A STUDY ON RECOGNIZING AND ANALYZING AEROBICS MOVEMENTS BY VIDEO FEATURE EXTRACTION

Jian WANG[1]

Recognizing and analyzing different aerobics movements has practical implications for teaching and learning in aerobics research. This paper captured skeletal data from aerobics videos using Kinect V2 camera and extracted features such as coordinates, velocity, acceleration, and vectors of joint points from the data. These features were fused together and used as input for the movement recognition and analysis algorithm. The bidirectional long short-term memory (LSTM) was combined with a self-attention (SA) mechanism to obtain the SA-BiLSTM algorithm. Tests were performed on the NTU RGB+D dataset and a custom-created aerobics dataset. The results showed that compared to single features, the fused feature exhibited the best recognition and analysis performance. The SA-BiLSTM algorithm achieved an accuracy of 91.27% in cross-subject and 96.58% in cross-view, respectively, surpassing methods like ST-GCN. In terms of recognizing and analyzing aerobics movements, the SA-BiLSTM algorithm method achieved an average accuracy of 94.03%, while the LSTM and BiLSTM methods only reached 81.06% and 90.11%. These results demonstrate the reliability of the SA-BiLSTM algorithm in action recognition and analysis, making it applicable in practical situations.

Keywords: video, aerobics, movement recognition, skeleton data, self-attention mechanism

1. Introduction

Aerobics is a sport that combines dance, gymnastics, music, and other elements [1]. It improves body coordination and flexibility through various movements. Loved by the general public, it is also a competitive sport [2]. The identification and analysis of aerobics movements play an important practical role in helping learners better grasp the key points of movements. The influence of technological development has led to an increasing prevalence of storing information through videos in daily life. Consequently, video-based movement recognition and analysis have gradually emerged as a prominent area of research [3]. By utilizing features such as RGB and skeletal points, it is possible to achieve the recognition of different movements, which finds extensive applications in fields like security monitoring and medical care [4]. Liu et al. [5] designed a method for human motion state recognition based on micro-electro-mechanical system sensors

[1] Physical Education Department, Shandong Women's University, Jinan, Shandong 250300, China, e-mail: jk7wrz@126.com

and Zigbee network. By comparing it with existing research, they found that this method increased efficiency by 10% and accuracy by approximately 15%. Gao et al. [6] developepd a deep learning-based approach that combines deep video features and red, green and blue (RGB) features to classify different behaviors using deep learning algorithms. Experimental results showed that the method achieved an average classification accuracy of 85.79%. Hu et al. [7] designed a network called MV2Flow for recognizing movements in videos. Experiments conducted on UCF-101 and HMDB-51 demonstrated that this method had higher efficiency and comparable accuracy. Jaouedi et al. [8] presented an approach to identify human movements by combining sequential visual features with motion paths. It was evaluated on UCF Sports Action, UCF101, and KTH datasets and achieved competitive results. Currently, there have been some applications of video-based movement recognition methods in sports activities [9], but limited research specifically focused on aerobics has been conducted. Therefore, this paper conducted research on the recognition and analysis of aerobics movements. The Kinect V2 was utilized to capture skeleton data from videos, followed by feature extraction. Subsequently, an approach based on long short-term memory (LSTM) neural network was designed, and its effectiveness in movement recognition and analysis was proven through experiments on a dataset. This study presents a novel methodology for movement recognition and analysis, which can be applied to practical aerobics learning to improve teaching and learning efficiency.

2. Feature extraction from aerobics videos

In traditional movement recognition methods, the RGB sequence of videos is often used as a feature. However, this high-dimensional feature is easily influenced by background and lighting information, leading to low accuracy in movement recognition. With the development of technologies such as sensors and depth cameras [10], skeleton data has emerged as a new feature that has attracted increasing research attention in movement recognition.

