Electronics, Communications and Networks
A.J. Tallón-Ballesteros et al. (Eds.)

doi:10.3233/FAIA231231

Multi-Skeleton Joint Behavior Recognition Method Based on a Dual-Stream Network

Hongda MOU [a,b], Lianggui TANG [a,b1], Jing RAN [a], Xue WANG [a,b], Song PENG [a,b], Xiuling ZHU [a] and Chan TANG [a]
[a] *College of Artificial Intelligence, Chongqing Technology and Business University, China*
[b] *Chongqing Key Laboratory of Intelligent Perception and Blockchain Technology, China*

Abstract. Skeleton joint data, as a representation closely related to human actions, has received significant attention in the field of human behavior recognition in recent years. However, current methods for behavior recognition using skeleton data face several challenges, such as insufficient accuracy and reliance on a single modality. To address these challenges, a multi-stream network approach for behavior recognition using multiple skeleton joints is proposed. Firstly, a top-down approach is employed to extract the coordinates of 133 key skeleton joints from the human body, which are then represented as heatmaps. Subsequently, in the PoseRgb model, RGB data is used as input for the RGB channel to capture spatial features, while the heatmap representations of skeleton data are utilized as input for the Pose channel to capture temporal features. Leveraging the fusion capabilities of heatmaps, these two channels are merged using lateral connections, ultimately yielding behavior recognition results. Experimental results demonstrate that on widely-used UCF101 and HMDB51 datasets, the accuracy reaches 99.05% and 86.1%, respectively. Compared to the Temporal Shift Module (TSM) network, there is a 5% and 15% improvement in accuracy, respectively. The proposed method effectively combines spatial and temporal information in videos, resulting in more accurate behavior recognition.

Keywords. Behavior recognition; Bone data; Heat map; Spatial features; Time series Features; Multimodal

1. Introduction

In action recognition, a key challenge is interpreting video data. Various feature representation modalities have been explored in existing research. For instance, the RGB[1-2] modality not only contains contextual information about the scene but is also easy to capture, making it the most commonly used data modality. However, RGB data still suffers from the problem of containing excessive redundant information, including a large amount of unnecessary details. The optical flow[3] modality builds upon RGB to capture motion dynamics, and although it has shown significant improvements in recognition, extracting optical flow requires substantial time and computational resources. Skeletal joints[4-6], as a representation of human actions, have gained importance in video understanding. Skeletal joints can express motion information

[1] *Corresponding author: TLG@CTBU.EDU.CN.*

efficiently while focusing on the motion itself, giving them a unique advantage. in recent years, there has been a growing recognition of the importance of the skeletal data modality. Skeletal modality not only effectively conveys motion information but also offers advantages such as lightweight representation and resilience to factors like lighting and background context.

Currently, Graph Convolutional Networks (GCN)[5] are one of the most popular methods for action recognition based on skeletal data. The general equation for graph convolution can be defined as a nonlinear function H:

$$H^{(l+1)} = f(H^l, A) \tag{1}$$

Where $H^0 = X, X \in R^{(N*D)}$ is the input for the first layer, N is the number of nodes in the graph, D is the dimension of each node's feature vector, A is the adjacency matrix, and the function f represents the specific form of the model.

From the above formula, it can be seen that GCN directly processes the coordinates of human joints, and therefore its recognition capability is significantly affected by the distribution of coordinate offsets. At the same time, GCN operates on irregular skeletal graphs, making it difficult to integrate with other information (such as optical flow and RGB data) for complementary information. In addition, since GCN treats each person's joint as a node, its complexity scales linearly with the number of people, which limits its applicability in multi-person scenarios.

2. Related Work

In the field of deep learning for video understanding, the main focus lies in extracting and analyzing features from three data modalities: RGB images, optical flow information, and skeletal data. Currently, research on these three data modalities has become a primary direction in the field of action recognition.

RGB data modality: As the most readily available data format, it initially garnered significant attention in research and applications. In reference[7], a dual-channel approach utilizing RGB images was proposed. In this approach, the contextual channel receives downsampled frames, which are half of the original spatial resolution, while the focus channel receives the central region, maintaining the original resolution. However, researchers soon realized that using RGB modality alone could only capture spatial information, whereas action recognition requires the fusion of both motion and temporal information. To address this issue, reference[8] employed the LSTM gated network model to learn temporal information in the data. The specific approach involved first using a Convolutional Neural Network (CNN) to extract features from images and then using the CNN's output as input for LSTM, thereby capturing information across both time and space, ultimately enhancing recognition accuracy. Additionally, 3D networks naturally incorporate temporal information compared to 2D networks. In reference[9], it is elucidated how to leverage 3D networks by expanding already designed 2D networks into 3D. This approach not only avoids additional network structure design but also significantly improves accuracy, as 3D networks inherently possess temporal information.

