Int&Int: A Two-Pathway Network for Skeleton-Based Action Recognition

Xiangyuan Qi, Zhen He, Qiang Wang
Department of Control Science and Engineering
Harbin Institute of Technology, Harbin, China
xiangyuan202103@163.com, hezhen@hit.edu.cn, wangqiang@hit.edu.cn

***Abstract*—Human action recognition has recently attracted a lot of researchers, while most of them mainly focus on the local dynamics or global information alone and cannot focus on both the intensity and the integrity of the action. In this work, we propose a two-pathway *Int&Int* network(*Intensity&Integrity*) for skeleton-based action recognition to satisfy both aspects, where the great complementarity between the two pathways further enhances the performance. Besides, for *Integrity* pathway, we apply the uniform sampling strategy. For *Intensity* pathway, we introduce the intensity-dependent sampling, where a clip composed of consecutive frames around the frame with the largest motion intensity is sampled. Moreover, we explain various definitions of the motion intensity containing different semantic information based on the extracted 2D human poses. For each pathway, the poses are represented by a 3D heatmap volume and 3D-CNNs of both pathways have the same architecture. Late fusion is used to ensemble them. The model is evaluated on two action recognition datasets, FineGYM-99 and HMDB-51, and it achieves superior performance on both of them. The code has been shown at https://github.com/SarahQi666/Int-and-Int.**

***Index Terms*—Skeleton-based action recognition, Two-pathway network, Intensity-dependent sampling**

I. Introduction

Human action recognition has been an attractive topic for decades. Simonyan and Zisserman [1] proposed a two-stream framework where the spatial stream used video frames as input, whilst the temporal stream used optical flow. [2] introduced the strategies of trajectory-constrained sampling and pooling to encode deep features into effective descriptors. [3] used 3D CNNs trained on a large scale supervised video dataset to learn the spatiotemporal features from raw videos in an end-to-end learning framework. As an extension of C3D [3], [4] conducted a ConvNet architecture searching across multiple dimensions. The model in [5] involved a slow pathway which operated at low frame rate to capture spatial semantics, and a fast pathway which operated at high frame rate to capture motion at fine temporal resolution.

However, these networks mainly focused on local dynamics of the action. Having noticed the problem, TSN [6] extracted short snippets over a long video sequence, where the samples were distributed uniformly. [7] was designed to learn and reason about temporal dependencies between video frames at multiple time scales. [8] was a form of temporal aggregation of features sparsely sampled over the whole video using feature map aggregation techniques. ActionVLAD [9] integrated the classic two-stream framework [1] with learnable spatiotemporal feature aggregation, which used a vocabulary of action words into a single video-level fixed-length vector. [10] learned video representations using 3D CNNs with long-term temporal convolutions. They increased the temporal extent of representations at the cost of decreased spatial resolution. In [11], they decomposed spatiotemporal convolutions into depthwise-separable temporal convolutions.

In this work, we proposed a two-pathway *Int&Int* framework for skeleton-based action recognition. *Intensity* pathway pays attention to the motion intensity of the action. *Integrity* pathway is responsible for ensuring the integrity of the action. An overview of *Int&Int* framework is depicted in Fig.1. In particular, motivated by [12], we first extract 2D human poses from each frame in a video using a two-stage pose estimator(detection + pose estimation). As for sampling frames, for *Integrity* pathway, we apply the uniform sampling strategy as [6] has done. For *Intensity* pathway, which our work mainly focuses on, we compute the motion intensity of each frame in the video based on 2D human poses. Then, we sample a clip composed of consecutive frames around the most intense frame, since motions with large amplitude and quick changes have more action-specific semantic information for identifying the action category intuitively. After sampling of two pathways, the poses are represented by stacks of heatmaps of skeleton joints along the temporal dimension to form a 3D heatmap volume [12]. For each pathway, we use a 3D-CNN to classify the 3D heatmap volumes. Finally, late fusion is used to recognize action [13].

