The European Journal of Technical and Natural Sciences 2025, No 4–5
ISSN 2414-2352

PREMIER Publishing
ppublishing.org

Section 2. Computer science

DOI:10.29013/EJHSS-25-4.5-8-17

SKELETON-BASED HUMAN ACTION RECOGNITION USING CNN+SOFTMAX WITH MULTI-DIMENSIONAL CONNECTED WEIGHTS

Avazjon Rakhimovich Marakhimov [1],
Kabul Kadirbergenovich Khudaybergenov [2, 3]

[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

Cite: *Marakhimov A.R., Khudaybergenov K.Kh. (2025). Skeleton-Based Human Action Recognition Using Cnn+Softmax With Multi-Dimensional Connected Weights. European Journal of Technical and Natural Sciences 2025, No 1.* https://doi.org/10.29013/EJHSS-25-4.5-8-17

Abstract

Skeleton-based human activity recognition through closed-circuit television surveillance systems has garnered substantial attention within the artificial intelligence research domain, primarily attributed to the rich feature representation inherent in skeletal data. Contemporary machine learning approaches predominantly employ joint-coordinate representations of human anatomical structure, resulting in suboptimal understanding of motion pattern classification. This work introduces a novel methodology utilizing SoftMax classification enhanced with multi-dimensional connected weights for improved human action categorization accuracy. Our approach emphasizes skeletal edge point analysis as discriminative features and develops a skeleton-driven algorithmic framework that extracts robust deep feature representations from skeletal point vectors through convolutional neural network architectures integrated with the proposed multi-dimensional weighted SoftMax classifier. Empirical validation conducted on established human action recognition benchmarks, including PennAction and CSL datasets, demonstrates the superior performance of our proposed methodology.

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

I. Introduction

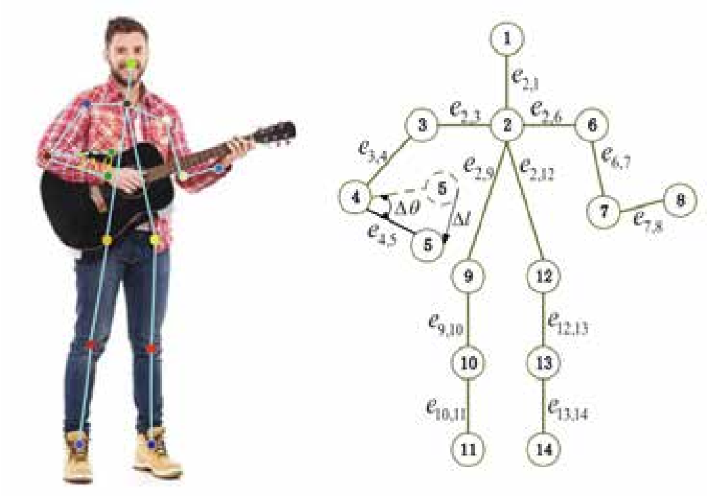
Human activity recognition constitutes a fundamental component across diverse computer vision domains, encompassing surveillance infrastructures (Nguyen T. V., Mirza B., 2017), behavioral pattern analysis (Minhas R., Baradarani A., Seifzadeh S., Wu Q. J., 2010), and autonomous robotic systems (Zhao D., Shao L., Zhen X., Liu Y., 2013). Current deep learning methodologies for activity recognition primarily concentrate on extracting intricate spatiotemporal characteristics from video data streams (Tran D., Wang H., Torresani L., Ray J., Le Cun Y., Paluri M., 2018). Over recent years, skeletal modeling of human subjects – obtained through hardware solutions like Kinect sensors (Zhang Z., 2012) or via computational pose estimation frameworks – has attracted considerable research focus within activity recognition studies, facilitated predominantly by progress in human pose detection methodologies (Cao Z., Simon T., Wei S.-E., Sheikh Y., 2017). While skeletal modeling offers benefits of data compactness and robustness against environmental complexities, the efficient derivation of discriminative patterns from temporal skeletal sequences presents ongoing challenges (Cao C., Zhang Y., Zhang C., Lu H., 2017).

