Multimedia Tools and Applications
https://doi.org/10.1007/s11042-023-15334-9

# Skeleton-based human action recognition by fusing attention based three-stream convolutional neural network and SVM

Fang Ren[1] · Chao Tang[1] · Anyang Tong[1] · Wenjian Wang[2]

Received: 10 November 2022 / Revised: 27 March 2023 / Accepted: 6 April 2023

**Abstract**

This work proposes a method, aiming the 3D skeleton sequence, for the human action recognition by fusing the attention-based three-stream convolutional neural network and support vector machine. The traditional action recognition methods primarily employ RGB video as input. However, RGB video has issues with respect to large data volume, low semanticity, and ease of making the model interfered by irrelevant information such as the background. The efficient and advanced human action information contained in the 3D skeleton sequence facilitates human behavior recognition. First, the information of 3D coordinates, temporal-difference information, and spatial-difference information of joints are extracted from the raw skeleton data, and the above information is input into the respective convolutional neural networks for pre-training. Then, the pre-trained network model extracts the feature containing the spatial-temporal information. Finally, the mixed feature vectors are input into the support vector machine for training and classification. Under the X-View and X-Sub benchmarks, the accuracy on the open dataset NTU RGB+D is 92.6% and 86.7% respectively, demonstrating that the method proposed for incorporating multistream feature learning, feature fusing, and hybrid model can improve the recognition accuracy.

**Keywords** Skeleton-based human action recognition · Convolutional neural network · Attention mechanism · Support vector machine · Spatial-temporal feature

## 1 Introduction

Recently, owing to the development of deep learning and graphics processing unit, computer vision techniques are used to handle a variety of tasks in daily life [4–8] and medical field [11, 27, 28, 34, 57, 60, 64]. As an active topic in the field of computer vision, human action recognition has an extensive range of applications, such as the intelligent video surveillance systems [67], human-computer interaction [69], driverless [66], and sports body analysis

✉ Chao Tang
tangchao@hfuu.edu.cn

[1] School of Artificial Intelligence and Big Data, Hefei University, Hefei, China

[2] School of Computer and Information Technology, Shanxi University, Taiyuan, China

Published online: 23 May 2023

[21]. The accuracy of human action recognition models is generally based on the action representations. Therefore, several researchers have focused on the efficient extraction of the robust action features.

The previous research on human action recognition is based on RGB data [1, 12, 51]. RGB data-based research finds extensive applications owing to the RGB data being ubiquitous in daily life. However, the data has the problems of large data volume, low semantic meaning, and is easily disturbed by the viewpoint, illumination, and complex background. Thanks to the continuous development and application of sensor technology, such as Microsoft's Kinect sensor, several researchers have explored the human action recognition methods based on the depth data and 3D skeleton sequence data. Compared with the RGB video recorded by the traditional 2D cameras, the skeleton sequence data, as an advanced feature, contains rich information about the human body structure and can provide an efficient and robust representation for describing the human action with complex contents [62]. Therefore, the feature representation methods based on the skeleton sequence data have recently been utilized for human action recognition [18, 53].

For the previous skeleton-based human action recognition methods, the researchers primarily have conducted experiments by manually designing and extracting the descriptors of joints for human actions. Then machine learning algorithm is used to classify actions. However, manual feature has a limited ability to characterize human actions, hence the model recognition accuracy is insufficient. Representative works include [22, 36, 49, 52, 56]. In the field of deep learning, skeleton-based human action recognition methods mainly include recurrent neural network (RNN), convolutional neural network (CNN) and graph convolutional network (GCN). As a kind of non-Euclidean data, human skeleton topology needs to be rearranged to construct a data form for RNN and CNN training. In addition, RNN-based methods tend to deal with time series data, but lacks the ability of spatial modeling. GCN-based methods can handle non-Euclidean data, but it needs to consume a lot of computing resources. CNN-based methods not only have powerful high-level feature learning capability but also have the characteristics of few parameters and fast calculation speed.

There is redundant joint information in human actions. Therefore, focusing the model on the critical joint information is beneficial for human behavior recognition. The CNN is able to extract advanced features. Furthermore, the machine learning algorithm can solve several problems [2, 3, 63] with a rigorous theoretical foundation and powerful generalization capabilities. As a consequence, we combine the CNN with the machine learning algorithm as a hybrid model for skeleton-based human action recognition.

The main contributions of this paper are writon in bullet points as follows.

(1) A new traversal order of skeleton joints has been adopted to model the human body, which is conducive for retaining the correlation between the adjacent joints and for an efficient and adequate extraction of the co-occurring features of the joints.

(2) Based on the joint traversal order proposed in this paper, three streams of joint sequence information with diversity and complementarity have been constructed, which is conducive for mining spatial-temporal features.

(3) This paper proposes an attention-based three-stream CNN (A3SCNN) model which can hierarchically extract the joint spatio-temporal features with robustness, besides focusing on the critical joints in the action.

(4) This paper proposes a hybrid model for skeleton-based human action recognition by fusing A3SCNN and SVM. A3SCNN, as a feature extractor, is performed to obtain spatio-temporal features. We choose SVM to classify actions. Compared with the black-box property of neural networks, the SVM has the support of a solid and rigorous

mathematical theory which is beneficial for improving the generalization ability of the model.

## 2 Related work

The above methods [22, 36, 49, 52, 56] are traditional ones that rely on the manually designed behavioral features. The features have limited ability to represent actions, which in turn leads to poor model generalization. Alternatively, the deep learning has achieved far better results in RGB-based data than with the traditional methods, hence several researchers have also attempted to combine the deep learning with the skeleton sequence data. In this section, we review and discuss some respective deep learning methods for skeleton-based human action recognition.

**RNN-based methods** The RNN and its variants, viz., the long short-term memory network (LSTM), are special recurrent neural networks proposed to solve the gradient problem arising from the long-time dependence of the data and backpropagation [32]. To meet the input requirements of the RNN model, the skeleton sequence information needs to be converted into the form of a time series of joints. One line of works [14, 31] focus on the skeleton data form with joint co-occurrence features. Other works attempt various techniques, such as attention mechanism [31], regularization techniques [68] and novel model structure [25, 39, 65] to improve the model's ability to extract features. Furthermore, Pan et al. [37] have proposed a multi-level LSTM model based on skeleton sequences. The model first inputs the data of each joint and its parent joints into a fine-grained subnet, and then structures and fuses the features of the upper and lower body, separately. Considering the correlation between the joints, Shen et al. [42] have proposed a skeleton human behavior recognition method based on a complex network for the skeleton feature extraction combined with the LSTM. Although the RNN can adequately extract the time-domain information of the skeleton sequence data, it is difficult to extract the high-level features of the skeleton sequence data owing to the insufficient extraction of the space domain information of the skeleton sequence data. However, RNNs has poor spatial modeling capability. Besides, RNNs also has problem of difficulty in training, owing to the gradient disappearance and gradient explosion.

**GCN-based methods** The human skeleton is a natural topological graph, and the vertices and edges represent the joints and bones, respectively. Researchers tried to apply GCN to modeling human skeleton. Yan et al. [59] have constructed spatial temporal graph of a skeleton sequence and proposed the spatial-temporal GCN (ST-GCN), where a series of spatio-temporal graphs convolution blocks is stacked for spatial-temporal modeling. Upon the baseline, researchers pay attention to adjacency powering which is used for multi-scale modeling [16, 26, 33, 58]. In addition, attention mechanism embedded in the backbone network [46, 61] and multi-stream framework [44, 45] improve the model performance. ST-GCN and its variants has already achieved encouraging results and became the mainstream method for skeleton-based human action recognition [17]. However, above methods based on GCN spend a lot of computational resource in matrix operations and are subject to limitations in robustness and scalability.

**CNN-based methods** Similar to RNN-based method, the skeleton sequence information needs to be converted into a 2D pseudo-image form to meet the input requirements of the

2D-CNN. Thus, several works [20, 24, 29] try applying CNN to the skeleton sequence data. This stream framework takes a hierarchical approach towards learning the co-occurrence features of the joints, and hence the joint features at different levels are gradually aggregated. Different from above methods, other forms of 2D input have been used for CNN. Caetano et al. [9] have adopted a novel form of the joint representation, viz., tree structure reference joints images. The constructed skeleton sequence information has been fed into a self-built CNN for training and classification. Besides, Ding et al. [13], have used skeleton-based square grid approaches to describe the human actions, and used CNN for the action classification. In order to better extract the temporal and spatial features, some works [30, 38] tried to construct 3D input for 3D-CNN. Furthermore, Duan et al. [15] has constructed novel data structure, namely 3D heatmap volume, and used 3D-CNN model to extract features. This method achieves the state-of-the-art on skeleton-based human action recognition benchmarks and reflashes CNN. Compared with the RNN-based method, CNN-based method has excellent advanced information extraction capabilities and can learn the advanced features efficiently and easily. In addition, CNN-based method has relatively fewer parameters and faster calculation than GCN-based method. Concomitantly, the process of network model training is less prone to problems such as the overfitting, gradient explosion, or gradient disappearance.