The main contributions of our work lie in three folds:

- A two-pathway *Int&Int* framework is proposed which concentrates on the local and global information of the action, where there is great complementarity in the two pathways.
- A clip whose position depends on the most intense frame is sampled as for *Intensity* pathway, where there is a clear physical meaning in the sampling process.
- The motion intensity of each frame is defined in $3 + 2$ ways containing different semantic information.

The remainder of this paper is organized as follows. Section II provides a detailed description of our proposed *Int&Int* network. After that, Section III introduces different kinds of

This work is supported by the National Natural Science Foundation of China (No. 61876054)

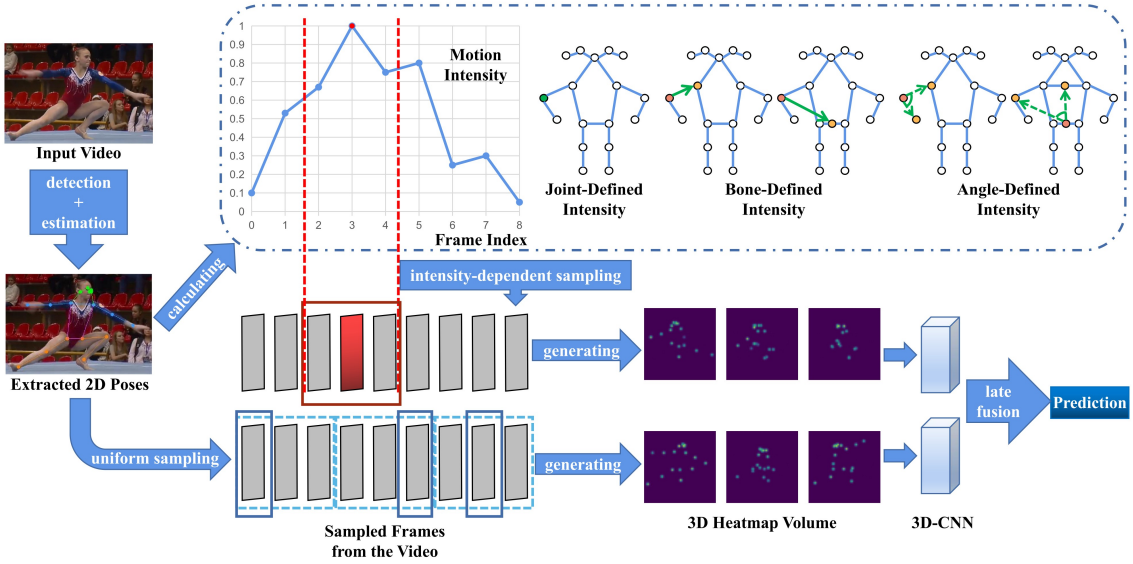


Fig. 1: An overview of *Int&Int* framework.

definitions of the motion intensity. Then, Section IV shows experiments on public action datasets, followed by an analysis and comparison of experimental results. Finally, the conclusion can be found in Section V.

II. *Int&Int* Network

We propose *Int&Int* network, a two-pathway approach for action recognition. In the following parts. We begin with 2D human poses extraction from each frame in the video. Subsequently, we introduce the designs of two pathways separately. Finally, we describe the late fusion of them.

A. Poses Extraction

Poses extraction is composed of the detector Faster-RCNN [14] and the 2D Top-Down pose estimator HRNet [15], which is able to maintain high resolution representations through the whole process, based on the research that 2D poses have better quality compared to 3D poses [12]. 2D poses annotations extracted with the Top-Down pipeline will play critical roles in both the sampling and the representation of heatmaps.

