
Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 5, pp. 2053–2060, DOI 10.12694/scpe.v26i5.4857

HUMAN BEHAVIOR RECOGNITION IN COMPLEX SCENES BASED ON DEEP LEARNING

RUIFENG HONG*

Abstract. In order to overcome some of the problems that have been encountered in the past, for example, due to their reliance on manual feature extraction and limitation of model generalization, this paper proposes a new method to identify people's behaviour in complicated situations. Based on Convolutional Neural Networks (CNN), this method has been proposed for the automatic extraction of a large number of datasets. In addition, the Long Short Term Memory (LSTM) network is used to capture the long-run dependence in time order. Lastly, we use the soft max classifier to classify the various actions of people. Experiments show that the CLT network is able to achieve a high performance of 97.5% over 13 different types of people, outperforming CNN alone, LSTM, and BP models on the DaLiAc dataset, demonstrating superior performance in human behavior recognition and classification. The accuracy, recall, and F1 score evaluation indicators of the CLT net model are the highest, while all indicators of the BP model are the lowest, indicating that the CLT net model has good stability and reliability in recognizing and classifying different human behaviors.

Key words: Human behavior recognition, Deep learning, Convolutional neural network, Long Short Term Memory Network

1. Introduction. With the continuous breakthroughs in computer vision research and computer hardware performance, the idea of machines possessing partial visual abilities of the human eye has become possible. As a core topic in the field of computer vision, human behavior recognition has been active at the forefront of research, playing a key role in intelligent security, medical assistance, smart education, human-computer interaction, and gradually changing people's lives [1]. The traditional video surveillance methods mainly rely on manual monitoring, which inevitably results in a significant waste of manpower and material resources. Due to the limitations of the human body, the monitor cannot maintain a high level of attention at all times during work. When the monitor becomes tired or briefly leaves the monitoring screen, it is difficult to handle emergencies in a timely manner, ultimately leading to missed or false detections in the monitoring process [2]. Secondly, the video streams generated by traditional monitoring methods only include simple storage and playback functions, and there is still a possibility of false or missed detections when manually reviewing historical events [3].

With the fast developing of smart hardware and computing techniques, it is becoming more and more important to substitute for conventional human surveillance with artificial intelligence, complete the recognition and analysis of human behavior and actions in the monitoring area, and combine the advantages of computer vision technology such as real-time, efficiency, and accuracy with video surveillance to build intelligent video surveillance has become an inevitable development trend [4]. Deep learning based human behavior recognition in complex scenarios mainly includes two aspects: motion object detection and human behavior recognition. By collecting video data of human behavior actions through cameras, using object detection methods to identify and detect moving targets, extracting behavioral action features of moving targets and completing classification, the current human behavior state can be determined, providing timely and effective information for staff [5-6]. However, despite the enormous potential of deep learning in behavior recognition, it still faces many challenges. In complicated situations, for instance, the influence of environmental noise and disturbance on the recognition capability of the model will be influenced. Furthermore, the behavior patterns of each person are not uniform enough, so that it is difficult to generalize the model. This paper is intended to investigate the approach of human behaviour identification in complicated situations, especially for complicated background, illumination, multi-object occlusions, etc.

*Guangzhou Civil Aviation College, Guangdong, China (hongrf@qq.com)