Optical flow modality data: Compared to RGB, optical flow data inherently focuses on capturing motion dynamics. Reference[10] not only introduced optical flow modality but also combined it with LSTM to capture temporal information. On the other hand, reference[3] proposed a two-stream network that doesn't require the model

to explicitly learn temporal features. Instead, they use optical flow early in the network to extract features, designing two identical backbone networks, one for RGB and one for optical flow modality, leveraging the complementary nature of these two modalities, significantly enhancing action recognition accuracy. However, optical flow extraction poses a significant challenge in terms of computational power and time. Reference[11] took a different approach from the two-stream network; they didn't employ two identical backbone networks to process the two data modalities. Instead, they ingeniously designed a fast pathway and a slow pathway to better handle fused information features, taking into account the differences between the two modalities, achieving improved fusion results.

Skeletal data modality: Due to its ability to perfectly represent human motion postures, skeletal data has become an excellent data source. In recent years, researchers have explored various methods for analyzing skeletal data, such as Convolutional Neural Networks (CNNs), Long Short-Term Memory networks (LSTMs), and others. However, the results have not always been satisfactory. For example, studies[12-14] explored the use of CNNs, LSTMs, and similar methods for skeletal data analysis, but the results were not significantly promising.Reference[15] proposed a method based on d-CNN, which initially maps the skeleton sequences into pseudo-images based on manually designed transformations. Then, the heatmaps are aggregated over the temporal dimension to create 2D inputs with color encodings or learning modules. Despite this method being well-designed, there are still issues of information loss during the aggregation process, leading to subpar recognition performance.Currently, graph convolution-based methods have shown the best performance in action recognition using skeletal modality. Reference[5] represents the human skeleton as a directed acyclic graph to effectively integrate skeletal and joint information. Then, they utilize Graph Convolutional Networks (GCNs) to separately learn spatial and temporal features from the skeletal modality. Although this approach has achieved some success, GCNs still have several limitations, resulting in lower recognition accuracy.

3. Model Construction

3.1. Overall framework of the network

The overall network framework of this article is divided into two main parts, as shown in Figure 1. Data preprocessing consists of two sub-stages: first, extracting RGB frame images from the video; second, generating 3D heatmaps through pose estimation and heatmap construction.

The PoseRgb network model consists of two parallel channels. The first channel takes RGB data as input and utilizes the RGB channel based on ResNet-50 (a Residual Network architecture with 50 layers)[16] for excellent spatial feature extraction. The second channel takes 3D heatmaps containing 133 skeleton nodes as input and uses the pose channel, also based on ResNet-50, to capture temporal features. Data fusion between these two channels is achieved through lateral connections. Finally, after passing through pooling layers and fully connected layers, the prediction results are obtained. Detailed network parameters are listed in Table 1, where T represents the number of frames, and S^2 represents the image size.

Table 1. Detailed parameter table for PoseRgb network

stage	Rgb pathway	Pose pathway	out ($T \times S^2$)
data layer	stride $8,1^2$	stride $32,4^2$	Rgb: 8×224^2 Pose: 32×56^2
stem layer	conv $1 \times 7^2, 64$ stride $1,2^2$ 1×3^2 maxpool stride $1,2^2$	conv $1 \times 7^2, 64$ stride $1,1^2$ 1×3^2 maxpool stride $1,1^2$	Rgb: 8×56^2 Pose: 32×56^2
res_2	$\begin{bmatrix} 1 \times 1^2, 64 \\ 1 \times 3^2, 64 \\ 1 \times 1^2, 256 \end{bmatrix} \times 3$	N.A.	Rgb: 8×56^2 Pose: 32×56^2
res_3	$\begin{bmatrix} 1 \times 1^2, 128 \\ 1 \times 3^2, 128 \\ 1 \times 1^2, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 3 \times 1^2, 32 \\ 1 \times 3^2, 32 \\ 1 \times 1^2, 128 \end{bmatrix} \times 4$	Rgb: 8×28^2 Pose: 32×28^2
res_4	$\begin{bmatrix} 3 \times 1^2, 256 \\ 1 \times 3^2, 256 \\ 1 \times 1^2, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} 3 \times 1^2, 64 \\ 1 \times 3^2, 64 \\ 1 \times 1^2, 256 \end{bmatrix} \times 6$	Rgb: 8×14^2 Pose: 32×14^2
res_5	$\begin{bmatrix} 3 \times 1^2, 512 \\ 1 \times 3^2, 512 \\ 1 \times 1^2, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 3 \times 1^2, 128 \\ 1 \times 3^2, 128 \\ 1 \times 1^2, 512 \end{bmatrix} \times 3$	Rgb: 8×7^2 Pose: 32×7^2

