Computer Science and Information Systems 19(3):1427–1444 https://doi.org/10.2298/CSIS220322041H

DRN-SEAM: A Deep Residual Network Based on Squeeze-and-Excitation Attention Mechanism for Motion Recognition in Education

Xinxiang Hua

College of Marxism, Zhengzhou University of Science and Technology
Zhengzhou, 450015 China
zxcvfdsa5024@foxmail.com

Abstract. In order to solve the shortcomings of the traditional motion recognition methods and obtain better motion recognition effect in education, this paper proposes a residual network based on Squeeze-and-Excitation attention mechanism. Deep residual network is widely used in various fields due to the high recognition accuracy. In this paper, the convolution layer, adjustment batch normalization layer and activation function layer in the deep residual network model are modified. Squeeze-and-Excitation (SE) attention mechanism is introduced to adjust the structure of network convolution kernel. This operation enhances the feature extraction ability of the new network model. Finally, the expansibility experiments are conducted on WISDM(Wireless Sensor Data Mining), and UCI(UC Irvine) data sets. In terms of F1, the value exceeds 90%. The results show that the proposed model is more accurate than other state-of-the-art posture recognition models. The proposed method can obtain the ideal motion recognition results.

Keywords: motion recognition; Deep residual network; Squeeze-and-Excitation; attention mechanism, education.

1. Introduction

A large number of education videos have been collected in the process of education training and teaching. Accurate recognition of education motions in the videos can prevent accidental injuries and protect the health of students. Therefore, it is of great significance to construct an excellent motion recognition method [1-3].

At present, a variety of portable devices have been developed rapidly, such as smart bracelets and smart-phones, etc,. These emerging adaptive mobile applications can use the collected big data by embedded sensors to conduct motion recognition and behavior analysis. For example, the medical system can use motion recognition to effectively monitor the movement behavior. This not only guides medical staffs to perform the correct treatment, but also solves the shortage of hospital staff. In rehabilitation training, motion recognition and behavior analysis can assist patients in rehabilitation activities, analyze patients' movements and behaviors, and monitor the elderly for ensuring safety. In education, behavior recognition technology as an auxiliary means can be used to analyze various data of students. Effective data analysis improves the score of students, thus improving the overall competitive level [4-6].

Related researches have been striving to achieve the following two goals: enhancing the recognition accuracy and reducing the reliance on engineering features. However, they are difficult to achieve, because the difficulty of motion recognition lies in the great variety of specific motion patterns. In other words, the motion trajectory pattern of different individuals completing the same action is not exactly the same, so it is difficult for engineering features to fully and correctly express the motion trajectory pattern in each action process resulting in the unsatisfactory recognition results.

In this paper, the convolution layer, adjustment batch normalization layer and activation function layer are improved to built a new deep residual network model. Meanwhile, Squeeze-and-Excitation (SE) attention mechanism is introduced to adjust the structure of network convolution kernel. The structure of this paper is as follows. Section 2 introduces the related works. In section 3, the proposed motion recognition method is displayed. Experiments and analysis are shown in section 4. There is a conclusion in section 5.

2. Related Works

This section briefly overviews previous studies on posture recognition based on deep learning, then introduces CNN applications in image classification. In order to achieve the above two goals, some researchers propose the feature extraction method based on deep learning.

Deep learning is organized hierarchically with the latter layer processing the output of the former layer [7,8]. It is a neural network that uses multiple nonlinear information processing layers to extract and classify features. Two classical network structures commonly used in deep learning are convolutional neural network (CNN) and recurrent neural network (RNN). CNN is a deep neural network (DNN) with feature extraction capability. It stacks several convolution operations to create a feature map that becomes progressively more abstract. It automatically extracts the valid features from raw data, and no longer depends on prior knowledge. This method can not only enhance the accuracy and generalization ability of the model, but also form an end-to-end model. It reduces the complexity of the model training. RNN can process serial information [9-11]. The LSTM (long short term memory) [12] and the other extended forms of RNN contain a memory module which can simulate the time dependence in a time series, this way can better handle the time sequence dependence information.

