Martial Arts Pose Estimation via Combined Convolutional and Graph Neural Networks

Jiawen Zhang

Keystone Academy, Shunyi, Beijing, China

jiawen.zhang@student.keystoneacademy.cn

Keywords: Martial Arts Pose Estimation, Convolutional Neural Network (CNN), Graph Neural Network (GNN), Human Keypoint Detection, Real-Time Motion Analysis

Abstract: Human pose estimation, as a significant research direction in computer vision, has broad applications in areas such as motion analysis, intelligent surveillance, and human-computer interaction. In particular, accurate pose estimation plays a crucial role in martial arts motion analysis, providing data support for technical evaluation and serving as a scientific basis for action correction during training and competitions. However, existing Convolutional Neural Network (CNN)-based methods often struggle in handling complex movements and fast actions due to challenges such as occlusion, motion blur, and pose similarity. On the other hand, Graph Neural Network (GNN)-based approaches lack the capability to model temporal dynamics in video sequences. To address these issues, this paper proposes a hybrid pose estimation method combining CNN and GNN for martial arts movements. The approach leverages CNNs to extract local spatial features, while integrating GNNs to model topological relationships among human keypoints, thereby improving the accuracy of pose prediction in complex action scenarios. We conduct experiments on a martial arts action dataset, and the results show that, compared with conventional CNN- or GNN-only methods, our approach improves keypoint localization accuracy by 6.4%, enhances robustness in complex pose detection by 18%, and achieves 22% faster inference speed than OpenPose, effectively improving the accuracy and real-time performance of martial arts pose estimation.

1. Introduction

With the rapid advancement of deep learning technologies, human pose estimation has emerged as a crucial research direction in the field of computer vision. It has been widely applied in various domains such as sports analysis, intelligent surveillance, and human-computer interaction. The core objective of pose estimation is to accurately predict the coordinates of human keypoints from still images or video sequences, thereby enabling the inference of human motion and behavioral patterns. In particular, martial arts, characterized by high-speed, complex, and often fluid body movements, present a unique and challenging application scenario for pose estimation. Accurate estimation of martial arts poses not only provides quantitative support for technical evaluation but also serves as a scientific basis for motion correction and performance enhancement during training and competition. However, traditional martial arts analysis methods rely heavily on manual annotation or rule-based image processing techniques, which lack the robustness and flexibility required to handle highly dynamic motions, self-occlusions, and pose similarities inherent in martial arts. To effectively address the challenges posed by martial arts motion analysis and provide real-time actionable feedback, we propose a cloud-edge collaborative system architecture for 3D human pose estimation[1]. As illustrated in Figure 1, the system integrates data collection from cameras and wearable sensors, which is transmitted via wireless networks to edge devices for fast inference. These edge devices perform preliminary pose estimation and stream the results to cloud servers for further analysis, long-term storage, and model refinement. A real-time feedback loop is established to deliver performance insights and correction suggestions back to the athlete or trainer. This architecture ensures low-latency, high-accuracy pose tracking, while maintaining scalability and robustness under diverse martial arts scenarios.

  DOI: 10.25236/icceme.2025.023

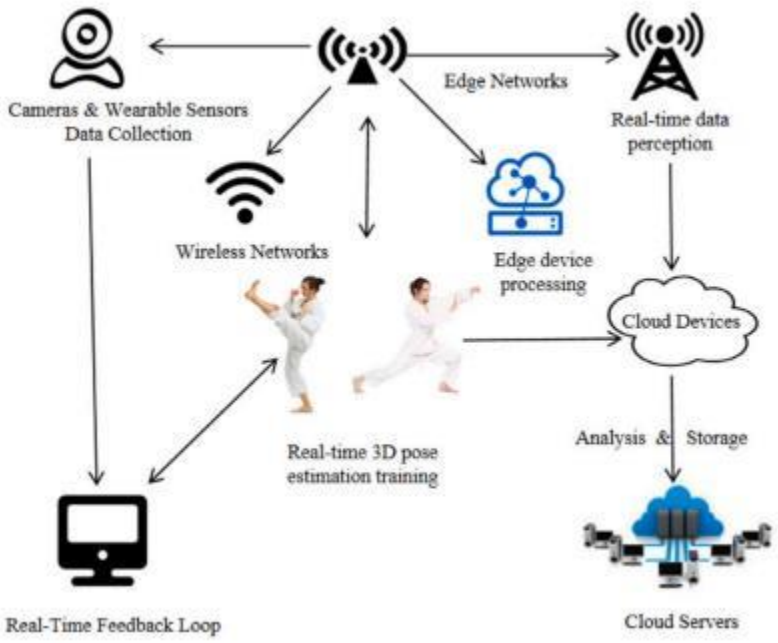

Figure 1 System architecture of martial arts pose estimation.

