ST-GCN-PAM
Pairwise Adjacency Matrix on Spatial-Temporal Graph Convolutional Network for Two People Action Recognition

The 27th IEEE International Conference on Image Processing (ICIP 2020)

Chao-Lung Yang, Aji Setyoko, Hendrik Tampubolon, Kai-Lung Hua

National Taiwan University of Science and Technology

Action recognition

Provides a very usefull information which difficult to extract:

- personality and psychological state.

Wide range of applications:

- intelligent video surveillance,
- environmental home monitoring,
- video storage and retrieval,
- intelligent human–machine interfaces,
- and identity recognition.

One important type of real-world information extraction.

e.g. daily action

Typing

Reading

Take off bag

e.g. mutual action(two person Interaction)

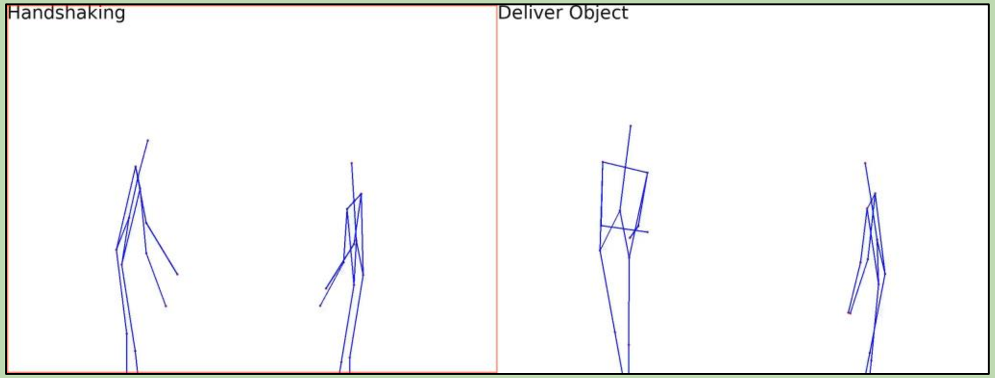
Shaking hands

Hugging

Two-people interaction recognition (TPIR)

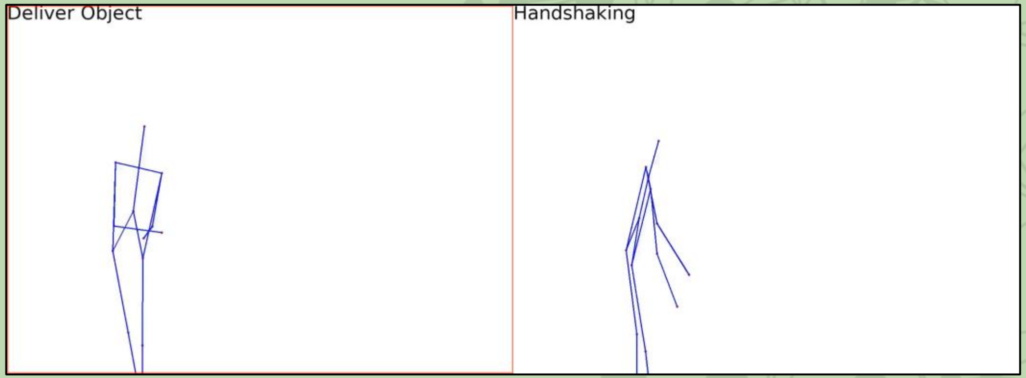
- Open research.
- Model developed for TPIR can serve as a primitive model toward more complex action recognition (e.g. , Multiple Activity Recognition, Collective Activity Recognition, etc).
- The state-of-the-art of TPIR is general action recognition.
  - Solve TPIR problem by refering only from each people in the videos.
  - We notice that it might be less performed since we have to extract the interaction feature.

Two-person interaction



Handshake Deliver Object

An interaction only by one person



Handshake Deliver Object

ST-GCN for Skeleton-Based Action Recognition

- ST-GCN is the state-of-the-art of action recognition which have speciality in:
  - Light.
  - Real-time performance.
  - High performance on large scale dataset.
- ST-GCN
  - Achieve outstanding performance on general action recognition (single action recognition, human-object interaction recognition,etc.)
  - However, there are no graph connection that represent an interaction (Might be less superior to TPIR Case).
  - ST-GCN *does not extract the interaction feature*.
  - In fact, they detect the mutual action by only averaging each actor in the action input.
- Possible solution
  - Providing a new graph connection between actors allow the model to extract interaction feature
  - Expected to be enhancing the performance of ST-GCN on the TPIR problem

Yan, Sijie, Yuanjun Xiong, and Dahua Lin. "Spatial temporal graph convolutional networks for skeleton-based action recognition." In Thirty-second AAAI conference on artificial intelligence. 2018.

