INFORMATICA, 2026, Vol. 0, No. 0, 1–24

DOI: https://doi.org/10.15388/26-INFOR645

An Efficient Cascade Neural Network for Human Action Recognition

Jianying XIONG[1], Zikang FAN[1], Ning LIU[2], Keyun XIONG[1], Leiyue YAO[1,*]

[1] *College of Intelligent Medicine and Information Engineering, Jiangxi University of Chinese Medicine, Nanchang, China*
[2] *Hanlin Hangyu (Tianjin) Industrial Co., Ltd., Tian Jin, China*
e-mail: leiyue_yao@163.com

Received: January 2026; accepted: August 2026

Abstract. Calculating motion features frame by frame and organizing them into a 3D matrix is a typical CNN-based solution for human action recognition (HAR). With the widespread use of consumer electronics, reducing computational costs and enabling efficient edge-side action recognition have become a research hotspot. In this paper, we extract action key frames via a well-designed algorithm to reduce computational overhead, so that the proposed method can be deployed on mobile electronic devices. Then we construct local and global motion features from these key frames and feed them into a cascade neural network for action recognition. The primary contributions include three aspects. First, the strategic adoption of key frames is introduced to greatly reduce the number of input parameters. The number of key frames can be adjusted to adapt to the temporal scales of different actions. Second, multiple origin points are adopted to construct motion matrices with larger dimensions than those constructed using a single origin point. Thus, deeper neural networks can be employed to achieve higher recognition accuracy. Third, a cascade neural network is proposed for action prediction, which leverages global and local information to achieve better efficiency and accuracy. Experimental results on UTKinect-Action3D, Florence-3D and our self-built HanYue-3D datasets demonstrate that our method achieves accuracy and efficiency competitive with state-of-the-art (SOTA) approaches. Moreover, the flexibility of the proposed method enables users to readily balance effectiveness and efficiency, making it well-suited for resource-constrained mobile devices.

Key words: key frame extraction, cascade neural network, multiple origin points, skeleton-based action recognition, CNN-based action recognition..

1. Introduction

Human action recognition (HAR) has a wide range of industrial applications, such as video retrieval, video captioning, virtual reality and human-computer interaction (Sun *et al.*, 2022). Consequently, many researchers have proposed various solutions, including wearable device-based methods (Yu *et al.*, 2024; Wang *et al.*, 2025), ambient-based

*Corresponding author.

methods (Jain *et al.*, 2023; Wang *et al.*, 2024), monocular camera based methods (Zhang *et al.*, 2025; Zhou *et al.*, 2023), and the depth sensor based methods (Hu *et al.*, 2024). Vision-based methods have grown increasingly popular due to their low cost and extensive use in surveillance systems (Hussain *et al.*, 2024). In particular, methods leveraging depth sensors have proven to be highly effective in representing actions, primarily due to their reliance on skeleton motion data.

Regardless of the type of sensor used, the core goal is to find an effective representation of human actions. Before the emergence of deep learning, HAR representations mainly depended on handcrafted features, such as spatiotemporal interest points (STIP) (Laptev, 2025), histogram of oriented gradients (HOG) (Dalal and Triggs, 2005), silhouettes (Kurban and Yildirim, 2024), etc. The common solution for HAR is "handcrafted features + support vector machine (SVM)" (Herath *et al.*, 2017).

In recent years, with the advancement of deep learning techniques, the accuracy of HAR has continued to improve. Many classical global-feature representations, such as motion history image (MHI) (Yang *et al.*, 2025), static history images (SHI) (Zhang *et al.*, 2016), and optical flow (OF) (Ullah *et al.*, 2019) emerged at that time. The combination of global features with various CNNs was the mainstream HAR paradigm in the early deep learning era. Although global feature-based methods perform well for short-duration action recognition, they suffer from motion trajectory overlap, which causes the loss of spatial and temporal information. To address this issue, the motion energy images (MEI) (Abdelbaky and Aly, 2020) assigns differentiated weights to frames according to their temporal positions. Thus, the losses of motion information can be partly reduced. However, this approach is still not the final solution.

Furthermore, some researchers try to extract context information among frames, such as recurrent neural network (RNN) (Zhang *et al.*, 2026; Giveki, 2024), long short term memory (LSTM) network (Majd and Safabakhsh, 2020). With the ever increasing computational power, some researchers used 3D-CNNs (Jeyanthi *et al.*, 2024; Arif *et al.*, 2021) to extract both temporal and spatial features for HAR. 3D-CNN-based methods have achieved remarkable improvements in recognition accuracy and have been validated for their effectiveness and robustness across diverse applications. However, their high computational cost introduces substantial challenges in model training and fine-tuning. Thus, these methods are difficult to deploy on mobile devices with limited computing resources (Karim *et al.*, 2024).

To effectively and efficiently extract the spatial and temporal motion information of human actions, more and more researchers utilized skeleton/joint motion data in HAR (Zhang, 2019; Li *et al.*, 2023; Xie *et al.*, 2025). The skeleton/joint-based methods have three merits. First, skeleton motion information is strongly related to human subjects. The usage of skeleton motion data can dramatically reduce the unnecessary input data, such as background, illumination, actor's appearance, etc. Second, skeleton data provided by depth sensor contains plentiful 3D-coordinate information which is valuable for human action quantifying or encoding. Moreover, some specific data augmentation algorithms for human actions can be put forward more easily via these quantified data (floating-point values). Third, the quantified/encoded human action can be represented by a 3D float

matrix in small size. Thus, a shallow neural network can also achieve expected accuracy with high efficiency, which is helpful for the deployment in mobile devices.

With the popularity of consumer electronic products, how to reduce the computational requirements and perform efficient behaviour recognition at the edge has become a hotspot in HAR. In this paper, we have introduced an innovative method that enhances efficiency by extracting key action frames. *Firstly*, an algorithm is put forward to extract a certain number of key frames from a long video. Thus, two major problems can be solved in this step. (1) The temporal scale of the human actions can be normalized by fixing the number of key frames, and samples with uniform temporal scales are critical for neural network training. (2) The usage of key frames can dramatically reduce the input parameters, which would greatly improve the efficiency. *Secondly*, a float matrix with flexible scale is propounded to store the quantified motion data of human actions. Compared with motion images of fixed scale, our proposed matrix can be scaled up to better match deeper neural networks, which may achieve higher accuracy. *Thirdly*, the data augmentation strategies that were used in our previous work (Yao *et al.*, 2020) were continued and improved. These strategies were specially designed for human action representation, which are explainable and can be easily expanded by setting different parameters. *Fourthly*, the multi-scale learning theory is introduced to further improve the recognition accuracy of actions at different time scales.

The main contributions of this paper are summarized as follows:

1) Using keyframes instead of all frames for action recognition is the core contribution of this paper. It not only meets the demand of reducing input parameters in consumer electronics to achieve model lightweighting, but also simultaneously addresses the requirement of time dimension normalization.
2) By constructing large-scale motion feature maps, the model can accommodate deeper convolutional layers and consequently deliver superior prediction accuracy. This allows CNN-based method to achieve competitive performance, obviating the need to deploy state-of-the-art, yet computationally intensive architectures such as Vision Transformers, and thereby exhibits stronger adaptability for consumer electronics.
3) A cascaded neural network is put forward to further enhance the accuracy of HAR. This architecture fully exploits the robust anti-interference capability of local features as well as the abundant contextual semantic cues embedded in global features. By introducing confidence-based regulation, the method achieves effective complementarity between local and global representations while maintaining low power overhead.

This paper is structured as follows: Section 2 provides an overview of the latest research on HAR. Section 3 details the key frame extraction algorithm and HAR algorithm. Section 4 describes our experiments and the results. Finally, Section 5 concludes the paper and discusses future work.

