SAGN: Semantic Adaptive Graph Network for Skeleton-Based Human Action Recognition

Ziwang Fu
Beijing University of Posts and Telecommunications
Beijing, China
fzw150355@163.com

Feng Liu*
East China Normal University
Shanghai, China
lsttoy@163.com

Jiahao Zhang
East China Normal University
Shanghai, China
zjh20000218@163.com

Hanyang Wang
East China Normal University
Shanghai, China
faceeyes@126.com

Chengyi Yang
Shanghai University of International Business and Economics
Shanghai, China
hifipsysta@163.com

Qing Xu
Beijing University of Posts and Telecommunications
Beijing, China
xqing@bupt.edu.cn

Jiayin Qi*
Shanghai University of International Business and Economics
Shanghai, China
qijiayin@139.com

Xiangling Fu*
Beijing University of Posts and Telecommunications
Beijing, China
fuxiangling@bupt.edu.cn

Aimin Zhou
East China Normal University
Shanghai, China
amzhou@cs.ecnu.edu.cn

ABSTRACT

With the continuous development and popularity of depth cameras, skeleton-based human action recognition has attracted people's wide attention. Graph Convolutional Network (GCN) has achieved remarkable performance. However, the existing methods do not better consider the semantic characteristics, which can help to express the current concept and scene information. Semantic information can also help with better granularity classification. In addition, most of the existing models require a lot of computation. What's more, adaptive GCN can automatically learn the graph structure and consider the connections between joints. In this paper, we propose a relatively less computationally intensive model, which combines semantic and adaptive graph network (SAGN) for skeleton-based human action recognition. Specifically, we mainly combine the dynamic characteristics and bone information to extract the data, taking the correlation between semantics into the model. In the training process, SAGN includes an adaptive network so that we can make attention mechanism more flexible. We design the Convolutional Neural Network (CNN) for feature extraction on the time dimension. The experimental results show that SAGN achieves the state-of-the-art performance on NTU-RGB+D 60 and NTU-RGB+D 120 datasets. SAGN can promote the study of skeleton-based human action recognition. The source code is available at https://github.com/skeletonNN/SAGN.

CCS CONCEPTS

• **Computing methodologies** → **Activity recognition and understanding**.

KEYWORDS

Skeleton-Based Human Action Recognition; Semantic Information; Data fusion; Adaptive GCN

ACM Reference Format:
Ziwang Fu, Feng Liu, Jiahao Zhang, Hanyang Wang, Chengyi Yang, Qing Xu, Jiayin Qi, Xiangling Fu, and Aimin Zhou. 2021. SAGN: Semantic Adaptive Graph Network for Skeleton-Based Human Action Recognition. In *Proceedings of the 2021 International Conference on Multimedia Retrieval (ICMR '21), August 21–24, 2021, Taipei, Taiwan.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3460426.3463633

1 INTRODUCTION

With the development of artificial intelligence and the improvement of compute capability, the research on skeleton-based human action recognition [22] has become a hot and very challenging research issue. Human action recognition has a wide range of applications, such as video supervision [34], intelligent monitoring [29] and human-computer interaction [3, 41]. Existing action recognition methods usually use RGB video or skeleton data. RGB video contains intuitive scene information, but it is easily affected by lighting, occlusion, etc. Skeleton data is a topological representation of human joints and bones. With the continuous development of depth sensors and human pose estimation technology, we can easily obtain accurate human skeleton data. The skeleton joint data represents the three-dimensional coordinates of the human body

*Corresponding author.

ICMR '21, August 21–24, 2021, Taipei, Taiwan

ACM ISBN 978-1-4503-8463-6/21/08…$15.00
https://doi.org/10.1145/3460426.3463633

joint points in the video frame, including the spatial structure information of the skeleton and the dynamic information of the time sequence. At the same time skeleton joint data is a good way to avoid light, shelter and complex background interference.

Early skeleton-based action recognition methods [5, 32] are mainly characterized by manual design of features. However, these manual features only work well on some specific datasets and may not be portable to other datasets. With the development and excellent performance of deep learning methods in other computer vision tasks, Convolutional Neural Networks (CNN) [9, 13], Recurrent Neural Networks (RNN) [12, 31, 35] and Graph Convolutional Networks (GCN) [11, 20, 24, 26, 36, 39] are starting to emerge. The skeleton sequence is the natural time-series sequence of joint nodes, and RNN is more suitable for processing time series data. Therefore, there are more skeleton action recognition methods based on RNN and its variants (for example, Long Short-Term Memory (LSTM) and Gated Recycler Unit (GRU)). The CNN network model usually converts the bone data on a slice into a pseudo image, and then uses CNN to extract the corresponding deep learning features from the pseudo image. When CNN processes skeleton data series, it usually needs to combine RNN model. The combination of the temporal context information of RNN and the abundant spatial information of CNN can often achieve better results than the single structural model. In the past two years, many scholars have begun to apply GCN to the skeleton-based action recognition. GCN is the fusion of graph theory and convolutional neural network, and its essential purpose is to extract the relational features of topology graphs. The human skeleton sequence is a natural topology structure, and the GCN network model is more suitable to describe the spatial and temporal topological information between the key points of the skeleton. Therefore, GCN has an advantage over RNN when handling skeleton data tasks. For CNN, the extraction of features in the temporal scale has a very good effect on the existing models, so in this paper, we mainly adopt the combination of GCN and CNN.