Research Objective

- Build two people interaction recognition by:
  - **Enhance ST-GCN performance on recognizing TPIR problem.**

- Contribution:
  - Focused on developing a graph-based deep learning model to solve the TPIR which involved two-person interaction.
  - Propose PAM that is able to capture the pairwise relationship of two graphs on TPIR in which the performance of the ST-GCN can be enhanced.
  - The proposed model outperforms the state-of-the-art methods by validating on NTU RGB+D 60 and NTU RGB+D 120 datasets.

Dataset

- Large Scale dataset. Provide skeleton in 3D coordinate. 60 action and 11 two-people interaction.

NTU-RGB-D 60[4]

- And extension of NTU-RGB-D 60, with 120 action and 27 action for two-people interaction.

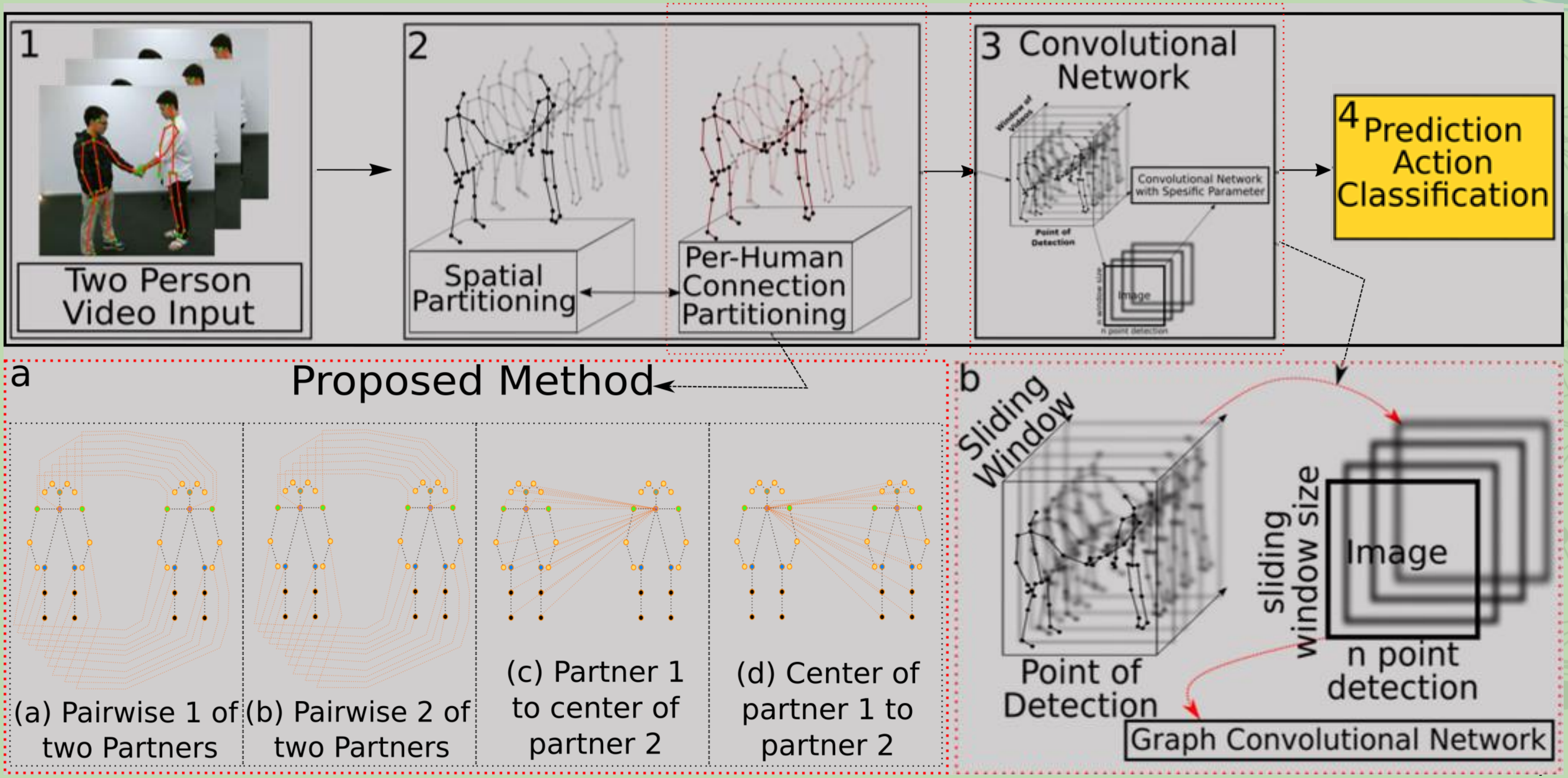
NTU-RGB-D 120[5]

- Youtube Video.
- 6 selected action used.

