The Austrian Journal of Technical
and Natural Sciences, No 9–10
ISSN 2310-5607

PREMIER Publishing
ppublishing.org

DOI:10.29013/AJT-25-9.10-68-74

SPATIO-TEMPORAL LATENT FEATURES FOR SKELETON-BASED HUMAN ACTION RECOGNITION USING GCN+SOFTMAX CLASSIFIER

Avazjon Marakhimov [1], ***Kabul Khudaybergenov*** [2,3], ***Zakhriddin Mominov*** [4]

[1] Tashkent State Technical University, Tashkent, Uzbekistan;

[2] Kimyo International University in Tashkent, Tashkent, Uzbekistan;

[3] Research Institute for the Development of Digital Technologies
and Artificial Intelligence, Tashkent, Uzbekistan;

[4] Tashkent State University of Economics, Tashkent, Uzbekistan

Cite: *Avazjon M., Kabul K., Zakhriddin M. (2025). Spatio-temporal latent features for skeleton-based human action recognition using GCN+SoftMax classifier. Austrian Journal of Technical and Natural Sciences 2025, No 9–10.* https://doi.org/10.29013/AJT-25-9.10-68-74

Abstract

Human action recognition through skeletal analysis represents a fundamental challenge with significant implications for real-world applications. Contemporary approaches frequently depend on singular skeletal sequence representations, potentially limiting their capacity to comprehensively encode the multifaceted characteristics inherent in human actions. This work introduces LFHAR (Latent Features for Human Action Recognition), an innovative architectural framework that leverages diverse spatio-temporal latent encodings to enhance action feature extraction. The proposed representations model the temporal progression of skeletal configurations while incorporating both joint-level and limb-level motion patterns. The methodology employs a graph-based transformation for individual skeletal frames within temporal sequences, subsequently organizing the extracted graph features into spatio-temporal matrices. Experiments on benchmark datasets validates the robustness and invariance properties of the LFHAR framework. The approach achieves notable performance gains, with accuracy improvements of 2.7% and 2.1% on the NTU-RGB+D 60 and NTU-RGB+D 120 datasets, respectively, substantiating its effectiveness in advancing skeleton-based action recognition.

Keywords: *Invariant representations, Latent features, Skeleton-based action recognition, Spatio-temporal graph network*

1. Introduction

Human action recognition constitutes a critical component across diverse application domains, encompassing human-computer interaction systems, automated surveillance, robotic systems, medical applications, and immersive virtual environments. Research in this domain has explored multiple data modalities, spanning RGB video streams, depth-based recordings, and skeletal joint representations (Sun et. al., 2022). The skeleton-based paradigm has emerged as particularly compelling due to its inherent robustness against environmental variations including background clutter, attire differences, and illumination changes, while simultaneously offering high-fidelity 3D spatial coordinates of anatomical landmarks. Historical approaches to skeletal data acquisition relied on motion capture infrastructures employing body-mounted sensors, multi-view camera arrays, or infrared tracking systems within constrained laboratory settings (Ahmad et. al., 2021).

Recent breakthroughs in computer vision methodologies have catalyzed fundamental transformations across diverse sectors, encompassing medical diagnostics, financial modeling, predictive analytics, and visual computing. Deep learning architectures have particularly revolutionized automated hierarchical feature learning from visual data, enabling sophisticated pattern recognition capabilities that obviate traditional manual feature design. These advances have yielded remarkable achievements in visual recognition tasks including categorical classification, object localization, and semantic segmentation, often matching or exceeding human-level accuracy. Consequently, these technological developments have facilitated robust skeletal pose estimation directly from video streams through deep learning pipelines, eliminating dependencies on specialized sensor hardware (Cheng et. al., 2020).

The extraction of discriminative action-specific features from skeletal trajectories constitutes a fundamental challenge in skeleton-based recognition systems. Traditional approaches relied on handcrafted feature engineering, transforming skeletal sequences into compact representations suitable for conventional classifiers including K-Nearest Neighbor algorithms, Random Forest ensembles, or Hidden Markov Models.

2. Related works

To address the aforementioned constraints, this research introduces a novel architectural framework for skeleton-based action recognition, designated as LFHAR (**L**atent **F**eatures for **H**uman **A**ction **R**ecognition). The LFHAR architecture comprises three principal components: action representation, latent feature extraction, and human action prediction modules. Within the action representation component, we formulate five distinct image-based spatio-temporal action latent features that encode action sequences from complementary perspectives, thereby mitigating the inherent limitations associated with singular action representation schemes. Although numerous action representation paradigms exist, our proposed quintet of representations strategically emphasizes fundamental motion dynamics to achieve distinctive characterization of individual actions.