**Other methods** Different from above deep neural networks, Transformer was originally designed to solve problems related to natural language processing (NLP) [40, 41]. Transformer has achieved great success in the NLP field, which has also driven its exploration in the field of human action recognition [10, 35]. Compared to traditional CNN models, Transformer is a simple and extensible framework that can obtain long-term dependent information and global information. Besides, attention module can further improve the performance of Transformer. Similar to GCN-based method, the Transformer consumes a lot of computing resources.

## 3 Method

The general framework of the proposed attention-based three-stream convolutional neural network, which is fused with the SVM for the skeleton-based human action recognition method, is displayed in Fig. 1. Our method can be divided into three main parts, viz., multi-stream data construction module, feature extraction module, and classification module.

First, we need to construct three streams data with spatial-temporal feature in 2D pseudo-image form to meet the input requirements. We perform the time domain differencing

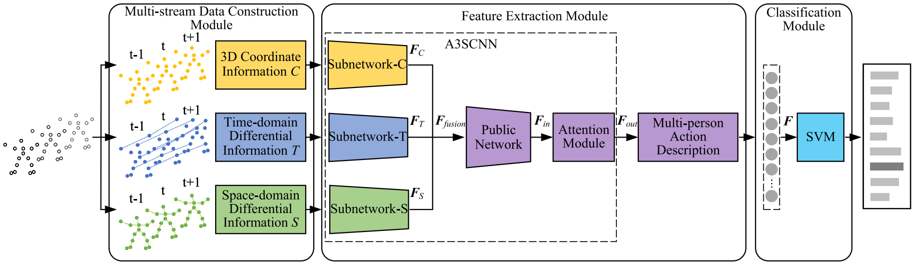

Multi-stream Data Construction Module
t-1
t
t+1
3D Coordinate Information $C$
Time-domain Differential Information $T$
Space-domain Differential Information $S$
Feature Extraction Module
A3SCNN
Subnetwork-C
Subnetwork-T
Subnetwork-S
$F_C$
$F_T$
$F_S$
$F_{fusion}$
Public Network
$F_{in}$
Attention Module
$F_{out}$
Multi-person Action Description
Classification Module
$F$
SVM

**Fig. 1** Framework of the proposed method

and space domain differencing on joint 3D coordinate information *C* to obtain the joint time-domain differential information *T*, and joint space-domain differential information *S*, respectively. Joint 3D coordinate information *C*, joint time-domain differential information *T* and joint space-domain differential information *S* are employed to characterize the skeleton-based human action.

Then, *C*, *T* and *S* are input to the sub-network of the corresponding feature extraction module, for spatial-temporal feature extraction. The three obtained feature vectors are then concatenated by channel dimension. The spliced feature vectors are input to the public network of the feature extraction module for extracting the global features. Subsequently, the attention module allows the network model to focus on the critical features. The feature vectors of each human body in the action are feature fused to achieve the representation of the multi-person interactive actions.

Finally, the fused feature vectors from feature extraction module have been employed to train and classify the SVM to complete the human behavior recognition based on the 3D skeleton sequence data.

### 3.1 Multi-stream data construction module

Owing to the error of the camera acquisition, plenty of useless information forms in the raw skeleton dataset, and hence, the pre-processing of the skeleton data is required. Furthermore, it is necessary to construct the multi-stream skeleton sequence data to meet the input requirements of the method in this paper, i.e., a four-dimensional array $R^{P \times M \times N \times V}$ where *P* is the number of human bodies in the action, *M* the number of frames in the action sequence, *N* the number of joints required to describe the human body, and *V* the coordinate dimension of the joints.

#### 3.1.1 Description of human skeleton

The human skeleton can be partitioned into the torso, left arm, right arm, left hand, right hand, left leg, right leg, left foot, and right foot. A natural correlation between the body parts and the skeleton data can be seen as a sequence of adjacent joints with certain dependency. At the same time, the co-occurrence of joints can describe human behavior to a certain extent. For example, the interaction of the joints of the left and right hands can describe the behavior of “clapping”, and the interaction of the joints of the trunk, and the left and right legs, can describe the behavior of “sitting”.

Therefore, describing the human skeleton by using the joints rationally, enables the efficient tapping of the correlation and co-occurrence of the joints, and extract the spatial-temporal characteristics of the joints. The original human body joints of the NTU RGB+D dataset are displayed in Fig. 2. We propose a new human skeleton joints description, which fully reflects the natural correlation of joints. For the 3D coordinates of the joints of the human body, i.e., *V=3* , we define

$$J_i = (x, y, z) \tag{1}$$

where $i = (1, 2, ..., 25)$ is the joint of the human body. The original traversal order of joints breaks the connection between the adjacent joints, and therefore, not conducive for extracting the spatial-temporal features of the joints. Thus, our method adjusts the traversal order of the joint points and re-describes each part of the human body.

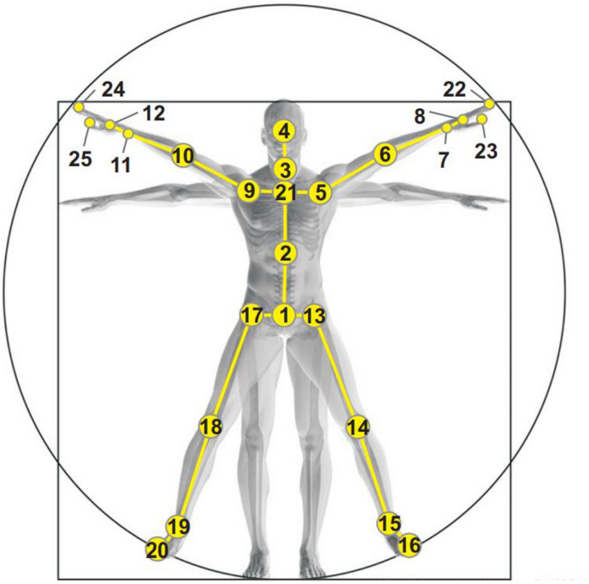


**Fig. 2** The original human body joints

$J_{arms}$, $J_{trunk}$, $J_{leftleg}$, and $J_{rightleg}$ are the joints forming the arms, body, and the left and right legs, respectively. They are formulated as

$$J_{arms} = \{J_{24}, J_{25}, J_{12}, J_{11}, J_{10}, J_{9}, J_{21}, J_{5}, J_{6}, J_{7}, J_{8}, J_{23}, J_{22}\} \tag{2}$$

$$J_{trunk} = \{J_{4}, J_{3}, J_{2}, J_{1}\} \tag{3}$$

$$J_{leftleg} = \{J_{17}, J_{18}, J_{19}, J_{20}\} \tag{4}$$

$$J_{rightleg} = \{J_{13}, J_{14}, J_{15}, J_{16}\} \tag{5}$$

Accounting for the connection between the limbs and the trunk, the human skeleton is modeled in this paper using the 3D coordinates of 29 joints, i.e., *N=29*. We formulate it as

$$\begin{aligned} Person = & \{J_{arms}, J_{body}, J_{leftleg}, J_{body}, J_{rightleg}\} \\ = & \{J_{24}, J_{25}, J_{12}, J_{11}, J_{10}, J_{9}, J_{21}, J_{5}, J_{6}, \\ & J_{7}, J_{8}, J_{23}, J_{22}, J_{4}, J_{3}, J_{2}, J_{1}, J_{17}, J_{18}, \\ & J_{19}, J_{20}, J_{4}, J_{3}, J_{2}, J_{1}, J_{13}, J_{14}, J_{15}, J_{16}\} \end{aligned} \tag{6}$$

Considering the different lengths of each action sequence in the skeleton dataset, we make the model maximally light, besides preserving the integrity of the action sequences. We have randomly selected 32 frames in each 3D skeleton sequence to construct the sample data, i.e., *M=32*.

### 3.1.2 Constructing three-stream skeleton data

According to the human skeleton description proposed in 3.1.1, 32 frames of skeleton sequence information have been randomly selected from the action sequences using the equal interval random selection method for constructing the multi-stream skeleton sequence information, which satisfies the input requirements of the A3SCNN.

(a) We define $J_i^t = (x, y, z)$ as the coordinate of a 3D joint of a human body in the action sequence. Further, the 3D coordinate information of a human skeleton sequence with $M$ action frames and $N$ joints can be described as

$$C = \left\{ J_i^t \mid i = (1, 2, ..., N), t = (1, 2, ..., M) \right\} \tag{7}$$

where $C$ is a $M \times N \times V$ array.