2. Related Work

2.1. *Key-Frame-Based Global Feature*

The widespread adoption of depth sensors, including the Kinect v1.0/v2.0 and the Orbbec Astra, has significantly enhanced the precision of human action recognition, as depth images furnish 3D data, thereby presenting an effective approach to pose estimation challenges. Moreover, the majority of classical methods relying on global features greatly benefit from the low-noise data supplied by depth sensors. MHI and EMI, which are generated from skeleton motion data, have replaced those derived from body boundaries. For example, Phyo *et al.* (2019) proposed the colour skeleton motion history image (Colour Skl-MHI), which encodes skeleton frames with colour information to construct a novel global motion representation for action recognition and similar action discrimination. Li *et al.* (2018) similarly employed colour encoding to generate the improved joint trajectory map (IJTM), which was then projected onto the xy, xz, and yz planes for feature extraction. Then, 3 LSTM networks and 3 CNNs were adopted to extract temporal and spatial features for further action recognition. However, despite improvements, JTMs generated from all frames still struggle with motion information loss due to overlapping motion trajectories. Thus, researchers attempt to use a limited number of key frames to generate motion trajectories. Wang *et al.* (2022) extracted key frames by the maximum redundancy coefficient. Yang *et al.* (2023) adopted the autoregressive moving average (ARMA) algorithm to find out key frames. Tan *et al.* (2022) identified key frame in each video segment via a global max pooling operation in the temporal dimension. Dong *et al.* (2022) selected key frames by a hard attention guided frame sampling model. Xia and Xin (2024) claimed that not all frames are positive to HAR, and proposed a key frame sampling module based on rewards and pseudo-labels.

These key frame-based methods have two advantages: 1) Due to the sparsity of key frames, the side effects brought by trajectory overlaps can be minimized. 2) The key frame-based methods can effectively eliminate redundant frames, reduce the length of the motion and uniform the temporal scale of action samples. However, it should be clearly acknowledged that the adverse influence of motion trajectory overlaps is not fully resolved due to the inherent limitations of the global features.

2.2. *Skeleton-Based Local Feature*

Compared with global features, local features are less sensitive to scale, observation angle, and occlusion (Zhang *et al.*, 2025). Therefore, they are more suitable for action recognition in real environment. Consequently, the skeleton-based local features emerged. Du *et al.* (2015) were the first to store a joint coordinates (x, y, z) in the three channels (R, G, B) of a colour pixel. Thus, an action can be encoded as a colour image, and HAR problem is transformed to a well-farmed image classification problem. Inspired by this, many variations have emerged. In Yang *et al.* (2018), a tree structure skeleton image (TSSI) was proposed to improve the action representation of literature (Du *et al.*, 2015). This new

structure enhances the preservation of spatial relations among the joints. These innovative methods have introduced novel quantification techniques for action videos, effectively addressing the issue of information loss during the encoding process. However, there are at least three drawbacks that require careful attention. First, raw three-dimensional coordinate data of human joints cannot be directly adopted for model input, as such coordinate values are susceptible to shifts induced by the depth camera's placement. In our prior studies (Yao *et al.*, 2020), we set the coordinate of the SpineBase joint in the initial frame as the relative coordinate origin. This preprocessing strategy substantially improves the generalization capacity of the proposed method and simultaneously elevates its prediction accuracy. Second, nearly all existing motion-image-based approaches uniformly resize visual frames to fixed dimensions to satisfy the input dimensional constraints of convolutional neural networks (CNNs) during model training. Nevertheless, such simplistic cropping or geometric warping operations frequently inject extraneous noise into encoded motion representations and distort the inherent temporal motion characteristics of human actions. To address this limitation, abundant research efforts have focused on temporal normalization for action sequences, which constitutes a mainstream research direction in motion timing analysis. Representative solutions either supplement missing frames via frame interpolation algorithms or discard redundant frames to align sequence lengths. Third, encoded motion images are compact in size. Although their low resolution leads to good computational efficiency, the insufficient pixel granularity makes it impossible to support multi-layer deep convolution operations, which becomes the bottleneck for optimizing accuracy. With the continuous improvement of hardware computing performance, even for consumer electronics, it has become a key requirement to construct a scalable human motion data representation paradigm, to achieve an ideal balance between computational efficiency and recognition performance.

To advance the research on HAR utilizing quantified action representation and CNNs, we introduce a novel method capable of concurrently addressing high efficiency, ease of fine-tuning, few-shot learning, and multi-scale learning requirements. A more detailed description is provided in Section 3.

3. Proposed Method

Figure 1 shows the general block of our method. The key frames of an action video are automatically selected by a specific strategy, so that not all frames need to be taken into consideration. Thus, the number of input parameters is reduced by an order of magnitude. Then, the key frames are quantified and stored in a flexible data structure which is named as Dense Joint Motion Matrix (DJMM) (Yao *et al.*, 2021). We used different origin points to calculate the joints' motion features and generate DJMMs. Thus, the more origin points are selected, the bigger scale of the DJMM will be, and vice versa. By this step, it is easy to find a balance between effectiveness and efficiency. Furthermore, a data augmentation module is designed to mitigate the issue of insufficient training samples, which effectively improves the generalization performance of the deep neural network (DNN).

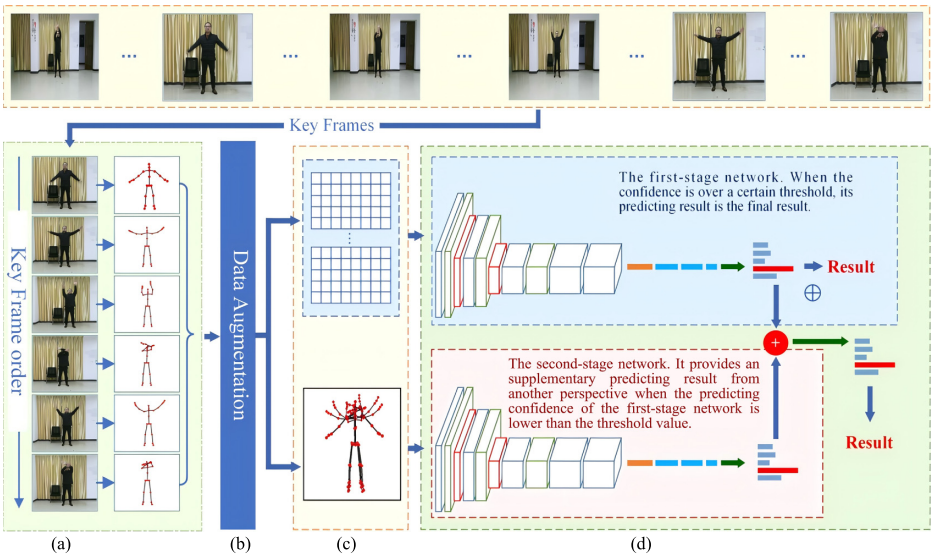

Fig. 1. The general block diagram of our proposed method. (a) Key frame selecting module. (b) Data augmentation strategy. (c) Action representation (including both local and global features). (d) A fusing deep neural network for action recognition.

Finally, a multi-scale DNN was put forward to recognize human actions with different temporal scales. Notably, the four modules of our method are independently decoupled and can be integrated into any skeleton-based HAR framework to boost recognition performance. Detailed description of the 4 parts are as follows:

3.1. *Depth Image and Human Skeleton*

A depth image provides 3D coordinates of every pixel, which makes it easier for researchers to perform spatial calculations across pixels. Furthermore, the human joints detected by depth sensors are closely related to the subjects being studied. Therefore, researchers can concentrate solely on the joints and rigid bones, disregarding other pixels in the depth image. Therefore, the skeleton-based HAR methods always have superiority in efficiency.

For instance, Kinect v2.0 offers data on a total of 25 joints, all of which can be tracked or estimated by its embedded software. Figure 2 shows the 25 joints and those used as origin points to calculate relative motion features.

3.2. *Key Frame Extraction*

Inspired by the principle of comic strips, where a small set of key frames suffices to convey a complete human action, we adopt a key frame-based strategy for efficient action representation. This observation can lead to two key insights. First, if an action's key frames can be extracted from a video, the whole action can be precisely represented. Second, the

Fig. 2. Twenty-five joints of the human skeleton used in our method. The joints labelled in red are used as origin points for feature extraction.

key frames typically correspond to moments of intense physical activity. Therefore, factors such as displacement and changes in body geometry can serve as measurable criteria for identifying key frames.

Our proposed method involves calculating joint displacement and joint angle variations frame by frame, thereby quantifying the action intensity within a specific frame. Equation (1) shows the way to calculate displacement of a joint between 2 frames.

$$D^i = \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2 + (z_i - z_{i-1})^2}, \tag{1}$$

where D^i means the displacement of a certain joint of the ith frame, and the value range of i is $(0, x]$, x stands for the total number of an action's frames. Thus, the motion intensity of a certain frame can be quantified as the total displacement of the whole 25 joints, as equation (2) shows:

$$\overbrace{D}^{i} = \sum_{n=0}^{24} D_n^i, \tag{2}$$

where $\overbrace{D}^{i}$ means the total displacements of all the 25 joints in the ith frame.

