Hindawi
Journal of Electrical and Computer Engineering
Volume 2023, Article ID 9780325, 1 page
https://doi.org/10.1155/2023/9780325

Retraction

Retracted: A Multimodal Information Fusion Model for Robot Action Recognition with Time Series

Journal of Electrical and Computer Engineering

Received 19 December 2023; Accepted 19 December 2023; Published 20 December 2023

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] X. Zhang, H. Li, and M. Qian, "A Multimodal Information Fusion Model for Robot Action Recognition with Time Series," *Journal of Electrical and Computer Engineering*, vol. 2022, Article ID 7270412, 11 pages, 2022.

Hindawi
Journal of Electrical and Computer Engineering
Volume 2022, Article ID 7270412, 11 pages
https://doi.org/10.1155/2022/7270412

Research Article

A Multimodal Information Fusion Model for Robot Action Recognition with Time Series

Xiaozhi Zhang, Hongyan Li, and Mengjie Qian

Information Engineering Department, Hebei Vocational University of Technology and Engineering, Xingtai 054000, China

Correspondence should be addressed to Mengjie Qian; polaris119@hdu.edu.cn

Received 9 May 2022; Revised 20 May 2022; Accepted 31 May 2022; Published 16 June 2022

Academic Editor: Xuefeng Shao

The current robotics field, led by a new generation of information technology, is moving into a new stage of human-machine collaborative operation. Unlike traditional robots that need to use isolation rails to maintain a certain safety distance from people, the new generation of human-machine collaboration systems can work side by side with humans without spatial obstruction, giving full play to the expertise of people and machines through an intelligent assignment of operational tasks and improving work patterns to achieve increased efficiency. The robot's efficient and accurate recognition of human movements has become a key factor in measuring robot performance. Usually, the data for action recognition is video data, and video data is time-series data. Time series describe the response results of a certain system at different times. Therefore, the study of time series can be used to recognize the structural characteristics of the system and reveal its operation law. As a result, this paper proposes a time series-based action recognition model with multimodal information fusion and applies it to a robot to realize friendly human-robot interaction. Multifeatures can characterize data information comprehensively, and in this study, the spatial flow and motion flow features of the dataset are extracted separately, and each feature is input into a bidirectional long and short-term memory network (BiLSTM). A confidence fusion method was used to obtain the final action recognition results. Experiment results on the publicly available datasets NTU-RGB + D and MSR Action 3D show that the method proposed in this paper can improve action recognition accuracy.

1. Introduction

The advancement of technology has given rise to human-robot interaction systems. Human-robot interaction refers to the ability to communicate and interact between humans and machines, and the prerequisite for achieving this activity is that the robot is able to accurately read the human language, behavior, action intentions, etc. [1–3]. Currently, the carriers of human-machine interaction are mostly intelligent robots. Intelligent robots are used in many fields such as the service industry, industry, and agriculture. Robots are moving into a new stage of human-robot collaborative operation led by the new round of information science and technology. Unlike the existing traditional robots that need to use isolation fences to maintain a certain safety distance from people, the new generation of human-machine collaboration systems can work side by side with humans without spatial barriers, giving full play to the expertise of people and machines by intelligently assigning operational tasks and improving work patterns to achieve increased efficiency. For example, in future factories, robots will be responsible for completing repetitive, dangerous, and difficult tasks, while humans will be freed to focus on dynamic planning or work that requires flexibility and toughness. The key to building a human-robot collaboration system is to achieve a more intelligent human-robot interaction. Traditional human-computer interaction usually uses a keyboard, mouse, and other tools to operate, which is obviously clumsy, restricts people's range of activities, and does not really help people to work easily. If people can use voice, gestures, or actions to control the robot to perform relevant operations, it can greatly ease the hard work of people at work. Whether the robot is controlled by voice, gesture, or body movement, the essential requirement is that the robot

be able to recognize human commands. In the field of motion-based robotics, accurate and rapid motion recognition is the key to smooth human-robot interaction.

Human action recognition (HAR) is to parse the human activity behavior from the input data and then determine the specific action category [4–6]. Initially, the idea of HAR is to extract the spatiotemporal features of each image frame in the video. The extracted features are input to the classifier in the form of feature vectors for the training of the model. The feature vectors from the test dataset are fed into the trained model to output the human action classes. The key to this approach is feature extraction. Traditional manual feature extraction is mostly used by hand. In the face of complex and variable action sequences, this approach cannot fully express spatiotemporal information and has certain limitations [7]. With the introduction of deep learning methods, the feature extraction method has been changed from a manual extraction method to automatic extraction. In deep learning methods, the main data used are two data types, RGB video and skeletal data. RGB video consists of multiple consecutive RGB images, and each frame is played for the same amount of time. In the early stages of research on HAR based on deep learning, many deep neural networks have been successful in image classification tasks [8–10]. Image data contains only a spatial dimension; video data also has a temporal dimension on top of that. Therefore, for video data, both intraframe spatial features and interframe temporal features need to be extracted. The two kinds of information together characterize the human action information. Because image classification networks are not directly applicable to action classification tasks, the core problem of HAR in RGB-based videos is how to extract spatiotemporal features from data. Reference [11] uses a convolutional neural network (CNN) feature spatiotemporal extraction for action recognition for each video frame individually. Reference [12] used 3D CNN for action recognition. The experimental results are general. Reference [13] deepens the number of network layers of 3D CNN and proposes an improved 3D CNN model. The model uses 3D convolution to model both visual and motion information with powerful generalization ability and subsequently becomes a general video feature extractor. With the rapid development of sensor technology as well as human pose recognition methods [14], high-precision skeletal data is becoming more and more readily available. Skeletal data is the description of the different positions of each joint in the spatial dimension during human movement. Compared with RGB video data, skeletal data is characterized by not focusing on environmental factors such as color, background, and brightness, but only on human posture and position. Therefore, skeletal information is more robust and robust to changes in viewpoint, body proportions, movement speed, clothing texture, and background [15]. In addition, skeletal data is smaller in magnitude compared to image data, which greatly reduces the time complexity of model training. Based on these advantages, skeletal data is well suited for HAR. The field started to use skeletal data extensively. Initially, action recognition based on skeletal data also tried to use manual feature extraction methods, and with the popularity of deep learning algorithms, automatic skeletal data feature extraction methods based on deep learning algorithms were proposed one after another. The commonly used approaches for modeling skeletal data can be classified as Recurrent Neural Network (RNN) based [16], CNN based [17], Graph Convolutional Network (GCN) based [18], etc. Reference [19] uses Long Short-Term Memory (LSTM) to mine the information in the data. Reference [20] introduced 3D CNN to improve the rate of standing action recognition, but it is not suitable for small data sets. The unimodal behavior recognition method can correctly recognize some actions, but it is difficult to represent human behavior accurately and comprehensively in complex scenes. To solve this problem, some researchers have tried to fuse features from different modalities to exploit their complementarity to achieve better recognition results. Reference [21] fused three different modalities, RGB, RGB-D, and 3D coordinate information and fed the merged data into a multiclassification support vector machine. Reference [22] proposed a network that captures multimodal correlations at arbitrary timestamps, and the network performed well in long-video action recognition. Reference [23] improved image description, skeleton flow, and inertial sensor data with feature fusion, respectively. The fusion results show an improvement of up to 4% in accuracy over feature recognition alone.

