JT-MGCN: Joint-temporal Motion Graph Convolutional Network for Skeleton-Based Action Recognition

KYUNG HEE UNIVERSITY

PerCV Perception & Computer Vision Lab.

Suekyeong Nam, Seungkyu Lee

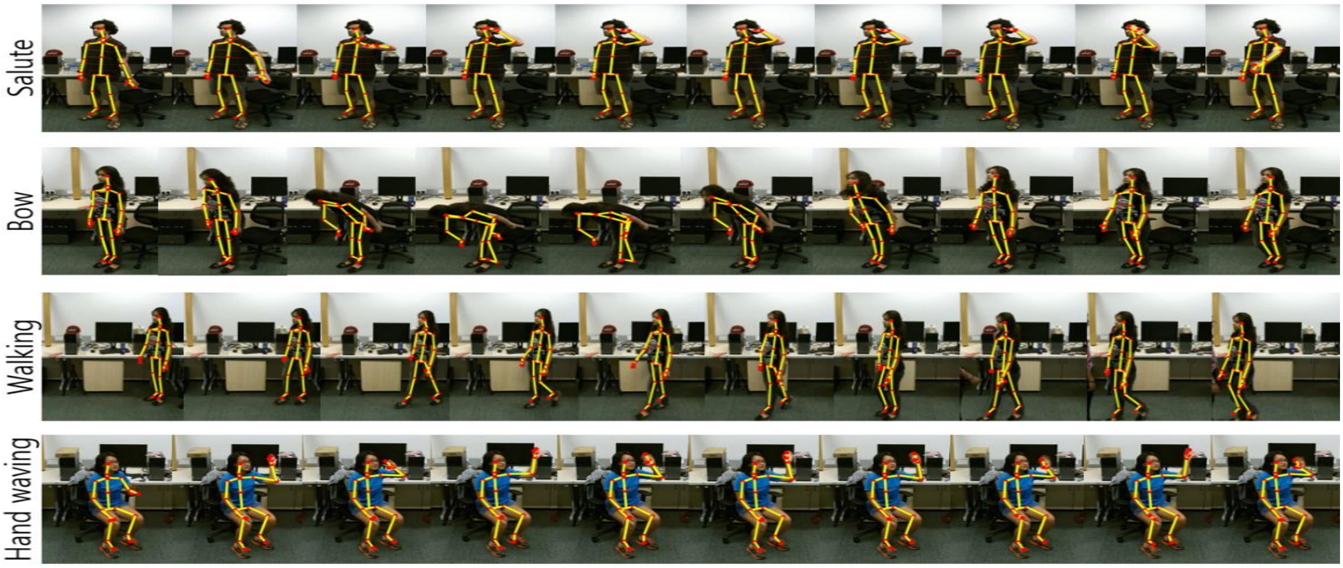
Kyung Hee University, Republic of Korea

ICPR 2020 25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION Milan, Italy 10 | 15 January 2021

Problem

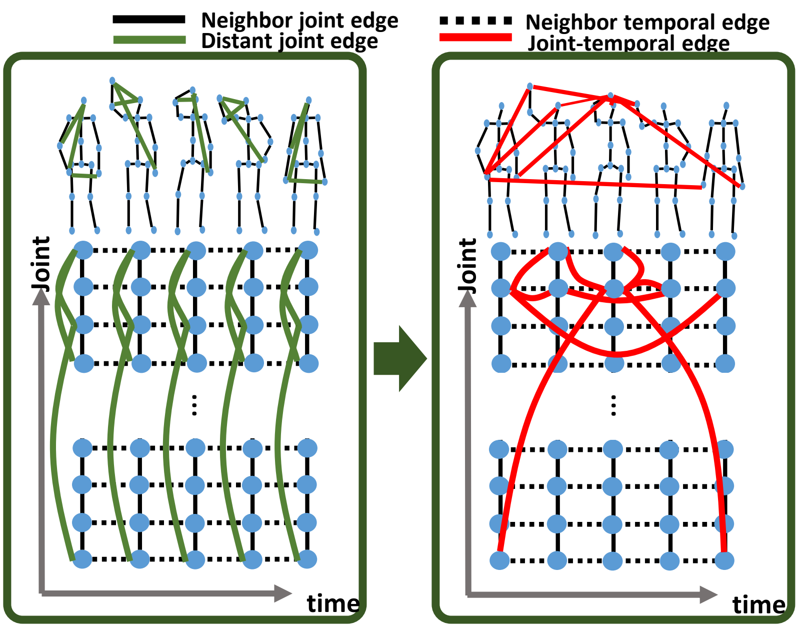
- Temporal variations and correlations in human actions.

1. Some actions consist of unique symmetric or repeated motion patterns.
2. Some actions can be characterized by only certain temporal motion segments or combination of such motion segments rather than entire temporal motions



Based on our observations, action can be described better if temporal correlations of optimal joint pairs are learned.

Proposed Algorithm



2S-AGCN[1], Shi et al. *CVPR* 2019.