B. Two Pathways

1) Intensity-dependent sampling in Intensity pathway: A complete action includes beginning, middle and end processes. Thus, motion features are non-uniformly distributed along the time axis. Instead of sampling from a random temporal window [5], our intensity-dependent sampling can avoid missing the salient features of the motion. Using the coordinates (x, y) of skeleton joints from the extracted 2D poses annotations, we first compute the motion intensity of each frame in the video. We formulate it as $SV = \{SV^0, SV^1, \cdots, SV^{\text{L-1}}\}$, where L is the number of frames in the video, $\{SV^{\text{t}} \geq 0 \mid t = 0, 1, \cdots, L-1\}$ and $\{t \mid t = 0, 1, \cdots, L-1\}$ means the index of corresponding frame. In Section III, SV will be defined specifically as $\{SV_{Jnf}, SV_{Jap}, SV_{Bnf}, SV_{Bap}, SV_{Anf}, SV_{Aap}\}$. After getting SV of the video, we select the largest element in SV as SV_{max}, which corresponds to the most intense frame. Subsequently, we choose indices of the consecutive frames to sample as below. T is the length of the clip, t_{intense} is the index of SV_{max} in SV, and $t_{\text{start}} = t_{\text{intense}} - \frac{T}{2}$. The indices we choose are $\{t_{\text{start}}, ..., t_{\text{start}} + T - 1\}$.

2) Uniform sampling in Integrity pathway: We apply the uniform sampling strategy [6] to *Integrity* pathway in order to maintain the global information of the video. To be specific, in case there are n frames we need to sample, one input video is divided into n segments of equal length and a frame is randomly selected from each segment.

3) 3D heatmap volumes: After sampling, 2D poses extracted from the sampled frames are reformulated into a 3D heatmap volume [12] for both pathways separately. For a single frame, a 2D pose is represented as a heatmap of size $K \times H \times W$, where K is the number of joints, H and W are the height and width of the frame. We generate a joint heatmap $\boldsymbol{J}$ by composing K gaussian maps centered at every joint, which are based on joint coordinates and confidence:

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

where (x_k, y_k) and c_k are respectively the location and confidence score of the k-th joint, σ controls the variance of gaussian maps. The representation of $\boldsymbol{J}_{kij}$ depends on the distance from the point (i, j) to (x_k, y_k). In the multi-person case of one frame, the k-th gaussian maps of all persons are directly accumulated. For all sampled T frames, we stack all heatmaps along the temporal dimension to obtain a 3D heatmap volume of size $K \times T \times H \times W$. Subjects-centered cropping [12] is also adopted to reduce the redundancy of 3D heatmap volumes.

4) 3D-CNN of two pathways: Using the pseudo heatmap volumes as the input, as shown in Table I, we demonstrate the architecture of 3D-CNN for skeleton-based action recognition [12], where $T \times S^2, C$ denote the dimensions of kernels for temporal, spatial, channel sizes and GAP denotes global average pooling. We choose SlowOnly [5], obtained by inflating the ResNet [16] layers in the last two stages from 2D to 3D, to instantiate the backbone. Because we have already extracted skeleton from the video using HRNet, the block ResNet2 has been discarded. Note that both pathways have the same architecture, and they train the individual losses respectively.

TABLE I: Architecture of One Pathway

Stage	Pathway	Output Size $T \times S^2$
Data Layer	$32, 4^2$	32×56^2
Stem Layer	Conv $1 \times 7^2, 32$ Stride $1, 1^2$	32×56^2
ResNet3	$\begin{bmatrix} 1 \times 1^2, 32 \\ 1 \times 3^2, 32 \\ 1 \times 1^2, 128 \end{bmatrix} \times 4$	32×28^2
ResNet4	$\begin{bmatrix} 3 \times 1^2, 64 \\ 1 \times 3^2, 64 \\ 1 \times 1^2, 256 \end{bmatrix} \times 6$	32×14^2
ResNet5	$\begin{bmatrix} 3 \times 1^2, 128 \\ 1 \times 3^2, 128 \\ 1 \times 1^2, 512 \end{bmatrix} \times 3$	32×7^2
GAP	GlobalAveragePooling	#Classes

C. Late Fusion

As the final step of the whole framework, the predictions of two pathways are combined in a late fusion manner as a form of ensembling. In particular, the softmax scores of the two pathways are added to obtain the fused score and predict the corresponding action category.