The integration of skeletal information within activity recognition architectures has achieved broad adoption, with human joint positioning typically organized as temporal sequences, pseudo-imagery, or graph-based structures. Investigators have implemented diverse neural network configurations to derive reliable spatiotemporal characteristics from these input representations, encompassing recurrent neural networks (RNNs) (Liu J., Shahroudy A., Xu D., Wang G., 2016), convolutional neural networks (CNNs) (Hou Y., Li Z., Wang P., Li W., 2016), and graph neural networks (GNNs) (Li M., Chen S., Chen X., Zhang Y., Wang Y., Tian Q., 2019). This investigation specifically targets skeletal pseudo-imagery as input modality, utilizing CNNs enhanced with multi-dimensional connected weights for activity classification. We observe that prevailing CNN-based approaches generally overlook the examination of limb segment dynamics (skeletal edge motion), focusing primarily on joint positional data (Li C., Hou Y., Wang P., Li W., (2017).

Body segment kinematics constitute a pivotal factor in distinguishing human activities within skeletal sequences; however, extracting skeletal edge characteristics directly from joint positional data through neural network architectures poses considerable computational challenges. To address this constraint, we introduce an innovative skeletal modality termed skeleton edge motion, which augments representation learning by incorporating limb segment kinematic analysis via CNNs enhanced with multi-dimensional connected weights. Figure 1 depicts our methodology, where the proposed modality encodes both the angular variations $\Delta\theta$ of anatomical segments and the positional shifts $\Delta\theta$ of their corresponding joints. We combine this novel modality with conventional joint coordinates throughout the skeleton point vectors within the pseudo-image framework. As evidenced in Figure 2(b), our proposed approach exhibits adaptability across diverse CNN configurations. Furthermore, recognizing the inherent organization of skeletal pseudo-images – where horizontal dimensions encode joints from discrete frames (spatial data) and vertical dimensions represent individual joints across temporal progressions (temporal data)—we have developed a specialized CNN framework for optimal feature extraction. Our proposed architecture incorporates dual convolutional components: 1) a spatial processing layer employing an $1\times k$ convolutional kernel, and 2) a temporal processing branch implementing a $k\times 1$ convolutional kernel. The comprehensive skeleton edge motion network is assembled through sequential concatenation of multiple such CNN modules, as depicted in Figure 2(b). Additional specifications concerning the CNN architecture are elaborated in Section 3.2.

The temporal organization of data constitutes a critical element in comprehending interdependencies within intricate activity sequences. Consequently, multiple reasoning architectures have been developed for human activity recognition (Si C., Jing Y., Wang W., Wang L., Tan T., 2018). A compelling demonstration of temporal sequencing significance can be witnessed in activity pairs such as “standing up” and “sitting down,” where the primary discriminating characteristic is the chronological progression of video frames. A remarkable observation is that both algorithmic models and human evaluators often erroneously classify these activity pairs when the temporal sequence is reversed – a problem designated as the "arrow of time in videos" (Wei D., Lim J. J., Zisserman A., Freeman W. T., 2018). Based on these observations, we introduce a novel SoftMax approach for skeleton-based human activity recognition.

Figure 1. *Depiction of skeletal points serving as discriminative feature elements in human activity recognition applications. These points constitute predetermined linkages between anatomical landmarks that correspond to distinct body regions, for instance, $e_{4,5}$ denotes the forearm, whereas $e_{3,4}$ corresponds to the upper arm*



This manuscript is organized as follows: Section 2 reviews pertinent literature in the domain. Section 3 details the framework and methodology of our proposed approach for human activity recognition. Section 4 reports experimental results from our performance assessment. Section 5 offers concluding observations and prospective research directions. The primary contributions of this study can be encapsulated in three fundamental aspects: 1) the presentation of a novel skeletal input point vector for human activity classification; 2) the formulation of CNN architecture with SoftMax incorporating multi-dimensional connected weights which facilitates the extraction of comprehensive spatiotemporal representations. We perform efficacy assessments of the proposed methodology on two established human activity recognition benchmarks, PennAction (Zhang W., Zhu M., Derpanis K. G., 2013) and CSL (Zhang J., Zhou W., Xie C., Pu J., Li H., 2016), with experimental outcomes validating the effectiveness of our approach.

2. Related Works

In this section, we present reviews of previous works on human action classification, particularly in machine learning and deep learning methods utilizing human skeleton information.