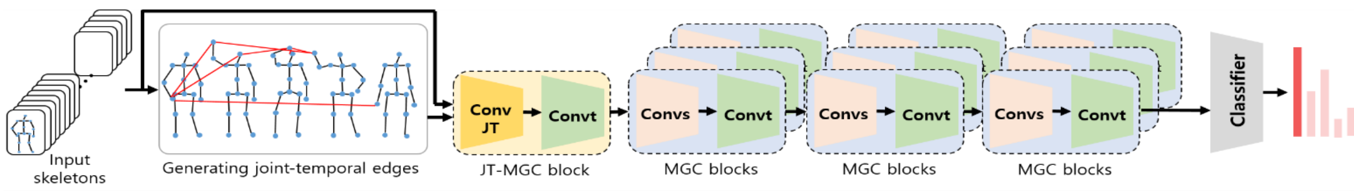
Our proposed algorithm

Previous work:
2S-AGCN trains distant joint edges, but same edges are applied to all temporal frames.

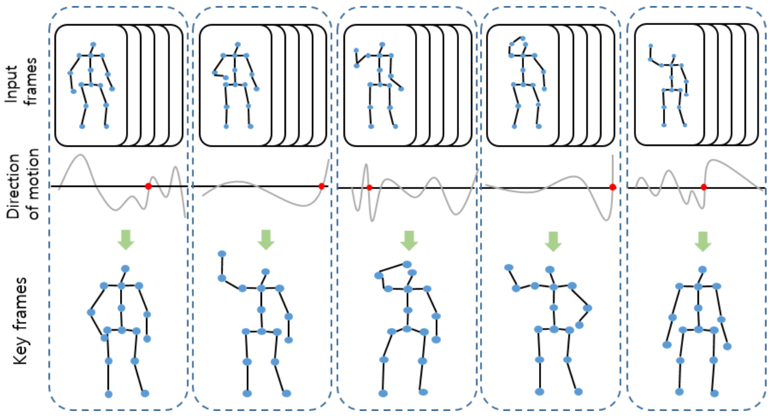
Proposed algorithm:
We makes **distant joint-temporal edges** as shown in the left figure.

Ex) 'drinking' action.
Right hand of first frame shows strong correlation with head of second frame

Overall Frameworks



1) Frame sampling

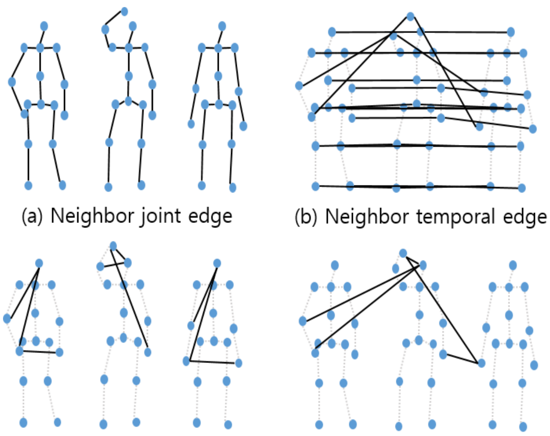


For F frame images and N body joints, we need (N × F) × (N × F) adjacent matrix!

- Perform frame sampling to extract fixed number of key frames that play critical role in characterizing action types.

Reduce computational cost!

