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

CONTENTS

1

Research Status

Our Framework

Contributions

Research Status

- Mainly focus on the local or global information **alone**.
- Cannot focus on both the intensity and the integrity.

local dynamics	global information
Two-Stream[1]	TSN[6]
TDD[2]	TRN[7]
C3D[3]	TLE[8]
Res3D[4]	ActionVLAD[9]
Slowfast[5]	LTC[10]
	Timeception[11]

	Methods
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_x0002_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, "Action_x0002_vlad: 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.

1

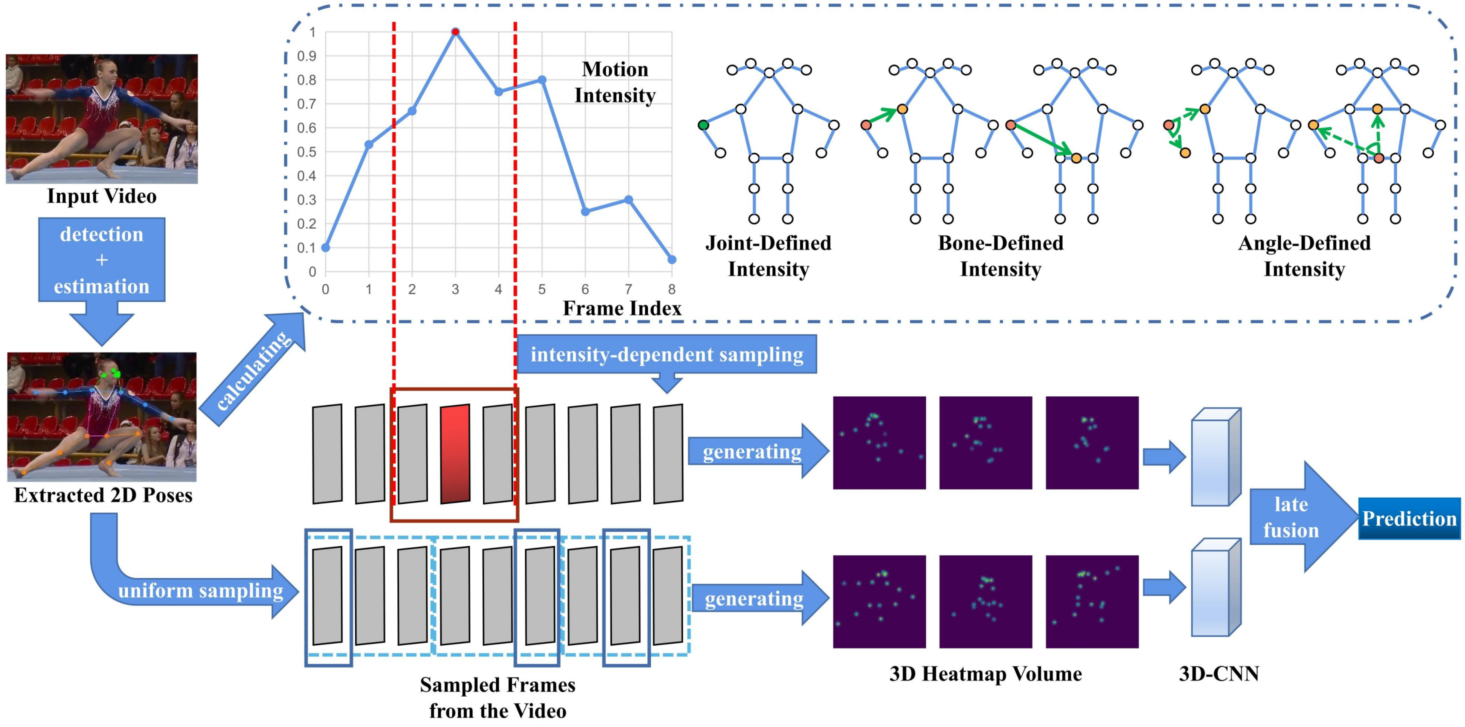
Research Status

Our Framework

Contributions

Our Framework

We propose a two-pathway *Int&Int* network(*Intensity&Integrity*) for skeleton-based action recognition to satisfy both aspects.



Contributions

1

Research Status

Our Framework

Contributions

1) 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.

2) 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.

3) The motion intensity of each frame is defined in 3 + 2 ways containing different semantic information.