By analyzing the existing related studies, the following problems exist. One is that the action recognition effect of multimodal feature fusion is better than that of unimodal feature recognition under a large probability. Second, the different fusion strategies of multimodal feature data can affect the final recognition efficiency. Third, models trained based on multimodal feature fusion tend to take more time and are less suitable for real-time HCI systems. Fourth, even in the context of multifeature fusion as input data, the choice of the classifier is significant for global action recognition results. On the other hand, considering the following difficulties in action recognition in HCI systems, such as interference of environmental background, uncertainty of ambient light intensity, and interference of multiple people, all of them can bring impact on the accuracy of recognition. As the executor of the action, a human has strong discretion and flexibility. For example, in the simplest hand-waving action, the same person in different moments of the execution of the action will also have differences, including waving the hand amplitude and waving speed. Different people have different heights, body types, and distances from the camera. This can lead to different recognition results or even false recognition for people of different body types doing the same action in different positions. Human-robot interaction should have real-time requirements, such as the use of action to control the robot; the basic requirement is that with the execution and recognition of the action, the robot should respond quickly and in real time. However, as the variety of actions increases, the amount of computation also increases, and the timeliness of interactions cannot be guaranteed. In order to solve the above problems, improve the accuracy of action recognition, and meet the demand for real-time action recognition as much as possible, this paper proposes an action recognition model with multimodal

feature fusion of time series and applies the model to robot interaction. The dataset's spatial flow features and motion flow features are extracted separately in this study, and each feature is input into a BiLSTM. A confidence fusion method is used to obtain the final action recognition results. Experiment results on publicly available datasets NTU-RGB + D and MSR Action 3D show that the method proposed in this paper can improve action recognition accuracy when compared to other methods. Furthermore, in this paper, the recognition model is ported to a robot to evaluate the robot's performance in HAR, including recognition accuracy and response time. The effectiveness of the method described in this paper is demonstrated through experiments on a publicly available dataset.

2. Related Knowledge

2.1. Principle of Robot Operation. Robots usually include hardware systems and software systems. The hardware system mainly has cameras, sensors, servos, development boards, communication modules, bodies, and limbs. The lower limbs of limbs are replaced by four wheels. These four wheels can perform forward, backward, left, right, and stop movements. There are four degrees of freedom. The upper limbs can swing back and forth to achieve flat lifting, lifting, bending, and grasping objects. There are 6 degrees of freedom. The development board is STM32, which uses the mainstream Cortex core, a rich software package, a wide range of chip models, rich and reasonable peripherals, reasonable power consumption, reasonable price, and a strong user base. The robot designed in this paper mainly uses the STM32 microcontroller development board. The integrated development environment of STM32 is STM32CubeIDE, and the programming language is C/C++ language. This IDE has a peripheral configuration, code generation, code compilation, and debugging functions for STM32 microcontrollers and microprocessors. ST officially provides libraries for various peripherals of STM32, and the use of ready-made device libraries simplifies the work of project building. Meanwhile, for program development, ST encapsulates library functions for peripherals, and developers do not need to spend much effort to understand the structure of STM32 internal registers. The software design flow of the robot is shown in Figure 1.

2.2. Principle of Motion Recognition. Usually, the recognition of multiple motions can be achieved through model training, thus enabling flexible control of the robot. There are two phases to action recognition: training and recognition. During the training phase, action features are extracted and feature data are fed into the model to train the action recognition classifier, which is represented by the action library in Fig. Various evaluation metrics can be used to calculate the performance of various aspects of the classifier in order to quantify the performance of the trained classifier. During the recognition phase, the sample data to be measured is subjected to feature extraction. To obtain recognition results, action recognition is performed using the trained model. Based on the results of the recognition, the corresponding commands are sent to the robot's operating system, which controls the robot's response. The flow of action recognition is shown in Figure 2.

3. Algorithm

3.1. Algorithmic Framework. HAR serves as the foundation for applications like computer vision and human-computer interaction. Its main purpose is to enable the computer to recognize different human actions, such as raising the left hand, raising both hands, and lifting the left foot. The essence of HAR is the process of having the computer classify the video from the set of categories already given, determine which category of action is present in the input unknown video, and give the result of the judgment. People perform many kinds of action behaviors in the course of their work and life. Such as running, jumping, rowing, and dancing, the movements made by people are different in various situations. Different combinations of movements made by upper and lower limbs demonstrate the activities of people. In order to recognize these movements, this paper proposes a recognition model by fusing multifeature fusion and a deep learning algorithm. Figure 3 depicts the framework of the proposed action recognition method in this paper.

The model is divided into four stages: multiple sequence sampling, feature extraction, classification recognition, and result fusion. The first stage is to perform multiple sequence sampling for each video segment. The feature extraction stage uses CNN to extract spatial flow and motion flow features from video frame sequences and optical flow image sequences, respectively. The classification recognition stage is to input the collected data of multiple modalities into the BiLSTM network for time-series feature modeling. In the fourth stage, the motion classifiers of each modality are fused with confidence to obtain the final experimental results.