Kinetics-Dataset[24]

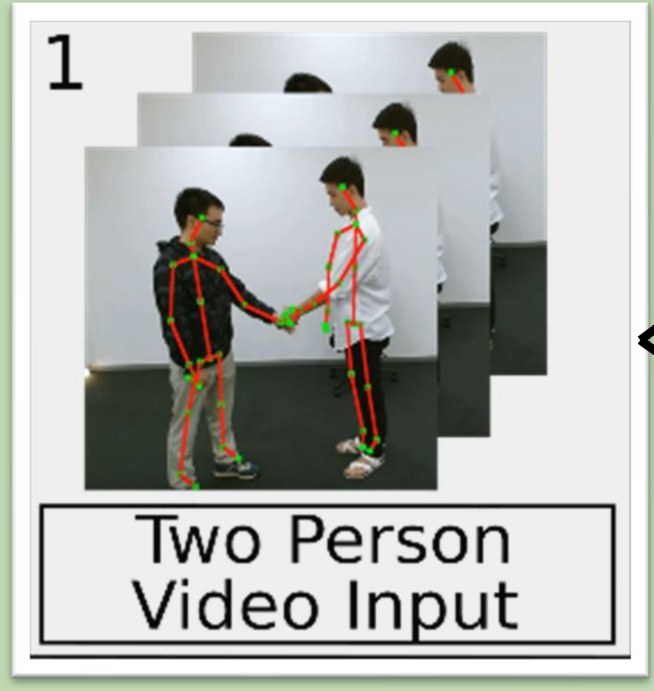
[4] A. Shahroudy, J. Liu, T. Ng, and G. Wang, "NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27-30 June 2016 2016, pp. 1010-1019, doi: 10.1109/CVPR.2016.115.
[5] J. Liu, A. Shahroudy, M. L. Perez, G. Wang, L. Duan, and A. K. Chichung, "NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding," IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1-1, 2019, doi: 10.1109/TPAMI.2019.2916873.
[24]Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., ... & Natsev, P. The kinetics human action video dataset. arXiv 2017. arXiv preprint arXiv:1705.06950.

Proposed Framework

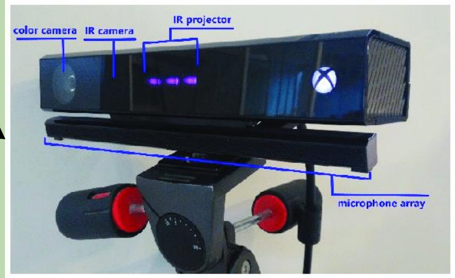


Proposed Framework

Skeleton-Extraction



- Used for Kinetics.
- Input:
  - RGB Images or video
- Outputs:
  - 2D Coordinate for each joint.
  - 1D confidence score for each joint detection.

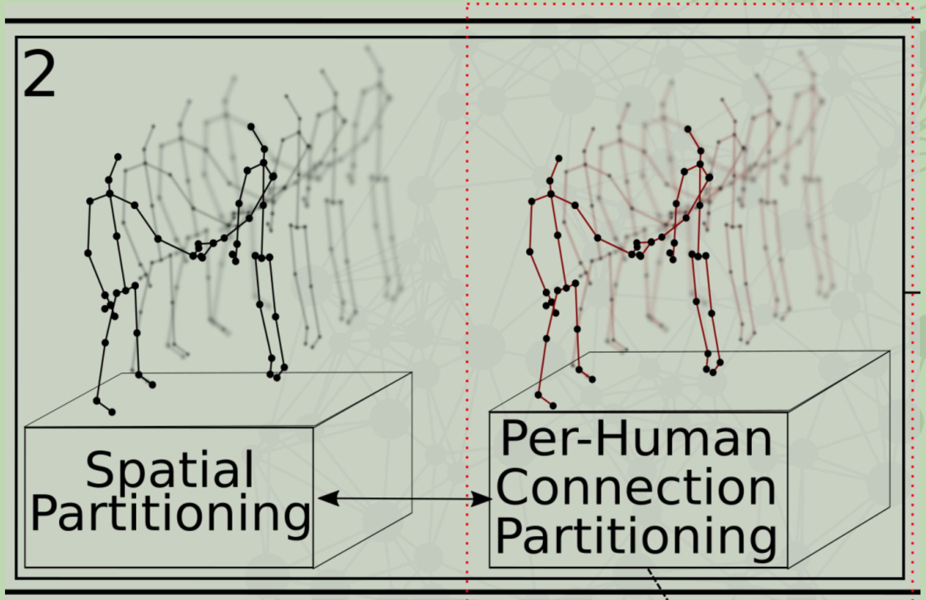


- Not used, only used the dataset which already provide the coordinate.
- 3D coordinate for each joint detection. (2D and 1D depth)
- Directly using skeleton data.
- Used NTU-RGB+D 60 and NTU-RGB+D 120

