

Skeleton-based Large-Scale Joint Matrix and Dual Long-Short Cascade Network for Human Action Recognition

Yitong ZHOU, Lei Yue YAO, Chao ZENG, Qing YE*

College of Computer Science, Jiangxi University of Chinese Medicine, Nanchang 330006, China
13576255881@163.com, lei yue_yao@163.com, zc0201082023@163.com,
jxyeqing@163.com (*Corresponding author)

Abstract: The skeleton-based methods have received extensive attention in human action recognition. However, the simple skeleton sequence representation is insufficient for extracting the rich spatio-temporal information, as many two-branch networks typically incur high computational costs. This study introduces a spatio-temporal representation method for a Large Scale Joint Matrix (LSJM), which can capture rich spatial semantic information and improve the recognition performance in deep convolutional neural networks. Furthermore, this paper proposes a Dual Long-Short Cascade Network (DLSCNet), which can fully exploit the temporal information related to Long-Short time scales and which features an increased computational efficiency. This approach is evaluated and validated on two benchmark datasets and one self-collected dataset, namely Florence-3D, UTKinect-3D, and HanYue-3D. The experimental results demonstrate that this approach can effectively support large neural networks and improve the computational efficiency while maintaining the recognition accuracy.

Keywords: human skeleton, action recognition, skeleton sequence representation, spatio-temporal information, Dual Long-Short Cascade Network.

1. Introduction

Human action recognition (HAR) based on deep learning has attracted extensive attention and is widely used in areas such as human-machine interaction, video surveillance, and intelligent healthcare. The availability of accurate and low-cost depth cameras (e.g. Intel RealSense, Orbbec Astra) combined with real-time skeleton estimation algorithms (Diaz-Rojas & Myowa, 2024) has simplified the extraction of human skeletal data from video (Chen et al., 2025). 3D skeletal data provides reliable joint coordinates that capture human motion trajectories. Compared with RGB images and depth maps, skeleton data is more robust to variations in clothing, background, and lighting, and handles occlusion more effectively (Sun et al., 2022).

Many previous methods based on handcrafted skeleton features (Nan et al., 2024) can perform well only on simple or single-type actions, and generalize poorly. The features obtained by these methods are shallow, typically targeted towards specific datasets, and rely on the researcher's previous knowledge. So, this work focuses on deep learning for skeleton-based action recognition. Although Convolutional Neural Networks (CNNs) can capture spatial information effectively in action sequences, typical CNN architectures (e.g., VGG, ResNet) have large model sizes. Significantly, the receptive field gradually expands as the number of CNN layers increases. Restricted by the small number of joints

in the skeleton data, the size of the constructed input matrix is much smaller than it is required by the network. If up-sampling is applied to obtain the appropriate dimensions, it leads to noise and non-interpretable semantic information in the data. Inspired by Liu et al. (2019), a large-scale input matrix for adapting deep CNN network models was constructed. This method selects multiple key joints as roots to construct a tree-structured skeleton image, and obtains multiple sequences through depth-first traversal. It then semantically encodes each joint channel of these temporally ordered 2D grids with motion information such as relative displacement and velocity.

Meanwhile, multi-scale two-stream architectures have achieved remarkable success in video action recognition. Human motion inherently includes both long-stride and short-stride motion, and the visual tempo helps estimate motion trajectories (Chang & Jazayeri, 2018). Yet the visual tempo is often overlooked and temporal multi-scale modeling has not been fully explored in skeleton-based methods. Therefore, fast-motion features are explicitly separated from slow-motion features to better capture temporal information in skeleton sequences. This idea is inspired in part by SlowFast (Feichtenhofer et al., 2019), which is motivated by biological studies of the primate visual system. However, its parallel dual-pathway design with lateral connections significantly increases computational complexity.

Studies have found that visual information is processed in a cascade from the primary to higher visual cortex (Van Essen & Maunsell, 1983). According to the Reverse Hierarchy Theory in cognitive neuroscience, humans recognize objects from global to local (Hochstein & Ahissar, 2002). Based on these findings, a cascade connection strategy is proposed. This coarse-to-fine process reflects human motion perception, where a global judgement is formed first and local details are incorporated only when needed. To address the high computational cost of parallel dual-stream networks, a cascade dual-branch model is introduced.

This work delivers three main contributions. First, it proposes a Large-Scale Joint Matrix (LSJM), a unified representation for 3D skeleton data with rich spatial semantic encoding. By organizing joint motion features into a structured large-scale matrix, LSJM improves the compatibility of large-scale deep neural networks, thereby mitigating overfitting. Second, it introduces a multi-frequency temporal sampling strategy that explicitly separates fast-motion and slow-motion features. The dual-scale temporal modeling allows the network to capture both short-stride and long-stride motion within a unified temporal framework. Finally, a cascade dual-branch network is presented, in which global-to-local processing improves the computational efficiency without compromising the recognition accuracy and aligns with the hierarchical motion perception.

The remainder of this paper is organized as follows. Section 2 reviews the background of the proposed method. Section 3 describes the architecture of the proposed network in detail. Then, the results of the experiments which were carried out are presented in Section 4. Finally, Section 5 concludes this paper and it also outlines possible future research directions.

2. Related Work

The significant research is reviewed from two perspectives: the input representation of skeleton sequences, and multi-stream networks for action recognition.

2.1 3D skeleton Representation

Skeleton-based action recognition often converts 3D sequences into pseudo-images for CNN

processing. This format combines spatial and temporal information and is widely used for feature learning.

Ke et al. (2017) connect the skeleton joints for each frame in series and use four reference joints to calculate the relative positions of the other joints. The sequence-generated grey-scale images are then concatenated and transformed into three segments, thus preserving the spatial-temporal features. However, this image-based transformation may introduce discrepancies and spatial information loss. Li et al. (2017) used a translation scale-invariant image mapping method to divide the bone sequence into five parts according to the physical structure of the human body, and then the coordinates of the joint sequence (x,y,z) corresponding to the (R, G, B) of the image for each frame in the frame series. While this approach includes more spatial structural information, the skeletal joints are still isolated from each other, and there are external dependencies between the parts of the human body. Yang et al. (2018) proposed a new skeleton sequence arrangement, using the centre of the spine as the root node to transform the conventional skeleton into a tree-structured skeleton with depth-first traversal, which best preserves the spatial relationships between the joints in the original skeleton structure.

