attention has been the majority and has given significant impetus to later studies. For example, ST-GCN (Yan, Xiong, and Lin 2018) and AT-GCN (Sheng and Li 2021) use the local attention mechanism to capture the local association information in the skeleton sequence. Later, with the rise of transformer and other attention mechanisms in Natural Language Processing (NLP), many scholars introduced them into behavior recognition tasks, such as KA-AGTN (Liu et al. 2022b), ACT (Mazzia et al. 2022), HGCT (Bai et al. 2022), CSCMFT (Liu et al. 2023), etc. The idea of transformer is adopted. And the attention mechanism represented by transformer focuses on the global state information.

In addition to the transformer model, excellent GCN models such as CTR-GCN (Chen et al. 2021c) and EfficientNet (Song et al. 2022) also choose to adopt a more global focus on attention information. This also makes sense biologically, as people are often used to adopting a top-down strategy (Lange and Lappe 2006) when judging what category an action belongs to, which shows the importance of global information. The STC-P attention mechanism we deployed

in FRF-GCN also adopts the idea of global attention.

Model Architecture

The model of FRF-GCN proposed in this paper is shown in Figure 1, and its main component structures are specified by following subsections.

Front-Rear Dual Fusion Strategy

To balance the number of parameters and performance of the state of the art models, this paper designs a dual fusion strategy, combining both front fusion and rear fusion. The idea is straightforward: before inputting the information into the network, the four streams (i.e. joint, bone, joint motion and bone motion) are fused in pairs, which the joint and bone information are fused, as well as the joint motion and bone motion information, as shown in Figure 1. Subsequently, the two fusion streams are input into the network using a dual-stream model.

The front fusion stage is to splice the two sets of information mentioned earlier in terms of channel dimensions, after which it is fed into the GCN_TCN_unit shown in Figure 1. Unlike the conventional two-stream GCN model, the adjacency matrix in FRF-GCN is based on different focuses on input data characteristics. After the front fusion stage, the adjacency matrix A_{JB} is selected based on the fact that the joint and bone fusion flow information is more focused on large-amplitude movements (Figure 2, left). And to focus more on small-amplitude movements as well as the motion of the joint points themselves, we also select adjacency matrix A_{JBM} (Figure 2, right) based on the features that the joint motion and bone motion fusion streams provide. The features obtained from the two tributaries enter a rear fusion phase. This process can be described as the following equation.

$$f_{out} = \alpha f_c\left(GAP\left(f_{JB}\right)\right) + \beta f_c\left(GAP\left(f_{JBM}\right)\right) \quad (1)$$

The final output scores of the FRF-GCN model are shown in equation (1). α and β are set to 0.6 and 0.4, respectively, as the backward fusion weights after a small grid search. f_c is the fully connected layer, GAP is the average pooling layer, f_{JB} is the behavioral discriminant score of joint plus bone fusion flow, and f_{JBM} is the behavioral discriminant score of joint motion plus bone motion fusion flow. This reduces the model's parameter count by half and enhances performance through targeted adjacency matrices compared to pure front fusion. Relevant experiments are presented in Table 4.

Targeted Spatial Graph Convolution

In the model of 2S-AGCN (Shi et al. 2019b), the original adjacency matrix can be represented as the following equation. Where I stands for the joint point itself, A_{ske} records the physical connections inherent in the body.

$$A = I + A_{ske} \quad (2)$$

Motivated by previous research (Fang et al. 2022), we studied whether different adjacency matrices should be used for different information streams. Through extensive experimentation, we found that considering relationships between all joints may lead to misjudgments by the model. However, using only identity matrix I and learnable masks is not sufficient for the fused information after front fusion. Moreover, existing attention mechanisms often prioritize actions with larger variations by the connections learned with adjacency matrix, which may not be suitable for subtle actions with smaller variations. Additionally, using a fixed adjacency matrix for different information streams is not always optimal, as it introduces more computilities and overlooks data characteristics.

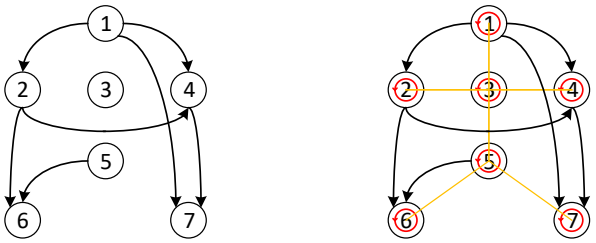

Figure 2: Illustration of targeted adjacency matrix, the left is A_{JB}, contains only the newly learned relationships between joints, and the right is A_{JBM}, includes both the inherent connections between joints (the orange connections) and the motion patterns of individual joints themselves (the red connections), best view in the color mode.