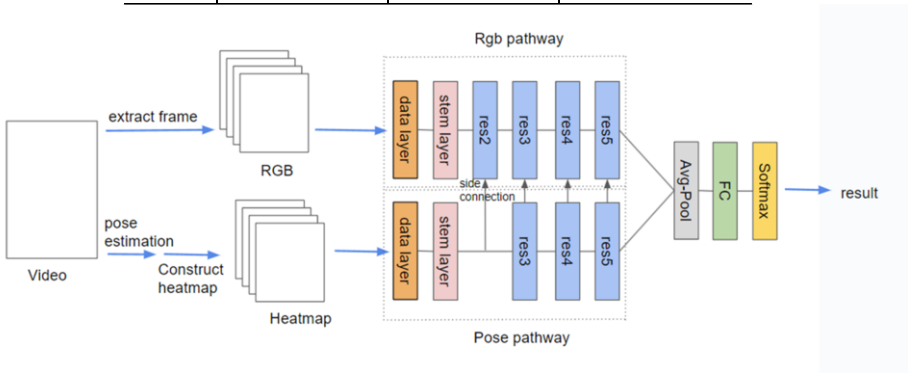
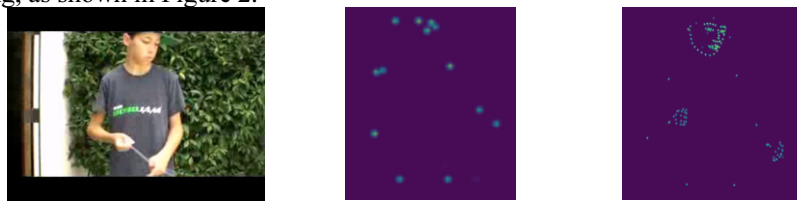


Figure 1. Overall network architecture

3.2. Pose Estimation

Currently, in the research on human skeleton keypoint detection, a common approach involves using a pose estimation scheme that includes 17 key points. While this can roughly capture the human pose, many skeleton nodes in areas such as the face and hands can exhibit significant differences when the body performs different actions. To address this issue, this paper employs the HRnet[17] network model for 2D pose estimation, extending the number of skeleton nodes to 133, thereby covering all skeletal details in areas like the face and hands. This measure ultimately achieves more

comprehensive information capture, providing more accurate data for subsequent training, as shown in Figure 2.



(a)original frame (b)17 bone nodes (c)133 bone nodes

Figure 2. Skeletal Node Visualization Comparison Chart

The algorithm for extracting the pose estimation of the skeletal nodes in the video is as follows:

Algorithm 1 2D Pose Estimation Process Based on HRnet Model

Input:Video sample V, pre-trained weight file for HRnet model.

Output:The coordinate set of 133 skeletal nodes for each frame of the video sample $\{(x_k, y_k), k \in [1,133]\}$.

step1. Perform a sampling operation on input video V to obtain a collection of video frames $S \in \{S_1,\ldots,S_k\}$, where K represents the number of sampled frames;

step2. Load the pre-trained model's weights;

step3. for $k \leftarrow 1$ to HRnet do;

step4. Obtain a set of skeletal node coordinates that can represent the human body pose.

After the pose extraction process by Algorithm 1, the video can effectively remove spatial information and focus on human motion information. This allows for efficient learning of temporal motion-related features through the Pose channel. The original video and heatmap final results are shown in Figure 3.

