Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 14, No. 4, 2026, 577-585.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

Efficient Frequency-Aware Skeleton Action Recognition for Elderly Monitoring

Fateme Naghavi and Kourosh Kiani*

Electrical and Computer Engineering Department, Semnan University, Semnan, Iran.

Article Info

Article History:
Received 06 September 2025
Revised 25 February 2026
Accepted 01 July 2026

DOI:10.22044/jadm.2026.16786.2810

Keywords:
Human Activity Recognition, Elderly Daily Activity, Monitoring daily activities, CNN, Spatio_Temporal Feature, Fast Fourier Transform.

Corresponding author: Kourosh.kiani@semnan.ac.ir (K. Kiani).

Abstract

Human Activity Recognition (HAR) based on skeletal representations has become a promising solution for elderly monitoring due to its robustness to illumination changes and privacy preservation. However, achieving high recognition accuracy while maintaining real-time performance on resource-constrained edge devices remains a significant challenge. This paper presents a lightweight skeleton-based HAR framework for elderly activity monitoring. The proposed model integrates multi-view motion representations, including absolute joint coordinates, short-term motion, long-term motion, and spatial relationship features. To enhance temporal motion discrimination, we introduce a Temporal Frequency Attention (TFA) module that performs Fast Fourier Transform (FFT) exclusively along the temporal dimension and adaptively recalibrates feature responses according to their frequency-domain energy distributions. The frequency-enhanced representations are subsequently processed by multi-scale depth-wise temporal convolutions to capture motion patterns across different temporal ranges with minimal computational overhead. The resulting architecture contains only 0.37M trainable parameters and is specifically designed for efficient deployment on edge platforms. Owing to its favorable accuracy–efficiency trade-off, the proposed framework provides a practical solution for real-time elderly activity monitoring in resource-constrained environments.

1. Introduction

The rapid growth of the global elderly population has increased the demand for intelligent healthcare technologies capable of continuously monitoring daily activities and detecting abnormal behaviors [1]. Age-related physical decline significantly elevates the risk of falls, mobility impairments, and other health complications, making timely activity monitoring a critical component of modern elderly care systems [2]. Accurate recognition of daily activities such as walking, sitting, standing, eating, and transitional movements can provide valuable insights into an individual's physical condition, functional independence, and overall well-being. Consequently, human action activity has become an important research topic in healthcare, ambient assisted living, and smart home environments [3].

Among various sensing modalities, skeleton-based activity recognition has gained considerable attention due to its ability to represent human motion in a compact and privacy-preserving manner [4,5]. Unlike RGB-based approaches [6,7], skeletal representations are largely insensitive to illumination variations, background clutter, clothing appearance, and camera viewpoints. Furthermore, skeleton sequences contain explicit structural information regarding human body joints and their temporal evolution, making them particularly suitable for activity analysis in indoor healthcare environments. The availability of large-scale benchmark datasets such as ETRI-Activity3D has further accelerated the development of skeleton-based Human Activity Recognition

(HAR) methods [8-12]. Recent advances in skeleton-based activity recognition have been dominated by graph neural networks (GNNs), spatiotemporal graph convolutional networks, transformer architectures, and hybrid deep learning frameworks [13–23]. These approaches have achieved remarkable recognition performance by modeling complex spatial and temporal dependencies between body joints. However, their success is often accompanied by substantial computational complexity, increased memory consumption, and high inference latency. Such requirements may be acceptable in server-based environments but pose significant challenges for real-world elderly monitoring systems, where models must operate continuously on resource-constrained edge devices with strict real-time constraints [24].

In addition to computational complexity, another limitation of many existing methods is their predominant focus on temporal modeling in the time domain. Human activities inherently contain characteristic motion frequencies that reflect repetitive movement patterns and temporal dynamics. For example, activities such as walking, exercising, or performing daily routines exhibit distinct frequency signatures that may not be fully captured by conventional temporal convolutions alone. Despite the growing success of deep learning approaches, the explicit exploitation of frequency-domain information remains relatively underexplored in lightweight skeleton-based HAR systems, particularly for edge-oriented elderly monitoring applications. To address these challenges, this paper proposes a lightweight skeleton-based activity recognition framework that combines efficient temporal modeling with frequency-aware feature enhancement. The proposed architecture first integrates multiple complementary skeletal representations, including absolute joint coordinates, short-term motion, long-term motion, and spatial relationship features. A Temporal Frequency Attention module is then introduced to analyze motion dynamics in the frequency domain by applying Fast Fourier Transform exclusively along the temporal dimension. The extracted frequency responses are used to adaptively recalibrate feature representations, enabling the network to emphasize discriminative motion patterns while preserving computational efficiency. Subsequently, multi-scale depth-wise temporal convolutions capture activity dynamics at different temporal resolutions, providing an effective balance between recognition performance and computational cost.

