PAIRWISE ADJACENCY MATRIX ON SPATIAL TEMPORAL GRAPH CONVOLUTION NETWORK FOR SKELETON-BASED TWO-PERSON INTERACTION RECOGNITION

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

National Taiwan University of Science and Technology, Taiwan

ABSTRACT

Spatial-temporal graph convolutional networks (ST-GCN) have achieved outstanding performances on human action recognition, however, it might be less superior on a two-person interaction recognition (TPIR) task due to the relationship of each skeleton is not considered. In this study, we present an improvement of the ST-GCN model that focused on TPIR by employing the pairwise adjacency matrix to capture the relationship of person-person skeletons (ST-GCN-PAM). To validate the effectiveness of the proposed ST-GCN-PAM model on TPIR, experiments were conducted on NTU RGB+D 120. Additionally, the model was also examined on the Kinetics dataset and NTU RGB+D 60. The results show that the proposed ST-GCN-PAM outperforms the-state-of-the-art methods on 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 the 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.

Index Terms— skeleton-based action recognition, spatial-temporal graph convolution network, two-person-interaction, pairwise adjacency matrix

1. INTRODUCTION

The usage of the human skeleton is now getting more and more attention in the area of human action recognition because of the robustness of adapting to the illumination change and scene variation [1, 2]. The vertex of human skeleton can be easily obtained by using the pose estimation such as OpenPose [3] or by accurate depth sensors [4, 5]. Many skeleton-based human action recognition models were proposed by applying either handcrafted features [6] or the deep learning method to train the classifier for recognition. Conventionally, the handcrafted features such as joint location or the joint angles are used to represent the human body [7]. However, the performance of handcrafted approach is not satisfied [1]. Currently, the deep learning model is more popular.

Within deep learning based methods, many works used graph-based methods which present a human skeleton as a graph for the input of deep learning model. The challenge lies in how to generalize a graph model into Convolution Neural Network (CNN) to be able to process arbitrarily structured graphs. Kipf and Welling first proposed a method called graph convolution network (GCN) to combined the graph model with the CNN to represent spatial data [8]. Later on, based on the GCN sturcture, a method named spatial-temporal graph convolutional networks (ST-GCN) was proposed to combine the temporal features of data [9]. Shi et al. continued to utilize ST-GCN model to perform the human action recognition [1, 10, 11].

Presently, ST-GCN is mainly focused to solve single action recognition but not two-person interaction recognition (TPIR) which involves two people to perform the mutual actions. The examples of TPIR can be illustrated in Fig. 1. As can been seen, Fig. 1 (i)(ii) show human hugging interaction and their skeletons while Fig. 1 (iii)(iv) show two people are hand-shaking. Similar to single human action recognition, TPIR aims to recognize the mutual action among two people.

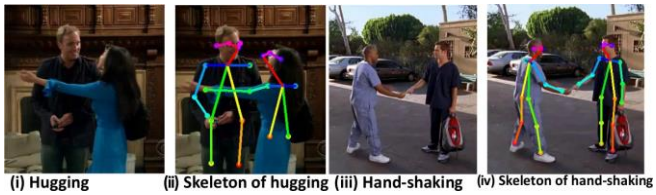

Figure 1 An example of two-person interactions

In literature, several approaches have been developed to solve TPIR problem [12]. For instance, Sener and Ikizler-Cinbis proposed a framework utilizing two-person descriptors by an embedded representation integrated with spatial multiple learning to captures the relationship of person-person [13]. Another work utilized the body-pose features incorporated with joint features and plane features to perform a multiple instance learning, considered as one kind of supervised learning [14]. However, most of the previous works including the papers mentioned above all used a region of interest based on image-based approach.

In this paper, we tried to enhance the original ST-GCN to be capable to capture the relationship between two persons for TPIR problem. A new data presentation called pairwise adjacency matrix (PAM) was proposed to contain the intra-inter bidirectional relationship of a pair of two graphs which are constructed by two human skeletons. The proposed PAM is added into ST-GCN to represent the person-to-person connection. By the backprorogation mechanism of ST-GCN, the model is trained to classify the two-person interaction.