2. Literature Review. Deep learning based human behavior recognition is a process in which machines obtain human limb movement information from videos and complete action category judgment. It is a research hotspot in the field of computer vision and a key technology in video understanding. It has huge application prospects in smart homes, surveillance security, smart healthcare, education, and other fields [7]. Although many algorithms have emerged in this field, there are still problems such as difficulty in determining the starting time of actions, poor recognition performance in cases of occlusion or low resolution, and insufficient real-time recognition. Wang et al. developed a meta-learning framework named MetaTTE, which uses a well-designed DED (a data-pre-processing module and a coder decoding net module) to continually deliver an accurate journey time estimation [8]. Fuentes et al. proposed an approach to tracking the behaviour of an individual cow by using the image data as an input to identify the movement. In this paper, we use an image sequence as an input, which is used to recognize the hierarchy of activities that are divided into parts and single actions. These areas of concern are then entered into the tracing and identifying mechanism so that the system can continually follow every individual on the spot and give them a unique identifier. Through this approach, the behaviour of cows can be continually monitored and statistically analyzed to assess behavioral changes in time [9]. Jiao et al. presented a new approach to detect the variation of facial features in movement, and made a comparison between them and the former one. Experiments indicate that the algorithm is more accurate by 1. 68%, and the precision of the modified moving historical image is raised by 14.8%. The proposed approach has been successfully applied to the identification of human motion [10].

In order to improve the precision of human behaviour identification, we propose a new approach, called CLT network, which is based on temporal and spatial characteristics fusion. Using Convolutional Neural Networks (CNN), this approach utilizes Long Short Term Memory (LSTM) net to capture the intrinsic temporal relations of the data. The classification of human behaviors is then done using a softmax classifier within this framework.

3. Method.

3.1. CNN Model. Convolutional neural networks have achieved remarkable success in many fields such as object detection, facial recognition, and speech recognition, there is still no widely accepted architecture for their use in sequence signal classification [11]. To address this gap, the author developed a CNN model tailored for human behavior recognition, inspired by the LeNet-5 architecture. Unlike LeNet-5, this CNN model is designed to handle sequential data as input. After each max pooling layer, additional batch normalization and activation layers (using Leaky ReLU) are incorporated [12]. The CNN structure includes a number of critical elements as shown in Figure 3.1: Sequential Entry Level, Collapse Level, 3 CNN Characteristic Extracting Levels (Convolution, Max Pooling, Batch Normalization, and Leaky ReLU Activation Layer), Unfolding, Leveling, Full Connection, and Soft Max Classifying Layer [13]. The feature extraction layers are crucial for the model, with the convolutional layers extracting essential features from human behavior data [14]. The max pooling layers serve to compress the data and reduce dimensionality, while the batch normalization layers ensure that the extracted features are normalized. The Leaky ReLU activation layers introduce non-linearity, aiding in the effective mapping of features post-normalization. Lastly, the full connection level reduces the loss of the data in the process of extracting the characteristic, and makes the ultimate classification of the person's actions by soft max.

The proposed method can not only increase the rate of convergence, but also reduce the "gradient dispersion" and improve the stability of the training model [15]. The Leaky Relu Trigger Function is used to solve the problem that Relu has a negative input and a constant output of zero, whereas the first derivative is also zero, which leads to no change in neural parameters and no learning. The definition is shown in equation 3.1:

$$f(x) = \begin{cases} x, x \geqslant 0 \\ s \times x, x < 0 \end{cases} \tag{3.1}$$

Among them, s is a non-negative number not smaller than 1. The Leaky Relu activation function becomes Relu when s is set to zero. The softmax classification layer is shown in equation 3.2:

$$s(x_i) = \frac{e^{x_j}}{\sum_{j=1}^{K} e^{x_j}}, i = 1, 2, \cdots, K \tag{3.2}$$