Methodology

Feature Engineering

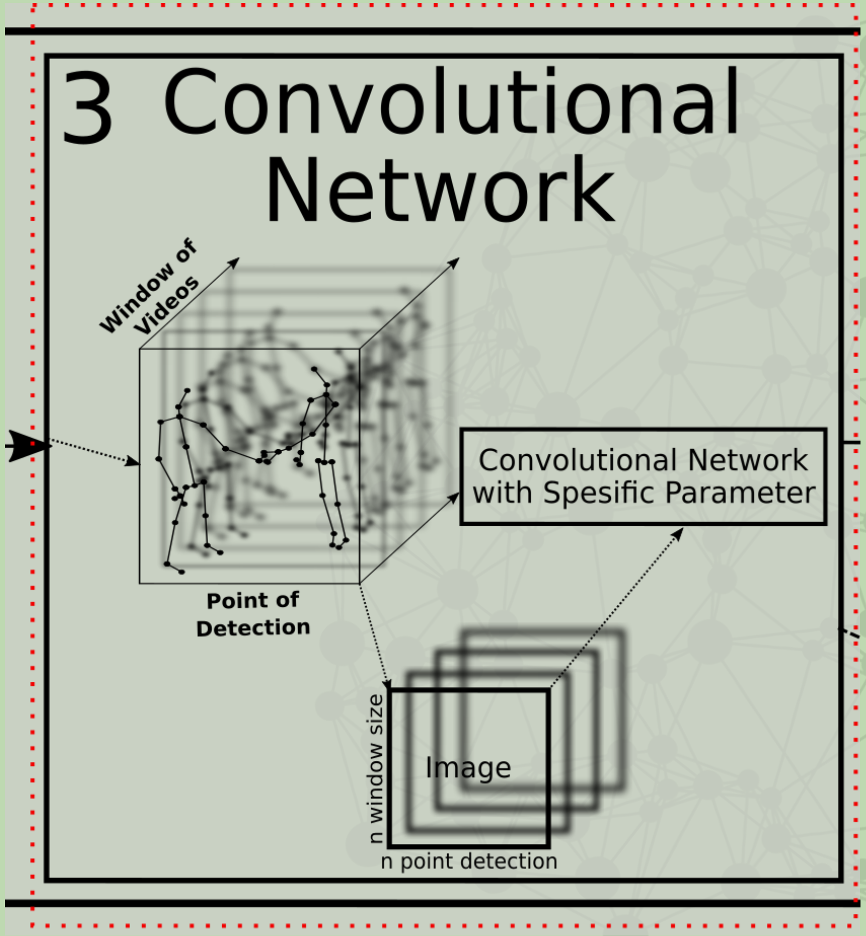
- **Spatial Partitioning**, to capture the relationship for each joint from the same people.
  - There are no representation to capture the relationship of two people interaction.
- **Introduce Per-Human Connection Partitioning.**
  - To capture interaction feature.



Feature extraction

General flow:

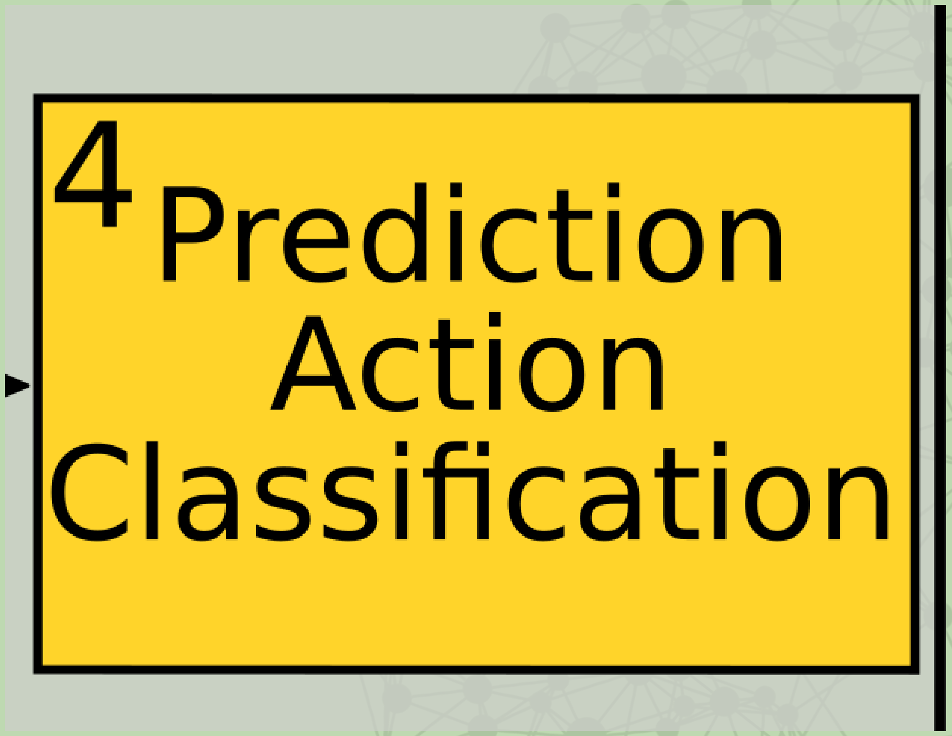
- The coordinate data which is 2D or 3D will be convolving to each other to get the spatial feature.
- The temporal feature is extracted by convolving through the same joints in consecutive frames.
- By combining each joint from every frame and stack for each frame, an image like data will be produced.