Most of the existing work [33] does not consider semantic features. Semantics can help understand the current specific information and pinpoint the specific joint. For example, feet and heads have different meanings. At the same time, the semantic information can help accurately distinguish between two actions, such as two actions with similar ranges of variation. If combined with the semantic information, two actions can have a better granularity classification. In addition, the current model has a large computational cost, which is not conducive to the implementation of the product. The previous work of GCN was more about fixed graph structure, which lacked flexibility and might lose implicit correlation. A fixed network structure is not optimal for different types of actions of different samples.

Therefore, To address the limitations of the above methods, we propose a semantic and adaptive graph network (SAGN) with relatively little computation. Figure 1 shows the overall framework. Figure 3 shows the number of parameters and the accuracy of different models on NTU-RGB+D 60(X-Sub). Specifically, we fuse the dynamic feature information (position information, motion information and velocity information) with the bone information (bone information and bone information based on velocity difference) and add the semantic information (joint type and frame index). The whole information fusion is carried out by means of splicing. The fused data is put into the three-layer adaptive GCN for learning, highlighting its adaptability and dynamic learning of graph structure. And the adjacency matrix of graph convolution is obtained by self-learning. The structure learned through the adaptive GCN is spliced with the semantic information of the frame index. In the process of CNN, three layers are designed, which are mainly used for feature extraction of temporal scale. The features of the nodes are pooled into a frame, and the semantic features of the frame index are fused in the way of splicing to learn advanced features. In the whole process, we test the model on two public datasets, NTU-RGB+D 60 [23] and NTU-RGB+D 120 [14], and get a better result. We verify the effectiveness of the three innovation points of semantic information, adaptive GCN and the fusion method of splicing through ablation experiments. The proposed model achieves the best result considering the number of parameters and accuracy.

We summarize our three main contributions as follows:

- We propose a semantic and adaptive graph network (SAGN) for skeleton-based human action recognition, which innovatively combines dynamic feature information, bone information and semantic information through splicing and adaptively learns the graph structure dynamically for each sample.
- We design two modules, three-layer adaptive GCN and temporal scale based triple CNN, to efficiently learn the spatial structure and temporal structure of the skeleton sequence.
- We achieve an absolute advantage in balancing the number of parameters and accuracy. At the same time, we conduct experiments on two public datasets and the proposed SAGN model achieves the best performance with an order of magnitude smaller model size.

2 RELATED WORK

With the development of deep learning, the current research on skeleton-based action recognition is mainly divided into the following three categories:

RNN forms a recursive join within its structure by taking the output of the previous moment as the input of the current moment. Hong and Liang [31] proposed a two stream RNN structure to model temporal and spatial characteristics of skeleton data. Chunyu and Baochang [35] used attention RNN and CNN models to improve complex space-time modeling. To solve the problem of gradient disappearance in RNN, Shuai and Wanqing [12] proposed an independent cyclic neural network, which can be used to process longer sequences. [17] considered that not all joints are useful for behavior analysis, global Contex-aware attention is added to the LSTM network to selectively focus on the information-rich joints in the skeleton sequence.

CNN can effectively and easily learn advanced features by virtue of its excellent advanced information extraction capability. However, CNN usually focuses on image tasks. Therefore, in order to meet the input requirements of CNN, skeleton data will be converted into pseudo images, so that the pseudo images have spatio-temporal information at the same time. Bo and Mingyi [9] used a translation-invariant image mapping strategy, first dividing the human skeleton joints of each frame into five major parts based on the body object structure, and then mapping these parts into 2D

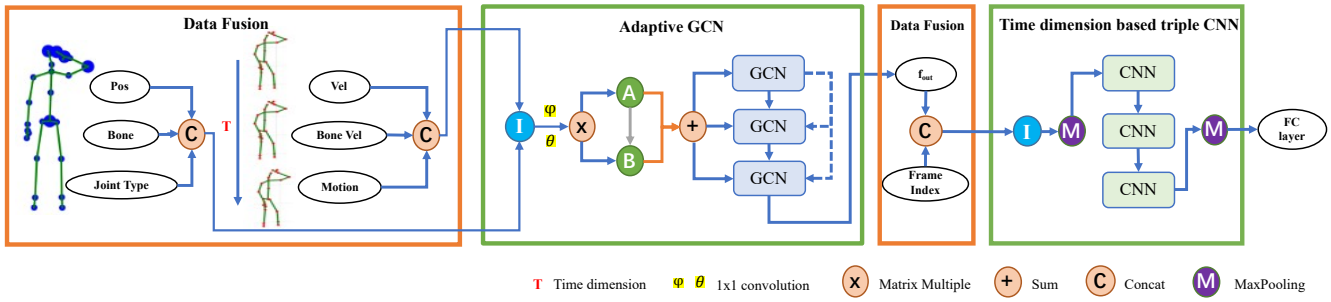

Figure 1: Framework of the proposed semantic and adaptive graph network (SAGN). In the data fusion part, we fuse dynamic feature information, bone information and the joint types in the semantic information, and the fusion is carried out through splicing. In the adaptive GCN part, we designed a three-layer graph convolution structure, and learned the graph structure of each sample using the idea of residuals. We splice the learned graph structure and frame index in semantic information to get new fusion data. In the time dimension based triple CNN part, we design a three-layer convolution operation with two MaxPooling.

form. Yanshan and Rongjie [13] came up with shape-motion representaion from geometric algebra, which addresses the importance of joints and bones and makes full use of the information provided by skeleton sequences. CNN has outstanding advantages in feature extraction, so this paper draws on CNN's advanced feature extraction capability.