There are several ways to obtain skeleton data: (1) manually converting RGB videos and calculating the coordinates of human joints for each frame; (2) using models such as Openpose [11], HigherHRNet [12], Alphapose [13], etc., to detect keypoint features of the human body using RGB videos as input; (3) directly capturing the three-dimensional coordinates of human joint movements using 3D motion sensors like Kinect.

The method of capturing skeleton data using Kinect has the advantages of being non-contact and more efficient. Therefore, this paper is based on Microsoft Kinect V2 to collect skeleton data from aerobics videos. The specific equipment includes a computer, a Kinect V2 camera, and a tripod. The camera parameters are shown in Table 1.

Table 1

Kinect V2 parameters

Parameter	Value
Depth image resolution	512*424 px
Color Image Resolution	1920*1090 px
Frame rate	30 Hz
Depth distance	0.5 m-4.5 m
Joint point	25

The 25 joint points collected by the Kinect V2 camera are presented in Fig. 1 and Table 2.

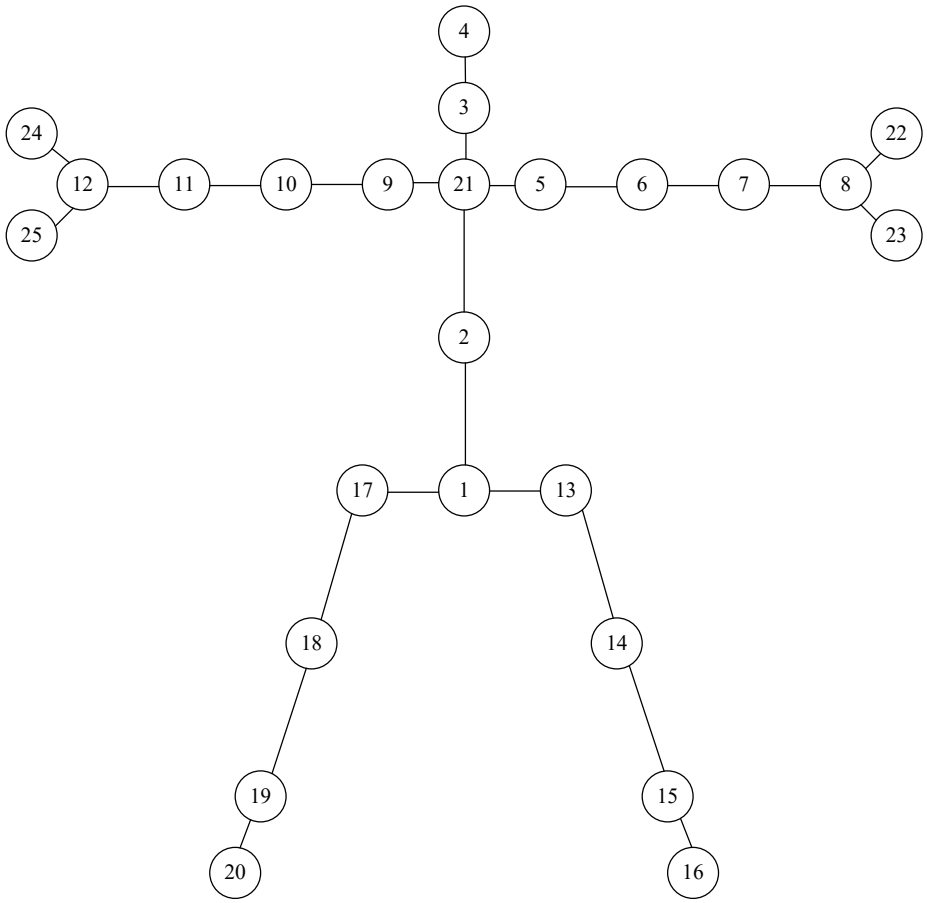

Fig. 1. Kinect V2 joint points

Table 2

Names corresponding to the joint points

1	Base of the spine
2	Middle of the spine
3	Neck
4	Head
5	Left shoulder
6	Left elbow
7	Left wrist
8	Left hand
9	Right shoulder
10	Right elbow
11	Right wrist
12	Right hand
13	Left hip
14	Left knee
15	Left ankle
16	Left foot
17	Right hip
18	Right knee
19	Right ankle
20	Right foot
21	Spine
22	Tip of the left hand
23	Left thumb
24	Tip of the right hand
25	Right thumb