To validate the effectiveness of the proposed ST-GCN-PAM model on TPIR, experiments were conducted based on NTU RGB+D 120, Kinetics dataset and NTU RGB+D 60. The results were compared with the single-person ST-GCN as baseline and graph-based LSTM model [15] as the benchmark method. The results shows the proposed method outperforms the mentioned methods on TPIR. The contributions of this work are summarized as follows:

- This work focused on developing a graph-based deep learning model to solve the TPIR which involved two-person interaction.
- The propose PAM 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.

2. PAIRWISE ADJACENCY MATRIX ON ST-GCN

Our model follows the original work of ST-GCN on action recognition [9]. The major difference between our work and ST-GCN is a new graph construction to represent the connection of two people in one frame for the streaming video. Fig. 2 describes the flow of the proposed framework. The top of Fig. 2 illustrates four stages of the proposed framework from left to right. 1) 2D or 3D joint coordinates of skeleton joints are generated from two-person human interaction video. 2) the model constructing the adjacency matrix which defines the pairwise adjacency among all of joints in a skeleton. Then, the model will construct spatial graph features based on the adjacency matrix for each skeleton. 3) a sliding window is used to obtain a temporal feature at the same time, hence, the spatial and temporal features can be treated as an image as the input of CNN. 4) the model employs a SoftMax classifier to identify the interaction action.

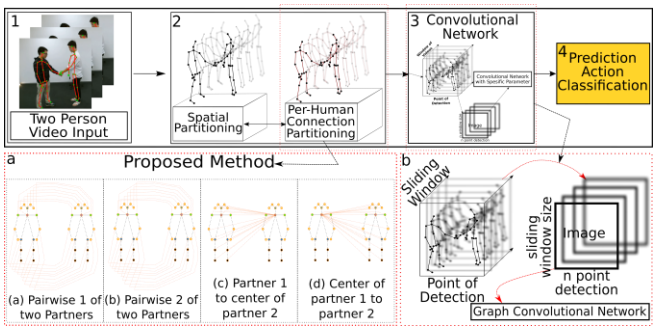

Figure 2 Overview of the proposed ST-GCN-PAM framework

The bottom of Fig. 2 from left to right illustrates a) how to construct the adjacency by the proposed methods, and b) how ST-GCN works, respectively. In a), basically, two adjacency matrices were proposed and compared. (a) and (b) are cases where the pairwise of two partners links every corresponding joint among the skeletons. (c) and (d) are to link the center of one skeleton with all joints of another skeleton. In b), by following the settings of ST-GCN [9], the sliding window is used to collect temporal features in the corresponding video frames. For example, a 18-joint skeleton with 150 sliding window will construct a 18×150 matrix containing the spatial and temporal features of human interaction across video frames. In this work, the training procedure was done using an end to end manner with a backpropagation procedure. The following subsections describes the technical details of the procedures.

2.1. Graph Construction

The human skeleton can be represented as a 2D or 3D coordinates of each human joint in each frame as a sequence of vectors. We can utilize the skeleton information by employing a spatial-temporal graph to model the structured data among all skeleton joints. Principally, this procedure of creating single feature vector per-frame was followed by the data processing in ST-GCN [9].

The undirected spatial-temporal graph can be used to feature both inter-frame and intra-body connections for each human body skeleton. The human skeleton sketch in Fig. 3 visualizes the examples of inter-frame and intra-body connections. The left sketch in Fig. 3 presents an example of the constructed spatial-temporal skeleton graph (inter-frame connection) where a joint is represented as a vertex and their corresponding connections across frames are represented as spatial edges (yellow lines). The color dots of the right sketch in Fig. 3 presents an example of intra-body connections defined in [9].

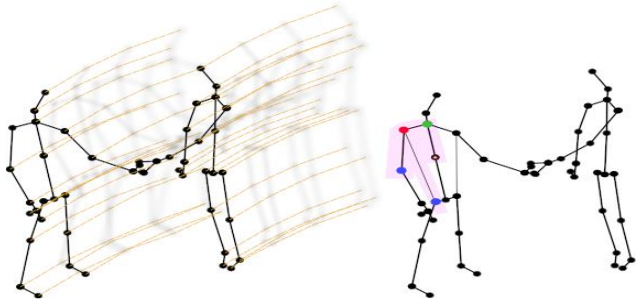

Figure 3 Examples of inter-frame (left) and intra-body (right) connections

In addition, we follow the original ST-GCN [9] to find the feature map on the spatial dimension. The output value of the feature map for a single-channel at spatial location x can be written as:

$$f_{out}(x) = \sum_{h=1}^{K} \sum_{w=1}^{K} f_{in}\big(\boldsymbol{p}(\boldsymbol{x}, h, w)\big) \cdot \boldsymbol{W}(h, w) \quad (1)$$

where f_{in} denotes the feature map; $\boldsymbol{x}$ denotes the vertex location of the graph; $\boldsymbol{p}$ denotes the sampling area of the convolution at $\boldsymbol{x}$ vertex; both h and w are the number of nodes in skeleton; K is the kernel size. In this work, $\boldsymbol{p}$ is defined as 1 distance among the neighbor vertexes to the target vertex. $\mathbf{W}$ is the weight function which provides a weight vector in the input channel dimension that has the same operation with the original convolution by giving a fixed weight vector. However, the number of node is varied mentioned in [9]. The mapping function is then employed to simplify the labelling process by transform a varied vector to a fixed vector automatically.

Then, the mapping function in ST-GCN is defined to map the weight vector into the adjacency matrix by the partitioning strategy. In [9], the best partitioning strategy is spatial configuration which is applied in this work. The example of this mapping strategy is shown in the right skeleton of Fig. 3. The enclosed curve (the pink area) of the body skeletons will be divided into three subsets. The first subset is the root node (the red circle). The second is the centrifugal group (the green circle) which are a set of nodes having a closer distance to the center of gravity (the yellow dot) than the root node. Third subset is centripetal group (the blue circle) which are a set of nodes having longer distance than the root node [9].

2.2. Implementation

The implementation of graph convolution is not as straightforward as other types of convolution network. In detail, the shape of the feature map before processed is actually $C \times T \times N$. Where C, T, N represents the number of channels, temporal length (windows size), and the number of vertexes respectively. In this work, N is set to be 18 points for the input from OpenPose [16] and 25 points for the input from NTU-RGB-D 120. Original Graph Convolution [8] formulated GCN as:

$$f_{out} = \Lambda^{-1/2}(A + I)\Lambda^{-1/2} f_{in} W \quad (2)$$

then follows ST-GCN [9, 11] with considering the normalization part $\Lambda^{-1/2}(A+I)\Lambda^{-1/2}$ where A is the adjacency matrix, I denotes the identity matrix and Λ diagonal node degree matrix. In equation (2), W is the attention map (weighted function) in the convolution operation. The equation (1) and (2) are combined and transformed into equation (3):

$$f_{out} = \sum_k^{Kv} W_k (f_{in} A_k) \otimes L_k \quad (3)$$

where K_v represent the kernel size of the spatial dimensions. In our case, the spatial configuration which has $K_v = 3$ (kernel sizes) is used. A_k represent the adjacency matrix as $\mathrm{A_k} = {\Lambda_k}^{-1/2}(A_k + I)\Lambda_{\mathrm{k}}^{-1/2}$. This part is used to extract the connected vertexes to their corresponding weight vector W_k. In the implementations, α is introduced to avoid empty rows in $\mathrm{A_k}$ which caused by a non-connected graph. In $\Lambda_k^{ii} = \sum_j (\bar{A}_k^{ij}) + \alpha$. α is set to 0.001 following [10]. L_k denotes a learnable importance weighting or attention map that indicates the importance of each vertex, and $\otimes$ denotes dot product.

2.3 Pairwise Adjacency Matrix

The proposed PAM was inspired from the combination of learnable edge importance weighting in [9], a pairwise adjacency matrix in [17], and a join graph in [18, 19]. A pairwise adjacency is used to create a connection between two graphs. On the other hand, a join graph is used as a rule of connection between two entity graphs. This combination creates a feature to represent a relationship between two peoples in frame t. We call this connection as pairwise connection with the pairwise matrix as the mapping function.

Fig. 2a visualizes the proposed graph. A mapping function is used to create a connection between vertex and between vertex in corresponding frames [9, 11]. We employed the same mapping function to construct a connection between two different skeletons. By using this method, two features: the connection of the corresponding frame and the connection between people can be generated.

