IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 26, 2019, accepted December 17, 2019, date of publication December 23, 2019,
date of current version January 2, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2961770

Action Recognition Using Attention-Joints Graph Convolutional Neural Networks

TASWEER AHMAD[1,2], HUIYUN MAO[3], LUOJUN LIN[1], AND GUOZHI TANG[1]
[1]School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510000, China
[2]Department of Electrical Engineering, COMSATS University Islamabad, Sahiwal Campus, Sahiwal 57000, Pakistan
[3]School of Computer Science, South China University of Technology, Guangzhou 510000, China

Corresponding author: Tasweer Ahmad (tasveer28@yahoo.com)

This work was supported in part by the Chinese Scholarship Council (CSC), in part by the Natural Science Foundation of Guangdong Province (GD-NSF) under Grant 2017A030312006, in part by the National Key Research and Development Program of China under Grant 2016YFB1001405, in part by the NSFC under Grant 61673182 and Grant 61771199, in part by the Foundation of Guangdong Science and Technology Department (GDSTP) under Grant 2017A010101027, and in part by the Guangzhou Science, Technology and Innovation Commission (GZSTP) under Grant 201704020134.

ABSTRACT Human skeleton contains significant information about actions, therefore, it is quite intuitive to incorporate skeletons in human action recognition. Human skeleton resembles to a graph where body joints and bones mimic to graph nodes and edges. This resemblance of human skeleton to graph structure is the main motivation to apply graph convolutional neural network for human action recognition. Results show that the discriminant contribution of different joints is not equal for different actions. Therefore, we propose to use attention-joints that correspond to joints significantly contributing to the specific actions. Features corresponding to only these attention-joints are computed and assigned as node features of the graph. In our method, node features (also termed as attention-joint features) include the i) distances of attention-joints from the center-of-gravity of human body, ii) distances between adjacent attention-joints and iii) joints flow features. The proposed method gives a simple but more efficient representation of skeleton sequences by concatenating more relative distances and relative coordinates to other joints. The proposed methodology has been evaluated on single image Stanford 40-Actions dataset, as well as on temporal skeleton-based action recognition PKU-MDD and NTU-RGBD datasets. Results show that this framework outperforms existing state-of-the-art methods.

INDEX TERMS Human action recognition, attention-joints, graph convolutional neural network.

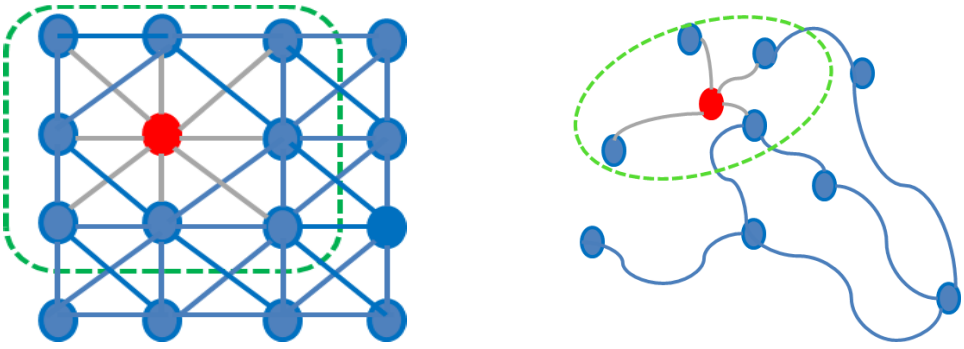
I. INTRODUCTION

Human action recognition in videos has numerous practical applications such as video surveillance, video content analysis, health-care and entertainment. In literature, different modalities have been investigated for action recognition in videos, such as RGB-images, optical flow/warped optical flow and body skeletons. In [1] and [2] human actions are recognized using spatial and temporal two-stream network. Further, a skeleton-based approach for human action recognition has been exercised in [3], [4]. This approach has achieved early success in action recognition, because human skeletons are invariant to illumination and appearance. Human skeletons can be represented in the form of graphs, however the direct application of Convolutional Neural Networks (CNN) on human skeletons is not so intuitive. Specifically, owing to Graph Convolutional Neural Networks (GNN) that CNN is applicable on non-euclidean domains such as graphs of arbitrary nodes and edges. Over the years, GNN has been successfully applied to many fields such as image and text classification [5], object recognition [6] and human activity recognition [7]. Graph convolution neural networks are such powerful models that a randomly initialized two layer GNN can produce useful feature representation of nodes in a network [8].

In graph convolutional network, graph statistics (i.e. graph features) are exploited by properly devising graphlet kernels [9], which promises the occurrences of various graphlets (i.e. subgraphs) on a graph. Intuitively, graphs belonging to a particular class should have specific features conditioned on that class. Such distinctive graph features are beneficial for classification tasks.

Convolutional neural networks perform well on dealing with euclidean data, e.g. images, voice or videos, however,

The associate editor coordinating the review of this manuscript and approving it for publication was Zheng Xiao.



(a) Convolution operation in euclidean domain (b) Convolution operation on a graph

FIGURE 1. Illustration of Euclidean convolution vs Graph convolution [24].