(b) To describe the spatial-temporal and co-occurrence properties of the human skeleton sequence, we construct the time-domain differential information $T$ of the joint sequence. $T$ represents the motion information of the same joint between adjacent frames, which can reflect the spatial-temporal characteristics of the joint motion. It can be formulated as

$$T = \{J_1^{t+1} - J_1^t, J_2^{t+1} - J_2^t, ..., J_N^{t+1} - J_N^t\} \tag{8}$$

where $T$ is a $M - 1 \times N \times V$ array.

(c) Furthermore, we construct the space-domain differential information $S$ of the joint sequence. $S$ represents the distance information between the adjacent joints in the same frame, which can reflect the spatial characteristics of the action. It can be formulated as

$$S = \{J_1^t - J_2^t, J_2^t - J_3^t, ..., J_{N-1}^t - J_N^t\} \tag{9}$$

where $S$ is a $M \times N - 1 \times V$ array.

We perform zero filling on the missing dimensions of $T$ for unifying the dimensions of the three skeleton sequence information, $S$. All of them obtain the three-dimensional array of $M \times N \times V$. Thus, this paper constructs the three skeleton sequence information with diversity and complementarity and uses them as the input of the A3SCNN.

## 3.2 Feature extraction module

This module first employs the constructed multi-stream data to pre-train the A3SCNN. Further, the feature extraction is performed using the pre-trained A3SCNN. Finally, we adopt a multi-person feature fusion strategy to achieve the description of the multi-person actions.

### 3.2.1 A3SCNN

We take the original skeleton sequence data through the multi-stream data construction module to obtain the joint 3D coordinate information $C$, joint time-domain differential information $T$, and joint space-domain differential information $S$. The above three skeleton sequence information is fed into the feature extraction network in parallel for the feature extraction, including a series of operations such as the convolution, pooling, and feature fusion. Then the multi-person feature fusion is performed to obtain the representation of the multi-person interactive actions. Finally, the prediction of action categories is accomplished through the fully connected layer. The network model designates Cross Entropy Loss as the loss function to calculate the difference between the true and predicted probability distributions. The loss value is calculated as

$$Loss = -\sum_{i=1}^{n} p(x_i) \ln(q(x_i)) \tag{10}$$

where $n$ denotes the sample category, $p(x_i)$ and $q(x_i)$ are the true and predicted probability distributions, respectively, corresponding to the variable $x_i$. The structure of the A3SCNN is displayed in Fig. 3.

**Three-stream convolutional neural network** The first stage can learn the point-level representation of 3D coordinates with a $1 \times 1$ (Conv1) and $3 \times 1$ (Conv2) convolution layers. We select the rectified linear unit (ReLU) for the activation function to accelerate the training and prevent the vanishing of the gradient which is defined as

$$f(x) = \max(0, x) \tag{11}$$

The standard convolution procedure is expressed as

$$y_{i,j}^{k} = f^{m_h * n_w}\left(\sum_{i \in m_h} \sum_{j \in n_w} x_{i,j}^{k} w_{i,j}^{k} + b^{k}\right) \tag{12}$$

where $x_{i,j}^{k}$ denotes the value of the point *(i, j)* of layer $k$, $w_{i,j}^{k}$ the weights on the convolution kernel, $b^k$ the bias value on the convolution kernel, $m_h$ and $n_w$ the height and width of the receptive field, respectively, $y_{i,j}^{k}$ the output value of the point *(i, j)* of layer $k$, and $f^{m_h * n_w}(\cdot)$ the sparsification using the ReLU activation function following a convolution operation with a convolution kernel size $m_h \times n_w$.

The second stage performs the feature learning on the joint sequence, which contains two convolution layers. First, the joint and channel dimensions are swapped by transposing the feature map. The general process is denoted as

$$X^{h \times c \times w} \leftarrow Transpose(X^{h \times w \times c}) \tag{13}$$

This stage contains two convolution layers with a kernel size $3 \times 3$ and channels (32, 64), respectively, and a Maxpooling layer with stride 2. The three skeleton sequence data are processed through their respective sub-networks for the feature extraction, and three feature vectors $F_C$, $F_T$, and $F_S$ are obtained. They are described as

$$F_C \leftarrow Subnetwork\text{ - C}(C) \tag{14}$$

$$F_T \leftarrow Subnetwork\text{ - T}(T) \tag{15}$$

$$F_S \leftarrow Subnetwork\text{ - }S(S) \tag{16}$$

The third stage inputs the multi-view feature vector into the public network for the global spatial-temporal feature learning. The above feature vectors of the three different views are concatenated by the channel dimension to obtain $F_{fusion}$ as

$$F_{fusion} \leftarrow Concat(F_C, F_T, F_S) \tag{17}$$

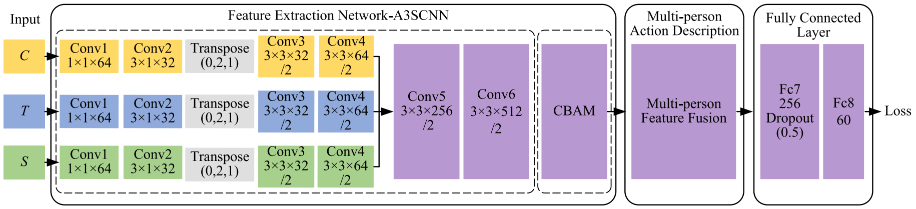


**Fig. 3** The structure of the A3SCNN

Then, all the subsequent convolution layers extract the global spatial-temporal features. The obtained vector $F_{in}$ is denoted as

$$F_{in} \leftarrow PublicNetwork(F_{fusion}) \tag{18}$$

$F_{in}$ is passed through convolutional block attention module (CBAM), allowing the model to focus on the critical joint features. The obtained vector $F_{out}$ is denoted as

$$F_{out} \leftarrow CBAM(F_{in}) \tag{19}$$

The feature vectors of the different human bodies are fused to represent the multi-person interactive actions. Through several comparison experiments, the strategy of max feature fusion achieves the best results. The obtained vector $F$ is described as

$$F \leftarrow MaxFusion(F_{out}) \tag{20}$$

Finally, the extracted feature vectors are input to the fully connected layer to calculate the loss value for backpropagation. Fc7 employs the Dropout strategy, which lets the neurons deactivate with a certain probability for alleviating the overfitting phenomenon of the network. By comparing multiple experimental groups, the best classification results have been obtained for the inactivation probability 0.5.

**Attention mechanism** The attention plays a vital role in the human perceptual system. The human visual system selectively focuses on certain salient parts through a series of local observations, instead of processing the whole scene at once. Skeleton sequence information contains temporal and spatial information about each joint. However, only the joint information that is useful for the action classification is of interest. The traditional convolutional neural networks cannot focus on the critical joint information, and different joints contribute differently to the action recognition. This work adds an attention mechanism after extracting the global spatial-temporal features of joints, for focusing the network on the critical joint information and ignoring the redundant joint information. Thus, the network can focus on the critical joint information and extract more robust spatial-temporal features of joints. We choose to place the integration position of the attention mechanism after extracting the global spatial-temporal features of the joints, by considering the different contributions of different joints in each frame of the skeleton sequence for action recognition. This is to reduce the influence of the locally optimal joints on the action recognition and thus achieve the global optimum.

We have used the CBAM, which can assign different weights to the extracted features, to indicate the importance of the features [54]. The CBAM innovatively proposes to combine the channel and spatial attention mechanisms with a serial structure. Owing to its lightweight and end-to-end characteristics, it can be seamlessly integrated into convolutional neural networks, besides effectively improving the performance of convolutional neural networks. The structure of the CBAM is displayed in Fig. 4.