Methodology

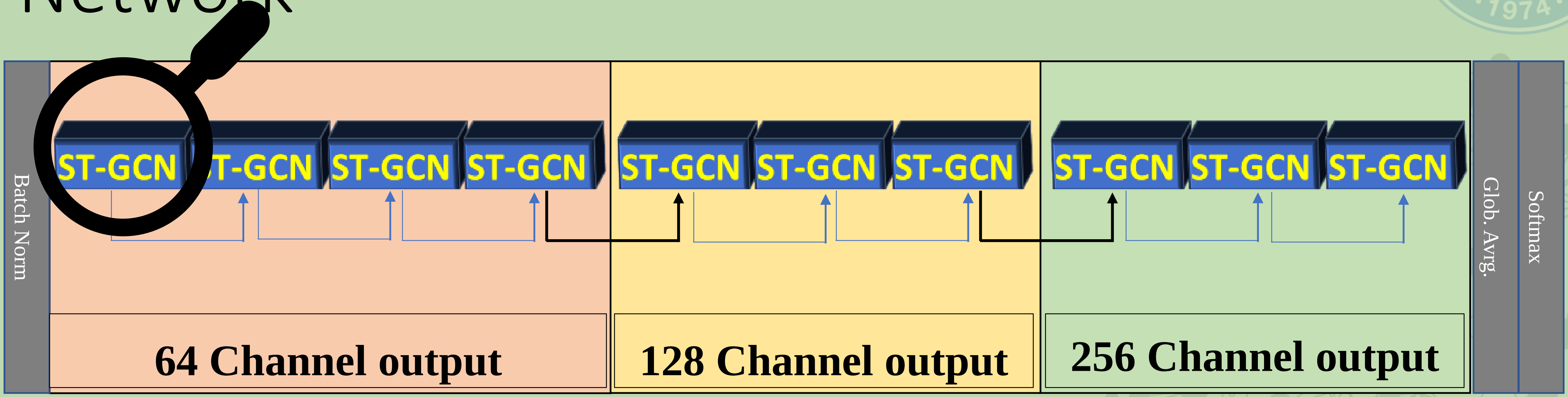
Feature Representation

- Extracted feature is classisified into different class.
- Using SoftMax Classifier