Motion recognition can be generally divided into the following two main steps. The first step is the segmentation of time series. The current mainstream method is to use a fixed length sliding window to segment the entire motion time series into equal segments [13-15]. For example, most studies adopt the time series segmentation method on WISDM data set, that is, fixed window length and fixed moving direction. The second step is to extract the valid features from the obtained original segment. Feature extraction is the most critical part in the whole project, which will directly affect the overall recognition accuracy of the model. In the previous recognition algorithms, engineering features are often used [16-19]. Although this method can also show excellent performance, it requires rich domain knowledge. It can also be time-consuming for domain experts to find good engineering features. The common engineering features in relevant studies include spectral entropy, autoregression coefficient and fast Fourier transform coefficient.

The classical solutions are template matching method, hidden Markov model and support vector machine, etc,.

Many researchers combine the deep learning method with motion classification. Qamar [20] trained deformation postures by combining R-FCN with HyperNet network. Ahmad [21] designed a new convolutional neural network, which could effectively complete the classification of motion images. Lin [22] proposed a novel matching R-CNN framework based on mask R-CNN to prefect motion detection, posture estimation, segmentation and retrieval. Yan [23] adopted a sparse algorithm to extract the spatial and temporal features of sports motion, and then used the neural network to establish sports motion recognition model. Wen [24] extracted the energy diagram and motion descriptor of sports movements, and established the sports movement recognition model by using the support vector machine. The above researches all use a deep convolutional neural network to recognize and classify sports images. In order to improve the recognition accuracy, the convolutional layer number in CNN is usually modified to improve the recognition and classification performance of the model.

However, the deep convolutional neural network still has some problems:

1. with the deepening of the deep learning network, the stacked effect of the network is not good;
2. If the network is more complex, it will result in some problems such as gradient dispersion or gradient explosion in the training process.

This paper draws on the advantages of ResNet in solving the gradient dispersion problem in deep network training and proposes a new deep residual network to improve the performance of action recognition and classification. The deep residual network is composed of residual blocks as shown in figure 1.

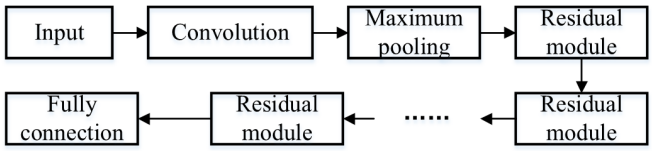

Fig. 1. The deep residual network structure

Each residual block can be expressed as:

$$y_i = h(x_i) + F(x_i, w_i). \tag{1}$$

$$x_{i+1} = f(y_i). \tag{2}$$

Where F is the residual function, f is the ReLU function. W_i is the weight matrix. x_i and y_i are the input and output of the $i - th$ layer, respectively.

The definition of residual function F is:

$$F(x_i, w_i) = w_i \cdot \sigma(B(w_i') \cdot (B(x_i))). \tag{3}$$

$B(x_i)$ is batch normalization. $(\cdot)$ denotes convolution. $\sigma(x) = max(x, 0)$.

The basic idea of residual learning is a branch of the gradient propagation path. For CNN, this idea is first introduced into the Inception model with a parallel form. Residual networks share some similarities with Highway Network. It is connected through residual blocks and shortcuts. The gradient loss problem associated with increasing layers in ResNet is mitigated. However, the output of each path in the Highway Network is controlled by the gate function learned in the training stage.

Unlike the convolutional layer in traditional CNN, the residual units in ResNet are not stacked. Instead, there are some shortcut connections from the input to the output in each convolutional layer. Using identity mapping as shortcut connection reduces the complexity of residual networks and makes deep networks be quickly trained. A residual network can be thought of as a collection of many paths, rather than a very deep architecture. However, all network paths in the residual network have different lengths. There is only one path through all the residual units. In addition, all of these signal paths do not propagate gradients, which is why the residual network is optimized and trained faster. With the network layer increasing, the accuracy rate does not decrease. The structure of the residual block is shown in figure 2.