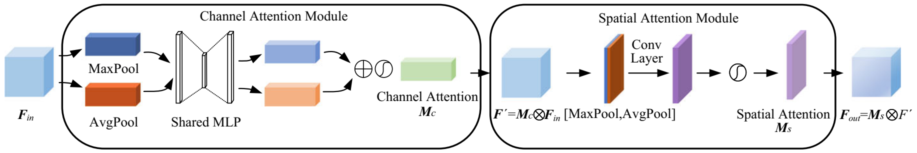


**Fig. 4** The structure of the CBAM. ⊕ denotes element-wise multication, ⊗ denotes element-wise summation

The CBAM is a serial architecture and contains two sub-modules, viz., the channel attention module and the spatial attention module. The workflow of the CBAM is shown below. The initial step is to input the extracted feature vectors $F_{in}$ to the channel attention module and assign different weights to the different channels. First, the global max pooling and global average pooling operations have been employed to aggregate the channel information of the feature vector $F_{in}$ to obtain $F^c_{\max}$ and $F^c_{avg}$. Then, $F^c_{\max}$ and $F^c_{avg}$ are sequentially passed through a shared multilayer perceptron. The number of neurons in the hidden layer of this multi-layer perceptron (MLP) is half the number of neurons in the input layer. Finally, the feature vectors, after passing through the MLP, are summed element by element and fed into the Sigmoid activation function to obtain the weight vector $M_c$ for each channel. The element-by-element multiplication of $M_c$ and $F_{in}$ yields the feature vector $F'$ after the channel attention module. The Sigmoid activation function is defined as

$$\begin{aligned} Sigmoid(x) &= \frac{1}{1+e^{-x}} && (21) \\ M_c(F_{in}) &= Sigmoid(MLP(AvgPool(F_{in})) + MLP(MaxPool(F_{in}))) && (22) \\ F' &= M_c \otimes F_{in} && (23) \end{aligned}$$

The second step is to input the feature vector into the spatial attention module. First, the global max pooling and global average pooling operations are employed to aggregate the spatial information of the feature vector $F'$ to obtain $F^s_{\max}$ and $F^s_{avg}$. Further, they are concatenated and convolved by a standard convolutional layer. The convolution kernel size of this convolution layer is $3 \times 3$ and the activation function is ReLU, i.e., $f^{3\times3}$. Finally, the feature vector obtained after the convolution operation is input to the Sigmoid activation function to obtain the weight vector on the spatial pixels $M_s$. The feature vector $F_{out}$ obtained after the CBAM is obtained by multiplying $M_s$ and $F'$ element by element. The calculation process for both the modules is given as

$$\begin{aligned} M_s(F') &= Sigmoid(f^{3\times3}([AvgPool(F'); MaxPool(F')])) && (24) \\ F_{out} &= M_s \otimes F' && (25) \end{aligned}$$

The network learns the global co-occurrence features of all joints. The critical joints play an important role for the human behavior classification tasks. The channel attention module enables the model to focus on those important joints and can build the dependencies between the non-adjacent joints, which helps to improve the model performance. For example, in the action of "put on a shoe", the joints that form the arm and leg are not adjacent. The channel attention module in the network focuses on the joints that form the arms and legs, and applies greater weight to the relevant feature information, while the spatial attention module can focus on the spatial-temporal information of the critical joints in the action and focus more on the critical features. Thus, the CBAM allows the network to focus on the critical spatial-temporal feature information of the critical joints in the action to improve the generalization ability of the model.

### 3.2.2 Multi-person action description

Human actions include the multi-person interactions, such as handshakes and hugs, and single-person actions. We have adopted a multi-person feature fusion strategy to achieve the description of the multi-person actions. The skeleton sequence information of the interactive action is a four-dimensional array of $P \times M \times N \times V$, where $P$ is the number of human bodies in the interaction action. The feature extraction network performs the feature extraction on the skeleton sequence information of different human bodies in the interactive actions. The feature vector for each person is described as

$$F_{outi} = [x_1^i, x_2^i, ..., x_{2048}^i] \tag{26}$$

where $i = (1, 2, ..., P)$. The feature vector $F$ with spatial-temporal characteristics, obtained after the fusion of multi-person features, is described as

$$\begin{aligned} F &= [x_1, x_2, ..., x_{2048}] \\ &= MaxFusion(F_{out1}, F_{out2}, ..., F_{outP}) \\ &= [\max(x_1^1, x_1^2, ..., x_1^P), \max(x_2^1, x_2^2, ..., x_2^P), \\ &\quad ..., \max(x_{2048}^1, x_{2048}^2, ..., x_{2048}^P)] \end{aligned} \tag{27}$$

The structure of the multi-person action description module is displayed in Fig. 5. The feature fusion strategy of this module belongs to late fusion, which has the following two advantages compared to the early fusion [23]. First, the scalability of the model is excellent, and it can be extended to classify the actions of the varying number of people. Secondly, the skeleton information of different people in the multi-person interaction behaviors is extracted by a feature extraction network with shared parameters, and hence, no additional parameters are added to the model, which makes the model efficient and lightweight.

### 3.3 Multi-classification SVM

The feature vectors $F$ that can describe the 3D skeleton sequences have been obtained, and they are trained and predicted using a multiclassification SVM. The SVM is a classical machine learning model for the binary classification tasks and is a supervised algorithm [48]. SVM maps samples from a linearly indivisible low-dimensional space to a linearly divisible high-dimensional space through a kernel function, and achieves the classification by finding the optimal segmentation hyperplane. The generally used kernel functions are the linear kernel, radial basis function (RBF) and polynomial kernel. The RBF, also known as Gaussian kernel, has the advantages of a strong local characterization, high flexibility, high applicability, and few computational parameters. We have finally selected RBF as the kernel function through comparison experiments. The formula for RBF is described as $K(x_i, x_j) = \exp(-\gamma \left\| x_i - x_j \right\|^2)$. $\gamma$ is an artificially set parameter. $x_i$ and $y_i$ are feature vectors. After

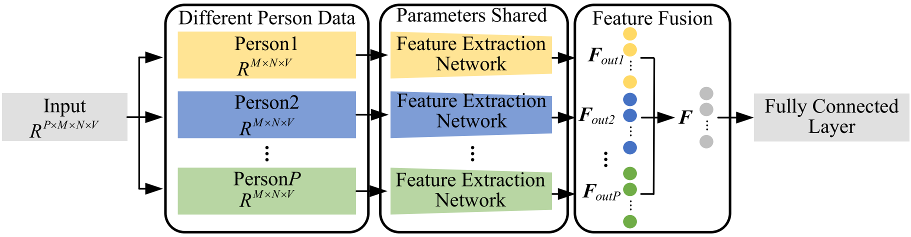


**Fig. 5** The structure of the multi-person action description module

introducing kernel function, the problem that SVM needs to solve is described as

$$\begin{aligned} &\min_{w^{mn},b^{mn},\varepsilon_i^{mn}} \tfrac{1}{2}\|w^{mn}\|^2 + Cost\sum_{i=1}^{60}\varepsilon_i^{mn} \\ &s.t. y_i[(w^{mn} \bullet \phi(F)) + b^{mn}] \geq 1 - \varepsilon_i^{mn} \\ &\varepsilon_i^{mn} \geq 0, m = 1, 2, ...60, n = 1, 2, ..., 60 \end{aligned} \tag{28}$$

where $w$ is the normal vector, which determines the direction of the hyperplane, an $b$ is the displacement term that determines the distance between the hyperplane and the origin. Further, $y_i \in \{m, n\}, i = 1, 2, ...56880$, kernel function $\phi$ maps the input low-dimensional sample $F$ to the high-dimensional space, $\varepsilon_i$ denotes the slack variables, and $Cost$ is the regularization parameter. The implementation of our method is displayed in Algorithm 1.

```
Input: $D = \{(x^{(i)}, y^{(i)}) \mid i = 1, ..., 56880\}$, SVM, RBF, $Cost$, $Epoch$
Output: Action categories $Prediction$
obtain $D_{train}$ and $D_{test}$ under the benchmark.
initialize A3SCNN parameters $w$.
for $i=1,2,...,Epoch$ do
    train A3SCNN with $D_{train}$.
    test A3SCNN with $D_{test}$.
    obtain $Accuracy_i$ using (29).
    obtain parameters $w_i$.
    if $i > 1$ and $Accuracy_i > Accuracy_{i-1}$
        update $w = w_i$
end
obtain $F$ using (20) with parameters $w$.
obtain $Prediction$ using (28).
```

**Algorithm 1** The implementation steps of this method

# 4 Experiments

## 4.1 Dataset

The NTU RGB+D dataset from Nanyang Technological University, Singapore, has been captured simultaneously by three Microsoft Kinect V2 cameras at different angles. This dataset contains 56880 skeleton sequences in 60 action categories performed by 40 volunteers. Out of a total of 60 action categories, the first 50 action categories are single-person actions, and the last ten action categories are two-person interactions, with each human body containing 25 3D joint coordinates. The original paper [39] of the dataset recommends two benchmarks. The first one is the (1) cross-subject (X-Sub), and under this benchmark, 40 volunteers are split into two groups, with the first group having 40320 skeleton sequences as the training samples and the second group having 16560 skeleton sequences as the test samples. The second one is the (2) cross-view (X-View), and this benchmark uses the skeleton sequences recorded by cameras 2 and 3 for the training samples, totaling 37,920, and the skeleton sequences recorded by camera 1 for the test samples, totaling 18,960. The specific action categories in the NTU RGB+D dataset are shown in Table 1.