3.2. Multimodal Feature Extraction. CNN is used to extract spatial flow features and motion flow features in images. The computation is too large due to processing the images frame by frame. Considering the coherence of motion, all images are sampled to reduce the computational effort. Therefore, it is necessary to choose a feasible and efficient sampling scheme. In this paper, the scheme is adopted as follows: the video is cropped into D clips and each clip is sampled uniformly for M frames. Each video clip X_M is used to train the model. The key frames of each clip $X_M = \{Z_1, Z_2, \ldots, Z_i\}$. To generate multimodal data, the key frames and optical flow images are fed into the CNN for deep feature extraction. Figure 4 depicts the CNN flow for feature extraction.

3.3. BiLSTM-Based Feature Classification Recognition. The video content is continuously changing from moment to moment, and therefore, the pattern of change between video frames derives more information. The information within and between each frame of the video makes the classification more accurate. CNN can only process one input at a time, and the previous input has no bearing on the next. However,

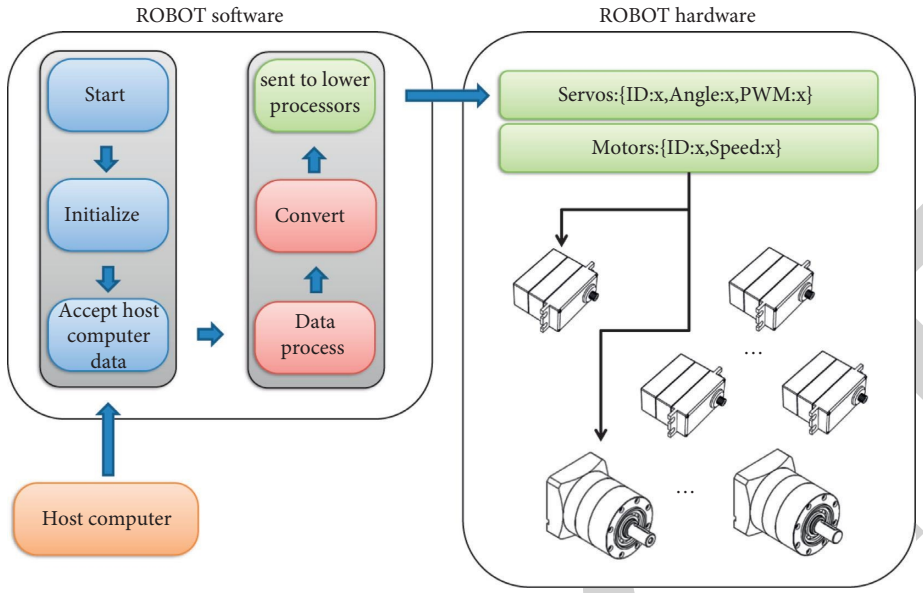

Figure 1: Principle of robot operation.

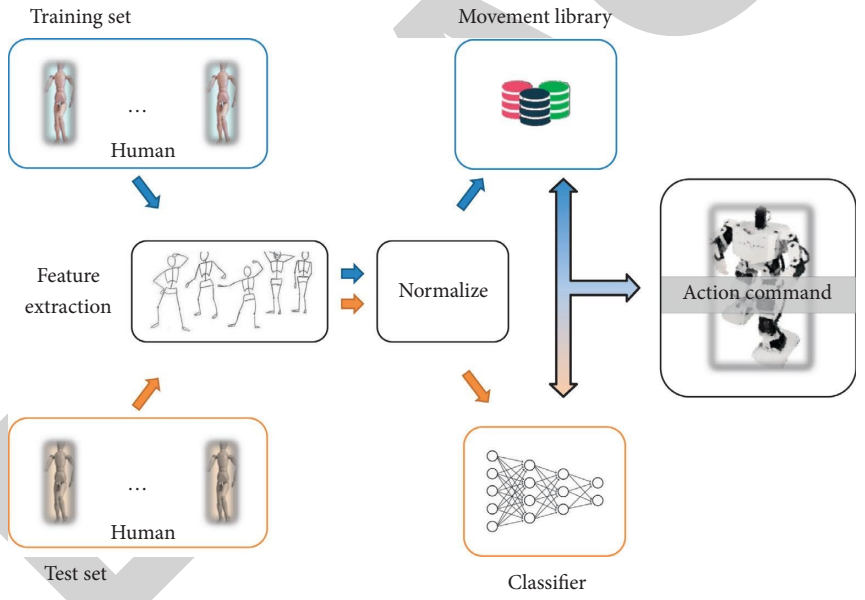

Figure 2: Action recognition process.

some tasks require improved handling of sequential information; i.e., the previous input is related to the subsequent input. RNNs excel at dealing with constantly changing data. The LSTM is an improved RNN introduced in [24]. This network solves the vanishing and exploding gradient problem by remembering long-distance prior information. Since LSTM is not easily parallelized during training, an improved LSTM, namely, BiLSTM [25] was proposed. The BiLSTM model is able to extract effective features from sequential data with complex structure, learn past and future information, and derive labels for the current time. The special cellular units of this network can learn long-term dependencies without preserving invalid contextual information.

The RNN cells are linked to form a loop that can use sequential information to perform the same task for each element of the sequence. It takes as input an arbitrary embedding sequence $x = (x_1, x_2, \ldots, x_T)$, which consists of a hidden cell h and an output y. T represents the final time step. The RNN's hidden state ht is calculated for each time step t based on the previously hidden state h_{t-1} and the current input at x_t, and the hidden and output layers are calculated as follows.

$$\begin{aligned} h(t) &= \mathrm{sig}\left(W_1 x_t + W_2 h_{t-1}\right), \\ y(t) &= g\left(V h_t\right), \end{aligned} \tag{1}$$

where W_1 and W_2 are the weight matrices of the network, and sig and sof are the sigmoid and softmax activation functions, respectively. The activation function is calculated as follows.

