Multi-hop Modulated Graph Convolutional Networks for 3D Human Pose Estimation

Jae Yung Lee
jaeyung.lee@kt.com
I Gil Kim
igil.kim@kt.com

AI2XL, Institute of Convergence Technology
Korea Telecom
Seoul, South Korea

Abstract

Graph convolutional networks (GCNs) have recently been applied to 3D human pose estimation (3D HPE) from 2D body joints. GCN-based 3D HPE has achieved promising performance in modeling the relationships between body parts. However, vanilla graph convolution only considers the relationship between neighbouring joints at a one-hop distance. Some recent approaches have utilised high-order graph convolution to model long-range dependency. They exploit the adjacency matrix of one-hop neighbouring joints, but the method cannot capture the long-range dependency. To solve this problem, we propose the multi-hop modulated GCN (MM-GCN) for 3D HPE. The unique adjacency matrix of each hop distance is derived, and aggregate features of nodes at various hop distances are modulated to capture the long-range dependency. Thus, the proposed network can model a wide range of interactions between body joints more adequately than does the vanilla graph approach. Moreover, we investigate the impact of combination with affinity modulation (AM) because AM adjusts the graph in a GCN. Our experiments, and an ablation study conducted on two standard benchmarks demonstrate the effectiveness of the proposed network, showing that our MM-GCN outperforms some recent state-of-the-art techniques.

1 Introduction

3D human pose estimation (3D HPE), which aims to localise the 3D position of human body joints in a camera coordinate system from single or multiple images, is an important area of study in human–computer interaction (HCI), gesture recognition, and human behaviour recognition. However, 3D HPE remains challenging due to its ill-posed nature, which means multiple valid 3D body configurations can be projected onto the same 2D pose in the image space. Thus far, the performance of 3D HPE has been greatly improved by the rapid development of deep neural network solutions, which have achieved performance superior to that of classical approaches which use handcrafted features. Most existing 3D pose estimation methods use an end-to-end pipeline [14, 20, 25, 26, 30] via a convolutional neural network (CNN) from an image, or a two-stage pipeline [4, 6, 16, 22, 31]. Two-stage approaches for 3D pose estimation have shown great promise, outperforming end-to-end models. For

example, Martinez *et al*. [16] designed a simple fully connected network with residual connections for estimating 3D poses from 2D joint detections, and they achieved state-of-the-art 3D HPE performance.

In recent years, interest in the adoption of graph convolutional networks (GCNs) [3, 9, 29] for 3D pose estimation [9, 23, 23, 31, 35, 36] has rapidly increased. Because a 2D human skeleton can be represented as a graph, in a GCN for 3D HPE, the nodes in the graph are body joints, and the edges in the graph are connections between neighbouring body joints. Moreover, GCNs repeatedly transform and aggregate the features of neighbouring nodes to obtain more powerful feature representations. Zhao *et al*. [31] proposed SemGCN, which learns to capture semantic information by multiplying a learnable mask by a given graph, yielding improved performance in 3D pose estimation while using a much smaller number of parameters. In addition, Zou *et al*. [35] proposed a modulated GCN that adjusts the graph structure in a GCN so that it can model additional edges beyond the human skeleton. However, GCNs designed for 3D HPE have a potential limitation; that is, they perform graph convolutions only on the one-hop neighbours of each node, and hence they lack the ability to capture long-range dependencies. Some works [1, 35, 36] have attempted to alleviate this limitation. For example, MixHop was proposed in [1] by concatenating the feature representations of multi-hop neighbours via a sparsified neighbourhood mixing, which leverages a graph convolutional layer that mixes the powers of the adjacency matrix. Zou *et al*. [36] proposed a high-order GCN for 3D pose estimation based on MixHop that fused the features of these multi-hop neighbouring joints, and Quan *et al*. [35] leveraged residual connections to help mitigate the over-smoothing problem.

To address this limitation, this study introduces a multi-hop modulated GCN (MM-GCN) for 3D HPE. First, we derive the unique adjacency matrix of each hop distance, whereas in previous works, graphs of k-hop neighbouring nodes were computed by the kth power of the adjacency matrix of the one-hop distance. For the aggregation method, the features of these multi-hop neighbours are modulated and fused by a learnable matrix to consider the long-range dependency, which is designed such that the features are more heavily weighted as the hop distance becomes smaller. The effectiveness of our approach is validated by a comprehensive evaluation with a rigorous ablation study and comparisons with state-of-the-art techniques on standard 3D benchmarks. The performance of our approach matches that of state-of-the-art techniques on Human3.6M [12] and MPI-INF-3DHP [17] using only 2D joint coordinates as inputs. Furthermore, we show the visual results of MM-GCN, which demonstrate the effectiveness of our approach qualitatively.

2 Related Work

3D Human Pose Estomation Recently, state-of-the-art 3D HPE approaches have taken advantage of deep neural networks, and they can be roughly divided into two categories. The first category of approaches directly predicts the 3D pose from the image [17, 19, 20, 26, 33, 34]. For example, Zhou *et al*. [34] integrated a 3D depth regression subnetwork into a state-of-the-art 2D detector. Pavlakos *et al*. [20] proposed a fine discretisation of the 3D space around the subject and trained a CNN to predict the per-voxel likelihood for each body joint. Sun *et al*. [26] designed a simple integral operation to relate and unify a heatmap representation and joint regression.