Spatial-Temporal Graph Convolutional Network

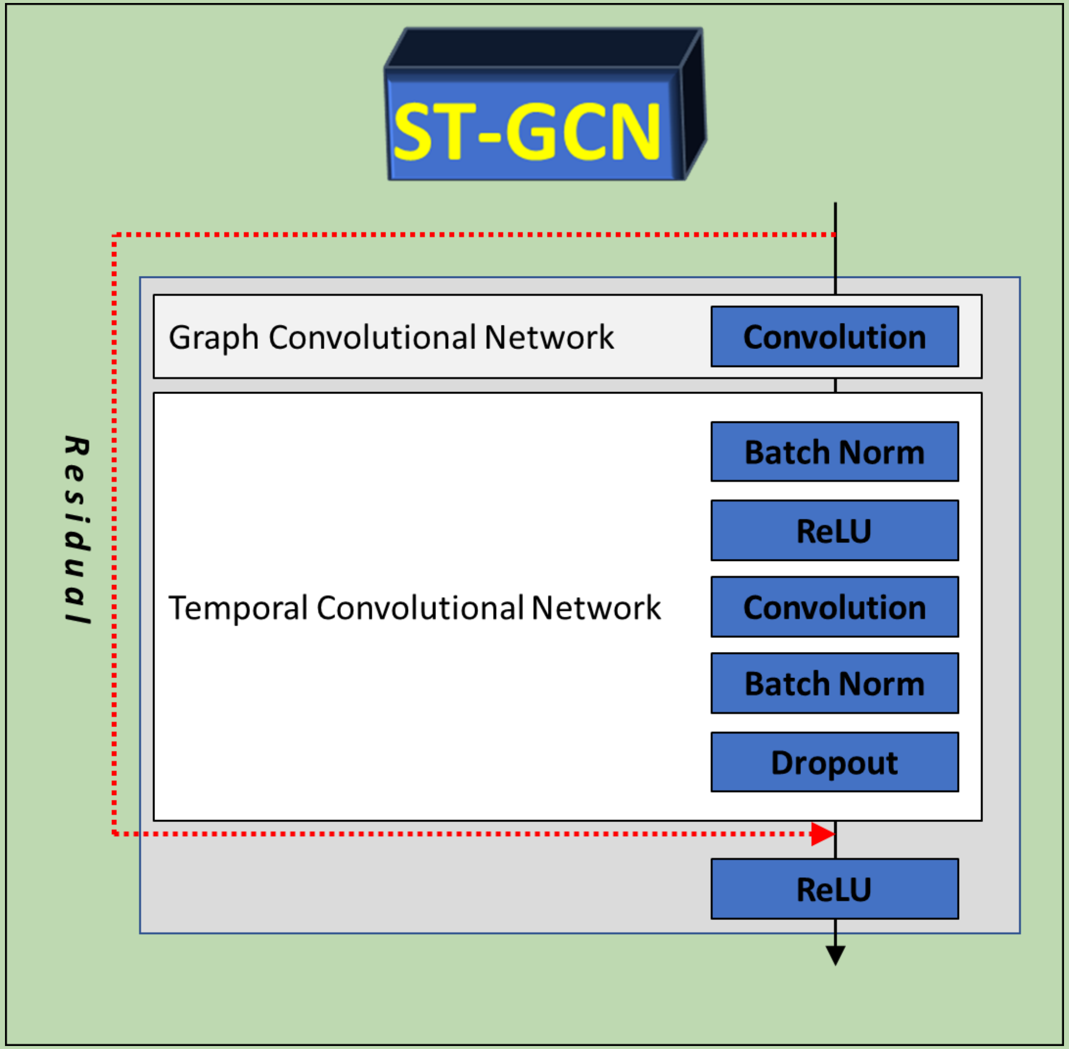
Network



- Batch Normalization(BN) is used before ST-GCN.
- Global Average Pooling at the end of ST-GCN before SoftMax.
- Dropout mechanism is used by factor 0.5.
- The stride parameter is set to 2 for the 4th layer and the 7th layer. (all using 1 stride).

Spatial-Temporal Graph Convolutional Network

Network



- 2 BN, 2 Convolution, 2 Relu, Residual, and Dropout.
- **The modification of this work only on Graph Convolution part.**

Skeleton Graph Convolutional Network

1

$$Y = M \circ \tilde{A} X W$$

2

$$f_{out}(v_i) = \sum_{v_j \in \beta_i} \frac{1}{Z_{ij}} f_{in}(v_j) . w((l_i(v_j))),$$

3

- **Uni-Labelling strategy.**
- $f_{out} = \Lambda^{-\frac{1}{2}} (A + I) \Lambda^{-\frac{1}{2}} f_{in} \boldsymbol{W}$
- A represent the adjacency for intra conection, I for self-connection.

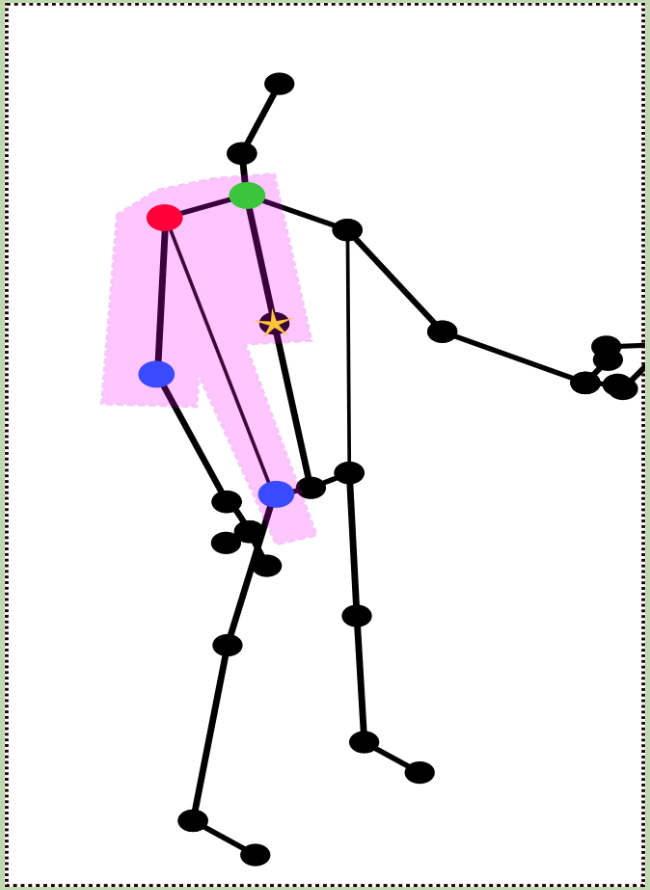
4

- **Spatial configuration partitioning**
- $f_{out} = \sum_{k}^{Kv} W_k (f_{in} A_k)$
- $\boxed{\mathrm{A_k} = \Lambda_k^{-\frac{1}{2}} (\bar{A}_k) \Lambda_k^{-\frac{1}{2}}}$
- $\sum_k A_k = A + I$; k = each subset, A_k, K_v= 3

5

- M are a weight matrix for each adjency matrix in $\tilde{A}$ or A_k.
- $\Lambda_k^{ii} = \sum_j (\bar{A}_k^{ij}) + \alpha$, with α = 0.01 to avoid empety rows.
- $\circ$ Hadamard product (element-wise product),
- $M \in \mathbb{R}^{nxn}$ is the trainable weights for edges,
- $W \in \mathbb{R}^{nxd_{out}}$ is the trainable weights for for vertexes.