Motion analysis research indicates joint displacement and joint angle dynamics should both be considered to improve key frame extraction accuracy. Certain motions such as boxing, clapping and waving feature dramatic local joint rotation yet slight overall displacement. Our method computes the 19 annotated joint angles shown in Figure 3 for each frame.

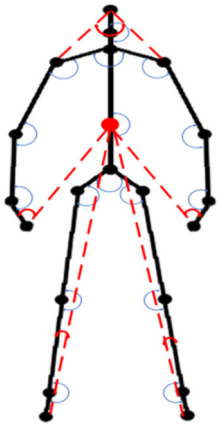

Fig. 3. The joint angles used in the proposed method.

The angle constructed by the three adjacent joints in a certain frame can be calculated using equation (3).

$$\theta^i = \cos(\overrightarrow{J_m J_n}, \overrightarrow{J_l J_n}) = \frac{\overrightarrow{J_m J_n} \cdot \overrightarrow{J_l J_n}}{|\overrightarrow{J_m J_n}| \times |\overrightarrow{J_l J_n}|}, \tag{3}$$

where θ^i stands for a certain joint angle in ith frame, J_m, J_n and J_l are the 3 adjacent joints, $\overrightarrow{J_m J_n}$ and $\overrightarrow{J_l J_n}$ are two vectors which are from joint J_m to J_n, and from joint J_l to J_l respectively. Thus, the joint angle variation that compares to its previous frame can be represented as $\overline{\theta}^i = |\theta^i - \theta^{i-1}|$. Then, the total variation of the 19 joint angles can be obtained by equation (4).

$$\overbrace{\overline{\theta}}^{i} = \sum_{z=0}^{18} \overline{\theta}_z^i, \tag{4}$$

where z is the order of the joint angle, i is the frame order and the value range of i is $(0, x]$, x stands for the total number of an action's frames.

After quantification, key frames are selected by ranking motion displacement or joint angle variations in descending order and retaining the top-N entries. A weighting strategy is required to jointly incorporate both criteria.

$$W^i = \frac{\overbrace{D}^{i}}{MAX(\overbrace{D}^{1}, \overbrace{D}^{2}, \ldots, \overbrace{D}^{x})} + \frac{\overbrace{\overline{\theta}}^{i}}{MAX(\overbrace{\overline{\theta}}^{1}, \overbrace{\overline{\theta}}^{2}, \ldots, \overbrace{\overline{\theta}}^{x})}, \tag{5}$$

where W^i is the weighted quantified value, and key frames can also be selected via the reverse order of $W^i s$.

The pseudo-code of the key frame extracting algorithm was shown as Algorithm 1.

Algorithm 1 The key frame extracting algorithm based on skeleton motion sequence

Input: N_{kf}, // The number of key frames to be extracted
$F_{All} = getSampleTotalFrames()$; //Obtain the total frame count of the currently processed video sample.
$Arr\overbrace{D} = \text{new } Array()$; //An array used to record frame-wise displacement values.
$Arr\overbrace{\theta} = \text{new } Array()$; //An array used to store frame-wise angular variations.
$ArrW = \text{new } Array()$; // An array for storing frame-wise quantified motion metrics.
for($i = 1; i < F_{All}$; i++){
$\overbrace{D}^{i} = 0$;
$\overbrace{\theta}^{i} = 0$;
for ($n = 0; n < 25$; n++){
$D_n^i = calculateJointDisplacement()$; //Calculate the displacement of a single joint in the i-th frame.
$\overbrace{D}^{i} += D^i$; // $\overbrace{D}^{i}$ denotes the cumulative displacement of all joints in the i-th frame.
}
for ($z = 0; z < 18$; z++){
$\overline{\theta}^i = calculateJointAngleVariation()$; //Calculate the variation of a single angle in the i-th frame.
$\overbrace{\theta}^{i} += \overline{\theta}^i$;// $\overbrace{\theta}^{i}$ denotes the parameter representing the aggregate variation of all angles in the i-th frame.
}
$Arr\overbrace{D}.push(\overbrace{D}^{i})$;
$Arr\overbrace{\theta}.push(\overbrace{\theta}^{i})$;
}
$Arr\overbrace{D} = getDescendingSort(Arr\overbrace{D})$;
$Arr\overbrace{\theta} = getDescendingSort(Arr\overbrace{\theta})$;
for($i = 1; i < F_{All}$; i++){
$W_D^i = Arr\overbrace{D}[i] / Arr\overbrace{D}[1]$;
$W^i\overbrace{\theta} = Arr\overbrace{\theta}[i] / Arr\overbrace{\theta}[1]$;
$W^i = W_D^i + W^i\overbrace{\theta}$
$ArrW.push(W^i)$;
}
$ArrW = getDescendingSort(ArrW)$;
$ArrKF = ArrW[0, N_{kf} - 1]$; // obtain the frame order for the top N_{kf} *entries in ArrKF*

3.3. *Action Representation*

In our approach, we select explainable features possessing significant physical and kinematic implications. These features encompass four quantified motion characteristics and one global motion image. In the following sections, the detailed information about motion features' calculating and the final action representation will be elaborated.

3.3.1. *Motion Feature 1: Joint Displacement*

This feature captures the spatial changes of a joint over a specific duration. Equation (1), which determines whether a frame is key, can also serve as a feature to represent actions. Notably, D represents the joint displacement in the 3D coordinate system. To obtain a larger-scale motion matrix, D is projected onto the x, y, and z axes in this study.

3.3.2. *Motion Feature 2: Joint Moving Direction*

The physical or kinematic meaning of this feature is recording the direction variation of a joint in a certain frame's duration. It contains both spatial and temporal information. The context information is essential to describe an action.

According to the vector's translation in-variance, the motion direction in $(j-i)$ frames' duration can be projected on plan xy, xz and zy, and their values can be calculated by the law of cosines.

$$\alpha^{i,j} = \begin{cases} \alpha_{xy}^{i,j} = \frac{x}{\sqrt{\Delta x^2+\Delta y^2}}, \\ \alpha_{xz}^{i,j} = \frac{xz}{\sqrt{\Delta x^2+\Delta z^2}}, \\ \alpha_{zy}^{i,j} = \frac{z}{\sqrt{\Delta z^2+\Delta y^2}}, \end{cases} \tag{6}$$

where $\{x, y, z\}$ is the position of a certain joint in the ith frame, while $\{\Delta x, \Delta y, \Delta z\}$ represents the joint's displacement along the x, y, and z-axes between the ith and jth frames.

3.3.3. *Motion Feature 3: Joint Motion Velocity*

This feature indicates the motion intensity of a certain joint in a period. From the physics perspective, it can be used to discriminate similar actions. Based on the data D, the velocity V can be readily computed using equation (7).

$$V = \frac{D}{0.03333 \times n}, \tag{7}$$

where 0.03333 is the duration time of one frame (the Kinect v2 record video at 30 fps, which indicates there are 30 frames in a second), and n is the number of interval frames.

3.3.4. *Motion Feature 4: Joint Angle in a Certain Frame*

In the geometrical view, the set of joint angles of a certain frame can be partly considered as a description of human pose, while an MHI-like global feature can be quantified by storing the angles in a float-point array by time sequence.

As it is shown in Figure 3, there are two types of joint angles. One is constituted by 3 adjacent joints and their related bones, and another one is constituted by the designed joints and red dotted lines (such as Head, HandTipLeft, HandTipRight, FootLeft and FootRight).

3.3.5. *K-DJMM: The Data Structure of The Proposed Action Representation*

In many works, the joints' coordinates, even the raw coordinates, are the only adopted data to describe an action. Our previous works (Yao *et al.*, 2020) have proved 3 rules for action quantification and representation.

(1) The relative coordinates are more suitable than raw coordinates to represent human actions. Raw coordinates are highly variable and may be influenced by the camera's installation position, leading to unintended consequences. Consequently, methods relying on raw coordinates are inherently dataset-oriented, often resulting in poor performance when the dataset varies or is applied in real-world scenarios.

(2) The necessary pre-processing of the coordinates is helpful to improve the accuracy. Despite CNNs' exceptional feature extraction capabilities, the calculation of essential motion features, including speed, orientation, and joint angles, among others, can furnish CNNs with additional valuable data, ultimately leading to enhanced detection accuracy.