2) Generating 4 types of edges

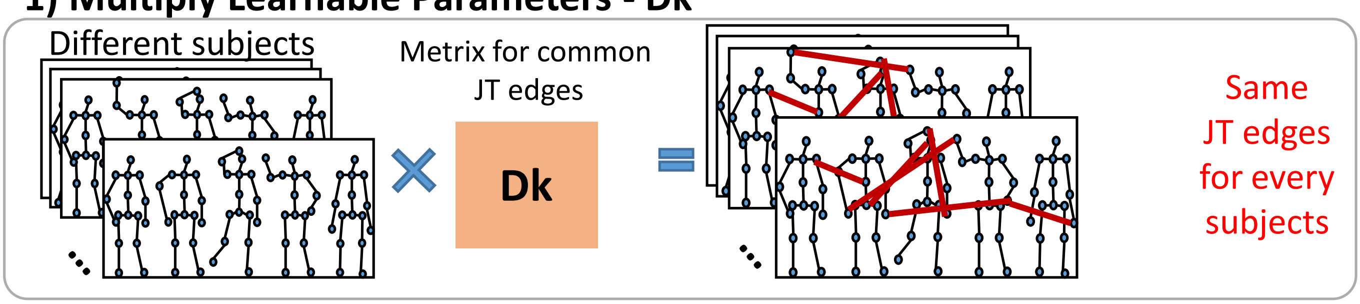


(a) Neighbor joint edge (b) Neighbor temporal edge

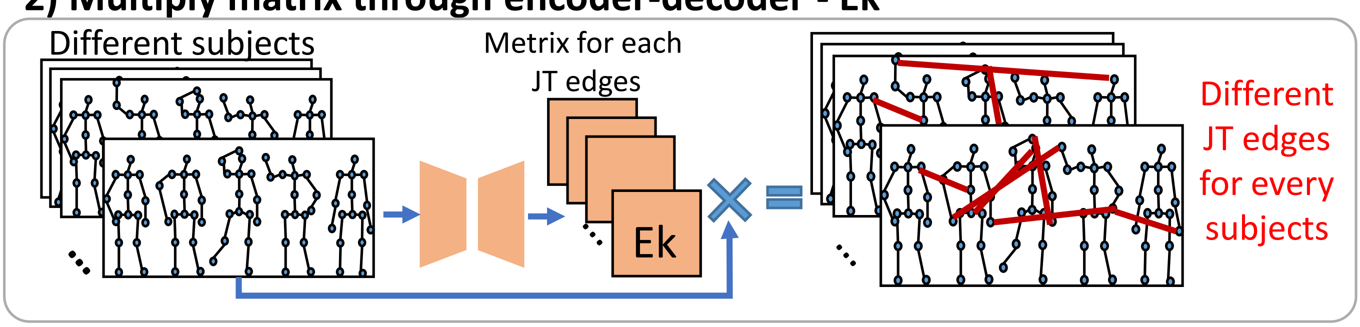
(c) Distant joint edge (d) Joint-temporal edge

- Neighbor joint edge: Spatial relationship connecting each pair of neighbor joints.
- Neighbor temporal edge: Temporal correlation of respective joint over time.
- Distant joint edge: Additional connections connecting two joints located away to each other within single time frame.
- **Joint-temporal edge**: Connection over time and space without any structural limitation.

1) Multiply Learnable Parameters - Dk



2) Multiply matrix through encoder-decoder - Ek



Experiment

Using NTU RGB+D Action Recognition Dataset

- 56,880 action samples.
- containing RGB videos, depth map sequences, 3D skeletal data, and infrared videos.
- 60 classes with daily action, medical conditions, mutual conditions.

Using Kinectics-skeleton Dataset

- about 300,000 RGB videos
- 400 classes with daily action.
- obtained the coordinates of skeleton using open source, OpenPose [2], in rgb video.

Results

method	CS(%)
(a)+(b)	81.5
(a)+(b)+(c)	88.5
(a)+(b)+(c)+(d)	**90.4**

Method	Year	CS(%)	CV(%)
PA-LSTM [18]	2016	62.9	70.3
ST-LSTM+TS [13]	2016	69.2	77.7
STA-LSTM [23]	2017	73.4	81.2
VA-LSTM [27]	2017	79.4	87.6
MTLN [5]	2017	79.6	84.8
ST-NBMIM [25]	2018	80.0	84.2
MTCNN [6]	2018	81.1	87.4
ST-GCN [26]	2018	81.5	88.3
DPRL+GCNN[24]	2018	83.5	89.8
SR-TSL [22]	2018	84.8	92.4
AS-GCN [10]	2019	86.8	94.2
2S-AGCN [20]	2019	88.5	95.1
DGNN [19]	2019	89.9	96.1
JT-MGCN(proposed)	**2019**	**90.40**	**95.78**

- It has improved accuracy about 2% compared to baseline, 2S-AGCN with NTU dataset.

For demonstrating how JT-edges are effected to classify action, we shuffled each samples along the temporal axis and input to each pre-trained JT-MGCN models.

	Method	CS(%)
Without shuffling	(a)+(b)+(c)	88.5
	(a)+(b)+(c)+(d)	90.4
With shuffling	(a)+(b)+(c)	13.4
	(a)+(b)+(c)+(d)	**8.8**

Better

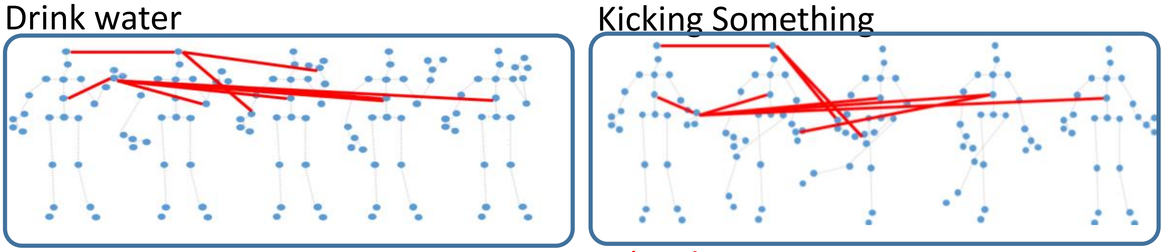
Worse

- If the orders of human action are mixed, it means different action.
- JT-edges make easy to extract temporal relationships than the previous one.

JT edges of Top 10 highest weight value of (Dk+Ek) on the body skeleton

Drink water

Kicking Something



Same edges!

Between different actions!

The overall characteristics of the samples, generally become more pronounced.

References

[1] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, “Actionalstructural graph convolutional networks for skeleton-based action recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3595–3603.
[2] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime multi-person 2d pose estimation using part affinity fields,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7291–7299.