Mapping Function for Spatial configuration partitioning



- The adjacency matrix is built based on the distance to the gravity center of the skeleton. The nodes will be labeled as follow:
- $l_{ti}(v_{tj}j) = \begin{cases} 0 & if\ r_j = r_i \\ 1 & if\ rj < r_i \\ 2 & if\ rj > ri \end{cases}$

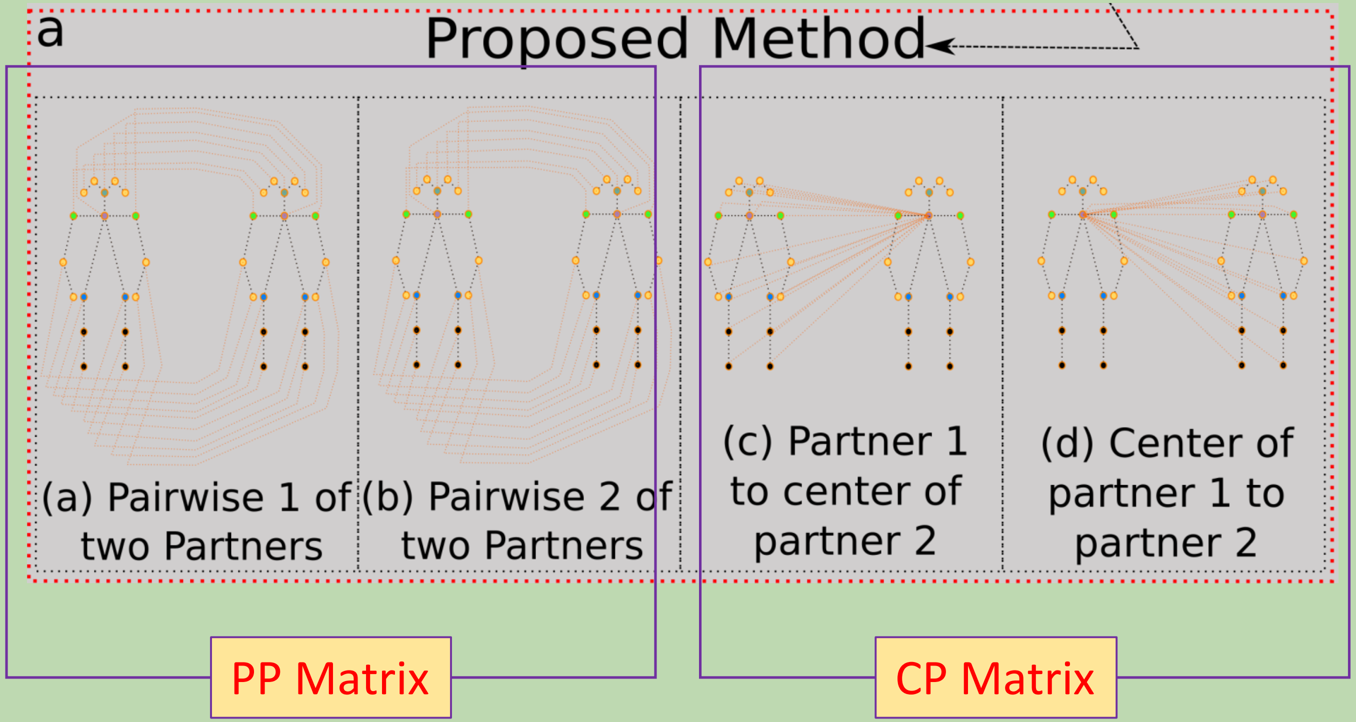
Root node

Centripetal node

Centrifugal node

- r_i is the average distance from the gravity center to joint i over all frames in the training set.

Pairwise-Graph-Connectivity



The proposed PAM was inspired by the combination of:

- learnable edge importance weighting in [11],
- a pairwise adjacency matrix in [34],
- a join graph in [41, 42]

Matrix proposed:

- all joints to all corresponding joints (PP)
- all joints in the first skeleton to the center of gravity on the second skeleton and vice versa (CP)
- The combinations (PCP)

$$f_{out(n,c,s,v,w)} = \sum_{n=0}^{2} f_{in(b,n,c,v,w)} \cdot (\boldsymbol{w}_{(n,v,w)} * l_{(n,v,w)})$$

Evaluation Procedure

NTU RGB+D 60 [4]

- **Cross-Subject (CS)** :
  - Training: 40,320 data
  - Testing: 16,560 data
  - based on actor of the video (one subset for training, the rest for validation).
- **Cross-View(CV)**:
  - Training: 37,920 clips
  - Testing: 18,960 clips.
  - based on camera view (2 and 3 for training, 1 for validation).