Human Action Recognition

Current methodologies for human activity recognition have exhibited remarkable effectiveness utilizing both video sequences (Zolfaghari M., Singh K., Brox T., 2018) and skeletal data annotations (Du W., Wang Y., Qiao Y., 2017). For deriving significant feature representations from skeletal data, investigators have deployed various deep learning frameworks: 1) sequential architectures such as Long Short-Term Memory networks (Zhu W., Lan C., Xing J., Zeng W., Li Y., Shen L., Xie X., 2016) and Gated Recurrent Units (Song S., Lan C., Xing J., Zeng W., Liu J., 2017); 2) convolutional neural networks (Zhang B., Yang Y., Chen C., Yang L., Han J., Shao L. 2017); and 3) graph neural networks (Shi L., Zhang Y., Cheng J., Lu H., 2019). To optimize the exploitation of both video and skeletal information, investigators have primarily utilized CNN-based methodologies for skeletal data processing. Multiple investigations have augmented skeletal information with heat map representations (Newell A., Yang K., Deng J., 2016) to encode morphological movements through pseudo-images. Graph-based structural architectures (Shi L., Zhang Y., Cheng J., Lu H., 2019) have been constructed to integrate skeletal data with video sequences, simultaneously improving pose estimation precision and CNN-based activity recognition capabilities. Transfer learning strategies (He K., Zhang X., Ren S., Sun J., 2016) have been deployed to concurrently (enhance joint detection robustness and activity classification performance. Investigators have also examined human intention (Xu B., Li J., Wong Y., Kankanhalli M. S., Zhao Q., 2019), which can be derived from environmental context and functions as a valuable indicator for activity recognition. Visual appearance information from action-related objects has been utilized to guide attention toward relevant anatomical regions during recognition. Generally, video data are typically integrated with skeletal information through convolutional architectures to improve recognition performance.

Moving beyond basic joint positional data, investigators have examined more advanced representational architectures by developing sophisticated joint-derived features. Joint-to-joint distances have been utilized to characterize skeletal configurations and subsequently integrated with conventional three-dimensional interest points for comprehensive information extraction. The temporal evolution of human postures has been characterized through joint trajectory representations (Devanne M., Wannous H., Berretti S., Pala P., Daoudi M., 2014). Joint co-occurrence matrices (Li C., Zhong Q., Xie D., Pu S., 2018) have been leveraged as discriminative action descriptors. In comparison with distance-based metrics, the angular relationships between anatomically connected skeletal segments exhibit robustness to scaling variations. Consequently, joint angular similarity metrics (Ohn-Bar E., Trivedi M., 2013) have been proposed for activity differentiation. These angular relationships enable the detection of informative skeletal landmarks throughout video sequences.

Notwithstanding substantial progress based on these representational architectures, existing methodologies have primarily depended on joint-centric approaches, including joint kinematics (Vemulapalli R., Arrate F., Chellappa R., 2014) and articular angles (Ofli F., Chaudhry R., Kurillo G., Vidal R., Bajcsy R., 2014). Human anatomical segments contain considerable informational value regarding skeletal structure, necessitating additional exploration into segment-based representations in activity classification. Recent studies examining inter-segment relationships have primarily concentrated on architectural advances, such as Lie group formulations within CNNs (Huang Z., Wan C., Probst T., Van Gool L., 2017) and graph-based NN (Shi L., Zhang Y., Cheng J., Lu H., 2019). In this investigation, we present an innovative convolutional neural architecture with multi-dimensional connected weights to better characterize the dynamics of human anatomical segments for activity classification.