To alleviate these issues, FRF-GCN performs targeted adjacency matrix selection for different information flows. This approach increases the utilization of effective computility and reduces the chances of redundant information interfering with action judgments.

For joint and bone fusion streams, both joint and bone information represent absolute positional information. The coordinates of the data indicate the positions of the respective joints or bones. In this case, it is important to focus on the newly generated relationships, as illustrated in the left diagram of Figure 2.

Thus, we utilize an adjacency matrix A_{JB} to represent the fused information of joint and bone. A_{JB} represents the parameterized initial physical topology relationship (A) as shown in the left side of Figure 2. The parameters in the matrix indicates whether there is a connection between two articulations and the strength of the connection. The parameters are updated each time by back propagation and there is no restriction on the parameters in A_{JB} to maintain its learning capability. The formula for the joint and bone fusion flow is shown in equation (3), where A_{JB} is initialized by A. Where K_v follows the setting of ST-GCN (Yan, Xiong, and Lin 2018), which is set to 3, and W_k is the corresponding weight.

$$f_{out} = \sum_{k=1}^{K_v} W_k f_{in}\left(A_{JBk} + \mathrm{P}_k\right) \quad (3)$$

For the joint motion and bone motion fusion streams, both fused information belong to relative position information,

which is different to joint and bone information. Moreover, the fusion flow information of joint motion and bone motion itself focuses on motion information, adding I and A_{ske} can make it fully exploit small-amplitude movements. A_{JBM} is shown in the right of Figure 2, and its calculation formula is as in Eq. (4),

$$f_{out} = \sum_{k=1}^{K_v} W_k f_{in} \left(A_{JBMk} + P_k\right) \tag{4}$$

$A_{JBMk} = I + A_{ske} + A_{JBk}$, I represents the movement of the joint itself, and A_{ske} is the normalized relationship matrix between the joints. P_k is a unique graph learned for each sample, obtained by calculating the similarity between two vertices by two normalized embedding Gaussian functions (θ and φ) and then normalizing, as in equations. Where θ and φ are realized by two 1*1 convolutions, N is the number of joints in a single human skeleton, i and $j \in [1, N]$. The computation process of f can be expressed as Equation (6).

$$f\left(v_i, v_j\right) = \frac{e^{\theta(v_i)^T \varphi(v_j)}}{\sum_{j=1}^{N} e^{\theta(v_i)^T \varphi(v_j)}} \tag{5}$$

$$P_k = softmax\left(f_{in}^T W_{\theta k}^T W_{\varphi k} f_{in}\right) \tag{6}$$

In theory, large-scale movements often prioritize newly generated relationships. Small-scale movements typically emphasize the motion of individual joints or the motion of initial physical connections. The joint and bone fusion streams using A_{JB} are better able to focus on the behaviors that serve as criteria for large-scale movements. Meanwhile, the fusion of joint motion and bone motion streams using A_{JBM} is more attentive to behaviors that serve as criteria for small-scale movements. These two types of information complement each other in the rear fusion stage, resulting in improved fusion performance. The relative experiments could be seen in experiments section.

Multi-Field Temporal Depth-Point Convolution

To address the issues with conventional temporal convolution in terms of parameter explosion and limited temporal receptive field, FRF-GCN replaces it with a multi-scale depth-point convolution, which combines depth-wise and point-wise convolutions with different receptive field sizes achieved through dilated convolutions. This allows for a more comprehensive extraction of temporal information. The number of parameters in the whole model drops dramatically due to the effective splitting of the convolution process, achieving lightweight design.

The multi-scale depth-point temporal graph convolution used in FRF-GCN was called MD-TGCN. MD-TGCN is shown in Figure 3 and consists of three branches, regular depth-point convolution, large field of view depth-point convolution, and residual connection. The regular depth-point convolution captures the local detail information, the large field of view depth-point convolution captures the global information, and the residual join is added with the information from the other two branches after stitching to optimize the training process.