2

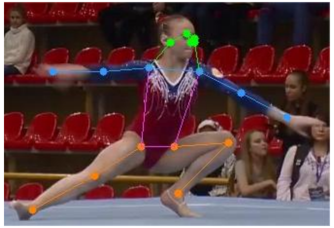
Poses Extraction

Two Pathways

Late Fusion

Poses Extraction

Input Video



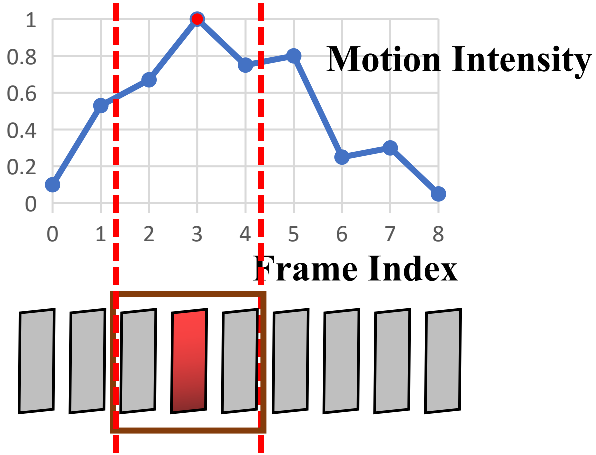
Extracted 2D Poses

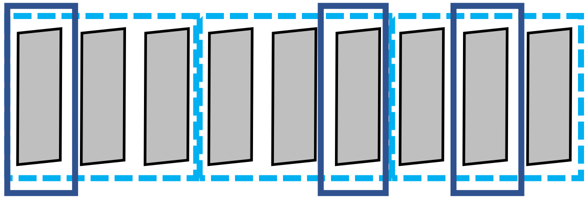
Detection + Estimation = Poses Extraction

Detection → **Faster-RCNN**

Estimation → **HRNet**

Two Pathways





Sampled Frames
from the Video

Intensity-dependent sampling in Intensity pathway

- Motion features are **non-uniformly** distributed along the time axis.
- Motions with large amplitude and quick changes have more action-specific semantic information for identifying the action intuitively.
- Intensity-dependent sampling can avoid missing the salient features:

1) extracted 2D poses --> motion intensity of each frame in the video
2) select the largest element corresponding to the most intense frame
3) sample a clip composed of consecutive frames around the most intense frame

Uniform sampling in Integrity pathway

- Maintain the global information of video:

1) one input video is divided into n segments of equal length (n frames to sample)
2) one frame is randomly selected from each segment

Two Pathways

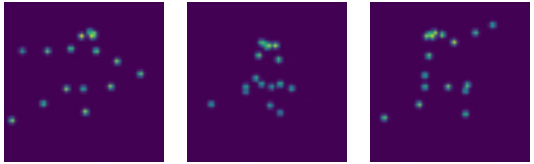
2

- Poses Extraction
- Two Pathways
- Late Fusion

3D heatmap volumes

- We generate a joint heatmap $\boldsymbol{J}$ by composing K gaussian maps centered at every joint, which are based on joint coordinates and confidence:

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



3D Heatmap Volume

- 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 $K \times T \times H \times W$, where K is the number of joints, H and W are the height and width of the frame.
- Subjects-centered cropping is also adopted to reduce the redundancy of 3D heatmap volumes.

3D-CNN of two pathways

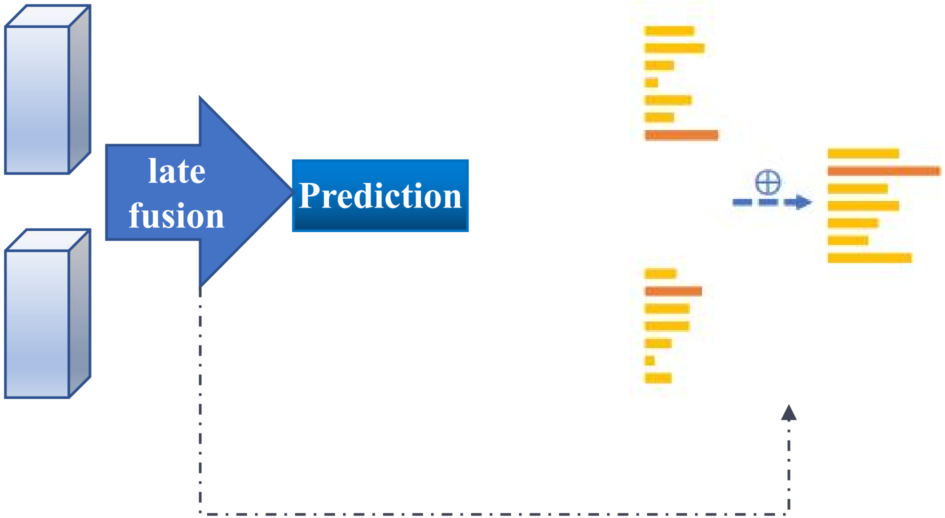
- We use the pseudo heatmap volumes as the input.
- $T \times S^2$, C denote the dimensions of kernels for temporal, spatial, channel sizes.
- We choose SlowOnly, obtained by inflating the ResNet layers in the last two stages from 2D to 3D, to instantiate the backbone.
- Both pathways have the same architecture, and they train the individual losses respectively.