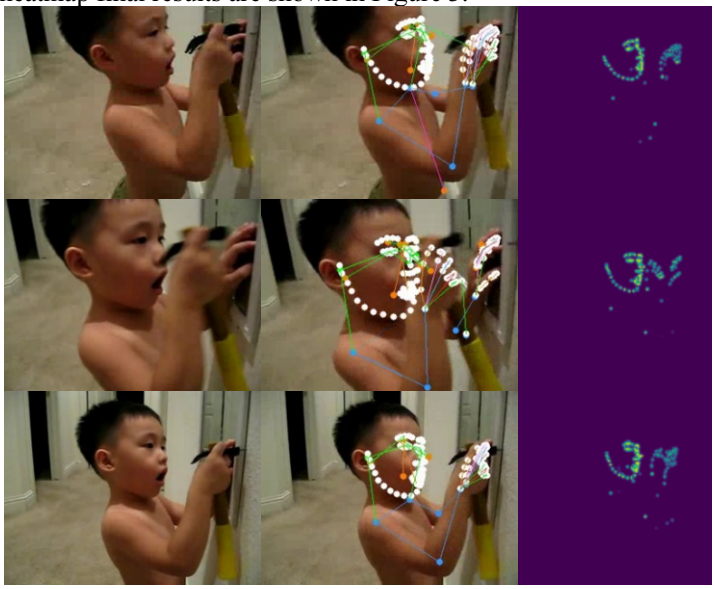


Figure 3. Pose Extraction Algorithm Visualization

3.3. Construct heat map

In the PoseRgb model proposed in this paper, there are two input data modalities. One is the RGB data modality, which can be directly extracted from the video. The second modality is the skeletal joint data extracted from the 2D pose, with specific steps as follows.

First, the 2D top-down pose estimator from reference[17] is used for pose extraction. The pose extraction result is represented as a heatmap of $K \times H \times W$, where K is the number of skeletal joints, and H and W are the height and width of the image. Skeletal joint definition is as follows: (x_k, y_k, c_k) for the K-th joint coordinate, (x_k, y_k) for the K-th joint's coordinate,and C_k is for the confidence score. confidence score. The skeletal joint heatmap G_{kij} is obtained by calculating the Gaussian distribution maps for K joints.

$$G_{kij} = \frac{1}{\sqrt{2\pi}} \times e^{-\frac{(i-x_k)^2+(j-y_k)^2}{2\times\sigma^2}} \left(c_k > \upsilon\right) \tag{2}$$

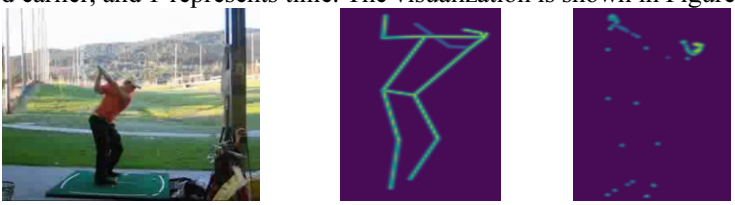
Where σ^2 represents the variance, and υ is the threshold.

Based on the Gaussian distribution, we can also obtain limb heatmaps H_{kij}.

$$H_{kij} = \frac{1}{\sqrt{2\pi}} \times e^{-\frac{D\left((i,j),seg\left[a_k,b_k\right]\right)^2}{2\times\sigma^2}} \left(a_k > \upsilon, b_k > \upsilon\right) \tag{3}$$

Where the function D is defined as the distance from point (i, j) to the line segment $\left[(x_{ak}, y_{ak}),(x_{bk}, y_{bk})\right]$, and a_k 、 b_k are the two ends of a joint.

Finally, stack all the skeletal joint heatmaps along the time dimension to obtain a three-dimensional heatmap, with $T \times K \times H \times W$ size. Here, H, W, and K are explained as mentioned earlier, and T represents time. The visualization is shown in Figure 4.



(a)original frame (b) limbs (c) bone nodes

Figure 4. Visualization of limb and bone nodes

3.4. Space-time channel fusion

Although both the RGB and Pose channels use ResNet50 as their base network architecture, these two channels have differences in their respective focuses. The characteristic of the RGB channel is that the input RGB images carry richer spatial information and possess semantic spatial features. Therefore, during runtime, it requires a slower frame rate and lower resolution to capture spatial semantic information. In other words, the RGB channel is used for extracting static spatial

features. To achieve this goal, larger strides and fewer temporal dimension convolution operations are employed.

From an abstract perspective, the RGB channel can be seen as a spatial cube, primarily responsible for extracting spatial information with fewer temporal dimensions. In contrast, the Pose channel places more emphasis on the temporal dimension and can be regarded as a temporal cube, where time information is more prominent. This concept is illustrated in Figure 5. H and W represent the height and width of the image, while Time and Channel represent the number of temporal channels and convolutional kernel channels respectively.