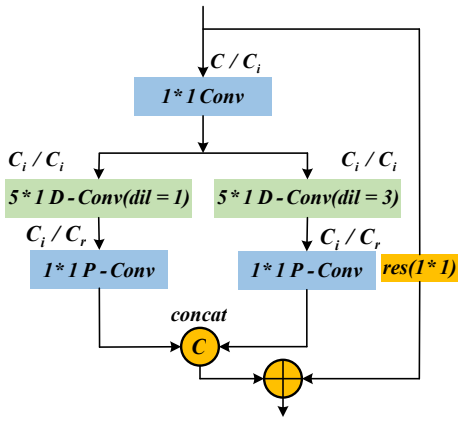

Figure 3: Temporal convolution flow chart, where C_i is the number of embedded channels, C_r is the number of single-branch output channels, $D-Conv$ and $P-Conv$ represent depth-wise convolution and point-wise convolution, respectively.

The depth-point convolution model with multiple fields of view operates on the input sequence as in equation (7) and (8), where f_{in} is the input data, f_{D1} is the depth-wise convolution with an expansion factor of 1, f_{D2} is the expansion depth-wise convolution with an expansion factor of 3, f_P is the $1*1$ point-wise convolution, $\oplus$ represents the splicing in the channel dimension, f_{D-P} is the combined output of the depth-point convolution, Res is the residual connection.

$$f_{D-P} = f_P\left(f_{D1}\left(f_{in}\right)\right) \oplus f_P\left(f_{D2}\left(f_{in}\right)\right) \tag{7}$$

$$f_{out} = Res\left(f_{D-P}, f_{in}\right) \tag{8}$$

STC-P Attentional Mechanisms

In addition to the intuitive spatio-temporal features, the hidden channel features should not be ignored in the extraction of attention. As the number of channels changes, during the learning process of the model, many channels start to contribute less to the overall performance, but expend considerable computilities to re-learn their weights, especially in convolutional layers with a large number of channels. To alleviate this impunity, we propose the Spatial-Temporal-Channel Parallel Attention Module (STC-P), which integrates channel-level attention based on the Squeeze-and-Excitation Network concept (Hu, Shen, and Sun 2018).

The specific operations are illustrated in Figure 4. The extraction of spatio-temporal attention (Song et al. 2022) is shown in the upper left corner of Figure 4. However, an additional branch is introduced to extract channel attention. The whole process is simple and is divided into three major steps as shown in the bottom left corner of Figure 4: T-Pooling, S-Pooling, and FC. The resulting channel attention map is multiplied element-wise with the input feature map to obtain the final channel attention-weighted feature map. The final output is obtained via a fusion module.

Therefore, the STC-P module completes the task of simultaneously obtaining attention scores in three dimensions: spatial, temporal, and channel. The learned spatial features

