SKELETON-BASED HUMAN ACTION RECOGNITION USING SPATIO-TEMPORAL LATENT FEATURES WITH GCN MODEL

Marakhimov Avazjon

Department of Information Processing and Control Systems, Tashkent State Technical University, Tashkent, Uzbekistan, DSc, professor;

Khudaybergenov Kabul

Kimyo International University in Tashkent, Uzbekistan, PhD.

E-mail: kabul.kudaybergenov@gmail.com

*DOI: **https://doi.org/10.5281/zenodo.19829095***

Abstract: *In this work we present LFHAR (Latent Features for Human Action Recognition), a novel architecture that utilizes multiple spatio-temporal latent representations to improve action feature extraction. The approach applies graph-based transformations to individual skeletal frames in temporal sequences, then arranges the derived graph features into spatio-temporal matrices. The method produces substantial performance improvements, achieving accuracy increases of 2.7% and 2.1% on the NTU-RGB+D 60 and NTU-RGB+D 120 datasets, respectively, confirming its efficacy in improving skeleton-based action recognition.*

Keywords: *Latent features, Skeleton-based action recognition, Spatio-temporal graph network, action classification, Deep Learning.*

1. Introduction

Human action recognition represents an important research area with numerous practical applications [1], including human-computer interaction, robotics, virtual reality, and intelligent video surveillance systems. Action recognition based on skeletal data focuses on classifying action types using skeletal representations of human bodies, with its effectiveness demonstrated through notable achievements in recent studies. Skeletal data comprises sequences of three-dimensional joint coordinates that preserve the inherent topological structure of the human body [2]. In skeleton-based action recognition, traditional methods treat individual human body joints as separate feature-carrying units and establish spatiotemporal relationships through manually designed strategies[3-4].

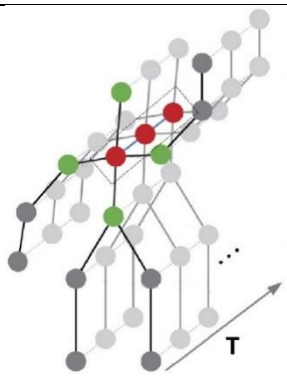

Fig. 1. Skeleton Spatio-temporal Graph. The spatial graph consists of green vertices and the temporal graph consists of red vertices.

2. Problem formulation

Skeleton-based actions consist of hundreds or thousands of continuous 3D skeleton poses connected in the temporal domain. First, how can we model these skeleton actions into a Spatio-Temporal Graph Neural Network? In the spatial domain, let the skeleton pose of frame t be $P^{(t)} \in \mathbb{R}^{N\times C}$, where N is the number of joints contained in the skeleton and C are the 3D coordinates of each joint. In fact, the skeleton pose essentially corresponds to a skeleton spatial graph G_S, as shown in Fig. 1, in which joints are regarded as vertices (green nodes) and bones as edges (black lines). G_S can be further formulated as an undirected graph $G_S\{V, E_{spa}\}$, which consists of $N = |V|$ vertices $V = v_1,...,v_N$ and $M = |E_{spa}|$ edges $E_{spa} = \{e_1,...,e_M\}$. The strength of spatial correlation between vertices is encoded into the adjacency matrix $A \in \mathbb{R}^{N\times N}$. In the temporal domain, as shown in Fig. 2, the vertices (red nodes) representing the joints form a linear sequence. Thus, the temporal correlation of the joints is used to construct temporal edges (blue lines) $E_{tem} = \{e_1,...,e_T\}$. $T = |E_{tem}|$ represents the number of frames. Ultimately, skeleton-based actions are formulated as a Spatial-Temporal Graph $G_{ST} = \{V, \{E_{spa}, E_{tem}\}\}$.

On this basis, let a set of poses belonging to one action class be $\{P, y\}$, where $P = \{P^{(1)},...,P^{(T)}\} \in \mathbb{R}^{N\times C\times T}$ denotes the initial action feature representation embedded by Spatial-temporal Graph G_{ST}, T is the number of frames, $y \in \{0,1\}^D$ is the one-hot vector and represents the class-label with D possible categories. Then, we define an overall model $F(\cdot)$. Finally, the output class $\hat{y}$ generated by the recognition model is formulated as:

$$\hat{y} = F(P, \mathbf{W}) \tag{1}$$

where W represents the trainable parameters of the overall model, and our goal is to dynamically update W to make $\hat{y}$ and y as close as possible. In this paper, we will pay more attention to the optimization of $F(\cdot)$ and further we detailed approach.