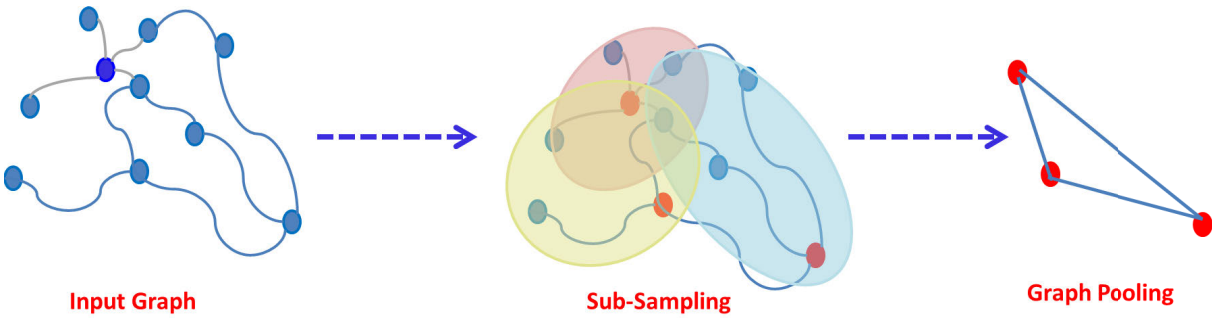


FIGURE 2. Sub-sampling and pooling operation on graph, (better perceived in color).

it is more likely to fail to non-euclidean data. To address this, graph convolutional networks also known as geometric CNNs, can be applied for node classification and link prediction in non-euclidean space such as social networking, molecular biology and brain-signal processing. Empirically, graph convolution networks extract high-level features from graphs and are hence suitable for application such as human action recognition via skeletons with body joints corresponding to nodes and bones between joints corresponding to edges, respectively. An analogy from conventional spatial convolution to graph convolution is illustrated in Figure 1, where the image pixels are represented as graph nodes and their spatial relationships are delineated as graph edges. Analogously, spatial convolutional kernels are extended to graph convolutional kernels to compute the sum-of-product over neighboring nodes. The pooling operation is defined in the form of graph coarsening and partitioning. Balanced cuts and heavy edge matching (HEM) are the techniques used for graph pooling. The graph sub-sampling and pooling operations are explained in Figure 2.

From spatial convolution, it is clear that a graph may contain redundant or noisy edges. Therefore, it is pragmatic that we use attention mechanism to emphasize the significant nodes, while suppressing redundant nodes. This concept is utilized in a way that attention-joints contribute more to final action recognition, while redundant nodes may lead to noisy or false prediction. For instance, the attention-nodes of actions like drinking and phoning, are from body arms, head and neck as illustrated in Figure 3.

Contribution from our work is summarized as follow: 1) We discover and develop the most relevant attention-joints for some actions. 2) We utilize normalized distances and joints-flow based features for such attention-nodes. 3) A new attention-joints graph convolutional neural network is designed for skeleton-based action recognition, which achieves state-of-the-art performance on three public benchmarks.

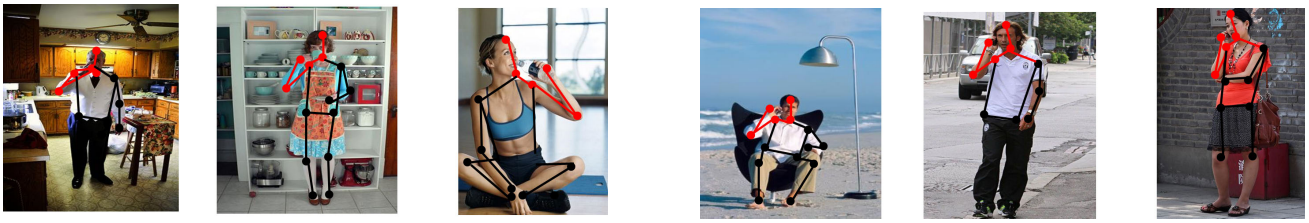
The rest of paper is organized as follows: Section II describes literature work, and Section III explains proposed methodology. Experimental details and results are introduced and analyzed in Section IV.

II. RELATED WORK

In literature, the problem of skeleton based action has been addressed by i) Convolutional neural network and ii) Graph Convolutional networks.

A. CONVOLUTIONAL NEURAL ACTION RECOGNITION

An approach for action recognition in single image using body parts was developed in [10]. In [4], a path signature-based approach has been proposed for action recognition using body skeleton. A raw skeleton coordinates and skeleton motion based action recognition technique is discussed in [11]. An end-to-end convolutional co-occurrence



(a) Does all the joints are important for "drinking action"?

(b) Does all the joints are important for "phoning action"?

FIGURE 3. Illustration of attention-joints for action predication on Stanford 40-Actions dataset images (better perceived in color).

methodology was investigated in [12]. [13] incorporated multiple modalities (e.g. RGB, depth, Flow and IR) in convolutional neural network for action recognition.

B. GRAPH NEURAL ACTION RECOGNITION

Over the past few years, convolutional neural networks have been generalized from Euclidean domain (image or audio) to non-Euclidean domain as graphs, [5], [11], [14]. The preliminary work on graph convolution network are found in [9]. This seminal work was extended in [15] by introducing gated recurrent unit in graph convolutional neural networks. An edge-conditioned convolution method was proposed in [16], where the convolutional filters were conditioned on edge labels. In [8], the authors set forth graph CNN to solve the problem of semi-supervised learning. Applying convolutional neural networks on graphs has two perspective, i) spectral perspective, the convolutional filters and pooling operations are applied in spectral domain [14], [17], ii) spatial perspective, that the convolutional filters are applied directly on graph nodes and the neighbors, [5]. To apply convolution network in spectral domain, a spectral convolution layer is devised as in [5]. In [16], authors generalize the convolution filters guided by Euclidean grids from arbitrary graphs with varying nodes and edges. Reference [18] establishes a depth-wise separable graph convolution that surpass the performance over other graph convolution and geometric convolution networks. Reference [19] presents a mathematical formulation for action recognition in skeleton-based spatio-temporal graph convolution.