The resulting framework is specifically designed for deployment in elderly monitoring systems operating on edge devices. By combining lightweight convolutional operations with frequency-aware temporal modeling, the proposed method achieves strong recognition performance while maintaining low computational requirements and high inference speed.

The main contributions of this work are summarized as follows:

1. We propose a lightweight skeleton-based HAR architecture tailored for real-time elderly activity monitoring on resource-constrained edge devices.
2. We introduce a Temporal Frequency Attention module that performs frequency-domain analysis through temporal FFT and adaptively enhances discriminative motion representations.

Extensive experiments on the ETRI-Activity3D datasets demonstrate that the proposed framework achieves a favorable balance between recognition accuracy, computational efficiency, and model compactness. The results highlight its suitability for real-time elderly activity monitoring on resource-constrained edge devices.

The remainder of this paper is organized as follows. Section 2 reviews related work in skeleton-based activity recognition and frequency-aware deep learning methods. Section 3 describes the proposed framework in detail. Section 4 presents the experimental setup and results. Finally, Section 5 concludes the paper and discusses future research directions.

2. Related Work

Skeleton-based action recognition has recently emerged as a prominent research area within computer vision and human–computer interaction. Prior studies in this domain can be broadly categorized into two main directions: general human motion modeling approaches and elderly-focused activity recognition methods.

2.1. General Human Action Recognition Approaches

Early works introduced graph convolutional networks to represent skeleton data as spatiotemporal graphs, achieving significant success in action recognition. However, these models suffered from challenges such as high computational complexity and limited receptive fields. To address these limitations, more efficient models such as Shift-GCN [25] were proposed, leveraging shift operations and lightweight

convolutions to reduce complexity while enabling more flexible spatiotemporal feature aggregation**.** With the advent of transformers, researchers further explored modeling long-range temporal dependencies and spatial correlations among joints through self-attention mechanisms. Approaches such as TranSkeleton [17] and Hyperformer [23] successfully integrated skeletal topology into transformer-based architectures, overcoming the constraints of conventional GCNs while improving both accuracy and efficiency. Other methods, including SkateFormer [20], ST-TR [18], and STTFormer [19], introduced multi-branch spatiotemporal attention modules and higher-order joint relation modeling, achieving state-of-the-art performance on large-scale benchmarks such as NTU RGB+D. This research trend highlighted the potential of combining the strengths of GCNs and transformers—or designing lighter attention-based modules—to achieve a better balance between accuracy, computational efficiency, and generalization.

2.2. Elderly-Oriented Activity Recognition Methods In response to the global rise in the aging population and the growing need for daily activity monitoring, dedicated datasets such as ETRI-Activity3D have been introduced [1]. This dataset, which contains over 100,000 samples covering a wide range of elderly daily activities, has spurred extensive research in this field. Several studies have leveraged multimodal data (RGB, depth, and skeleton) to improve recognition robustness. For instance, FSA-CNN [1] by employing four parallel streams with adaptive mechanisms, demonstrated increased resilience to spatiotemporal variations. Similarly, approaches such as ESE-FN [26] utilized modality- and channel-level attention mechanisms to achieve more effective multimodal feature fusion. Other works focused on hybrid model designs. For example, THGTF-Net [27] combined GCNs and transformers to jointly exploit joint- and bone-level representations, thereby addressing challenges such as inter-class motion similarity and human–object interactions. Furthermore, recent studies have moved toward multimodal hybrid architectures that integrate geometric, motion, and semantic features, enabling more accurate recognition of daily living activities and even unseen ADLs.

2.3. Summary

Overall, prior research can be summarized into three major trajectories:

- Optimization of GCN-based architectures to reduce complexity and enhance flexibility.
- Utilization of transformer models to capture long-term dependencies and higher-order joint relationships.
- Development of elderly-focused, multimodal, and hybrid architectures incorporating attention mechanisms to manage human–object interactions and motion similarity.

Despite substantial progress, an open challenge remains: designing a lightweight and efficient model that achieves a favorable trade-off between accuracy and computational cost, while ensuring feasibility for real-world, real-time applications, particularly in elderly care scenarios.