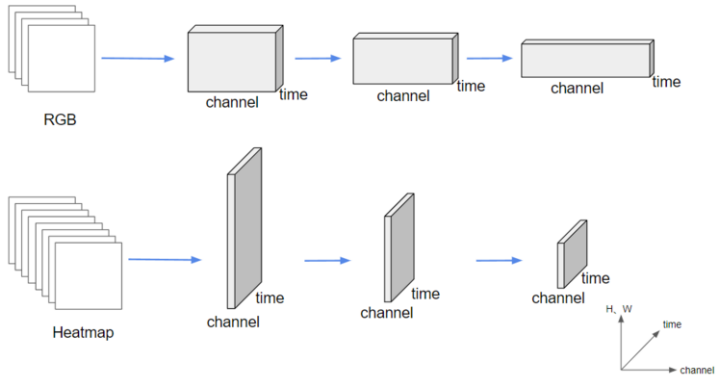


Figure 5. Dual channel focus map

Due to the situation where there is no mutual knowledge about the representation features learned by the Pose channel and the RGB channel, to address this issue, we introduce lateral connections to achieve information fusion between these two channels. In this way, we can better interlace the features of the Pose channel and the RGB channel, thereby enhancing the overall model's performance.

First, we define the feature vectors of the Pose channel F_{Pose} and the feature vectors of the RGB channel F_{RGB}, where T represents the length of the time sequence, and H and W represent height and width, respectively. S denotes the factor by which we adjust the feature vectors. We consider scaling for both the time sequence and dimensions.In practical implementation, we adjust the feature vectors of the Pose channel F_{Pose} to have the same size $F_{\text{Pose\_resized}}$ as the feature vectors of the RGB channel by applying 3D convolution operations. Specifically, we first adjust the Pose feature vector using an appropriate convolution kernel to match its size with the feature vector of the RGB channel. In other words, we perform the following two steps:

$$F_{\text{Pose\_resized}} = Conv3D(F_{Pose}, kernel\_pose)$$

Where kernel_pose represents the 3D convolution kernel used for resizing.

$$F_{RGB\_Fused} = F_{RGB} \oplus F_{Pose\_resized}$$

Here, $\oplus$ represents the concatenation operation along the channel dimension.

An example of this process is shown in Figure 6.

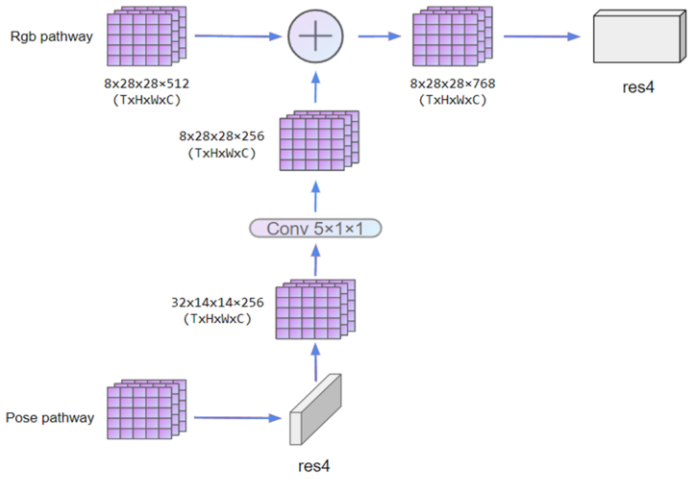


Figure 6. Example diagram of side connection fusion

4. Experimental results

The experimental environment is Ubuntu 18.04.5, the hardware platform is NVIDIA A100, and the CUDA platform version is 11.0. The program is designed and implemented using the PyTorch deep learning framework. The training parameters for the dataset are set as follows: 40 training epochs, an initial learning rate of 0.001 (which is reduced by a factor of 1/10 at the 20th and 30th epochs), and a batch size of 16.

4.1. Data set

This paper validates the proposed algorithm's effectiveness on four classic action recognition datasets: UCF101[18], HMDB51[19], NTU60[20], and Charades[21].

The UCF101 dataset consists of 101 categories and 13,320 videos. These videos are user-uploaded from YouTube and contain actions such as fast motion, interactions between multiple people, and more.