(3) It is necessary to define several temporal scales for action samples. Because the durations of different actions can vary significantly (for example, a 'punch' action may be completed in 10∼15 frames, while a 'sit down' action may take 30 ∼60 frames), only one uniformed scale may result in motion information loss or motion variation.

In this paper, we followed the above three rules and addressed a small-scale issue to further improve the flexibility and accuracy of our method. In most research endeavours, the action representation typically consists of an image or a float matrix. Given that existing algorithms or depth cameras can provide at most 25 joints, the resulting motion image or motion matrix is relatively small in size. With continuous advances in computing hardware, lightweight representation models have hit inherent performance bottlenecks. Their low-dimensional design, once a key advantage, now restricts accuracy improvements.

To meet all the requirements of the above analysis, a multi-original-point idea was raised to generate the motion matrices with flexible scales. As shown in Figure 4(a), displacement, direction, velocity, and angle are calculated on a frame-by-frame basis.

According to literature (Le *et al.*, 2018), the relative origin point should be selected at the joint that exhibits minimal movement during action. Therefore, the 8 joints, such as Neck, SpineShoulder, ShoulderLeft, ShoulderRight, SpineBase, SpineCenter, HipLeft, HipRight, are adopted as the candidates of the relative origin points. As it is shown in Figure 4.

Using multiple relative original points has two advantages. First, one motion feature can be calculated multiple times. Thus, the scale of feature dimension can also be enlarged. Second, one motion feature based on different relative original points can partly reflect the geometrical information of the human body in a certain frame, which is also useful for the CNN to learn deep features.

In most of the "skeleton + CNN"-based methods, the dimensions of the motion representation are 'joints × features × frames'. For example, with four motion features, a 45-

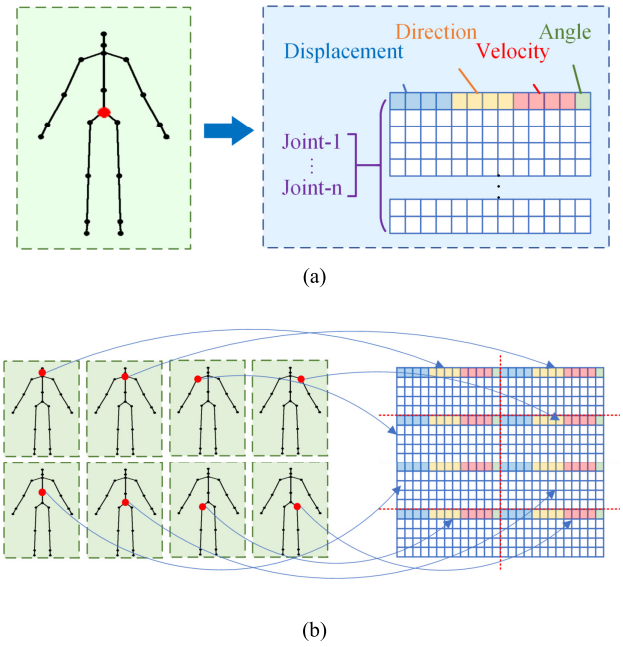

Fig. 4. The multi-scale motion matrix (MSMM) which consists of the features calculated by different origin points.

frame action yields a representation of dimensions $25 \times 4 \times 45$. If only 10 key frames are extracted, the dimensionality shrinks to $25 \times 4 \times 10$, which is too small to support deep convolutional operations. However, if we take all the relative origin points to calculate the 4 motion features and use their variations, the scale can be enlarged. For example, the joint displacement of a joint in 3D coordinate can be reflected in three planes. Thus, one displacement feature can be divided into 4 features. Furthermore, if all the 8 relative origin points are adopted, the final scale of the action representation matrix can be enlarged to $(25 \times (4 \times 4) \times (10 \times 8))$, which can be re-sized to $(200 \times 160 \times 1)$.

3.3.6. *The Cascade Network for HAR*

In our method, an action is represented by a local feature form—MSMM, and a global feature form—SMHI. As Figure 1(d) shows, the MSMM is used as the input of the first-stage neural network because of its small size and the innate merits of local features. If the prediction confidence of the first-stage network exceeds a certain threshold, the prediction result is deemed final. However, if the predicting confidence is lower than the threshold value, the second-stage neural network starts working. The SMHI is fed into the second-stage network, which produces a supplementary prediction from a global semantic perspective. The final confidence score is computed as the average of the two stages' outputs, and the class with the highest final confidence is selected as the recognition result.

The advantages of this network structure can be summarized in two aspects. First, the network uses both local feature and global feature. Local feature utilization in the first-

stage network mitigates inherent limitations of global feature. In the second-stage network, global feature-based predictions offer semantic enhancements, boosting accuracy when first-stage confidence is lower than a specific threshold. Second, by adjusting the threshold of prediction confidence in the first-stage network, a clear distinction can be made between effectiveness and efficiency.

4. Experimental Results and Evaluation

The method was implemented using the TensorFlow-gpu v2.3 and Keras. The experiments were performed on a desktop equipped with an NVIDIA GTX 4090 GPU, an Intel Core i7-13700K processor running at 3.70 GHz, and 64 GB of RAM operating at 3200 MHz.

We conducted three sets of experiments to evaluate the key frame extraction performance on two public datasets (Florence 3D Actions (Seidenari *et al.*, 2013) and UTKinect-Action3D dataset (Xia *et al.*, 2012), and one self-collected dataset. Recognition accuracy and computational efficiency are used as core metrics to assess the overall performance of the proposed method. The performance of our model was compared with typical methods. The results demonstrated that our proposed model achieved competitive accuracy and surpassed other methods in terms of efficiency.

4.1. *Dataset*

Florence-3D was collected using a Kinect camera. It includes 9 activities: wave, drink from a bottle, answer a phone, clap, tighten laces, sit down, stand up, read a watch, and bow. During acquisition, 10 subjects were asked to perform the above actions 2 or 3 times. This resulted in a total of 215 activity samples, with each action class containing approximately 20 to 30 samples. For each subject in this dataset, 15 joints were recorded.

The videos in UT-3D were captured using a single stationary Kinect. The dataset includes 10 types of actions: walking, sitting down, standing up, picking up, carrying, throwing, pushing, pulling, waving hands, and clapping hands. There are 10 subjects, and each subject performed each action twice. However, the 'carry' action group contains only 19 samples due to the failure to capture skeleton information for one sample. For each subject, 20 joints were recorded, 15 of which were used in our method.

HanYue-3D is a self-collected dataset. The dataset was gathered utilizing a Kinect v2.0 camera. It includes 15 simple action types: make a phone call, drink, wave hands, look at a watch, pat dust off clothes, fall, push a chair, jump in place, stand up, stand still, stand clap, walk, sit, sit still, and sit clap. Nine participants were instructed to execute each of the 15 activities three to four times. The 3D coordinates of all 25 joints, as detected by the Kinect v2.0 sensor, were meticulously documented. In total, 413 samples were collected, and each action type is represented by 35–37 samples.

4.2. *Key Frame Extraction Evaluation*

In our proposed method, two features, displacement and joint angle variation, are used to extract action key frames. It should be noted that other features can be easily added