3. Proposed Method

3.1. Overall Architecture

The proposed network is designed as a lightweight yet effective framework for skeleton-based human activity recognition, with a particular emphasis on elderly activity monitoring. Unlike many recent approaches that rely on graph neural networks or transformer architectures, the proposed method adopts a lightweight convolutional design while incorporating frequency-domain motion analysis to enhance temporal feature representation. After preprocessing and feature construction, the four complementary skeletal representations, namely normalized joint coordinates, short-term motion, long-term motion, and spatial relational features, are concatenated along the feature dimension to form a unified input representation. This early-fusion strategy allows the network to jointly exploit spatial structure and motion dynamics while maintaining a compact architecture. An overview of the proposed architecture is illustrated in Figure 1.

3.2. Data Preprocessing Prior

To feature extraction, all skeleton sequences undergo a preprocessing pipeline to mitigate inter-subject variability and ensure a consistent spatio-temporal representation, following the protocol established in [1]. To achieve temporal uniformity, each sequence is standardized to a fixed length. For sequences longer than the target length, a random temporal sampling strategy is employed, whereby a fixed number of frames are randomly selected from the original sequence while preserving their chronological order. Specifically, the sampled frame indices are sorted after selection to maintain the temporal progression of the activity. For shorter sequences, frame repetition (padding) is applied by replicating the final available frame until the desired sequence length is reached.

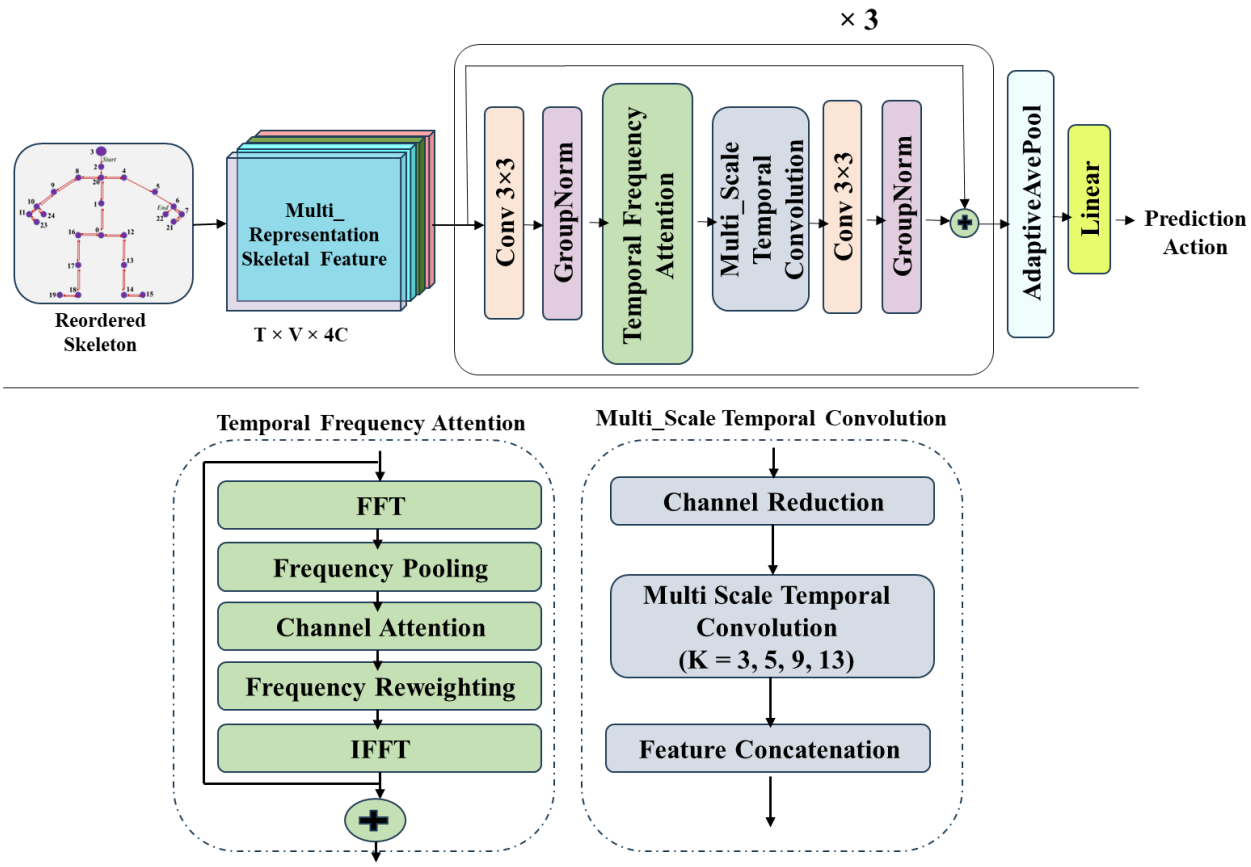

Figure 1. Architecture of the proposed frequency-aware multi-scale temporal network. Multi-view skeletal features are processed through residual convolutional blocks equipped with Temporal Frequency Attention (TFA) and Multi-Scale Temporal Convolution (MSTC) modules, followed by adaptive average pooling and linear classification. T, V, and C denote the temporal length, number of reordered joints, and feature dimensionality, respectively.