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

2

Late Fusion

The softmax scores of the two pathways are added to obtain the fused score and predict the corresponding action category.

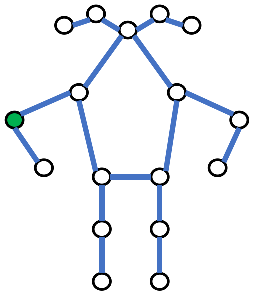


Joint-Defined

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

- calculate 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:



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

2) Using average position across the frames

- 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:

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

3

Bone-Defined

The final value of bone-defined motion intensity is the arithmetic average of 2 × 2 values (2 ways to calculate×2 types of bone features) which have been normalized.

1) Using corresponding bone-vector in the next frame

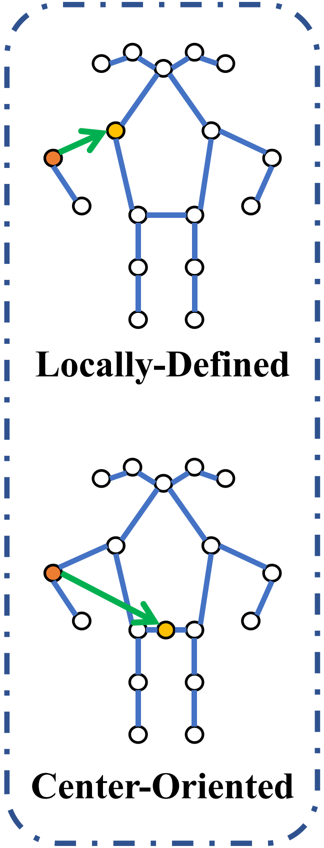
- calculate the angle between bone-vector $B^t = \overrightarrow{J^t v^t}$ and $B^{t+1} = \overrightarrow{J^{t+1} v^{t+1}}$
- B^t_{ij} means the i-th bone-vector of the j-th person in the t-th frame:

$$SV^t_{Bnf} = \frac{\sum_{j=0}^{M-1} \sum_{i=0}^{K-1} \left(1 - \frac{B^t_{ij} \cdot B^{t+1}_{ij}}{|B^t_{ij}||B^{t+1}_{ij}|}\right)}{M}$$

2) Using average bone-vector across the frames

- calculate the angle between bone-vector $B^t = \overrightarrow{J^t v^t}$ and $\bar{B} = \overrightarrow{\bar{J}\bar{v}}$
- $\bar{B}_{ij}$ means the i-th average bone-vector of the j-th person across all frames:

$$SV^t_{Bap} = \frac{\sum_{j=0}^{M-1} \sum_{i=0}^{K-1} \left(1 - \frac{B^t_{ij} \cdot \bar{B}_{ij}}{|B^t_{ij}||\bar{B}_{ij}|}\right)}{M}$$



3

Joint-Defined

Bone-Defined

Angle-Defined

JBA-Defined

Angle-Defined

The final value of angle-defined motion intensity is the arithmetic average of 2×2 values (2 ways to calculate×2 types of angular features) which have been normalized.

1) Using corresponding angle in the next frame

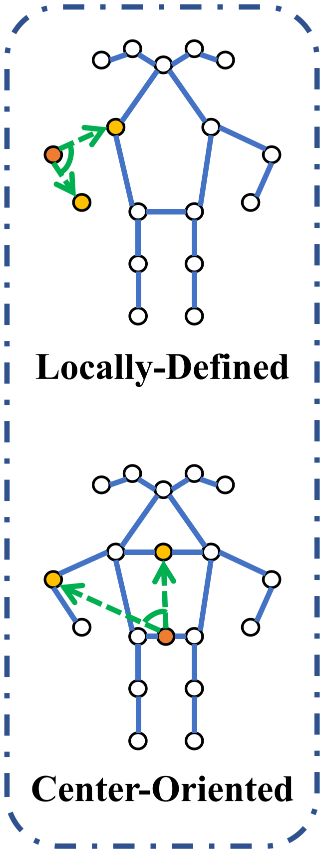
- calculate the difference between corresponding angles of adjacent frames
- $A_{ij}^t = 1 - \cos\theta_{ij}^t$ where θ_{ij}^t means the i-th angle of the j-th person in the t-th frame:

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

2) Using average angle across the frames

- calculate the difference between the i-th angle in the t-th frame and 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:

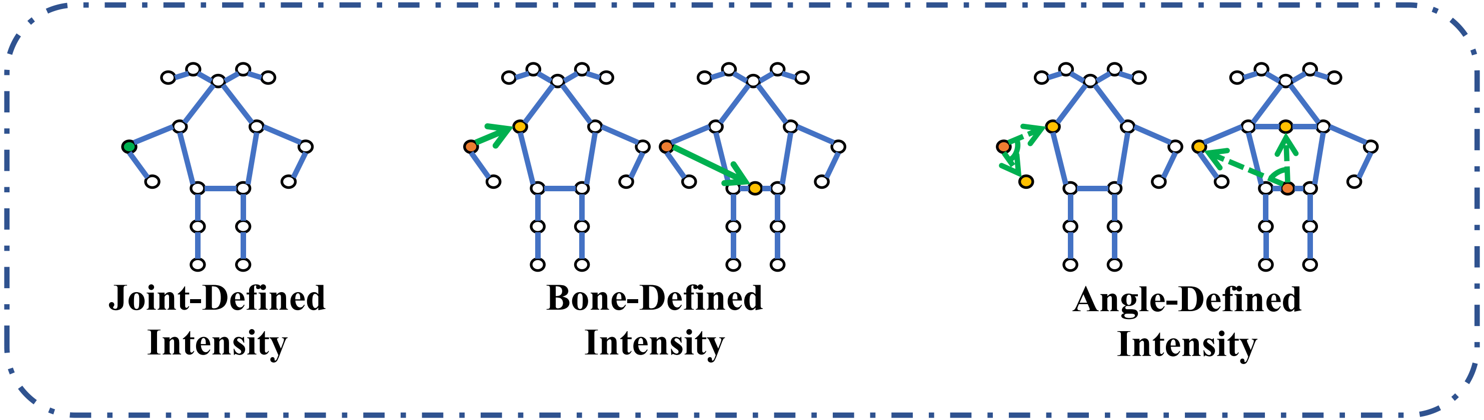
$$SV_{Aap}^t = \frac{\sum_{j=0}^{M-1}\sum_{i=0}^{K-1}|A_{ij}^t - \bar{A}_{ij}|}{M}$$



JBA-Defined

- The values of motion intensity calculated in three definitions need to be normalized because three definitions have different dimensions.
- Two kinds of methods to ensemble them:

1) Arithmetic average
2) Weighted average: utilize the normalized two-pathway accuracy we get from each definition as the weight of the corresponding definition



4

Datasets

1) FineGYM-99:

- 29K videos of 99 fine-grained gymnastic action classes

2) HMDB-51:

- 51 distinct action categories, each containing at least 101 clips for a total of 6766 video clips extracted from a wide range of sources

Settings

1) Pytorch framework, 4 NVIDIA GTX 3090 GPUs

2) For FineGYM-99:

- learning rate: 0.4, weight decay: 0.0003, learning rate adjustment: cosine annealing strategy, number of epochs: 80, model: trained from scratch, ResNet blocks setting is (4,6,3)

3) For HMDB-51:

- learning rate: 0.01, weight decay: 0.0001, learning rate adjustmentis: step strategy, number of epochs: 12, model: pretrained on Kinetics400, ResNet blocks setting is (3,4,6)

4) For both datasets:

- batch size: 16, optimizer: SGD, momentum: 0.9, the length of sampled clip: 48

The code has been shown at https://github.com/SarahQi666/Int-and-Int.

4

Datasets&Settings

Comparisons with SOTA Methods

Ablation Study

Comparisons with SOTA Methods

- 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.

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%**

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%**

- For FineGYM-99, we report the Top-1 accuracy. For HMDB-51, we report average accuracy over three splits.

4

Datasets&Settings

Comparisons with SOTA Methods

Ablation Study

Ablation Study

1) Single-pathway *vs.* Two-pathway

- The two-pathway architecture almost consistently outperforms the single-pathway architectures.

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%

2) Randomly sampling *vs.* Intensity-dependent sampling

- Randomly sampling: the central frame in the sampled clip is not the most intense frame of the video but a frame randomly selected.
- Intensity-dependent sampling consistently outperforms sampling from a random temporal window.

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%**

4

Datasets&Settings

Comparisons with SOTA Methods

Ablation Study

Ablation Study

3) Intensity defined through motion prior knowledge

- Various definitions place emphasis on different semantic information:

a) The positions of joints carry the most direct and obvious information of motion.

b) The lengths and directions of bones contain higher-order information, which have been proven to be effective in 2s-AGCN.

c) 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.

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%

- 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.

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:**

a) reduce the computational complexity of our network and further improve the performance

b) conduct research on the applications such as security, human-computer interaction, physical training, medical rehabilitation, and entertainment

Thanks for your attention!

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