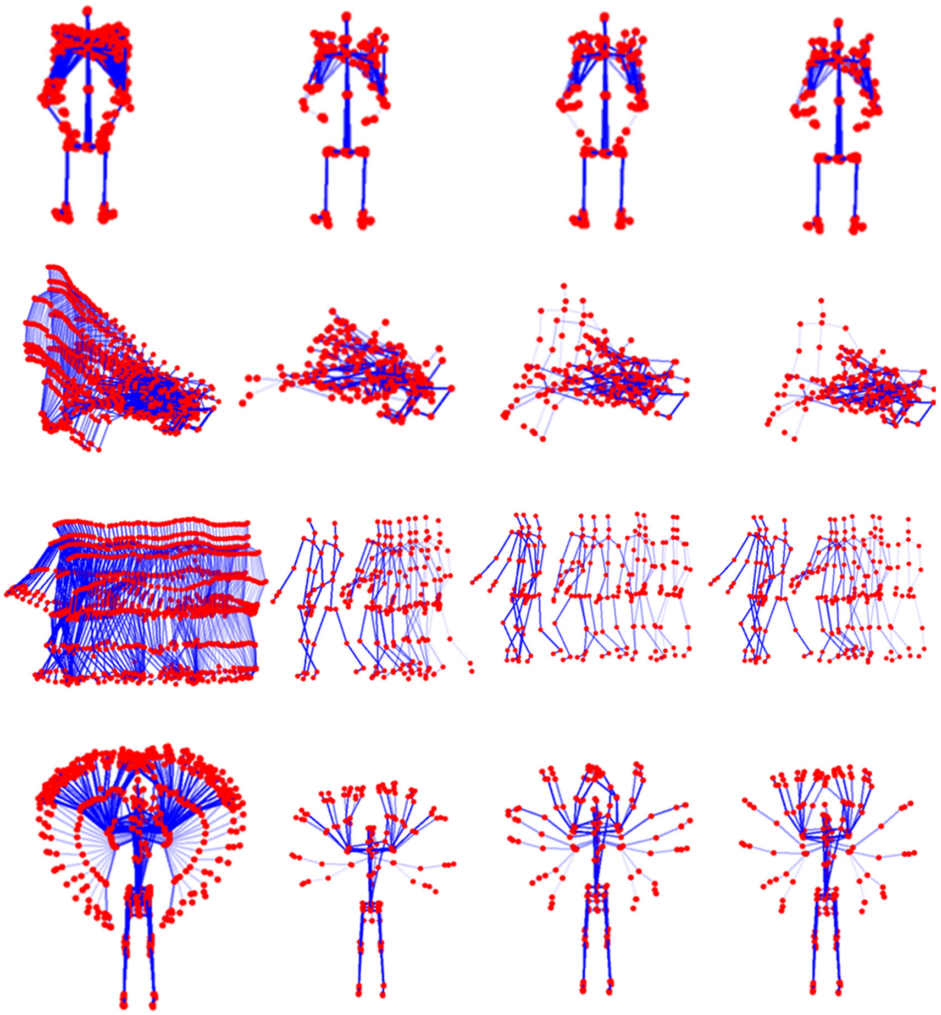

Fig. 5. The motion history images of 4 typical actions, where the SMHIs were generated by 10 key frames. From horizontal perspective, the 1st line is “sitting and clapping”, the 2nd line is “falling”, the 3rd is “walking”, the 4th line is waving. From vertical perspective, (a) MHI (b) SMHI generated via displacement feature. (c) SMHI generated via Angle variation feature. (d) SMHI generated via displacement and angle variation features.

or replace the features here we used to improve the precision of key frame extraction or balance the effectiveness and efficiency.

In this part of experiments, all action samples were depicted using a sparse motion history image (SMHI), created from 10 key frames. Figure 5 takes 4 typical actions as examples to show the universality of the proposed key frame extracting method. These 10 key frames and SMHIs were generated through displacement, joint angle variations, and a combination of both. From the experiments, 2 conclusions can be drawn as follows:

(1) The SMHI is capable of accurately capturing the motion tendency of an action. Hence, it has been demonstrated that the utilization of key frames for action recognition is technically viable. By leveraging key frames, which are significantly fewer in number than the total frames, the computational effort can be notably diminished.

(2) Although the three types of SMHIs share high similarity, the one generated through the combination of displacement and joint angle variation outperforms the others. For instance, in the "falling" scenario (line 2 of Figure 5), the key frames derived from the displacement feature solely concentrate on the final position, whereas in the "walking" example (line 3 of Figure 5), those from the joint angle variation feature exhibit insufficient continuity. Therefore, in subsequent experiments, we solely utilize the SMHI produced by displacement and joint angle variation (SMHI-D + A).

4.3. *Evaluation and Comparison*

In this section, 5 different types of experiments are conducted to evaluate the effectiveness and efficiency of our proposed method carefully.

4.3.1. *The Evaluation of Effectiveness of SMHI*

Just as its original version—MHI, SMHI is also a kind of global feature. Due to the fixed size of motion image, the input parameters that are based on MHI and SMHI will be the same. Therefore, there is no difference in the efficiency between MHI-based and SMHI-based methods. Hence, this section focuses solely on evaluating the effectiveness of SMHI-based methods and comparing it with MHI-based results. In the experiments, 80% of the samples are used for training, while 20% of them are used for testing. However, for SMHI, 80% of the samples are used for extracting key frames and generating SMHI, while in testing stage, all the samples are used for generating MHI and testing. Notably, to ensure experimental consistency and rigour, a stratified 80/20 train-test split is performed at the category level for each dataset, rather than a global random split applied to all samples collectively. The sample IDs for each partition are recorded and fixed. All experiments are conducted using the identical set of training samples for model training and the identical set of test samples for evaluation. Table 1 records the detailed results based on DenseNet121.

Table 1 demonstrates that all four motion image types can supply the neural network with essential motion data for HAR tasks. Furthermore, SMHI-D + A exhibited the best overall performance in experiments, confirming the efficacy of our key frame extraction algorithm. Experimental results indicate that 10 key frames of SMHI are optimal. Hence, in subsequent experiments, we will exclude 5 and 15 key frames of SMHI-D + A, focusing on SMHI-D + A in the following sections that stands for SMHI-D + A generated by 10 key frames.

4.3.2. *The Evaluation of Effectiveness of MSMM*

Compared with global features, local features offer numerous advantages, especially when occlusion occurs. In our previous work (Yao *et al.*, 2021), motion features were computed

Table 1
Comparison of HAR statistics among MHI and different types of SMHIS.

DataSet	Criterion	Accuracy (%)		
		Florence-3D	UT-3D	HanYue-3D
MHI	ACC	**82.50%**	**64.10%**	61.04%
	TOP-3	87.50%	**94.87%**	88.31%
	TOP-5	92.50%	**94.87%**	96.10%
SMHI-D + A 5 key frames	ACC	80.00%	58.97%	53.25%
	TOP-3	90.00%	84.62%	83.11%
	TOP-5	97.50%	**94.87%**	93.51%
SMHI-D + A 10 key frames	ACC	**82.50%**	61.54%	**62.34%**
	TOP-3	**92.50%**	**94.87%**	**89.61%**
	TOP-5	**100.0%**	**94.87%**	**97.40%**
SMHI-D + A 15 key frames	ACC	**82.50%**	58.97%	61.04%
	TOP-3	**92.50%**	87.17%	89.61%
	TOP-5	95.00%	**94.87%**	94.81%

frame by frame to construct a 3D floating-point matrix called the Dense Joint Motion Matrix (DJMM), which served as the input to a self-defined CNN. The small size and flexibility of the motion matrix endow our previous works with great superiority in time efficiency and competitive achievements in accuracy. However, the bottleneck shows when further improvement is conducting. While the compact size of the motion matrix enhances efficiency, it poses a significant challenge in employing deeper neural networks to boost accuracy. In the proposed method, this bottleneck is broken by utilizing MSMM. Based on Figure 4(b), the subsequent experiments calculate MSMMs using the method of Mean Squared, which involves 8 relative origin points. Additionally, to facilitate multi-scale learning, the softmax layer in conventional CNNs is substituted by the SPP layer.

In the MSMM group of Table 2, it can be indicated that the test accuracies are generally superior to those in the DJMI group, though the networks also suffered from overfitting problems. In the experiments, there are 2 networks, VGG19 and Densenet121, which should be paid special attention to. Within the DJMI group, VGG19 in the DJMIM subset achieved a training accuracy of only 26.41%, suggesting that DJMM is not suitable for VGG19 or other deep networks. However, the MSMM group achieved a peak value of 100% in the training stage, demonstrating that a larger-scale MSMM is a superior choice for deep neural networks.

Table 2 also reveals that test accuracies are generally lower than those achieved during the training phase. This indicates that the neural networks exhibit overfitting. This pattern is particularly pronounced in early sequential architectures such as VGG16 and VGG19, as deeper network structures are more prone to overfitting. Furthermore, it is worth noting that data augmentation alone is insufficient to fully mitigate the overfitting problem. For example, as Table 3 shows, in the VGG19 model with DJMM input (tensor dimensions: 23 joints $\times$ 7 features $\times$ 9 key frames), the tensor size is reduced to $(\ldots, 1, \ldots)$ at the third convolutional block—a dimension too small to sustain subsequent convolutional operations.

Table 2
Comparison of detection accuracy between DJMM and MSMM in the training and testing stages which are conducted on HanYue-3D dataset.