To eliminate the influence of global positioning, joint coordinates are translated relative to the root joint (base of the spine) in the central frame of each sequence. This transformation aligns all samples within a common coordinate system, thereby reducing the model's sensitivity to the subject's absolute location within the sensor's field of view. Furthermore, to ensure robustness against anthropometric variations such as body size and stature, the aligned coordinates are normalized using a skeletal scale factor derived from the spinal length in the reference frame. This normalization renders the representation invariant to physical differences while preserving the essential geometric relationships between joints.

To optimize the exploitation of spatial dependencies through convolutional operations, the original joint ordering provided by the dataset is reorganized according to the anatomical topology of the human body, as illustrated in Figure 2. Instead of following the default indexing scheme, joints are rearranged along continuous kinematic chains representing the head, torso, upper limbs, and lower limbs.

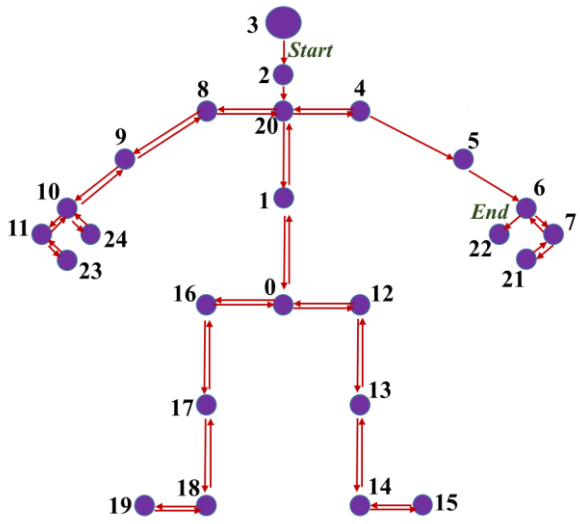

Figure 2. Sequential joint ordering from joint 3 to joint 22, following the indicated arrows.

This reordering transforms the skeletal representation into a topology-preserving sequence

in which adjacent positions correspond to anatomically connected body parts. Consequently, neighboring elements in the reordered representation exhibit meaningful physical relationships, enabling convolutional kernels to capture local structural patterns more effectively.

3.3. Multi-Representation Skeletal Feature Construction

To provide a richer description of human motion, four complementary skeletal feature representations adopted from prior work [1] are utilized: normalized joint coordinates, short-term motion, long-term motion, and spatial relational features. These representations are concatenated along the feature dimension to construct a unified multi-representation input, which serves as the input to the proposed network.

The first representation consists of the normalized three-dimensional joint coordinates. This stream preserves the instantaneous spatial configuration of the human body and provides detailed information about body posture at each time step. Such spatial cues are particularly important for distinguishing activities characterized by similar motion dynamics but different body poses. The joint representation is defined as:

$$F_{Jo\,\mathrm{int}}(t,i) = J_t^i \tag{1}$$

where $J_t^i = [\mathrm{x}, \mathrm{y}, \mathrm{z}]$ denotes the three-dimensional coordinates of joint i at frame t. The second representation captures short-term temporal dynamics by computing frame-to-frame joint displacements throughout the sequence. This representation emphasizes local motion variations and reflects instantaneous changes in joint positions, effectively highlighting rapid movements and subtle transitions between consecutive postures. To preserve a consistent sequence length, zero padding is applied to the final frame.

$$F_{Vel}(t,i) = J_{t+1}^i - J_t^i \tag{2}$$

While short-term motion descriptors are effective for capturing local dynamics, many daily activities performed by older adults evolve over longer temporal intervals. Therefore, a third representation is generated by computing joint displacements over a wider temporal window spanning five frames. Compared with frame-to-frame differences, this long-term motion representation suppresses small fluctuations and measurement noise while providing a more stable description of overall movement direction and progression. Consequently, it facilitates the recognition of activities characterized by gradual or sustained body motion.

$$F_{Long}(t,i) = J_{t+5}^i - J_t^i \tag{3}$$

The fourth representation explicitly models structural relationships within the human skeleton. Following the joint reordering procedure described previously, spatial differences are computed between adjacent joints along the anatomically organized skeletal sequence. This representation encodes local limb geometry, relative joint orientation, and body-part connectivity, thereby preserving important structural information that may not be fully captured by absolute joint coordinates alone.

$$F_{Bone}(t,i) = J_t^{i+1} - J_t^i \tag{4}$$

where i and $i+1$ denote two consecutive joints in the reordered skeletal sequence. Finally, the four feature representations are concatenated along the channel dimension to:

$$F_{input} = Concat(F_{Jo\,\mathrm{int}} + F_{Vel} + F_{Long} + F_{Bone}) \tag{5}$$

which is subsequently fed into the proposed network for feature learning and action classification.