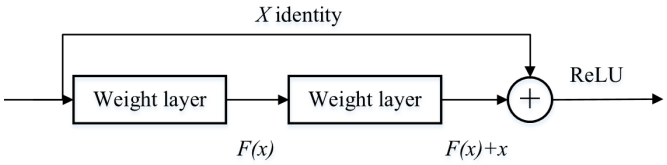

Fig. 2. Residual block structure

Most state-of-the-art methods for action recognition rely on a two-stream architecture that processes appearance and action independently. Aiming at the action features in the context of complex surveillance, and color, texture, edge and other features are extracted manually. The feature process is cumbersome and the classification accuracy is low, so this paper proposes a residual network based on Squeeze-and-Excitation attention mechanism. The recognition performance of the model with shallow network layers is not ideal, while the residual block is introduced into the residual network. When the layer of the model is increased, the residual block can solve the degradation problem well. Therefore, we focus on these problems and propose a deep residual network based on Squeeze-and-Excitation attention mechanism for posture recognition in basketball motion.

The rest of this paper is organized as follows. In Section 2, the proposed posture recognition method is introduced. Section 3 elaborates the detailed experiment process. Finally, a summary alongside with the future research direction is provided in Section 4.

3. Proposed DRN-SEAM Model

At present, most researchers select CNN-based methods to extract the features of motion images [25]. However, if there are many sports postures, CNN network layers are

relatively few, which directly affects the feature learning ability of the network. Then, researchers improve the network by increasing the layer number of the deep convolutional neural network such as GoogleNet and ResNet. Generally, if the layer is deeper in the image recognition and classification model, then the model recognition performance will be better [26]. The flow chart of the proposed method in this paper is shown in figure 3.

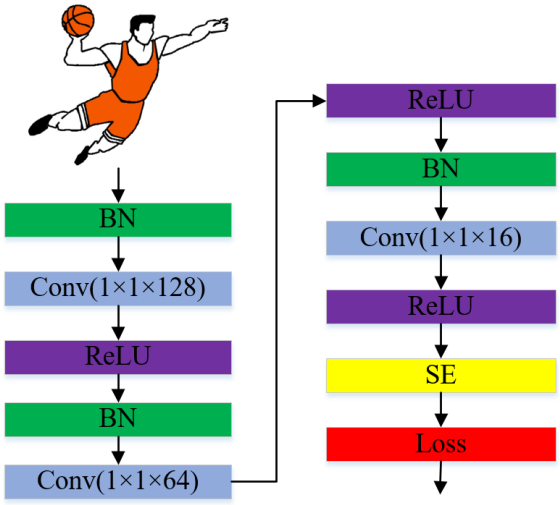

Fig. 3. Flow chart of proposed method

3.1. Modified residual neural network

During the deep convolutional neural network training, the weight of a certain layer changes, and the output feature map of that layer also changes accordingly. The weight of the next layer needs to be relearned, and the network weight of each subsequent layer will be affected. Adding activation function in ResNet can improve the non-linear capability of network model construction. ReLU is used as the activation function in the deep residual network. When $x > 0$, the gradient of ReLU function is always 1, the gradient is not attenuated, which alleviates the problem of gradient dispersion. Assuming that no activation function is used, the network can only be constructed through linear mapping. Even if there are many convolutional layer networks, each layer in the whole network is equivalent, but the convolutional feature map will not change much. So the network should use an activation function.

With the increase of the convolutional neural network layers, the convergence speed of the network will drop sharply and the gradient dispersion will appear in the training process. The Batch Normalization (BN) is an effective solution to this problem, the detailed explanation is shown in reference [27,28]. The specific solution is to normalize the input signals at the same layer. The formula is as follows:

$$\hat{x} = \frac{X - E(x)}{\sqrt{Var(x) + \varepsilon}}. \tag{4}$$

Where, $\hat{x}$ is the activation value of the network normalization. x is the activation value of a certain layer in the network. $E(x)$ is the average value. $Var(x)$ is the variance, and ε is the minimum value. The BN algorithm formula is as follows:

$$y^k = r^k \hat{x}^k + \beta^k. \tag{5}$$

Where, each neuron x^k has the parameters r and β. In this way, when $r^k = \sqrt{Var[x^k]}$, $\beta^k = E[x^k]$, the original learning features in a certain layer can be maintained. The parameters r and β can be reconstructed to restore the initial network learning feature distribution. BN layer is a normalized neural network activation method, it adds batch normalization algorithm to normalize the input signal of each layer, stabilizes its data distribution, and sets a higher learning rate during training to make the network convergence speed and training speed faster. Figure 4 shows the sequence of "convolutional layer+BN layer+ReLU layer" in the traditional residual network.