Typical CNN	DJMM ACC(%)		MSMM ACC(%)	
	Training	Testing	Training	Testing
LeNet-5	100%	66.23%	100%	**70.13%**
VGG16	86.35%	42.86%	100%	**45.45%**
VGG19	26.41%	31.17%	100%	**38.96%**
Densenet121	100%	71.43%	100%	**81.82%**
ResNet50	100%	66.23%	100%	**70.13%**
ResNet50V2	99.41%	67.53%	100%	**68.83%**
ResNet101	100%	59.74%	99.70%	**66.23%**
ResNet101V2	100%	67.53%	100%	**72.73%**
ResNet152	99.11%	61.04%	100%	**63.64%**
ResNet152V2	100%	59.74%	100%	**71.43%**

Table 3
Output shapes of each layer of VGG19.

Layer(type)	OutputShape
input_1(InputLayer)	[(None, 23, 7, 9)]
block1_conv1(Conv2D)	(None, 23, 7, 64)
block1_conv2(Conv2D)	(None, 23, 7, 64)
block1_pool(MaxPooling2D)	(None, 12, 4, 64)
block2_conv1(Conv2D)	(None, 12, 4, 128)
block2_conv2(Conv2D)	(None, 12, 4, 128)
block2_pool(MaxPooling2D)	(None, 6, 2, 128)
block3_conv1(Conv2D)	(None, 6, 2, 256)
block3_conv(Conv2D)	(None, 6, 2, 256)
block3_conv3(Conv2D)	(None, 6, 2, 256)
block3_pool(MaxPooling2D)	**(None, 3, 1, 256)**
block4_conv1(Conv2D)	(None, 3, 1, 512)
…	…

To further prove this conclusion, the training processes are visualized and displayed in Figure 6. It is shown that the MSMM can help neural networks reach training purposes more quickly. As depicted in Figure 6(a), the network attained a training convergence at approximately 65 epochs, whereas in Figure 6(b), the network reached a training convergence at around 50 epochs. Furthermore, an analysis of the training curves of VGG16 and VGG19 reveals that adopting large-scale motion matrices is an effective strategy to alleviate prevalent deep learning obstacles and prevent neural networks from suffering overfitting induced by small-scale tensors. Additionally, the arbitrarily configurable count of relative origin points enables flexible tuning of the MSMM dimension. This characteristic is compatible with both shallow and deep neural network architectures and facilitates a favourable trade-off between computational efficiency and model performance.

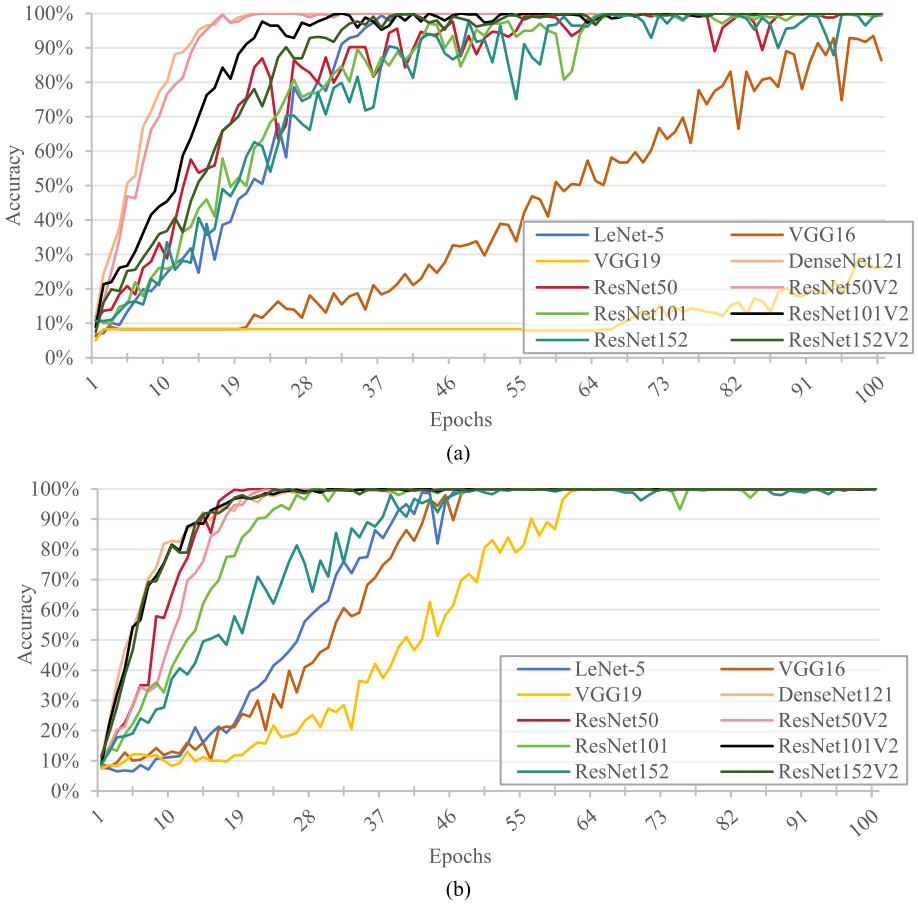

Fig. 6. The training curves of the typical neural networks evaluated on HanYue-3D by using the DJMMs and MSMMs. (a) The training curves based on the DJMMs. (b) The training curves based on the MSMMs.

4.3.3. *Evaluation and Comparison of The Proposed Cascade Neural Network Structure*
The above experiments explained and proved 3 key points. 1) Is it feasible to use key frames to represent human actions? 2) How many key frames are the best choice to represent an action? 3) Is the proposed data structure—MSMM a better representation for describing human actions? However, the overfitting problem caused by insufficient training samples remains unresolved. Furthermore, the proposed method's final HAR model has yet to undergo evaluation. In this section, we present the detailed solutions for addressing these two issues.

In our experiments, we developed a data augmentation program to simulate actions performed by subjects of different heights. The height scale factor was adjusted within [0.85, 1.15] with a step size of 0.01. As a result, the number of training samples was increased by 30 times. Thus, there are 6 450 samples of Florence-3D, 5 970 samples of UT-3D, and 115 500 samples of HanYue-3D respectively. Assume that an adult's height is

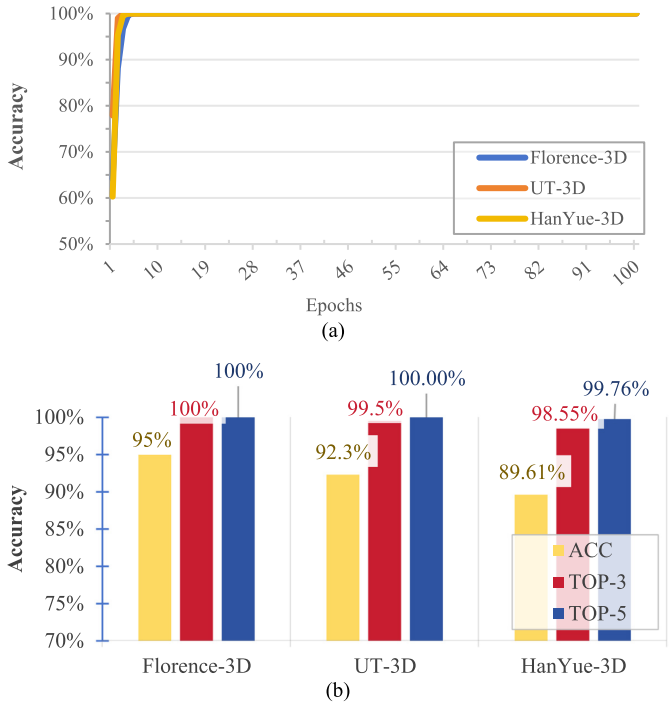

Fig. 7. The training and testing results of DenseNet121 that evaluated on Florence-3D, UT-3D and HanYue-3D respectively.

about 1.7 meters, then the generated person's height is among [1.45, 1.96], which matches the height of people in the real world. Furthermore, based on the methodology of our proposed data augmentation technique, additional strategies can readily be employed to create more action samples, including altering motion directions and adjusting motion speeds. The threshold of the first stage is defined as 0.8.