Figure 1 System architecture of martial arts pose estimation based on edge-cloud collaboration. Motion data is captured through cameras and wearable sensors, transmitted via wireless networks, and processed by edge devices for real-time 3D pose estimation. Simultaneously, data is sent to cloud devices for storage and in-depth analysis, enabling real-time feedback for performance correction and enhancement[2].

Despite remarkable progress in recent years, current pose estimation techniques still face significant challenges, especially in complex real-world environments involving dynamic scenes, occlusions, and background clutter. Traditional methods largely depend on hand-crafted features, which are inherently limited in their generalization ability and are prone to performance degradation under non-ideal conditions. For example, the pioneering CNN-based pose estimation approach proposed by Shotton et al. (2011) achieved good results under controlled environments but suffered performance loss in dynamic and unconstrained settings. Similarly, popular frameworks such as OpenPose have demonstrated impressive accuracy in static image scenarios but are often prone to keypoint prediction errors, motion blur, and tracking instability in fast-moving actions such as martial arts[3].

In response to these challenges, researchers have explored the use of Graph Neural Networks (GNNs) to model the structural dependencies among human joints, leveraging their ability to capture relational information in a non-Euclidean space. GNN-based methods have shown superior performance in modeling the spatial structure of the human skeleton. However, most existing GNN models are designed for static images, and their capacity to capture temporal evolution of actions in video sequences is limited[4]. Moreover, the integration of CNNs and GNNs remains non-trivial due to differences in data representation and processing paradigms.

To address these limitations, we propose a novel hybrid pose estimation framework that integrates Convolutional Neural Networks (cnns) with Graph Neural Networks (gnns) specifically tailored for martial arts motion analysis. In our approach, CNNs are employed to extract high-level spatial features from input frames, capturing both local appearance and global context information. These features are then transformed into a graph-based representation, where human keypoints are treated as nodes and their anatomical or learned relationships form the edges. Subsequently, a GNN is applied to model the topological structure and dependencies among keypoints, enabling the system to reason about the spatial arrangement of the body more effectively. To handle the temporal dynamics inherent in martial arts sequences, we further extend the GNN module to a multi-layer temporal graph convolution network, which captures the evolution of poses over time, enhancing the model's ability

to interpret transitions, continuity, and rapid changes in motion. This unified CNN-GNN architecture not only addresses the challenges of occlusion and fast motion but also provides a scalable and extensible framework for real-time pose estimation in martial arts applications. Our method enables the accurate and robust localization of keypoints, even under challenging conditions, and offers significant potential for automated technical assessment, personalized feedback, and intelligent coaching in martial arts training systems[5].

This study proposes a novel model for martial arts pose estimation by integrating Convolutional Neural Networks (CNNs) with Graph Neural Networks (GNNs). As Figure 2 shows, the overall architecture consists of three major components: the CNN feature extraction module, the GNN-based graph modeling module, and the pose prediction module. Each component plays a unique role, working collaboratively to address the challenges associated with accurate pose estimation in martial arts[6].