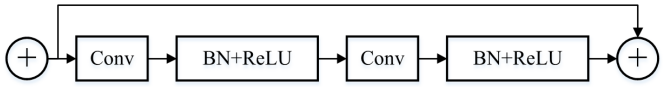

Fig. 4. The sequence of traditional residual blocks

The sequence of traditional residual blocks is defective in deep convolution ResNet, for example, the input of identical blocks is transferred to the deep network from two paths.

The right path indicates that the feature map goes through the convolutional layer and then to BN and ReLU. The input feature map has not been normalized first, so the existence of BN layer is not meaningful. According to the above defects, a new nonlinear branch "BN layer+convolutional layer+ReLU layer "is adopted in this paper to arrange the identical block structure. As shown in figure 5, the network structure is still the same as the traditional residual block network structure (figure 3), while the typical residual block in ResNet is composed of three convolutional layers. In this paper, the new residual block arrangement method not only preserves the identity mapping of the left path, but also maintains the learning ability of the right nonlinear network path.

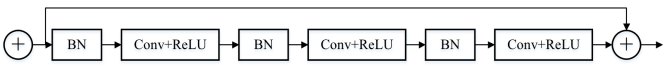

Fig. 5. The sequence of modified residual blocks

3.2. Squeeze-and-Excitation attention mechanism

The attention mechanism is a weighted change for object data that uses top information to guide the bottom-up feed-forward process. The attention model of the human brain is a resource allocation model. At any given moment, attention is always focused on one focal point of the image, while the rest is invisible. In recent years, many attempts have been made to apply attention to deep neural networks. Therefore, the attention mechanism is further located for the discriminative site features. The Squeeze-and-Excitation (SE) algorithm [26] is used for the improved network. The performance of the network model is improved due to the precise modeling of the interaction between the channels for the convolution feature. A mechanism for network models to calibrate features enables networks to selectively enhance valuable feature channels and inhibit useless feature channels from the perspective of global information.

The SE Network modules are shown in figure 6. To ensure the sensitivity of valuable information after adding the network, and make the valuable features be effectively used in the subsequent network layer, accurate modeling of the dependency relationship between channels can be achieved, the features are redefined using the Squeeze, Excitation and Reweight:

1. Squeeze(global information embedding). In order to solve the problem of channel dependence, spatial dimension compression features are used. Each two-dimensional eigenvector is represented by a variable with global spatial information, and the output dimension matches the input channel number. The formula is expressed as follows:

$$Z_c = F_{sq}(U_c) = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} u_c(i,j). \tag{6}$$

2. Excitation (The adaptive recalibration). In order to utilize the information gathered in the compression operation, the dependency of the channel is captured comprehensively. The nonlinear interaction and non-exclusive relationship between learning channels must be satisfied.

$$s = F_{ex}(z,w) = \sigma(g(z,w)) = \sigma(w_2\delta(W_1 z)). \tag{7}$$

3. Reweight. The weights generated after the Excitation are multiplied by the original features.

$$\tilde{X} = F_{scale}(u_c, S_c) = S_c \cdot U_c. \tag{8}$$

After the structures "BN layer+convolutional layer+ReLU layer", the SE algorithm is introduced to the new residual block. The dynamic features of the network are recalibrated to improve the performance of the network, and the soft attention mechanism is successfully applied to the deep network. Embedding the SE module into the new residual module is as shown in figure 7. Table 1 shows the main structure of the original ResNet50 network and the modified structure.

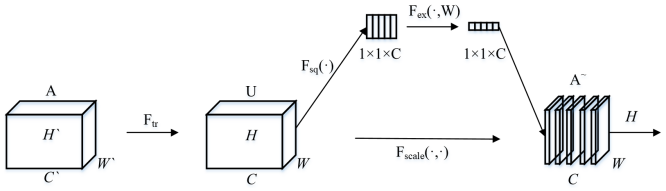

Fig. 6. SE model

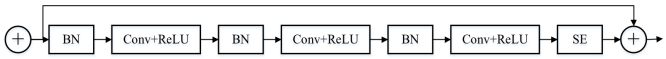

Fig. 7. The residual block with SE module

4. Experiment and Analysis

To verify the performance of the proposed model (DRN-SEAM), we conduct experiments on UCI, WISDM data set and real basketball motion images. Experimental environment is GPU GTX1080, Memory 16 GB, Windows 10 system, MATLAB7a, Tensroflow. It uses stochastic gradient descent (SGD) to optimize the network parameters. The loss function adopts mutual entropy loss. In order to improve the efficiency, the model training is divided into 200 mini-batches. The learning rate is set as 0.0001. All models are trained with 1000 epochs.