The second category includes approaches that consist of two stages [4, 6, 7, 8, 11, 15, 16, 21, 22, 24], 2D joint locations are first extracted using a 2D pose detector, and then a

lifting network is employed to regress 3D poses from 2D detections. Our approach belongs in this category. Martinez *et al*. [16] introduced a simple yet effective method that predicted 3D keypoints purely based on 2D detections. Fang *et al*. [7] further extended this approach through pose grammar networks. Zhou *et al*. [33] directly embedded a kinematic object model into a deep neural network for general articulated object pose estimation. Ci *et al*. [6] enhanced the representation capability of a GCN by introducing a locally connected network. Zou *et al*. [36] designed a high-order GCN model for 3D pose estimation based on MixHop in a bid to capture the long-range dependencies between distant body joints. By contrast, our MM-GCN derives the adjacency matrix of body joints at various distances, which provides a more powerful representation of the GCN by modeling long-range dependencies between body joints. We also introduce the method of aggregating features.

Graph Convolutional Networks Generalising CNNs to inputs with graph-like structures is an important topic in the field of deep learning. The principle of constructing GCNs on graphs is generally based on one of two perspectives: the spectral perspective and the spatial perspective. Our proposed MM-GCN is most related to spatial GCN, because the convolution filters of our MM-GCN are applied directly to the graph nodes and their neighbours. Sami *et al*. [1] learned neighbourhood mixing relationships by repeatedly mixing feature representations of neighbours at various distances through the powers of the graph adjacency matrix. Bai *et al*. [2] exploited a high-order GCN for skeleton-based action recognition, but their high-order adjacency matrix was constructed by summing the mixed powers of the original adjacency matrix. Zou *et al*. [36] introduced the aggregation method to fuse the multi-order feature representations.

3 Muti-hop Modulated Graph Convolutional Networks

3.1 Previous GCN for 3D HPE

A vanilla GCN for 3D HPE was introduced in [29]. By letting $G = fV, Eg$ denote a graph where V is the set of N nodes and E are the edges which connect two body joints, the edges can be represented by an adjacency matrix $\mathbf{A} \; 2 \; f0,1g^{\mathrm{NxN}}$ whose *(i,j)*-th entry is equal to the weight of the edge between neighbouring nodes i and j, and zero otherwise. Each node i is associated with a D-dimensional feature vector $\mathbf{h}_i \; 2 \; \mathrm{R}^{\mathrm{D}}$. The collection of features of all nodes can be written as matrix $\mathbf{H} \; 2 \; \mathrm{R}^{\mathrm{DxN}}$. A graph convolution layer updates the features of each node via the equation below:

$$\mathbf{H}^{\emptyset} = s\left(\mathbf{WH}\tilde{\mathbf{A}}\right) \tag{1}$$

where $\tilde{\mathbf{A}}$ is the symmetrically normalized version of $\mathbf{A}$ with self-connections [29], $\mathbf{W} \; 2 \; \mathrm{R}^{\mathrm{D}^{\emptyset}\mathrm{xD}}$ is a learnable weight matrix that transforms the feature dimension from D to $D^{\emptyset}$, followed by an activation function $s(\;)$ such as $\mathrm{RELU}(\;) = \max(0, \;)$ [18], and $\mathbf{H}^{\emptyset} \; 2 \; \mathrm{R}^{\mathrm{D}^{\emptyset}\mathrm{xN}}$ is the updated feature matrix, where the i-th column of $\mathbf{H}^{\emptyset}$ is $\mathbf{h}_i^{\emptyset}$. Eq. (1) can be equivalently written as below:

$$\mathbf{h}_i^{\emptyset} = s\left(\sum_{j \, 2 \, \tilde{N}(i)} \mathbf{W}\mathbf{h}_j \tilde{a}_{ij}\right) \tag{2}$$

where $\tilde{a}_{ij}$ is the *(i,j)*-th entry of $\tilde{\mathbf{A}}$, and these transformed node representations are gathered to node i from its neighbouring nodes $j \; 2 \; \tilde{N}(i)$ and $\tilde{N}(i) \quad N(i) \; [\; fig$. This vanilla graph

convolution has one limitation (it shares a weight matrix $\mathbf{W}$ for each node), and this limitation makes various relational patterns difficult to learn. To solve this problem, Zou *et al.* [35] proposed weight modulation and affinity modulation (AM) while retaining a small model size. The weight modulation of the modulated GCN is as shown below:

$$\mathbf{H}^{l} = s\left((\mathbf{M} \quad (\mathbf{W}\mathbf{H}))\,\tilde{\mathbf{A}}\right) \tag{3}$$

where $\quad$ is an element-wise multiplication, $\mathbf{M} \in \mathbb{R}^{D^{l} \times N}$ is a learnable modulation matrix and set of all modulation vectors, and its i-th column is the modulation vector $\mathbf{m}_i \in \mathbb{R}^{D^{l}}$. The modulation vectors are multiplied by the updated feature vectors of different nodes. Thus, Eq. (3) can be rewritten as follows:

$$\mathbf{h}_i^{l} = s\left(\sum_{j \in \tilde{N}(i)} (\mathbf{m}_i \quad \mathbf{W})\,\mathbf{h}_j \tilde{a}_{ij}\right) \tag{4}$$