The function $F(\cdot)$ incorporates a Graph Convolutional Network architecture to process the spatio-temporal graph representations derived from the five latent feature streams. For each latent representation, the GCN operates through a series of graph convolutional layers that aggregate neighborhood information while preserving the skeletal topology. Specifically, given an input feature matrix $H^{(l)} \in \mathbb{R}^{N \times d_l}$ at layer l, where N represents the number of nodes and $d_l$denotes the feature dimension, the graph convolution operation is formulated as

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

where $\tilde{A} = A + I_N$ is the adjacency matrix with added self-connections, $\tilde{D}$ is the corresponding degree matrix with

$$\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}, W^{(l)} \in \mathbb{R}^{d_l \times d_{l+1}}$$

represents the learnable weight matrix, and $\sigma(\cdot)$ denotes the activation function. This spectral-based convolution effectively captures both local joint relationships and global skeletal patterns through successive message passing operations across the graph structure.

The multi-layer GCN architecture in $F(\cdot)$ progressively refines the feature representations through Lstacked convolutional layers, followed by a readout operation for global feature aggregation. The readout function $R(\cdot)$ combines node-level features into a graph-level representation through

$$h_G = R(\{h_i^{(L)} \mid i \in V\}),$$

where $h_i^{(L)}$represents the final hidden state of node i, and V denotes the vertex set. This can be implemented as a permutation-invariant pooling operation such as

$$h_G = \frac{1}{N}\sum_{i=1}^{N} h_i^{(L)}$$

for mean pooling or $h_G = \max_{i \in V}(h_i^{(L)})$ for max pooling. The aggregated features $h_G \in \mathbb{R}^d$ are subsequently processed through fully connected layers $\phi(\cdot)$ with dropout regularization, yielding the final action classification scores as

$$\hat{y} = \text{SoftMaxMCW}(\phi(h_G)),$$

where SoftMaxMCW represents the multi-connected weight softmax classifier that enhances the discriminative power of the model through enriched weight connectivity patterns.

3. Experiments

Table 1 shows the performance comparison of LFHAR with existing methods on the NTU-RGB+D 60 dataset. The LFHAR method demonstrates higher recognition accuracy on benchmark datasets compared to current approaches that combine GCN with attention mechanisms. The findings show a 1.5% performance gain over 2s-AGCN, which applies reinforcement learning for selecting joints dynamically. By comparison, our method employs multiple invariant representations that provide complete action encoding, allowing the system to adjust to dynamic changes without needing complex training processes.

Table 1. Comparison of accuracy between LFHAR-GCN-SoftMaxMCW and the other human action recognition methods on dataset.

Method	**Accuracy**	
	NTU-RGB+D 60	NTU-RGB+D 120
MST-GCN	87.5	88.8
Ta-CNN	85.7	87.3
ST-GCN	87.2	88.3
AS-GCN	95.2	94.2
EfficientGCN	92.8	96.1
RA-GCN	93.5	93.6
AGC-LSTM	92.1	95.0
2s-AGCN	95.4	95.1
LFHAR-GCN-SoftMaxMCW	96.9	97.2

4. Conclusion

This work proposes a framework for skeleton-based action recognition consisting of three key components: action representation, feature extraction, and action prediction modules. For feature extraction, a GCN+SoftMaxMCW architecture is applied to process both individual descriptors and their combined representations for extracting relevant features and performing classification.

References

1. Sun Z., Ke Q., Rahmani H., Bennamoun M., Wang G., Liu J. Human action recognition from various data modalities: A review IEEE Trans. Pattern Anal. Mach. Intell. (2022)

2. Ahmad T., Jin L., Zhang X., Lai S., Tang G., Lin L.. Graph convolutional neural network for human action recognition: A comprehensive survey. IEEE Trans. Artif. Intell., 2 (2) (2021), pp. 128-145

3. Cheng K., Zhang Y., He X., Chen W., Cheng J., Lu H., Skeleton-based action recognition with shift graph convolutional network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 183–192.

4. Aouaidjia K., Zhang C. and Pitas I., Spatio-temporal invariant descriptors for skeleton-based human action recognition, Inf Sci (NY), 700, 121832, doi: 10.1016/j.ins.2024.121832 (2025).