4.1. Performance evaluation indexes

In this paper, Precision (P), Recall (R), F1 Measure, average accuracy are used to evaluate the performance.

$$Precision(i) = T(i)/(T(i) + F(i)). \tag{9}$$

$$Recall(i) = T(i)/D(i). \tag{10}$$

$T(i)$ represents the number of i-th correct recognized motions. $F(i)$ represents the number of i-th incorrect recognized motions. $D(i)$ represents the number of i-th motion samples. F value is the weighted average between accuracy and recall, as shown in equation (11).

Table 1. Comparison of two network convolution structures

ResNet50	New network structure
$[1 \times 1, 64; 3 \times 3, 64; 1 \times 1, 256] \times 3$	$[1 \times 1, 128; 3 \times 3, 128; 1 \times 1, 128] \times 3$
$[1 \times 1, 128; 1 \times 1, 128; 1 \times 1, 512] \times 4$	$[1 \times 1, 256; 1 \times 1, 256; 1 \times 1, 256] \times 4$
$[1 \times 1, 256; 1 \times 1, 256; 1 \times 1, 1024] \times 6$	$[1 \times 1, 512; 1 \times 1, 512; 1 \times 1, 512] \times 6$
$[1 \times 1, 512; 1 \times 1, 512; 1 \times 1, 2048] \times 3$	$[1 \times 1, 1024; 1 \times 1, 1024; 1 \times 1, 1024] \times 3$

$$F = \frac{(\vartheta^2 + 1) \cdot P \cdot R}{\vartheta^2 (P + R)}. \tag{11}$$

When $\vartheta = 1$, $F = F1$, namely,

$$F1 = \frac{2 \cdot P \cdot R}{P + R}. \tag{12}$$

Average accuracy=correct recognized motions/the total number of samples.

4.2. Datasets description and the compared methods

The experiment datasets in the experiment are WISDM, UCI and real basketball data sets. The WISDM data contains Walking, Jogging, Upstairs, Downstairs, Sitting, Standing [30]. We only select 4526 samples in this datasets, and randomly select 80% of the dataset for training, and retained 20% as the test set. The UCI dataset includes a group of 30 volunteers within an age bracket of 19-48 years. Each person performed six activities (Walking, Walking_upstairs, Walking_downstairs, Sitting, Standing, Laying). The obtained dataset has been randomly partitioned into two sets, where 80% of the volunteers was selected for generating the training data and 20% the test data [31]. The real basketball data is composed of ten football students, each student conducts 4 actions (Jumping, Shooting, Falling back, Defending) with a total of 40 samples. Then, the data sets are enhanced through rotation, translation, scaling. Finally, we obtain a total of 1000 samples. They are collected by professional PE student belonging to in-house data with mobile phone. Similarly, we randomly select 80% of the dataset for training, and retained 20% as the test set. Some sample images from UCI and real basketball data are shown in figure 8 and figure 9.

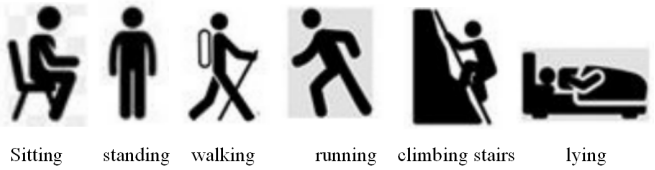

Fig. 8. The sample images in dataset

We compare the DRN-SEAM model with one classical algorithm ResNet50 and other three state-of-the-art motion recognition methods including LTPCNN [32], TFDF [33], DEAPP [34].