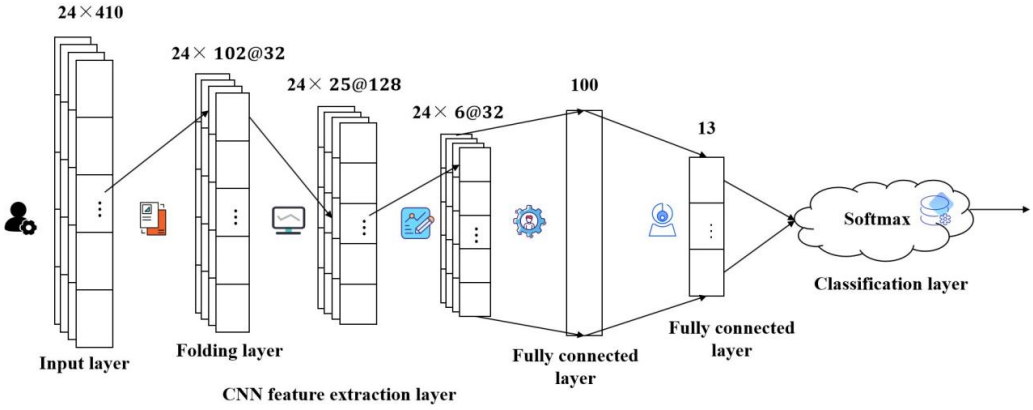

Fig. 3.1: Structure of CNN Model

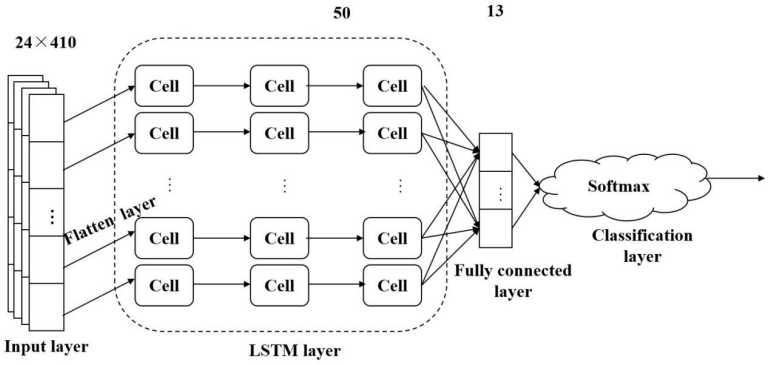

Fig. 3.2: Structure of LSTM Model

Among them, Xi is the character order of the extracted behavioral data, and K is the quantity of human behaviour. The Soft Max Function's classification result indicates the probability that an input sample will be categorized into different classes, and the total probability is 1 [16].

3.2. LSTM Model. LSTM was developed by SCHMIDHUBER and HOCHREITER in 1997 to improve Recurrent Neural Network (RNN). The key parts of LSTM networks include sequential entry and LSTM. The sequential entry level is capable of inputting sequential or temporal information into the net, and LSTM level is able to study the long term dependence among sequential data time steps, which efficiently resolves the RNN gradient disappearing. Because LSTM is an effective way to deal with and predict time signals, the LSTM is used as a characteristic filter in CLT network model [17]. The LSTM model is illustrated in Figure 3.2, which consists of sequential input, flat, LSTM, full connection, and soft max.

From Figure 3.2, we can see that the sequence entry level has a sample size of 24x410 x 1. Then, the multi-dimension data is smoothed out as an input to LSTM level. The LSTM level has 50 hidden cells, and the full connection level has 13 hidden cells. At last, we apply soft max classifier to classify the various actions of people. The LSTM layer's units provide temporal dependence and temporal properties of the input data [18]. The LSTM network achieves long-term control of a unit, which is then used for classification and prediction of temporal signals. Cellular function is primarily implemented by logical gates, input gates, and output gates. Figure 3.3 illustrates the inner architecture of the LSTM layer cells.

The LSTM layer is able to learn the weights such as the weight of the input, the value of the recurrence and the value of the error. Matrices W, R, and B represent a sequence of input weights, recurrence weights,

