Integrating Human Parsing and Pose Network for Human Action Recognition

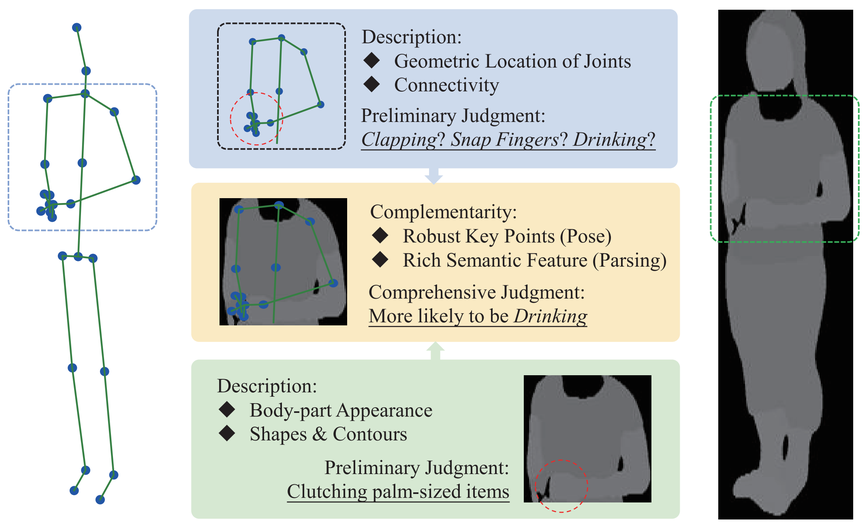
Runwei Ding[1*], Yuhang Wen[2*], Jinfu Liu[2], Nan Dai[3], Fanyang Meng[4], Mengyuan Liu[1†]

[1] Shenzhen Graduate School, Peking University
[2] Sun Yat-sen University
[3] Changchun University of Science and Technology
[4] Peng Cheng Laboratory

* Contribute Equally † Correspondence: nkliuyifang@gmail.com

https://github.com/liujf69/IPP-Net-Parsing

Motivation



Skeletons lack the ability to depict the appearance of human body parts
RGB features are prone to being influenced by various sources of noise
Can we explore a new modality that incorporates body-part appearance depiction while remaining noiseless and robust?

Our affirmative response is inspired by the **Human Parsing** task.
By focusing on these semantic parts, human parsing explicitly and effectively eliminates action-irrelevant details, while retaining crucial extrinsic features of the human body.

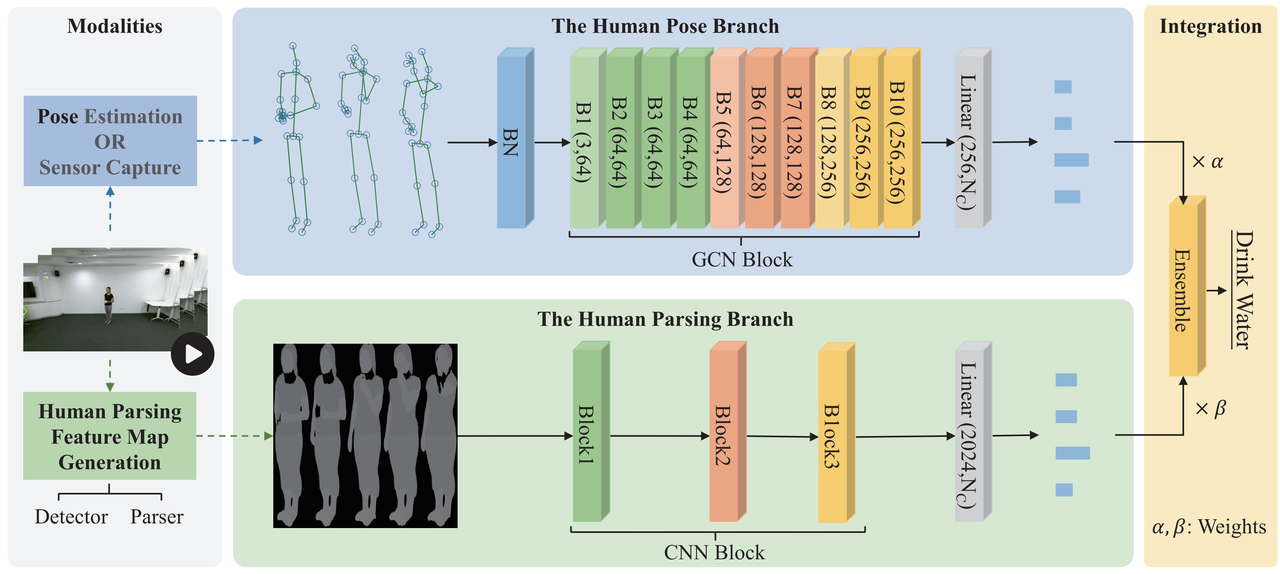
Contributions

1. We advocate to leverage human parsing feature map as a new modality for human action recognition task, which is appearance-oriented depictive and also action-relevant.