Empirically, it is found that there is complementarity in nodes and edges to model skeleton-based action recognition using GCN. Reference [21] discusses graph node convolution and graph edge convolution to model the complementarity using shared intermediate layers. In [22], the authors present a skeleton-based action recognition using a graph regression GCN to model the spatio-temporal variations in data. Reference [23] introduced a novel representation of skeleton data as a directed acyclic graph (DAG) based on the kinematic dependency between joints and bones of human body. A recent detailed survey on graph convolution network for different applications is presented in [24].

III. PROPOSED METHODOLOGY

The design of our proposed framework for action recognition is shown in Figure 4. Our framework consists of two main parts, i) attention network and ii) graph convolutional network.

A. ATTENTION NETWORK

We investigate residual attention network to extract attention-joints from the human body. The main motivation for residual attention network is stacking a large number of attention modules in a residual manner. The attention network takes raw RGB-images as input and generates the attention masks. These attention-masks are element-wise multiplied with the skeleton images to identify attention joints. Mathematically, element-wise multiplication of attention-mask with the input image is defined as follows,

$$X_{i,c} = X_{i,c} * M_{i,c} \tag{1}$$

where i denotes the spatial index and c denotes the channel index of a pixel in the masks. Thus, residual-attention network emphasizes the most important regions in an input RGB-image and suppresses less important regions of that image, pertaining to some action. A major advantage of residual-attention network is a large reduction in network parameters compared to residual network parameters, [20].

In residual attention network, each block consists of two branches, i) trunk branch and ii) mask branch, where the trunk branch is designed to learn target-oriented features and can be implemented by any existing CNN architecture. In this paper, we used VGG-16 to implement the trunk branch of the residual attention network. Different from the trunk branch, the mask branch is implemented in the bottom-up and top-down structure to learn the attention mask $M_i(x)$. The mask branch is the main contributor to superior performance of residual attention network for action recognition, which works as a feature selector to enhance the most informative part and suppress redundant part of the features obtained from the trunk branch. The architecture of residual-attention network is shown in Figure 5.

In residual-attention network, trunk branch features adaptively change mask branch attention. Feature map from each channel is normalized using spatial attention and then sigmoid operation is performed to obtain soft-mask related to spatial information. In this paper, we use spatial attention mathematically described as

$$f(x_{i,c}) = 1/(1 + exp(-(x_{i,c} - \mu_c)/\sigma_c)) \tag{2}$$

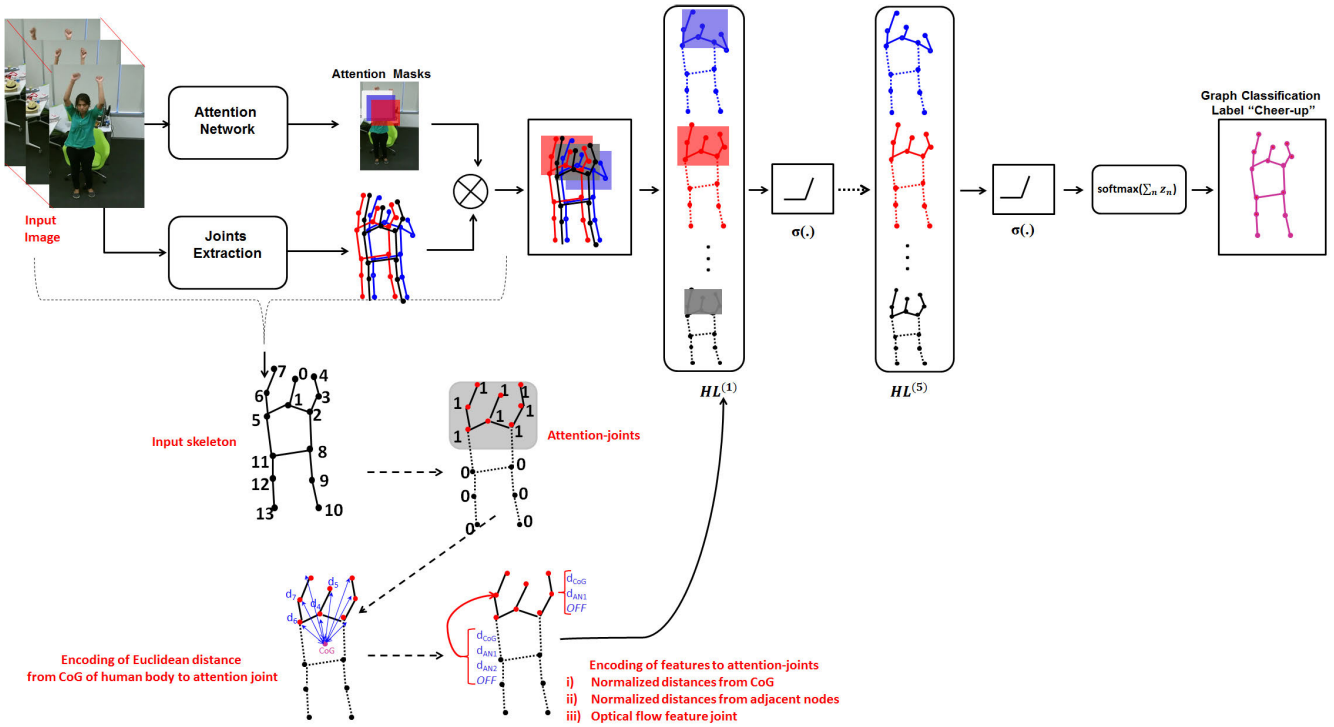


FIGURE 4. Block diagram of our proposed attention-joints graph convolutional neural network. Attention network is utilized to extract attention-joints of input skeleton. Then only the features associated with attention-joints are fed into graph convolutional network for classification. $HL^{(1)}$ to $HL^{(5)}$ are the hidden layers from 1 to 5, performing convolutional+pooling operations.