GCN: The human 3D skeleton data is the topology of nature. Sijie and Yuanjun [36] first proposed a new model based on skeleton action recognition - spatio temporal graph convolutional network (ST-GCN). The Action Structural Graph Convolutional Network (AS-GCN) [11] proposed by Maose and Siheng could not only identify human actions, but also output the next possible pose of the target by using the multi-tasking learning strategy. 2s-AGCN [26] proposed an adaptive double-flow graph convolutional network structure, which allows the addition of new connections other than natural connections to dynamically adjust the graph structure to better fit the model hierarchy. NAS [20] improves graph structure by using polynomials to obtain a larger receptive field, and provides a cross-entropy method with import-mixing (CEIM) search strategy. Directed acyclic graph neural network (DGNN) [24] could not only extract the information of nodes and bones, but also extract the directional correlation information between them. SGN [39] put forward semantic information for the first time and integrates joint information and semantic information. MS-G3D [19] proposes a unified space-time graph convolution operator G3D, which uses dense edges across space and time as jump connections to directly propagate information in the space-time graph. 2s-Shift-GCN [1] is an extension of shift convolution operator on graph structure data. ST-TR [21] is based on the dual-flow Transformer based model. Self-attention is used to model the dependency relationship between joints in both spatial and temporal dimensions. In the future, we will conduct further research on skeleton action recognition based on visual Transformer.

The above three methods are the mainstream ones based on skeleton behavior recognition at present, but they all have some problems. Although the method based on RNN has obvious advantages in the representation of time information, it has the problems of high optimization difficulty and easy to lose the original node information. Although the CNN method can extract specific multi-scale local patterns from different time intervals, it has the problem of too large number of parameters and too high requirement for calculation. The method based on GCN benefits from the great advantages of non-Euclidean data modeling and is more advantageous than the former two methods. In our work, we combine semantics and adaptive GCN to learn spatial characteristics and CNN to extract features.

3 SEMANTIC ADAPTIVE GRAPH NETWORK

For a given skeleton sequence, we divide the skeleton data into three parts for data fusion. The three parts are dynamic feature information, bone information and semantic information. Dynamic characteristics include joints information, motion information and velocity difference information. Bone data includes bone information and bone information based on velocity difference. Semantic information includes joint type and frame index. Different from SGN [39], these data are fused by splicing.

We propose a semantic adaptive graph network (SAGN) for skeleton-based human action recognition. The whole structure of our model is shown in figure 1, which is divided into three parts. Next, we will describe the above work in detail.

In detail, for a given skeleton sequence, we define all the joints as a set $S = \{X_t^v | t = 1, 2, 3, \ldots T; v = 1, 2, 3, \ldots V\}$. T represents the total number of frames in the sequence and V represents the total number of joint points. X_t^v represents the joint v at time t.

3.1 Data fusion

The skeleton data is divided into three parts, which are dynamic characteristic information, bone information and semantic information.

Dynamic features are physical features that indicate the concept of motion and position information of an object. In this study, we selected the coordinate information of the joint $P_t^v = (x_t^v, y_t^v, z_t^v)$. The joint coordinate can indicate the space position in the t frame. Secondly, the motion information is obtained by calculating the difference between two adjacent frames of the same joint. We define the coordinate of the joint v in frame t as $M_t^v = (x_t^v, y_t^v, z_t^v)$, and the

coordinate of the joint v in frame $t+1$ as $M_{t+1}^v = (x_{t+1}^v, y_{t+1}^v, z_{t+1}^v)$. And we define the motion information as:

$$M_t^v = M_{t+1}^v - M_t^v = (x_{t+1}^v, y_{t+1}^v, z_{t+1}^v) - (x_t^v, y_t^v, z_t^v) \quad (1)$$

The velocity information represents the difference between the previous $T-1$ frame and the next $T-1$ frame. The specific formula is as:

$$V = P_{1:T}^v - P_{0:T-1}^v \quad (2)$$

Similar to the work of SGN [39], we embed the above three kinds of information(taking joint information as an example):

$$\widetilde{P_t^v} = \sigma(W_2\sigma(W_1 P_t^v + b_1) + b_2) \quad (3)$$

$\widetilde{P_t^v}$ represents joint features after embed. Motion information and velocity information are also encoded in this way as $\widetilde{M_t^v}$ and $\widetilde{V}$.

Bone information has been proven to be effective in previous work [26], so bone information is included in this paper. The bone information takes the center of gravity of the human body as the source joint $B_t^v = (x_t^v, y_t^v, z_t^v)$ and other nodes as the target joint $B_t^k = (x_t^k, y_t^k, z_t^k)$. The bone information is calculated by the difference between the source joint and the target joint. The specific formula is as follows:

$$B_t^v = B_t^k - B_t^v = (x_t^k - x_t^v, y_t^k - y_t^v, z_t^k - z_t^v) \quad (4)$$

Based on the bone information of the velocity difference, the difference information of the former T-1 frame and the subsequent T-1 frame of the bone information is calculated as $B = B_{1:T}^v - B_{0:T-1}^v$. Similar to the encoding above, the two messages as $\widetilde{B_t^v}$ and $\widetilde{B}$.