**Table 1** Action category of the NTU RGB+D dataset

| Action number | Action category | Action number | Action category | Action number | Action category |
|---|---|---|---|---|---|
| 1 | drink water | 21 | take off a hat/cap | 41 | sneeze/cough |
| 2 | eat meal | 22 | cheer up | 42 | staggering |
| 3 | brush teeth | 23 | hand waving | 43 | falling down |
| 4 | brush hair | 24 | kicking something | 44 | headache |
| 5 | drop | 25 | reach into pocket | 45 | chest pain |
| 6 | pick up | 26 | hopping | 46 | back pain |
| 7 | throw | 27 | jump up | 47 | neck pain |
| 8 | sit down | 28 | phone call | 48 | nausea/vomiting |
| 9 | stand up | 29 | play with phone/tablet | 49 | fan self |
| 10 | 鼓掌 clapping | 30 | type on a keyboard | 50 | punch/slap |
| 11 | reading | 31 | point to something | 51 | kicking |
| 12 | writing | 32 | take a selfie | 52 | pushing |
| 13 | tear up paper | 33 | check time (from watch) | 53 | pat on back |
| 14 | put on jacket | 34 | rub two hands | 54 | point finger |
| 15 | take off jacket | 35 | nod head/bow | 55 | hugging |
| 16 | put on a shoe | 36 | shake head | 56 | giving object |
| 17 | take off a shoe | 37 | wipe face | 57 | touch pocket |
| 18 | put on glasses | 38 | salute | 58 | shaking hands |
| 19 | take off glasses | 39 | put palms together | 59 | walking towards |
| 20 | put on a hat/cap | 40 | cross hands in front | 60 | walking apart |

## 4.2 Training details

The experimental environment and configurations are listed in Table 2. The network has been trained using the Adam optimizer. We train the model 700 epochs in total and the batch size is set to 64. The learning rate is initialized to 0.0001 and falls exponentially by every epoch at a rate of 0.99. To alleviate the problem of overfitting, we have appended the dropout after Conv4, Conv5, Conv6, and Fc7 with a dropout ratio of 0.5. To improve the nonlinear expression of the model and speed up the model training [19], we have appended Batch Normalization after Conv1, Conv3, Conv5, and Fc7. For the multi-classification SVM, we choose RBF as the kernel function and set the regularization parameter *Cost* to 1 after several comparative experiments.

Since the accuracy has an intuitive expression of the generalization ability of the model, we choose the accuracy as the evaluation index of the model performance. The Model accuracy represents the proportion of the number of correctly predicted samples, to the number of all predicted samples. The accuracy formula is given as

$$Accuracy = \frac{TP}{Total} \times 100\% \tag{29}$$

where *TP* denotes the number of correctly classified samples and *Total* denotes the total number of samples classified.

**Table 2** Experimental environment and configurations

| Device | Version |
|---|---|
| Operating system | Windows10 |
| Python version | 3.8 |
| Torch version | 1.9.0 |
| CPU | Inter i9-11900K |
| GPU | NVIDIA GeForce RTX3080 |
| Storage | Samsung 16G×2 DDR4 |

### 4.3 Ablation study

We have verified the validity of the proposed method on the open dataset NTU RGB+D. We have conducted an extensive ablation on the effect of different components in the model on the recognition, including the human skeleton description method, attention module, feature fusion strategy, classification algorithm, and multi-stream data.

**Human skeleton description method** We test the effect of the two different modeling methods on the model accuracy. According to Table 3, the accuracy of the human skeleton description method proposed in this paper has been improved by 3.5% and 4.2% under the X-Sub and X-View benchmarks, respectively. Our method considers the correlation and co-occurrence between the joints, and therefore, has a higher recognition accuracy.

**Attention mechanism** The ablation experiments have been performed for the attention mechanism and the multiclassification SVM module to demonstrate the contribution of different modules for this method. According to Table 4, the attention mechanism and SVM can further improve the model accuracy. Particularly, owing to the X-Sub benchmark, the addition of the attention mechanism increases the model accuracy by 0.2%. Further, the model extracted features are fed into the SVM for training and testing classification, which further improves the accuracy by 0.4%. Similarly, in the X-View benchmark, the addition of the attention mechanism improves the model accuracy by 0.4%, and the SVM module further improves the accuracy by 2.0%. The ablation experiments demonstrate that the attention mechanism can effectively improve the model accuracy. Using the trained convolutional neural network as a feature extractor, the extracted features train the SVM. Finally, the classification task has been completed using a multi-classification SVM, and the above operations can effectively improve the accuracy of the model. The machine learning algorithm SVM has been supported by the rigorous mathematical theory and has an excellent classification effect. However, the deep learning model has exceptional feature extraction ability. The above idea of using the deep learning model to extract the features,

**Table 3** Effect of different human skeleton description methods on recognition accuracy

| Methods | X-Sub(%) | X-View(%) |
|---|---|---|
| Original | 83.2 | 88.4 |
| Ours | **86.7** | **92.6** |

Bold values indicate the best results

**Table 4** Comparison of different methods

| Methods | X-Sub(%) | X-View(%) |
|---|---|---|
| 3SCNN | 86.1 | 90.2 |
| 3SCNN+CBAM | 86.3 | 90.6 |
| 3SCNN+CBAM+SVM | **86.7** | **92.6** |

Bold values indicate the best results

and then using the machine learning model, to complete the classification task has certain applicability.

**Feature fusion strategy** The multi-person action description module is for extracting the skeleton sequence information of each human body in the multi-person interaction action by the convolutional neural network for feature extraction. Further, it performs the feature fusion on the extracted feature vector of each person. The feature fusion strategies include Concat, Mean, and Max. The experimental results with respect to the comparison of the effect of different feature fusion strategies on the model recognition accuracy, are displayed in Table 5. We find that the Max strategy is significantly better than the Concat and Mean strategies. The Max strategy can preserve the integrity of the high-level features to the maximum extent.

**Classification algorithm** There are several excellent algorithms for machine learning. We have selected different learning algorithms to complete the classification task, for comparing the generalization ability of such algorithms. The algorithms chosen for the comparison include eXtreme gradient boosting (XGBoost), artificial neural network (ANN), and random forest (RF). Before training and testing, the extracted feature vectors have been pre-processed by standardization. The formula is given as

$$x' = (x - \mu)/\sigma \tag{30}$$

where $\mu$ is the mean value and $\sigma$ is the standard deviation. According to Table 6, the SVM has the highest accuracy rate of 92.6%. Owing to the rigorous mathematical theory, the SVM algorithm shows the most robust generalization ability on the current dataset.

**A3SCNN** Ablation experiments have been performed using different data streams on the proposed method in this paper for comparing the contribution of multiple data streams with the method in this paper. The data stream includes joint 3D coordinate information $C$, joint time-domain differential information $T$, and joint space-domain differential information $S$. The experimental results are listed in Table 7. Experimental results demonstrate that the

**Table 5** Influence of feature fusion strategy on recognition accuracy

| Strategies | X-Sub(%) | X-View(%) |
|---|---|---|
| Concat | 86.1 | 91.5 |
| Mean | 86.3 | 92.1 |
| Max | **86.7** | **92.6** |

Bold values indicate the best results

**Table 6** Comparison of different classification algorithms

| Methods | X-Sub(%) | X-View(%) |
|---|---|---|
| XGBoost | 86.0 | 90.4 |
| ANN | 86.3 | 90.6 |
| RF | 86.5 | 91.1 |
| SVM | **86.7** | **92.6** |

Bold values indicate the best results

accuracy of two-stream data $C$+$S$ is 2.2% and 1.3% higher than that of the single-stream data $C$ for X-Sub and X-View benchmarks, respectively. The accuracy of the two-stream data $C$+$T$ compared to the single-stream data $C$ has been improved by 4.8% and 5.3% for the X-Sub and X-View benchmarks, respectively. The joint time-domain differential information $T$ and joint space-domain differential information $S$ contain rich human behavior features, which can effectively improve the accuracy of the model. The accuracy of $C$+$T$ is 2.6% and 4.2% higher than that of $C$+$S$ in both the X-Sub and X-View benchmarks, respectively, which shows that the joint time-domain differential information $T$ contains more human behavioral features than that of the joint space-domain differential information $S$. The three-stream data $C$+$S$+$T$ has the highest accuracy under two benchmarks, viz., the X-Sub and X-View, with 86.7% and 92.6%, respectively. The above experiments fully demonstrate the effectiveness of the method given in this paper.