The data captured by the Kinect V2 camera was transferred to the computer, and then the Kinect for Windows SDK 1.6 development toolkit was used to obtain three-dimensional coordinate information of human body skeletal points. Since there are currently no dedicated video data available for aerobics movement recognition and analysis research, this study collected data from 50 aerobics athletes while they performed five different movements using the Kinect V2 camera. The movements are shown in Fig. 2.

Fig. 2. Five aerobics movements

In terms of video feature extraction, in addition to the three-dimensional coordinates of human skeletal points, this paper also constructed the following features to more effectively represent skeleton data.

(1) Joint point speed: The velocity of a joint point can be expressed by the difference in coordinates between adjacent frames. It is assumed that the three-dimensional coordinate of the i-th skeletal point is j_i^{t+1} at the $t+1$-th moment and j_i^t at the t-th moment. Its movement velocity can be written as:

$$v_i = j_i^{t+1} - j_i^t. \tag{1}$$

(2) Joint point acceleration: It is used to represent the change speed of the joint point movement trajectory. It is assumed that the speed of the i-th skeletal point is v_i^{t+1} at the $t+1$-th moment and v_i^t at the t-th moment. Then, the acceleration can be written as:

$$a_i = v_i^{t+1} - v_i^t. \tag{2}$$

(3) Joint point vector: According to the principle of human force generation, when performing a movement, two adjacent joints rotate around an axis. Therefore, according to the distance with the central point of the human body, the joint point closer to that point is denoted as j_s, and the farther joint point is denoted as j_e. The joint point vector is defined as q_i:

$$q_i = j_s - j_e. \tag{3}$$

(4) Suppose that there are N joint points, a video has T frames, the extracted features include joint point coordinate J_i^t, joint point speed V_i^t, joint point acceleration A_i^t, and joint point vector Q_i^t, and the dimension of every kind of

feature is $N \times T \times 3$. Two fully connected (FC) layers realize feature mapping. Taking J_i^t as an example, the fused feature obtained after mapping is:

$$\tilde{J}_i^t = \sigma\left[W_2\left(\sigma(W_1 J_i^t + b_1)\right) + b_2\right], \tag{4}$$

where W_1 and W_2 are weight matrices of two FC layers, b_1 and b_2 are the bias of two FC layers, and σ is the ReLU activation function. By following this logic, the resulting video skeletal features can be expressed as:

$$F_i^t = \sigma\left[W\left(concat\left(\left(\tilde{J}_i^t + \tilde{V}_i^t\right), \left(\tilde{V}_i^t + \tilde{A}_i^t\right), \tilde{Q}_i^t\right)\right) + b\right], \tag{5}$$

where $\tilde{\mathrm{J}}_\mathrm{i}^\mathrm{t}$, $\tilde{\mathrm{V}}_\mathrm{i}^\mathrm{t}$, $\tilde{\mathrm{A}}_\mathrm{i}^\mathrm{t}$ and $\tilde{\mathrm{Q}}_\mathrm{i}^\mathrm{t}$ are the features obtained after mapping joint point coordinate J_i^t, speed V_i^t, acceleration V_i^t, and vector Q_i^t, $concat$ stands for the concat operation, W and b are the weight and bias of the network layer, and σ stands for the rectified linear unit (ReLU) activation fucntion. The FC layer dimensionally increases the concatenated features, enhancing the adaptability of the data and enabling a more effective representation of skeletal data features.