Semantic information includes the joint type and the frame index. We splice the information obtained. These two semantics describe the spatial structure and the temporal structure, which are very representative. Joint type can help identify specific joint during classification. For example, when two movements change from the bottom to the top, knowing the joint type can make a better distinction between the two movements to improve the classification accuracy. Frame index can help to obtain important frame information throughout the training process. For example, for the same motion of two joints, the range of change is different when the frame index is different in size. The extraction of these two semantic information is similar to SGN [39], but we change the addition into splicing in the process. Splicing will not lose some feature information and can integrate features more effectively. At the same time, splicing may lead to excessive number of parameters, so we added a layer of convolution after splicing to keep the number of parameters within a certain range. In the GCN stage, only the semantic information of joint type is used, and in the CNN stage, the frame index information is used. We carry out one-hot coding for these two parts. The resulting vector is encoded as $\widetilde{S}$ and $\widetilde{T}$.

The data mentioned above are stitched together to get the fused data(BN represents BatchNorm):

$$I = \sigma(BN(W_1[\widetilde{P_t^v}, \widetilde{M_t^v}, \widetilde{V}, \widetilde{B_t^v}, \widetilde{B}, \widetilde{S}]) + b_1) \quad (5)$$

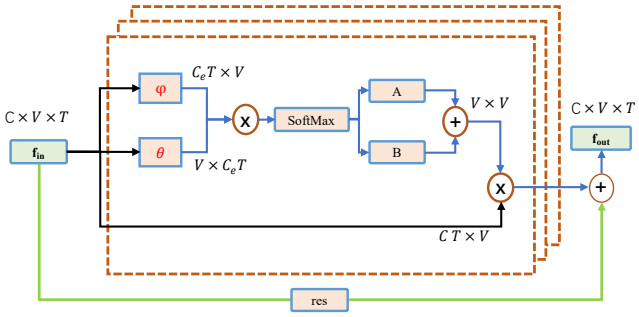

Figure 2: The working mechanism of adaptive GCN. There are two types of adjacency matrices, θ and φ changes the dimensions of the input data f_{in} from $C \times V \times T$ to $V \times C_e * T$ and $C_e * T \times V$ through transformation. The adjacency matrix A and B are obtained by matrix multiplication. Among them, the dimension of B is the same as that of B, but other parameters in the B network are optimized for training without any constraints and can be any element. The result of adding A and B is multiplied by the original data to get the graph structure, and the final f_{out} is obtained through the residual structure. The dimension of f_{out} is the same as the dimension of the input data f_{in}.

3.2 Adaptive GCN

The fusion data I is put into the adaptive GCN. The adjacency matrix A is calculated according to the fused data, and the adjacency matrix is obtained by the following formula:

$$A = softmax(\theta(I)^T * \varphi(I)) \quad (6)$$

where θ and φ denote two transformation functions. A learns a graph for each sample. The specific process is shown in figure 2. θ changes the data dimension of I from $C \times V \times T$ to $V \times C_t * T$ and φ changes the data dimension of I to $C_t * T \times V$ and then it multiplies these two matrices together to get an $V \times V$ similar matrix A. A_k^{ij} represents the similarity between joints v_i and v_j.

After the adjacency matrix is obtained, GCN can be carried out. The operation of graph convolution is similar to that of 2S-AGCN [26], and the formula is as follows:

$$f_{out} = \sum W_k I(A + B) \quad (7)$$

Where A represents the adjacency matrix just obtained. B is a $V \times V$ matrix, it doesn't have any constraints, it can be any element. B will train with other parameters in the network to optimize the instrument. B can not only strengthen the connection between joints, but also make the connection between joints that are not related. For example, the posture of hugging, the movements of the two arms are similar, and there is a certain correlation. Using this random parameter B, the joints that have no natural structural association of the human body can be connected together.

In this paper, we design a three-layer series GCN, which utilizes the residual. We learn the unique adjacency matrix for each sample, and put the matrix into the parameter list for training. The whole process embodies its self-adaptability.

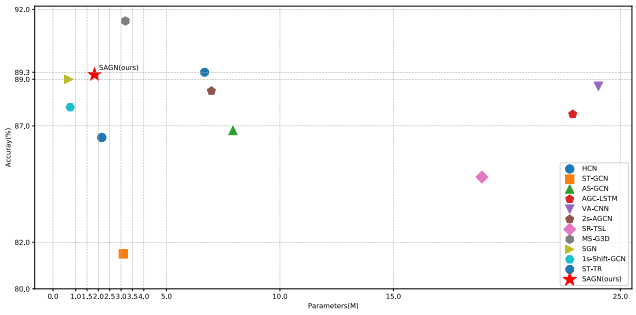

Figure 3: : Comparisons of different methods on NTU-RGB+D 60 (X-Sub) in terms of accuracy and the number of parameters. The proposed SAGN model achieves the best performance with an order of magnitude smaller model size.