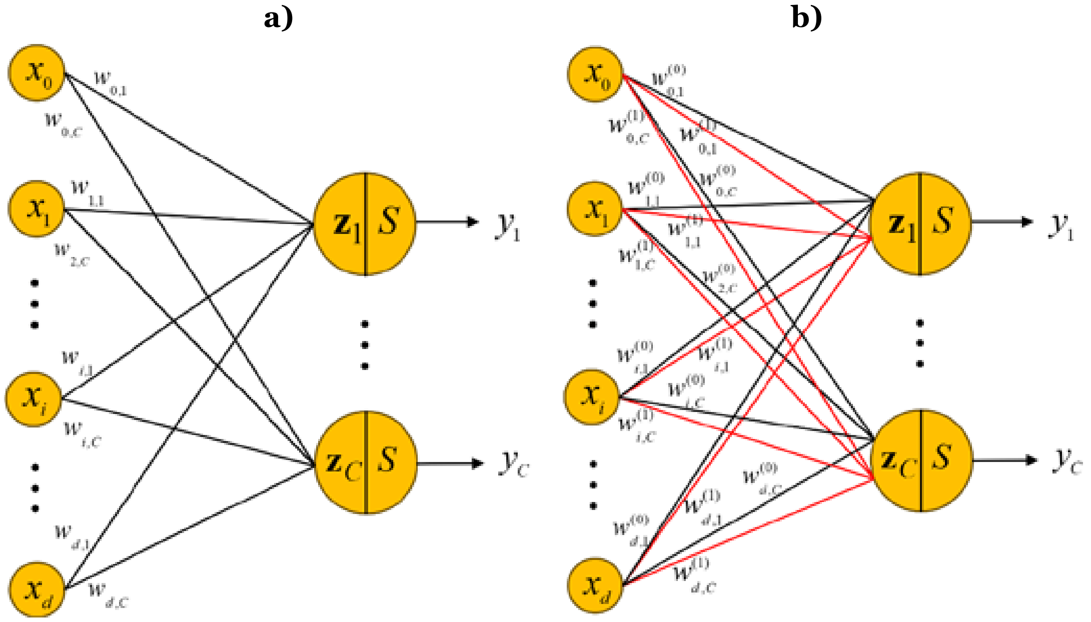
3. Model Description of CNN With Multi-Dimensional Connected Weights

This section presents the architectural framework for CNN Soft Max MCW and its corresponding training procedure. To accomplish this, we revisit the network architecture of the conventional SoftMax activation function and the CNNSoftMaxMCW framework for comparative analysis, where we employ 2-dimensional connection weights for the CNN Soft Max MCW model to maintain clarity (Fig. 2 (b)).

$$\mathrm{S}(\mathbf{x}) = \frac{e^{\mathbf{w}_c^T\mathbf{x}}}{\sum_{j=1}^{C} e^{\mathbf{w}_j^T\mathbf{x}}} \qquad (2)$$

Here, we consider the case when the models are built for C – class classification problem. In this case, the traditional SoftMax activation function has $\mathbf{x} = (x_0, x_1, ..., x_d)$ – input vector, where $x_0 = 1$, and $\mathbf{y} = (y_1, y_2, ..., y_C)$ – prediction output as vector, $\mathbf{w} = (w_0, w_1, ..., w_d)$ – weight parameters, where d is the dimension of the input vector or input space, C – is the number of classes in learning dataset. In general, both the traditional SoftMax and the proposed SoftMaxMCW activation functions have C – neurons, which are equal to the number of classes, where all the input values are multiplied by their corresponding weights and further processed by the SoftMax activation function. However, for the SoftMaxMCW activation function with multi-dimensional connection weights, the SoftMax activation function is different, which is defined later.

Figure 2. *(a) Structure of standard SoftMax activation function. (b) Structure of the CNNSoftMaxMCW model with 2-dimensional connected weights, where* S *is the SoftMax activation function as eq.(2)*



A traditional SoftMax activation function consists of a single connected weight between every input x_i and hidden unit z_j, and outputs, which is illustrated in Fig. 2a, where only one connected weight from input x_i to the hidden neuron is permitted with a connected weight $w_{i,j}$, and input connections from more than one weight coefficients are permitted. In general, in a traditional SoftMax activation function, connection between two units is provided by a single real number which is considered a scalar value. This indicates every input value has its own assigned weight parameter. These inputs enter SoftMax through input layer's units and are distributed from input layer's units to the output to calculate the probability of belonging for each class. In general, the standard SoftMax activation function can be determined as below.

Summation block for the first neuron,

$$z_1 = w_{0,1} + w_{1,1}x_1 + \ldots + w_{d,1}x_d = \sum_{i=0}^{d} w_{i,1}x_i, \quad (3)$$

For the second neuron,

$$z_2 = w_{0,2} + w_{1,2}x_1 + \ldots + w_{d,2}x_d = \sum_{i=0}^{d} w_{i,2}x_i, \quad (4)$$

...