III. Motion Intensity

In this section, we define the motion intensity of each frame in $3 + 2$ ways based on the motion prior knowledge. (3: joint-defined, bone-defined, angle-defined; 2: JBA-defined(arithmetic average), JBA-defined(weighted average))

A. Joint-Defined

In terms of joint-defined motion intensity, there are two ways to calculate its value of each frame: using corresponding position in the next frame and using average position across the frames. The final value of joint-defined motion intensity is the arithmetic average of values calculated and normalized in two ways respectively.

1) Using corresponding position in the next frame: After getting the coordinates of skeleton joints from the extracted 2D poses annotations, we then acquire the motion intensity of the frame whose index is t through calculating the distance between corresponding positions of adjacent frames. J_{ij}^t means the position of the i-th keypoint of the j-th person in the t-th frame. K is the number of keypoints of the skeleton. M is the number of people in the frame:

$$SV_{Jnf}^t = \frac{\sum_{j=0}^{M-1} \sum_{i=0}^{K-1} |J_{ij}^{t+1} - J_{ij}^t|}{M} \tag{2}$$

2) Using average position across the frames: In this way, we do not use the inter-frame information but utilize information within a frame. Particularly, we calculate the distance between the i-th keypoint in the t-th frame and the average position of the keypoint. $\bar{J}_{ij}$ means the average position of the i-th keypoint of the j-th person across all frames:

$$SV_{Jap}^t = \frac{\sum_{j=0}^{M-1} \sum_{i=0}^{K-1} |J_{ij}^t - \bar{J}_{ij}|}{M} \tag{3}$$

B. Bone-Defined

The pose estimator we adopt is pre-trained on COCO-keypoint [17]. The two ways to calculate the motion intensity of each frame is similar to that in joint-defined, while we define two types of bone features of skeleton: locally-defined bones and center-oriented bones. In particular, a bone-vector has a one-to-one correspondence with a keypoint J, and goes from J to another keypoint v. The difference between two types of bone features lies in the definition of v. For locally-defined bones, v is one adjacent neighbor of J. For center-oriented bones, v is the body center joint, which is the middle of the hip. The final value of bone-defined motion intensity is the arithmetic average of 2×2 values(2 ways to calculate$\times$2 types of bone features) which have been normalized.

1) Using corresponding bone-vector in the next frame: We acquire the motion intensity of the frame whose index is t through calculating the angle between bone-vector $B^t = \overrightarrow{J^t v^t}$ and $B^{t+1} = \overrightarrow{J^{t+1} v^{t+1}}$. B_{ij}^t means the i-th bone-vector of the j-th person in the t-th frame:

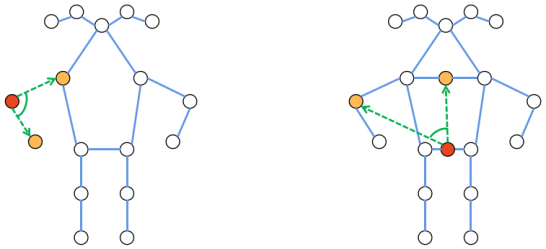
$$SV_{Bnf}^t = \frac{\sum_{j=0}^{M-1} \sum_{i=0}^{K-1} \left(1 - \frac{B_{ij}^t \cdot B_{ij}^{t+1}}{|B_{ij}^t||B_{ij}^{t+1}|}\right)}{M} \tag{4}$$

2) Using average bone-vector across the frames: In this way, we calculate the angle between bone-vector $B^t = \overrightarrow{J^t v^t}$ and $\bar{B} = \overrightarrow{\bar{J}\bar{v}}$, where $\bar{J}$ and $\bar{v}$ are average positions of endpoints of the bone-vector across all frames. $\bar{B}_{ij}$ means the i-th average bone-vector of the j-th person across all frames:

$$SV_{Bap}^t = \frac{\sum_{j=0}^{M-1} \sum_{i=0}^{K-1} \left(1 - \frac{B_{ij}^t \cdot \bar{B}_{ij}}{|B_{ij}^t||\bar{B}_{ij}|}\right)}{M} \tag{5}$$