In detail, there will be two matrices for each skeleton to connect to another skeleton. The first matrix comes from the connections from all joints to the all corresponding joints. This kind of the connection is based on the relationship that we believe has a high probability when two people performs mutual action. The second matrix comes from the connections between all joints in the first skeleton to the center of gravity on the second skeleton. This connection aims to obtain a feature representation of mutual action recognition.

In the original ST-GCN [9, 11], the graph calculation is based on the number of vertex c. However, in this work, the graph calculation is based on the number of people n. Then, follows a bipartite graph connection, the calculation after the convolution operation of graph convolutional network [17] is shown as equation (4):

$$f_{out\,(n,c,s,v,w)} = \sum_{n=0}^{2} f_{in(b,n,c,v,w)} \cdot \left(\boldsymbol{w}_{(n,v,w)} * l_{(n,v,w)}\right) \quad (4)$$

where f_{in} is the output of graph convolution, $\boldsymbol{w}$ is the weighted function shown in equation (1), and l is the mapping function to make the $\boldsymbol{w}_{(n,v,w)} * l_{(n,v,w)}$ is learnable for GCN. $b, c,$ stands for a batch number, and channel size at each convolution location, respectively. v, w represents the number of vertexes.

3. EXPERIMENTS

To validate the propoased method, the state-of-the-art methods of action recognition were compared with the proposed ST-GCN-PAM in mutual action on NTU RGB+D 60, NTU RGB+D 120, and Kinetics-Skeleton datasets. For each dataset, two experiements were conducted. One used all of data including single and two-person actions, and another only used for data with mutual actions. The experiments were conducted based on the same computer specifications: using PyTorch deep-learning framework on two-GPU RTX 1080Ti 128GB memory.

3.1. Dataset

NTU RGB+D 120 [5] is the extended version of NTU RGB+D 60 [4] which is the most widely used of an action recognition dataset. This dataset contains 114.480 action clips in 120 classes. Benchmark evaluation of this dataset are Cross Subject (CS) and Cross View (CV) stated in the paper [5]. For CS, the data was divided into a training set which comes from 53 subjects and will test by the other 53 subjects. For CV, the data was divided by camera ID, the data with the even camera was reserved for training, while the data with the odd camera was for validation.

The version of Kinetics-Skeleton data used in this work has 400 action class from 300k YouTube video clips. The OpenPose [3] was used to extract the skeleton data. Table 1 shows the specified actions used for training and validation.

Table 1. Specific actions used from the Kinetics dataset

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

3.2. Experimental Setting

In general, we applied Stochastic Gradient Descent (SGD) as the optimization strategy with Nesterov Momentum (0.9) in GCN. The loss function to backpropagate gradients is Cross-Entropy with weight decays 0.0001.

For the NTU RGB+D 120 (and 60) dataset, data was formatted to have 2 people and to have 300 frames in each video. If there is just one person, the second people coordinate are filled by zero. If the length of the video frame is less than 300, it was repeated until it reaches 300. For the training option of GCN, this dataset is set to have a learning rate of 0.1 and run with 80 epochs with 10 dividends on 60 and 70 epochs.

For the Kinetics dataset, data was formatted to have150 frames in every sample where two person presents in each frame. The 150 frames sample is obtained by the same data augmentation mode in NTU RGB+D. The learning rate is set to have 0.1 with 10 dividends on epoch 45 and 55. The maximum learning epoch is 65.

3.3. Result and Discussion

NTU RGB+D 120 datasets. The ST-GCN-PAM was trained and tested on a mutual action subset of NTU RGB + D 120 dataset and was compared to the state-of-the-art methods. The proposed method was also compared to the original ST-GCN [9]. Also the recognition accuracies incorporated with pairwise of two-partner matrix only (PP), partner-1 to the center of the partner-2 matrix (CP), and employed both of the two types matrices (PCO) were investigated.

Table 2 shows the results. As can be seen, the proposed ST-GCN-PAM outperforms the other methods on mutual action tasks

[15, 20-22]. Based on these experimental results, using the pairwise of two-partner (PP) yields more significant improvement than utilizing only the center of partner-1 to partner-2 and vice-versa (CP). However, incorporating both PP and CP (PCP) mutually yields the better result. Therefore, for the rest of experiments, the PCP matrix was used for comparison.