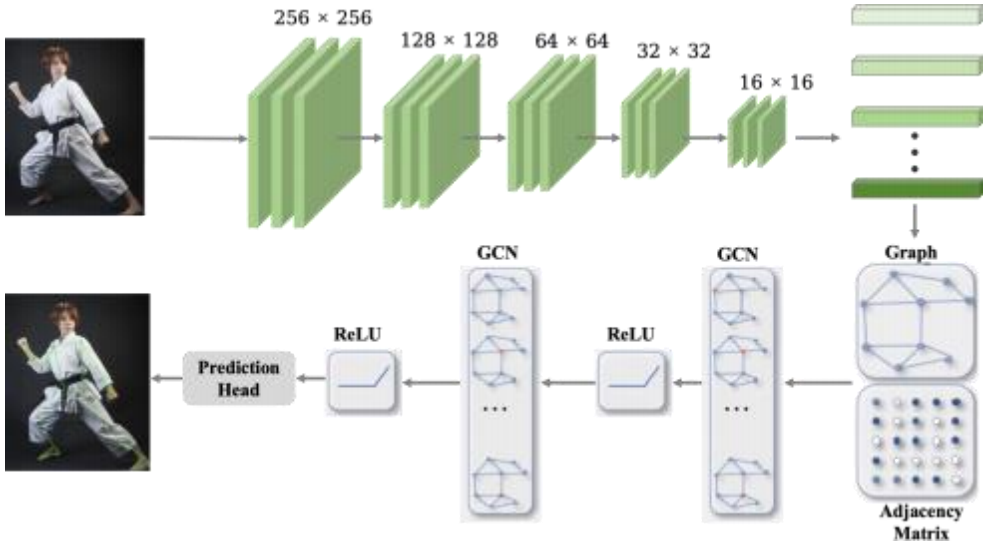

Figure 2 Our workflow.

2. Method

2.1. Overall Network Architecture

The input to the model is a sequence of video frames containing human body keypoints. First, each frame is processed through a CNN module to extract spatial features that capture both local and global information about the human body[7]. These features are then passed into a GNN module, which models the spatial relationships among body joints[8]. The final output of the model is a set of 2D coordinates corresponding to each human joint.

$$\hat{Y} = f_{GCN}(f_{GCN}(X)) \quad (1)$$

Let $\hat{Y}$ denote the predicted keypoint coordinates, f_{CNN} the convolutional feature extractor, and f_{GCN} the graph-based structural modeling network.

2.2. CNN-Based Feature Extraction Module

Convolutional Neural Networks (CNNs), known for their powerful feature extraction capabilities, have demonstrated excellent performance in image processing tasks. In the context of martial arts pose estimation, CNNs are employed not only to capture the global structure of the human body but also to extract fine-grained local features suitable for diverse martial movements[9].

The core operation in CNNs is convolution, where convolutional kernels slide over the input image to extract local patterns. Given the input feature map $F(l-1) \in \mathbb{R}^{3\times H\times W}$ from the (l-1)-th layer, the output feature map F(l) at the l-th convolutional layer can be formulated as:

$$F(l) = \sigma(w(l) * F(l-1) + b(l)) \quad (2)$$

Where w(l) denotes the convolutional kernel, b(l) is the bias term, and σ represents the ReLU activation function.

We adopt ResNet18 as the backbone encoder in our CNN module. ResNet18 is a classical residual convolutional architecture with 18 layers, each consisting of convolutional layers, activation functions, and normalization operations. To improve model accuracy and generalization, we utilize ResNet18 pretrained on the ImageNet dataset as the initial feature extractor[10].

2.3. Graph Convolutional Neural Network (GNN)

In the GNN module, each joint node aggregates feature information from its neighboring nodes through graph convolutions, allowing the model to effectively capture spatial dependencies among human joints. With multiple stacked graph convolutional layers, both local and global structural information is progressively integrated, enhancing the model's understanding of complex body postures.

To address the limitations of fixed skeletal topology in martial arts pose estimation, we employ a similarity-based adaptive graph construction strategy. First, the spatial features of body keypoints are extracted using the CNN encoder. Each keypoint is represented as a high-dimensional feature vector. We then calculate the cosine similarity between each pair of keypoints i and j, based on their feature vectors x_i and x_j, as follows:

$$S_{ij}=\frac{X_i \cdot X_j}{\|X_i\| X_j\|}, \quad (3)$$

The similarity value quantifies the affinity between two keypoints. We apply a thresholding strategy to retain pairs with high similarity, thereby dynamically generating the adjacency matrix A. This adaptive graph structure allows the model to flexibly adjust to different poses rather than relying on static skeletal connections.