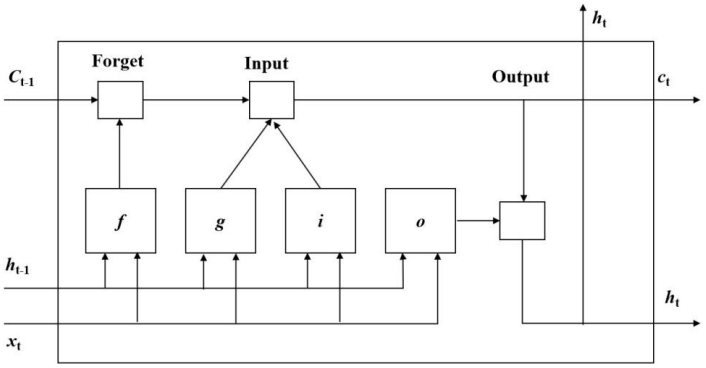

Fig. 3.3: Internal structure of LSTM cells

and deviations, as illustrated in formula 3.3:

$$W = \begin{bmatrix} W_i \\ W_f \\ W_g \\ W_o \end{bmatrix}, R = \begin{bmatrix} R_i \\ R_f \\ R_g \\ R_o \end{bmatrix}, b = \begin{bmatrix} b_i \\ b_f \\ b_g \\ b_o \end{bmatrix} \tag{3.3}$$

The output of the cell status and the output of the hidden state at the time t is given by the expressions 3.4 and 3.5:

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{3.4}$$

$$h_t = o_t \odot \sigma_c(c_t) \tag{3.5}$$

Among them, $\odot$ is Hadamard product (multiplication of vector elements); σ_c is the hyperbolic tangent function(tanh) state activation function.

At time t in Figure 3.3, forget f_t, activate input i_t, output o_t, activate candidate unit input g_t as shown in Equation 3.6 and Equation 3.9:

$$f_t = \sigma_g(W_f x_t + R_f h_{t-1} + b_f) \tag{3.6}$$

$$i_t = \sigma_g(W_i x_t + R_i h_{t-1} + b_i) \tag{3.7}$$

$$o_t = \sigma_g(W_o x_t + R_o h_{t-1} + b_o) \tag{3.8}$$

$$g_t = \sigma_c(W_g x_t + R_g h_{t-1} + b_g) \tag{3.9}$$

Use h_{t-1} and x_t as input information for the current time step in network training. After passing through the gate activation function , these pieces of information ultimately result in an output value between [19].

The larger the forget gate activation ft is, the less the previous cell state output ct−1 is forgotten. Conversely, the larger the input gate activation it is, the more the candidate input gt is represented, allowing more information to be written into the cell state at the current time. Together, the forget gate ft and input gate it determine how much of the new input information is incorporated into the current cell state output ct. Additionally, the activation of the output gate ot dictates the current hidden state output ht. This combined control mechanism enables the model to capture long-term dependencies in human behavior data over time steps.