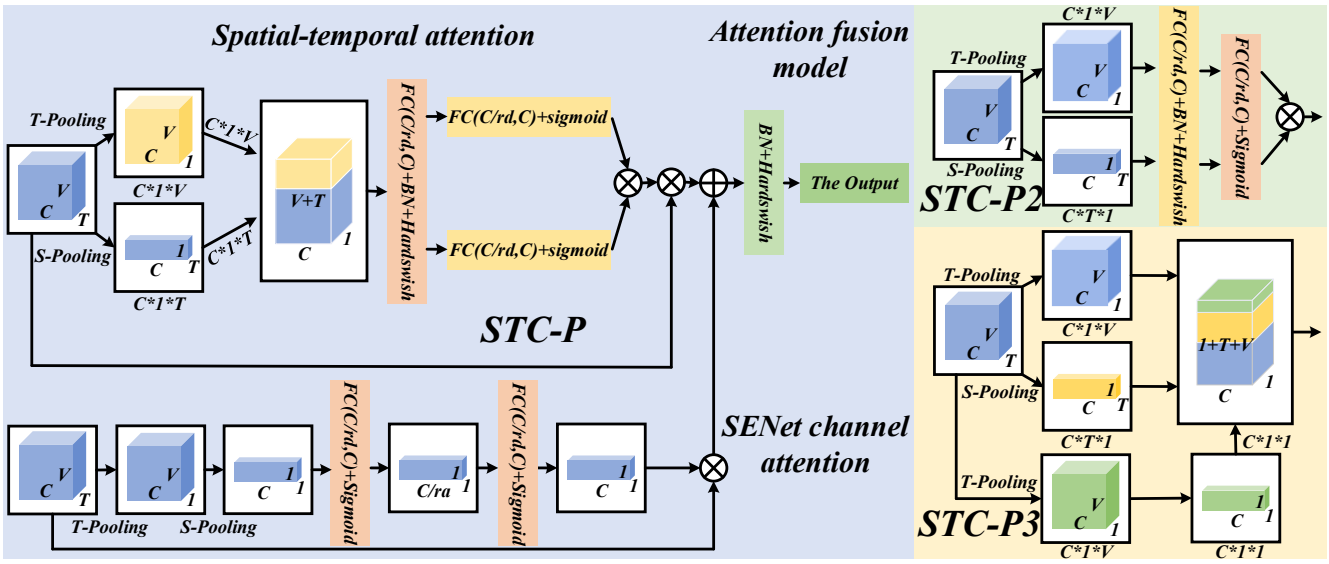

Figure 4: The STC-P attention module, "$Hardswish$" refers to the activation function used, and rd, ra denote reduction factors, temporal pooling and spatial pooling are denoted as T-pooling and S-pooling, respectively. The variants of STC-P were also be shown, where STC-P2 is in the green area and STC-P3 is in the yellow area, only shows how they differ from the STC-P.

of the STC-P module are connected to the prior features learned by the preceding spatial convolutional neural network, which makes the learned spatial attention map more accurate. During the experiment, it was found that the best results were achieved after adding STC-P only to layers 5, 6 and 7. The other cases caused different degrees of performance loss. Based on the experimental results, it may be due to the close relationship between the channel attention mechanism and the distribution of the number of channels in the convolutional layers. The function of the STC-P attention module can be represented by the following equation,

$$f_{st} = \theta\left(\left(pool_t\left(f_{in}\right) \oplus pool_v\left(f_{in}\right)\right) \cdot w\right) \tag{9}$$

$$f_{stj} = f_{in} \odot \left(\sigma\left(f_{st} \cdot w_t\right) \otimes \sigma\left(f_{st} \cdot w_v\right)\right) \tag{10}$$

$$f_c = \sigma\left(w_k\left(w_l \cdot pool_{tv}\left(f_{in}\right)\right)\right) \odot f_{in} \tag{11}$$

$$f_{out} = \emptyset\left(BN\left(f_{stj} \oplus f_c\right)\right) \tag{12}$$

where f_{stj} is the spatial-temporal joint attention fraction, f_c is the channel attention fraction, and f_{out} is the three parallel attention features of the final output. w, w_t and w_v are the corresponding weights, which are updated with backpropagation, and θ , σ and $\emptyset$ are activation functions.

Additionally, we designed two other variants of STC-P, referred to as STC-P2 and STC-P3, to compare their performance with the proposed STC-P attention mechanism. This comparison aims to validate the importance of the joint relationship between spatial, temporal, and channel attention for action recognition. The design diagrams are shown in the figure 4 right.

Experiments

In this section, we conducted experimental evaluations of the proposed FRF-GCN model on three large-scale datasets. We also performed extensive ablation experiments to validate the effectiveness of the proposed components. To reduce the complexity of the experiments, unless otherwise stated, the experiments in the ablation study were conducted only on the CS evaluation metric of the NTURGB-D 60 dataset for validation purposes.

Datasets

NTU RGB+D 60 NTU RGB+D 60 dataset (Shahroudy et al. 2016), which is a large-scale skeleton dataset used for human action recognition models. It contains a total of 56,880 action sequences, 60 action categories, performed by different people, of which 40 actions are daily behaviors, 9 health-related actions, and 11 two-person interaction behaviors. The dataset is evaluated using two different evaluation protocols: Cross-Subject and Cross-View.

NTU RGB+D 120 The NTU RGB+D 120 dataset (Liu et al. 2019) is a supplement to the NTU RGB+D 60 dataset. The evaluation is conducted using two protocols: Cross-Subject and Cross-Setup. In the Cross-Subject evaluation, the 106 participants are divided into training and testing sets, with each set containing 53 subjects. In the Cross-Setup evaluation, action sequences from even-numbered setup IDs are used for training, while action sequences from odd-numbered setup IDs are used for testing.

Kinetics-Skeleton 400 Kinetics-Skeleton 400 (Carreira and Zisserman 2017) is a very large dataset for behavior recognition. It contains 400 actions, each video lasts about 10 seconds, and includes both indoor and outdoor, which not only has a wide variety of actions but also interacts with the scene, making it challenging to correctly identify behavior categories. The data were processed in the same way as 2S-AGCN (Shi et al. 2019b), and the experiments provided the Top-1 and Top-5 accuracies of the model.