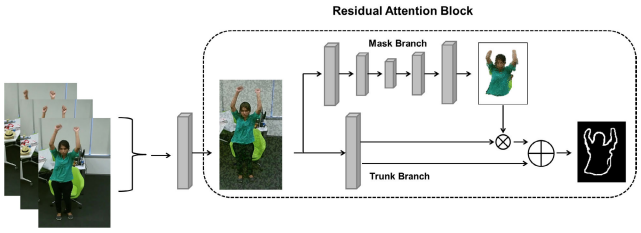


FIGURE 5. Residual-attention block.

where i and c represent spatial positions and channels respectively. Also, μ_c and σ_c corresponds to the mean and standard deviation of feature-map for the c-th channel and x_i denotes the feature vector at the i-th spatial position.

B. SKELETON GRAPH FORMULATION

The human skeleton and its associated joints can be represented in 2D or 3D coordinates in each frame. In retrospective study, human skeletons are represented as a single feature vector [19] or as a spatial-temporal graph [25]. The joints within one frame are connected via edges according to the connectivity of human body structure to give an undirected graph [25]. The terms nodes and joints are interchangeable in our work. We formulate a graph $G = (N, E)$ for a skeleton where N denotes a set of graph nodes $N = \{n_1, n_2, \ldots, n_k\}$ and E denotes the set of edges between nodes, defined as ordered pairs, $E = \{(n_1, n_2), (n_2, n_3), \ldots, (n_{k-1}, n_k)\}$. Mathematically, the overall graph can be represented as, $G = \{n_1, n_2, \ldots, n_k | (n_1, n_2), (n_2, n_3), \ldots, (n_{k-1}, n_k)\}$.

C. ATTENTION JOINTS ENCODING

To keep a simple architecture, only 14-joints of the human body are considered. The feature encoding for attention joints is illustrated in Figure 4. As shown in this figure, the body joints of the input skeleton are enumerated from 0 to 13. The attention-joints extracted by residual-attention network are labeled as "1" on the graph node, whereas other joints are labeled as "0".

As discussed in Section I that the node labeling procedure include three types of features, i) weighted distances of attention-nodes from the body-center, d_{CoG} ii) distances between neighboring attention-nodes, d_{AN} and iii) flow features of each attention-joint, namely OFF. The first kind of features, d_{CoG}, is defined as the Euclidean distances from body vertex or center-of-gravity (CoG) of human body to the attention-joints. These distances, d_{CoG}, are the weighted distances, as some joints of the human body are well-articulated and contribute more to the final action prediction. For example, as shown in Fig 4, the joints 3 and 4 are more agile to move as compared to joint 2, therefore are weighted more than joint 2. Likewise joints 6 and 7 are well-articulated than joint 5. The same observations can also be made for the other pairs of joint (9, 10) and (12, 13), as they are more dexterous compared to joints 8 and 11 respectively.

In our work, the second features associated with attention-joints are the normalized Euclidean distances of an attention-joint from neighboring-joints, denoted as $d_{AN1}...d_{ANn}$. If the attention-node is connected to one node only, it has just one distance, d_{AN1}, if connected to two

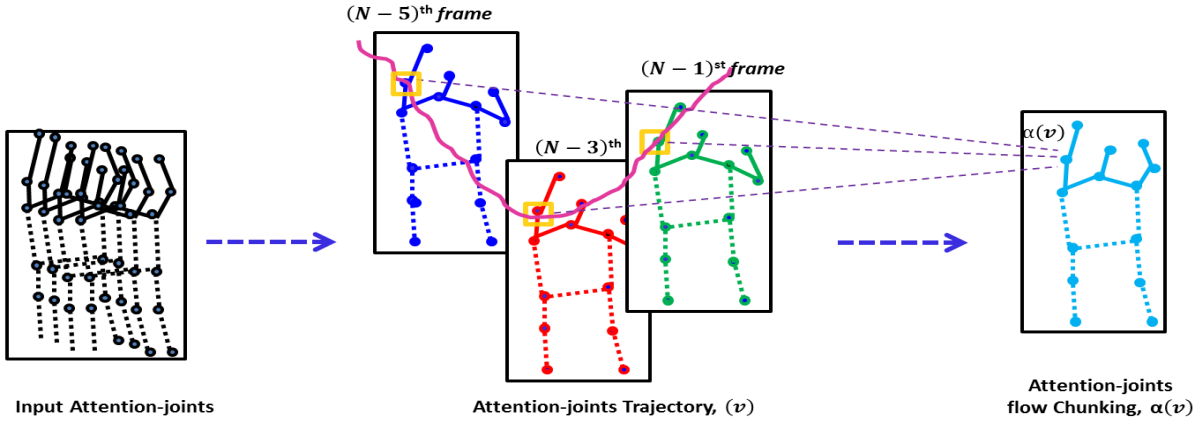


FIGURE 6. Aggregation of attention-joints flow over frames $N-1$, $N-3$ to $N-5$.

attention-joints, it has d_{AN1} and d_{AN2} and so on. This concept of adjacent node distances, d_{AN}, is shown in Figure 4, where node "4" has only one adjacent neighbor and one corresponding adjacent distance, d_{AN1}, while node "3" has two adjacent neighbors and hence two corresponding adjacent distances, d_{AN1}, and d_{AN2}.