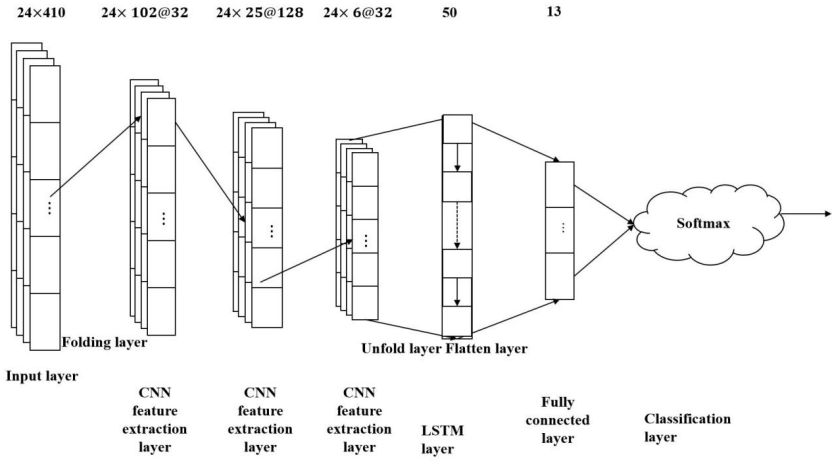

Fig. 3.4: CLT net model structure

3.3. CLT net human behavior recognition model. Since the INS can be viewed as a time sequence, the conventional SVM based on artificial characteristics might result in sub-optimal information usage and insufficient recognition of complicated human activities. In order to solve this problem, we present a novel CLT Net, which combines spatial-temporal and spatial characteristics to identify people's behaviour. The CLT Net model is a combination of CNNs that are good at automatic extraction from data, and LSTMs that can effectively capture time dependence in time sequence data. The proposed model uses the same architecture as CNN, except that the LSTM layer is used instead of a full connection layer. The CLT Net Model's parameters and functional specs match those of the CNN and LSTM modules. The CLT Net architecture, as shown in Figure 3.4, consists of a sequential entry level, a folding level, a CNN characteristic extracting layer (consisting of convolutions, Max Pooling, Batch Normalization, Leaky ReLU), unfolding, flat, LSTM, full connection, and soft max [20].

A CLT network model is employed to classify a person's behaviour in this way. Firstly, a CNN module is used to extract the input person's behavioral data series. Then, the 2D data characteristics are compressed to one dimension and transferred to the LSTM layer, where temporal feature filtering takes place. Next, a fully connected layer maps the selected features to the label space using a weight matrix. Lastly, we use soft max level to classify and select the class that has the best forecast probability as the forecast tag of the input data sample. During training, the model adjusts its parameters by comparing the predicted categories from forward propagation with the true labels of the samples. The error is used to backpropagate through the network, with the loss function and optimizer continually refining the weights and bias terms, allowing the model to improve and ultimately achieve optimal performance.

3.4. Experimental verification.

3.4.1. Experimental Dataset. The author conducted research on human behavior recognition based on wearable sensor data, and the experiment used the publicly available DaLiAc (Daily Life Activities) dataset. The dataset was collected by placing four 6-axis inertial sensor nodes on the subjects' right buttock, chest, right wrist, and left ankle. Each sensor node includes a three-axis accelerometer and a three-axis gyroscope. The accelerometers have a range of ±6g, and the gyroscopes, particularly the one on the wrist, chest, and hip sensor nodes is $\pm 500(\circ)/s$, the gyroscope range of the ankle sensor node is $\pm 2000(\circ)/s$, and the data sampling frequency is $204.7Hz^2$. A total of 19 healthy subjects participated in the data collection experiment (8 females, 11 males, age 26 ± 8 years, height $177 \pm 11cm$, weight $75.2 \pm 14.2kg$, deviation ± mean), and a total of 13 activities were collected. The activities and corresponding tags are shown in Table 3.1.

Table 3.1: Activities and Corresponding Tags

Activity Description	Label
Sit still	1
Lie flat	2
Stand	3
Wash dishes	4
Vacuum cleaner	5
Sweep the floor	6
Walk	7
Go upstairs	8
Go downstairs	9
Treadmill running	10
Test bike riding (50W)	11
Test bike riding (100W)	12
Skipping rope	13

Table 3.2: Experimental Parameter Settings

Parameter	Set up
Initialization of CNN layer weight coefficients	Kaiming method
Initialization of LSTM layer weight coefficients	Orthogonal method
Initialization of fully connected layer weight coefficients	Kaiming method
Optimizer	Adam optimizer
Loss function	Cross baking
Initial learning rate	0.001
Sample sequence size	24× 410
Number of samples in the training set	20088
Number of test set samples	2232
Training rounds	20
Batch size	500
Leaky Relu factor	0.1

3.4.2. Experimental operating environment. All of the author's models were trained and tested on a computer equipped with a Core i5-6500U CPU @ 3.20 GHz and 16 GB of RAM. The system ran Windows 10 Professional 64-bit, and the models were developed using the Matlab 2020b Deep Learning Toolbox framework.