The feature vectors $F_C$, $F_T$, $F_S$, and $F_{out}$ appearing in each level of the A3SCNN, for the testing samples, are visualized using t-distributed stochastic neighbor embedding (t-SNE) under the X-View benchmark. We have found that the network model can hierarchically extract the features, and the features will gradually aggregate to facilitate the classification. The dimensions of $F_C$, $F_T$, and $F_S$ are $64 \times 8 \times 8$, and the dimension of $F_{out}$ is $512 \times 2 \times 2$. The visualization results are shown in Fig. 6.

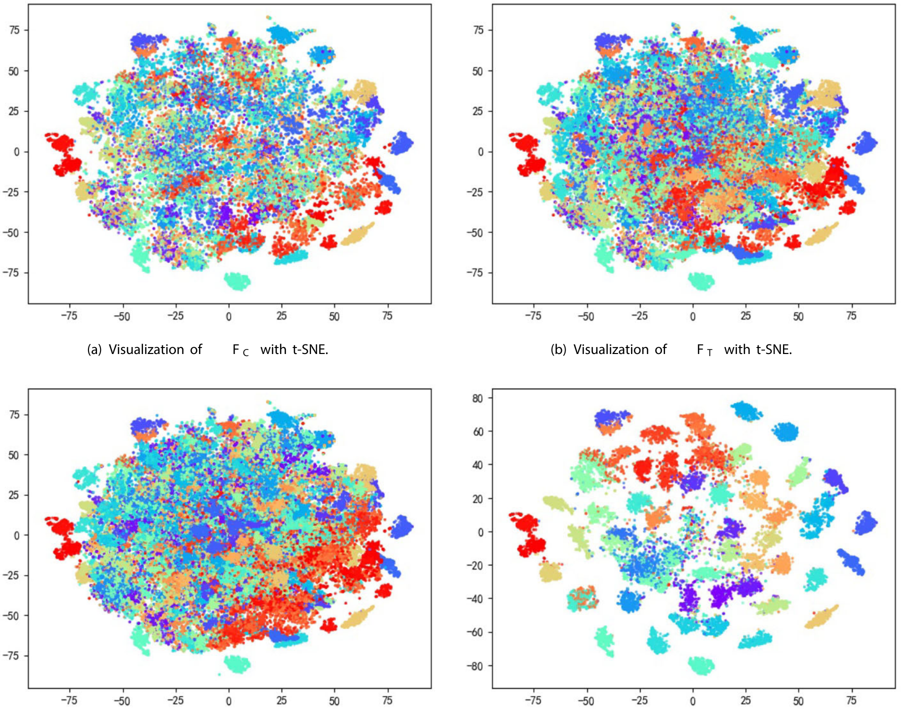
Using the three-stream data $C$, $T$, and $S$ proposed in this paper, a multi-person feature max fusion strategy has been adopted. The feature extraction network A3SCNN has been pre-trained under the two different benchmarks, and the variation curves of the training loss and testing accuracy that were obtained, are shown in Fig. 7.

The confusion matrix obtained by employing the method in this paper under the X-View benchmark is displayed in Fig. 8, and the darker color indicates higher accuracy. From the confusion matrix, the model can fully extract the co-occurrence features of the joints and accurately identify the actions such as “throw”, “stand up”, and “sit down”. However, the accuracy is not high for the nuanced and similar actions such as “reading” and “writing”, “play with phone” and “type on the keyboard”. The accuracy rate is modest when the

**Table 7** Recognition accuracy of the different data stream

| Data Stream | X-Sub(%) | X-View(%) |
|---|---|---|
| C | 81.4 | 86.8 |
| C+S | 83.6 | 88.1 |
| C+T | 86.2 | 92.3 |
| C+S+T | **86.7** | **92.6** |

Bold values indicate the best results



(a) Visualization of $F_C$ with t-SNE.

(b) Visualization of $F_T$ with t-SNE.

(c) Visualization of $F_S$ with t-SNE.

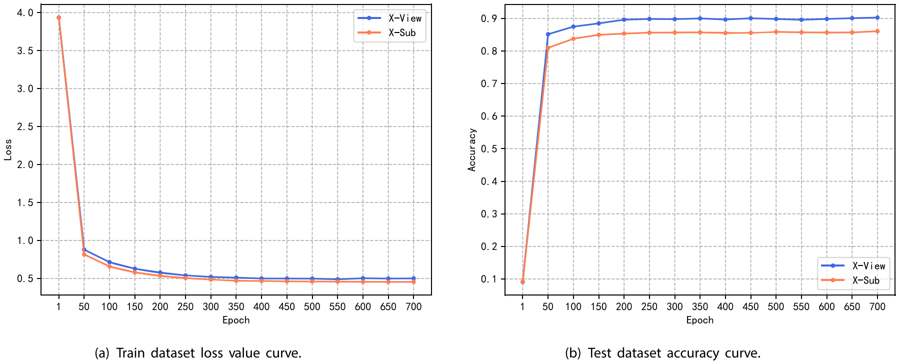
(d) Visualization of $F_{out}$ with t-SNE.

**Fig. 6** Visualization of different feature vectors with t-SNE

movements are similar. Furthermore, owing to the multi-person action description module, the model is highly accurate in recognizing the multi-person interactions such as “shaking hands” and “walking towards”.

## 4.4 Comparison with the classical methods

Here, we compare the method of this paper with the classical network model, as listed in Table 8. The CNC-LSTM model first transforms the human skeletal features using the network coding, and then uses the LSTM for the human behavior recognition. Compared with this model, our method proposes an end-to-end feature extraction network A3SCNN and employs the SVM for action classification, which is efficient and accurate. The HCN model is an end-to-end, hierarchical, co-occurrence, and feature learning framework for the skeleton-based human behavior recognition. Compared with that, the three-stream framework of this paper’s approach can describe richer joint features. Furthermore, the embedded attention module facilitates the network model to focus on the spatial-temporal features of the critical joints. Compared with the two-stream CNN model, the three-stream data of our method contains additional action features, which has a higher recognition accuracy. The ST-TSL model proposes a novel model with spatial reasoning and temporal stack learning (ST-TSL) for long-term skeleton-based action recognition. Compared with that, A3SCNN model has excellent advanced information extraction capabilities and can learn the advanced features efficiently and easily. The ST-GCN model views the human skeleton



(a) Train dataset loss value curve.

(b) Test dataset accuracy curve.

**Fig. 7** Train dataset loss value curve and Test dataset accuracy curve

as the graph structure data and uses the graph convolutional networks for the human behavior recognition. Compared with that, our method has the advantages from the need of a small number of parameters such as the high efficiency and ease of scaling up. The DPRL+GCNN model proposes a deep progressive reinforcement learning (DPRL) method to extract representative frames from action videos and uses graph-based convolutional network model for action recognition. Compared with that, the attention mechanism in our method can extract richer action features effectively and the SVM has excellent classification performance.

In addition, we compare our work with the GCN-based methods in FLOPs and Params, as listed in Table 9. As we mentioned in the Related Work, the GCN-based methods have huge amount of parameters and involve matrix calculation, which requires a lot of computing resources. PoseConv3D is a 3D-CNN model and still requires a lot of calculation. To sum

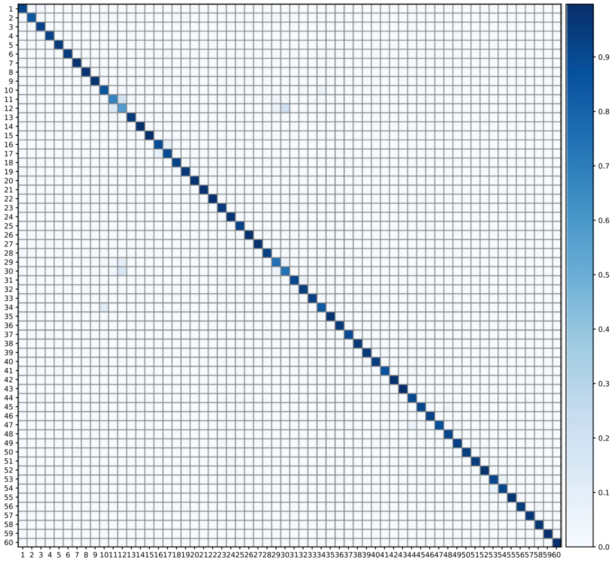


**Fig. 8** Confusion matrix of the proposed method on NTU RGB+D dataset

**Table 8** Comparison with classical networks

| Methods | X-Sub(%) | X-View(%) |
|---|---|---|
| CNC-LSTM [43] | 83.3 | 91.8 |
| HCN [24] | 86.5 | 91.1 |
| Two-stream CNN [23] | 83.2 | 89.3 |
| ST-TSL [47] | 84.8 | 92.4 |
| ST-GCN [59] | 81.5 | 88.3 |
| DPRL+GCNN [50] | 83.5 | 89.8 |
| Ours | **86.7** | **92.6** |

Bold values indicate the best results

up, our method has the characteristics of few parameters and fast calculation speed while maintaining excellent recognition rate, which has great advantages for the application of human skeleton behavior recognition tasks in reality.