The last feature associated with an attention-node is the flow features of attention-joints (OFF), where the joints optical flow contain the temporal information of attention-joints over a sequence of frames. We compute three levels of flow features for joints including: i) the joints flow between two consecutive frames, N and $N-1$; ii) joints-flow between current frame and third-last frame, N and $N-3$, and iii) joints-flow between current frame and fifth-last frame, N and $N-5$. This concept is depicted in Figure 6. The two reasons for computing these three-levels of joints-flow between frames, namely, the first that joint-flow between consecutive frames may have spurious motion and resulting in noisy joint-flow and the most of the actions may prolong over 3-5 frames, so the joints-flow over 5-frames can better model such temporal relationship between frames.

The encoding of above-mentioned features is shown in Figure 4, where each attention-node feature vector contains d_{CoG}, d_{AN} and OFF features. The feature vectors are encoded only for attention-nodes. The skeleton of an input frame contains information of all nodes and edges, where the outputs of attention-nodes are labeled as "1" and node features are associated only to attention-nodes. The final output of a graph is labeled as a class-label during training, like a supervised learning fashion.

D. IMPLEMENTATION OF GRAPH CNN

The implementation of geometric graph-based convolutional neural network is challenging compared to Euclidean 2D convolutional neural network. To implement graph convolution on human skeleton, the joints within a single frame are represented as an adjacency matrix A and self-connections are represented as identity matrix, I. The graph convolution propagation rule from a layer $H^{(l)}$ to $H^{(l+1)}$ is defined as,

$$H^{(l+1)} = \sigma(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^{(l)}W^{(l)}) \tag{3}$$

where $\tilde{A}$ is the adjacency matrix, $\tilde{D}$ is the degree matrix of $\tilde{A}$. The factor $\tilde{D}^{-0.5}\tilde{A}\tilde{D}^{-0.5}$ is used to normalize nodes with large degree and called normalized adjacency matrix. $H^{(l)}$ is the feature matrix at previous layer, and $W^{(l)}$ is the weight matrix at previous layer. The activation function σ is introduced as a non-linear ReLU function, $max[0, x]$.

An important property of convolutional neural network is pooling, which is implemented as graph down-sampling or coarsening in graph neural networks. The graph pooling is implemented in such a way that similar node features are sub-sampled and pooled together to create global invariance with multiple layers. Non-linear multi-scale coarsening, graph partitioning, and heavy edge matching (HEM) are the commonly used graph pooling strategies. The graph pooling operation is explained in Figure 2.

IV. EXPERIMENTS

A. DATASET

1) STANFORD 40-ACTIONS

The Stanford 40-Actions dataset [26] contains 9,532 images of 40 different human actions of diverse categories, varying from brushing teeth, fishing, fixing a car, holding an umbrella. Each action class contains about 180-300 images, obtained from Google, Bing and Flicker databases. In the dataset, there is significant within-class variance for each action due to varying body pose, appearance and background clutter.

2) PKU-MMD

PKU-MMD [27] is a large scale 3D human action dataset for action recognition and understanding. The dataset contains depth maps, skeleton joints, infrared sequences and RGB videos. There are 20,000 action instances in the dataset, performed by 66 different subjects in three camera views. The dataset contains 51 action categories, with cross-subject

(CS) and cross-view (CV) settings, where we evaluated our methodology only for cross-subject setting. For cross-subject setting, the dataset is provided with a data split of 57-subjects for training and 9-subjects for testing. The training set has total of 944 videos while testing set has 132 video samples.

3) NTU-RGBD

NTU-RGBD [28] is a large-scale dataset that contains 56,000 action clips from 60 different action classes. To capture each action, three cameras are mounted at same height but with three different angles: -45, 0, +45. The dataset contains the joints location, detected by Kinect depth sensor. Each frame contains 25-joints for each subject; however, for this research we used only 14 significant joints on the human body, as presented in Figure 4. The cross-view and cross-subject benchmark are provided for NTU-RGBD dataset where cross-subject setting includes 39,889 training clips and 16,390 testing clips.

B. IMPLEMENTATION DETAILS

Our proposed architecture contains five consecutive stack of convolution-pooling layers, followed by a SoftMax layer. The spatial kernel size for first three layers is fixed as (5, 5), while for last two layers is kept as (3, 3) with a stride of one and padding of two. For all pooling layers, global average pooling (gap) is used with pooling ratio of 0.5. The first 3-convolutional layers have 32 output channels while last two convolutional layers have 64-output channels. During training, the batch_size for our experiment is set to 64, with the initial learning_rate of 0.005, which was subsequently reduced down by a factor of 1/10 after every 1/3 of iterations. The gamma was fixed at 0.99 and weight-decay was set to 0.00001. Drop-out with a probability of 0.5 was used to avoid over-fitting. Stochastic gradient descent (SGD) was selected as an optimizer, with momentum of 0.9. The cross-entropy loss was used as the loss function to back-propagate the gradients. We implemented our model in Pytorch-geometric [30] with CUDA 9.0. The experiments were run for a maximum of 200 epochs, using two NVIDIA TITAN X GPU with 24-GB RAM.

For extracting graph and input to graph convolutional network, we followed the procedure of pytorch-geometric dataset creation, where it is provided with graph-adjacency, graph-indicator, graph-labels and node-labels for experimentation.

C. BASELINE MODEL

We define our baseline model as having graph convolutional-pooling layers without attention network. Our baseline architecture is excited with 14-body joints skeleton and with all three modalities, d_{CoG}, d_{AN} and OFF. The performance of baseline architecture is enlisted in Table 1. Moreover, our baseline architecture uses the same hyper-parameters setting as that of attention-joints architecture.