For the C^{th}-neuron,

$$z_C = w_{0,C} + w_{1,C}x_1 + \ldots + w_{d,C}x_d = \sum_{i=0}^{d} w_{i,C}x_i, (5)$$

$$\text{SoftMax}(z) = = \left\{ \frac{e^{z_1}}{\sum_{j=1}^{C} e^{z_j}}, \frac{e^{z_2}}{\sum_{j=1}^{C} e^{z_j}}, \ldots, \frac{e^{z_C}}{\sum_{j=1}^{C} e^{z_j}} \right\}, \quad (6)$$

where $\mathbf{x} = (x_0, x_1, \ldots, x_d) \in R^{d+1}$ – input vector with $x_0 = 1$, $\mathbf{W} \in R^{(d+1)\times C}$ – weight matrix or weight parameters, $w_{0,j}$ – threshold value and SoftMax is a activation function, which is used to obtain output predictions as the classification results.

Model structure of CNNSoftMaxMCW model

The SoftMaxMCW activation function has multiple connected weights between every input node and a computation node, and output (SoftMax) blocks, which is depicted in (Fig. 2b), where multiple connection parameters from input unit to the hidden neurons are permitted. Here the notion of computation node can be considered as a hidden neuron in general meaning when it refers to classifier layer at the end of NN and CNN models. This means that every input unit has its own weight vector $\mathbf{w}^{(h)}$ of coefficient parameters, which assumes multiple connected weights between every input unit and summation block. Analogously, sensory signals are input to the SoftMaxMCW activation function through input layer and these inputs are propagated from input to the output to calculate probabilities for each class. Now we can introduce formulas for the proposed SoftMaxMCW activation function.

Suppose that there are H – dimensional connected weights between every input unit x_i and summation blocks $\mathbf{z}_j$. The SoftMaxMCW activation function also consists of C -neurons the same as the traditional SoftMax model for C classes in a learning dataset. Then the following equations can be obtained.

Summation block for the first neuron:

$$z_1^{(1)} = w_{0,1}^{(1)} + w_{1,1}^{(1)} x_1 + \dots + w_{d,1}^{(1)} x_d = \sum_{i=0}^{d} w_{i,1}^{(1)} x_i, \tag{7}$$

$$z_1^{(2)} = w_{0,1}^{(2)} + w_{1,1}^{(2)} x_1 + \dots + w_{d,1}^{(2)} x_d = \sum_{i=0}^{d} w_{i,1}^{(2)} x_i, \tag{8}$$

...

$$z_1^{(H)} = w_{0,1}^{(H)} + w_{1,1}^{(H)} x_1 + \dots + w_{d,1}^{(H)} x_d = \sum_{i=1}^{d} w_{i,1}^{(H)} x_i. \tag{9}$$

For the second neuron:

$$z_2^{(1)} = w_{0,2}^{(1)} + w_{1,2}^{(1)} x_1 + \dots + w_{d,2}^{(1)} x_d = \sum_{i=0}^{d} w_{i,2}^{(1)} x_i, \tag{10}$$

$$z_2^{(2)} = w_{0,2}^{(2)} + w_{1,2}^{(2)} x_1 + \dots + w_{d,2}^{(2)} x_d = \sum_{i=0}^{d} w_{i,2}^{(2)} x_i, \tag{11}$$

...

$$z_2^{(H)} = w_{0,2}^{(H)} + w_{1,2}^{(H)} x_1 + \dots + w_{d,2}^{(H)} x_d = \sum_{i=1}^{d} w_{i,2}^{(H)} x_i. \tag{12}$$

For the C^{th} -neuron:

$$z_C^{(1)} = w_{0,C}^{(1)} + w_{1,C}^{(1)} x_1 + \dots + w_{d,C}^{(1)} x_d = \sum_{i=0}^{d} w_{i,C}^{(1)} x_i, \tag{13}$$

$$z_C^{(2)} = w_{0,C}^{(2)} + w_{1,C}^{(2)} x_1 + \dots + w_{d,C}^{(2)} x_d = \sum_{i=0}^{d} w_{i,C}^{(2)} x_i, \tag{14}$$

...