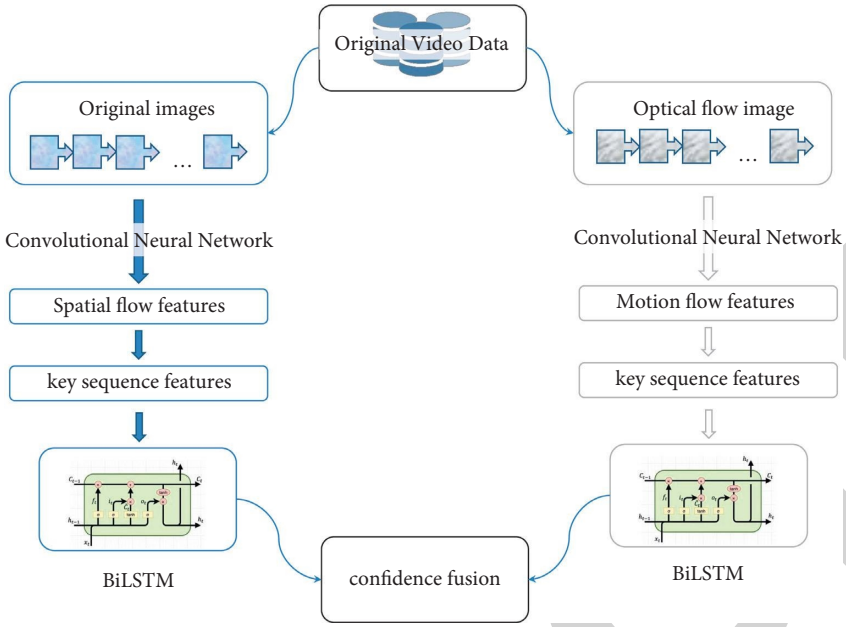

Figure 3: HAR framework.

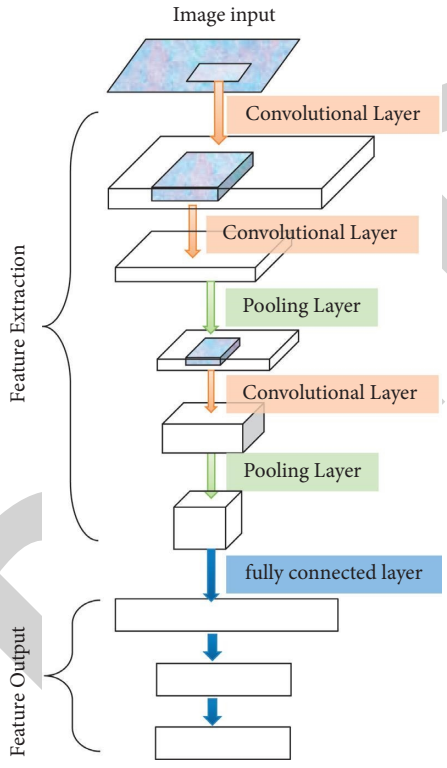

Figure 4: CNN feature extraction schematic.

$$\begin{aligned} \operatorname{sig}(z) &= \frac{1}{1+e^{-z}}, \\ \operatorname{sof}(z_m) &= \frac{e^{z_m}}{\sum_k e^{z_k}}. \end{aligned} \tag{2}$$

LSTM introduces input, output, and forget gate to control the update of hidden and storage units based on RNN. The formula in LSTM is defined as follows.

$$\begin{aligned} i_t &= \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right), \\ f_t &= \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right), \\ \hat{c}_t &= \tanh\left(W_c \cdot \left[h_{t-1}, x_t\right] + b_c\right), \\ c_t &= f_t \cdot c_{t-1} + i_t \cdot \hat{c}_t, \\ o_t &= \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right), \end{aligned} \tag{3}$$

where i_t, f_t, o_t, and c_t are the cell's input, forget, output gates, and output at time point t, respectively. At time point t, the input variables and hidden variables are x_t and h_t, respectively. The sigmoid activation function is represented by σ, and the weight vector and bias vector are represented by $\mathbf{W}$ and $\mathbf{b}$. Figure 5 depicts the structure of an LSTM cell, which consists primarily of three basic gates and a cell.

The input gate, forget gate, and output gate in Figure 5 control the storage and updating of the LSTM memory module's cell cells. Which part of the information after updating is stored in the cell state is mainly determined by the input gate. Which part of the legacy information of the cell state is forgotten is determined by the forget gate. Which part of the updated cell state is output is controlled by the output gate.

BiLSTM is a combination of two directional LSTMs, forward and backward. BiLSTM aims to model past and future context dependencies. Unlike the LSTM network, the BiLSTM network has two parallel structures in both propagation directions, and the forward and backward passes of each layer are executed first from the front and back parts of the input sequence in the same way as other neural networks operate, so the BiLSTM network can preserve the sequence information from the front and back two different directions. It is able to fully take into account contextual information. Since there are two LSTM layers in the network, the vector formulation is adjusted as follows:

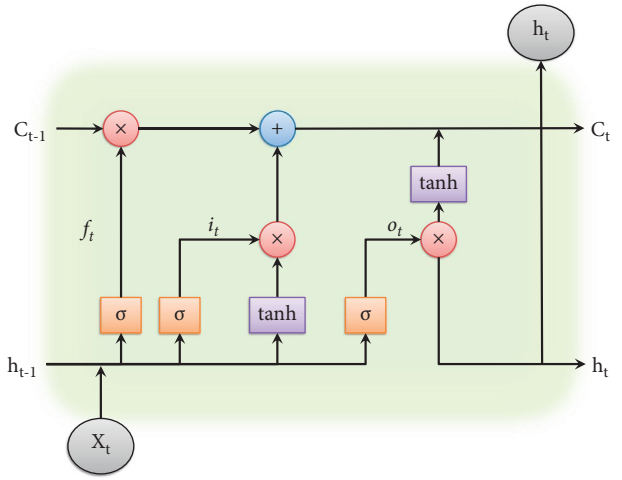

Figure 5: LSTM network structure.