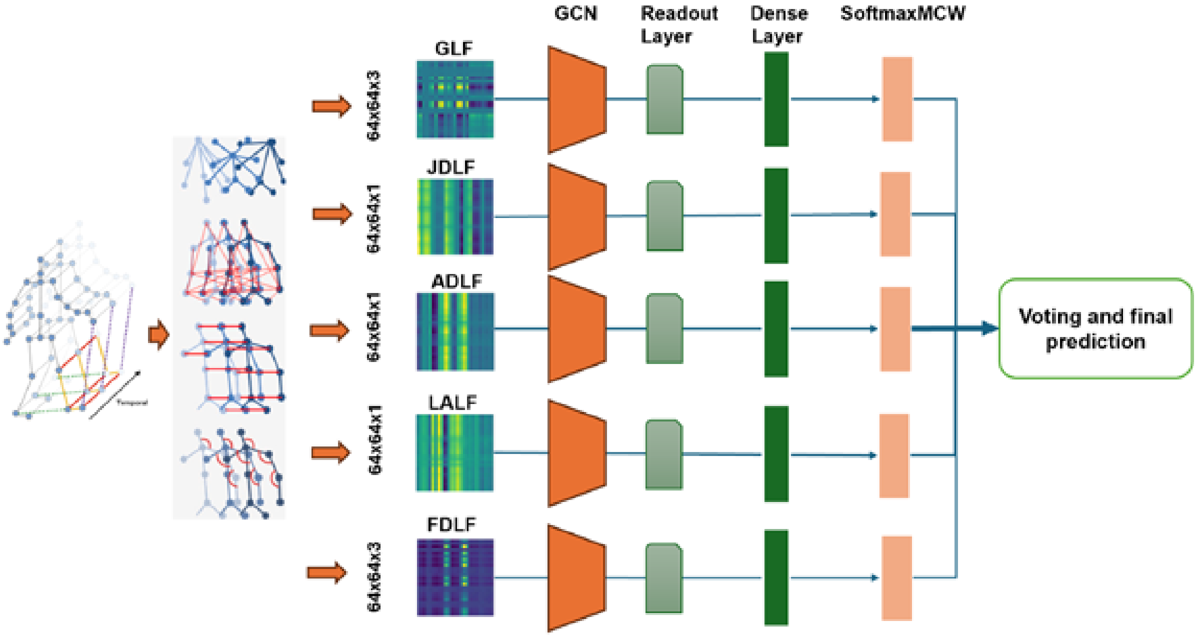
In examining human kinematic patterns, action differentiation emerges through temporal evolution of skeletal configurations and the directional characteristics of articular and limb trajectories. Directional motion attributes can be encoded through two complementary modalities: angular variations between interconnected limb segments and spatial displacement metrics between joint coordinates (Xin et. al., 2022). Furthermore, temporally similar actions executed at varying velocities manifest subtle disparities in both angular measurements and inter-joint spatial relationships, necessitating the development of velocity-invariant representational schemes to ensure robust action characterization across different execution speeds.

For encoding the temporal evolution of skeletal poses, we introduce a spatio-temporal Graph Latent Features (GF) representation, manifested as an RGB feature map derived from a graph matrix wherein columns encode individual skeletal graph representations across the three Cartesian dimensions. This representational scheme achieves coordinate invariance by emphasizing temporal variations in topological joint relationships rather than absolute positional coordinates. The underlying rationale for this approach stems from the over-smoothing phenomenon inherent in GCN-based architectures, where representational efficacy becomes contingent upon network depth (Liu et. al., 2025). In contrast, consolidating the complete graph sequence into a unified, fixed-dimensional image representation ensures computational stability.

Recognizing that graph representations inadequately capture joint motion dynamics, we propose three complementary distance-based latent features to encode spatio-temporal joint and limb displacement patterns throughout the action sequence. Each descriptor manifests as a single-channel grayscale image derived from distance matrices, with columns encoding inter-joint or inter-limb distances. The Joint Distance Latent Features (JDLF) encode intra-frame joint-to-joint distances across the temporal sequence. The Adjacent Distance Latent Features (ADLF) capture inter-frame joint displacements between consecutive temporal instances, providing visual encoding of action velocity that enables recognition of identical actions performed at varying speeds, thereby conferring velocity invariance. The Limbs Angle Latent Features (LALF) quantify angular variations between adjacent limb segments, additionally encoding joint directional information through temporal angular changes. To construct a comprehensive motion representation, these three distance latent features are integrated into a tri-channel Fusion Distance Latent Feature (FDLF). These four distance-based latent features exhibit view invariance through their exclusive reliance on relative distances between anatomical landmarks. Moreover, they normalize actions of heterogeneous temporal durations into fixed-dimensional image representations, ensuring frame-count invariance.

