DOI: https://doi.org/10.5281/zenodo.16949180

SKELETON-BASED HUMAN ACTION RECOGNITION USING TRANSFORMER MODEL WITH SOFTMAX WITH MULTI-DIMENSIONAL CONNECTED WEIGHTS

Avazjon Rakhimovich Marakhimov[1], Kabul Kadirbergenovich Khudaybergenov[2,3]
[1]DSc, professor, Tashkent State Technical University, Tashkent, Uzbekistan, avaz.marakhimov@yandex.ru;
[2]PhD, Associate professor, Kimyo International University in Tashkent, Tashkent, Uzbekistan, kabul85@mail.ru;
[3]PhD, Associate professor, Research Institute for the Development of Digital Technologies and Artificial Intelligence, kabul85@mail.ru.

Abstract. *Skeleton-based human action recognition (HAR), particularly from CCTV surveillance footage, has garnered significant interest within the artificial intelligence community. The skeletal modality provides a robust, high-level representation of human motion. Prevailing methods in this domain predominantly rely on a joint-centric approach, modeling the human body as a set of coordinate points. However, this representation often fails to fully capture the rich structural and kinematic relationships essential for accurate motion classification. To address this limitation, we propose a novel method termed SoftMax with Multi-Dimensional Connected Weights. This approach enhances classification by explicitly modeling the informative connections between body joints, represented as skeletal edges. We develop an end-to-end deep learning framework that learns discriminative spatio-temporal representations directly from sequences of skeleton point vectors using Convolutional Neural Networks (CNNs). Results demonstrate that our approach achieves state-of-the-art performance, underscoring the effectiveness of leveraging skeletal edge information and advanced classification techniques for human action recognition.*

Keywords: *SoftMax, machine learning, action classification, skeleton motion, human action recognition, convolution, deep learning.*

1. Introduction

Human action recognition constitutes a fundamental component of various computer vision applications, such as surveillance systems [1], human behavior analysis [2], and human-robot interaction [3]. Contemporary deep learning methods for action recognition primarily focus on learning complex spatiotemporal features from video data [4]. In recent years, skeletal representations of human motion, obtained either through hardware sensors like Kinect [5] or vision-based pose estimation algorithms, have gained considerable research interest due in part to progress in human pose estimation [6]. Although skeletal data offers benefits such as compactness and robustness to background clutter, effectively capturing discriminative patterns from sequential skeleton data remains a challenging task [7].

Skeletal data has been widely adopted in action recognition systems, with human joint coordinates typically organized into sequences, pseudo-images, or graph structures. A variety of neural network architectures have been employed to learn

spatiotemporal features from these representations, including recurrent neural networks (RNNs) [8], convolutional neural networks (CNNs) [9], and graph neural networks (GNNs) [10]. In this work, we focus on skeleton-based pseudo-images as input representations and propose a CNN-based framework enhanced with multi-dimensional connected weights for action classification. We observe that current CNN-based approaches often overlook kinematic relationships between joints—represented as skeleton edges—and predominantly rely on joint coordinate information [11].

2. Methods

2.1. The proposed model. The proposed architecture, illustrated in Figure 1, features a dual-branch design for the parallel processing of skeletal and video data streams. This framework integrates deep supervision to optimize modality fusion, addressing a key limitation of conventional methods that often rely on a single modality or simplistic feature concatenation. Diverging from such approaches, which frequently neglect nuanced cross-modal interactions, our model introduces several key innovations. The skeletal branch employs a convolutional neural network (CNN) coupled with a self-attention mechanism and a multiconnected SoftMax layer, facilitating the dynamic modeling of inter-keypoint dependencies and the extraction of complex spatial features. Concurrently, the video processing branch implements a slow-fast architecture to enable multi-temporal-scale analysis, thereby enhancing the capture of rapid motion patterns. Empirical evaluations demonstrate that the proposed network achieves state-of-the-art performance, offering superior multimodal integration and significantly improving the accuracy and computational efficiency of human action recognition.

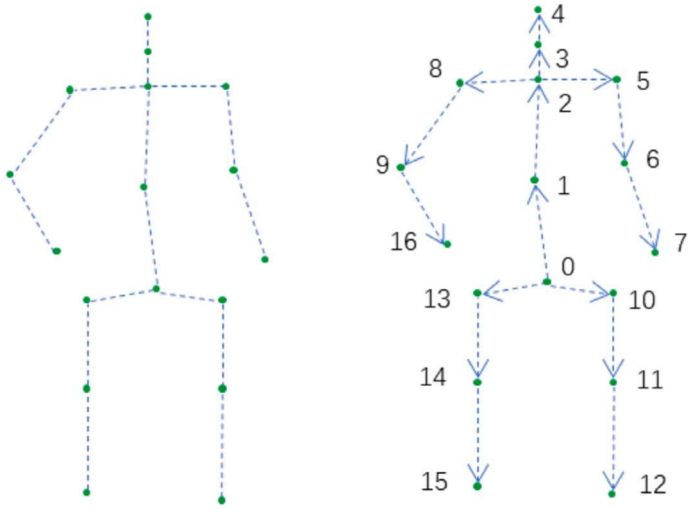

Fig. 1. Representations of keypoints.