Figure 7 records the training and testing results of DenseNet121 that evaluated on Florence-3D, UT-3D and HanYue-3D respectively. From the results, three points can be concluded. *First*, the smoothness of the curves indicates that MSMM is a good data structure for action recognition. *Second*, the high accuracies achieved in both training and testing stages have proven that an action can be adequately represented by several key frames, with other frames considered as redundant input that can be discarded. *Third*, either multi-scale learning or data augmentation has positive influence on HAR accuracy improvement (Xin *et al.*, 2024). Moreover, since the nature of the proposed method is CNN, other prior technologies, including attention mechanisms, batch normalization and etc., can further contribute to accuracy improvement.

4.3.4. *Comparison With Other Classical or Latest Methods*

In the previous experiments, the proposed method has demonstrated its capacity and adaptability. However, the comprehensive evaluation of our network's architecture has not been conducted.

Table 4
Comparison of HAR accuracy between the proposed method and other typical methods.

Method	Acc(%)		
	Florence-3D	UT-3D	HanYue-3D
Dataset creators' method (Seidenari *et al.*, 2013)	82.20%	–	–
Dataset creators' method (Xia *et al.*, 2012)	–	90.92%	–
ST-GCN (Yan *et al.*, 2018)	85.00%	87.18%	87.01%
2s-AGCN (Shi *et al.*, 2019)	90.00%	92.30%	84.42%
ShiftGCN (Cheng *et al.*, 2020)	87.50.%	92.30%	79.22%
Hyperformer (Zhou *et al.*, 2023)	85.00%	89.74%	79.22%
DJMI + ZfNet + Data augmentation	92.50%	91.84%	–
DJMI + ZfNet + LSTM + Data augmentation (Yao *et al.*, 2020)	93.77%	94.23%	–
Multi-scale + ZfNet + Data augmentation (Yao *et al.*, 2021)	94.74%	**94.74%**	83.87%
MHI + DenseNet121	82.50%	64.10%	61.04%
SMHI + DenseNet121	82.50%	61.54%	62.34%
DJMM + DenseNet121	–	–	71.43%
MSMM + DenseNet121	–	–	81.82%
MSMM + DenseNet121 + Data augmentation	95.00%	92.30%	83.12%
Our method	**95.00%**	92.30%	**89.61%**

In this section, the two-stage cascade structure of our neural network and data augmentation strategy were all adopted to show the best performance and superiority by comparing it with other typical methods. To further verify the outstanding performance of the proposed framework, Table 4 compares its performance with classical CNN-based methods, key frame-based methods, and our previous works across multiple evaluation dimensions. For all comparative approaches listed in the table, we strictly follow the hyper-parameter configurations reported in their original publications. Their evaluation metrics are directly extracted from published literature if available; for methods lacking official reported results, we replicate the experiments under a unified dataset split to obtain corresponding performance values.

For additional performance improvement of the proposed framework, we conducted exhaustive analysis on all mispredicted samples. Among these samples, several "wave hand" actions were incorrectly identified as the "answer phone" category.

For an intuitive visualization of such misclassification cases, Figure 8 selects 3 representative key frames of the samples and the corresponding JTMs of the two action categories. Since the Florence-3D dataset lacks annotations for left and right hand joints, the resultant JTMs extracted from these two distinct movements exhibit extremely similar feature distributions. Although this limitation can be effectively alleviated by introducing separate left- and right-hand joint coordinates to reconstruct new discriminative joint temporal maps (JTMs), neither the DJMM nor the MSMM framework achieves reliable recognition performance on publicly available datasets. Specifically, samples misclassified by the DJMM model cannot be correctly distinguished by the MSMM either.

Furthermore, regarding the efficiency improvement based on the DJMM method, a detailed analysis and proof have been conducted in our previous work (Yao *et al.*, 2020), (Yao *et al.*, 2021). The concept of MSMM is based on DJMM. Therefore, we will not repeat the proof of its efficiency improvement here.

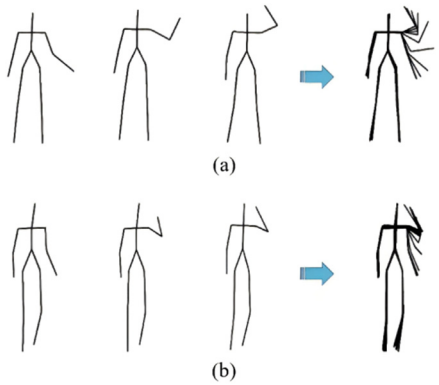

Fig. 8. The JTMs of two similar samples in the Florence-3D dataset. (a) "wave left hand", Sample ID: 1_30_2. (b) "answer phone", Sample ID: 2_151_8.

5. Conclusions

Inspired by the narrative logic of comic strips, where a small number of key frames suffice to convey a complete action, we propose an efficient HAR method based on key frame selection. This paper introduces an efficient method for HAR, leveraging the use of key frames. Initially, our solution identifies frames containing significant movements and geometric changes as key frames, selecting the top N to represent an action. It is the core step of the whole method. Two major issues are addressed. 1) The temporal scales of different actions are normalized to the same size. 2) The complexity of the representation of an action video has been significantly reduced. Then, a sparse motion history image (SMHI) and four per-joint motion features, displacement, velocity, motion direction, and joint angle, are computed and fed into a cascade neural network. The cascade network mechanism further enhances the efficiency of the network. Finally, to use a deeper network for higher accuracy purposes, multiple origin points are adopted in our method to gain the motion matrix with a larger size. The number of origin points can be flexibly defined for high accuracy or high-efficiency purposes.

Although the proposed method exhibited remarkable performance in the experiments, two aspects should be continually refined in future work. One is the key frame selecting algorithm, and the other is the design of motion features. These two components form the foundation of the HAR framework, and further accuracy improvements can be achieved through more precise key frame selection and finer-grained motion feature quantification.

Conflict of Interest Statement

No author associated with this paper has disclosed any potential or pertinent conflicts that may be perceived to have an impending conflict with this work.

Author Statement

Jianying Xiong: method, supervision, reviewing, **Zikang Fan:** paper writing, data visualization. **Ning Liu:** data collecting, reviewing, **Keyun Xiong:** editing, experimental implementation, **Leiyue Yao:** supervision, reviewing and validation.

Funding

This research was supported by the National Natural Science Foundation of China under Grant 62366023.

References

Abdelbaky, A., Aly, S. (2020). Human action recognition using short-time motion energy template images and PCANet features. *Neural Computing and Applications*, 32(16), 12561–12574.

Arif, S., Wang, J., Siddiqui, A.A., Hussain, R., Hussain, F. (2021). Bidirectional LSTM with saliency-aware 3D-CNN features for human action recognition. *Journal of Engineering Research*, 9(3), 115–133.

Cheng, K., Zhang, Y., He, X., Chen, W., Cheng, J., Lu, H. (2020). Skeleton-based action recognition with shift graph convolutional network. In: *Conference on Computer Vision and Pattern Recognition*, pp. 180–189.

Dalal, N., Triggs, B. (2005). Histograms of oriented gradients for human detection. In: *IEEE Conference on Computer Vision and Pattern Recognition*, San Diego, 2005, pp. 886–893.

Dong, W., Zhang, Z., Song, C., Tan, T. (2022). Identifying the key frames: an attention-aware sampling method for action recognition. *Pattern Recognition*, 130, 108797.

Du, Y., Fu, Y., Wang, L. (2015). Skeleton based action recognition with convolutional neural network. In: *IEEE Asian Conference on Pattern Recognition*, pp. 579–583.

Giveki, D. (2024). Human action recognition using an optical flow-gated recurrent neural network. *International Journal of Multimedia Information Retrieval*, 13(3), 29.

Herath, S., Harandi, M., Porikli, F. (2017). Going deeper into action recognition: a survey. *Image and Vision Computing*, 60, 4–21.

Hu, Z., Xiao, J., Li, L., Liu, C., Ji, G. (2024). Human-centric multimodal fusion network for robust action recognition. *Expert Systems with Applications*, 239, 122314.

Hussain, A., Khan, S.U., Khan, N., Bhatt, M.W., Farouk, A., Bhola, J., Baik, S.W. (2024). A hybrid transformer framework for efficient activity recognition using consumer electronics. *IEEE Transactions on Consumer Electronics*, 70(4), 6800–6807.

Jain, V., Gupta, G., Gupta, M., Sharma, D.K., Ghosh, U. (2023). Ambient intelligence-based multimodal human action recognition for autonomous systems. *ISA Transactions*, 132, 94–108.