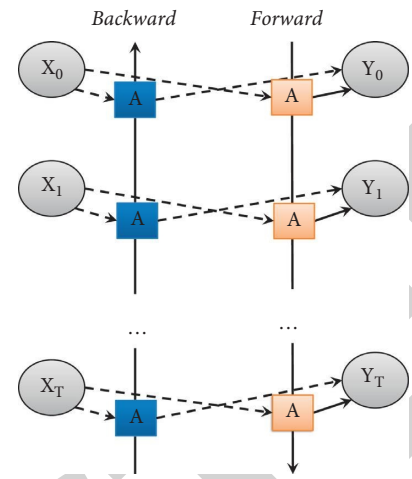

Figure 6: BiLSTM network structure.

$$h_{ft} = H\left(W_{xh_f}x_t + W_{h_fh_f}h_{f_{t-1}} + b_{h_f}\right),$$
$$h_{bt} = H\left(W_{xh_b}x_t + W_{h_bh_b}h_{b_{t-1}} + b_{h_b}\right), \tag{4}$$

where $h_b \in R^d$ and $h_f \in R^d$ represent the backward and forward outputs, respectively. The final output $y_t = [h_{f_t}, h_{b_t}]$ is a concatenation of these two parts and $y_t \in R^{2d}$. The combination of the forward and backward layers is used as a single BiLSTM layer.

In the process of action recognition based on video data, BiLSTM is used to capture the temporal dependence of different key sequence frames and learn the contextual information of action categories with different time spans. In the paper, the input is taken from the final pooling layer of the Inception-V3 network and fed to the BiLSTM network to learn the encoding information in different video sequences. The output of the last layer of the key subsequence is characterized as

$$\mathbf{x}' = \left[\mathbf{x}'^1, \mathbf{x}'^2, \ldots, \mathbf{x}'^{n_i}\right]. \tag{5}$$

In equation (5), n_i represents the number of frames in key subsequence. The sequence features obtained after the BiLSTM network are defined as

$$r = \text{BiLSTM}(\mathbf{x}'). \tag{6}$$

The multimodal features of the key sequences are obtained and fed into each classifier separately to obtain the corresponding class scores:

$$s = \text{soft max}\left(r^i\right). \tag{7}$$

To visually display the structure of the BiLSTM model, Figure 6 visually describes the structure information of the model.

3.4. Result Fusion Strategy. The different features extracted from the data have a complementary effect on each other. In the video data used in this paper, spatial flow features, motion flow features, and time-series features are complementary. Choosing an appropriate fusion strategy for the classification results obtained from different features can significantly improve the accuracy of action recognition. The confidence level of a classifier is an important parameter for measuring the effectiveness of the classification task, as it determines the refuse-recognition threshold and is crucial in the integration of multiple classifiers. The confidence level used in this paper is

$$z_v(x) = (1-\alpha)\left(P_v^{\max}(x) - P_v^{\text{sub max}}(x)\right) + \alpha\left(P_v^{\max}(x) - \frac{1}{n-1}\sum_{j=1}^{c-1} p_{v,j}(x)\right) \tag{8}$$
$$s.t. \quad p_{v,j}(x) \neq P_v^{\max}(x),$$

where $v \in \left\{\text{softmax}_s^i, \text{softmax}_m^i\right\}$. softmax_s^i and softmax_m^i denote the classifiers for RGB streams and optical streams, respectively. The classifier's confidence level in distinguishing a class to which sample x belongs is expressed as $z_v(x)$. $P_v^{\max}(x)$ is the probability that classifier v distinguishes between two categories to which sample x belongs. $P_v^{\text{sub max}}(x)$ is the likelihood that classifier v can distinguish between two categories to which sample x belongs. $(1/c-1)\sum_{j=1}^{c-1} p_{v,j}(x)$ denotes the mean of the probability that classifier v discriminates a category to which sample x belongs. $p_{v,i}(x)$ is the probability that classifier v discriminates a category j to which sample x belongs; $\alpha \in [0,1]$ is the confidence parameter. c represents the number of categories.

Multimodal classifier fusion recognition of action categories is determined. Firstly, the predicted category score vectors $y_s(x)$ and $y_m(x)$ and confidence levels $z^s(x)$ and $z_m(x)$ for the original frame sequence and optical flow image sequence are obtained. Finally, the individual classifier scores and confidence levels are weighted and fused to obtain the final results for human action classification as follows:

$$y(x) = y_s(x) * z_s(x) + y_m(x) * z_m(x). \tag{9}$$

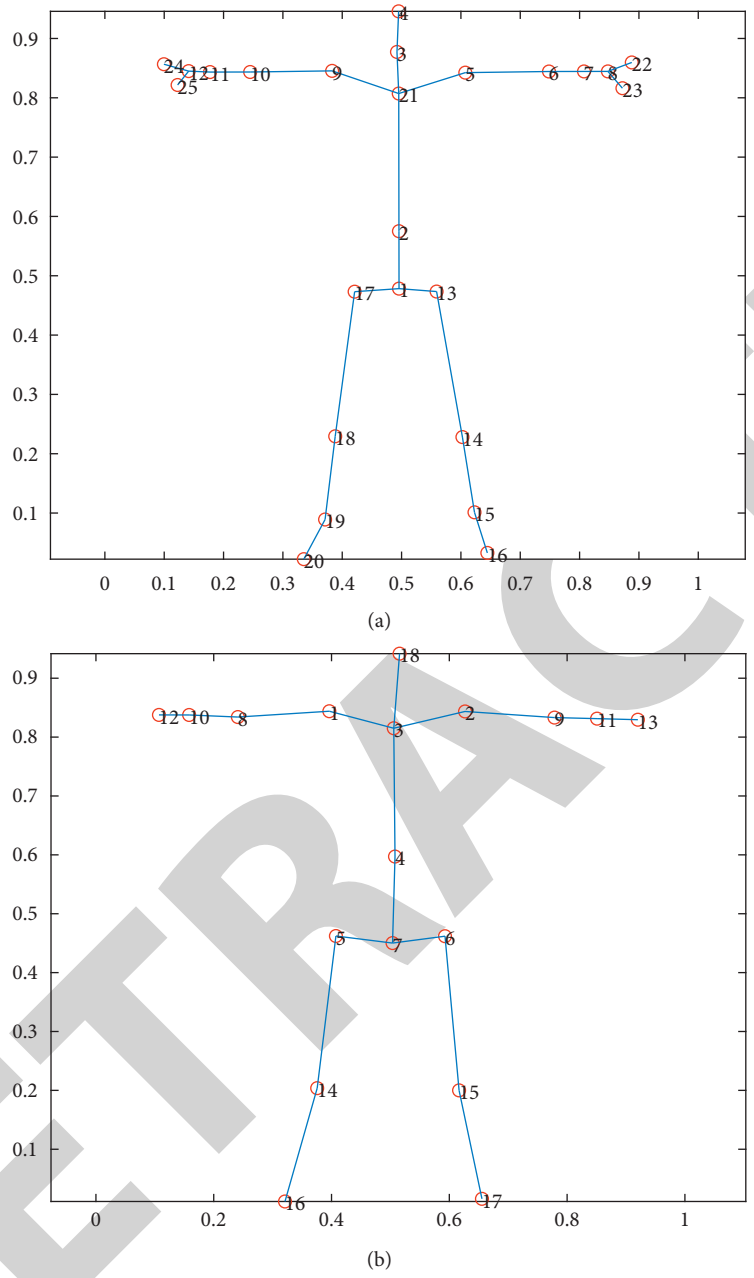

Figure 7: Schematic diagram of the joints of the two data sets. (a) NTU-RGB + D. (b) MSR Action 3D.

The category score vector is noted as $y(x) = (q_1, q_2, \ldots, q_c)$, and the action category label of sample x is

$$Y(x) = \arg\max_j q_j. \tag{10}$$

4. Experiment

4.1. Experimental Data Set. In this paper, to validate the performance of the used action recognition method, two publicly available datasets, NTU-RGB + D and MSR Action 3D, are used. 56880 samples and 60 action categories are available in the NTU-RGB dataset. 320 samples and 16 action categories are available in the MSR Action 3D dataset. There are 320 samples and 16 action categories in the MSR Action 3D dataset. The dataset can be divided into three subsets, two of which are mainly simple actions and the third one is complex actions, and the number of categories in all three subsets is 8. Figure 7 lists the joint diagram of the two datasets.

4.2. Experimental Parameters and Environment. The settings of each parameter of the method in this paper during the experiments are shown in Table 1. The environment used in the experimental process is described in terms of both hardware and software. The hardware environment is as

Table 1: Experimental parameter settings.

Parameters	Value
Number of clips D	30
Image frame size	$280 * 280$
Convolutional layer	5
Pooling layer	3
Inception module	3
Hidden layer dimension	2048
Dropout	0.3
Initial learning rate	0.001
Batch size	32
Weight decay factor	0.00001

Table 2: Accuracy of action recognition with different fusion methods.

Data set\fusion method	Mean fusion	Confidence fusion
NTU-RGB+D	0.8524 ± 0.0346	0.8683 ± 0.02154
MSR action 3D	0.9262 ± 0.0653	0.9529 ± 0.03447

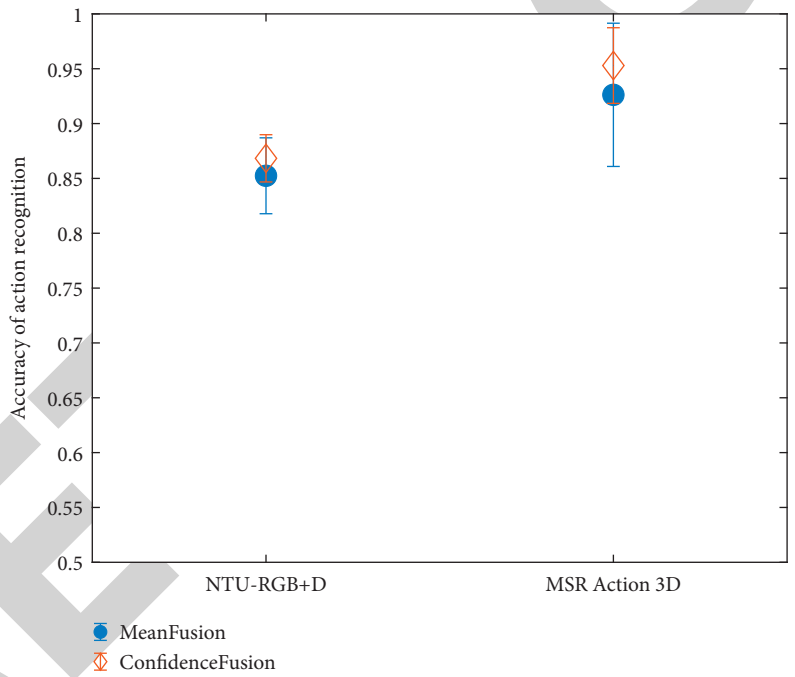

Figure 8: Accuracy of action recognition with different fusion methods.

follows: the computer processor is i7, the GPU used is NVIDIA 3090, the memory is 32G, and the SSD is 1T. The software environment is as follows: 64-bit Windows 10 stable version, Mini-Conda, TensorFlow 1.8.

4.3. Experimental Discussion. To compare the performance of each fusion method, mean fusion and confidence fusion are used for comparison in this paper. The results based on the action recognition method used in this paper using the above two result fusion methods on two datasets are shown in Table 2 and Figure 8. During the experiments, the training samples were 75% of the total samples and the test samples were 25% of the total samples. The accuracy rate was used to assess the algorithm's effectiveness in action recognition. Each method was run 20 times to take the average value to get the experimental results.

The experimental results in Table 2 and Figure 8 show the different accuracy rates of this same method on the two datasets. The recognition based on the MSR Action 3D dataset is significantly better than that based on the NTU-RGB + D dataset. This indicates that the method used is more suitable for a dataset of the type MSR Action 3D. Furthermore, the experimental results obtained by the confidence-based fusion method are significantly better than those obtained by the mean fusion method on both datasets.