2.2. Human Skeletal Keypoints. The acquisition of human skeletal data is scenario-dependent; it is either provided directly within benchmark datasets or must be extracted from raw video using pose estimation algorithms, a process detailed in the experimental section. The raw skeletal keypoint data is structured as a three-dimensional tensor of dimensions (I, V, T), where the first dimension corresponds to the spatial coordinate channels (x, y, and, if applicable, confidence score), V denotes

the number of defined human keypoints (which is dataset-specific), and T represents the temporal length of the sequence, determined by the frame sampling strategy.

As depicted in Figure 2, two primary representations are employed for skeletal sequences: keypoint and bone representations. The keypoint representation utilizes the spatial coordinates of each keypoint directly as node features, emphasizing their absolute positional information. In contrast, the bone representation defines the keypoint with index 0 as a root node. For all subsequent keypoints, a vector is computed that represents the directed skeletal segment from the parent keypoint to the current one. This vector is set to zero for the root node. This approach explicitly models the relational topology of the human skeleton graph.

To computationally formalize the bone representation, a predefined adjacency matrix W is constructed. The elements of W are defined such that for any directed edge connecting a parent node to its child, the corresponding matrix element is set to -1. For example, given two connected keypoints, p_1 (parent) and p_2 (child), in a single frame, the directed bone vector e_2 is calculated as:

$$e_2 = p_2 - p_1$$

This operation can be efficiently implemented for the entire graph via the matrix multiplication $E = \mathbf{W}^T P$, where P is the matrix of keypoint coordinates.

$$e_2 = p_2 - p_1 = (x_2 - x_1, y_2 - y_1, z_2 - z_1)^T \quad (1)$$

Concurrently, the element at the matrix coordinate (2, 1) within W is set to -1 to encode this specific directed relationship. The resulting bone representation, computed as the matrix product P·W, retains the same dimensional structure as the original keypoint tensor PP. The final input to the network is constructed by concatenating the keypoint (P) and bone (P·W) representations along the channel dimension:

$$I = Concat(P, P*W) \quad (2)$$

where $I \in R^6 \times V \times T$ denotes the concatenation operation, P is the keypoint representation, and P·W represents the bone linkages. This formulation effectively integrates absolute spatial positions with the inherent topological structure of the human skeleton, providing a more comprehensive input for subsequent processing.

To capture elementary motion dynamics, the temporal difference for each feature across the temporal dimension T is computed. This differential V, representing the change from the previous timestep, is formulated as:

$$V_t = I_t - I_{t-1}$$

yielding a resultant tensor $V \in R^6 \times V \times (T-1)$. Subsequently, the original input I and the temporal differential V are each processed through separate 1×1 convolutional layers to project their channel dimensionality to 64, enriching their feature representations. The refined feature tensors are thus given by:

$$\tilde{I} = \mathrm{ReLU}(W_2(\mathrm{ReLU}(W_1 I))) \quad (3)$$

$$\tilde{V} = \mathrm{ReLU}(W_4(\mathrm{ReLU}(W_3 V))) \quad (4)$$

Subsequently, the enriched features from these components are fused through element-wise summation to produce an augmented representation:

$$Z = V' + I' \quad (5)$$

where $Z \in R^{64} \times V \times T$. Following this enhancement of the input's representational capacity, it is critical to incorporate structural and temporal context, such as spatial keypoint indexing and temporal ordering, into the feature ensemble. To this end, one-hot encodings of the spatial (J) and temporal (T) indices are generated. These encodings are then projected into a higher-dimensional latent space using a process analogous to that in Eqs. (1) and (2), which involves two 1×1 convolutional layers for feature refinement. This yields enriched representations $J \in R^{64} \times V \times T$ and $T' \in R^{128} \times V \times T$.

Finally, these contextual features are concatenated with the motion-augmented features Z along the channel dimension:

$$Z' = Concat\left(Z, \tilde{J}, \tilde{T}\right) \tag{6}$$

This integration consolidates motion, spatial structure, and temporal context into a unified representation. This concludes the input encoding process, resulting in a tensor Z' with a channel dimension of 256. This final encoded representation is then propagated to the self-attention convolutional module for subsequent skeletal feature extraction.