3. Movement recognition and analysis method design

The skeleton data contains certain temporal relationships; therefore, this paper proposes a recognition and analysis approach for aerobics movements based on the LSTM neural network. LSTM is an improved type of recurrent neural network (RNN), which has extensive applications in processing time series data such as water resource prediction [14] and trajectory prediction [15]. However, LSTM can only consider the contextual information before the current time step and is not comprehensive enough in learning features. In order to process the preceding and succeeding frame information in skeletal features, this paper adopts a bidirectional LSTM (BiLSTM) [16]. By using both forward and backward LSTMs, it obtains results from both directions and then utilizes a softmax layer for recognizing different movements, thus achieving better performance in movement recognition.

The self-attention (SA) mechanism can capture more useful information by allocating weights. Therefore, this paper incorporates the SA mechanism into BiLSTM and uses a multi-layer perceptron to calculate the similarity weights of the BiLSTM output layer. Finally, the flowchart of the recognition and parsing method for aerobics movements based on video feature extraction is displayed in Fig. 3.

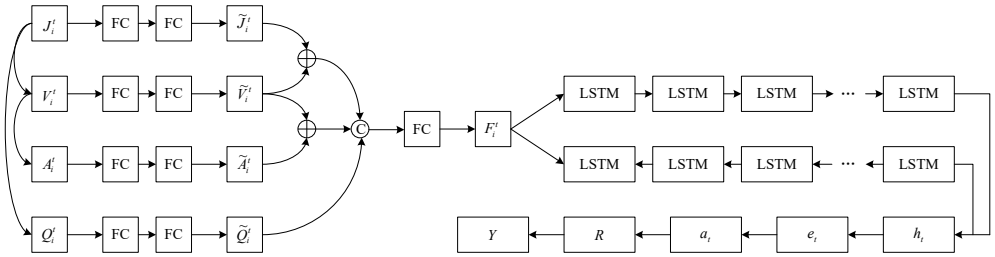

Fig. 3. The flowchart of the movement recognition and parsing method based on SA-BiLSTM

As shown in Fig. 3, two FC layers are used to map the four types of features into a high-dimensional space, obtaining fused features as the input of BiLSTM. After learning from forward and backward LSTMs, the output of BiLSTM is fed into the SA module, continuously updating connection weights until meeting accuracy requirements. Finally, the parsed results of aerobics movement recognition are output.

4. Experiment and analysis

4.1 Experimental settings

The algorithm was designed and implemented using the PyTorch framework, with Python 3.10 as the programming language. The learning rate of the algorithm was set to 0.01, epoch = 300, dropout = 0.2, and batch size = 64. The optimizer adopted was stochastic gradient descent (SGD), and the cross-entropy loss function was also used. The experimental datasets used are as follows.

(1) NTU RGD+D [18]: It is a dataset based on Kinect V2, containing the collection of 60 classes of movements from 40 individuals, with a total of 56,880 data samples. The dataset is divided into two validation methods: cross-subject (CS) and cross-view (CV). In terms of CS, the data of 20 movement performers were used as the training set, while the data of the remaining 20 movement performers were used as the test set. In terms of CV, the data collected by two sensors were used as the training set, while the data collected by the remaining one sensor was used as the test set. The sample distribution for both methods is shown in Table 3.

Table 3

The distribution of the NTU RGB+D dataset

	cross-subject	cross-view
Training set	40,320	37,920
Testing set	16,560	18,960

(2) Self-built set: it is the dataset of aerobics movements collected by Kinect V2 earlier, including five movements from fifty people. There are totally 2,000 samples. The dataset was split into an 80% training set and a 20% testing set.

For the evaluation of the effect of movement recognition and analysis effect, the accuracy was used. Its equation is:

$$Accuracy = \frac{\sum_{k=1}^{C} T_k}{\sum_{k=1}^{C} (T_k + F_k)}, \tag{19}$$

where C stands for the quantity of sample classes, T_k stands for the quantity of samples correctly identified in class k, and k stands for the quantity of samples incorrectly identified in class k.