The AM of the modulated GCN is as below:

$$\mathbf{A}_{add} = \mathbf{A}_{skeleton} + \mathbf{Q} \tag{5}$$

where $\mathbf{A}_{add}$ is a modulated adjacency matrix $\mathbf{A}_{skeleton} \in \mathbb{R}^{N \times N}$ is an adjacency matrix whose element values are 1 if the corresponding pair of body joints are naturally connected, and zero otherwise. $\mathbf{Q} \in \mathbb{R}^{N \times N}$ is a learnable matrix. Liu *et al.* [15] found that decoupling the transformations of self-connections and other edges can significantly improve the 3D HPE performance, as shown below:

$$\mathbf{H}^{l} = s\left(\mathbf{W}^{(0)}\mathbf{H} + \mathbf{W}^{(1)}\mathbf{H}\hat{\mathbf{A}}\right) \tag{6}$$

where $\mathbf{W}^{(0)} \in \mathbb{R}^{D^{l} \times D}$ and $\mathbf{W}^{(1)} \in \mathbb{R}^{D^{l} \times D}$ are the weight matrices corresponding to the self and neighbour transformations, respectively, and $\hat{\mathbf{A}}$ is the symmetrically normalised version of $\mathbf{A}$ without self-connections. We adopt this method as our baseline. Zou *et al.* [36] exploited a high-order adjacency matrix to consider the relationship of neighbouring nodes at a given k-hop distance. An adjacency matrix of k-hop $\hat{\mathbf{A}}^{k}$ is calculated by the k-th power of $\mathbf{A}$, and then symmetrical normalisation is applied. In this paper, we introduce the MM-GCN. First, we define the simple adjacency matrix of the k-hop distance. Then, we propose the MM-GCN to merge the updated features of each hop distance.

3.2 Multi-hop Modulated GCNs

The MM-GCN consists of two components. One of the components derives the adjacency matrix according to the k-hop distance. The vanilla GCN considers only one-hop neighbours to define the adjacency matrix $\mathbf{A}$. In previous works [1, 2, 36], the adjacency matrix for k-hop neighbouring nodes $\mathbf{A}^{k}$ was computed by the k-th power of the adjacency matrix $\mathbf{A}$. Because $\mathbf{A}^{k}$ represents the relationships of all the neighbouring nodes up to a distance of k hops, it is difficult to merge the features of these multi-hop neighbours. Thus, we propose the simple and novel adjacency matrix for multi-hop neighbouring nodes of the human skeleton, and it is represented as $\mathbf{A}_k^{l}$.

For our proposed adjacency matrix $\mathbf{A}_k^{l}$, as shown in Fig. 1, the neighbouring joints at various numbers of hops from the pelvis, such as the centre joint, are represented by coloured

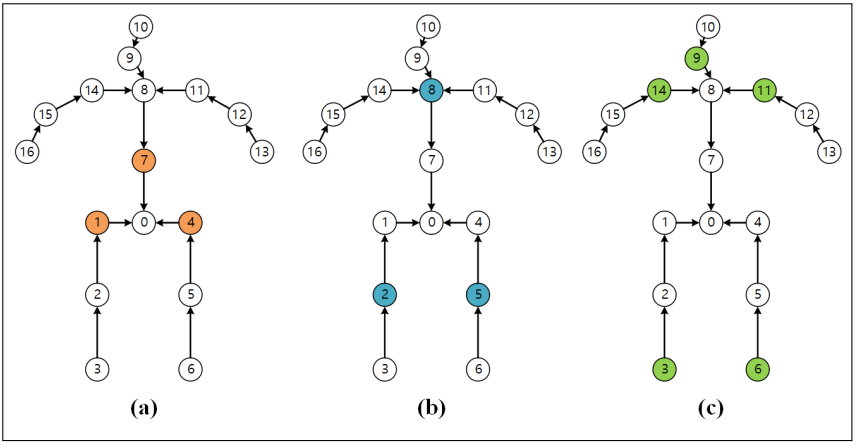

Figure 1: Illustration of neighbouring joints at a multi-hop distance from the pelvis. (a) Neighbouring joints one hop from the pelvis, (b) Neighbouring joints 2-hops from the pelvis, and (c) Neighbouring joints 3-hops from the pelvis.

circles. Fig. 1 (a) illustrates the one-hop neighbouring joints from the centre joint, and the neighbouring joints are indicated by orange circles. Because no single joint can be defined within one hop, the proposed adjacency matrix $\mathbf{A}_1^{\emptyset}$ is equal to the conventional adjacency matrix $\mathbf{A}$. Moreover, Figs. 1 (b) and (c) illustrate the 2-hop neighbouring joints from the centre joint and the 3-hop neighbouring joints from the centre joint, respectively, and these are indicated by blue circles and green circles, respectively. Because the proposed adjacency matrix $\mathbf{A}_k^{\emptyset}$ represents the relationships between neighbouring joints, except middle joints up to a distance of k hops, the relationships between the adjacency matrices $\mathbf{A}_k^{\emptyset}$ of each hop have low correlations with each other. Thus, it can provide a flexible modeling structure for learning the long-range relationships between body joints.