$$z_C^{(H)} = w_{0,C}^{(H)} + w_{1,C}^{(H)} x_1 + \dots + w_{d,C}^{(H)} x_d = \sum_{i=1}^{d} w_{i,C}^{(H)} x_i. \tag{15}$$

Next, input vector Z for proposed model can be obtained as below,

$$Z = \left(\mathbf{z}_1, \dots, \mathbf{z}_C\right), \tag{16}$$

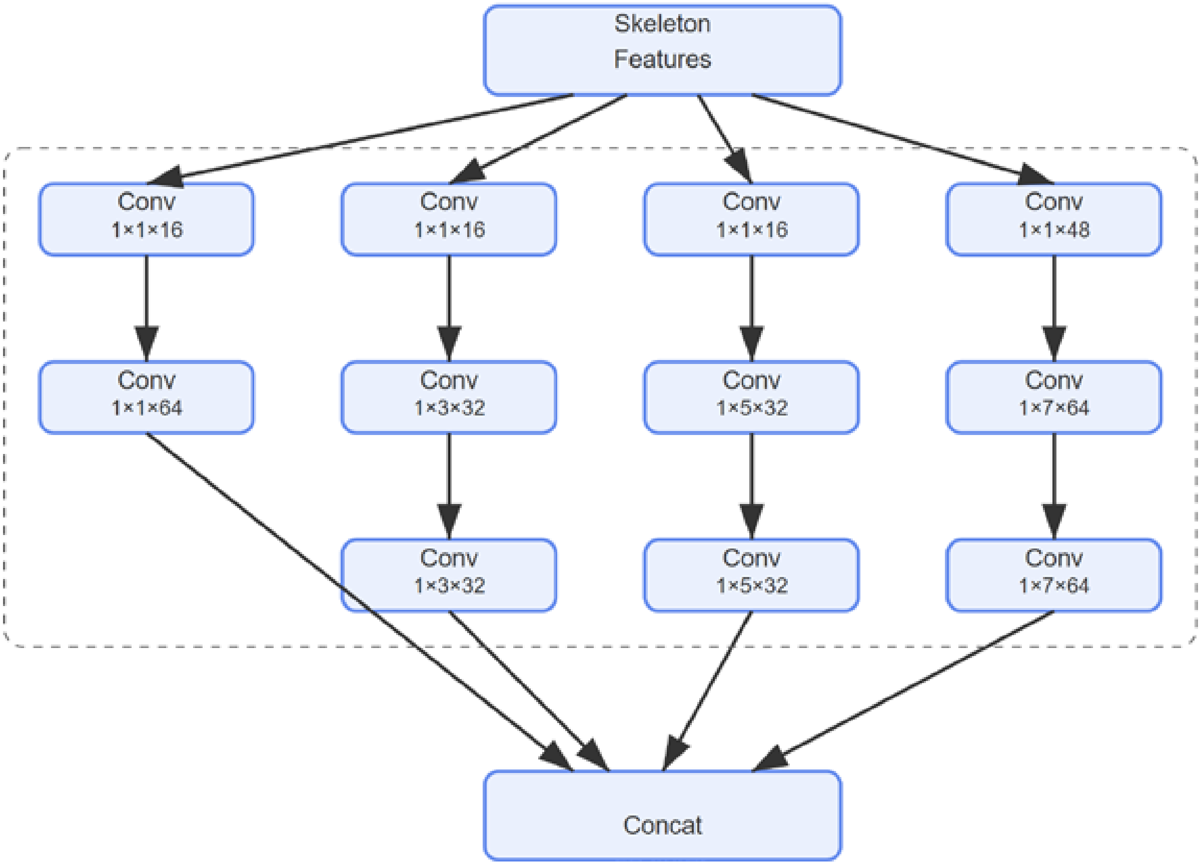
where $\mathbf{z}_j = (z_j^{(1)}, z_j^{(2)}, \dots, z_j^{(H)}),\ j = 1, 2, \dots, C,$

hence, we obtain a SoftMax activation function with multi-dimensional connected weights as below,