4.2 Result analysis

Four types of features were selected for fusion. Therefore, the impact of different features on the recognition and analysis effectiveness of the SA-BiLSTM algorithm was compared using the NTU RGB+D dataset. Table 4 shows the obtained results.

Table 4

The accuracy of the SA-BiLSTM algorithm under different features

	CS/%	CV/%
J_i^t	85.64	91.37
$J_i^t + V_i^t$	86.77	92.46
$J_i^t + V_i^t + A_i^t$	87.83	92.57
$J_i^t + V_i^t + Q_i^t$	88.02	92.91
F_i^t	91.27	96.58

It can be seen from Table 3 that when using joint point coordinate J_i^t only, the accuracy of the algorithm was not high for both CS and CV, 85.64% and 91.37%, respectively. When using $(J_i^t + V_i^t)$, the accuracy of the algorithm was 86.77% for CS, which was 1.13% higher than that using J_i^t only, and 92.46% for CV, which was 1.09% higher than that using J_i^t only. Then, the contribution of A_i^t and Q_i^t to the algorithm accuracy improvement was compared. When using $(J_i^t + V_i^t + A_i^t)$, compared to using $(J_i^t + V_i^t)$, the accuracy was 1.06% higher for CS and 0.11% higher for CV. When using $(J_i^t + V_i^t + Q_i^t)$, compared to using $(J_i^t + V_i^t)$, the accuracy was 1.25% higher for CS and 0.45% higher for CV, suggesting that Q_i^t contributed greater to the accuracy improvement. When using F_i^t as the feature, the algorithm achieved an accuracy of 91.27% and 96.58% for CS and CV respectively,

which were higher than those using the other features. The results showed that it was reliable to use the fusion of these four features as the input of the SA-BiLSTM algorithm as it could effectively enhance the recognition and analysis effect of the algorithm for movements.

Then, the SA-BiLSTM algorithm was compared to the other movement recognition and analysis methods using the NTU RGB+D dataset, including the end-to-end spatio-temporal attention (STA)-LSTM [19], spatiotemporal graph convolutional network (ST-GCN) [20], two-stream adaptive graph convolutional network (2s-AGCN) [21], and directed graph neural network methods [22] (Table 5).

Table 5

The comparison results of the movement recognition and analysis methods

	CS/%	CV/%
STA-LSTM	73.40	81.20
ST-GCN	81.50	88.30
2s-AGCN	88.50	95.10
DGNN	89.90	96.10
SA-BiLSTM	91.27	96.58

From Table 4, it can be observed that our method achieved an accuracy improvement of 17.87% compared to the STA-LSTM method, 9.77% compared to the ST-GCN method, 2.77% compared to the 2s-AGCN method, and 1.37% compared to the DGNN method in the field of CS; while in the field of CV, the SA-BiLSTM method showed an accuracy improvement of 15.38% compared to the STA-LSTM method, 8.28% compared to the ST-GCN method, 1.48% compared to the 2s-AGCN method, and 0.48% compared to the DGNN method. These results demonstrated the superiority of the SA-BiLSTM approach in movement recognition and analysis.

Finally, the effectiveness of the SA-BiLSTM algorithm for recognizing and analyzing aerobics movements was verified using a self-built dataset. Additionally, how the improvements made to LSTM affect the results was analyzed, as depicted in Fig. 4.