Table 2. Comparison of performance to the state-of-the-art methods on Mutual Action Subset of NTU-RGBD + 120 dataset

Model	Mode	CS	CV
ST-LSTM [20]	MA	63.0	66.60
GCA-LSTM [21]	MA	70.60	73.70
FSNET [22]	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.

Table 3. Comparison of performance to the state-of-the-art methods on NTU-RGBD + 120 dataset

Model	Mode	CS	CV
ST-GCN [9]	MH	78.7	79.26
	AD	74.6	71.95
Js-AGCN [11]	MH	72.0	72.43
	AD	74.0	70.22
Bs-AGCN [11]	MH	79.28	74.08
	AD	75.23	70.83
2s-AGCN [11]	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

Table 3 shows the comparison of the state-of-the-art methods of action recognition methods [9, 11] which trained on the whole data of NTU RGB+D 120 then tested on all action data tests (AD) including single and mutual actions and also tested on the subset of NTU RGB+D 120 including the mutual actions only (MH). As can been seen, the proposed ST-GCN-PAM achieved the best result on MH test data. It means that ST-GCN-PAM with PCP matrix is able effectively perform the mutual action recognition task.

The purpose of testing on AD is to show the capability of the proposed ST-GCN-PAM on single action problem, although it was designed for solving the two-person action problem. The result shows 2s-AGCN has the better performance than ST-GCN-PAM. However, ST-GCN-PAM still obtained the comparable accuracy on single person action. Becasue 2s-AGCN utilizes two-stream of joints skeleton and body-skeleton is learned either individually or uniformly, it is believed that this setting gives the beneficial information to the model. Therefore, incorporating the proposed PAM with 2s-AGCN can be considered as the further future work.

NTU RGB+D 60 Dataset. ST-GCN-PAM was also evaluated with NTU RGB+D 60 on mutual action tasks. Notice that the evaluation was conducted on CV Protocol only because the other previous works only evaluated on CV. Table 4 shows the comparison of ST-GCN-PAM with the benchmark method F2CS [23]. The precision and recall were evaluated in this experiement by following [23]. The resut shows ST-GCN-PAM outperforms F2CS in most of the cases, for example, the action of punching, kicking, pushing, point finger, touch other's, hand-shaking, and walking-toward. Particularly, in the action of walking toward, our method gains significant improvement by 97% of precision and 94% of recall. Moreover, an interesting finding shows the proposed model has more precision on the action of hugging, and 'giving something', but less recall. Based on this experimental result, the robustness of the proposed is confirmed.

Table 4. Comparison of performance in Cross View (NTU RGB+D 60) to mutual action with F2CS model [23]

Action	F2CS		ST-GCN-PAM	
	Prec.	Rec.	Prec.	Rec.
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 smth.	90	**95**	**94**	90
Touch other's	95	94	**99**	**95**
Handshaking	96	97	**99**	**98**
Walking twrd.	76	77	**97**	**94**

Kinetics Dataset. Please notice that this dataset is wild video, and also a multi-person action problem. It means this dataset is not specific to two-person interaction or mutual action such as NTU RGB+D dataset. The comparison can be seen in Table 5. As we can see, ST-GCN-PAM significantly outperforms the baseline ST-GCN in both Top-1 and Top-5. Although, ST-GCN-PAM is less superior than 2s-AGCN model, the result is still comparable. As mentioned in the result of NTU RGB+D 120 datasets, we consider applying PAM on 2s-AGCN as our future work.

Table 5. Comparison of performance to the state-of-the-art model on kinetics dataset (Multi-Person Action)

Model	Top-1	Top-5
ST-GCN [9]	24.98	43.53
2s-AGCN [11]	44.96	90.34
ST-GCN-PAM(Ours)	41.68	88.91

4. CONCLUSION

In this study, 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 Kinetics dataset by achieving 41.68% in Top-1 and 88.91% in Top-5 and the result is still comparable with the most recent 2s-AGCN model. In the future, the proposed PAM matrix can be incorporated with any graph-based model such as 2s-AGCN to enhance the performances in terms of accuracy. Moreover, applying on the most recent multi-stream model is also an open direction for investigation.
Code is available at *https://github.com/ajisetyoko/mutual-action* .