The HMDB51 dataset is collected from various sources, with the majority of clips sourced from movies and a small portion from public databases like Prelinger Archives, YouTube, and Google Videos. The dataset comprises 6,766 clips categorized into 51 action classes, with each class containing at least 100 clips.

The NTU60 dataset consists of 60 action classes and 56,880 video samples. The dataset includes RGB videos, depth map sequences, 3D skeletal data, and infrared (IR) videos for each sample. Each dataset was captured using three Kinect V2 cameras simultaneously. The RGB videos have a resolution of 1920x1080, while the depth maps and IR videos have a resolution of 512x424. The 3D skeletal data contains the 3D coordinates of 25 body joints for each frame.

Charades is a dataset comprising 9,848 videos of everyday indoor activities, collected through Amazon Mechanical Turk. 267 different users were presented with a sentence that included objects and actions from a fixed vocabulary, and they recorded a video performing that sentence (similar to a game). The dataset includes 66,500

temporal annotations for 157 action classes, 41,104 labels for 46 object classes, and 27,847 textual video descriptions.

4.2. Analysis of results

4.2.1. Spatio-temporal fusion experiment results

In this section, experiments were conducted, and mainstream methods in the field of action recognition were compared as control experiments for the four datasets mentioned above. Here are brief descriptions of these methods:TSM Model: The TSM (Temporal Shift Module) model uses 2D convolution operations and simulates 3D modeling by shifting in time. This approach effectively enhances model performance through operations in the temporal dimension, achieving high-performance video understanding at low cost;ST-GCN Network Model: The ST-GCN network model employs Temporal Convolutional Networks (TCN) for convolutional operations in the temporal dimension and utilizes Graph Convolutional Networks (GCN) for convolutional operations in the spatial dimension. This combination of temporal and spatial information achieves high recognition accuracy;LFB Network: In current recognition.

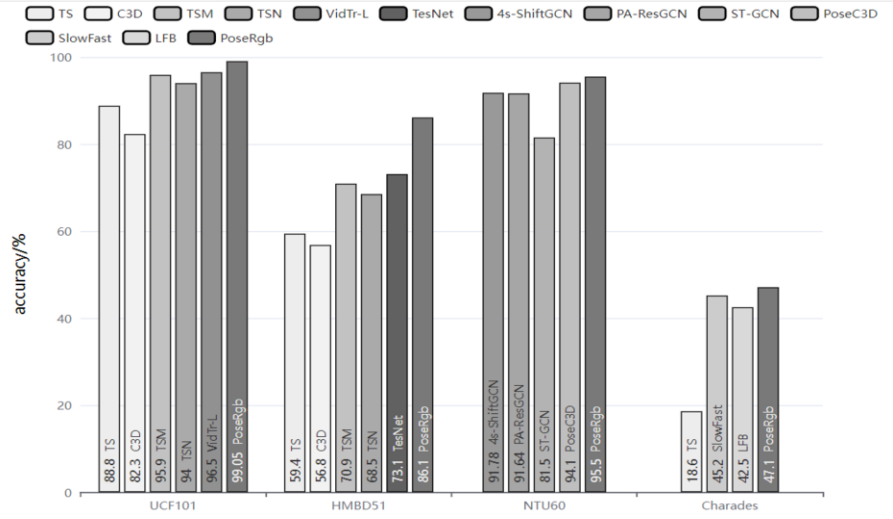


Figure 7. The result of space-time fusion experiment

The results are shown in Table 2, and the comparisons are depicted in Figure 7. The experimental results indicate that the highest accuracy was achieved on the UCF101 dataset, reaching 99.05%. For the other datasets, there were also improvements compared to other methods.

Table 2. Comparison of Different Methods on Four Datasets

dataset	method	accuracy/%
UCF101	Two-Stream[3]	88.8
	C3D[22]	82.3
	TSM[23]	95.9
	TSN[24]	94.0
	VidTr-L[25]	96.5
	PoseRgb	99.05
HMDB51	Two-Stream[3]	59.4
	C3D[22]	56.8

	TSM[23]	70.9
	TSN[24]	68.5
	TesNet[26]	73.1
	PoseRgb	86.1
NTU60	4s-ShiftGCN[27]	91.78
	PA-ResGCN[27]	91.64
	ST-GCN[28]	81.5
	PoseC3D[29]	94.1
	PoseRgb	95.5
Charades	Two-Stream[3]	18.6
	SlowFast[11]	45.2
	LFB[30]	42.5
	PoseRgb	47.1