In this study, another component of the proposed method is a modulation method to merge the aggregate features of each hop. First, the updated feature matrix $\mathbf{H}^{\emptyset}$ is derived by simply applying the adjacency matrix $\mathbf{A}_k^{\emptyset}$ to Eq. (6) as follows:

$$\mathbf{H}^{\emptyset} = s\left(\sum_{k=1}^{K} w_k \quad \left(\mathbf{W}^{(0)}\mathbf{H} + \mathbf{W}^{(1)}\mathbf{H}\hat{\mathbf{A}}_k^{\emptyset}\right)\right) \tag{7}$$

where $w_k \; 2 \; \mathrm{R}^{\mathrm{D}^{\emptyset}\mathrm{xN}}$ is a learnable modulation matrix to model the relationships between the aggregate features of each hop distance, $\mathbf{W}^{(0)}$ and $\mathbf{W}^{(1)}$ are the weight matrices corresponding to the self and neighbour transformations, respectively, and $\hat{\mathbf{A}}_k^{\emptyset}$ is the symmetrically normalised version of $\mathbf{A}_k^{\emptyset}$ without self-connections. However, the learnable modulation matrix w_k is insufficient to model the relationships between the aggregate features of each hop distance, because w_k is only utilised to aggregate the features of the k-hop distance. To consider the long-range dependency, we assume that the feature representation of the k-hop distance is influenced by the merged features up to the *(k+1)*-hop distance $\mathbf{C}_{k+1}$, as follows:

$$\mathbf{C}_k = I_k \quad \left(\mathbf{W}^{(0)}\mathbf{H} + \mathbf{W}^{(1)}\mathbf{H}\hat{\mathbf{A}}_k^{\emptyset}\right) + (1 \quad I_k) \quad \mathbf{C}_{k+1} \tag{8}$$

where $I_k \; 2 \; \mathrm{R}^{\mathrm{D}^{\emptyset}\mathrm{xN}}$ is a learnable modulation matrix to model the relationships between the

aggregate features of the k-hop distance $\left(\mathbf{W}^{(0)}\mathbf{H} + \mathbf{W}^{(1)}\mathbf{H}\hat{\mathbf{A}}_k^{\emptyset}\right)$ and the merged features up to the *(k+1)*-hop distance $\mathbf{C}_{k+1}$. If $0 < k < K$, $\mathbf{C}_k$ is calculated using the bi-linear modulation between the aggregate features of the k-hop distance and $\mathbf{C}_{k+1}$. Otherwise, $\mathbf{C}_k$ is equal to $\left(\mathbf{W}^{(0)}\mathbf{H} + \mathbf{W}^{(1)}\mathbf{H}\hat{\mathbf{A}}_k^{\emptyset}\right)$. Thus, the MM-GCN is designed such that the features become more heavily weighted as the hop distance becomes shorter, and it can be represented as follows:

$$\mathbf{H}^{\emptyset} = s\left(I_1 \quad \left(\mathbf{W}^{(0)}\mathbf{H} + \mathbf{W}^{(1)}\mathbf{H}\hat{\mathbf{A}}_1^{\emptyset}\right) + (1 \quad I_1) \quad \mathbf{C}_2\right) \tag{9}$$

3.3 Network Architecture

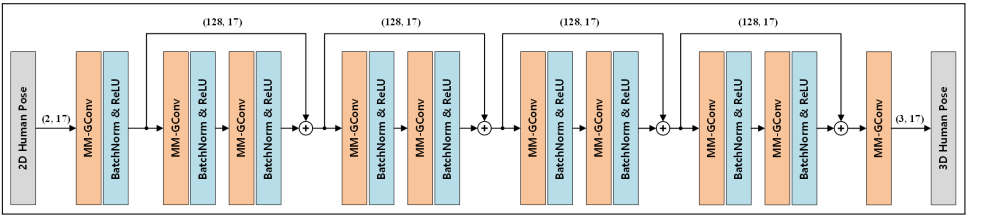

Figure 2: Network architecture of proposed MM-GCN for 3D HPE. (D, N) indicates feature channels and number of body joints, respectively.

Fig. 2 shows the architecture of the MM-GCN for 3D HPE. The input of our MM-GCN is a set of 2D keypoints generated via an off-the-shelf 2D pose detector, and we use two multi-hop modulated graph convolutional (MM-GConv) layers as building blocks and apply a skip connection, such as a residual block. All MM-GConv layers are followed by batch normalisation and ReLU activation except for the last one. The 3D pose is generated by the last layer of the network.

4 Experiments

4.1 Setting

Datasets The Human3.6M dataset contains 3.6 million 3D human poses by 11 professional actors and corresponding images captured by a high-speed motion capture system using four different cameras. Each actor performs 15 everyday activities. Following previous work [16], we used standard normalisation to pre-process the 2D and 3D poses before feeding them into our model. The MPI-INF-3DHP is a recent 3D human pose dataset constructed using a motion capture system with both indoor scenes and complex outdoor scenes. It includes eight actors performing eight activities each. In contrast to Human3.6M, it covers more action classes, ranging from walking and sitting to challenging exercise poses and dynamic actions. To demonstrate the generalisability of our model quantitatively, we evaluated our model on the testing set of MPI-INF-3DHP after the model was trained on Human3.6M. The test split was made up of approximately 3K images from six subjects performing seven actions.

Evaluation Protocols Two standard protocols were exploited to evaluate our model on Human3.6M. We used five subjects (S1, S5, S6, S7, and S8) for training and two subjects

(S9 and S11) for testing under both Protocol #1 and Protocol #2. Under Protocol #1, we reported the mean per-joint position error (MPJPE), which computes the average Euclidean distance between the predicted 3D joint positions and ground truth. Under Protocol #2, we reported the Procrustes-aligned MPJPE (PA-MPJPE), where MPJPE is computed after rigid alignment of the prediction with respect to the ground truth. Both error metrics were measured in millimetres, and lower values indicated better performance. For MPI-INF-3DHP, a 3D extension of the percentage of correct keypoints (3DPCK) and the area under the curve (AUC) were adopted as the evaluation metrics.