J+B	J_M+B_M	Acc(%)	Param(M)
I	$I + A_{ske}$	89.57	2.38
$I + A_{ske}$	I	90.03	2.38
$I + A_{ske}$	$I + A_{ske}$	90.19	2.38
$I + A_{ske}$	A_{JB}	90.45	2.40
A_{JB}	$I + A_{ske}$	91.19	2.40
A_{JB}	A_{JB}	91.18	2.42
A_{JBM}	A_{JB}	91.10	2.42
A_{JBM}	A_{JBM}	90.90	2.42
A_{JB}**(ours)**	A_{JBM}**(ours)**	**91.29**	**2.42**

Table 1: Selection and comparison of adjacency matrix for different fusion flows

Attention mechanism	Acc(%)	Param(M)
STC-P	**91.29**	**2.42**
STC-P2	90.95	2.42
STC-P3	91.18	2.44
ST-joint-Att (Song et al. 2022)	89.72	2.42

Table 2: Performance Comparison of Different Attention Mechanisms(ntu60 cs)

Model	Acc(%)	Param(M)
Conventional TCN	90.13	6.94
MD-TGCN	**91.29**	**2.42**

Table 3: Comparison between different temporal convolutions

Category	Acc(%)	Param(M)
Joint	88.55	1.21
Bone	89.63	1.21
Joint_motion	86.37	1.21
Bone_motion	86.23	1.21
BF-GCN	91.56	4.84
FF-GCN	90.03	1.21
Joint+Bone	89.93	1.21
Joint_motion+Bone_motion	86.79	1.21
FRF-GCN	**91.29**	**2.42**

Table 4: Comparison between different fusion strategies

Implementation Details

All experiments were conducted using the PyTorch deep learning framework. The stochastic gradient descent (SGD) with Nesterov momentum (0.9) was employed as the optimization strategy, with a batch size of 56. The cross-entropy function was selected as the loss function for backpropagation gradients. A weight decay of 0.0003 and a learning rate of 0.1 were set. For NTU60,120 and Kinetics-Skeleton, the total epochs were set to 50, 60, and 65, respectively. The two-phase auto-attenuation renditions were 30, 40, 30, 50, 45, and 55, respectively.

Unlike most of the previous studies, FRF-GCN uses a preprocessing method of data interpolation to supplement the skeleton sequence to 300 frames. Defaults to two individuals per sample, supplemented with zeros if there is only one. Additionally, a warm-up strategy (He et al. 2016) was applied during the first 5 epochs to enhance training stability. All experiments were conducted on RTX 3080 GPUs.

Ablation Studies

Adjacency Matrix Targeting In Table 1, the effect of using A_{JBM} for all information is worse than using A_{JB} for all information, which shows that the establishment of more relations does not necessarily make the recognition better, and the selection of a suitable adjacency matrix is very important. The best performance is obtained with the targeted adjacency matrix used in FRF-GCN, which fully reflects the complementary nature of the concerns generated by using targeted adjacency matrices, again in agreement with our analysis. Because P_k is a plot unique to each sample and exists by default, it is no longer shown in the Table 1.

Comparison of the Effects of Different Attention Mechanisms As mentioned in Model architecture, STC-P2 reduces the fusion step of temporal and spatial features, while keeping the same number of parameters in the attention module. However, this modification results in a certain degree of performance degradation, indicating the importance of joint processing of temporal and spatial information. STC-P3 fuses temporal, spatial and channel attention and then extracts the attention between the three separately. The recognition rate decreases slightly compared with STC-P, and there is a slight increase in the number of model parameters, which indicates that channel attention plays a moderating role in the learning of joint spatial-temporal attention. But if the weight of the channel attention score is too large, it will hinders the normal learning of spatial-temporal attention and makes the performance decrease.

Compared to ST-joint-Att, STC-P significantly improves the performance of FRF-GCN while almost not increasing the parameter count.

The Temporal Depth-Point Convolutional Layer As shown in Table 3, after replacing the conventional temporal convolution with multi-field depth-point convolution, the number of parameters of the model is reduced to about 35% of the original one and the performance is further improved. This demonstrates the effectiveness of fusing different temporal sensory fields and the efficiency of depth-point convolution.

Comparison Among Different Fusion Strategies Table 4 shows the comparison between different fusion strategies, where BF-GCN refers to a pure backward fusion strategy and FF-GCN refers to a pure forward fusion strategy.

From Table 4, we can conclude that the performance of these two forward fusion stream information is complementary. FRF-GCN makes a trade-off between performance and number of parameters, and consumes only half the number of parameters of the former while losing very little accuracy compared to BF-GCN, which further reflects the advantages of the front-rear dual fusion strategy.