3.4. Proposed Network Architecture

The backbone network consists of a series of residual convolutional blocks with progressively increasing channel dimensions. Each block employs standard convolutional operations for local feature extraction, together with residual connections to facilitate stable optimization and efficient information propagation. To enrich temporal modeling, the deeper stages of the network incorporate a Temporal Frequency Attention (TFA) module.

Unlike conventional temporal modeling approaches that operate solely in the time domain, TFA transforms temporal features into the frequency domain using a one-dimensional Fast Fourier Transform applied exclusively along the temporal axis. Frequency-domain responses are then adaptively recalibrated through a lightweight attention mechanism, enabling the network to emphasize informative motion frequencies while suppressing less relevant temporal components. The refined frequency representation is subsequently transformed back to the temporal domain and integrated with the original features through a residual formulation.

Following frequency-aware refinement, temporal dependencies are further modeled using a multi-scale temporal convolution module composed of multiple depthwise convolutional branches. Each

branch employs a different temporal receptive field, allowing the network to simultaneously capture short-, medium-, and long-range motion patterns. The outputs of all branches are fused through a pointwise convolution, producing a compact multi-scale temporal representation while maintaining low computational complexity.

After feature extraction, global average pooling is applied to aggregate joint- and time-level information into a compact feature vector for each subject. Finally, a dropout layer and a fully connected classifier produce the activity prediction.

Overall, the proposed architecture combines lightweight convolutional modeling, multi-scale temporal feature extraction, and frequency-domain attention within a unified framework. This design achieves an effective balance between recognition capability and computational efficiency, making it particularly suitable for real-time elderly activity monitoring on edge devices.

4. Experimental Results

In this section, we evaluate the performance of the proposed model.

4.1. Datasets

4.1.1. ETRI activity3D

The ETRI-Activity3D dataset was collected using Kinect v2 sensors and comprises three data modalities: RGB video, depth maps, and skeleton sequences. The skeleton sequences capture the 3D positions of 25 body joints of the tracked individuals. This dataset includes a total of 55 action categories, of which 52 are derived from observations of daily activities (such as eating, cleaning, and reading) performed by elderly individuals, while the remaining three categories pertain to human-robot interactions (including waving, pointing, and beckoning). Among these actions, five involve mutual interactions, such as handshaking and hugging. To select the activities, the homes of 53 elderly individuals over 70 years old were closely monitored, and their daily activities from morning to night were documented. Subsequently, 55 frequently occurring activities were identified.

The study involved 100 participants, consisting of 50 elderly individuals and 50 young adults. The ages of the elderly participants ranged from 64 to 88 years, with an average age of 77 years. In contrast, the young adults were in their 20s, with an average age of 23 years. Among the elderly participants, there were 17 men and 33 women, while among the young adults, there were 25 men and 25 women. The training set was composed of data from 67 subjects, whereas the test set included data from the remaining 33 subjects. The subject IDs in the test set are {3, 6, 9, 12, …, 99}, with all other IDs designated for the training set.

Figure 3. Examples of daily actions in ETRI-Activy3D

4.2. Implementation Details

The proposed model was implemented in PyTorch and trained on a single NVIDIA H100 GPU. Optimization was performed using the AdamW optimizer with an initial learning rate of 3×10^{-4} and a weight decay of 1×10^{-4}. The learning rate was dynamically adjusted using a cosine annealing schedule throughout training. The network was trained for 150 epochs with a mini-batch size of 8. To improve generalization and reduce overfitting, a dropout layer with a dropout rate of 0.3 was applied before the final classification layer. Cross-entropy loss was adopted as the training objective. All convolutional layers were followed by Group Normalization and ReLU activation functions. For reproducibility, random seeds were fixed for all experiments, and deterministic training settings were enabled. Following the standard protocol, all skeleton sequences were temporally normalized to a fixed length of 100 frames before feature extraction. Model selection was performed based on the highest validation accuracy, and the corresponding checkpoint was retained for final evaluation.