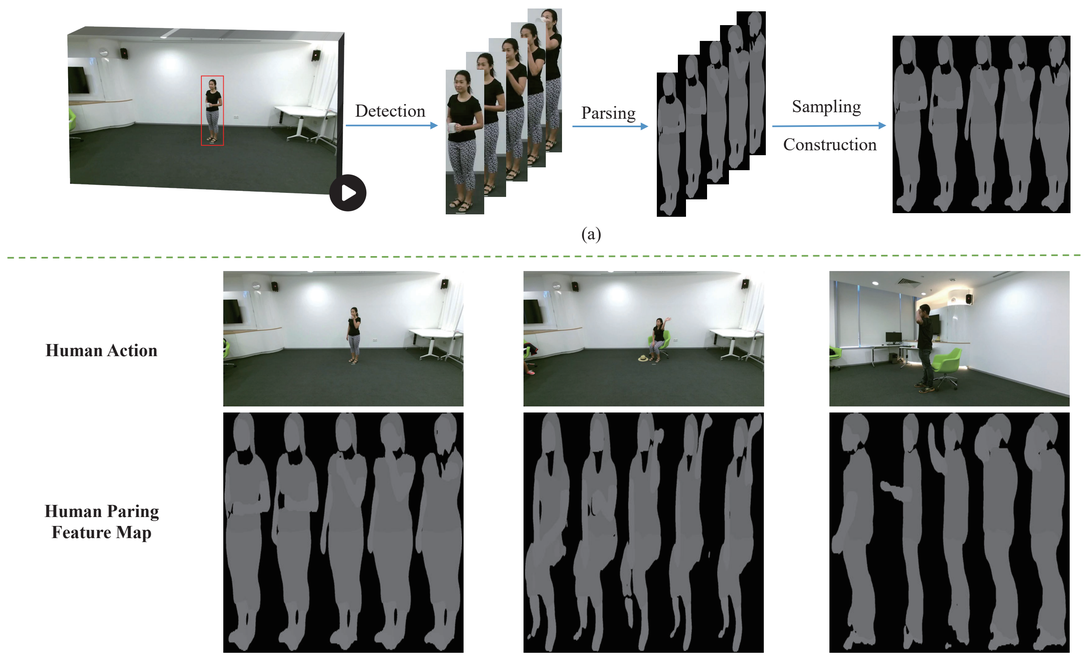
2. We propose a framework called Integrating Human Parsing and Pose Network (IPP-Net), which is the first to effectively integrates human parsing feature maps and pose data for robust human action recognition.

3. Extensive experiments on benchmark NTU RGB+D and NTU RGB+D 120 datasets verify the effectiveness of our IPP-Net, which outperforms most existing action recognition methods.

IPP-Net



In our proposed Integrating Human Parsing and Pose Network, we incorporate two primary branches, specifically the human pose branch and the human parsing branch.
The human pose branch utilizes GCN to model the skeleton data, while **the human parsing branch** employs CNN to extract deep features of the human parsing feature maps. Subsequently, the outcomes from these two branches are integrated via a late ensemble to get predictions.



Experimental Results

Table 1. Accuracy comparison with state-of-the-art methods on NTU-RGB+D and NTU-RGB+D 120 dataset.

Type	Method	Source	NTU 60 (%)		NTU 120 (%)	
			X-Sub	X-View	X-Sub	X-Set
Pose	Shift-GCN [4]	CVPR'20	90.7	96.5	85.9	87.6
	DynamicGCN [28]	MM'20	91.5	96.0	87.3	88.6
	DSTA-Net [19]	ACCV'20	91.5	96.4	86.6	89.0
	MS-G3D [15]	CVPR'20	91.5	96.2	86.9	88.4
	MST-GCN [3]	AAAI'21	91.5	96.6	87.5	88.8
	CTR-GCN [2]	ICCV'21	92.4	96.8	88.9	90.6
	GS-GCN [33]	CICAI'22	90.2	95.2	84.9	87.1
	PSUMNet [22]	ECCV'22	92.9	96.7	89.4	90.6
	InfoGCN [5]	CVPR'22	93.0	97.1	89.8	91.2
	STSA-Net [16]	Neurocomputing'23	92.7	96.7	88.5	90.7
Multi-Modality	VPN [6]	ECCV'20	93.5	96.2	86.3	87.8
	LST [27]	arXiv'22	92.9	97.0	89.9	91.1
	Ours (J+B+P)		93.4	96.8	89.4	91.2
	Ours		**93.8**	**97.1**	**90.0**	**91.7**

Table 2. Accuracy of different modalities on NTU-RGB+D and NTU-RGB+D 120.