Table 3: Accuracy of action recognition obtained by different methods.

Dataset\Method	RNN	LSTM	Reference [28]	Reference [29]	Reference [30]	Proposed
NTU-RGB+D	0.8284 ± 0.0234	0.8303 ± 0.0132	0.8389 ± 0.0207	0.8521 ± 0.0265	0.8612 ± 0.0201	0.8683 ± 0.0215
MSR action 3D	0.9297 ± 0.0354	0.9265 ± 0.0268	0.9432 ± 0.0193	0.9542 ± 0.0213	0.9460 ± 0.0321	0.9529 ± 0.0344

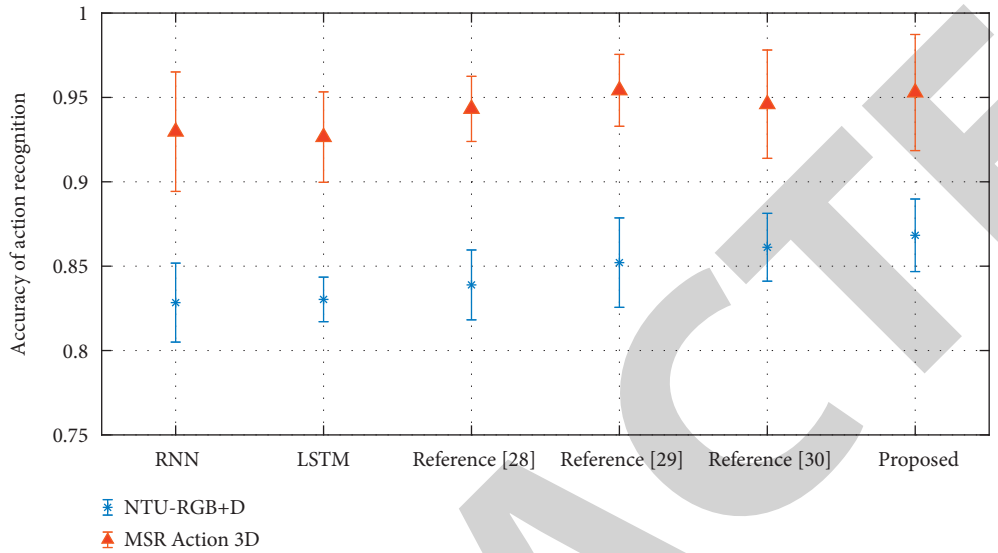

Figure 9: Comparison of the accuracy of action recognition obtained by different methods.

For the NTU-RGB + D dataset, the confidence-based recognition accuracy is improved by 1.87% compared to the mean-based fusion. For the MSR Action 3D dataset, the confidence-based recognition accuracy is improved by 2.88% compared to the mean-based fusion. Moreover, the confidence-based variance is smaller than the variance of the mean-based fusion on both datasets, which indicates that the confidence-based approach is more stable. The experimental data and the analysis conclusions can conclude that the HAR accuracy obtained by the confidence-based fusion approach is higher and more stable.

To further validate the superiority of the methods in this paper over other methods, the comparison methods chosen in this paper are mainly RNN [26], LSTM [27], graph CNN based reference [28], dual-stream CNN based reference [29], and two-dimensional graph convolution-based reference [30]. The accuracy of action recognition obtained by various methods on two publicly available datasets is shown in Table 3 and Figure 9.

From the experimental results shown in Table 3 and Figure 9, it can be seen that, overall, the experimental results obtained based on the MSR Action 3D dataset exceed 0.9 with good results, no matter which method they are based on. This indicates that excellent action recognition results require not only excellent recognition algorithms but also data sets carrying rich information. For the NTU-RGB + D dataset, the accuracy obtained for traditional unimodal algorithms such as RNN, LSTM, and graphical CNNs is significantly lower than that of multimodal correlation algorithms. Therefore, the recognition accuracies obtained by [29], [30], and the proposed method in the table are greater than the previous three methods. Among the multimodal recognition algorithms, the proposed method obtains the highest recognition rate, which proves the superiority of the method in this paper. For the MSR Action 3D dataset, the unimodal algorithms RNN, LSTM, and graph CNN, the accuracy obtained is significantly lower than the multimodal correlation algorithm. The recognition accuracy obtained by [29], [30], and the proposed method in the table is greater than the previous three methods. However, for the MSR Action 3D dataset, the recognition performance of [29] is better than that of [30], and the recognition rate obtained by the proposed method is still the highest. This further proves the superiority of the method in this paper. The experimental results on both datasets show the effectiveness of the method in this paper because the multimodal features based on spatial flow and motion flow combined with the confidence classification result fusion mechanism can effectively improve the action recognition accuracy of the method.

5. Conclusion

With the development of artificial intelligence technology, robots have been widely used in many fields such as industry, agriculture, and the service industry. Robots are often used to perform repetitive, long-duration, heavy, or dangerous tasks that humans cannot perform, freeing humans to focus on dynamic planning or tasks that require flexibility and toughness. Traditional robots have limitations in assisting people in their work, as they are unable to assist humans without spatial or temporal barriers. In order to achieve a more intelligent human-robot interaction, more suitable

ways for humans to manipulate machines need to be urgently explored. Therefore, the research of robotics based on HAR has emerged. To improve the efficiency of human-robot interaction, the accuracy and real time of machine recognition of people's actions become critical. Considering that multiple features can comprehensively characterize data information, to improve the accuracy of action recognition, this study separates the dataset's spatial flow features and motion flow features, inputs each feature separately into BiLSTM, and uses the confidence fusion method to obtain the final action recognition results. The experimental results on two publicly available datasets show that the method in this paper has good recognition results. The datasets recognized by the method in this paper are all fixed-bit acquisition datasets. It ignores the fact that many vision devices are currently mounted on mobile robots. The camera is moving while the human body is moving, so the platform is subsequently improved accordingly. In addition, multi-view and more modal features can be considered for action recognition, thus improving recognition accuracy.

Data Availability

The labeled dataset used to support the findings of this study is available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This work was supported by the Hebei Vocational University of Technology and Engineering.