NTU RGB+D 120 [5]

- Cross-Subject:
  - Each group consists of 53 subjects.
- Cross-Setup:
  - Training: Camera ID with event number
  - Testing: Camera ID with odd number.
  - 16 camera setup each subset.

[4] A. Shahroudy, J. Liu, T. Ng, and G. Wang, "NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27-30 June 2016 2016, pp. 1010-1019, doi: 10.1109/CVPR.2016.115.
[5] J. Liu, A. Shahroudy, M. L. Perez, G. Wang, L. Duan, and A. K. Chichung, "NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding," IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1-1, 2019, doi: 10.1109/TPAMI.2019.2916873.

Evaluation Procedure

- **Selected Action from Kinetics-Dataset.**
  - 400 action class from 300k YouTube video clips
  - extract the skeleton data with OpenPose

No.	Action Name	No.	Action Name
1.	Hugging	**5.**	Massaging person's head
2.	Massaging back	**6.**	Haking hands
3.	Massaging Feet	**7.**	Slapping
4.	Massaging Legs	**8.**	Tickling

[24]Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., ... & Natsev, P. The kinetics human action video dataset. arXiv 2017. arXiv preprint arXiv:1705.06950.

Experimental Settings

In general,

- Optimization
  - Stochastic Gradient Descent (SGD) + Nesterov Momentum (0.9) in GCN
- Loss function
  - Cross-Entropy + weight decays 0.0001.

NTU RGB+D 120 and NTU RGB+D 60

- All data (single / two person) is formated to have 2 people and 300 frames in each video.
- If there is just one person, the second person coordinates are filled by zero.
- If the length of the video frame is less than 300 -> repeat until reach 300.
- Learning rate: 0.1 For the training option of GCN
- Epoch: 80 (10 dividends on 60 and 70 epochs)

Experimental Settings

For the Kinetics dataset	
	300 frames in every sample where two-person presents in each frame.
	The 300 frames sample is obtained by the same data augmentation mode in NTU RGB+D
	Learning rate: 0.1
	Epoch: 65 (10 dividends on epoch 45 and 55.)

NTU-RGB+D 120 (1/2)

Model	Mode	CS	CV
ST-LSTM [45]	MA	63.0	66.60
GCA-LSTM [46]	MA	70.60	73.70
FSNET [20]	MA	61.20	69.70
LSTM-IRN [15]	MA	77.70	79.60
ST-GCN-PAM(PP)	MA	80.17	85.56
ST-GCN-PAM(CP)	MA	78.93	82.87
ST-GCN-PAM(PCP)	MA	83.28	88.36

PP=Pairwise of two partners; CP=partner-1 to the center of partner-2 and vice versa; PCP = use both PP and CP; MA = trained and tested on mutual actions only.

NTU-RGB+D 120 (2/2)

Model	Mode	CS	CV
ST-GCN [11]	MH	78.7	79.26
	AD	74.6	71.95
Js-AGCN [12]	MH	72.0	72.43
	AD	74.0	70.22
Bs-AGCN [12]	MH	79.28	74.08
	AD	75.23	70.83
2s-AGCN [12]	MH	76.91	80.34
	AD	79.55	78.90
*ST-GCN-PAM(Ours)	MH	82.1	80.91
	AD	73.87	76.85
MH = Tested on mutual action subset only; AD=Tested on all actions label, *PCP			

NTU-RGB+D 60

Action	F2CS		ST-GCN-PAM	
	Precision	Recall	Precision	Recall
Punching	90	91	**97**	**92**
Kicking	88	86	**96**	**95**
Pushing	82	80	**89**	**82**
Pat on back	**88**	**91**	84	90
Point finger	92	83	**99**	**91**
Hugging	88	**91**	**95**	89
Giving something	90	**95**	**94**	90
Touch other’s	95	94	**99**	**95**
Handshaking	96	97	**99**	**98**
Walking toward	76	77	**97**	**94**

Kinetics dataset

Model	Top-1	Top-5
ST-GCN [11]	24.98	43.53
2s-AGCN [12]	44.96	90.34
ST-GCN-PAM(Ours)	41.68	88.91

Conclussion

- An enhancement of ST-GCN was proposed by employing PAM to be able to capture the relationship between the two-person skeletons.
- The proposed ST-GCN-PAM outperforms the-state-of-the-art on TPIR or mutual action of NTU RGB+D 120 by achieving 83.28% (cross-subject) and 88.31% (cross-view) accuracy.
- The model is also superior to original ST-GCN on the multi-human action of the Kinetics dataset by achieving 41.68% in Top-1 and 88.91% in Top-5.

Thank You