C. Angle-Defined

For angle-defined motion intensity, we also define two types of angular features of skeleton: locally-defined angles and center-oriented angles [18], as shown in Fig.2. In particular, an angle has a one-to-one correspondence with a keypoint J. For locally-defined angles, J is the target keypoint to calculate the angular features and v_1 and v_2 are endpoints in the skeleton, where v_1 and v_2 are adjacent neighbors of J. For center-oriented angles, v_1 is the target keypoint to calculate the angular features and J and v_2 are endpoints, where v_1 is



(a) Locally-Defined (b) Center-Oriented

Fig. 2: Two types of angular features. Red point: target joint. Yellow point: anchor joint.

the middle of the hip and v_2 is the middle of shoulders. The final value of angle-defined motion intensity is the arithmetic average of 2×2 values (2 ways to calculate$\times2$ types of angular features) which have been normalized.

1) Using corresponding angle in the next frame: The motion intensity of the frame whose index is t is acquired through calculating the difference between corresponding angles of adjacent frames. $A_{ij}^t = 1 - \cos\theta_{ij}^t$ where θ_{ij}^t means the i-th angle of the j-th person in the t-th frame:

$$SV_{Anf}^t = \frac{\sum_{j=0}^{M-1}\sum_{i=0}^{K-1}|A_{ij}^{t+1} - A_{ij}^t|}{M} \tag{6}$$

2) Using average angle across the frames: We calculate the difference between the i-th angle in the t-th frame and the average angle. $\bar{A}_{ij} = 1 - \cos\bar{\theta}_{ij}$ where $\bar{\theta}_{ij}$ means the i-th average angle of the j-th person across all frames:

$$SV_{Aap}^t = \frac{\sum_{j=0}^{M-1}\sum_{i=0}^{K-1}|A_{ij}^t - \bar{A}_{ij}|}{M} \tag{7}$$

D. JBA-Defined

The values of motion intensity calculated in three definitions need to be normalized because three definitions have different dimensions. After that, there will be two kinds of methods to ensemble them:

- Arithmetic average.
- Weighted average. We utilize the normalized two-pathway accuracy we get from each definition as the weight of the corresponding definition.

IV. Experiments

In this section, we first give a brief introduction of two datasets. Then, the experiment settings are stated. Moreover, we show the recognition results and the comparisons with SOTA methods. Finally, we conduct the ablation study to prove the superiority of our method.

A. Datasets

We evaluate our model on human action datasets FineGYM-99 [19] and HMDB-51 [20]. FineGYM-99 is a new proposed fine-grained dataset built on top of gymnastic videos. In particular, it provides 29K videos of 99 fine-grained gymnastic action classes. HMDB-51 contains 51 distinct action categories, each containing at least 101 clips for a total of 6766 video clips extracted from a wide range of sources. 2D human poses of these datasets are extracted from video clips through Faster-RCNN with the ResNet50 backbone and HRNet pretrained on COCO-keypoint. We use the preprocessed annotations provided by [21]. All samples only include skeleton information of humans and are irrelevant to the background.

TABLE II: Comparisons with SOTA Methods on FineGYM-99

Method	Accuracy
TSN(RGB) [6] [19]	74.8%
TSN(RGB+Flow) [6] [19]	86.0%
TRN(RGB) [7] [19]	79.9%
TRN(RGB+Flow) [7] [19]	87.4%
ActionVLAD [9] [19]	69.5%
Int&Int(Ours)	**95.6%**

TABLE III: Comparisons with SOTA Methods on HMDB-51

Method	Accuracy
Two-Stream(fusion by SVM) [1]	59.4%
Two-Stream(fusion by averaging) [1]	58.0%
TDD [2]	63.2%
C3D [3] [8]	56.8%
TSN(RGB+Flow) [6]	68.5%
TSN(RGB+Flow+Warped Flow) [6]	69.4%
TLE(FC-Pooling) [8]	68.8%
ActionVLAD [9]	66.9%
LTC(RGB+Flow) [10]	64.8%
Int&Int(Ours)	**69.6%**