LTPCNN: The main idea of the method is the action mapping image classification via convolutional neural network (CNN) based approach. Firstly, we project the raw frames onto three orthogonal Cartesian planes and stack the results into three still images (corresponding to the front, side, and top views) to form the Depth Motion Maps (DMMs). Secondly, Local Ternary Pattern (LTP) is introduced as an image filter for DMMs, thus to

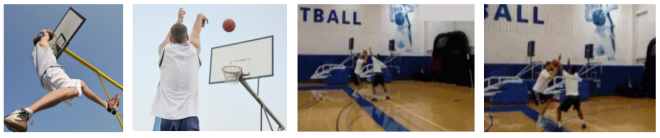

Fig. 9. The sample images in basketball data

improve the distinguishability of similar actions. Finally, we apply CNN to action recognition by classifying corresponding LTP-encoded images.

TFDF: It proposed a shoulder motion recognition optimization method based on the maximizing mutual information from multiclass CSP selected spatial feature channels and wavelet packet features extraction.

DEAPP: It proposed a novel edge-aware end-to-end deep network method, which used the edge-aware pooling module to improve contour accuracy and captured video sequences using multi-scale pyramid pooling layer spatial-time context feature.

We use the related code to transform the non-uniform dataset as matrix format to train and test this model.

4.3. Average accuracy performance test

Average accuracy testing will be conducted on three data sets in this experiment. We test all the categories and get the test results. Assume $a_i(1 \leq i \leq k)$ is the i-th accuracy rate, the average accuracy is $acc_{average} = \frac{a_1+\cdots+a_k}{k}$, k is the action class. The experimental results are shown in tables 2-4. It can be seen that the average accuracy of the DRN-SEAM recognition model in this paper is higher than other methods on the three different data sets.

In the WISDM data set, DEAPP has the best average accuracy in relevant studies due to the use of CNN. And it also adds the additional mathematical statistics features by manual extraction in the dense layer. However, the accuracy of the DRN-SEAM model is higher than that of the DEAPP in the absence of any artificial features. The recognition rate of the proposed DRN-SEAM is higher than that of other methods. The experimental results also show that the feature extraction ability of the Squeeze-and–Excitation structure is higher than the normal CNN and deep separable convolution.

Table 2. Results with different methods on WISDM

Method	Average accuracy (%)
ResNet50	86.07
LTPCNN	86.78
TFDF	86.57
DEAPP	96.52
DRN-SEAM	98.49

Table 3. Results with different methods on UCI

Method	Average accuracy (%)
ResNet50	77.25
LTPCNN	78.82
TFDF	84.53
DEAPP	96.35
DRN-SEAM	98.67

Table 4. Results with different methods on basketball data

Method	Average accuracy (%)
ResNet50	74.86
LTPCNN	76.47
TFDF	82.17
DEAPP	86.54
DRN-SEAM	91.37

Table 5, table 6 and table 7 display the detailed classification results of each action on WISDM, UCI and basketball data, and compare them with other methods. It can be seen from the two tables, jogging, walking, and standing are the easiest to recognize on WISDM. The accuracy of all the methods is more than 90%. Since the changes in the three actions have the biggest difference than other actions. Downstairs and downstairs are difficult to recognize. Because the two actions are the easiest to confuse. However, the recognition accuracy of the DRN-SEAM is bigger than 90%. Figures 10,11,12 are the statistical analysis for these data sets, which also shows that the proposed method has better result.

Table 5. Classification accuracy on WISDM with different methods (%)

Motion type	ResNet50	LTPCNN	TFDF	DEAPP	DRN-SEAM
downstairs	63.48	51.64	87.35	80.12	92.16
jogging	92.54	94.83	97.98	98.25	99.26
sitting	82.96	83.58	83.74	98.61	98.76
standing	93.87	95.87	93.45	92.72	98.87
upstairs	71.55	66.78	72.34	84.38	91.85
walking	83.79	84.65	98.61	97.81	99.32

In order to observe the performance of DRN-SEAM model in a more detailed way, the accuracy rate, recall rate and F1 value of the six motions are counted respectively as shown in table 8, table 9 and table 10.

By recording the accuracy of each network on the test set after each epoch training, the accuracy is obtained as shown in figure 13 and figure 14.