3.3 Temporal scale based triple CNN

We denote the result obtained by adaptive GCN as f_{out}. We fuse the graph features f_{out} and frame index $\widetilde{T}$ to get the new data $I = [f_{out}, \widetilde{T}]$. The way of fusion is also through splicing. Dimensions are then changed, similar to formula (5). The resulting data uses the spatial pooling layer to convert the joint dimensions to 1.

In the CNN process, we set three layers, in which the first and third layers are used to extract features, the second layer is used to enhance dimensions, and finally the whole process is followed by a time pooling layer to convert the temporal scale to 1. Finally, we design the full connection layer with Softmax for classification output.

4 EXPERIMENTS

4.1 Datasets

In this paper, we use two current mainstream human skeleton datasets: NTU-RGB+D 60 and NTU-RGB+D 120.

The NTU-RGB+D 60 is a public 3D dataset of human actions.This dataset contains 60 actions captured simultaneously by three Microsoft Kinect V2 cameras. This dataset collects 25 joint points, 17 camera placement combinations, and consists of 56,880 action videos. The total of 40 actors are included to perform action classification. [23] provides two benchmarks for partitioning dataset: X-Sub and X-View.

- X-Sub represents that the actions of the training set and the verification set come from different actors, where the sample size of the training set is 40,320 and the sample size of the test set is 16,560.
- X-View indicates that 37,920 videos from Camera 1 and 3 are used as the training set, and 18,960 videos from Camera 2 are used as the verification set.

NTU-RGB+D 120 dataset is an extension of the NTU-RGB+D 60, and it is also a public dataset for 3D human action recognition. The data and camera placement combinations are expanded from 17 to 32, the action categories are expanded from 60 to 120, the number of actors is expanded to 106, and the action segments are expanded to 114,480, while the joint points remain unchanged. [14] provides two benchmarks for partitioning dataset: X-Sub and X-View. X-Sub divides this dataset into training set (63,026 videos) and verification set (50,919 videos) according to the different participants in the video. X-View divides this dataset according to the video number. The even numbered videos are taken as the training set (7,582 videos), and the odd numbered videos are taken as the test set (38,122 videos).

4.2 Implementation Details

In the data fusion stage, the data is encoded from 3 to 64 dimensions. The encoded multi-modal data is spliced, and the data dimension obtained by splicing is very high. Therefore, Formula 5 is used to reduce the dimension to 128. Since the stitching of the semantic information of frame index is after the obtained graph structure, the coding of frame index is directly from 3 dimension to 256 dimension. In the Adaptive GCN, the dimensions of the data are 128,256 and 256 respectively. After that, the graph structure and the semantic information of frame index obtained will be spliced, and the spliced result will be pooled into 1 dimension on the joint dimension. During the ime dimension based triple CNN, the dimensions of the data are 256,512 and 512 respectively, and the whole process is followed by a time pooling layer to convert the temporal scale to 1. Finally, we use the full connection layer to output with Softmax, and the output dimension is the corresponding number of categories.

During the experiment, we set the number of method epoch to 120, and set the batch size of one epoch to 64, with an initial learning rate of 0.001, and continuously decreasing during the iteration. When the number of iterations is 60,90,110, the learning rate is multiplied by 0.1. In order to save computing resources and improve computational efficiency, we choose Adam to optimize the model, where the weight decay is 0.0001. In order to prevent overfitting, Dropout is added to the training and set to 0.2. In the whole process, the activation function used is the Relu function. Cross entropy loss is used to train the networks. All the models are trained by a Tesla k80 GPU.

In data preprocessing, we discard the missing data and denoise the existing data. Denoising includes two parts : frame length-based and spread-based. In the process, if a frame contains two people, we divide a frame into two frames so that each sequence is one person. In addition, we randomly divide the skeleton sequence into 20 segments, and randomly select a town from each segment to obtain a new sequence of 20 frames.

4.3 Ablation Study

4.3.1 Effectiveness of Bone and Semantic information. Bone information can indicate changes in relative position and velocity of different joints and source joints. Semantic information can categorize actions more precisely. This paper selects two semantic information of joint type and frame index. In order to verify the validity of the two types of information, we design an ablation experiment to prove it. We perfrom various experiments on the NTU-RGB+D 60 dataset. We use *wo* to represent without and *w* to represent with, where the semantic information is abbreviated as *SI*, and the bone information is abbreviated as *BI*. Table 1 shows the comparisons.

From table 1, We have three main observations as follows:

Table 1: Effectiveness of bone and semantic information on NTU-RGB+D 60 dataset in terms of accuracy(%), *BI* represents bone information and *SI* represents semantic information.

Method	Params(M)	X-Sub	X-View
SAGN w BI wo SI	1.66	88.0	93.1
SAGN w SI wo BI	1.81	87.5	93.1
SAGN w BI w SI	**1.83**	**89.2**	**94.2**

Table 2: Effectiveness of Splicing on NTU-RGB+D 60 dataset in terms of accuracy(%), *P* represents dynamic characteristic information, *B* represents bone information and *S* represents semantic information.