B. Experiment Settings

Our experiments are conducted using Pytorch [22] framework. We train our model on 4 NVIDIA GTX 3090 GPUs in parallel with the Pytorch distributed package. For FineGYM-99, we choose SGD as our optimizer, with a learning rate of 0.4, momentum of 0.9, and weight decay of 0.0003. A cosine annealing strategy of learning rate adjustment is used. The number of epochs is 80. And the model is trained from scratch. For HMDB-51, we choose SGD as our optimizer, with a learning rate of 0.01, momentum of 0.9, and weight decay of 0.0001. A step strategy of learning rate adjustment is used. The number of total epochs is 12. And the model has been pretrained on Kinetics400 for HMDB-51. For both datasets, we train our model with batch size 16 and the length of sampled clip is 48. The architecture of our model is shown in Table I, while the setting of the ResNet blocks is (3,4,6) for HMDB-51.

C. Comparisons with SOTA Methods

In Table II and Table III, we compare our results with prior works for action recognition. For FineGYM-99, we report the

TABLE IV: Single-Pathway *vs*. Two-Pathway

Sampling	HMDB-51											
	3 Splits			*Split 1*			*Split 2*			*Split 3*		
	Int&Int	Intensity Pathway	Integrity Pathway	Int&Int	Intensity Pathway	Integrity Pathway	Int&Int	Intensity Pathway	Integrity Pathway	Int&Int	Intensity Pathway	Integrity Pathway
Random	**69.3%**	63.7%	68.6%	**70.2%**	64.4%	69.2%	**69.0%**	63.0%	68.6%	**68.6%**	63.7%	68.0%
Joint-Defined	**69.6%**	63.9%	68.6%	**70.5%**	64.5%	69.2%	**69.3%**	63.5%	68.6%	**69.0%**	63.5%	68.0%
Bone-Defined	**69.0%**	64.2%	68.6%	69.1%	63.9%	**69.2%**	**68.7%**	63.7%	68.6%	**69.2%**	65.0%	68.0%
Angle-Defined	**69.5%**	63.4%	68.6%	**69.3%**	63.9%	69.2%	**70.3%**	62.6%	68.6%	**68.8%**	63.6%	68.0%
JBA(Arithmetic)	**69.3%**	63.6%	68.6%	68.6%	63.4%	**69.2%**	**70.1%**	63.2%	68.6%	**69.2%**	64.3%	68.0%
JBA(Weighted)	**69.2%**	63.9%	68.6%	**69.6%**	64.0%	69.2%	**69.7%**	64.5%	68.6%	**68.2%**	63.1%	68.0%

Top-1 accuracy. For HMDB-51, we report average accuracy over three splits. We can see that although these prior works all use at least videos as the input which contain much more information than our skeleton-based method, our results achieve superior performance and outperform the results of SOTA methods.

D. Ablation Study

1) Single-pathway vs. Two-pathway: To validate the great complementarity of two pathways in our framework, we conduct experiments of *Int&Int* model, *Intensity* pathway and *Integrity* pathway respectively on HMDB-51. In Table IV, we report average accuracy over three splits and Top-1 accuracy of each split respectively. We see that the two-pathway architecture *Int&Int* almost consistently outperforms the single-pathway architectures which only concentrate on intensity or integrity.

2) Randomly sampling vs. Intensity-dependent sampling: The input sampled from a random temporal window may not capture the salient features of the motion. To validate this, we conduct experiments of our two-pathway model on FineGYM-99 and HMDB-51. When sampling from a random temporal window, the central frame in the sampled clip is not the most intense frame of the video but a frame randomly selected. In testing, we adopt a fixed random seed when sampling to make sure the test results are reproducible. In Table V, we report the Top-1 accuracy for FineGYM-99. For HMDB-51, we report average accuracy over three splits and Top-1 accuracy of each split respectively. We see that our intensity-dependent sampling consistently outperforms sampling from a random temporal window.