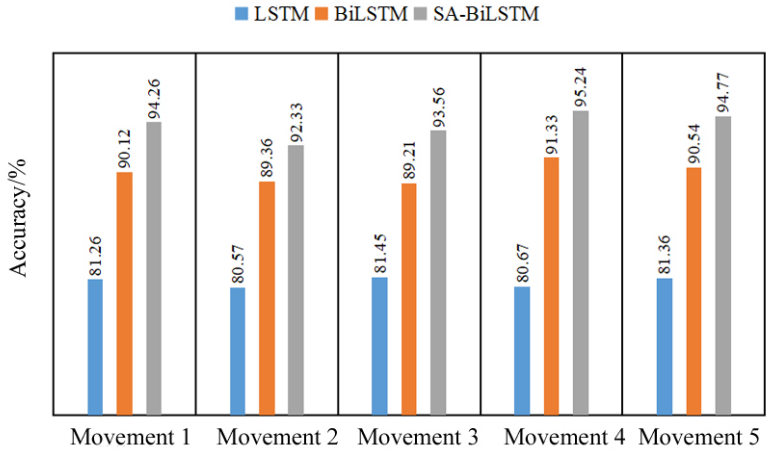

Fig. 4. The analysis of the recognition and analysis effect for the self-built set

From Figure 4, firstly, in the recognition and analysis of different actions, LSTM has the lowest accuracy at around 80%, indicating poor performance in learning features from aerobics videos. Subsequently, the BiLSTM algorithm shows some improvement with an accuracy of around 90%, suggesting that forward and backward learning helps enhance the action recognition and analysis performance of the LSTM algorithm. The accuracy of the SA-BiLSTM algorithm significantly improved after incorporating the SA module, with both exceeding 90%. Taking movement 1 as an example, the accuracy of the SA-BiLSTM algorithm increased by 13% compared to the LSTM algorithm and by 4.14% compared to the BiLSTM algorithm. Calculations showed that the LSTM, BiLSTM, and SA-BiLSTM algorithms achieved average accuracies of 81.06%, 90.11%, and 94.03% respectively for the custom-created dataset, proving the reliability of the improvement made in this paper to the LSTM algorithm.

5. Conclusion

This paper conducted research into recognizing and analyzing aerobics movements through video feature extraction. An SA-BiLSTM algorithm was designed, and experiments were performed using the NTU RGB+D dataset as well as a custom-created dataset. The results showed that the feature fusion technique adopted in this article could effectively improve the accuracy of movement recognition and analysis, and the SA-BiLSTM algorithm achieved an accuracy of 91.27% and 96.58% for CS and CV in the NTU RGB+D dataset, which was better than the ST-GCN method and the other methods. It was also found that the SA-BiLSTM algorithm achieved an average accuracy of 94.03% for recognizing different aerobics movements in the self-built set. These results suggested the

effectiveness of the SA-BiLSTM algorithm, which can be further promoted and applied in practice.

REFERENCES

[1]. *S. Bai, L. Chen and L. Zhao*, "Research on the evolution of movement difficulty of competitive aerobics based on digital image processing", J. Intell. Fuzzy Syst., no. 3, 2021, pp. 1-7.

[2]. *C. Chen and X. Zhu*, "Application research on information security of aerobics information digital system based on Internet of things technology", J. Intell. Fuzzy Syst., no. 1, 2021, pp. 1-8.

[3]. *W. Luo and B. Ning*, "High-Dynamic Dance Motion Recognition Method Based on Video Visual Analysis", Sci. Programming, **vol. 2022**, 2022, pp. 1-9.

[4]. *Q. Ye, Z. Tan, C. Qu and L. Zhang*, "Human motion recognition using three-dimensional skeleton model based on RGBD vision system", J. Phys. Conf. Ser., **vol. 1754**, no. 1, 2021, pp. 1-5.

[5]. *Q. Liu*, "Human motion state recognition based on MEMS sensors and Zigbee network", Comput. Commun., **vol. 181**, 2022, pp. 164-172.

[6]. *P. Gao, D. Zhao and X. Chen*, "Multi-dimensional data modelling of video image action recognition and motion capture in deep learning framework", IET Image Process., **vol. 14**, no. 7, 2020, pp. 1257-1264.