Algorithm	Cross Subject(%)	Cross View(%)	X-sub120	X-set120	Param(M)
ST-GCN(Yan, Xiong, and Lin 2018)	81.5	88.3	–	–	3.10
AS-GCN(Li et al. 2019)	86.8	94.2	–	–	9.50
DGNN(Shi et al. 2019a)	89.9	96.1	–	–	26.24
2S-AGCN(Shi et al. 2019b)	88.5	95.1	–	–	6.94
SGN(Zhang et al. 2020)	89.0	94.5	79.2	81.5	0.69
4S-Shift-GCN(Cheng et al. 2020)	90.7	96.5	85.9	87.6	2.76
MS-G3D(Liu et al. 2020)	91.5	96.2	86.9	88.4	6.40
MS-AAGCN(Shi et al. 2020)	90.0	96.2	–	–	3.77
Dynamic-GCN(Ye et al. 2020)	91.5	96.0	87.3	88.6	14.40
AdaSGN(Shi et al. 2021)	90.5	95.3	85.9	86.8	2.05
SEFN(Kong, Deng, and Jiang 2021)	90.7	96.4	86.2	87.8	34.7
Graph2Net(Wu, Wu, and Kittler 2021)	90.1	96.0	86.0	87.6	0.9
MST-GCN(Chen et al. 2021d)	91.5	96.6	87.5	88.8	12.00
FR-AGCN(Hu et al. 2022)	90.5	95.8	86.6	87.0	13.88
EfficientGCN-B0(Song et al. 2022)	90.2	94.9	86.6	85.0	0.29
SMotif-GCN+TBs(Wen et al. 2022)	90.5	96.1	87.1	87.7	–
ASE-GCN(Xiong et al. 2022)	89.4	96.2	–	–	6.00
4s-ACE-Ens(Qin et al. 2022)	91.6	96.3	88.2	89.2	5.80
ML-STGNet(Zhu et al. 2022)	91.9	96.2	88.6	90.0	5.76
2M-STGCN(Zhang et al. 2023)	90.8	96.2	–	–	–
4s STF-Net(Wu, Zhang, and Zou 2023)	91.1	96.5	86.5	88.2	6.80
FRF-GCN(ours)	**91.3**	**96.5**	**87.1**	**88.4**	**2.42**

Table 5: Performance comparison with various methods on NTURGB+D dataset and NTURGB+D 120 dataset

Algorithm	Top-1	Top-5
ST-GCN(Yan, Xiong, and Lin 2018)	30.7	52.8
AS-GCN(Li et al. 2019)	34.8	56.5
2S-AGCN(Shi et al. 2019b)	36.1	58.7
MS-G3D(Liu et al. 2020)	38.0	60.9
MST-GCN(Chen et al. 2021d)	38.1	60.8
SMotif-GCN+TBs(Wen et al. 2022)	37.8	60.6
ASE-GCN(Xiong et al. 2022)	36.9	59.7
ML-STGNet(Zhu et al. 2022)	38.9	62.2
2M-STGCN(Zhang et al. 2023)	39.0	61.6
STF-Net(Wu, Zhang, and Zou 2023)	36.1	58.9
FRF-GCN(ours)	**37.9**	**60.7**

Table 6: Performance comparison with various methods on Kinetics-Skeleton dataset

Comparisons With SOTA Methods

Apart from the direct comparison with other methods, Table 5 highlights three methods that deserve our special attention: ST-GCN, MS-G3D, and EfficientGCN. ST-GCN, has become a baseline model for many subsequent research studies and has had a significant impact. Our FRF-GCN improves accuracy by 9.8% and 8.2% relative to ST-GCN on two different evaluation criteria, CS and CV, respectively, while reducing model complexity by about 22%. MS-G3D, as one of the state-of-the-art methods in the industry. It achieves slightly higher recognition accuracy compared to FRF-GCN. However, FRF-GCN has significantly fewer parameters, achieving comparable performance with approximately 38% of the parameter count.

EfficientGCN, as one of our baselines, to maintain variable consistency, we will only compare FRF-GCN with EfficientGCN-B0 (without composite scaling strategy). From Table 5, it can be observed that although our model has a higher parameter count compared to EfficientGCN-B0, FRF-GCN outperforms it in terms of performance. Moreover, FRF-GCN's input stream information is easier to obtain compared to EfficientGCN-B0, and the forward fusion process is simpler and more practical. It is worth noting that not only does FRF-GCN exhibit competitive performance compared to multiple current SOTA methods, but it also has a lightweight overall model and a lower parameter count.