Caetano et al. (2019a) proposed SkeleMotion, a new representation of skeleton joint sequences based on motion information. It creates joint motion structures from inter-frame differences and uses them to encode temporal dynamics by computing magnitude and orientation values. Caetano et al. (2019b) used SkeleMotion in combination with the reference joints and tree-structure skeletons. Li et al. (2018) used two features, the original skeleton sequence and the skeleton motion, independently fed into the network in a two-stream mode, with the inter-frame time difference of each joint used for the skeleton motion. However, recognizing an action includes both process and result, the motion cannot be reflected by the inter-frame difference alone, and the change from the first frame, i.e. the cumulative frame difference, should also be added. Other representations include fine spatio-temporal coded skeleton images (Liu et al., 2019) and rectangular joint displacement maps (Zhu et al., 2019). Two earlier works introduced a Dense Joint Motion Image (DJMI) to encode motion in each

pixel (Yao et al., 2021) and a 3D floating-point matrix that captured spatio-temporal information effectively but was prone to overfitting in deep networks (Yao et al., 2020).

2.2 Multi-stream Networks for Motion Information

Multi-scale motion information is widely used in action recognition. Javed et al. (2022) proposed a multi-stream CNN model for multi-view HAR recognition using both depth and skeleton data. Unlike the traditional dual-stream architectures, Feichtenhofer et al. (2019) proposed the SlowFast network dual-channel network, which is based on visual tempos of fast and slow. Zhang et al. (2019) designed a DTPN to sample video frames at different frame rates to construct frame pyramids for input video of arbitrary length to capture fast and slow-paced temporal information. Nevertheless, the approach of predefined tempos in the input pyramid is not extensible, and this method requires multiple frames, especially with large sample rates. Thus, Yang et al. (2020) proposed the Temporal Pyramid Network (TPN), which can be flexibly integrated into the backbone network as an assistant module in a plug-and-play manner and utilizes the hierarchical structure within the backbone network to form features of different depths thereby modelling various visual rhythms. However, limited by the network’s temporal modelling ability, it cannot take advantage of low-level features, and the fine-grained temporal dynamics in the higher levels is underutilized. Jie et al. (2020) integrated TPN into SlowFast networks to capture multilevel features.

Since the aforementioned RGB-based multi-tempo methods are not designed for skeleton data, research has focused on developing

specialized networks. Yang et al. (2019) suggested a lightweight dual-feature dual-motion network (DDNet) based on a skeleton, which was designed using a three-stream CNN architecture, in which the input features were slow-motion features, fast-motion features, and joint collection distance (JCD) features. Both Yin et al. (2023) and Liu et al. (2025) improved DDNet, Yin et al. (2023) changed the two-scale information to three-scale information, i.e., slow, average speed and fast. In application domains demanding real-time performance, such as human-computer interaction, recent works by Chang et al. (2023) and Qi et al. (2024) show that computational efficiency is a key challenge. This motivates the proposed cascade dual-branch network, which improves efficiency without reducing accuracy.

3. Proposed Method

3.1 Overall Framework

The overall structure of the proposed method is shown in Figure 1. Firstly, as shown in Figure 1 (a), a large-scale 3D floating-point matrix (LSJM) is proposed to solve the problem of a small number of human skeleton sequences being unable to adapt to large-scale networks. It converts the original joint coordinates into feature sequences with physical meaning through tree-structured skeletons and feature selection (subsection 3.2). With its special arrangement, rich joint correlation and spatial semantic information can be captured. Then, a large-scale 3D floating-point matrix (LSJM) (subsection 3.3) is used in order to arrange these feature sequences into input representations suitable for network learning. Then, as shown in Figure 1 (b), a Dual Long-Short Cascade Network (DLSCNet) is proposed to achieve an efficient computation. It

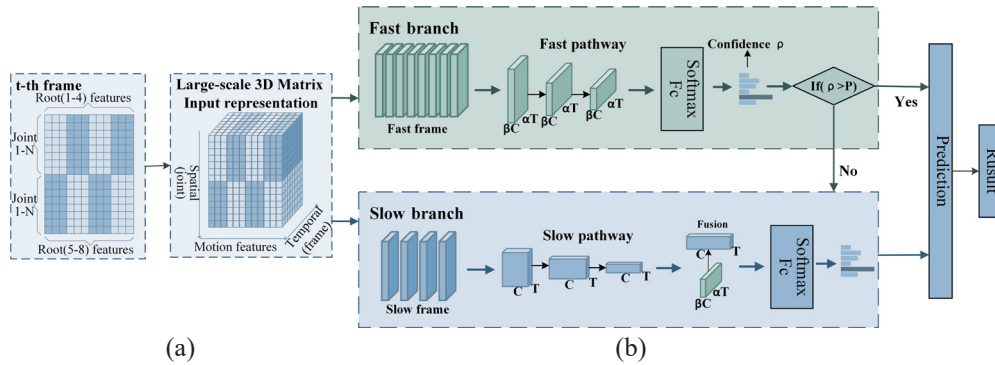


Figure 1. The overall flowchart of the proposed method: (a) Skeleton-based large-scale 3D matrix input representation; (b) A cascade two-branch network for action recognition

uses dual scale motion modeling (3.4) to sample and generate fast and slow samples with different time resolutions from a unified input. Thus, instantaneous changes and global semantics can be captured separately. Finally, the cascade dual branch network (subsection 3.5) integrates the fast and slow branches through a confidence-threshold mechanism. Based on the predicted confidence, the network decides whether the slow branch is activated, enabling the transition from efficient feed-forward computation to fine fusion. This completes the end-to-end computation from input representation to action recognition.

3.2 Depth-first Traversal for Tree-structured Skeletons and Feature Selection

21 joints of the human skeleton were selected, which are modelled in Figure 2(a). In skeleton-based action recognition, each frame's order of skeleton joint chains is predefined and fixed. Normally, traditional skeleton sequences are arranged in the order of joint connections (1-2-3-...-19-20-21). This arrangement ignores dependencies between joints and also adds false connections between body joints that are not tightly connected. For example, joint 21 (upper spine point) is more strongly spatially correlated with the joints on the left arm (9, 10, 11, 12) than joint 8 (right hand). In CNN, neighbouring joints in the joint chain order are learned first in lower levels, meaning that more spatial relations are shared between neighbouring joints. Finding a well-ordered chain can help the model get more spatial semantic information during feature extraction.

Traditional skeletal joints were transformed into a tree-structured skeleton to mimic the motion dependencies between physically connected joints inspired by Yang et al. (2018). Figure 2(b) shows a tree structure constructed based on the spine centre. The edges between joints represent the spatial dependence between joints. The fewer the edges, the more relevant the joint connection is. Then, depth-first traversal of the tree-structured skeleton is conducted to ensure that each node obtains the contextual information from child and bi-parent nodes. The traversal sequence containing 41 joints is obtained as shown in Figure 2(c): (2, 21, 3, 4, 3, 21, 5, 6, 7, 8, 7, 6, 5, 21, 9, 10, 11, 12, 11, 10, 9, 21, 2, 1, 13, 14, 15, 16, 15, 14, 13, 1, 17, 18, 19, 20, 19, 18, 17, 1, 2).