Jeyanthi, A., Visumathi, J., Genitha, C.H. (2024). Enhanced two-stream Bayesian hyper parameter optimized 3D-CNN inception-v3 based drop-convLSTM2D deep learning model for human action recognition. *Information Technology and Control*, 53(1), 53–70.

Karim, M., Khalid, S., Aleryani, A., Khan, J., Ullah, I., Ali, Z. (2024). Human action recognition systems: a review of the trends and state-of-the-art. *IEEE Access*, 12, 36372–36390.

Kurban, O.C., Yildirim, T. (2024). A comparative analysis of multi-biometrics performance in human and action recognition using silhouette thermal-face and skeletal data. *Neural Networks: The Official Journal of the International Neural Network Society*, 170, 1–17.

Laptev, I. (2025). On space-time interest points. *Computer Vision*, 64, 107–123.

Le, T.M., Inoue, N., Shinoda, K. (2018). A fine-to-coarse convolutional neural network for 3D human action recognition. arXiv preprint. arXiv:1805.11790.

Li, C., Hou, Y., Wang, P., Li, W. (2018). Multiview-based 3-D action recognition using deep networks. *IEEE Transactions on Human-Machine Systems*, 49(1), 95–104.

Li, X., Kang, J., Yang, Y., Zhao, F. (2023). A lightweight attentional shift graph convolutional network for skeleton-based action recognition. *International Journal of Computers Communications & Control*, 18(3).

Majd, M., Safabakhsh, R. (2020). Correlational convolutional LSTM for human action recognition. *Neurocomputing*, 396, 224–229.

Phyo, C.N., Zin, T.T., Tin, P. (2019). Deep learning for recognizing human activities using motions of skeletal joints. *IEEE Transactions on Consumer Electronics*, 65(2), 243–252.

Seidenari, L., Varano, V., Berretti, S., Del Bimbo, A., Pala, P. (2013). Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 479–485.

Shi, L., Zhang, Y., Cheng, J., Lu, H. (2019). Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: *Conference on Computer Vision and Pattern Recognition*, pp. 12018–12027.

Sun, Z., Ke, Q., Rahmani, H., Bennamoun, M., Wang, G., Liu, J. (2022). Human action recognition from various data modalities: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3), 3200–3225.

Tan, K.S., Lim, K.M., Lee, C.P., Kwek, L.C. (2022). Bidirectional long short-term memory with temporal dense sampling for human action recognition. *Expert Systems with Applications*, 210, 118484.

Ullah, A., Muhammad, K., Del Ser, J., Baik, S.W., de Albuquerque, V.H.C. (2019). Activity recognition using temporal optical flow convolutional features and multilayer LSTM. *IEEE Transactions on Industrial Electronics*, 66, 9692–9702.

Wang, L., Yan, Y., Huang, D., Pan, Y., Cheang, C.F., Luo, K., Li, J. (2025). AHARNet: adaptive human activity recognition model for multimodal consumer electronics with different computation resources. *IEEE Transactions on Consumer Electronics*, 71(2), 5847–5855.

Wang, Z., Shen, K., Wang, D., Shen, H., Huang, K. (2024). Human body parsing in thermal InfraRed domain. *IEEE Transactions on Consumer Electronics*, 70(4), 6420–6429.

Wang, Y., Feng, T., Zheng, Y. (2022). Human action recognition using a depth sequence key-frames based on discriminative collaborative representation classifier for healthcare analytics. *Computer Science and Information Systems*, 19(3), 1445–1462.

Xia, L., Xin, W. (2024). Multi-stream network with key frame sampling for human action recognition. *Journal of Supercomputing*, 80, 11958–11988.

Xia, L., Chen, C.C., Aggarwal, J.K. (2012). View invariant human action recognition using histograms of 3D joints. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 20–27.

Xie, Q.L., Lu, W., Yang, W., Xiong, K., Zhang, L., Yao, L. (2025). Recognizing a complex human behaviour via a shallow neural network with zero video training sample. *International Journal of Computers Communications & Control*, 20(5), 1–11.

Xin, C., Kim, S., Cho, Y., Park, K.S. (2024). Enhancing human action recognition with 3D skeleton data: a comprehensive study of deep learning and data augmentation. *Electronics*, 13(4), 747.

Yan, S., Xiong, Y., Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. In: *AAAI Conference on Artificial Intelligence*, pp. 7444–7452.

Yang, W., Zhou, Y.T., Xiong, J.Y., Zhang, S., Zhang, L., Yao, L. (2025). Human action recognition using explainable features and sparse motion history images. *Technical Gazette Tehnički Vjesnik*, 32(5), 1614–1623.

Yang, C., Mei, F., Zang, T., Tu, J., Jiang, N., Liu, L. (2023). Human action recognition using key-frame attention-based LSTM networks. *Electronics*, 12(12), 2622.

Yang, Z., Li, Y., Yang, J., Luo, J. (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.

Yao, L.Y., Yang, W., Huang, W. (2020). A data augmentation method for human action recognition using dense joint motion images. *Applied Soft Computing*, 97, 106713–106723.

Yao, L.Y., Yang, W., Huang, W., Jiang, N., Zhou, B.B. (2021). Multi-scale feature learning and temporal probing strategy for one-stage temporal action localization. *International Journal of Intelligent Systems*, 12(1), 1–21.

Yu, J., Cheng, X., Chen, H., Xu, Y. (2024). Pose-guided robust action recognition for outdoor internet of things. *IEEE Transactions on Consumer Electronics*, 17(4), 7032–7043.

Zhang, D. (2019). ATSN: attention-based temporal segment network for action recognition. *Technical Gazette Tehnički Vjesnik*, 2(26), 1664–1669.

Zhang, S., Chen, E., Qi, C., Liang, C. (2016). Action recognition based on sub-action motion history image and static history image. *MATEC Web of Conferences*, 56, 02006.

Zhang, Y., You, S., Karaoglu, S., Gevers, T. (2025). 3D human pose estimation and action recognition using fisheye cameras: a survey and benchmark. *Pattern Recognition*, 162, 111334.

Zhou, S., Xu, H., Bai, Z., Du, Z., Zeng, J., Wang, Y., Wang, Y., Li, S., Wang, M., Li, Y., Li, J., Xu, J. (2023). A multidimensional feature fusion network based on MGSE and TAAC for video-based human action recognition. *Neural Networks*, 168, 496–507.

Zhou, Y., Cheng, Z.Q., Li, C., Fang, Y., Geng, Y., Xie, X., Keuper, M. (2023). Hypergraph transformer for skeleton-based action recognition. arXiv preprint. arXiv:2211.09590.

Zhang, Y., You, S., Karaoglu, S., Gevers, T. (2025). 3D human pose estimation and action recognition using fisheye cameras: a survey and benchmark. *Pattern Recognition*, 162, 111334.

Zhang, Y., Zhao, B., Wang, Y. (2026). HML-STN: high-middle-low spatio-temporal network for RGB-D based human action recognition. *Signal, Image and Video Processing*, 20(3), 179.

J. Xiong is an associate professor in the field of computer science. She received the ME degree from Zhejiang University of Technology, China, in 2006, and the PhD degree from Jiangxi University of Finance and Economics, China, in 2013. Her research interests include information systems, information management, and service computing.

Z. Fan received the bachelor's degree from East China University of Technology in 2025, and is currently pursuing the master's degree at Jiangxi University of Chinese Medicine. His research interests include computer vision, deep learning, multi-modal medical image analysis, and intelligent diagnosis of traditional Chinese medicine tongue.

N. Liu is the director, general manager and secretary of the board, Beijing Hanlin Hangyu Technology Development Inc. He has long been engaged in the R&D, engineering and industrialization of intelligent manufacturing equipment for traditional Chinese medicine solid preparations. His research focuses on intelligent production lines, digital factory construction and real-time monitoring systems for pharmaceutical manufacturing. He leads industrial research projects on the upgrading of domestic pharmaceutical equipment and promotes industry-university-research cooperation for intelligent transformation of Chinese medicine manufacturing.

K. Xiong was born in September 1980 in Nanchang, China. He received the BE degree and is currently a lecturer. His research interests mainly include big data architecture and data mining.

L. Yao received the BE, ME, and PhD degrees in computer science from Nanchang University, China. He is currently a professor at the School of Intelligent Medicine and Information Engineering, Jiangxi University of Chinese Medicine. He has published several papers in international journals and conferences. His current research interests include vision-based human action recognition, image and video processing, massive data processing, distributed systems, and software engineering.