Table 5 also demonstrates the competitiveness of FRF-GCN on the NTURGB+D120 dataset compared to mainstream methods. NTU 120 is a larger dataset than NTU 60, which contains more subtle actions as well as similar actions. 4s-ACE-Ens and ML-STGNet employed ingenious design for this feature to improve performance on this dataset, enable them to outperform FRF-GCN on NTU 120. The performance results of each model on Kinetics-Skeleton 400 are shown in Table 6, and the results show that FRF-GCN is still superior for action classes with less regularity.

Conclusion

In this work, we have developed a novel lightweight model, FRF-GCN, for skeleton-based action recognition. It aligns well with the targeted adjacency matrices. As a result, FRF-GCN achieves state-of-the-art recognition results on three large-scale action recognition datasets. Although the model works better, FRF-GCN is not enough to discriminate extremely similar actions, the next work will continue to research in this direction, such as multimodal data fusion and the design of more complex attention mechanisms.

Acknowledgments

This work is supported by the Natural National Science Foundation of China (62071472), and the Fundamental Research Funds for the Central Universities (2020ZDPYMS26).

References

Bai, R.; Li, M.; Meng, B.; Li, F.; Jiang, M.; Ren, J.; and Sun, D. 2022. Hierarchical graph convolutional skeleton transformer for action recognition. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, 01–06. IEEE.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308.

Chen, T.; Zhou, D.; Wang, J.; Wang, S.; Guan, Y.; He, X.; and Ding, E. 2021a. Learning multi-granular spatio-temporal graph network for skeleton-based action recognition. In *Proceedings of the 29th ACM international conference on multimedia*, 4334–4342.

Chen, T.; Zhou, D.; Wang, J.; Wang, S.; Guan, Y.; He, X.; and Ding, E. 2021b. Learning multi-granular spatio-temporal graph network for skeleton-based action recognition. In *Proceedings of the 29th ACM international conference on multimedia*, 4334–4342.

Chen, Y.; Zhang, Z.; Yuan, C.; Li, B.; Deng, Y.; and Hu, W. 2021c. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, 13359–13368.

Chen, Z.; Li, S.; Yang, B.; Li, Q.; and Liu, H. 2021d. Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 1113–1122.

Cheng, K.; Zhang, Y.; He, X.; Chen, W.; Cheng, J.; and Lu, H. 2020. Skeleton-based action recognition with shift graph convolutional network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 183–192.

Fang, Z.; Zhang, X.; Cao, T.; Zheng, Y.; and Sun, M. 2022. Spatial-temporal slowfast graph convolutional network for skeleton-based action recognition. *IET Computer Vision*, 16(3): 205–217.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Hu, Z.; Pan, Z.; Wang, Q.; Yu, L.; and Fei, S. 2022. Forward-reverse adaptive graph convolutional networks for skeleton-based action recognition. *Neurocomputing*, 492: 624–636.

Kong, J.; Deng, H.; and Jiang, M. 2021. Symmetrical enhanced fusion network for skeleton-based action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(11): 4394–4408.

Lange, J.; and Lappe, M. 2006. A model of biological motion perception from configural form cues. *Journal of Neuroscience*, 26(11): 2894–2906.

Li, M.; Chen, S.; Chen, X.; Zhang, Y.; Wang, Y.; and Tian, Q. 2019. Actional-structural graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3595–3603.

Liu, J.; Shahroudy, A.; Perez, M.; Wang, G.; Duan, L.-Y.; and Kot, A. C. 2019. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10): 2684–2701.

Liu, Y.; and Wang, X. 2020. The analysis of driver's behavioral tendency under different emotional states based on a Bayesian Network. *IEEE Transactions on Affective Computing*.

Liu, Y.; Zhang, H.; Xu, D.; and He, K. 2022a. Graph transformer network with temporal kernel attention for skeleton-based action recognition. *Knowledge-Based Systems*, 240: 108146.

Liu, Y.; Zhang, H.; Xu, D.; and He, K. 2022b. Graph transformer network with temporal kernel attention for skeleton-based action recognition. *Knowledge-Based Systems*, 240: 108146.

Liu, Z.; Cheng, Q.; Song, C.; and Cheng, J. 2023. Cross-scale cascade transformer for multimodal human action recognition. *Pattern Recognition Letters*, 168: 17–23.

Liu, Z.; Zhang, H.; Chen, Z.; Wang, Z.; and Ouyang, W. 2020. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 143–152.

Mazzia, V.; Angarano, S.; Salvetti, F.; Angelini, F.; and Chiaberge, M. 2022. Action Transformer: A self-attention model for short-time pose-based human action recognition. *Pattern Recognition*, 124: 108487.