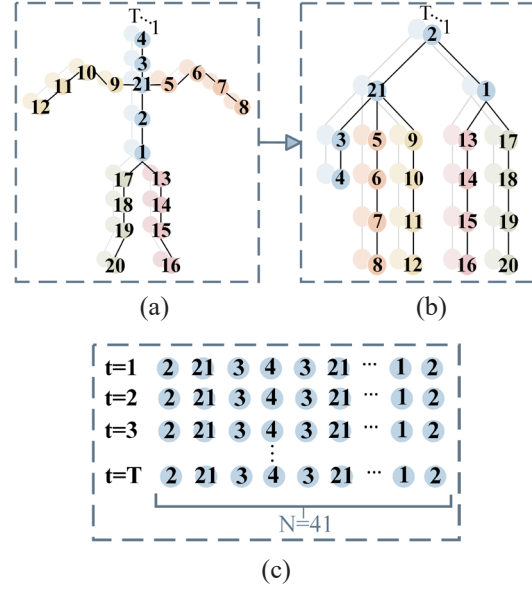


Figure 2. Human skeleton sequence chain with 21 key-points T frames

Each frame of the action video contains a row of joint chains, with a channel attached to each joint point. Usually, each channel represents a joint axis, and each column represents the value of the coordinates of a particular joint. In this paper, relative coordinates are used instead of directly using the 3D coordinates of the joints because the coordinates of the original data are affected by the distance between the camera and the monitored person. 3D motion features are extracted at each joint point to better explore the motion information. To avoid the influence of noise in the feature extraction process, only basic motion and geometric features are selected. This includes the direction of instantaneous motion, instantaneous velocity, instantaneous displacement between frames, the relative position to the first frame, total cumulative displacement, and acceleration.

It is constructed as follows:

1. **Relative position:** The relative position is expressed as the position of the joint coordinates of the current frame relative to the first frame. Equation (1) shows the calculation method for the relative position between two frames:

$$p = \{x', y', z'\} = \begin{cases} x' = x_j - x_0 \\ y' = y_j - y_0 \\ z' = z_j - z_0 \end{cases} \quad (1)$$

where $\{x', y', z'\}$ represents the relative position coordinates of the current j -th frame, $\{x_j, y_j, z_j\}$

represents the actual coordinates of the current j -th frame of the joint point. $i=0$ in $\{x_i, y_i, z_i\}$, represents the coordinates of the first frame used as a reference.

2. Instantaneous displacement between frames and total cumulative displacement:

The instantaneous displacement between frames indicates the displacement of the joint between neighbouring frames. The total cumulative displacement indicates the displacement of the joint from the first frame. Equation (2) provides the formula for displacement calculation:

$$D = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2 + (z_j - z_i)^2} \quad (2)$$

where the instantaneous displacement corresponds to $i=j-1$, while the cumulative displacement corresponds to $i=0$.

3. **Velocity and acceleration:** For frame i , the velocity vector v_i of the joint is $v_i = \frac{D_i}{\Delta t}$, where D_i denotes the displacement of the i -th frame, which can be taken as instantaneous or cumulative displacement to generate the instantaneous and cumulative velocities. $\Delta t = \frac{1}{FPS}$ is the time span between frames and FPS is the frame rate.

The acceleration vector is expressed in equation (3):

$$A_i = \frac{v_i - v_j}{(i - j) \cdot \Delta t} \quad (3)$$

where $j=0$ when calculating the motion acceleration in relation to the first frame. When $j=i-1$, instantaneous acceleration is calculated, that is, the acceleration between frames.

4. **Joint angle:** For movements with small displacements like gripping and waving, the change in joint angles is a significant geometric feature. Therefore, joint angles are incorporated as geometric features, as shown in Figure 3.

The joint angle consists of the space vector formed by three neighboring joint points. This can be calculated by using the cosine theorem as expressed equation (4):

$$\cos \theta = \cos(\overline{J_m J_1}, \overline{J_n J_1}) = \frac{\overline{J_m J_1} \times \overline{J_n J_1}}{|\overline{J_m J_1}| \times |\overline{J_n J_1}|} \quad (4)$$

where θ is the joint angle, while $\overline{J_m J_1}$ and $\overline{J_n J_1}$ are two space vectors which consist of three joints, namely J_m, J_1, J_n .

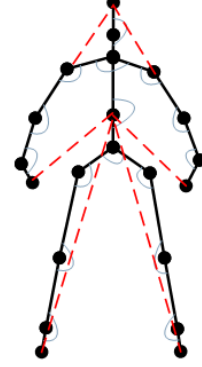


Figure 3. The angles of the joints of the human skeleton

5. Direction of instantaneous motion between frames:

The 3D motion vector of the joint between frames is projected on the XY plane, as equation (5):

$$\overline{d_1^{XY}} = (x_{i+1} - x_i, y_{i+1} - y_i) \quad (5)$$

According to the cosine theorem, θ_{xy} denotes the angle of this vector with respect to the x -axis and is calculated as:

$$\cos(\theta_{xy}) = \frac{\overline{d_1^{XY}} \cdot \overline{X}}{|\overline{d_1^{XY}}| \cdot |\overline{X}|} \quad (6)$$

where X is a unit vector on the x -axis, usually $(1,0)$. Ultimately, θ_{xy} can be calculated using equation (7):

$$\theta_{xy} = \tan^{-1} \left(\frac{y_{i+1} - y_i}{x_{i+1} - x_i} \right) \quad (7)$$

Therefore, the projected direction angles of the direction of instantaneous motion on the three planes are:

$$\begin{cases} \theta_{xy} = \tan^{-1} \left(\frac{y_{i+1} - y_i}{x_{i+1} - x_i} \right) \\ \theta_{xz} = \tan^{-1} \left(\frac{z_{i+1} - z_i}{x_{i+1} - x_i} \right) \\ \theta_{yz} = \tan^{-1} \left(\frac{z_{i+1} - z_i}{y_{i+1} - y_i} \right) \end{cases} \quad (8)$$

When constructing the feature matrix, the temporal-spatial information of the motion is considered, as well as the geometric features. It is also taken into account that the motion feature should contain information about the cumulative motion is relation to the first frame, as well as information about the inter-frame difference. This is because the factors that recognize and reflect the motion are related both to this process and to the final result of the motion.

3.3 Compact Representation of the Large-scale 3D Floating-Point Matrix

In most skeleton-based studies, the motion features of the joint sequences are quantified and stored in RGB images or grayscale images, which are used to represent the input data of the 3D skeleton. However, being restricted by the characteristics of the image, each pixel point has only three colour channels: R, G and B. The result is that the transformed image requires to be normalized to the range of (0,255), and many normalization methods lead to error representations. The images cannot be stored with high accuracy, which affects the action recognition rate with small inter-class differences. Limited once again by the large scale of common CNN network models (e.g. VGG, ResNet) and the small number of joints in the skeleton, the input scale of the constructed matrices is much smaller than the scale required for pre-trained networks.