Implementation Details The proposed model was implemented in PyTorch and optimised using the Adam optimiser. We trained our model for 200 epochs, setting the decay factor to 0.96 per four epochs with an initial learning rate of 0.01 and a batch size of 1024. Following previous work [35], the other configuration was divisionally set for the 2D ground truth and the 2D pose detection, because 3D pose regression from 2D detections is more challenging than that from 2D ground truth, as the former needs to deal with more uncertainty in the 2D space. For the 2D ground truth, we set the number of channels to 128. We obtained 2D pose detections using a cascaded pyramid network (CPN) [5]. We also set the number of channels to 384 to manage the detection errors. Following [9], we incorporated a non-local layer [28] and the Affinity Modulation [36] to improve the performance. In the ablation study, the non-local layer was excluded. Moreover, we used the 2D ground truth as input to bypass the influence from 2D pose detectors.

4.2 Ablation Study

In the ablation study, we used the 2D ground truth as input to the proposed MM-GCN to eliminate the extra uncertainty from the 2D pose detector. We started by investigating the effect of the various hop distances and the AM from [35] on model performance, where the AM was equal to Eq. (5). We also evaluated our model against [36] and [23], which are state-of-the-art multi-hop-based GCNs for 3D HPE. The results on the Human3.6M dataset are shown in Table 1. As can be seen, the proposed MM-GCN outperformed both high-order GCN and HOIF-Net under Protocols #1 and #2. Under Protocol #1, it did so while using a much smaller number of parameters.

Table 1: Ablation study on MM-GCN. Units of MPJPE and P-MPJPE are millimetres (mm).

Method	# of Hops	# of Channels	# of Parameters	MPJPE	P-MPJPE	Infer.Time	Complexity
Zou *et al.* [36]	2	128	1.20M	39.68	31.69	0.011s	$O((K+1)^2C^3)$
Ours	2	128	0.30M	39.20	30.86	0.007s	$O(2C^3)$
Zou *et al.* [36]	3	96	1.20M	39.52	31.07	0.013s	$O((K+1)^2C^3)$
Quan *et al.* [23]	3	96	1.20M	38.12	29.74	0.012s	$O((K+1)^2C^3)$
Quan *et al.* [23]	3	64	0.54M	39.78	31.26	0.009s	$O((K+1)^2C^3)$
Ours	3	128	0.32M	38.83	30.24	0.007s	$O(2C^3)$

In addition, the last column in Table 1 presents the complexity of various methods. The complexity of MM-GCN is equal to $O(2C^3 + KC^2)$, where C denotes channels and K denotes hops, and it can be rewritten as $O(2C^3)$. Moreover, the inference time of MM-GCN is approximately 0.007s.

Table 2 further compares our proposed MM-GCN with various GCN-based methods for 2D-to-3D pose estimation. We observe that our proposed MM-GCN outperforms the SemGCN with and without non-local modules. Note the non-local module [28] is designed to capture the non-local relationships between nodes, but the performance of the SemGCN with non-local modules is still worse than that of our approach. This demonstrates the great

advantage of our proposed MM-GCN. In addition, we report results on the effect of the AM, which significantly increases the performance of the proposed MM-GCN. Moreover, Table 2 shows that the MPJPE and P-MPJPE of our MM-GCN decrease by 1.42 mm and 0.93 mm, respectively, when the AM for the 3-hop model is applied. For the 2-hop model, the AM reduces the MPJPE and P-MPJPE by 0.84 mm and 1.22 mm, respectively, whereas the AM reduces the MPJPE and P-MPJPE of the modulated GCN by 0.58 mm and 0.29 mm, respectively. Because our MM-GCN derives the adjacency matrix for each hop distance and aggregates features using a learnable modulation matrix, the AM applied to each adjacency matrix provides more powerful feature representation.

Table 2: Performance comparison of proposed MM-GCN and various GCN-based methods.

Method	# of Channels	# of Parameters	MPJPE	P-MPJPE
SemGCN [31]	128	0.27M	42.14	33.53
SemGCN w/ Non-local [31]	128	0.43M	40.78	31.46
Modulated GCN w/o AM [35]	128	0.27M	38.83	30.35
Modulated GCN [35]	128	0.29M	38.25	30.06
Ours (2-hop)	128	0.30M	39.20	30.86
Ours (3-hop)	128	0.32M	38.83	30.24
Ours (2-hop) w/ AM	128	0.31M	38.36	29.64
Ours (3-hop) w/ AM	128	0.33M	37.41	29.31
Ours (4-hop) w/ AM	128	0.36M	36.10	28.76
Ours (5-hop) w/ AM	128	0.38M	35.63	27.55

Also, the below area in Table 3 shows the performance of MM-GCN according to various hop distances. The results show that as the k-hop distance increases, our MM-GCN aggregates features of neighbouring joints at various hop distances more efficiently. Thus, when neighbouring joints are at long hop distance but actually close in terms of Euclidean distance, our MM-GCN can efficiently model the relationship between neighbouring joints.

Table 3: Quantitative comparisons on Human3.6M under Protocol #1. All approaches take 2D ground truth as input.
Legend: (+) uses extra data from MPII dataset. (*) uses pose scales in both training and testing.