Method	Params(M)	X-Sub	X-View
SAGN P+B+S	1.66	88.7	93.5
SAGN [P, B, S]	**1.83**	**89.2**	**94.2**

- According to the experimental results of "SAGN w SI wo BI" and "SAGN w BI w SI", adding bone information during data fusion can increase 1.7% and 1.1% on the X-Sub and X-View settings. It turns out that bone information is very helpful for describing skeletal sequence, and the integration of bone information can describe the relative position of the joints.
- According to the experimental results of "SAGN w BI wo SI" and "SAGN w BI w SI", semantic information can help increase 1.2% and 1.1% on the X-Sub and X-View settings. The results show that adding semantic information in the training process can enhance the network understanding ability and improve the accuracy of the effect.
- According to the experimental results of "SAGN w BI wo SI" and "SAGN w SI wo BI", the accuracy of bone information is 0.5% higher than semantic information on X-Sub setting. The results show that the representation of bone information is higher in the process of network learning. Interestingly, the performance of the two information on X-View is the same, indicating that with the increase of training samples, the performance of the two is flat.

4.3.2 Effectiveness of Splicing. In this paper, we choose the splicing method to integrate the dynamic feature information, bone information and semantic information. Splicing can save complete information, and splicing features is more conducive to deeper network learning. In order to verify the effectiveness of splicing, we designed an ablation experiment. We use an additive method to express the fusion of information as $P + B + S$. We express the method of information fusion using splicing as $[P, B, S]$. And the experimental results obtained are shown in Table 2.

As shown in Table 2, compared with addition, splicing can increase by 0.5% and 0.7% on the X-Sub and X-View settings. The results show that some features are lost during addition. Splicing can represent feature information more broadly and accurately.

Table 3: Effectiveness of Adaptive GCN on NTU-RGB+D 60 dataset in terms of accuracy(%), *B* represents the adjacency matrix.

Method	Params(M)	X-Sub	X-View
SAGN wo B	1.83	89.1	94.0
SAGN w B	**1.83**	**89.2**	**94.2**

Table 4: Performance comparisons on NTU-RGB+D 60 with the X-Sub and X-View settings in terms of accuracy (%).

Method	Year	X-Sub	X-View
VA-LSTM [37]	2017	79.4	87.6
GCA-LSTM [17]	2017	74.4	82.8
Clips+CNN+MTLN [6]	2017	79.6	84.8
DPRL+GCNN [30]	2018	83.5	89.8
ElAtt-GRU [40]	2018	80.7	88.4
ST-GCN [36]	2018	81.5	88.3
SR-TSL [28]	2018	84.8	92.4
HCN [10]	2018	86.5	91.1
AGC-LSTM [27]	2019	87.5	93.5
AS-GCN [11]	2019	86.8	94.2
GR-GCN [4]	2019	87.5	94.3
2s-AGCN [26]	2019	88.5	95.1
VA-CNN [38]	2019	88.7	94.3
SGN [39]	2020	89.0	94.5
SAN [2]	2020	87.2	92.7
MS-G3D [19]	2020	**91.5**	**96.2**
DGCNN [25]	2020	89.9	96.1
2s Shift-GCN [1]	2020	89.7	96.0
ST-TR [21]	2020	89.3	96.1
SAGN(Ours)	2021	89.2	94.2

4.3.3 Effectiveness of Adaptive GCN. In the process of adaptive GCN, we add the B matrix, where the B matrix is mainly used to help strengthen the connection between the joints. In order to verify the effectiveness of B, We divide the model into without B and with B to observe the experimental results. The results are shown in Table 3.

As shown in Table 3, the adjacency matrix B with the dynamic learning graph structure can improve the accuracy of SAGN without changing the number of parameters, increasing by 0.1% and 0.2% in the X-Sub and X-View, respectively.

4.4 Comparison with the State-of-the-arts

We compare the proposed SGN with other state-of-the-art methods on the NTU-RGB+D 60, NTU-RGB+D 120 in Table 4, Table 5. As shown in table 4, the accuracy of SAGN on X-Sub and X-View evaluation criteria on NTU-RGB+D 60 dataset is 89.2 % and 94.2 %, respectively. The proposed SAGN has outstanding performance compared with the existing methods on the X-Sub setting. ElAtt-GRU [40] are one representative method for RNN-based methods, respectively. SAGN outperforms it by 8.5% in accuracy for the X-Sub, respectively. [28, 36] mix LSTM and GCN, or CNN and GCN

Table 5: Performance comparisons on NTU-RGB+D 120 with the X-Sub and X-View settings in terms of accuracy (%).

Method	Year	X-Sub	X-View
ST-LSTM [15]	2016	55.7	57.9
GCA-LSTM [17]	2017	58.3	59.2
Clips+CNN+MTLN [7]	2017	58.4	57.9
Two-Stream GCA-LSTM [16]	2017	61.2	63.3
RotClips+MTCNN [8]	2018	62.2	61.8
Body Pose Evolution Map [18]	2018	64.6	66.9
1s Shift-GCN [1]	2020	80.9	83.2
SGN [39]	2020	79.2	81.5
SAGN	2021	**82.1**	**83.8**

together. SAGN also outperforms them by 4.4% and 5.7% in accuracy for the X-Sub. Compared with [26] and [38], although the accuracy of the X-View setting is a little bit lower than them, the number of our parameters is very low. The number of parameters in [26] is $6.98M$, and the number of parameters in [38] is $24.03M$. Figure 3 shows the number of parameters and accuracy of different models on NTU-RGB+D 60(X-Sub). The SAGN proposed only has $1.83M$, which is more conducive to implementation. The proposed model achieves the best result considering the number of parameters and accuracy. Compared with SGN, our proposed method improves the X-Sub setting by 0.2%.