To solve the above problem, the motion features obtained from the joint chain are stored in a 3D floating-point matrix. The matrix representation generated by this special arrangement strategy will contain rich spatial and temporal motion information. It can provide suitable large-scale input data for the existing large-dimensional CNN models. The proposed approach is illustrated in Figure 4.

As shown in Figure 4(a), beside the tree- structure skeleton constructed based on the centre of the spine as mentioned in subsection 3.1, the left shoulder, the right shoulder, the left hip, the right hip, the upper end of the spine, the lower end of the spine, and the head as the root nodes respectively, were also selected. The eight tree-structure skeleton maps were then depth-first traversed so that eight sets of skeleton joint chains, each containing $N = 41$ number of joints, were obtained in each frame. C channels of motion feature information are attached to each

joint. The eight joint chains obtained from the t -th frame and the additional motion features are divided into upper and lower groups and arranged as in Figure 4(b). This constructed 2D grid slice contains rich information about the joint spatial structure in a single frame. Finally, as shown in Figure 4(c), single-frame slices are concatenated frame by frame, thus constructing a large-scale 3D floating-point matrix. The expression for the input matrix is:

$$A = N \times C \times T \quad (9)$$

where T is the total number of frames in an action sequence.

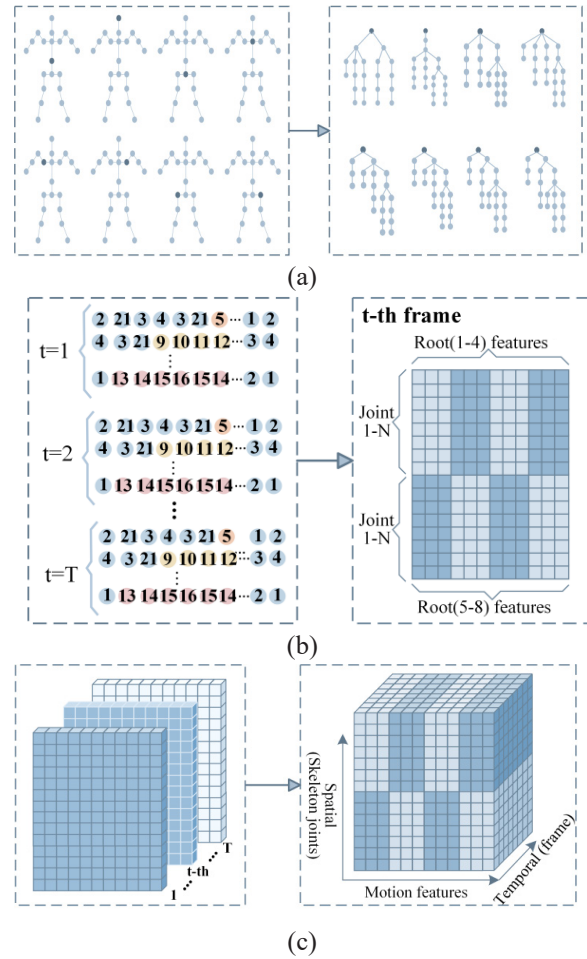


Figure 4. Input representation

3.4 Dual-scale Motion Involving Slow and Fast Motion

The motion features presented above utilize only basic physical motion information and geometric features and lack global motion information. Depending on their different ages, genders, and body sizes, people do not perform the same action at the same scale of motion. Some are faster, and some are slower. To complement the geometric descriptors, motion speed is used here as a

temporal indicator to enable multi-rate sampling. Since geometric features such as distances and angles remain insensitive to temporal frequency, introducing slow-fast sampling allows the model to explicitly capture the complementary temporal dynamics that single-rate representations would otherwise miss. So, two scales of motion features are introduced, i.e. fast motion and slow motion. The matrix slices are obtained by sampling at different frame rates to obtain the two-scale motion features. Conventional two-stream networks usually have a similar network structure and number of parameters. Here, the two-channel network for fast and slow motions utilizes convolutional models to capture information with different temporal changes. The difference between the two pathways lies mainly in the number of channels. The structure of the proposed dual-scale slow-fast pathways and their cascade fusion follows the architecture illustrated in Figure 1(b).

3.4.1 Slow Branch

The slow pathway is used for processing slow motion features and focuses more on spatial modeling capabilities. For the slow path, the matrix slices are sampled from the input sequence at a sampling rate λ . The value of λ will be verified in the subsequent experiments. If the number of frames after slow path sampling is F , the original sequence contains F/λ frames. The convolution model of the slow path can use any pre-trained model as the backbone network. The channel capacity of the slow pathway is denoted as C .

3.4.2 Fast Branch

The fast pathway is used for dealing with fast action motion features and it focuses more on temporal information changes. Unlike the slow path, it samples at a higher frame rate in steps. Thus, for the same action video, the number of frames sampled for the slow path is denoted by the αF frame. α is the frame sampling ratio between the fast and slow pathways. The channel capacity for the fast channel is set to β times that of the slow channel, i.e. βC ($\beta < 1$). This makes the fast track lighter and more efficient. For fast pathways, temporal down-sampling is avoided during feature extraction and global pooling is performed only before the final classification. This make it possible to achieve high-resolution features in the temporal dimension (enhancing the temporal modeling capabilities), while moderately reducing the spatial capacity in the spatial dimension.

3.5 Cascade Dual-branch Networks

Traditional fast-slow networks are fusing the two paths through lateral connections therefore two-stream networks that simultaneously process fast and slow visual rhythms are computationally expensive. This paper proposes a cascade fast-slow network to mimic the reception of visual information delivered cascade by cascade as the human brain senses movements. This cascade formulation also alleviates the computational burden typically observed for dual-branch models, leading to a more efficient processing pipeline. The input features are first sampled at a high frequency to obtain the fast-motion matrix slices, which are fed into the fast-channel network to obtain the corresponding fast-motion feature matrix. A Softmax layer is connected after the fully connected layer of the recognition network, where the activation function of Softmax is expressed as:

$$\text{Softmax}(z_i) = \frac{e^{z_i}}{\sum_i z_i} \quad (10)$$

where Z is an array of outputs from multiple neurons in the previous layer and denotes the i th element in Z .

The Softmax layer allows to obtain the probability values of each category and the maximum confidence value is taken. Then the obtained confidence level is compared with the predefined threshold value. If it is greater than the threshold value, the final recognition result is obtained directly. Otherwise, the input matrix is then sampled at a high frequency to obtain the slow motion features through the slow channel network. The previous fast motion features are fused into the slow pathway. The two feature categories are connected and then used as inputs to the fully connected classifier to obtain the final recognition result.

4. Experiments

4.1 DataSets

The proposed method is evaluated on three skeleton datasets: Florence-3D, UTKinect-3D, and a self-collected HanYue-3D.