2.3. Convolutional network module

In CNN, integrating a self-attention mechanism is essential for dynamically inferring the adjacency matrix and its corresponding edge weights. Unlike conventional CNNs that utilize a fixed, predefined adjacency matrix to model node relationships, the self-attention mechanism adaptively recalibrates these relational weights based on data-driven similarity measures. This facilitates more flexible feature propagation and information aggregation across the graph.

The mechanism operates by quantifying the pairwise affinities between joints (nodes) and assigning a weight to each pair. These weights dictate the influence between nodes during the message-passing phase. A key advantage of this approach is its ability to capture long-range dependencies; even joints without a direct physical connection can exhibit strong latent associations, which are amplified through self-attention. Consequently, the model can emphasize dynamic interactions between joints during critical action segments (e.g., periods of rapid motion) as well as subtle correlations between less active joints, thereby enhancing the overall understanding of action patterns.

The self-attention mechanism, widely adopted across domains, computes a representation for each feature as a weighted sum of all features. The general formulation is:

$$Attention\left(Query, Source\right) = \sum_{i=1}^{L} Similarity\left(Query, Key_i\right) \cdot Value_i \tag{7}$$

where Similarity is a trainable function that calculates affinity scores. In self-attention, the Query, Key, and Value are all derived from the same input source.

The specific architecture implemented here is designed within a neighborhood graph framework, customized for skeletal data. An initial undirected graph is first oriented into a task-specific directed graph to better model informational flow.

The input to this module is a feature tensor of dimensions C×T×V, which provides the sources for the Query, Key, and Value. The similarity matrix function f(x), corresponding to the Similarity term in Eq. (1), is defined as:

$$f(x) = softmax(\theta(x)^{\cdot} \ \phi(x)) \tag{8}$$

Here, θ(x) and ϕ(x) are linear transformations (implemented via 1×1 convolutions) applied to the Query and Key projections, respectively, and T denotes the transpose operation. The dimension C represents the number of input channels.

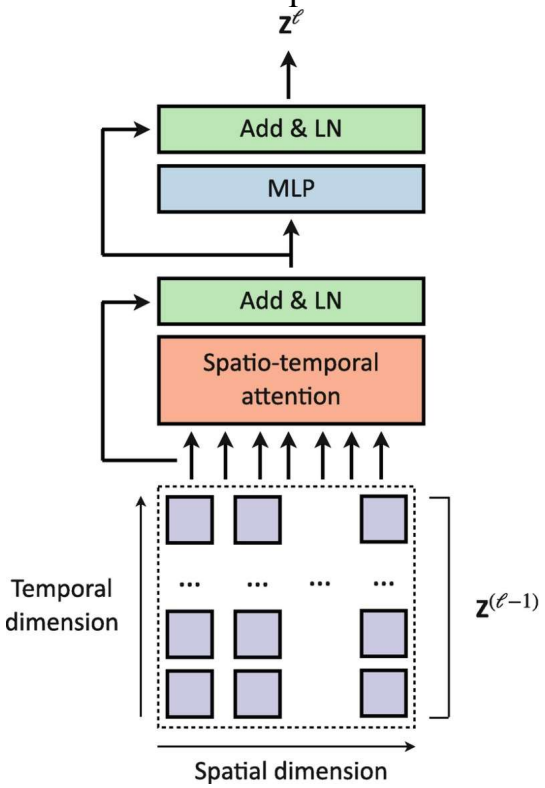

Fig. 2. Self-attention block.

The output of the self-attention block is formally defined by the operation:

$$y(x) = ReLU(h(f(x) \cdot g(x)) + x)$$

where x and y(x) represent the input and output feature maps, respectively. The function g(x) applies a 1×1 convolution to the Value pathway to enhance its representational capacity. The function h(x) employs another 1×1 convolution to project the aggregated features to the desired output dimension, and the term +x denotes a residual connection.

This self-attention module can be interpreted as a form of CNN. In standard CNNs, the adjacency matrix is a static, binary matrix (with values 0 or 1) defining connections between nodes. In this formulation, the learned similarity matrix f(x) functions as a dynamic, weighted adjacency matrix, where the values are continuous and data-dependent.

For skeletal data, the spatio-temporal self-attention module is designed to model both the spatial correlations between joints and their temporal kinematics. In contrast, for video data, the corresponding module focuses on integrating multi-scale temporal

features to capture the dynamics of motion at different speeds, thereby improving recognition robustness.

The adjacency matrix and its weights in this architecture are dynamically generated by the self-attention mechanism, confirming its characterization as a CNN. The resulting adjacency matrix is inherently asymmetric, indicating a directed graph structure.