Method	Dire.	Disc.	Eat	Greet	Phon.	Phot.	Pose	Purch.	Sit.	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Martinez *et al.* [16]	37.7	44.4	40.3	42.1	48.2	54.9	44.4	42.1	54.6	58.0	45.1	46.4	47.6	36.4	40.4	45.5
Zhao *et al.* [31]	37.8	49.4	37.6	40.9	45.1	41.4	40.1	48.3	50.1	42.2	53.5	44.3	40.5	47.3	39.0	43.8
Zhou *et al.* [34](+)	34.4	42.4	36.6	42.1	38.2	39.8	34.7	40.2	45.6	60.8	39.0	42.6	42.0	29.8	31.7	39.9
Liu *et al.* [20]	36.8	40.3	33.0	36.3	37.5	45.0	39.7	34.9	40.3	47.7	37.4	38.5	38.6	29.6	32.0	37.8
Ci *et al.* [6](+)(*)	36.3	38.8	29.7	37.8	34.6	42.5	39.8	32.5	36.2	39.5	34.4	38.4	38.2	31.3	34.2	36.3
Ours (2-hop) w/ AM	36.1	40.8	33.6	36.6	36.7	42.0	43.4	38.8	40.9	46.8	36.7	39.8	37.6	31.7	33.1	38.4
Ours (3-hop) w/ AM	37.7	39.4	33.7	35.9	37.5	40.9	41.9	34.3	38.6	43.2	36.6	36.7	37.8	30.8	33.4	37.4
Ours (4-hop) w/ AM	32.5	38.8	30.7	34.4	35.3	41.2	38.7	31.5	39.6	46.3	34.9	36.8	35.9	29.5	30.4	36.1
Ours (5-hop) w/ AM	34.6	39.6	31.3	34.7	33.9	40.3	39.5	32.2	35.4	43.5	34.0	35.1	36.9	29.7	31.4	35.6

4.3 Comparison with State-of-the-Art Techniques

We compared the MM-GCN with some state-of-the-art methods on Human3.6M under both Protocol #1 and Protocol #2. Following previous works [16], we used 2D poses detected by a pre-trained CPN [5] as the input. Some of the state-of-the-art methods utilised post-processing [35] , such as pose refinement. Thus, all experiments were evaluated without post-processing. The results are reported in Table 4 and Table 5. The proposed MM-GCN

Table 4: Quantitative comparisons on Human3.6M under Protocol #1. Boldface numbers indicate the best 3D pose estimation performance.

Method	Dire.	Disc.	Eat	Greet	Phon.	Phot.	Pose	Purch.	Sit.	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Martinez *et al.* [24]	51.8	56.2	58.1	59.0	69.5	78.4	55.2	58.1	74.0	94.6	62.3	59.1	65.1	49.5	52.4	62.9
Sun *et al.* [29]	52.8	54.8	54.2	54.3	61.8	67.2	53.1	53.6	71.7	86.7	61.5	53.4	61.6	47.1	53.4	59.1
Yang *et al.* [30]	51.5	58.9	50.4	57.0	62.1	65.4	49.8	52.7	69.2	85.2	57.4	58.4	60.1	43.6	47.7	58.6
Fang *et al.* [8]	50.1	54.3	57.0	57.1	66.6	73.3	53.4	55.7	72.8	88.6	60.3	57.7	62.7	47.5	50.6	60.4
Pavlakos *et al.* [21]	48.5	54.4	54.4	52.0	59.4	65.3	49.9	52.9	65.8	71.1	56.6	52.9	60.9	44.7	47.8	56.2
Zhao *et al.* [34]	47.3	60.7	51.4	60.5	61.1	49.9	**47.3**	68.1	86.2	55.0	67.8	61.0	42.1	60.6	45.3	57.6
Sharam *et al.* [27]	48.6	54.5	54.2	55.7	62.2	72.0	50.5	54.3	70.0	78.3	58.1	55.4	61.4	45.2	49.7	58.0
Zou *et al.* [33]	49.0	54.5	52.3	53.6	59.2	71.6	49.6	49.8	66.0	75.5	55.1	53.8	58.5	40.9	45.4	55.6
Quan *et al.* [23]	47.0	53.7	50.9	52.4	57.8	71.3	50.2	49.1	63.5	76.3	54.1	51.6	56.5	41.7	45.3	54.8
Liu *et al.* [15]	**46.3**	52.2	47.3	**50.7**	55.5	67.1	49.2	**46.0**	60.4	71.1	**51.5**	50.1	**54.5**	**40.3**	43.7	52.4
Zou&Tang [35]	48.2	51.6	47.8	51.8	53.1	61.5	50.4	48.4	60.5	**67.3**	52.2	49.0	55.3	40.9	42.6	52.4
Ours(3-hop)	46.8	**51.4**	**46.7**	51.4	**52.5**	**59.7**	50.4	48.1	**58.0**	67.7	51.5	**48.6**	54.9	40.5	**42.2**	**51.7**

Table 5: Quantitative comparisons on Human3.6M under Protocol #2. Boldface numbers indicate best 3D pose estimation performance.