Florence-3D: Collected with Kinect, containing 9 daily actions performed 2-3 times by 10

subjects, resulting in 215 sequences. Each skeleton has 15 joints.

UTKinect-3D: Captured with a stationary Kinect, containing 10 action categories performed twice by 10 subjects. Twenty joints were recorded, out of which 15 were used in this study.

HanYue-3D: A self-collected dataset recorded with Kinect v2, containing 15 action categories and 413 samples in total. Each skeleton includes 25 joints, out of which 21 were used in this study.

4.2 Training Implementation Details

All the experiments were implemented using TensorFlow-GPU and Keras on a workstation equipped with an NVIDIA GTX-3060 GPU and an Intel i7-11700KF CPU. The models were trained for 200 epochs with a batch size of 32, and the same training-testing split was applied across all experiments.

To ensure a consistent temporal length, all sequences were normalized at a fixed number of frames, and the optimal scale for each dataset was selected after generating multiple matrix sizes. Different datasets include different numbers of joints.

The proposed cascade dual-branch network is built upon ResNet50, which was fine-tuned to accommodate the proposed LSJM input and cascade structure. The cascade dual-branch network is also compared with a parallel two-branch variant and a short-pathway-only baseline model.

4.3 Results

In this subsection, the effectiveness of the key components was first validated, including the LSJM input representation, cascade network design, and data augmentation. Then, a comprehensive evaluation of the overall framework was conducted.

4.3.1 Effectiveness of LSJM Input Representation

To evaluate the influence of input matrix scales, the performance of the Small-Scale Joint Matrix (SSJM) was compared with that of the Large-Scale Joint Matrix (LSJM) across different CNN models on the Florence-3D, UTKinect-3D, and HanYue-3D datasets.

As shown in Table 1, the test accuracy varies considerably across different CNN models and input representations, with SSJM generally

Table 1. The comparison of test accuracy between LSJM and SSJM on three public datasets

Typical CNN		HanYue-3D	Florence-3D	UTKinect-3D
Resnet-50	LSJM	81.57%	92.50%	89.74%
	SSJM	71.05%	87.50%	79.48%
Resnet50-v2	LSJM	78.94%	89.99%	69.23%
	SSJM	73.68%	80.00%	61.53%
Resnet-101	LSJM	78.94%	89.99%	89.74%
	SSJM	71.05%	82.49%	79.48%
Resnet-101v2	LSJM	81.57%	89.99%	84.61%
	SSJM	73.68%	85.00%	82.05%
DenseNet-121	LSJM	65.78%	87.50%	
	SSJM			
Resnet-152	LSJM	75.00%	87.50%	89.74%
	SSJM	69.73%		82.05%
VGG16	LSJM	76.31%	94.99%	71.79%
	SSJM			69.23%
LeNet5	LSJM	80.26%		92.30%
	SSJM	76.31%		89.74%
AlexNet	LSJM		89.99%	87.17%
	SSJM		85.00%	74.35%
ZFNet	LSJM	75.00%		89.74%
	SSJM			

achieving a lower test accuracy than LSJM. The HanYue-3D LSJM performs better on ResNet-50 and ResNet-101v2, while the Florence-3D LSJM performs better on ResNet-50 and VGG16. The LSJM of the UTKinect-3D dataset performs well on ResNet-50, ResNet-152, LeNet5, and ZFnet. Based on this, ResNet-50 was selected as the basic model for subsequent cascade networks.

The training process visualization in Figure 5 demonstrates that LSJM enables the network to converge faster and achieve a better fitting. As shown in Figure 6, three different large-scale matrix arrangement strategies were also compared. The results in Table 2 indicate that the strategy simulating the physical arrangement of human joints (Strategy 2) achieved the best performance.

4.3.2 Analysis and Evaluation of the Proposed Cascade Networks

This subsection provides an ablation study on the proposed Dual Long-short Cascade Network using the Florence-3D dataset.

4.3.2.1 Fusion Method

Five feature fusion strategies for the fast and slow pathways were compared: Late fusion, Average fusion, Concatenation(Concat), weighted summation, and Multi-Scale Convolutional Weighted Summation (MCWS). As shown in Table 3, the fusion method employing MCWS achieves the highest recognition accuracy.

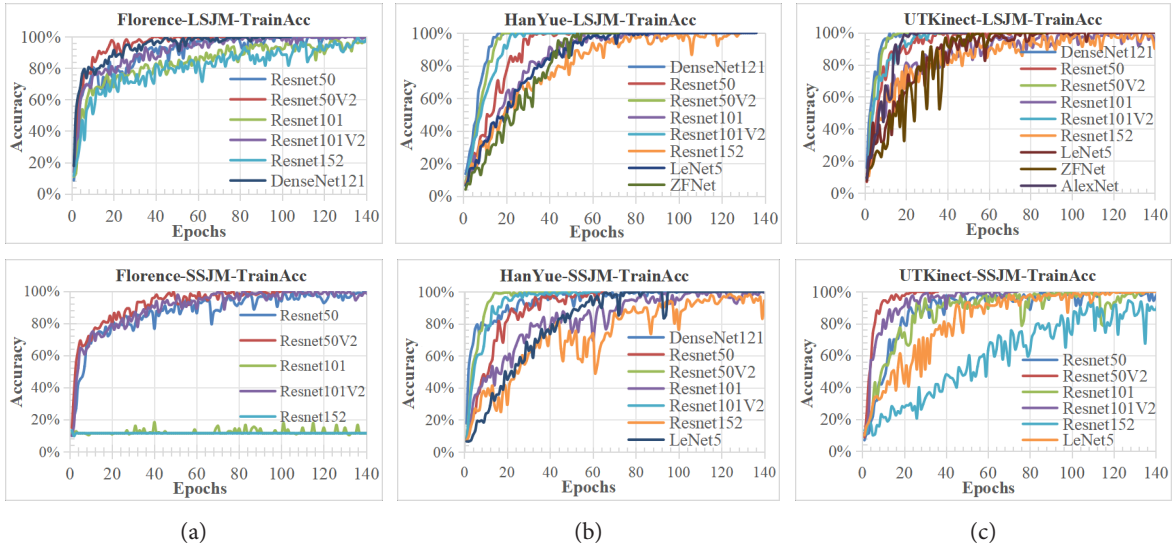


Figure 5. The training curves of the evaluated models using LSJM and SSJM on three public datasets

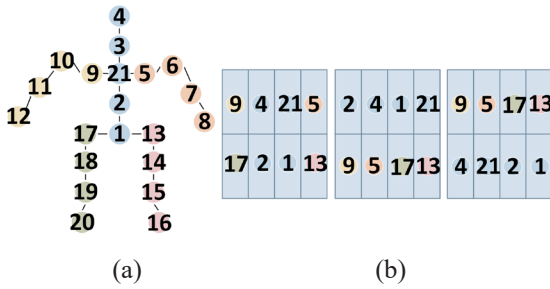


Figure 6. Different arrangement strategies

Table 2. The accuracy comparison for three arrangement strategies on the HanYue-3D dataset