Modality					NTU 60 (%)		NTU 120 (%)	
J	B	JM	BM	P	X-Sub	X-View	X-Sub	X-Set
✓					90.2	95.0	85.0	86.7
	✓				90.5	94.7	86.2	87.5
		✓			88.1	93.2	81.2	83.0
			✓		87.3	92.0	81.7	82.9
				✓	73.5	74.2	53.6	65.8
✓				✓	91.7	95.9	86.1	88.8
	✓			✓	91.8	95.9	87.5	89.8
✓	✓				92.2	96.2	88.7	90.2
✓	✓	✓	✓		92.4	96.5	89.0	90.5
✓	✓			✓	93.4	96.8	89.4	91.2
✓	✓	✓	✓	✓	**93.8**	**97.1**	**90.0**	**91.7**

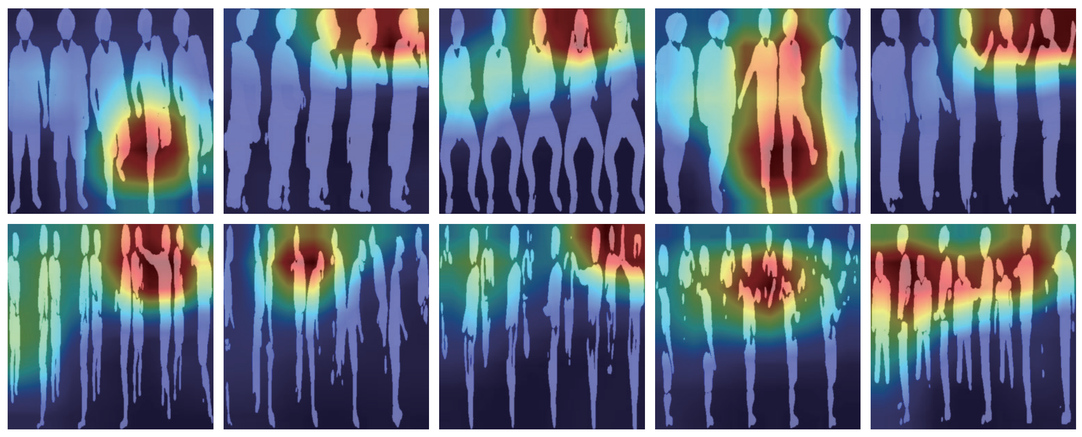
Table 3. Comparison of different numbers of frames in feature map construction.

#Frame	Parsing (%)	Ensemble (%)
3	46.7	89.8
4	50.4	89.9
5	53.6	**90.0**
6	55.6	89.9

Table 4. Accuracy of different CNN backbones in human parsing branch.

Backbone	Parsing (%)	Ensemble (%)
VGG11	49.0	89.8
VGG13	48.7	89.8
ResNet18	50.5	90.0
InceptionV3	53.6	**90.0**

Visualizations



This figure visually illustrates how human parsing feature maps in our IPP-Net help recognize actions by providing semantic information about body parts. The class activation maps indicates that the human parsing branch can focus on the most informative body parts.

Acknowledgement

This work was supported by

· the Basic and Applied Basic Research Foundation of Guangdong (No. 2020A1515110370)

· the National Natural Science Foundation of China (No. 62203476)

References

[1] Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.Y., Kot, A.C.: Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence 42(10), 2684–2701 (2020)
[2] Das, S., Sharma, S., Dai, R., Br emond, F., Thonnat, M.: Vpn: Learning video-pose embedding for activities of daily living. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020)
[3] Liang, X., Gong, K., Shen, X., Lin, L.: Look into person: Joint body parsing & pose estimation network and a new benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence 41(4), 871–885 (2019)
[4] Ultralytics: ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation (2022)

CICAI 2023 CAAI International Conference on Artificial Intelligence

July 22nd-23rd, 2023, Fuzhou, China