The feature extraction component employs specialized models for deriving features from the five latent representations (GLF, JDLF, ADLF, LALF, and FDLF). Given that these latent features manifest as texture-patterned feature maps (Fig. 1), we implement a shallow 1D CNN architecture optimized for pattern classification. Feature extraction proceeds independently for each latent representation through dedicated single-descriptor models for action prediction. Furthermore, features from all five representations undergo concatenation to construct a comprehensive and complementary action encoding via a fusion model. The prediction component generates six classification outputs: five from individual descriptor models and one from the fusion architecture. Final class determination employs a voting algorithm that implements majority consensus when multiple models produce concordant predictions.

Figure 1. *Architectural overview of the LFHAR framework, comprising three integrated modules*



The action representation module performs transformation of input skeletal sequences into five distinct image-based latent representations: Graph Latent Fea-

tures (GLF), Joint Distance Latent Features (JDLF), Adjacent Distance Latent Features (ADLF), Limbs Angle Latent Features (LALF), and Fusion Distance Latent Features (FDLF). The latent feature extraction module employs a GCN architecture followed by Readout layer operations. After then the output from this layer is processed by dense layer and for classification is used SoftMaxMCW (Marakhimov et. al., 2025). *The action prediction module generates classification scores from six feature streams – encompassing the five independent latent representations and their concatenated fusion – utilizing one-dimensional convolutional layers (Conv1D). A voting-based ensemble mechanism subsequently determines the final classification output through consensus aggregation of the multiple prediction streams.*

3. Methodology

The framework of LFHAR-GCN-SoftMaxMCW architecture is depicted in Fig. 1. It mainly consists of three separate stages: human skeleton action representation, latent feature extraction, and action classification.

We use *Human action representation, Skeleton graph matrix* and *Joints distance matrix* from work (Aouaidjia et. al., 2025). However, authors build very heavy models which is impossible to employ in real-time human action recognition. Thus we build our model based on (Marakhimov et. al., 2025) to get the faster human skeleton based action recognition.

3.1. Human action representation

We define the skeleton sequence as $Seq = \{S_k, k = 1,...,T\}$, where T is the number of frames, and the skeleton S_k as a graph $S_k = \{V, E\}$, $V = \{\{\{J_{i,j}\}_{i=1}^{N}\}_{j=1}^{3}\}, J_{i,j}$ is the coordinate *j* of the joint *i,* and *N* is the number of joints (vertices). We normalize the skeleton sequence by considering the "*middle of the spine*" joint of the first frame as the new origin of the coordinates system. In the rest of the coming sections, we use the term "*limb*" to refer to a skeleton bone that forms an angle with its adjacent bone.

3.1.1. Skeleton graph matrix

For each skeleton S_k represented as $N \times 3$ matrix, the symmetric adjacency matrix is defined as $A = \{a_{i,j}, i, j = 1,...,N\}$, where $a_{i,j} = 1$ if the joints i and j are connected, and $a_{i,j} = 0$, otherwise. The normalized adjacency matrix $\mathbf{A}'$ is calculated by dividing each row by the sum of its values. The graph $\mathbf{G}_k$ of a single skeleton $\mathbf{S}_k$ is obtained by:

$$\mathbf{G}_k = \mathbf{S}_k^t \mathbf{A}' = \begin{bmatrix} j_{11}\, j_{21} \cdots j_{N1} \\ j_{12}\, j_{22} \cdots j_{N2} \\ j_{13}\, j_{23} \cdots j_{N3} \end{bmatrix} \cdot \begin{bmatrix} a'_{11} & a'_{12} & \cdots & a'_{N1} \\ a'_{21} & a'_{22} & \cdots & a'_{N2} \\ \vdots & \vdots & \ddots & \vdots \\ a'_{N1} & a'_{N2} & \cdots & a'_{NN} \end{bmatrix} = \begin{bmatrix} g_{11}\, g_{21} \cdots g_{N1} \\ g_{12}\, g_{22} \cdots g_{N2} \\ g_{13}\, g_{23} \cdots g_{N3} \end{bmatrix}, \tag{1}$$

$$a'_{i,j} = \frac{a_{i,j}}{\sum_{j=1}^{N} a_{i,j}}, \tag{2}$$

where t represents the matrix transpose operation. The spatio-temporal graph representation of the sequence is given by the 3D graph matrix $M^{gr} \in R^{N \times 3 \times T}$ as:

$$\mathbf{M}^{gr} = \{\mathbf{G}_k^t\}_{k=1}^{T} = \left\{\left\{\left\{g_{i,j,k}\right\}_{i=1}^{N}\right\}_{j=1}^{3}\right\}_{k=1}^{T} = \begin{bmatrix} g_{1,j,1} & g_{1,j,2} & \cdots g_{N,j,T} \\ g_{21} & g_{2,j,2} & \cdots g_{N,j,T} \\ \vdots & \vdots & \ddots \quad \vdots \\ g_{N1} & g_{N,j,2} & \cdots g_{N,j,T} \end{bmatrix}, \tag{3}$$

where $j = 1,2,3$, $g_{i,j,k}$ are the features of the joint i of the coordinate k of the skeleton k. $\mathbf{M}_{gr}$ is a 3D matrix consists of three 2D matrices, corresponding to $j = 1,2,3$, which represents the three Cartesian axes. The construction of the graph matrix is illustrated visually in Fig. 2(a).

3.1.2. Joints distance matrix

One of the key factors for a robust motion representation is to capture the inter-joints distance evolution over time. For example, in the action '*clapping*', it is important to know how far the left-hand joint is moving from the right-hand joint. To construct a joint distance representation, we selected the 32 most informative pairs that have a higher changeable rate during the motion than the other pairs, where the joints involved in the selected pairs.

Given a skeleton S_k, we define the set of the selected pairs as: $Pair = \{p_i, i = 1,\ldots,P\}$, where P is the total number of selected pairs, and $p_i = (J^a, J^b)$ is the pair of the joints J^a and J^b. For each pair p_i, the Euclidean distance d_i is calculated between its two joints as: $d_i = \left\| J^a - J^b \right\|_2$. However, the distance between two joints in a skeleton of a shorter person is less than that of a taller person, due to the difference in limbs length. To normalize the distance d_i of the pair p_i for all body sizes, we divide the distance d_i by the body size z, where the normalized distance $d'_i = d_i / z$, and the body size $z = \sum_{i=1}^{N-1} L_i$, which is the sum of the lengths of all the skeleton limbs L_i, and the limb length is the distance between its joints. The joints distance matrix $M_{jdis} \in R^{P\times T}$ is defined as:

$$\mathbf{M}^{jdis} = \left\{ \left\{ \frac{d_i}{z} \right\}_{i=1}^{P} \right\}_{k=1}^{T} = \left\{ \left\{ d'_{i,k} \right\}_{i=1}^{P} \right\}_{k=1}^{T} = \\ = \begin{bmatrix} d'_{11} \; d'_{12} \cdots d'_{P1} \\ d'_{21} \; d'_{22} \cdots d'_{P2} \\ \vdots \quad \vdots \quad \ddots \quad \vdots \\ d'_{P1} \; d'_{P2} \cdots d'_{PN} \end{bmatrix}, \tag{4}$$

where $d'_{i,k}$ is the normalized distance of the pair p_i of the skeleton S_k.

3.1.3. Adjacent distance matrix

The adjacent distance matrix represents the temporal change in the joint coordinates values between each two consecutive frames. In other words, the distance between the joint in a frame k and its new position in the frame $k+1$. It also represents action velocity, where larger distances indicate faster movement of the joints. Given two consecutive skeletons S_k and S_{k+1} of the sequence, the adjacent Euclidean distance a_i, k of a joint i between two consecutive frames k and $k+1$ is written as: $a_{i,k} = \left\| J_{i,k} - J_{i,k+1} \right\|_2$. The Adjacent Distance Matrix $\mathbf{M}^{adis} \in R^{N\times(T-1)}$ is defined as follows:

$$\mathbf{M}^{adis} = \{\{a_{i,k}\}_{i=1}^{N}\}_{k=1}^{T-1} = \{\{\left\| J_{i,k} - J_{i,k+1} \right\|_2\}_{i=1}^{N}\}_{k=1}^{T-1} \tag{5}$$

$$\begin{bmatrix} \left\|J_{1,1} - J_{1,2}\right\|_2 & \left\|J_{1,2} - J_{1,3}\right\|_2 & \cdots & \left\|J_{1,T-1} - J_{1,T}\right\|_2 \\ \left\|J_{2,1} - J_{2,2}\right\|_2 & \left\|J_{2,2} - J_{2,3}\right\|_2 & \cdots & \left\|J_{2,T-1} - J_{2,T}\right\|_2 \\ \vdots & \vdots & \ddots & \vdots \\ \left\|J_{N,1} - J_{N,2}\right\|_2 & \left\|J_{N,2} - J_{N,3}\right\|_2 & \cdots & \left\|J_{N,T-1} - J_{N,T}\right\|_2 \end{bmatrix}$$