3.4.3. Experimental Parameters. Firstly, the human behavior data is divided into samples with a sliding window length of 410 (rounded to twice the sampling frequency), and there is 50% data overlap between adjacent windows. As a result, each sample sequence has a size of 24 × 410 (corresponding to 4 sensors with 6-axis data each). After segmenting the data, the samples were sorted, with the top 90% used as the training set and the remaining 10% as the testing set. The experimental parameters are detailed in Table 3.2. During the simulation experiments, all models were initialized with the same configuration to ensure a fair comparison, allowing for a more accurate assessment of the true performance of the CNN models, LSTM models, and CLT net models.

The CNN layers and fully connected layers utilize Kaiming initialization for their weights to speed up model convergence. For the LSTM layers, the weights are initialized using the orthogonal method. All models are optimized using the Adam algorithm, an adaptive moment estimation method that offers quicker convergence and lower memory usage. This approach also eliminates the need for a validation set during training.

4. Results and Discussion. To effectively showcase the generalization capability of the CLT net model, we calculated the macro precision, macro recall, and macro F1 score for the test results of the LSTM, CNN, CLT net, and traditional BP models. These metrics were computed by averaging the precision, recall, and

Table 4.1: Comparison of evaluation indicators for BP, LSTM, CNN, and CLT net models

Model	Macro precision	Macro recall rate	Macro F1 value
BP	0.5221	0.5127	0.5002
LSTM	0.7585	0.7117	0.7260
CNN	0.9510	0.9503	0.9501
CLT-net	0.9635	0.9605	0.9616

F1 score across all 13 categories of human behavior. The performance of the four models is summarized and compared in Table 4.1.

The overall average classification accuracy of BP, LSTM, and CNN models were 61.6%, 77.5%, and 96.3%, respectively. The author proposed the CLT net model, which achieved 97.5%, an improvement of 35.8, 20.1, and 1.1 percentage points, respectively. The CNN model can extract features of human behavior data, which represent the original human behavior data to the maximum extent possible. Using these features for human behavior recognition and classification has good performance. Compared to the LSTM model, the CNN model has a higher recognition rate. The LSTM model is only used for modeling temporal data to learn the correlation between data, and cannot achieve feature extraction. This also indicates that feature extraction is the key to classification and recognition, and the CNN feature extraction module is the most important component of the CLT net model. The precision assessment measure measures the proportion of the correct positive samples to the total expected positive, while the recall measure measures the proportion of correct recognized positive samples to the overall true positives. F1 is the harmonic mean of accuracy and recall. According to Table 4.1, the CLT net model exhibits the highest accuracy among the evaluated models, recall, and F1 score evaluation indicators, while all indicators of the BP model are the lowest, indicating that the CLT net model has good stability and reliability in recognizing and classifying different human behaviors.

5. Conclusion. The author presents a study on recognizing human behavior in complex environments using deep learning techniques, introducing a novel model called CLT net that leverages spatiotemporal feature fusion. The proposed approach combines Convolutional Neural Networks (CNNs) to extract the time dependence of time sequence data using Long Short-Term Memory (LSTM) networks, and uses the soft max classifier to classify the behavior. The experimental results on the DaLiAc dataset show that compared to LSTM, CNN, and BP models, the CLT net model converges faster and has better performance in human behavior recognition and classification. Subsequently, lightweight deep learning models will be constructed to optimize sensor based human behavior recognition methods and further improve feature recognition accuracy.