References

[1] M. Kraus, N. Wagner, Z. Callejas, and W. Minker, "The role of trust in proactive conversational assistants," *IEEE Access*, vol. 9, pp. 112821–112836, 2021.

[2] J. Pustejovsky and N. Krishnaswamy, "Situated meaning in multimodal dialogue: human-robot and human-computer interactions," *Traitement Automatique des Langues*, vol. 61, no. 3, pp. 17–41, 2020.

[3] M. Jarosz, P. Nawrocki, B. Śnieżyński, and B. Indurkhya, "MULTI-PLATFORM intelligent system for multimodal human-computer interaction," *Computing and Informatics*, vol. 40, no. 1, pp. 83–103, 2021.

[4] G. L. Sravanthi, M. V. Devi, K. S. Sandeep, A. Naresh, and A. P. Gopi, "An efficient classifier using machine learning technique for individual action identification," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 6, pp. 513–520, 2020.

[5] N. A. Othman and I. Aydin, "Challenges and limitations in human action recognition on unmanned aerial vehicles: a comprehensive survey," *Traitement du Signal*, vol. 38, no. 5, pp. 1403–1411, 2021.

[6] P. Gupta, A. Thatipelli, A. Aggarwal et al., "Quo vadis, skeleton action recognition?" *International Journal of Computer Vision*, vol. 129, no. 7, pp. 2097–2112, 2021.

[7] M. La Cascia, "3D skeleton-based human action classification: a survey," *Pattern Recognition*, vol. 53, no. 53, pp. 130–147, 2016.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neuralnetworks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.

[9] J. D. Bodapati, N. S. Shaik, and V. Naralasetti, "Deep convolution feature aggregation: an application to diabetic retinopathy severity level prediction," *Signal, Image and Video Processing*, vol. 15, no. 5, pp. 923–930, 2021.

[10] M. Vijayan, R. Mohan, and P. Raguraman, "Contextual background modeling using deep convolutional neural network," *Multimedia Tools and Applications*, vol. 79, no. 15-16, pp. 11083–11105, 2020.

[11] R. Sabitha, A. Aruna, S. Karthik, and J. Shanthini, "Enhanced model for fake image detection (EMFID) using convolutional neural networks with histogram and wavelet based feature extractions," *Pattern Recognition Letters*, vol. 152, pp. 195–201, 2021.

[12] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[13] H. Wu, X. Ma, and Y. Li, "Spatiotemporal multimodal learning with 3D CNNs for video action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1250–1261, 2022.

[14] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: r multi-person 2D pose estimation using Part Affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172–186, 2021.

[15] P. Pareek and A. Thakkar, "A survey on video-based human action recognition: recent updates, datasets, challenges, and applications," *Artificial Intelligence Review*, vol. 54, no. 3, pp. 2259–2322, 2020.

[16] M. Suleman, F. S. Hasan, Q. S. Tahir, S. Arfeen, S. I. Behlim, and B. Mirza, "Generation of sokoban stages using recurrent neural networks," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 3, pp. 466–470, 2017.

[17] L Zelenina, L. Khaimina, Khaimin, and D. Khripunov, "Convolutional neural networks in the task of image classification," *Mathematics and Information*, vol. 65, no. 1, pp. 19–29, 2022.

[18] H. A. Rashwan, M. A. Garcia, S. Abdulwahab, and D. Puig, "Action representation and recognition through temporal co-occurrence of flow fields and convolutional neural networks," *Multimedia Tools and Applications*, vol. 79, no. 45-46, pp. 34141–34158, 2020.

[19] S. Lotfi, M. Mirzarezaee, M. Hosseinzadeh, and V. Seydi, "Detection of rumor conversations in Twitter using graph convolutional networks," *Applied Intelligence*, vol. 51, no. 7, pp. 4774–4787, 2021.

[20] A. Abdelbaky and S. Aly, "Human action recognition using three orthogonal planes with unsupervised deep convolutional neural network," *Multimedia Tools and Applications*, vol. 80, no. 13, pp. 20019–20043, 2021.

[21] U. Zia, W. Khalil, S. Khan, I. Ahmad, and M. N. Khan, "Towards human activity recognition for ubiquitous health care using data from awaist-mounted smartphone," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 28, no. 2, pp. 646–663, 2020.

[22] T. Singh and D. K. Vishwakarma, "A deep multimodal network based on bottleneck layer features fusion for action

recognition," *Multimedia Tools and Applications*, vol. 80, no. 24, pp. 33505–33525, 2021.

[23] N. Yudistira and T. Kurita, "Correlation Net: spatiotemporal multimodal deep learning for action recognition," *Signal Processing: Image Communication*, vol. 82, Article ID 115731, 2020.

[24] J. Imran and B. Raman, "Evaluating fusion of RGB-D and inertial sensors for multimodal human action recognition," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 1, pp. 189–208, 2020.

[25] A. F. M. S. Saif, M. A. S. Khan, A. M. Hadi, R. P. Karmoker, and J. J. Gomes, "Silhouette pose feature-based human action classification using capsule network," *Journal of Information Technology Research*, vol. 14, no. 2, pp. 106–124, 2021.

[26] B. Solongontuya, K. J. Cheoi, and M. Kim, "Novel side pose classification model of stretching gestures using three-layer LSTM," *The Journal of Supercomputing*, vol. 77, no. 9, pp. 10424–10440, 2021.

[27] J. Imran and B. Raman, "Three-stream spatio-temporal attention network for first-person action and interaction recognition," *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, no. 2, pp. 1137–1152, 2021.

[28] B. Francesco and P. Alfredo, "A time based graph deep learning approach to gait recognition," *Pattern Recognition Letters*, vol. 126, pp. 132–138, 2019.

[29] Z. Ding, P. Wang, P. O. Ogunbona, and W. Li, "Investigation of different skeleton features for CNN-based 3Daction recognition," in *Proceedings of the IEEE International Conference on Multimedia & Expo Workshops*, pp. 617–622, IEEE Press, Hong Kong, China, May 2017.

[30] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based actionrecognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 956–965, AAAI Press, New Orleans, LA, U.S.A, February 2018.