We also compare the experimental results on the NTU-RGB+D 120 dataset, as shown in table 5. The accuracy of X-Sub and X-View on the dataset is 82.1 % and 83.8 %, respectively. Compared with SGN [39] method, it increased by 2.9 % and 2.3 % respectively.

5 CONCLUSION AND FUTURE WORK

In this work, we propose a semantic adaptive graph network (SAGN) for skeleton-based human action recognition. This method is mainly divided into three parts: data fusion, adaptive GCN and temporal scale based triple CNN. We fuse the dynamic feature information, bone information and semantic information by splicing. We design a workflow that dynamically learns the graph structure based on skeleton data and adds an adjacency matrix that enhances the correlation between joints. In the whole experiment process, this paper designs three innovation points: the integration of bone information, fusion data using splicing, adaptive graph network structure. We design ablation experiments for innovation points to prove the effectiveness and necessity. At the end of this paper, we compare this method with the State-of-the-arts and the results show that SAGN has achieved the state-of-the-art results on three benchmark datasets. However, this paper still has some deficiencies, and we plan to complete it in the future work. Firstly, this paper does not fully consider the connection between semantics. This paper does not elaborate on the different encoding methods of semantics. We plan to improve the different encodings of semantic information in future work. Secondly, the number of method parameters can be further reduced, and we will continue to improve the effectiveness of the method for different applications. Finally, we hope that this work can accelerate the development of action recognition.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant 82071171, in part by Beijing Municipal Natural Science Foundation under Grant L192026, in part by 2019, Digital Transformation in China and Germany: Strategies, Structures and Solutions for Ageing Societies, GZ 1570.

REFERENCES

[1] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu. 2020. Skeleton-Based Action Recognition With Shift Graph Convolutional Network. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 180–189. https://doi.org/10.1109/CVPR42600.2020.00026

[2] Sangwoo Cho, Muhammad Hasan Maqbool, Fei Liu, and Hassan Foroosh. 2019. Self-Attention Network for Skeleton-based Human Action Recognition. arXiv:1912.08435 [cs.CV]

[3] Chris Ellis, Syed Zain, Masood Marshall, F. Tappen, Joseph LaViola, and Rahul Sukthankar. 2012. Exploring the Trade-off Between Accuracy and Observational Latency in Action Recognition. *International Journal of Computer Vision* 101 (02 2012). https://doi.org/10.1007/s11263-012-0550-7

[4] Xiang Gao, Wei Hu, Jiaxiang Tang, Jiaying Liu, and Zongming Guo. 2019. Optimized Skeleton-based Action Recognition via Sparsified Graph Regression. (2019). arXiv:1811.12013 [cs.CV]

[5] J. Hu, W. Zheng, J. Lai, and Jianguo Zhang. 2015. Jointly learning heterogeneous features for RGB-D activity recognition. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5344–5352. https://doi.org/10.1109/CVPR.2015.7299172

[6] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid. 2017. A New Representation of Skeleton Sequences for 3D Action Recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 4570–4579. https://doi.org/10.1109/CVPR.2017.486

[7] Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. 2017. A New Representation of Skeleton Sequences for 3D Action Recognition. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Jul 2017). https://doi.org/10.1109/cvpr.2017.486

[8] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid. 2018. Learning Clip Representations for Skeleton-Based 3D Action Recognition. *IEEE Transactions on Image Processing* 27, 6 (2018), 2842–2855. https://doi.org/10.1109/TIP.2018.2812099

[9] Bo Li, Mingyi He, Xuelian Cheng, Yucheng Chen, and Yuchao Dai. 2017. Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep cnn. (2017). arXiv:1704.05645 [cs.CV]

[10] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. 2018. Co-occurrence Feature Learning from Skeleton Data for Action Recognition and Detection with Hierarchical Aggregation. (2018). arXiv:1804.06055 [cs.CV]

[11] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. 2019. Actional-Structural Graph Convolutional Networks for Skeleton-based Action Recognition. (2019). arXiv:1904.12659 [cs.CV]

[12] Shuai Li, Wanqing Li, Chris Cook, Ce Zhu, and Yanbo Gao. 2018. Independently Recurrent Neural Network (IndRNN): Building A Longer and Deeper RNN. (2018). arXiv:1803.04831 [cs.CV]

[13] Y. Li, R. Xia, X. Liu, and Q. Huang. 2019. Learning Shape-Motion Representations from Geometric Algebra Spatio-Temporal Model for Skeleton-Based Action Recognition. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*. 1066–1071. https://doi.org/10.1109/ICME.2019.00187

[14] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. 2020. NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 10 (Oct 2020), 2684–2701. https://doi.org/10.1109/tpami.2019.2916873

[15] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. 2016. Spatio-Temporal LSTM with Trust Gates for 3D Human Action Recognition. (2016).