[7]. *H. Hu, W. Zhou, X. Li, N. Yan and H. Li*, "MV2Flow: Learning Motion Representation for Fast Compressed Video Action Recognition", ACM T. Multim. Comput., **vol. 16**, no. 3s, 2020, pp. 1-19.

[8]. *N. Jaouedi, N. Boujnah and M. S. Bouhlel*, "Deep Learning Approach for Human Action Recognition Using Gated Recurrent Unit Neural Networks and Motion Analysis", J. Comput. Sci., **vol. 15**, no. 7, 2019, pp. 1040-1049.

[9]. *Z. Pan and C. Li*, "Robust basketball sports recognition by leveraging motion block estimation", Signal Process. Image, **vol. 83**, no. 10, 2020, pp. 1-6.

[10]. *M. T. K. Tsun, B. T. Lau and H. S. Jo*, "An Improved Indoor Robot Human-Following Navigation Model Using Depth Camera, Active IR Marker and Proximity Sensors Fusion", Robotics, **vol. 7**, no. 1, 2018, pp. 1-23.

[11]. *A. P. Yunus, N. C. Shirai, K. Morita and T. Wakabayashi*, "Time Series Human Motion Prediction Using RGB Camera and OpenPose", Int. Symp. Affect. Sci. Eng., **vol. ISASE2020**, 2020, pp. 1-4.

[12]. *A. M. Vukicevic, M. Djapan, V. Isailovic, D. Milasinovic, M. Savkovic and P. Milosevic*, "Generic compliance of industrial PPE by using deep learning techniques", Safety Sci., **vol. 148**, 2022, pp. 1-8.

[13]. *J. Y. Li*, "Fall Detection Technology Based on Alphapose Bone Key Points and GRU Neural Network", Comput. Sci. Appl., **vol. 11**, no. 4, 2021, pp. 840-848.

[14]. *Z. S. Khozani, F. B. Banadkooki, M. Ehteram, A. N. Ahmed and A. El-shafie*, "Combining autoregressive integrated moving average with Long Short-Term Memory neural network and optimisation algorithms for predicting ground water level", J. Clean. Prod., **vol. 348**, 2022, pp. 1-21.

[15]. *H. Tang, Y. Yin and H. Shen*, "A model for vessel trajectory prediction based on long short-term memory neural network", J. Mar. Eng. Technol., no. 3, 2019, pp. 1-10.

[16]. *N. Deng, H. Fu and X. Chen*, "Named Entity Recognition of Traditional Chinese Medicine Patents Based on BiLSTM-CRF", Wirel. Commun. Mob. Com., **vol. 2021**, no. 2, 2021, pp. 1-12.

[17]. *J. Shobana and M. Murali*, “An Improved Self Attention Mechanism Based on Optimized BERT-BiLSTM Model for Accurate Polarity Prediction”, Comput. J., **vol. 66**, no. 5, 2022, pp. 1279-1294.
[18]. *H. Basak, R. Kundu, P. K. Singh, M. F. Ijaz, M. Woźniak and R. Sarkar*, “A union of deep learning and swarm-based optimization for 3D human action recognition”, Sci. Rep., **vol. 12**, no. 1, 2022, pp. 1-17.
[19]. *S. Song, C. Lan, J. Xing, W. Zeng and J. Liu*,“An End-to-End Spatio-Temporal Attention Model for Human Action Recognition from Skeleton Data”, Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31, no. 1, 2016, pp. 1-7.
[20]. *S. Yan, Y. Xiong and D. Lin*, “Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition”, Proceedings of the AAAI Conference on Artificial Intelligence, 2018.
[21]. *L. Shi, Y. Zhang, J. Cheng and H. Lu*, “Two-Stream Adaptive Graph Convolutional Networks for Skeleton Based Action Recognition”, Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12018-12027.
[22]. *L. Shi, Y. Zhang, J. Cheng and H. Lu*, “Skeleton-Based Action Recognition With Directed Graph Neural Networks”, Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7904-7913.