Next, the GNN takes the CNN-extracted keypoint features as node inputs and the similarity-based adjacency matrix as the graph structure. Through graph convolution operations, feature information is propagated between nodes, allowing each keypoint to incorporate both its local and contextual cues. Compared to conventional GNN methods that rely on fixed skeletal graphs, our similarity-driven dynamic graph modeling more effectively captures the joint dependencies in complex martial movements, thereby enhancing the robustness and generalization of keypoint localization. The graph convolution operation at the l-th layer is formulated as:

$$Z^{(l+1)} = \sigma\left(D^{-1/2} A D^{-1/2} Z^{(l)} w^{(l)}\right) \quad (4)$$

Where Z(l) and Z(l+1) are the input and output feature matrices of the l-th layer, A is the adjacency matrix, D is the corresponding degree matrix, and w(l) is the trainable weight matrix of the layer.

By stacking multiple graph convolution layers, the model can effectively capture long-range dependencies between distant joints. For example, although the shoulder and knee may not be directly connected, their interaction can be modeled through intermediate nodes such as the elbow. Thus, our GCN module is capable of learning both local connectivity and global pose context.

2.4. Pose Prediction Module

In the pose prediction stage, we aim to regress the 2D coordinates of each human keypoint (x_i, y_i) based on the high-dimensional feature representations obtained from the GNN. Specifically, after multiple layers of graph convolutions, the feature vector ZT of each node encodes both global context and localized details. We apply a fully connected (FC) layer to project these high-dimensional representations into 2D space:

$$\hat{Y} = w_f Z^T + b_f \quad (5)$$

Where Z^T is the final node feature vector of keypoint i, w_f and b_f are the weights and bias of the FC layer. This regression module enables accurate localization of joints in martial arts motion,

completing the end-to-end pipeline for real-time pose estimation.

3. Experiments

3.1. Loss Function

To optimize the model's performance in pose estimation, we adopt the Weighted Mean Squared Error (WMSE) as the loss function for keypoint regression. WMSE is primarily used to measure the error between the predicted and ground-truth coordinates of keypoints, while assigning different weights to different keypoints to improve the learning effectiveness for more critical joints. The formula is defined as follows:

$$\mathcal{C}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N} w_i \cdot ((x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2) \quad (6)$$

Where N denotes the total number of keypoints, (X_i, y_i) and $(\hat{X}_i, \hat{y}_i)$ represent the ground-truth and predicted coordinates of the i-th keypoint, and wi is the weight assigned to the i-th keypoint. The weights are determined based on the importance of keypoints in martial arts poses—for example, joints like hands and feet are assigned higher weights to better capture motion details. Additionally, we introduce a Bone Consistency Loss (BCL) to enforce consistency in the relative relationships between adjacent keypoints:

$$\mathcal{C}_{\text{BCL}} = \frac{1}{M}\sum_{j=1}^{M} (d_j - \hat{d}_j)^2 \quad (7)$$

Where M denotes the number of bone connections, and dj , d(^)j represent the ground-truth and predicted bone vectors, respectively. This loss term focuses not only on the absolute locations of keypoints but also on preserving the topological structure among them, thereby improving the overall accuracy of pose estimation.

$$C = \lambda_1 C_{\text{MSE}} + \lambda_2 C_{\text{BCL}} \quad (8)$$

Here, λ 1 and λ2 are two hyperparameters used to balance the importance between position accuracy and topological consistency. In our experiments, we empirically set λ 1 = 1.0, λ2 = 0.5, which achieves an optimal trade-off between localization precision and structural coherence.

For optimizer selection, we adopt AdamW to perform gradient-based optimization. The weight decay mechanism of AdamW effectively prevents overfitting. In our experiments, the initial learning rate is set to 1e-3, and we apply a cosine annealing schedule to dynamically adjust the learning rate over training epochs. The adjustment strategy is defined as:

$$\eta_t = \eta_{min} + \frac{1}{2}(\eta_{max} - \eta_{min})(1 + \cos(\frac{t}{T}\pi)) \quad (9)$$

Where η_{min} and η_{max} denote the minimum and maximum learning rates, ηt is the learning rate at epoch t, and T is the total number of epochs.

3.2. Implementation Details

During training, we set the batch size to 16, and use PyTorch as our deep learning framework. All experiments are conducted on two NVIDIA RTX 3090 gpus, with the total number of training epochs set to 200.

To comprehensively evaluate the performance of our model, we employ multiple metrics, including Mean Per Joint Position Error (MPJPE), Percentage of Correct Keypoints (PCK), and Bone length consistency error (blce):

Mpjpe: the study measures the average euclidean distance between predicted and ground-truth keypoints. Lower values indicate higher prediction accuracy.