REFERENCES

[1] Dong, X. Q., Wang, X. C., Li, B. J., Wang, H. Y., & Chen, G. C. (2024). Mp-abr: a framework for intelligent recognition of abnormal behaviour in multi-person scenarios. Multimedia Tools and Applications, 83(18), 55605-55626.
[2] Wu, H., Han, Y., & Meng ZhangBihonegn Dianarose AbebeMolla Betelhem LegesseRuoyu Jin. (2023). Identifying unsafe behavior of construction workers: a dynamic approach combining skeleton information and spatiotemporal features. Journal of construction engineering and management, 149(11), 1-15.
[3] Ozdemir, C., Hoover, R. C., Caudle, K., & Braman, K. (2024). Tensor discriminant analysis on grassmann manifold with application to video based human action recognition. International Journal of Machine Learning and Cybernetics, 15(8), 3353-3365.
[4] Noor, T. H. (2023). Human action recognition-based iot services for emergency response management. Machine Learning and Knowledge Extraction, 5(1), 330-345.
[5] Reddy, G. V., Deepika, K., Malliga, L., Hemanand, D., Senthilkumar, C., & Gopalakrishnan, S., et al. (2023). Human action recognition using difference of gaussian and difference of wavelet. Big Data Mining and Analytics, 6(3), 336-346.
[6] Ye, Q., Tan, Z., & Zhang, Y. (2022). Human action recognition method based on motion excitation and temporal aggregation module. Heliyon, 8(11), 11401.
[7] Sun, B., Kong, D., Zhang, W., & Jia, W. (2022). Survey on human action recognition from depth maps, 27(6), 29.
[8] Wang, C., Zhao, F., Zhang, H., Luo, H., Qin, Y., & Fang, Y. (2022). Fine-grained trajectory-based travel time estimation for multi-city scenarios based on deep meta-learning. arXiv e-prints, 41(12), 3788-3817.
[9] Fuentes, A., Han, S., & Nasir, Muhammad FahadPark, JongbinYoon, SookPark, Dong Sun. (2023). Multiview monitoring

of individual cattle behavior based on action recognition in closed barns using deep learning. animals, 13(12), 76(9), 2667-2684.
[10] Jiao, C. (2022). Recognition of human body feature changes in sports health based on deep learning. Computational and Mathematical Methods in Medicine, 2022.
[11] Srihari, P., Harikiran, J., & Reddy, C. V. S. (2023). Effective framework for human action recognition in thermal images using capsnet technique. Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology, 45(6), 11737-11755.
[12] Gutoski, M., Lazzaretti, A. E., & Lopes, H. S. (2023). Unsupervised open-world human action recognition. Pattern analysis and applications: PAA, 26(4), 1753-1770.
[13] Kumar, R., & Kumar, S. (2024). A survey on intelligent human action recognition techniques. Multimedia Tools and Applications, 83(17), 52653-52709.
[14] Sam Slade Li Zhang Yonghong Yu Chee Peng Lim. (2022). An evo1ving ensemb1e mode1 of mu1ti stream convo1utiona1 neura1 networks for human action recognition in sti11 images. Neural Computing and Applications, 34(11), 9205-9231.
[15] Basak, H., Kundu, R., Singh, P. K., Ijaz, M. F., Woniak, M., & Sarkar, R. (2022). A union of deep learning and swarm-based optimization for 3d human action recognition. Scientific Reports, 12(1), 1-17.
[16] Berlin, Jeba, S., John, & Mala. (2022). Light weight convolutional models with spiking neural network based human action recognition. Journal of intelligent & fuzzy systems: Applications in Engineering and Technology, 39(1), 961-973.
[17] Ghosh, S. K., Rashmi, M., Mohan, B. R., & Guddeti, R. M. R. (2022). Deep learning-based multi-view 3d-human action recognition using skeleton and depth data. Multimedia Tools and Applications, 82(13), 19829-19851.
[18] Wu, Q., Huang, Q., & Li, X. (2022). Multimodal human action recognition based on spatio-temporal action representation recognition model. Multimedia Tools and Applications, 82(11), 16409-16430.
[19] Islam, M. S., Bakhat, K., Khan, R., Naqvi, N., Islam, M. M., & Ye, Z. (2022). Applied human action recognition network based on snsp features. Neural Processing Letters, 54(3), 1481-1494.
[20] Yongfeng, Q., Jinlin, H., & Xiaoxu, L. P. (2023). Semantic-guided multi-scale human skeleton action recognition. Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies, 53(9), 9763-9778.

Edited by: Hailong Li
Special issue on: Deep Learning in Healthcare
Received: Sep 9, 2024
Accepted: Oct 11, 2024