$$\mathrm{SoftMax}\left(Z\right) = \left\{ \frac{\sum_{h=1}^{H} e^{z_1^{(h)}}}{\sum_{j=1}^{C} \sum_{h=1}^{H} e^{z_j^{(h)}}}, \frac{\sum_{h=1}^{H} e^{z_2^{(h)}}}{\sum_{j=1}^{C} \sum_{h=1}^{H} e^{z_j^{(h)}}}, \dots, \frac{\sum_{h=1}^{H} e^{z_C^{(h)}}}{\sum_{j=1}^{C} \sum_{h=1}^{H} e^{z_j^{(h)}}} \right\}, \tag{17}$$

where, $h \in \left\{1, \dots, H\right\}$ – dimension of weight connections between every input node and the summation block. Likewise, to the traditional SoftMax activation function, the proposed SoftMax function has C – components for C – classes as the probability prediction of each class. To illustrate the proposed model, below is shown a case where the number of connected weights between every input node and summation block is taken as 2-dimentional weight connections, i.e. $H = 2$, (Fig. 2b).

Figure 2.



4. Experiments

We assess the efficacy of our proposed method on two widely recognized human action recognition benchmarks datasets: PennAction and CSL. This section begins with a concise description of these established datasets and our experimental methodology. Subsequently, we present a comprehensive series of experiments and comparative analyses between our proposed method and current leading approaches in the field. Finally, we conduct detailed ablation studies to isolate the contributions of individual components within our proposed framework and discuss potential avenues for future enhancement.

Datasets

PennAction. This dataset encompasses 15 distinct action categories, including “baseball pitch,” “bench press,” and “strum guitar,” comprising a total of 2326 video sequences obtained from YouTube. Images from video frames are annotated with 13 anatomical landmarks, although occlusion results in some landmarks being non-visible in certain frames. We employ the evaluation protocol established in (Rahmani H., Bennamoun M., (2017). [33], allocating 50% of the video sequences for model training and the remaining 50% for performance testing. This benchmark is characterized by significant challenges including complex body occlusions and substantial variations in subject scale.

CSL. This sign language corpus focuses on vocabulary commonly utilized in daily communication, including terms such as body, arm, leg and related concepts, encompassing 125,000 examples. The dataset has 500 distinct sign words, with each word performed by 50 different signers repeated 5 times. In accordance with the standardized evaluation methodology, we utilize samples from 36 signers for model training and reserve the remaining 14 signers for computational experiments.

Table 1. *Performance results on the PennAction dataset, utilizing skeletal data derived from pose estimation algorithms and pose recognition techniques*

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

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

Nguyen T. V., Mirza B. (2017). Dual-layer kernel extreme learning machine for action recognition, Neurocomputing – 260. – P. 123–130.

Minhas R., Baradarani A., Seifzadeh S., Wu Q. J. (2010). Human action recognition using extreme learning machine based on visual vocabularies, Neurocomputing – 73 (10–12). – P. 1906–1917.

Zhao D., Shao L., Zhen X., Liu Y. (2013). Combining appearance and structural features for human action recognition, Neurocomputing – 113. – P. 88–96.

Tran D., Wang H., Torresani L., Ray J., Le Cun Y., Paluri M. (2018). A closer look at spatio-temporal convolutions for action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, – P. 6450–6459.

Zhang Z. (2012). Microsoft kinect sensor and its effect, IEEE Multimedia – 19 (2). – P. 4–10.

Cao Z., Simon T., Wei S.-E., Sheikh Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, – P. 7291–7299.

Cao C., Zhang Y., Zhang C., Lu H. (2017). Body joint guided 3-d deep convolutional descriptors for action recognition, IEEE Transactions on Cybernetics – 48 (3). – P. 1095–1108.

Liu J., Shahroudy A., Xu D., Wang G. (2016). Spatio-temporal lstm with trust gates for 3d human action recognition, European Conference on Computer Vision – P. 816–833.

Hou Y., Li Z., Wang P., Li W. (2016). Skeleton optical spectra-based action recognition using convolutional neural networks, IEEE Transactions on Circuits and Systems for Video Technology – 28 (3). – P. 807–811.

Li M., Chen S., Chen X., Zhang Y., Wang Y., Tian Q. (2019). Actional-structural graph convolutional networks for skeleton-based action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, – P. 3595–3603.

Li C., Hou Y., Wang P., Li W. (2017). Joint distance maps based action recognition with convolutional neural networks, IEEE Signal Processing Letters – 24 (5). – P. 624–628.