TABLE 1. Performance of baseline architecture.

Dataset	mean Average Precision (mAP) in (%)	
	$d_{CoG} + d_{AN}$	$d_{CoG} + d_{AN} + OFF$
Stanford 40 Actions	75.4	-
PKU-MDD	-	86.1
NTU-RGBD	-	82.8

TABLE 2. Performance evaluation for stanford 40-actions dataset.

Dataset	mean Average Precision (mAP) in (%)		
	d_{CoG}	d_{AN}	$d_{CoG} + d_{AN}$
Stanford 40 Actions	78.2	69.5	**84.6**

D. EXPERIMENTAL RESULTS ON SINGLE IMAGE DATASET

Our proposed method is examined using single image Stanford 40-Action recognition dataset. For each image, human skeleton is extracted from the RGB-image using Deepercut method in [29]. Residual attention network is applied on RGB-images to extract attention regions on the input images. The attention region is mapped on the skeleton image to extract attention joints. For single image action recognition, it is challenging to define node features as there are no temporal details available. Therefore, we used normalized spatial distances, d_{CoG} and d_{AN}, to assign the node features. We need to drop the joint-flow features for attention-joints as this dataset is for only single image action recognition.

The validation of results for Stanford 40-Actions dataset have been enlisted in Table 2, where the first observation is made by considering only normalized euclidean distances from attention-joints, d_{CoG}. Then, second investigations incorporate normalized Euclidean distances of neighboring attention-joints, d_{AN}. The main motivation for investigating d_{CoG} and d_{AN} is that such normalized Euclidean distances are keep changing while performing different actions, therefore it is important to consider them for action recognition.

An interesting observation is that normalized euclidean distances, d_{CoG}, performs better than distances between adjacent nodes, d_{AN}, because d_{CoG} contains more significant details pertaining to the articulation of attention-joints. Subsequently, d_{CoG} and d_{AN}, are combined as node features and excited to the graph neural network. It turns out that d_{CoG} and d_{AN} contains complementary details and when excited together to graph convolutional neural network, raises the overall performance.

E. EXPERIMENTAL RESULTS ON TEMPORAL SKELETON DATASETS

We evaluated our proposed framework on two temporal skeleton-based action recognition datasets, PKU-MDD and NTU-RGBD. Both datasets contain spatial and temporal information and for each action video, RGB and skeleton information is provided. The joints of skeleton are already provided in these datasets, so we only need to identify the attention joints. Attention joints are identified by applying residual-attention network on RGB-images, using the method

TABLE 3. Performance evaluation for skeleton datasets.

Dataset	mean Average Precision (mAP) in (%)					
	d_{CoG}	d_{AN}	OFF	$d_{CoG} + OFF$	$d_{AN} + OFF$	$d_{CoG} + d_{AN} + OFF$
PKU-MDD	82.5	74.4	79.9	93.1	84.0	**95.5**
NTU-RGBD	80.9	78.1	79.3	87.9	83.3	**90.7**

TABLE 4. Performance comparison with contemporary methods.

Dataset	mean Average Precision (mAP) in (%)		
	Stanford 40 Action	PKU-MDD	NTU-RGBD
Attributes-Parts based [26]	65.1	-	-
Minimum-annotation effort [31]	82.6	-	-
Body-parts based [10]	83.4	-	-
Multi-modalities (RGB+depth+skeleton+IR) [27]	-	64.9	-
Raw-skeleton coordinate+motion [11]	-	92.2	83.3
Convolutional Co-occurrence [12]	-	92.6	86.5
TSN on RGB+depth+IR [13]	-	94.6	-
Spatial-temporal GNN [25]	-	-	81.5
Two-stream GNN [32]	-	-	85.4
Regression-based GNN [22]	-	-	87.5
Directed-edge GNN [23]	-	-	89.9
Our proposed	**84.8**	**95.5**	**90.7**

adopted in single image action recognition. After identifying attention-joints, each attention-joint is specified with node features.

The node features, d_{CoG} and d_{AN}, for PKU-MDD and NTU-RGBD are also defined in the similar fashion as that for Stanford 40-Action datasets. However, for temporal skeleton-based action recognition datasets, joints-flow between consecutive frames is introduced as an other important modality. The joints-flow between attention-joints is computed at three levels of frames ($N-1$, $N-3$, and $N-5$) in order to mitigate any spurious motion.

In Table 3, the performance of both PKU-MDD and NTU-RGBD datasets has been first evaluated for each individual modality, d_{CoG}, d_{AN} and OFF. It is noted that for both datasets d_{CoG} significantly performs better than other two modalities due to the reason that it, inherently, contains the articulation of attention-joints. Empirically, it turns out that the articulation of attention-joints (distances from body-center to attention-joint) contains very important clues for recognizing actions. Then it is revealed that joints-flow features OFF between frames perform better than the distances from neighboring joints, d_{AN}, signifying the fact that motion is also another important clue for action recognition. A further study also carried out when different modalities are unified as node features, such as $d_{CoG}+OFF$ and $d_{AN}+OFF$, and excited to graph CNN. It is published that unification of aforesaid three modalities exhibit their best when fused together. Empirically, it is noticed that the performance margin using $(d_{CoG} + OFF)$ and $(d_{CoG} + d_{AN} + OFF)$ is small, corroborating to the argument that $d_{CoG} + OFF$ contains significant clues as compared to combining d_{AN} with other modalities for skeleton based action recognition.