The computational complexity of the self-attention mechanism is primarily governed by the matrix multiplication required to compute the similarity matrix. To alleviate this cost, a more efficient spatial self-attention module was adopted. This design reduces computational burden by applying self-attention only in the spatial dimension, while temporal feature integration is handled by a parallel branch consisting of a simple 3×1 temporal convolution, which is then combined via the residual connection.

Empirical results indicate that the architecture in model achieves higher accuracy with reduced computational demands. Consequently, this design was selected for our final model. The channel dimension C_1 within the module was set to half the input channel size to further minimize computational overhead. The final output channels of the convolutional modules in the skeletal branch (Figure 1) are configured as 128, 256, 256, and 512.

3. Experiment

3.1. Datasets. We evaluate the performance of our proposed method on two widely-used benchmark datasets for human action recognition: PennAction and CSL. This section first provides a brief introduction to these datasets and outlines the experimental setup. We then present extensive experimental results and a comparative analysis against current state-of-the-art methods. Finally, we perform detailed ablation studies to examine the contribution of each component in our proposed framework and discuss potential directions for future improvement.

3.1. Datasets. PennAction. This dataset contains 2,326 video sequences, sourced from YouTube, representing 15 different action categories (e.g., "baseball pitch," "bench press," "strum guitar"). Each frame is annotated with the 2D coordinates of 13 human body joints; however, occlusions frequently result in missing joint annotations across frames. We adhere to the standard evaluation protocol outlined in [33], using 50% of the videos for training and the remaining 50% for testing. The dataset presents significant challenges due to frequent occlusions and large variations in subject scale and viewpoint.

Table 1.

Action recognition performance on the PennAction dataset. Results are reported for models using skeletal data extracted via pose estimation algorithms.

Method	Pose recognition (%)
Bilinear C3D	97.10
HDM	93.40
MDL	98.60
Heapmap	98.22
RPAN	97.40
SoftMax classifier	85.64

NN	90.23
CNN	91.25
CNNSoftMaxMCW	98.25

CSL. The CSL (Chinese Sign Language) dataset is a large-scale corpus containing 500 frequently used words, with each word performed 5 times by 50 different signers, resulting in a total of 125,000 video samples. In line with the established evaluation methodology, we partition the data at the signer level: samples from 36 signers are used for training, and samples from the remaining 14 signers are held out for testing, ensuring a person-independent evaluation.

Table 2.

Results on CSL dataset, skeleton obtained by pose estimation algorithm and pose recognition.

Method	Pose recognition (%)
Bilinear C3D	96.23
HDM	93.40
MDL	98.60
Heapmap	98.22
RPAN	97.40
SoftMax classifier	85.64
NN	90.23
CNN	91.25
CNNSoftMaxMCW	98.71

ACKNOWLEDGEMENTS

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

REFERENCES

[1] T.V. Nguyen, B. Mirza, Dual-layer kernel extreme learning machine for action recognition, Neurocomputing 260 (2017) 123–130.
[2] R. Minhas, A. Baradarani, S. Seifzadeh, Q.J. Wu, Human action recognition using extreme learning machine based on visual vocabularies, Neurocomputing 73 (10–12) (2010) 1906–1917.
[3] D. Zhao, L. Shao, X. Zhen, Y. Liu, Combining appearance and structural features for human action recognition, Neurocomputing 113 (2013) 88–96.
[4] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, M. Paluri, A closer look at spatiotemporal convolutions for action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6450–6459.
[5] Z. Zhang, Microsoft kinect sensor and its effect, IEEE Multimedia 19 (2) (2012) 4–10.
[6] Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh, Realtime multi-person 2d pose estimation using part affinity fields, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7291–7299.

[7] C. Cao, Y. Zhang, C. Zhang, H. Lu, Body joint guided 3-d deep convolutional descriptors for action recognition, IEEE Transactions on Cybernetics 48 (3) (2017) 1095–1108.
[8] J. Liu, A. Shahroudy, D. Xu, G. Wang, Spatio-temporal lstm with trust gates for 3d human action recognition, European Conference on Computer Vision (2016) 816–833.
[9] Y. Hou, Z. Li, P. Wang, W. Li, Skeleton optical spectra-based action recognition using convolutional neural networks, IEEE Transactions on Circuits and Systems for Video Technology 28 (3) (2016) 807–811.
[10] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, Q. Tian, Actional-structural graph convolutional networks for skeleton-based action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3595–3603.
[11] C. Li, Y. Hou, P. Wang, W. Li, Joint distance maps based action recognition with convolutional neural networks, IEEE Signal Processing Letters 24 (5) (2017) 624–628.