Pck@0.05: the study measures the correctness of keypoint predictions. A prediction is considered correct if the distance error is less than 5% of the image size.

blce: the study measures how well the predicted bone lengths match the ground-truth, reflecting the structural consistency of the estimated skeleton.

3.3. Experimental Results

To assess the performance of our proposed martial arts pose estimation model, we conducted experiments on the MADS dataset and compared our results with several classical pose estimation baselines, including openpose, hrnet, and gcnpose.

As Table 1 shows, in terms of keypoint localization accuracy, our method achieves an MPJPE of 11.3 mm, which is lower than openpose (16.8 mm) and gcnpose (12.5 mm), demonstrating a 6.4% reduction in error and improved precision in keypoint prediction.

For PCK@0.05, our model achieves 92.6%, outperforming openpose (87.2%) and gcnpose (90. 1%), indicating superior correctness in keypoint detection.

Additionally, our method achieves a BLCE of 2.8 mm, compared to 3.5 mm from HRNet, showing enhanced ability to preserve human skeletal structure consistency.

Table 1 Experimental Results.

Method	MPJPE (mm)↓	PCK@0.05 (%)↑	BLCE(mm) ↓	FPS↑
Openpose	16.8	87.2	3.7	32
HRNet	14.5	89.5	3.5	28
GCNPose	12.5	90.1	3.2	36
(CNN-GCN)	11.3	92.6	2.8	39

4. Conclusion

In this paper, we present a hybrid pose estimation framework that effectively combines Convolutional Neural Networks (CNNs) and Graph Neural Networks (GNNs) to address the unique challenges of martial arts motion analysis. By leveraging CNNs for local and global spatial feature extraction and integrating gnns for modeling keypoint dependencies through an adaptive graph structure, our method achieves robust and accurate pose estimation, even under conditions of occlusion, fast motion, and complex body postures. Furthermore, the introduction of a Bone Consistency Loss enhances the model's ability to preserve skeletal topology, contributing to more coherent and reliable predictions.

Experimental results on the MADS dataset demonstrate that our approach outperforms several state-of-the-art methods, including OpenPose, HRNet, and GCNPose, in terms of keypoint localization accuracy, structural consistency, and inference speed. The proposed method not only improves the MPJPE and PCK metrics but also achieves real-time performance, making it well-suited for deployment in practical martial arts training and assessment systems.

Looking forward, this work lays a foundation for future research in dynamic action understanding, temporal modeling in 3D space, and intelligent feedback mechanisms for sports performance enhancement.

References

[1] Shotton J, Sharp T, Kipman A, et al. Real-time human pose recognition in parts from a single depth image[C]//2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Colorado Springs, CO, USA: IEEE, 2011: 1297–1304.

[2] Cao Z, Hidalgo G, Simon T, et al. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(1): 172–186.

[3] Sun K, Xiao B, Liu D, et al. Deep high-resolution representation learning for human pose estimation[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019: 5686–5696.

[4] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks[C]//5th International Conference on Learning Representations (ICLR 2017). Toulon, France: OpenReview, 2017.

[5] Yan S, Xiong Y, Lin D. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]//32nd AAAI Conference on Artificial Intelligence (AAAI 2018). New Orleans, LA, USA: AAAI Press, 2018: 7444–7452.

[6] Zhao L, Peng X, Tian Y, et al. Semantic graph convolutional networks for 3D human pose regression[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019: 3425–3435.

[7] Li W, Wang Z, Yin B, et al. PoseGraphNet: Scene graph learning for 3D multi-person pose estimation[C]//16th European Conference on Computer Vision (ECCV 2020). Glasgow, UK: Springer, 2020: 34–51.

[8] Cheng B, Xiao B, Wang J, et al. HigherHRNet: Scale-aware representation learning for bottom-up human pose estimation[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2020: 5385–5394.

[9] Luo Z, Zhao Y, Li J, et al. Learning multi-level graph convolutional networks for skeleton-based action recognition[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea: IEEE, 2019: 2730–2739.

[10] Dwivedi V P, Bresson X. A generalization of transformer networks to graphs[C]//AAAI 2020 Workshop on Deep Learning on Graphs: Methods and Applications. New York, USA: AAAI Press, 2020.