F. ABLATION STUDY AND COMPARISON WITH THE STATE-OF-THE-ARTS

In convolutional networks, attention mechanism is used to emphasize the most contributing features and this concept is exploited by introducing attention-joints in our work. In our study, we investigate the contributions of attention-joints for action recognition using graph convolutional networks. We conduct experiments by removing all attention network and leading to a baseline network with only graph convolution-pooling layers. Comparing the results in Table 1 with Table 2 and 3, it reveals that attention-joint architecture performs better than baseline architecture. This empirical improvement in performance demonstrates the contribution of attention-joints for action recognition.

We compared our method with other state-of-the-art methods on three different action datasets. The results are given in Table 4. We used mean-Average Precision (mAP) as an evaluation metric for train-test splits of three datasets. mAP is calculated by using the following formula, $1/N * \sum_{i=1}^{N} AP_i$, where AP is corresponding average precision for each split of the dataset.

1) STANFORD 40-ACTIONS

In [26], authors jointly modeled the attributes and parts by using a sparse bases that entails the meaningful semantic information for action recognition and marked the performance up to 65.1%. In [31], a unique technique was devised by using minimum annotation efforts for action recognition, which greatly surpassed the performance up to 82.6%. Human body parts contain important clues of action recognition which were studied in [10], wherein the body-parts-based single image action recognition improved the performance up to 83.4%. The above-mentioned approaches for action recognition are based on convolutional neural network, while we first time addressed single image action recognition problem using attention-joints based graph CNN, and resulted in state-of-the-art performance of 84.8%.

2) PKU-MMD

The authors in [27] assembled and formulated PKU-MMD dataset and its performance was computed using multiple

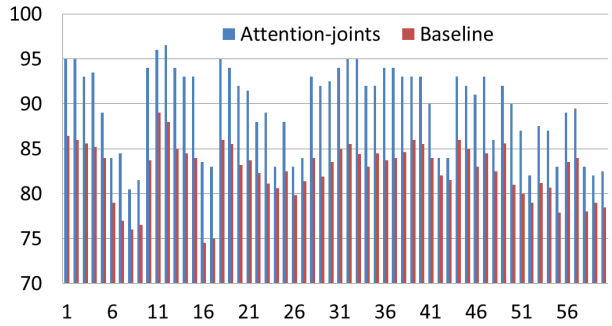


FIGURE 7. Comparison of performance of attention-joints architecture with baseline architecture for NTU-RGBD dataset.

modalities, such as RGB, depth, skeleton and IR-modality. Reference [11] proposed raw skeleton coordinate and skeleton motion for action recognition and prediction, where we included attention mechanism along with GNN in our work. The authors exhibits their results on PKU-MMD and NTU-RGBD datasets. Compared with an end-to-end convolutional co-occurrence feature learning proposed in [12], our method performs better by 2.9%. Reference [13] extended the temporal segment network using different modalities of RGB, depth and infra-red data, in order to investigate human action recognition. However, pre-training on UCF-101 dataset was required for this method, while no such pre-training is needed in our case and it still surpass the performance by 0.9%. All comparisons have been made for cross-subject settings.

3) NTU-RGBD

Reference [25] embedded spatial and temporal patterns of data in the graph using spatial-temporal graph convolutional network. Our proposed technique differ from this method in a sense that we accumulated the temporal details using joints-flow as attention-node features, rather than incorporating the sequential information by using complex temporal edges in graph. Reference [32] presents a two stream graph edge convolutional and node convolutional for skeleton based action recognition. Likewise [25], the sequential information is embedded in graph by using temporal graphs in [32]. Spatial and temporal details are fused in [22] by using graph regression-based convolutional neural network, where our framework surpass than this method by 3.2%. Reference [23] involve the directed graph structure for skeleton-based action recognition, where spatial and temporal informations are fused together by using two streams. Our approach just uses simple undirected edges and considerably improves the overall performance. In figure 7, a performance comparison of attention-joints with baseline architecture for NTU-RGBD dataset has been illustrated. The horizontal axis contains action IDs while vertical axis depicts performance in mAP.

V. CONCLUSION

In this paper, we present a novel idea of action recognition in skeleton images using attention-joints and graph convolutional neural network. First, we devise the spatial features of graph nodes as Euclidean distances and then introduce the temporal signatures of video sequence as flow features of attention-joints. In our proposed framework, the attention-joints are equipped with spatial-temporal features and excited as attention-nodes to graph neural network for action classification. The proposed methodology suppresses the noisy and spurious details incurred due to considering all graph nodes and edges. Extensive experiments show that our method is effective and has achieved state-of-the-art performance. In future work, LSTM and 3D graph convolutional networks will be investigated and exercised to better model spatial-temporal attributes for skeleton-based action recognition.

ACKNOWLEDGMENT

It is especially acknowledged to Prof. L. Jin for his tremendous help and support during this course of research. His scholastic guidance has also been endured while preparing this draft.

REFERENCES

[1] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 568–576.

[2] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, and D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 20–36.

[3] J. Liu, A. Shahroudy, M. L. Perez, G. Wang, L.-Y. Duan, and A. K. Chichung, "NNTU RGB+ D 120: A large-scale benchmark for 3D human activity understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.

[4] W. Yang, T. Lyons, H. Ni, C. Schmid, and L. Jin, "Developing the path signature methodology and its application to landmark-based human action recognition," 2017, *arXiv:1707.03993*. [Online]. Available: https://arxiv.org/abs/1707.03993

[5] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral networks and locally connected networks on graphs," 2013, *arXiv:1312.6203*. [Online]. Available: https://arxiv.org/abs/1312.6203

[6] Z. M. Chen, X. S. Wei, P. Wang, and Y. Guo, "Multilabel image recognition with graph convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 5177–5186.