Backbone	HanYue-3D		
	Strategy1 (ACC)%	Strategy2 (ACC)%	Strategy3 (ACC)%
Resnet-50	69.73%	81.57%	63.15%
Resnet-101	76.31%	78.94%	72.36%
DenseNet-121	63.15%	65.78%	64.47%
LeNet-5	75.00%	80.26%	77.63%
VGG16	68.42%	76.31%	71.05%

Table 3. Comparison of different fusion methods

Fusion Method	Acc. (%)	Top-3(%)	Top-5(%)	GFlops	Time-cost
Late	62.50	92.50	100	0.1185	11.57
Average	85.00	95.00	97.50	0.5414	25.02
Concat	85.00	95.00	97.50	0.5414	26.47
Weighted Sum	89.99	92.50	97.50	0.5414	27.15
MCWS	92.50	95.00	100	0.6839	27.52

4.3.2.2 Channel Capacity Ratio and Frame Sampling Ratio

The effects of the frame sampling ratio (α) and channel capacity ratio (β) were analyzed. Table 4(a) shows the accuracy and computational efficiency for different α values. $\alpha = 1/4$ achieves the best computational efficiency while maintaining a high accuracy. Table 4(b) shows that the best performance is achieved on a single branch when $\beta = 1/4$. It is worth noticing that from $\beta = 1/4$ to $\beta = 1/8$, although the computational complexity is decreased by 0.0585 GFlops, the accuracy drops significantly in the accuracy.

Table 4. Comparison of different combinations of channel capacity ratio and frame sampling ratio

α	Acc (%)	Top-3 (%)	Top-5 (%)	GFlops	Time-cost
1/2	90.00	95.00	97.50	0.6839	33.64
1/3	85.00	92.50	95.00	0.6803	33.62
1/4	90.00	95.00	95.00	0.6780	31.44
1/5	85.00	92.50	97.50	0.6768	31.92
1/6	77.49	95.00	100	0.6756	30.42

(a)

	Acc (%)	Top-3 (%)	Top-5 (%)	GFlops	Time-cost
Fast-only	87.50	92.50	97.50	0.6603	54.95
Slow-only	82.50	95.00	100	0.0335	60.70
$\beta = 1/2$	92.50	95.00	97.50	0.6839	29.71
1/4	95.00	97.50	97.50	0.5028	27.84
1/8	87.50	97.50	100	0.4443	27.65

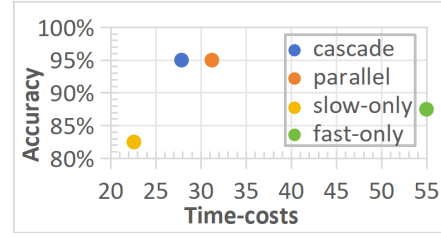
(b)

4.3.2.3 Parallel and Cascade

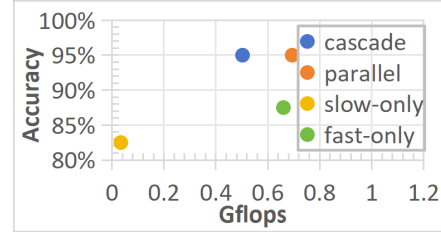
The accuracy and efficiency trade-offs of the dual-branch networks in parallel and cascade configurations were compared. Table 5 and Figure 7 demonstrate that the cascade structure effectively reduces the computational complexity and time consumption while maintaining the recognition accuracy.

Table 5. Accuracy vs. efficiency comparisons

	ACC (%)	GFlops	Time-cost
Slow-only	82.50%	0.0342	22.57
Fast-only	87.50%	0.6603	54.95
parallel	95.00%	0.6938	31.2
cascade	95.00%	0.5028	27.84



(a)

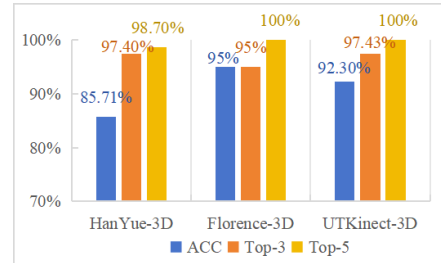


(b)

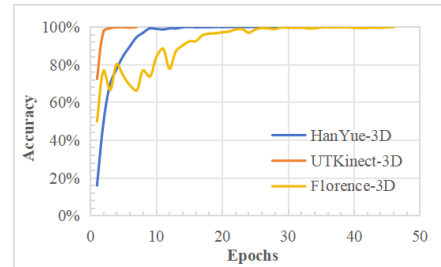
Figure 7. Comparison of parallel and cascade configurations

4.3.3 Data Augmentation

To mitigate over-fitting caused by the lack of sample data, data augmentation was employed. New samples were generated by scaling the original samples to mimic different body sizes, and Euler-angle rotation was applied to simulate movements from different perspective. Figure 8 shows the training and testing results on the Florence-3D, HanYue-3D and UTKinect-3D datasets.



(a)



(b)

Figure 8. The training and testing results

4.3.4 Comparison with Other Classical Networks or with the Latest Networks

LSJM, cascade networks (DLSCNet), and data augmentation strategies were integrated and they were compared with other existing methods. As

shown in Table 6, the proposed method achieved a 95%, 85.71%, and 92.3% accuracy on the Florence-3D, HanYue-3D, and UTKinect-3D dataset, respectively.

The confusion matrix in Figure 9 shows that for actions with similar motion patterns, such as standing up and sitting down that are difficult for many methods to distinguish, the proposed approach achieves a 100% accuracy on the

Florence-3D and UTKinect-3D datasets. Similarly, it works equally well for actions like answering a phone call and looking at one's watch on the HanYue-3D and Florence-3D datasets.

The precision, recall, and F1 scores reported in Table 7 further demonstrate that the proposed cascade network significantly reduces both false positives and false negatives across the three datasets compared with the slow-path-only baseline model.