Method	Dire.	Disc.	Eat	Greet	Phon.	Phot.	Pose	Purch.	Sit.	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Martinez *et al.* [24]	39.5	43.2	46.4	47.0	51.0	56.0	41.4	40.6	56.5	69.4	49.2	45.0	49.5	38.0	43.1	47.7
Sun *et al.* [29]	42.1	44.3	45.0	45.4	51.5	53.0	43.2	41.3	59.3	73.3	51.0	44.0	48.0	38.3	44.8	48.3
Fang *et al.* [8]	38.2	41.7	43.7	44.9	48.5	55.3	40.2	38.2	54.5	64.4	47.2	44.3	47.3	36.7	41.7	45.7
Pavlakos *et al.* [21]	34.7	39.8	41.8	**38.6**	42.5	47.5	38.0	36.6	50.7	56.8	42.6	39.6	43.9	32.1	36.5	41.8
Hossain&Little [10]	35.7	**39.3**	44.6	43.0	47.2	54.0	38.3	37.5	51.6	61.3	46.5	41.4	47.3	34.2	39.4	44.1
Zou *et al.* [33]	38.6	42.8	41.8	43.4	44.6	52.9	**37.5**	38.6	53.3	60.0	44.4	40.9	46.9	32.2	37.9	43.7
Quan *et al.* [23]	36.9	42.1	40.3	42.1	43.7	52.7	37.9	37.7	51.5	60.3	43.9	39.4	45.4	31.9	37.8	42.9
Liu *et al.* [15]	35.9	40.0	38.0	41.5	42.5	51.4	37.8	**36.0**	48.6	56.6	41.8	38.3	**42.7**	31.7	36.2	41.2
Zou&Tang [35]	36.6	40.1	37.7	41.7	41.0	46.4	38.4	36.8	48.5	**53.5**	42.1	37.1	43.5	32.0	34.9	41.0
Ours(3-hop)	**35.7**	39.6	**37.3**	41.4	**40.0**	**44.9**	37.6	36.1	**46.5**	54.1	**40.9**	**36.4**	42.8	**31.7**	**34.7**	**40.3**

performed the best on most of the tasks, and on average, under both Protocol #1 and Protocol #2, indicating that the MM-GCN is very competitive. Under Protocol #1, Table 4 shows that our proposed model achieved an improvement of 0.7 mm compared with the previous best result of 52.4 mm. Under Protocol #2, Table 5 shows that our model performed better than the previous best result with a 0.7 mm error reduction on average.

Table 6: Quantitative comparisons on MPI-INF-3DHP. Higher 3DPCK and AUC values indicate better performance.

Method	3DPCK	AUC
Yang *et al.* [30]	69.0	32.0
Pavlakos *et al.* [21]	71.9	35.3
Habibie *et al.* [10]	70.4	36.0
Wang *et al.* [27]	71.9	35.8
Quan *et al.* [23]	72.8	36.5
Liu *et al.* [15]	79.3	47.6
Ours	**81.6**	**50.3**

We also evaluated our MM-GCN on the testing set of MPI-INF-3DHP to test its generalisability across different datasets. Following [15], we used the 2D joints provided by the dataset as input. The results are shown in Table 6. As can be seen, our method achieved the best performance on all evaluation metrics.

Fig. 3 shows the visualisation results obtained by our MM-GCN on Human3.6M. It can accurately predict the 3D poses of different persons who are performing various actions, indicating the effectiveness of our proposed approach in tackling the 2D-to-3D pose estimation problem.

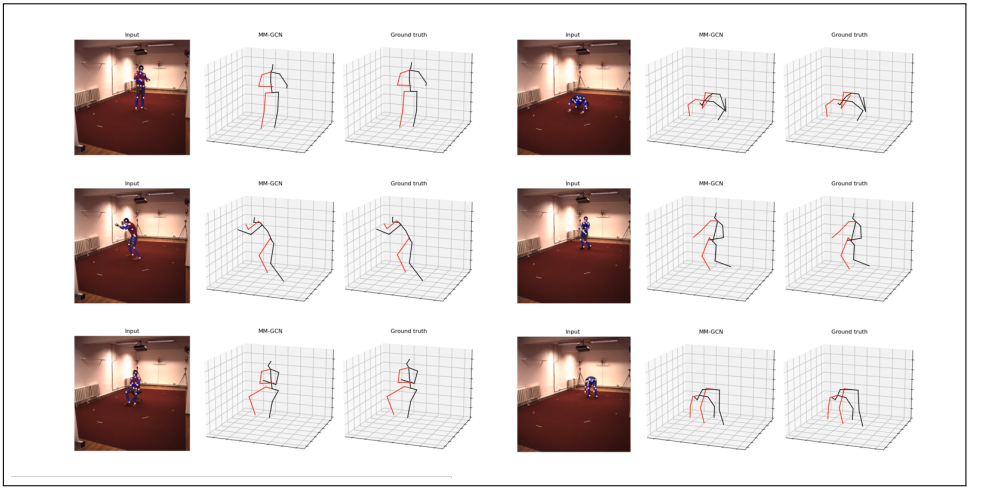

Figure 3: Qualitative results obtained by our MM-GCN on the Human3.6M test set.

5 Conclusion

In this paper, we introduced an MM-GCN for 3D HPE to effectively model long-range dependencies between each body part and its distant neighbours. We derived the adjacency matrix using a simple concept and modulated the features of each hop distance to aggregate multi-hop neighbours. We performed experiments and an ablation study to highlight the merits of our model and to demonstrate its competitive performance in comparison with state-of-the-art methods for 3D HPE.