On the UCF101 and HMDB51 datasets, the newly proposed method has demonstrated outstanding performance, surpassing classical two-stream networks, 3D convolution networks, and recent popular 2D network models. It achieved accuracy rates of 99.05% and 86.1%, respectively, showcasing its superiority on these two datasets. On the NTU60 dataset, which is specifically designed for action recognition using skeletal data, the method achieved an accuracy rate of 95.5%. This result also outperforms the current best models based on graph convolution networks and the PoseC3D model network.However, on the Charades dataset, which contains concurrent actions from real-life scenarios where temporal boundaries of actions in videos are often ambiguous, the model couldn't fully capture video features, resulting in only 18.6% accuracy for the two-stream network. Yet, through the fusion learning of skeletal and RGB modalities, the method improved accuracy to 47.1%, effectively enhancing recognition performance on these complex videos.

4.2.2. Temporal Scaling Experiments

Because the Pose channel and RGB channel have different characteristics, the input time sequences T for the channels differ by a factor of β. To explore the optimal value of β for achieving the best data fusion, experiments were conducted with β values in the set $\beta \in \{2,4,6,8,10\}$ on the UCF101 dataset.

Detailed results are presented in Table 3, and it can be observed that the model's performance reaches its peak when β = 4. Because β represents the spatiotemporal rate,, the difference in quantity between the heatmaps and RGB images. The reason β = 4 achieves optimal performance is that the Pose channel captures slow spatial changes associated with actions, while the RGB channel captures fast spatial changes.

Table 3. Time Series Multiplication Experiment Results

β	accuracy/%
2	98.58
4	99.05
6	98.90
8	98.55
10	97.57

4.3. Ablation experiment

In these experiments, the Pose channel, which uses 17 joint heatmaps as input, served as the base network. The experimental results are summarized in Table 4.When using only the Pose channel as input, the model achieved a final accuracy of 83.2%. This suggests that relying solely on temporal information is insufficient for achieving high

accuracy. However, when the Pose channel was fused with the RGB channel, accuracy increased significantly to 97.6%. This demonstrates the importance of fusing information from different sources. Building on the augmented skeletal nodes and combining the Pose and RGB channels, the final accuracy reached 99.05%. This highlights the combined effect of augmented skeletal nodes and multimodal fusion, greatly enhancing action recognition performance.

Table 4. Results of ablation experiment

method	accuracy/%
Pose pathway	83.2
Pose pathway + Skeleton Node Expansion	86.1
Pose pathway + Rgb pathway	97.6
Pose pathway + Rgb pathway + Skeleton Node Expansion	99.05

5. Conclusion

This paper has focused on the skeletal data modality and proposed the PoseRgb model, a multimodal dual-channel approach for skeletal action recognition. This method takes 3D heatmaps and original images as inputs, fusing multiple skeletal nodes and image data to enhance the effectiveness of action recognition and address the limitations of low recognition accuracy in graph convolution-based methods. Experimental results and comparative analyses demonstrate that the PoseRgb model outperforms traditional GCN models and single-modal models in terms of action recognition accuracy and efficiency. Future research can explore uncharted directions, such as comparing skeletal node extraction models and their impact on results. Additionally, efforts to optimize the network model for lightweight applications are needed to meet engineering-level requirements.

Acknowledgements

This work is partially supported by Major Scientific and Technological Research Project of Chongqing Education Commission(KJZD-M202000802)，The first batch of key industrial and information technology projects in Chongqing in 2022 (2022000537), Chongqing Business University's 2023 Graduate Research Innovation Project(yjscxx2023-211-183).

References

[1] KIM J,CHA S,WEE D, et al.Regularization on spatio-temporally smoothed feature for action recognition[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 12103-12112.

[2] JI S,XU W,YANG M, et al.3D convolutional neural networks for human action recognition[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2012,35(1): 221-231.

[3] SIMONYAN K,ZISSERMAN A.Two-stream convolutional networks for action recognition in videos[C] // Proceedings of the 27th International Conference on Neural Information Processing Systems.Cambridge: MIT Press,2014:568-576.

[4] YAN S,XIONG Y,LIN D.Spatial temporal graph convolutional networks for skeleton-based action recognition[EB/OL].[2019-01- 25]. https://arxiv. org/pdf/1801. 07455. pdf.