4.3. Evaluation Metric

To evaluate the performance of the proposed framework, classification accuracy was employed as the primary evaluation metric. Accuracy is one of the most widely used measures in Human Activity Recognition and represents the proportion of correctly classified samples among all test instances. It provides an overall assessment of the model's recognition capability across the entire dataset. The overall classification accuracy is calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

4.4. Comparison with State-of-the-Art Models

To assess the effectiveness of the proposed framework, its performance was compared with representative skeleton-based activity recognition methods on the ETRI-Activity3D dataset, as summarized in Table 1. It should be noted that the

ETRI-Activity3D dataset was specifically designed for elderly activity recognition, and many existing studies on this benchmark employ multimodal information, particularly the combination of RGB and skeletal modalities. Consequently, the number of directly comparable skeleton-only methods is relatively limited. For this reason, our comparisons focus on methods that report results using skeletal representations and can therefore provide a fair evaluation of the proposed approach.

As shown in Table 1, the proposed method achieves an accuracy of 90.1%, outperforming ESE-FN (88.1%) and approaching the performance of FSA-CNN (90.6%). More importantly, this performance is obtained using only 0.37 million parameters, which is substantially lower than FSA-CNN (2.7M parameters) and slightly lower than ESE-FN (0.42M parameters). Although the proposed model requires 6.3 GFLOPs, its computational complexity remains sufficiently low for practical deployment while providing improved recognition performance.

The observed performance gain can be attributed to the complementary feature representation strategy adopted in this work. Instead of relying solely on raw joint coordinates, the proposed framework simultaneously exploits four heterogeneous skeletal representations, namely joint coordinates, short-term motion, long-term motion, and spatial relational features. These representations provide complementary spatial and temporal information at different levels of abstraction and are fused before feature extraction. Consequently, the network receives a richer motion description without requiring a substantial increase in model capacity.

The ablation study presented in Table 2 further validates this observation. Starting from a lightweight spatial CNN baseline, the incorporation of temporal convolution improves the recognition accuracy from 87.4% to 88.7%. Subsequently, the introduction of the proposed frequency-aware temporal attention mechanism further increases the accuracy to 90.1%. These results demonstrate that each component contributes positively to the final performance and that the proposed architecture benefits from jointly modeling spatial, temporal, and frequency-domain information.

Table 1. Comparison of Proposed method's performance against current top models on ETRI-Activity3D.

Methods	X-Sub	Parameters(M)	FLOPs(G)
ESE_FN[20]	88.1	0.42	2.5
FSA-CNN[1]	90.6	2.7	3.3
Ours	**90.1**	0.37	6.3

Table 2. Ablation Study for Proposed Method with Different Components.

Baseline				Accuracy (%)
Spatial CNN	✓			87.4
Temporal CNN	✓	✓		89.6
Frequency Temporal Attention	✓	✓	✓	**90.1**

To evaluate the robustness of the proposed framework, the complete training and evaluation procedure was independently repeated ten times using different random initialization seeds while maintaining identical hyperparameters and experimental settings. The proposed model achieved an average accuracy of 90.10% ± 0.38%, indicating a highly stable behavior with very limited performance variation across runs.

From an efficiency perspective, direct inference-time comparisons with previously published methods are difficult because most studies do not report inference latency, hardware specifications, software environments, or implementation details. Since inference time is highly dependent on hardware configuration and software optimization, a fair comparison would require re-implementing all competing methods under identical experimental conditions.

Therefore, we report the inference time of the proposed model on our experimental platform. The average inference time is approximately 24 ms per sample, demonstrating that the proposed framework can process activity sequences with low latency while maintaining competitive recognition accuracy. This characteristic is particularly important for practical elderly-monitoring systems, where timely activity recognition is often more critical than achieving marginal improvements in classification accuracy.

Overall, although the proposed method is not intended to surpass large-scale transformer-based architectures in terms of raw recognition accuracy, it achieves an attractive balance between accuracy, model compactness, computational efficiency, stability, and inference time. These characteristics make it particularly suitable for real-world elderly activity monitoring applications, where

computational resources are often limited and reliable real-time operation is essential.