Qin, Z.; Liu, Y.; Ji, P.; Kim, D.; Wang, L.; McKay, R.; Anwar, S.; and Gedeon, T. 2022. Fusing higher-order features in graph neural networks for skeleton-based action recognition. *IEEE Transactions on Neural Networks and Learning Systems*.

Saleh, K.; Mihaita, A.-S.; Yu, K.; and Chen, F. 2022. Real-time Attention-Augmented Spatio-Temporal Networks for Video-based Driver Activity Recognition. In *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, 1579–1585. IEEE.

Shahroudy, A.; Liu, J.; Ng, T.-T.; and Wang, G. 2016. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1010–1019.

Sheng, W.; and Li, X. 2021. Multi-task learning for gait-based identity recognition and emotion recognition using attention enhanced temporal graph convolutional network. *Pattern Recognition*, 114: 107868.

Shi, L.; Zhang, Y.; Cheng, J.; and Lu, H. 2019a. Skeleton-based action recognition with directed graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7912–7921.

Shi, L.; Zhang, Y.; Cheng, J.; and Lu, H. 2019b. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12026–12035.

Shi, L.; Zhang, Y.; Cheng, J.; and Lu, H. 2020. Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. *IEEE Transactions on Image Processing*, 29: 9532–9545.

Shi, L.; Zhang, Y.; Cheng, J.; and Lu, H. 2021. Adasgn: Adapting joint number and model size for efficient skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13413–13422.

Song, Y.-F.; Zhang, Z.; Shan, C.; and Wang, L. 2020. Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition. In *proceedings of the 28th ACM international conference on multimedia*, 1625–1633.

Song, Y.-F.; Zhang, Z.; Shan, C.; and Wang, L. 2022. Constructing stronger and faster baselines for skeleton-based action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 45(2): 1474–1488.

Sun, Y.; Huang, H.; Yun, X.; Yang, B.; and Dong, K. 2022. Triplet attention multiple spacetime-semantic graph convolutional network for skeleton-based action recognition. *Applied Intelligence*, 52(1): 113–126.

Wen, Y.-H.; Gao, L.; Fu, H.; Zhang, F.-L.; Xia, S.; and Liu, Y.-J. 2022. Motif-GCNs with local and non-local temporal blocks for skeleton-based action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2): 2009–2023.

Wu, C.; Wu, X.-J.; and Kittler, J. 2021. Graph2Net: Perceptually-enriched graph learning for skeleton-based action recognition. *IEEE transactions on circuits and systems for video technology*, 32(4): 2120–2132.

Wu, L.; Zhang, C.; and Zou, Y. 2023. SpatioTemporal focus for skeleton-based action recognition. *Pattern Recognition*, 136: 109231.

Xin, W.; Liu, R.; Liu, Y.; Chen, Y.; Yu, W.; and Miao, Q. 2023. Transformer for Skeleton-based action recognition: A review of recent advances. *Neurocomputing*.

Xiong, X.; Min, W.; Wang, Q.; and Zha, C. 2022. Human skeleton feature optimizer and adaptive structure enhancement graph convolution network for action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(1): 342–353.

Yan, S.; Xiong, Y.; and Lin, D. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Yang, H.; Yan, D.; Zhang, L.; Sun, Y.; Li, D.; and Maybank, S. J. 2021. Feedback graph convolutional network for skeleton-based action recognition. *IEEE Transactions on Image Processing*, 31: 164–175.

Ye, F.; Pu, S.; Zhong, Q.; Li, C.; Xie, D.; and Tang, H. 2020. Dynamic gcn: Context-enriched topology learning for skeleton-based action recognition. In *Proceedings of the 28th ACM international conference on multimedia*, 55–63.

Zhang, H.; Liu, X.; Yu, D.; Guan, L.; Wang, D.; Ma, C.; and Hu, Z. 2023. Skeleton-based action recognition with multi-stream, multi-scale dilated spatial-temporal graph convolution network. *Applied Intelligence*, 1–15.

Zhang, P.; Lan, C.; Zeng, W.; Xing, J.; Xue, J.; and Zheng, N. 2020. Semantics-guided neural networks for efficient skeleton-based human action recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1112–1121.

Zhou, H.; Xiang, X.; Qiu, Y.; and Liu, X. 2023. Graph convolutional network with STC attention and adaptive normalization for skeleton-based action recognition. *The Imaging Science Journal*, 1–11.

Zhu, Y.; Shuai, H.; Liu, G.; and Liu, Q. 2022. Multilevel Spatial–Temporal Excited Graph Network for Skeleton-Based Action Recognition. *IEEE Transactions on Image Processing*, 32: 496–508.