5. REFERENCES

[1] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-Based Action Recognition with Directed Graph Neural Networks," in *Proceedings*

of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2019), Long Beach, CA, 2019, pp. 7912-7921.
[2] M. Fu *et al.*, "Human Action Recognition: A Survey," in *Proceedings of the 5th International Conference on Signal and Information Processing, Networking and Computers (ICSINC)*, Yuzhou, China, 2018, pp. 69-77: Springer Singapore.
[3] Z. Cao, T. Simon, S. Wei, and Y. Sheikh, "Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA 2017, pp. 1302-1310.
[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)*, 2016, pp. 1010-1019.
[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.
[6] M. J. Marín-Jiménez, E. Yeguas, and N. Pérez de la Blanca, "Exploring STIP-based models for recognizing human interactions in TV videos," *Pattern Recognition Letters*, vol. 34, no. 15, pp. 1819-1828, 1 November 2013 2013.
[7] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3d skeletons as points in a lie group," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 588-595.
[8] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," presented at the 5th International Conference on Learning Representations (ICLR-17), Toulon, France, April 24-26, 2017. Available: https://openreview.net/forum?id=SJU4ayYgl
[9] C. Li, Z. Cui, W. Zheng, C. Xu, and J. Yang, "Spatio-Temporal Graph Convolution for Skeleton Based Action Recognition," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, New Orleans, Louisiana, USA, 2018, pp. 3482-3489: AAAI Press.
[10] W. Zheng, P. Jing, and Q. Xu, *Action Recognition Based on Spatial Temporal Graph Convolutional Networks* (Proceedings of the 3rd International Conference on Computer Science and Application Engineering). Sanya, China: Association for Computing Machinery, 2019, p. Article 118.
[11] L. Shi, Y. Zhang, J. Cheng, and H.Lu, "Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition," in *CVPR 2019*, Long Beach, CA, USA, 2019.
[12] A. Stergiou and R. Poppe, "Analyzing human–human interactions: A survey," *Computer Vision and Image Understanding*, vol. 188, p. 102799, Nov 2019.
[13] F. Sener and N. Ikizler-Cinbis, "Two-person interaction recognition via spatial multiple instance embedding," *Journal of Visual Communication and Image Representation*, vol. 32, pp. 63-73, October 2015 2015.
[14] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Providence, RI, USA 2012, pp. 28-35.
[15] M. Perez, J. Liu, and A. C. Kot, "Interaction Relational Network for Mutual Action Recognition," *arXiv:1910.04963*,
[16] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields," presented at the 2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR 2017), Honolulu, HI, USA, July 21-26, 2017. Available: https://doi.org/10.1109/CVPR.2017.143 http://doi.ieeecomputersociety.org/10.1109/CVPR.2017.143
[17] S. Mehnaz and M. S. Rahman, "Pairwise compatibility graphs revisited," in *2013 International Conference on Informatics, Electronics and Vision (ICIEV)*, Dhaka, Bangladesh, 2013, pp. 1-6.
[18] D. M. Cardoso, M. A. A. de Freitas, E. A. Martins, and M. Robbiano, "Spectra of graphs obtained by a generalization of the join graph operation," *Discrete Mathematics*, vol. 313, no. 5, pp. 733-741, 2013/03/06/ 2013.
[19] G. Bergami, M. Magnani, and D. Montesi, "A Join Operator for Property Graphs," in *EDBT/ICDT Workshops*, Venice, Italy, 2017.
[20] J. Liu, A. Shahroudy, D. Xu, A. C. Kot, and G. Wang, "Skeleton-Based Action Recognition Using Spatio-Temporal LSTM Network with Trust Gates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 12, pp. 3007-3021, 09 November 2017 2018.
[21] J. Liu, G. Wang, P. Hu, L. Duan, and A. C. Kot, "Global Context-Aware Attention LSTM Networks for 3D Action Recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 3671-3680: IEEE.
[22] J. Liu, A. Shahroudy, G. Wang, L. Duan, and A. K. Chichung, "Skeleton-Based Online Action Prediction Using Scale Selection Network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1-1, 2019.
[23] T. M. Le, N. Inoue, and K. Shinoda, "A Fine-to-Coarse Convolutional Neural Network for 3D Human Action Recognition," *arXiv e-prints*, Accessed on: May 01, 2018Available: https://ui.adsabs.harvard.edu/abs/2018arXiv180511790L