TABLE V: Comparisons of Sampling Methods

Sampling	FineGYM-99	HMDB-51			
		3 Splits	*Split 1*	*Split 2*	*Split 3*
Random	95.3%	69.3%	70.2%	69.0%	68.6%
Intensity	**95.6%**	**69.6%**	**70.5%**	**70.3%**	**69.2%**

3) Intensity defined through motion prior knowledge: In Table VI, we discuss various definitions which place emphasis on different semantic information. The positions of joints carry the most direct and obvious information of motion. The lengths and directions of bones contain higher-order information, which have been proven to be effective in previous work [13]. Angles can maintain invariance against different human body sizes and discriminate actions sharing similar motion trajectories such as taking off glasses and taking off headphones. We conduct experiments using *Int&Int* model on FineGYM-99 and HMDB-51. We report the Top-1 accuracy for FineGYM-99. For HMDB-51, we report average accuracy over three splits and Top-1 accuracy of each split respectively. We can see that for the professional and flexible gymnastics, it is necessary to combine the information from three aspects because of the richness of fine-grained action information. For the actions in HMDB-51, which are generally simpler, the information from only one aspect is enough to capture the feature.

TABLE VI: Performance of Different Definitions

Sampling	FineGYM-99	HMDB-51			
		3 Splits	*Split 1*	*Split 2*	*Split 3*
Joint-Defined	95.0%	**69.6%**	**70.5%**	69.3%	69.0%
Bone-Defined	95.2%	69.0%	69.1%	68.7%	**69.2%**
Angle-Defined	95.5%	69.5%	69.3%	**70.3%**	68.8%
JBA(Arithmetic)	95.5%	69.3%	68.6%	70.1%	**69.2%**
JBA(Weighted)	**95.6%**	69.2%	69.6%	69.7%	68.2%

V. Conclusion

In this work, we propose a two-pathway *Int&Int* network for skeleton-based action recognition. The final model achieves superior performance on both of the datasets. In the future, on one hand, we intend to reduce the computational complexity of our network and further improve the performance. On the other hand, we will conduct research on the applications such as security, human-computer interaction, physical training, medical rehabilitation, and entertainment.

REFERENCES

[1] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in neural information processing systems*, vol. 27, 2014.

[2] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4305–4314.

[3] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[4] D. Tran, J. Ray, Z. Shou, S.-F. Chang, and M. Paluri, "Convnet architecture search for spatiotemporal feature learning," *arXiv preprint arXiv:1708.05038*, 2017.

[5] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6202–6211.

[6] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *European conference on computer vision*. Springer, 2016, pp. 20–36.

[7] B. Zhou, A. Andonian, A. Oliva, and A. Torralba, "Temporal relational reasoning in videos," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 803–818.

[8] A. Diba, V. Sharma, and L. Van Gool, "Deep temporal linear encoding networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 2329–2338.

[9] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, and B. Russell, "Actionvlad: Learning spatio-temporal aggregation for action classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 971–980.

[10] G. Varol, I. Laptev, and C. Schmid, "Long-term temporal convolutions for action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1510–1517, 2017.

[11] N. Hussein, E. Gavves, and A. W. Smeulders, "Timeception for complex action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 254–263.

[12] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2969–2978.

[13] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 026–12 035.

[14] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[15] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[17] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[18] Z. Qin, Y. Liu, P. Ji, D. Kim, L. Wang, B. McKay, S. Anwar, and T. Gedeon, "Leveraging third-order features in skeleton-based action recognition," *arXiv preprint arXiv:2105.01563*, 2021.

[19] D. Shao, Y. Zhao, B. Dai, and D. Lin, "Finegym: A hierarchical video dataset for fine-grained action understanding," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2616–2625.

[20] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *2011 International conference on computer vision*. IEEE, 2011, pp. 2556–2563.

[21] H. Duan, J. Wang, K. Chen, and D. Lin, "Pyskl: Towards good practices for skeleton action recognition," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 7351–7354.

[22] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.