[7] H. Martin, D. Bucher, E. Suel, P. Zhao, F. Perez-Cruz, and M. Raubal, "Graph convolutional neural networks for human activity purpose imputation," in *Proc. NIPS Spatiotemporal Workshop 32nd Annu. Conf. Neural Inf. Process. Syst. (NIPS)*, 2018.

[8] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*. [Online]. Available: https://arxiv.org/abs/1609.02907

[9] N. Shervashidze, S. Vishwanathan, T. Petri, K. Mehlhorn, and K. Borgwardt, "Efficient graphlet kernels for large graph comparison," in *Proc. Artif. Intell. Statist.*, 2009, pp. 488–495.

[10] Z. Zhao, H. Ma, and S. You, "Single image action recognition using semantic body part actions," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 3391–3399.

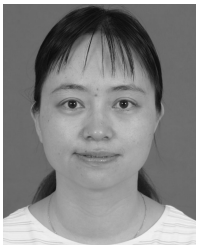
[11] C. Li, Q. Zhong, D. Xie, and S. Pu, "Skeleton-based action recognition with convolutional neural networks," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2017, pp. 597–600.

[12] C. Li, Q. Zhong, D. Xie, and S. Pu, "Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation," 2018, *arXiv:1804.06055*. [Online]. Available: https://arxiv.org/abs/1804.06055

[13] S. Ardianto and H. M. Hang, "Multi-view and multimodal action recognition with learned fusion," in *Proc. IEEE Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA-ASC)*, Nov. 2018, pp. 1601–1604.

[14] M. Henaff, J. Bruna, and Y. LeCun, "Deep convolutional networks on graph-structured data," 2015, *arXiv:1506.05163*. [Online]. Available: https://arxiv.org/abs/1506.05163
[15] Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel, "Gated graph sequence neural networks," 2015, *arXiv:1511.05493*. [Online]. Available: https://arxiv.org/abs/1511.05493
[16] M. Simonovsky and N. Komodakis, "Dynamic edge-conditioned filters in convolutional neural networks on graphs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3693–3702.
[17] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, "Convolutional networks on graphs for learning molecular fingerprints," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2224–2232.
[18] G. Lai, H. Liu, and Y. Yang, "Learning graph convolution filters from data manifold," 2018.
[19] C. Li, Z. Cui, W. Zheng, C. Xu, and J. Yang, "Spatio-temporal graph convolution for skeleton based action recognition," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018.
[20] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3156–3164.
[21] X. Zhang, C. Xu, X. Tian, and D. Tao, "Graph edge convolutional neural networks for skeleton-based action recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published.
[22] G. Xiang, H. Wei, T. Jiaxiang, L. Jiaying, and G. Zongming, "Optimized skeleton-based action recognition via sparsified graph regression," in *Proc. Proc. ACM Multimedia (ACM MM)*, 2019.
[23] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with directed graph neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 7912–7921.
[24] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," 2019, *arXiv:1901.00596*. [Online]. Available: https://arxiv.org/abs/1901.00596
[25] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018.
[26] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei, "Human action recognition by learning bases of action attributes and parts," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1331–1338.
[27] C. Liu, Y. Hu, Y. Li, S. Song, and J. Liu, "PKU-MMD: A large scale benchmark for continuous multi-modal human action understanding," 2017, *arXiv:1703.07475*. [Online]. Available: https://arxiv.org/abs/1703.07475
[28] A. Shahroudy, J. Liu, T. T. Ng, and G. Wang, "NTU RGB+ D: A large scale dataset for 3D human activity analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1010–1019.
[29] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "DeeperCut: A deeper, stronger, and faster multi-person pose estimation model," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 34–50.
[30] M. Fey and J. E. Lenssen, "Fast graph representation learning with PyTorch geometric," 2019, *arXiv:1903.02428*. [Online]. Available: https://arxiv.org/abs/1903.02428
[31] Y. Zhang, L. Cheng, J. Wu, J. Cai, M. N. Do, and J. Lu, "Action recognition in still images with minimum annotation efforts," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5479–5490, Nov. 2016.
[32] X. Zhang, C. Xu, X. Tian, and D. Tao, "Graph edge convolutional neural networks for skeleton based action recognition," 2018, *arXiv:1805.06184*. [Online]. Available: https://arxiv.org/abs/1805.06184

TASWEER AHMAD received the bachelor's degree in electrical engineering from the University of Engineering and Technology, Taxila, Pakistan, in 2007, and the master's degree from the University of Engineering and Technology, Lahore, Pakistan, in 2009. He is currently pursuing the Ph.D. degree with the South China University of Technology, China. He was an Instructor with the Government College University, Lahore, from 2010 to 2015, and with the COMSATS Institute of Information Technology, Sahiwal Campus, Pakistan, from 2015 to 2016. His current research interests include image processing, computer vision, and machine learning.



HUIYUN MAO received the Ph.D. degree from the South China University of Technology, Guangzhou, China, in 2011. She is currently a Lecturer with the College of Computer Science, South China University of Technology. She has authored over 20 scientific articles and patents. Her research interests include image processing, machine learning, and intelligent systems.

LUOJUN LIN received the B.S. degree in electronic and information engineering from Yunnan University, Kunming, China, in 2014. She is currently pursuing the Ph.D. degree with the School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China. Her research interests include deep learning and computer vision, especially facial attractiveness analysis.

GUOZHI TANG received the bachelor's degree in information technology from Yunnan University, China, in 2019. His research areas are deep learning and car license plate detection.