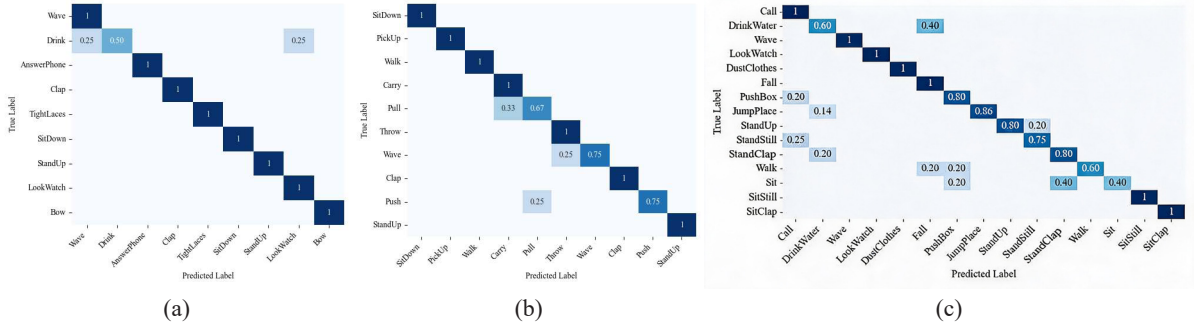


Figure 9. Confusion matrix of the proposed method

Table 6. Comparison of the proposed method with different models

Method	ACC(%)		
	Florence-3D	HanYue-3D	UTKinect-3D
DataSet creator's method	82.20		
DataSet creator's method			90.92
TMFAA+Multi-Scale+Data Aug (Zhu et al., 2019)	94.74	83.87	94.74
DJMI+ ZfNet (Yao et al., 2021)	92.50		91.84
SNV (Yang & Tian, 2014)			88.89
Lie group+CNN (Cai et al., 2020)	93.00		
SSJM+ResNet50	87.50	71.05	79.48
SSJM+ResNet101	82.49	71.05	79.48
LSJM+ResNet50	92.50	81.57	79.48
LSJM+ResNet101	89.99	78.94	89.74
LSJM+ResNet50+Data Aug	92.50	85.71	92.30
LSJM+ResNet101+Data Aug			89.74
The proposed method	95.00	85.71	92.30

Table 7. Detailed metrics of the slow-only baseline model and the proposed method

Dataset		Hanyue-3D	Florence-3D	UTkinect-3D
Slow-only	TP	4.00	3.67	3.10
	TN	70.73	34.78	34.30
	FP	1.13	0.78	0.80
	FN	1.13	0.78	0.80
	ACC	77.92	82.50	79.49
	Precision	77.44	82.96	79.50
	F1-Score	76.61	80.00	78.13
	Recall	77.60	81.11	78.33
Proposed (cascade-slow-fast)	TP	4.40	4.22	3.60
	TN	71.13	35.33	34.80
	FP	0.73	0.22	0.30
	FN	0.73	0.22	0.30
	ACC	85.71	95.00	92.31
	Precision	87.66	95.56	92.67
	F1-Score	84.96	93.83	91.59
	Recall	85.38	94.44	91.67

4.4 Discussion

4.4.1 On the Efficiency of LSJM in Mitigating Overfitting

The experimental results demonstrate that LSJM, as an input representation, effectively addresses the compatibility issues between SSJM and large-scale deep neural networks. The overfitting and sharp drop in test accuracy observed for SSJM within deep neural networks result from dimensional mismatch. As network depth increases, the stacked convolution and pooling layers cause the small-scale input matrix to shrink rapidly to an uncomputable size. This problem cannot be resolved by simply increasing the amount of data. LSJM ensures dimensional compatibility with complex model architectures by providing a higher-dimensional input. Its scalable design allows the input size to be modified by adjusting the root node combinations. This flexibility enables LSJM to adapt to neural networks of different depths, which is key to its effectiveness.

4.4.2 Theoretical Implications and Efficiency of the Dual Long-Short Cascade Network Design

The proposed cascade Dual Long-Short network design is supported by experimental data. The success of the multi-scale fusion strategy lies in the fact that it can integrate the characteristics of fast and slow pathways across different spatial-temporal scales. The selection of $\alpha = 1/4$ and $\beta = 1/4$ represents the optimal balance between precision and efficiency.

More importantly, the success of the cascade structure is closely related to the functional roles of its two pathways: the slow pathway captures global information, whereas the fast pathway focuses on local details. This mechanism aligns with the reverse hierarchy theory in cognitive neuroscience, which states that visual processing progresses from coarse global perception to fine-grained local analysis. DLSCNet simulates this process in the computational architecture. This biological plausibility provides a theoretical explanation for why the proposed approach achieves significant computational efficiency gains without sacrificing accuracy. This is not merely

an engineering optimization but a computational model inspired by natural intelligence.

5. Conclusion

This paper proposes a novel skeleton-based input representation, the Large Scale Joint Matrix (LSJM), and introduces the Dual Long-Short Cascade Network (DLSCNet). The effectiveness of the presented method is verified through experiments on three datasets.

The research results show that the LSJM can extract rich spatial semantic information while enhancing compatibility with large-scale deep neural networks and effectively mitigating overfitting. The empirical findings indicate that the design of DLSCNet is aligned with the Reverse Hierarchy theory in cognitive neuroscience. Based on this framework, motion-related visual information is processed progressively from the primary to higher visual cortex, supporting a global-to-local form of multi-scale temporal integration that parallels the dual-scale cascade strategy in the proposed model.

The main advantage of the model is its ability to significantly reduce computational costs while maintaining a high recognition accuracy. This approach can be applied to applications with strict real-time requirements, such as human-computer interaction.

Although the proposed method performs well, its frame sampling strategy is uniform, which may result in the loss of key information in sparse sampled segments. Future work could integrate key-frame extraction strategies to mitigate the potential loss of informative frames in single-branch sparse sampling, which may improve model performance and robustness.

Acknowledgement

This research was supported by the National Natural Science Foundation of China under Grant No. 62366023, the Scientific and Technological Projects of the Nanchang Science and Technology Bureau under Grant GJJ2202613.