5. Conclusion

In this study, a lightweight frequency-aware convolutional framework was developed for skeleton-based human activity recognition, with a particular emphasis on elderly activity monitoring. The proposed approach combines four complementary skeletal feature representations, including joint coordinates, short-term motion, long-term motion, and spatial relational information, within a unified learning framework. These heterogeneous representations provide complementary spatial and temporal cues that enable the network to capture both posture-related and motion-related characteristics of daily activities. The proposed architecture integrates convolutional feature extraction, multi-scale temporal modeling, and temporal frequency attention while maintaining a compact network structure. Experimental results on the ETRI-Activity3D dataset demonstrate that the framework is capable of achieving reliable recognition performance using a relatively small number of trainable parameters. The ablation study further confirms that each component contributes positively to the final performance, with additional gains obtained through the incorporation of temporal-frequency information. From an application perspective, the proposed framework is particularly relevant to elderly activity monitoring scenarios. In such environments, practical considerations including computational efficiency, privacy preservation, robustness, and deployment feasibility are often as important as maximizing recognition accuracy. Since the method operates solely on skeletal data, it avoids the privacy concerns associated with RGB-based monitoring systems while reducing computational requirements. Moreover, the measured inference latency of approximately 24 ms per sample indicates that the framework can support low-latency activity recognition in practical monitoring applications. Nevertheless, several limitations remain. The current framework relies exclusively on skeletal information and therefore does not exploit contextual information available in additional sensing modalities. Furthermore, the frequency attention module performs global frequency recalibration and may not fully capture activity-dependent spectral variations occurring at different temporal stages of an action sequence.

Future work will focus on developing more adaptive frequency modeling strategies capable of selectively emphasizing discriminative spectral patterns according to activity characteristics. In addition, lightweight structural reasoning mechanisms and efficient multimodal fusion strategies may be investigated to further improve recognition robustness while preserving the computational efficiency required for long-term deployment in elderly-care environments. Such extensions may contribute to more reliable activity monitoring, behavioral analysis, and health-support systems for independent living applications.

References

[1] J. Jang, D. Kim, C. Park, M. Jang, J. Lee, and J. Kim, "ETRI-activity3D: A large-scale RGB-D dataset for robots to recognize daily activities of the elderly," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020: IEEE, pp. 10990–10997.

[2] F. Mehmood, X. Guo, E. Chen, M. A. Akbar, A. A. Khan, and S. Ullah, "Extended multi-stream temporal-attention module for skeleton-based human action recognition (HAR)," *Computers in Human Behavior*, vol. 163, p. 108482, 2025.

[3] A. M. Ahmadi, K. Kiani, R. Rastgoo, "A Transformer-Based Model for Abnormal Activity Recognition," *Journal of Modeling in Engineering*, vol. 22, no. 76, pp. 213-221, 2024.

[4] R. Rastgoo, K. Kiani, S. Escalera, "A deep generative Skeleton-based dynamic hand gesture production model," *Multimedia Tools and Applications*, vol. 84, pp. 48589–48608, 2025.

[5] R. Mohammadizand, R. Rastgoo, "Skeleton-based Sign Language Generation Using a Transformer-based Generative Model," *Journal of AI and Data Mining*, 2026.

[6] R. Rastgoo, K. Kiani, "Face recognition using fine-tuning of Deep Convolutional Neural Network and transfer learning," *Journal of Modeling in Engineering*, vol. 17, no. 58, pp. 103-111, 2019.

[7] F. Bagherzadeh, R. Rastgoo, "Deepfake image detection using a deep hybrid convolutional neural network," *Journal of Modeling in Engineering*, vol. 21, no. 75, pp. 19-28, 2023.

[8] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI conference on artificial intelligence*, 2018, vol. 32, no. 1.

[9] N. Raisi, M. Rezaei, and B. Masoumi, "Attention-HAR: Advanced Human Activity Recognition Using a Deep Learning Model with an Integrated Attention Mechanism," *Journal of AI and Data Mining*, 2025.

[10] Z. Li, F. Li, and G. Hua, "Dynamic Graph Attention Network for Skeleton-Based Action Recognition," *Applied Sciences*, vol. 15, no. 9, p. 4929, 2025.

[11] K. Gedamu, Y. Ji, L. Gao, Y. Yang, and H. T. Shen, "Relation-mining self-attention network for skeleton-based human action recognition," *Pattern Recognition*, vol. 139, p. 109455, 2023.

[12] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1010–1019.

[13] H.-g. Chi, M. H. Ha, S. Chi, S. W. Lee, Q. Huang, and K. Ramani, "Infogcn: Representation learning for human skeleton-based action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 20186–20196.

[14] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 143–152.

[15] Z. Chen, S. Li, B. Yang, Q. Li, and H. Liu, "Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition," in *Proceedings of the AAAI conference on artificial intelligence*, 2021, vol. 35, no. 2, pp. 1113–1122.