## 5 Conclusion

In this paper, we proposed a skeleton-based human action recognition method by fusing the attention-based three-stream convolutional neural networks and SVM. First, to extract spatial-temporal features from skeleton data, we construct three data stream with diversity and robustness to train A3SCNN. The added attention module theoretically allows a better focus on the critical joint features, whereas the attenuation experiments of the attention module show the effectiveness of the module. A3SCNN has been employed as a feature extractor for extracting the spatial-temporal features of the joints. Further, the SVM has been utilized for the classification task. The method in this paper achieves appreciable results on the open dataset NTU RGB+D, with an accuracy of 86.7% under the X-Sub benchmark, and 92.6% under the X-View benchmark. Inspired by the work [15], we will explore the following aspects in future work. First, we will focus on constructing a data structure suitable for CNNs from human skeleton data. In addition, we will try to optimize CNNs structure to extract multi-scale spatial-temporal features for the nuanced and complex actions Classification. Multimodal feature fusion [55] is also a key research direction in the future.

**Table 9** Comparison with GCN-based methods

| Methods | FLOPs(G) | Params(M) |
|---|---|---|
| ST-GCN(CVPR2018) [59] | 16.3 | 3.1 |
| 2s-AGCN(CVPR2019) [45] | 37.2 | 6.9 |
| MS-G3D(CVPR2020) [33] | – | 6.4 |
| PoseConv3D(CVPR2022) [15] | 15.9 | 2.0 |
| Ours | **6.8** | **1.5** |

Bold values indicate the best results

**Acknowledgements** This work was supported in part by the National Natural Science Foundation (62076154, U21A20513), the Anhui Provincial Natural Science Foundation (2008085MF202), the University Natural Sciences Research Project of Anhui Province (KJ2020A0660), and the Open Project of Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Anhui University (MMC202003), Scientific Research Projects for Graduate Students in Anhui Universities (YJS20210564), Anhui Province Student Innovation Training Project (S202111059266, S202111059016).

**Data Availability** The datasets used or analyzed during the current study are available from the corresponding author on reasonable request.

## Declarations

**Conflict of Interests** The authors have no relevant financial or non-financial interests to disclose.

## References

1. Al-Faris M, Chiverton JP, Yang Y, Ndzi D (2020) Multi-view region-adaptive multi-temporal dmm and rgb action recognition. Pattern Anal Appl 23(4):1587–1602. https://doi.org/10.1007/s10044-020-00886-5
2. Bhatti UA, Huang M, Wang H, Zhang Y, Mehmood A, Di W (2018) Recommendation system for immunization coverage and monitoring. Human Vacc Immunotherap 14(1):165–171
3. Bhatti UA, Huang M, Wu D, Zhang Y, Mehmood A, Han H (2019) Recommendation system using feature extraction and pattern recognition in clinical care systems. Enterprise Inform Syst 13(3):329–351
4. Bhatti UA, Ming-Quan Z, Huo Q, Ali S, Hussain A, Yan Y, Yu Z, Yuan L, Nawaz SA (2021) Advanced color edge detection using clifford algebra in satellite images. IEEE Photonics J 13(2)
5. Bhatti UA, Nizamani MM, Huang M (2022) Climate change threatens Pakistan's snow leopards. Science 377(6606):585–586. https://doi.org/10.1126/science.add9065
6. Bhatti UA, Yan Y, Zhou M, Ali S, Hussain A, Huo Q, Yu Z, Yuan L (2021) Time series analysis and forecasting of air pollution particulate matter (pm2.5): an sarima and factor analysis approach. IEEE Access 9:41019–41031
7. Bhatti UA, Yuan L, Yu Z, Li J, Nawaz SA, Mehmood A, Zhang K (2021) New watermarking algorithm utilizing quaternion fourier transform with advanced scrambling and secure encryption. Multimed Tools Applic 80(9):13367–13387
8. Bhatti UA, Yu Z, Chanussot J, Zeeshan Z, Yuan L, Luo W, Nawaz SA, Bhatti MA, Ain QU, Mehmood A (2022) Local similarity-based spatial-spectral fusion hyperspectral image classification with deep cnn and gabor filtering. IEEE Trans Geosci Remote Sens 60:1–15
9. Caetano C, Brémond F, Schwartz WR (2019) Skeleton image representation for 3d action recognition based on tree structure and reference joints. In: 2019 32nd SIBGRAPI conference on graphics, patterns and images (SIBGRAPI). IEEE, pp 16–23
10. Chen J, Ho CM, Soc IC (2022) Mm-vit: multi-modal video transformer for compressed video action recognition. In: 22nd IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE Winter Conference on Applications of Computer Vision, pp 786–797
11. Dan Y, Jingbing L, Yangxiu F, Wenfeng C, Xiliang X, Bhatti UA, Baoru H (2021) A robust zero-watermarkinging algorithm based on phts-dct for medical images in the encrypted domain. Innovation in Medicine and Healthcare. Proceedings of 9th KES-InMed 2021. Smart Innovation, Systems and Technologies, pp 101–13
12. Dang LM, Min K, Wang H, Piran MJ, Lee CH, Moon H (2020) Sensor-based and vision-based human activity recognition: a comprehensive survey. Pattern Recogn, 108. https://doi.org/10.1016/j.patcog.2020.107561
13. Ding W, Ding C, Li G, Liu K (2021) Skeleton-based square grid for human action recognition with 3d convolutional neural network. IEEE Access 9:54078–54089
14. Du Y, Wang W, Wang L (2015) Hierarchical recurrent neural network for skeleton based action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 1110–1118
15. Duan H, Zhao Y, Chen K, Lin D, Dai B, Ieee Comp, S O C (2022) Revisiting skeleton-based action recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Conference on Computer Vision and Pattern Recognition, pp 2959–2968. https://doi.org/10.1109/cvpr52688.2022.00298

16. Feng D, Wu Z, Zhang J, Ren T (2021) Multi-scale spatial temporal graph neural network for skeleton-based action recognition. IEEE Access 9:58256–58265
17. Feng L, Zhao Y, Zhao W, Tang J (2022) A comparative review of graph convolutional networks for human skeleton-based action recognition. Artif Intell Rev, 4275–4305. https://doi.org/10.1007/s10462-021-10107-y
18. Han F, Reily B, Hoff W, Zhang H (2017) Space-time representation of people based on 3d skeletal data: a review. Comput Vis Image Underst 158:85–105
19. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning (ICML). PMLR, pp 448–456
20. Ke Q, Bennamoun M, An S, Sohel F, Boussaid F (2017) A new representation of skeleton sequences for 3d action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 3288–3297
21. Kennedy-Metz LR, Mascagni P, Torralba A, Dias RD, Perona P, Shah JA, Padoy N, Zenati MA (2020) Computer vision in the operating room: opportunities and caveats. IEEE Trans Med Robot Bion 3(1): 2–10
22. Koniusz P, Cherian A, Porikli F (2016) Tensor representations via kernel linearization for action recognition from 3d skeletons. In: European Conference on Computer Vision (ECCV). Springer, pp 37–53
23. Li C, Zhong Q, Xie D, Pu S (2017) Skeleton-based action recognition with convolutional neural networks. In: IEEE International Conference on Multimedia & Expo Workshops (ICMEW). IEEE, pp 597–600
24. Li C, Zhong Q, Xie D, Pu S (2018) Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. In: International Joint Conference on Artificial Intelligence (IJCAI)
25. Li S, Li W, Cook C, Zhu C, Gao Y (2018) Independently recurrent neural network (indrnn): building a longer and deeper rnn. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 5457–5466
26. Li MS, Chen SH, Chen X, Zhang Y, Wang YF, Tian Q, Soc IC (2019) Actional-structural graph convolutional networks for skeleton-based action recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Conference on computer vision and pattern recognition, pp 3590–3598
27. Li T, Li J, Liu J, Huang M, Chen Y-W, Bhatti UA (2022) Robust watermarking algorithm for medical images based on log-polar transform. Eurasip J Wireless Commun Network 2022:1. https://doi.org/10.1186/s13638-022-02106-6
28. Li Y, Li J, Shao C, Bhatti UA, Ma J (2022) Robust multi-watermarking algorithm for medical images using patchwork-dct. In: 8th International Conference on Artificial Intelligence and Security (ICAIS). Lecture notes in computer science, vol 13340, pp 386–399, https://doi.org/10.1007/978-3-031-06791-4_31
29. Liang D, Fan G, Lin G, Chen W, Zhu H (2019) Three-stream convolutional neural network with multi-task and ensemble learning for 3d action recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)
30. Lin Z, Zhang W, Deng X, Ma C, Wang H (2020) Image-based pose representation for action recognition and hand gesture recognition, 532–539
31. Liu J, Shahroudy A, Xu D, Wang G (2016) Spatio-temporal lstm with trust gates for 3d human action recognition. In: European Conference on Computer Vision (ECCV). Springer, pp 816–833
32. Liu A-A, Shao Z, Wong Y, Li J, Su Y-T, Kankanhalli M (2019) Lstm-based multi-label video event detection. Multimed Tools Applic 78(1):677–695. https://doi.org/10.1007/s11042-017-5532-x
33. Liu ZY, Zhang HW, Chen ZH, Wang ZY, Ouyang WL, Ieee (2020) Disentangling and unifying graph convolutions for skeleton-based action recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Conference on computer vision and pattern recognition, pp 140–149. https://doi.org/10.1109/cvpr42600.2020.00022
34. Liu W, Li J, Shao C, Ma J, Huang M, Bhatti UA (2022) Robust zero watermarking algorithm for medical images using local binary pattern and discrete cosine transform. Advances in artificial intelligence and security: 8th international conference on artificial intelligence and security, ICAIS 2022, Proceedings. Communications in computer and information science
35. Mazzia V, Angarano S, Salvetti F, Angelini F, Chiaberge M (2022) Action transformer: a self-attention model for short-time pose-based human action recognition. Pattern Recogn, 124