4. Experiments

Table 1 presents the comparative results between LFHAR and state-of-the-art approaches on the NTU-RGB+D 60 dataset. The proposed LFHAR method achieves superior recognition accuracy on benchmark datasets and outperforms existing methods that integrate GCN with attention mechanisms. The results indicate a 1.5% improvement over 2s-AGCN, which employs reinforcement learning for dynamic joint selection. In contrast, our approach utilizes multiple invariant representations that offer comprehensive action encoding, enabling adaptation to dynamic variations without requiring complex training procedures.

Table 1. *Comparison of accuracy between LFHAR-GCN-SoftMaxMCW and the other human action recognition methods on the NTU-RGB+D 60 and NTU-RGB+D 120 dataset*

Method	Accuracy	
	NTU-RGB+D 60	**NTU-RGB+D 120**
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

Table 2 presents a computational complexity analysis comparing the proposed LFHAR-GCN-SoftMaxMCW framework with existing action recognition methods, evaluated using Floating Point Operations Per Second (FLOPS) and inference time measured in sequences per second. The computational efficiency assessment reveals that the single-descriptor configuration achieves 7.8 FLOPS, positioning it as the second most efficient model in the comparison. The model, which processes the concatenated representations of all five descriptors, demonstrates a computational requirement of 31.6 FLOPS. This represents approximately a five-fold increase compared to the single-descriptor variant, which can be attributed to the additional computational overhead of processing the combined feature representations. Given that the complete framework employs multiple models operating in parallel, the overall computational cost is determined by aggregating the FLOPS values of all constituent models, yielding a cumulative value of 25.3 FLOPS for the entire system.

These computational metrics indicate that while the fusion approach incurs higher computational costs due to feature concatenation, the overall framework maintains reasonable efficiency when considering the performance gains achieved through multi-descriptor integration. The inference time measurements further support the practical viability of the proposed method for real-world action recognition applications.

Table 2. *Comparison of FLOPS and Inference time.*

Method	FLOPS	Inference time
ST-GCN	16.3	42.91
AS-GCN	14.5	18.50
EfficientGCN	17.5	23.4
RA-GCN	22.4	19.52
AGC–LSTM	9	30.3
2s-AGCN	11	12.90
LFHAR-GCN-SoftMax-MCW	25.3	17.50

5. Conclusion

This study presents a novel framework for skeleton-based action recognition that comprises three main components: action representation, feature extraction, and action prediction modules. The action representation module employs five image descriptors to encode skeleton sequences. One descriptor captures the spatio-temporal relationship variations among joints throughout the temporal dimension, while the remaining four descriptors encode the distance variations between joints and limbs over time. These descriptors function as latent features that maintain consistency across different poses, viewing angles, and movement velocities, thereby ensuring robust action representation. For the feature extraction component, we implement a GCN+SoftMaxMCW architecture that processes both individual descriptors and their combined representations to extract relevant features and perform classification. The action prediction module generates six class predictions, which are subsequently processed through a voting mechanism to determine the final action category.

We evaluated the proposed framework using four standard benchmark datasets. The experimental results, along with comprehensive ablation studies, confirm the efficacy of our approach. The comparative analysis with existing methods demonstrates that our framework achieves competitive performance in skeleton-based action recognition tasks. These findings suggest that the integration of multiple image descriptors with graph-based feature extraction provides an effective solution for robust action recognition.

Acknowledgements

The research is supported in part by Grants No. IL-5421101773 of the Uzbekistan Ministry for Innovative Development.

References

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

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

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

Xin W., Liu Y., Liu R., Miao Q., Shi C., Pun C.-M. (2023). Auto-learning-gcn: An ingenious framework for skeleton-based action recognition. Chinese Conference on Pattern Recognition and Computer Vision, PRCV, Springer, – P. 29–42.

Liu R., Liu Y., Wu M., Xin W., Miao Q., Liu X., Li L. (2025). SG-CLR: Semantic representation-guided contrastive learning for self-supervised skeleton-based action recognition. Pattern Recognit., – 162. – Article 111377.

Marakhimov, A.R., Khudaybergenov, K.K. (2025). Softmax Regression with Multi-Connected Weights. Computers, accepted.

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

submitted 23.09.2025;
accepted for publication 07.10.2025;
published 26.11.2025

Contact: kabul.kudaybergenov@gmail.com