[16] T. Chen et al., "Learning multi-granular spatio-temporal graph network for skeleton-based action recognition," in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 4334–4342.

[17] H. Liu, Y. Liu, Y. Chen, C. Yuan, B. Li, and W. Hu, "TranSkeleton: Hierarchical spatial–temporal transformer for skeleton-based action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 8, pp. 4137–4148, 2023.

[18] C. Plizzari, M. Cannici, and M. Matteucci, "Spatial temporal transformer network for skeleton-based action recognition," in *international conference on pattern recognition*, 2021: Springer, pp. 694–701.

[19] H. Qiu, B. Hou, B. Ren, and X. Zhang, "Spatio-temporal tuples transformer for skeleton-based action recognition," *arXiv preprint arXiv*:2201.02849, 2022.

[20] K. Kiani, S. Rezaeirad, "A new ergodic HMM-based face recognition using DWT and half of the face," in *5th Conference on Knowledge Based Engineering and Innovation (KBEI)*, 2019, pp. 531-536.

[21] J. Do and M. Kim, "Skateformer: skeletal-temporal transformer for human action recognition," in *European Conference on Computer Vision*, 2024: Springer, pp. 401–420.

[22] J. Zhang, Y. Jia, W. Xie, and Z. Tu, "Zoom transformer for skeleton-based group activity recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 12, pp. 8646–8659, 2022.

[23] W. Wu, C. Zheng, Z. Yang, C. Chen, S. Das, and A. Lu, "Frequency guidance matters: Skeletal action recognition by frequency-aware mixed transformer," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 4660–4669.

[24] Y. Zhou et al., "Hypergraph transformer for skeleton-based action recognition," *arXiv preprint arXiv:2211*.09590, 2022.

[25] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, "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.

[26] X. Shu, J. Yang, R. Yan, and Y. Song, "Expansion-squeeze-excitation fusion network for elderly activity recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 8, pp. 5281–5292, 2022.

تشخیص کارآمد فعالیت‌های اسکلتی با آگاهی از فرکانس برای پایش سالمندان

فاطمه نقوی و کوروش کیانی *

دانشکده برق و کامپیوتر، دانشگاه سمنان، سمنان، ایران.

ارسال ۱۴۰۴/۰۶/۱۵؛ بازنگری ۱۴۰۵/۰۳/۲۵؛ پذیرش ۱۴۰۵/۰۴/۱۰

چکیده:

تشخیص فعالیت‌های انسانی مبتنی بر بازنمایی‌های اسکلتی، به دلیل مقاومت در برابر تغییرات نورپردازی و حفظ حریم خصوصی، به‌عنوان راهکاری امیدبخش برای پایش سالمندان مطرح شده است. بااین‌حال، دستیابی به دقت بالای تشخیص و در عین حال حفظ عملکرد بلادرنگ در دستگاه‌های لبه‌ای با منابع محدود، همچنان یک چالش مهم به شمار می‌رود. در این مقاله، یک چارچوب سبک‌وزن مبتنی بر اسکلت برای تشخیص فعالیت‌های انسانی با هدف پایش فعالیت سالمندان ارائه می‌شود. مدل پیشنهادی، بازنمایی‌های چنددیدگاهی از حرکت را با یکدیگر ترکیب می‌کند که شامل مختصات مطلق مفاصل، حرکت کوتاه‌مدت، حرکت بلندمدت و ویژگی‌های روابط فضایی هستند. به‌منظور بهبود تفکیک حرکات در بُعد زمانی، یک ماژول توجه فرکانسی زمانی معرفی می‌کنیم که تبدیل فوریه سریع را منحصراً در امتداد بُعد زمانی انجام داده و پاسخ‌های ویژگی را به‌صورت تطبیقی، بر اساس توزیع انرژی آن‌ها در حوزه فرکانس، بازتنظیم می‌کند. سپس، بازنمایی‌های تقویت‌شده در حوزه فرکانس توسط کانولوشن‌های زمانی عمقی چندمقیاسی پردازش می‌شوند تا الگوهای حرکتی در بازه‌های زمانی مختلف با حداقل سربار محاسباتی استخراج شوند. معماری حاصل تنها دارای ۰.۳۷ میلیون پارامتر قابل آموزش است و به‌طور خاص برای استقرار کارآمد بر روی پلتفرم‌های لبه‌ای طراحی شده است.

کلمات کلیدی: تشخیص فعالیت‌های انسانی، فعالیت‌های روزمره سالمندان، پایش فعالیت‌های روزمره، شبکه عصبی کانولوشنی، ویژگی‌های فضایی-زمانی، تبدیل فوریه سریع.