[16] Jun Liu, Gang Wang, Ling-Yu Duan, Kamila Abdiyeva, and Alex C. Kot. 2018. Skeleton-Based Human Action Recognition with Global Context-Aware Attention LSTM Networks. (2018). arXiv:1707.05740 [cs.CV]

[17] J. Liu, G. Wang, P. Hu, L. Duan, and A. C. Kot. 2017. Global Context-Aware Attention LSTM Networks for 3D Action Recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3671–3680. https://doi.org/10.1109/CVPR.2017.391

[18] M. Liu and J. Yuan. 2018. Recognizing Human Actions as the Evolution of Pose Estimation Maps. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1159–1168. https://doi.org/10.1109/CVPR.2018.00127

[19] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang. 2020. Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 140–149. https://doi.org/10.1109/CVPR42600.2020.00022

[20] Wei Peng, Xiaopeng Hong, Haoyu Chen, and Guoying Zhao. 2019. Learning Graph Convolutional Network for Skeleton-based Human Action Recognition by Neural Searching. (2019). arXiv:1911.04131 [cs.CV]

[21] Chiara Plizzari, Marco Cannici, and Matteo Matteucci. 2020. Skeleton-based Action Recognition via Spatial and Temporal Transformer Networks. arXiv:2008.07404 [cs.CV]

[22] Bin Ren, Mengyuan Liu, Runwei Ding, and Hong Liu. 2020. A Survey on 3D Skeleton-Based Action Recognition Using Learning Method. (2020). arXiv:2002.05907 [cs.CV]

[23] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. 2016. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. (2016). arXiv:1604.02808 [cs.CV]

[24] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. 2019. Skeleton-Based Action Recognition With Directed Graph Neural Networks. (June 2019).

[25] L. Shi, Y. Zhang, J. Cheng, and H. Lu. 2019. Skeleton-Based Action Recognition With Directed Graph Neural Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7904–7913. https://doi.org/10.1109/CVPR.2019.00810

[26] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. 2019. Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition. (2019). arXiv:1805.07694 [cs.CV]

[27] Chenyang Si, Wentao Chen, Wei Wang, Liang Wang, and Tieniu Tan. 2019. An Attention Enhanced Graph Convolutional LSTM Network for Skeleton-Based Action Recognition. (2019). arXiv:1902.09130 [cs.CV]

[28] Chenyang Si, Ya Jing, Wei Wang, Liang Wang, and Tieniu Tan. 2018. Skeleton-Based Action Recognition with Spatial Reasoning and Temporal Stack Learning. (2018). arXiv:1805.02335 [cs.CV]

[29] Wataru Takano and Yoshihiko Nakamura. 2015. Statistical mutual conversion between whole body motion primitives and linguistic sentences for human motions. *International Journal of Robotics Research* 34 (09 2015), 1314–1328. https://doi.org/10.1177/0278364915587923

[30] Y. Tang, Y. Tian, J. Lu, P. Li, and J. Zhou. 2018. Deep Progressive Reinforcement Learning for Skeleton-Based Action Recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5323–5332. https://doi.org/10.1109/CVPR.2018.00558

[31] Hongsong Wang and Liang Wang. 2017. Modeling Temporal Dynamics and Spatial Configurations of Actions Using Two-Stream Recurrent Neural Networks. (2017). arXiv:1704.02581 [cs.CV]

[32] Lei Wang, Du Q. Huynh, and Piotr Koniusz. 2020. A Comparative Review of Recent Kinect-Based Action Recognition Algorithms. *IEEE Transactions on Image Processing* 29 (2020), 15–28. https://doi.org/10.1109/tip.2019.2925285

[33] Pichao Wang, Wanqing Li, Philip Ogunbona, Jun Wan, and Sergio Escalera. 2018. RGB-D-based Human Motion Recognition with Deep Learning: A Survey. (2018). arXiv:1711.08362 [cs.CV]

[34] Xiaogang Wang. 2013. Intelligent Multi-Camera Video Surveillance: A Review. *Pattern Recogn. Lett.* 34, 1 (Jan. 2013), 3–19. https://doi.org/10.1016/j.patrec.2012.07.005

[35] Chunyu Xie, Ce Li, Baochang Zhang, Chen Chen, Jungong Han, Changqing Zou, and Jianzhuang Liu. 2018. Memory Attention Networks for Skeleton-based Action Recognition. (2018). arXiv:1804.08254 [cs.CV]

[36] Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. (2018). arXiv:1801.07455 [cs.CV]

[37] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. 2017. View Adaptive Recurrent Neural Networks for High Performance Human Action Recognition from Skeleton Data. arXiv:1703.08274 [cs.CV]

[38] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng. 2019. View Adaptive Neural Networks for High Performance Skeleton-Based Human Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 8 (2019), 1963–1978. https://doi.org/10.1109/TPAMI.2019.2896631

[39] Pengfei Zhang, Cuiling Lan, Wenjun Zeng, Junliang Xing, Jianru Xue, and Nanning Zheng. 2020. Semantics-Guided Neural Networks for Efficient Skeleton-Based Human Action Recognition. (2020). arXiv:1904.01189 [cs.CV]

[40] Pengfei Zhang, Jianru Xue, Cuiling Lan, Wenjun Zeng, Zhanning Gao, and Nanning Zheng. 2018. Adding Attentiveness to the Neurons in Recurrent Neural Networks. In *Computer Vision – ECCV 2018*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.). Springer International Publishing, Cham, 136–152.

[41] Wenhao Zhang, Melvyn L. Smith, Lyndon N. Smith, and Abdul Farooq. 2016. Gender and Gaze Gesture Recognition for Human-Computer Interaction. *Comput. Vis. Image Underst.* 149, C (Aug. 2016), 32–50. https://doi.org/10.1016/j.cviu.2016.03.014