Si C., Jing Y., Wang W., Wang L., Tan T. (2018). Skeleton-based action recognition with spatial reasoning and temporal stack learning, in: Proceedings of the European Conference on Computer Vision, – P. 103–118.

Wei D., Lim J. J., Zisserman A., Freeman W. T. (2018). Learning and using the arrow of time, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, – P. 8052–8060.

Zhang W., Zhu M., Derpanis K. G. (2013). From actemes to action: A strongly-supervised representation for detailed action understanding, in: Proceedings of the IEEE International Conference on Computer Vision, – P. 2248–2255.

Zhang J., Zhou W., Xie C., Pu J., Li H. (2016). Chinese sign language recognition with adaptive hmm, in: 2016 IEEE International Conference on Multimedia and Expo, – P. 1–6.

Zolfaghari M., Singh K., Brox T. (2018). Eco: Efficient convolutional network for online video understanding, in: Proceedings of the European Conference on Computer Vision, – P. 695–712.

Du W., Wang Y., Qiao Y. (2017). Rpan: An end-to-end recurrent pose-attention network for action recognition in videos, in: Proceedings of the IEEE International Conference on Computer Vision, – P. 3725–3734.

Zhu W., Lan C., Xing J., Zeng W., Li Y., Shen L., Xie X. (2016). Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks, in: Thirtieth AAAI Conference on Artificial Intelligence, – P. 3697–3703.

Song S., Lan C., Xing J., Zeng W., Liu J. (2017). An end-to-end spatio-temporal attention model for human action recognition from skeleton data, in: Thirty-first AAAI Conference on Artificial Intelligence, – P. 4263–4270.

Zhang B., Yang Y., Chen C., Yang L., Han J., Shao L. (2017). Action recognition using 3d histograms of texture and a multi-class boosting classifier, IEEE Transactions on Image processing – 26 (10). – P. 4648–4660.

Shi L., Zhang Y., Cheng J., Lu H. (2019). Skeleton-based action recognition with directed graph neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, – P. 7912–7921.

Newell A., Yang K., Deng J. (2016). Stacked hourglass networks for human pose estimation, in: European Conference on Computer Vision, Springer, – P. 483–499.

Shi L., Zhang Y., Cheng J., Lu H. (2019). Two-stream adaptive graph convolutional networks for skeleton-based action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, – P. 12026–12035.

He K., Zhang X., Ren S., Sun J. (2016). Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, – P. 770–778.

Xu B., Li J., Wong Y., Kankanhalli M. S., Zhao Q. (2019). Interact as you intend: Intention-driven human-object interaction detection, in arXiv:1808.09796.

Devanne M., Wannous H., Berretti S., Pala P., Daoudi M. (2014). A. Del Bimbo, 3-d human action recognition by shape analysis of motion trajectories on riemannian manifold, IEEE Transactions on Cybernetics – 45 (7). – P. 1340–1352.

Li C., Zhong Q., Xie D., Pu S. (2018). Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation, in: International Joint Conferences on Artificial Intelligence, – P. 786–792.

Ohn-Bar E., Trivedi M. (2013). Joint angles similarities and hog2 for action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, – P. 465–470.

Vemulapalli R., Arrate F., Chellappa R. (2014). Human action recognition by representing 3d skeletons as points in a lie group, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, – P. 588–595.

Ofli F., Chaudhry R., Kurillo G., Vidal R., Bajcsy R. (2014). Sequence of the most informative joints (smij): A new representation for human skeletal action recognition, Journal of Visual Communication and Image Representation – 25 (1). – P. 24–38.

Huang Z., Wan C., Probst T., Van Gool L. (2017). Deep learning on lie groups for skeleton-based action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, – P. 6099–6108.

Shi L., Zhang Y., Cheng J., Lu H. (2019). Skeleton-based action recognition with directed graph neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, – P. 7912–7921.

Rahmani H., Bennamoun M. (2017). Learning action recognition model from depth and skeleton videos, in: Proceedings of the IEEE International Conference on Computer Vision, – P. 5832–5841.

submitted 15.09.2025;
accepted for publication 29.09.2025;
published 30.10.2025

Contact: kabul.kudaybergenov@gmail.com