References

[1] Sami Abu-El-Haija, Bryan Perozzi, Amol Kapoor, Nazanin Alipourfard, Kristina Lerman, Hrayr Harutyunyan, Greg Ver Steeg, and Aram Galstyan. Mixhop: Higher-order graph convolutional architectures via sparsified neighborhood mixing. In *Proc. International Conference on Machine Learning*, pages 21–29, 2019.

[2] Zhimin Bai, Hongping Yan, and Lingfeng Wang. High-order graph convolutional network for skeleton-based human action recognition. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 14–25, 2019.

[3] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

[4] Yujun Cai, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat Thalmann. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In *Proc. IEEE/CVF International Conference on Computer Vision*, pages 2272–2281, 2019.

[5] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 7103–7112, 2018.

[6] Hai Ci, Chunyu Wang, Xiaoxuan Ma, and Yizhou Wang. Optimizing network structure for 3d human pose estimation. In *Proc. IEEE/CVF International Conference on Computer Vision*, pages 2262–2271, 2019.

[7] Hao-Shu Fang, Yuanlu Xu, Wenguan Wang, Xiaobai Liu, and Song-Chun Zhu. Learning pose grammar to encode human body configuration for 3d pose estimation. In *Proc. AAAI conference on artificial intelligence*, 2018.

[8] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3d hand shape and pose estimation from a single rgb image. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10833–10842, 2019.

[9] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *Proc. International conference on machine learning*, pages 1263–1272, 2017.

[10] Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Gerard Pons-Moll, and Christian Theobalt. In the wild human pose estimation using explicit 2d features and intermediate 3d representations. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10905–10914, 2019.

[11] Mir Rayat Imtiaz Hossain and James J Little. Exploiting temporal information for 3d human pose estimation. In *Proc. European Conference on Computer Vision*, pages 68–84, 2018.

[12] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7): 1325–1339, 2013.

[13] Chen Li and Gim Hee Lee. Generating multiple hypotheses for 3d human pose estimation with mixture density network. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9887–9895, 2019.

[14] Sijin Li and Antoni B Chan. 3d human pose estimation from monocular images with deep convolutional neural network. In *Proc. Asian Conference on Computer Vision*, pages 332–347, 2014.

[15] Kenkun Liu, Rongqi Ding, Zhiming Zou, Le Wang, and Wei Tang. A comprehensive study of weight sharing in graph networks for 3d human pose estimation. In *Proc. European Conference on Computer Vision*, pages 318–334, 2020.

[16] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proc. IEEE International Conference on Computer Vision*, pages 2640–2649, 2017.

[17] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *Proc. International Conference on 3D Vision (3DV)*, pages 506–516. IEEE, 2017.

[18] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proc. International Conference on Machine Learning*, 2010.

[19] Sungheon Park, Jihye Hwang, and Nojun Kwak. 3d human pose estimation using convolutional neural networks with 2d pose information. In *Proc. European Conference on Computer Vision*, pages 156–169, 2016.

[20] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 7025–7034, 2017.

[21] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Ordinal depth supervision for 3d human pose estimation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 7307–7316, 2018.

[22] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7753–7762, 2019.

[23] Jianning Quan and A Ben Hamza. Higher-order implicit fairing networks for 3d human pose estimation. *Proc. British Machine Vision Conference*, 2021.

[24] Saurabh Sharma, Pavan Teja Varigonda, Prashast Bindal, Abhishek Sharma, and Arjun Jain. Monocular 3d human pose estimation by generation and ordinal ranking. In *Proc. IEEE/CVF international conference on computer vision*, pages 2325–2334, 2019.

[25] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *Proc. IEEE International Conference on Computer Vision*, pages 2602–2611, 2017.

[26] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *Proc. European Conference on Computer Vision*, pages 529–545, 2018.

[27] Jue Wang, Shaoli Huang, Xinchao Wang, and Dacheng Tao. Not all parts are created equal: 3d human pose estimation by modeling bi-directional dependencies of body parts. In *Proc. International Conference on Computer Vision*, 2019.

[28] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.

[29] Max Welling and Thomas N Kipf. Semi-supervised classification with graph convolutional networks. In *Proc. International Conference on Learning Representations*, 2016.

[30] Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy Ren, Hongsheng Li, and Xiaogang Wang. 3d human pose estimation in the wild by adversarial learning. In *Proc. IIEEE Conference on Computer Vision and Pattern Recognition*, pages 5255–5264, 2018.

[31] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N Metaxas. Semantic graph convolutional networks for 3d human pose regression. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3425–3435, 2019.

[32] Kun Zhou, Xiaoguang Han, Nianjuan Jiang, Kui Jia, and Jiangbo Lu. Hemlets pose: Learning part-centric heatmap triplets for accurate 3d human pose estimation. In *Proc. International Conference on Computer Vision*, pages 2344–2353, 2019.

[33] Xingyi Zhou, Xiao Sun, Wei Zhang, Shuang Liang, and Yichen Wei. Deep kinematic pose regression. In *Proc. European Conference on Computer Vision*, pages 186–201, 2016.

[34] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In *Proc. IEEE International Conference on Computer Vision*, pages 398–407, 2017.

[35] Zhiming Zou and Wei Tang. Modulated graph convolutional network for 3d human pose estimation. In *Proc. IEEE/CVF International Conference on Computer Vision*, pages 11477–11487, 2021.

[36] Zhiming Zou, Kenkun Liu, Le Wang, and Wei Tang. High-order graph convolutional networks for 3d human pose estimation. In *Proc. British Machine Vision Conference*, 2020.