REFERENCES

- Cai, L., Liu, C., Yuan, R. et al. (2020) Human action recognition using Lie Group features and convolutional neural networks. *Nonlinear Dynamics*. 99(4), 3253–3263. <https://doi.org/10.1007/s11071-020-05468-y>.
- Caetano, C., Sena, J., Bremond, F. et al. (2019a) SkeleMotion: A new representation of skeleton joint sequences based on motion information for 3D action recognition. In: *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance, 18-21 September 2019, Taipei, Taiwan*. New York, USA, IEEE. <https://doi.org/10.1109/AVSS.2019.8909840>.
- Caetano, C., Bremond, F. & Schwartz, W. (2019b) Skeleton image representation for 3D action recognition based on tree structure and reference joints. In: *2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images, 28-30 October 2019, Rio de Janeiro, Brazil*. New York, USA, IEEE. pp. 16-23.
- Chang, C. & Jazayeri, M. (2018) Integration of speed and time for estimating time to contact. *Proceedings of the National Academy of Sciences*. 115(12), E2879-E2887. <https://doi.org/10.1073/pnas.1713316115>.
- Chang, V., Eniola, R., Golightly, L. et al. (2023) An exploration into human-computer interaction: Hand gesture recognition management in a challenging environment. *SN Computer Science*. 4, Art. No. 441. <https://doi.org/10.1007/s42979-023-01751-y>.
- Chen, L., Li, J., Li, M. et al. (2025) Skeleton-based action recognition using multibranch adaptive graph convolutional network with pose refinement. *IEEE Transactions on Computational Social Systems*. 12(6), 4689 – 4699. <https://doi.org/10.1109/TCSS.2025.3566733>.
- Diaz-Rojas, F. & Myowa, M. (2024) Estimation of human body 3D pose for parent-infant interaction settings using Azure Kinect and OpenPose. *MethodsX*, 13, Art. No. 102861. <https://doi.org/10.1016/j.mex.2024.102861>.
- Feichtenhofer, C., Fan, H., Malik, J. et al. (2019) SlowFast networks for video recognition. *2019 IEEE/CVF International Conference on Computer Vision (ICCV), 27 October - 2 November 2019, Seoul, Korea (South)*. New York, USA, IEEE. <https://doi.org/10.1109/ICCV.2019.00630>.
- Hochstein, S. & Ahissar, M. (2002) View from the top: Hierarchies and Reverse Hierarchies in the Visual System. *Neuron*. 36(5), 791–804. [https://doi.org/10.1016/s0896-6273\(02\)01091-7](https://doi.org/10.1016/s0896-6273(02)01091-7).
- Javed, M. H., Yu, Z., Li, T. et al. (2022) Hybrid two-stream dynamic CNN for view adaptive human action recognition using ensemble learning. *International Journal of Machine Learning and Cybernetics*. 13, 1157–1166. <https://doi.org/10.1007/s13042-021-01441-2>.
- Jie, Z., Muqing, W. & Weiyao, X. (2020) A two-pathway convolutional neural network with temporal pyramid network for action recognition. In: *2020 IEEE 6th International Conference on Computer and Communications (ICCC), 11-14 December 2020, Chengdu, China*. New York, USA, IEEE. pp. 2448-2452.
- Ke, Q., Bennamoun, M., An, S. et al. (2017) A new representation of skeleton sequences for 3D action recognition. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 21-26 July 2017, Honolulu, HI, USA*. New York, USA, IEEE. pp. 4570–4579. <https://doi.org/10.1109/cvpr.2017.486>.
- Li, C., Zhong, Q., Xie, D. et al. (2018) Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. In: *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI), 13 – 19 July 2018, Stockholm, Sweden*. Washington, DC, USA, AAAI Press. pp. 786–792. <https://doi.org/10.24963/ijcai.2018/109>.
- Li, N. B., Dai, N. Y., Cheng, N. X. et al. (2017) Skeleton-based action recognition using translation-scale invariant image mapping and multi-scale deep CNN. In: *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), 10-14 July 2017, Hong Kong*. New York, USA, IEEE. pp. 601-604.
- Liu, J., Akhtar, N. & Mian, A. (2019) Skepxels: Spatio-temporal image representation of human skeleton joints for action recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 16-20 June 2019 Long Beach, United States*. New York, USA, IEEE. pp. 10–19.
- Liu, J., Gan, M., He, Y. et al. (2025) Multimodal multilevel attention for semi-supervised skeleton-based gesture recognition. *Complex & Intelligent Systems*, 11, Art. No. 189. <https://doi.org/10.1007/s40747-025-01807-x>.
- Nan, M., Trăscău, M. & Florea, A. (2024) Spatio-temporal neural network with handcrafted features for skeleton-based action recognition. *Neural Computing and Applications*. 36(16), 9221–9243. <https://doi.org/10.1007/s00521-024-09559-4>.

- Qi, J., Ma, L., Cui, Z. et al. (2024) Computer vision-based hand gesture recognition for human-robot interaction: A review. *Complex & Intelligent Systems*. 10(1), 1581–1606. <https://doi.org/10.1007/s40747-023-01173-6>.
- Sun, Z., Ke, Q., Rahmani, H. et al. (2022) Human action recognition from various data modalities: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 45(3), 3200 – 3225. <https://doi.org/10.1109/tpami.2022.3183112>.
- Van Essen, D. C. & Maunsell, J. H. (1983) Hierarchical organization and functional streams in the visual cortex. *Trends in Neurosciences*. 6, 370–375. [https://doi.org/10.1016/0166-2236\(83\)90167-4](https://doi.org/10.1016/0166-2236(83)90167-4).
- Yang, C., Xu, Y., Shi, J. et al. (2020) Temporal pyramid network for action recognition. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 13-19 June 2020, Seattle, WA, USA*. New York, USA, IEEE. pp. 588-597. <https://doi.org/10.1109/cvpr42600.2020.00067>.
- Yang, F., Wu, Y., Sakti, S. et al. (2019) Make skeleton-based action recognition model smaller, faster and better. In: *Proceedings of the 1st ACM International Conference on Multimedia in Asia (MMAsia), 15 – 18 December 2019, Beijing, China*. New York, USA, Association for Computing Machinery. <https://doi.org/10.1145/3338533.3366569>.
- Yang, X. & Tian, Y. (2014) Super normal vector for activity recognition using depth sequences. In: *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, 23–28 June 2014, Washington, DC, USA*. Washington, DC, USA, IEEE Computer Society. pp. 804–811.
- Yang, Z., Li, Y., Yang, J. et al. (2018) Action recognition with spatio-temporal visual attention on skeleton image sequences. *IEEE Transactions on Circuits and Systems for Video Technology*. 29(8), 2405–2415. <https://doi.org/10.1109/tcsvt.2018.2864148>.
- Yao, L., Yang, W. & Huang, W. (2020) A data augmentation method for human action recognition using dense joint motion images. *Applied Soft Computing*. 97(B), Art. No. 106713. <https://doi.org/10.1016/j.asoc.2020.106713>.
- Yao, L., Yang, W., Huang, W. et al. (2021) Multi-scale feature learning and temporal probing strategy for one-stage temporal action localization. *International Journal of Intelligent Systems*. 37(7), 4092–4112. <https://doi.org/10.1002/int.22713>.
- Yin, M., He, S., Soomro, T. A. et al. (2023) Efficient skeleton-based action recognition via multi-stream depthwise separable convolutional neural network. *Expert Systems with Applications*. 226(C), Art. No. 120080. <https://doi.org/10.1016/j.eswa.2023.120080>.
- Zhang, D., Dai, X. & Wang, Y.-F. (2019) Dynamic temporal pyramid network: A closer look at multi-scale modeling for activity detection. In: *Computer Vision – ACCV 2018 (14th Asian Conference on Computer Vision, 2-6 December 2018, Perth, Australia)*, part of the book series *Lecture Notes in Computer Science* (LNCS, vol. 11363). Berlin, Heidelberg, Springer. pp. 712-728. https://doi.org/10.1007/978-3-030-20870-7_44.
- Zhu, K., Wang, R., Zhao, Q. et al. (2019) A cuboid CNN model with an attention mechanism for skeleton-based action recognition. *IEEE Transactions on Multimedia*. 22(11), 2977–2989. <https://doi.org/10.1109/tmm.2019.2962304>.



This is an open access article distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.