36. Nguyen V-T, Nguyen T-N, Le T-L, Pham D-T, Vu H (2021) Adaptive most joint selection and covariance descriptions for a robust skeleton-based human action recognition. Multimed Tools Applic 80(18):27757–27783
37. Pan H, Chen Y (2019) Multilevel lstm for action recognition based on skeleton sequence. In: 2019 IEEE 21st international conference on high performance computing and communications; IEEE 17th International conference on smart city; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS). IEEE, pp 2218–2223
38. Ruiz AH, Porzi L, Bulo SR, Moreno-Noguer F (2017) 3d cnns on distance matrices for human action recognition, 1087–1095. https://doi.org/10.1145/3123266.3123299
39. Shahroudy A, Liu J, Ng T-T, Wang G (2016) Ntu rgb+d: a large scale dataset for 3d human activity analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 1010–1019
40. Shao Z, Han J, Marnerides D, Debattista K (2022) Region-object relation-aware dense captioning via transformer. IEEE Transactions on Neural Networks and Learning Systems. https://doi.org/10.1109/tnnls.2022.3152990
41. Shao Z, Han J, Debattista K, Pang Y (2023) Textual context-aware dense captioning with diverse words, 1–15. https://doi.org/10.1109/TMM.2023.3241517
42. Shen X, Ding Y (2022) Human skeleton representation for 3d action recognition based on complex network coding and lstm. J Vis Commun Image Represent 82:103386. https://doi.org/10.1016/j.jvcir.2021.103386
43. Shen X, Ding Y (2022) Human skeleton representation for 3d action recognition based on complex network coding and lstm. J Vis Commun Image Represent 82:103386. https://doi.org/10.1016/j.jvcir.2021.103386
44. Shi L, Zhang Y, Cheng J, Lu H (2019) Skeleton-based action recognition with directed graph neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 7904–7913. https://doi.org/10.1109/CVPR.2019.00810
45. Shi L, Zhang Y, Cheng J, Lu H (2019) Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 12018–12027. https://doi.org/10.1109/CVPR.2019.01230
46. Shi L, Zhang YF, Cheng J, Lu HQ (2020) Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. IEEE Trans Image Process 29:9532–9545. https://doi.org/10.1109/tip.2020.3028207
47. Si C, Jing Y, Wang W, Wang L, Tan T (2018) Skeleton-based action recognition with spatial reasoning and temporal stack learning. In: Proceedings of the European conference on computer vision (ECCV), pp 103–118
48. Singla M, Ghosh D, Shukla KK (2020) A survey of robust optimization based machine learning with special reference to support vector machines. Int J Mach Learn Cybern 11(7):1359–1385
49. Su B, Wu H, Sheng M, Shen C (2019) Accurate hierarchical human actions recognition from kinect skeleton data. IEEE Access 7:52532–52541
50. Tang Y, Tian Y, Lu J, Li P, Zhou J (2018) Deep progressive reinforcement learning for skeleton-based action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 5323–5332. https://doi.org/10.1109/CVPR.2018.00558
51. Tong A, Tang C, Wang W (2022) Semi-supervised action recognition from temporal augmentation using curriculum learning. IEEE Trans Circuits Syst Video Technol, 1–1. https://doi.org/10.1109/TCSVT.2022.3210271
52. Vemulapalli R, Chellapa R (2016) Rolling rotations for recognizing human actions from 3d skeletal data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 4471–4479
53. Wang L, Huynh DQ, Koniusz P (2020) A comparative review of recent kinect-based action recognition algorithms. IEEE Trans Image Process 29:15–28. https://doi.org/10.1109/TIP.2019.2925285
54. Woo S, Park J, Lee J-Y, Kweon IS (2018) Cbam: convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 3–19
55. Wu H, Ma X, Li Y (2022) Spatiotemporal multimodal learning with 3d cnns for video action recognition. IEEE Trans Circ Syst Video Technol 32(3):1250–1261. https://doi.org/10.1109/TCSVT.2021.3077512
56. Xia L, Chen C-C, Aggarwal JK (2012) View invariant human action recognition using histograms of 3d joints. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, pp 20–27
57. Xiliang X, Jingbing L, Dan Y, Yangxiu F, Wenfeng C, Bhatti UA, Baoru H (2021) Robust zero watermarking algorithm for encrypted medical images based on dwt-gabor. Innovation in Medicine

and Healthcare. Proceedings of 9th KES-InMed 2021. Smart Innovation, Systems and Technologies. https://doi.org/10.1007/978-981-16-3013-2_7

58. Xu W, Wu M, Zhu J, Zhao M (2021) Multi-scale skeleton adaptive weighted gcn for skeleton-based human action recognition in iot. Appl Soft Comput, 104
59. 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
60. Yangxiu F, Jing L, Jingbing L, Dan Y, Wenfeng C, Xiliang X, Baoru H, Bhatti UA (2021) A novel robust watermarking algorithm for encrypted medical image based on Bandelet-DCT. https://doi.org/10.1007/978-981-16-3013-2_6
61. Yu L, Tian L, Du Q, Bhutto JA (2022) Multi-stream adaptive 3d attention graph convolution network for skeleton-based action recognition. Appl Intell
62. Yue R, Tian Z, Du S (2022) Action recognition based on rgb and skeleton data sets: a survey. Neurocomputing 512:287–306. https://doi.org/10.1016/j.neucom.2022.09.071
63. Zeeshan Z, ul Ain Q, Bhatti UA, Memon WH, Ali S, Nawaz SA, Nizamani MM, Mehmood A, Bhatti MA, Shoukat MU (2021) Feature-based multi-criteria recommendation system using a weighted approach with ranking correlation. Intell Data Anal 25(4):1013–1029
64. Zeng C, Liu J, Li J, Cheng J, Zhou J, Nawaz SA, Xiao X, Bhatti UA (2022) Multi-watermarking algorithm for medical image based on kaze-dct. J Ambient Intell Humaniz Comput, https://doi.org/10.1007/s12652-021-03539-5
65. Zhang S, Yang Y, Xiao J, Liu X, Yang Y, Xie D, Zhuang Y (2018) Fusing geometric features for skeleton-based action recognition using multilayer lstm networks. IEEE Trans Multimed 20(9):2330–2343. https://doi.org/10.1109/TMM.2018.2802648
66. Zhang J, Lou Y, Wang J, Wu K, Lu K, Jia X (2021) Evaluating adversarial attacks on driving safety in vision-based autonomous vehicles. IEEE Internet Things J 9(5):3443–3456
67. Zheng Z, An G, Wu D, Ruan Q (2019) Spatial-temporal pyramid based convolutional neural network for action recognition. Neurocomputing 358:446–455. https://doi.org/10.1016/j.neucom.2019.05.058
68. Zhu W, Lan C, Xing J, Zeng W, Li Y, Shen L, Xie X (2016) Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. In: Proceedings of the AAAI conference on artificial intelligence, vol 30
69. Zhuang Q, Gan S, Zhang L (2022) Human-computer interaction based health diagnostics using resnet34 for tongue image classification. Comput Methods Programs Biomed 226:107096. https://doi.org/10.1016/j.cmpb.2022.107096

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.