[5] SHI L,ZHANG Y,CHENG J,et al.Skeleton-based action recognition with multi-stream adaptive graph convolutional networks[J]. IEEE Transactions on Image Processing 29 (2020):9532-9545.

[6] YAN S,XIONG Y,LIN D.Spatial temporal graph convolutional networks for skeleton-based action recognition[EB/OL].[2019-01-25].https://arxiv. org/pdf/1801. 07455. pdf.
[7] KARPATHY A,TODERICI G,SHETTY S,et al.Large-scale video classification with convolutional neural networks[C] // Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2014: 1725-1732.
[8] DONAHUE J,ANNE HENDRICKS L, GUADARRAMA S, et al.Long-term recurrent convolutional networks for visual recognition and description[C] // Proceedings of the IEEE conference on computer vision and pattern recognition. Piscataway: IEEE, 2015: 2625-2634.
[9] VAROL G,LAPTEV I,SCHMID C.Long-term temporal convolutions for action recognition[J].IEEE Transactions on Pattern Analysis and Machine Tntelligence,2017, 40(6): 1510-1517.
[10] WU Z, WANG X, JIANG Y G, et al. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification [C]// Proceedings of the 23rd ACM International Conference on Multimedia. New York: ACM, 2015: 461-470.
[11] FEICHTENHOFER C, FAN H, MALIK J, et al. SlowFast networks for video recognition [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2019: 6201-6210.
[12] HOU, YONGHONG, ZHAOYANG L, et al. Skeleton optical spectra-based action recognition using convolutional neural networks[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2016, 28(3): 807-811.
[13] SHI, LEI, YIFAN ZHANG, et al. Skeleton-based action recognition with directed graph neural networks [C]// Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 7912-7921.
[14] DU Y, WEI W, LIANG W. Hierarchical recurrent neural network for skeleton based action recognition [C]// Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2015: 1110-1118.
[15] CHOUTAS, VASILEIOS, et al. Potion: Pose motion representation for action recognition [C]// Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 7024-7033.
[16] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 770-778.
[17] SUN K, Xiao B, Liu D, et al. Deep high-resolution representation learning for human pose estimation [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 5693-5703.
[18] SOOMRO K, ZAMIR A R, SHAH M. UCF101: A dataset of 101 human actions classes from videos in the wild[EB/OL]. [2012-04-02]. https://arxiv.org/pdf/1212.0402.pdf.
[19] KUEHNE H, JHUANG H, GARROTE E, et al. HMDB: a large video database for human motion recognition [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. Piscataway: IEEE, 2011: 2556-2563.
[20] SHAHROUDY A, LIU J, NG T T, et al. Ntu rgb+ d: A large scale dataset for 3d human activity analysis [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. Piscataway: IEEE, 2016: 1010-1019.
[21] SIGURDSSON G A, GUPTA A, SCHMID C, et al. Charades-ego: A large-scale dataset of paired third and first person videos[EB/OL]. [2018-04-025]. https://arxiv.org/pdf/1804.09626.pdf
[22] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3d convolutional networks [C]// Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE, 2015: 4489-4497.
[23] LIN J, GAN C, HAN S. TSM: Temporal shift module for efficient video understanding [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2019: 7083-7093.
[24] WANG L, XIONG Y, WANG Z, et al. Temporal segment networks: Towards good practices for deep action recognition [C]// European Conference on Computer Vision. Springer:Cham, 2016: 20-36.
[25] ZHANG, YANYI, et al. Vidtr: Video transformer without convolutions [EB/OL]. [2021-10-15]. https://arxiv.org/pdf/2104.11746.pdf.
[26] HUANG, GUOXI, ADRIAN G B. Learning spatio-temporal representations with temporal squeeze pooling [C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2020: 1-9.
[27] CHENG, KE, YIFAN ZHANG, XIANGYU HE, et al. Skeleton-based action recognition with shift graph convolutional network [C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Piscataway: IEEE, 2020:183-192.

[28] DUAN H, WANG J, CHEN K, et al. Pyskl: Towards good practices for skeleton action recognition [C]// Proceedings of the 30th ACM International Conference on Multimedia. New York: ACM, 2022: 7351-7354.
[29] DUAN H, ZHAO Y, CHEN K, et al. Revisiting skeleton-based action recognition [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 2969-2978.
[30] WU C Y, FEICHTENHOFER C, FAN H, et al. Long-term feature banks for detailed video understanding [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 284-293.