It can be seen that the accuracy of DRN-SEAM and DEAPP during the training process has obvious advantages over the other two networks after 200 epochs. This indicates that the feature map extracted by the DRN-SEAM feature extraction module can be accu-

Table 6. Classification accuracy on UCI with different methods (%)

Motion type	ResNet50	LTPCNN	TFDF	DEAPP	DRN-SEAM
downstairs	97.45	97.65	98.24	98.87	99.86
jogging	93.77	97.88	98.67	100.00	100.00
sitting	91.85	92.36	94.87	95.87	99.87
standing	98.36	98.67	98.96	99.24	99.98
upstairs	98.66	98.67	98.56	99.41	100.00
walking	99.45	99.67	99.85	100.00	100.00

Table 7. Classification accuracy on basketball with different methods (%)

Motion type	ResNet50	LTPCNN	TFDF	DEAPP	DRN-SEAM
Jumping	81.32	87.32	88.34	90.23	92.54
shooting	82.46	88.47	89.75	89.67	93.57
Falling back	83.64	85.63	84.16	91.22	91.83

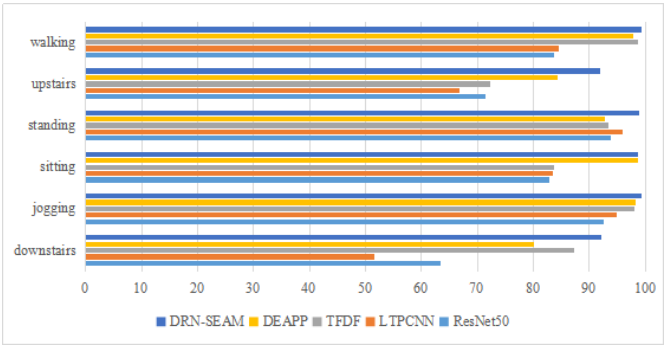

Fig. 10. Bar chart analysis for WISDM data set

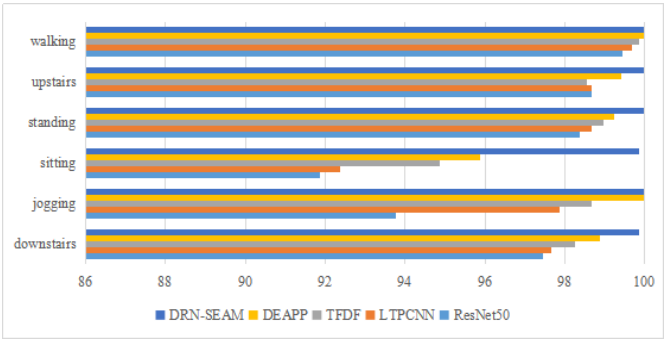

Fig. 11. Bar chart analysis for UCI data set

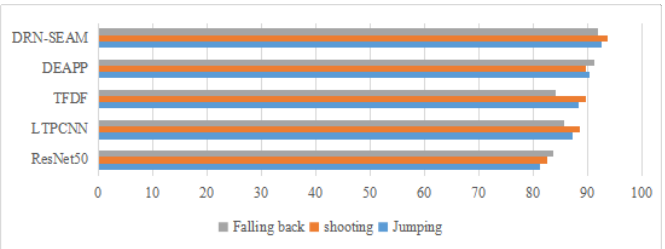

Fig. 12. Bar chart analysis for basketball data set

Table 8. Precision, recall and F1-score of action recognition on WISDM (%)

Method	Precision	Recall	F1
ResNet50	89.64	87.25	88.41
LTPCNN	96.35	94.65	96.85
TFDF	96.87	95.87	96.99
DEAPP	97.65	96.84	97.82
DRN-SEAM	98.26	97.98	98.96

Table 9. Precision, recall and F1-score of action recognition on UCI (%)

Method	Precision	Recall	F1
ResNet50	91.65	86.57	89.77
LTPCNN	96.87	97.21	98.24
TFDF	97.25	98.25	98.65
DEAPP	98.13	98.67	99.12
DRN-SEAM	99.69	99.54	99.57

Table 10. Precision, recall and F1-score of action recognition on basketball (%)

Method	Precision	Recall	F1
ResNet50	82.54	76.21	77.59
LTPCNN	89.32	82.14	83.65
TFDF	90.54	83.46	85.28
DEAPP	91.23	84.55	86.95
DRN-SEAM	92.87	91.32	90.77

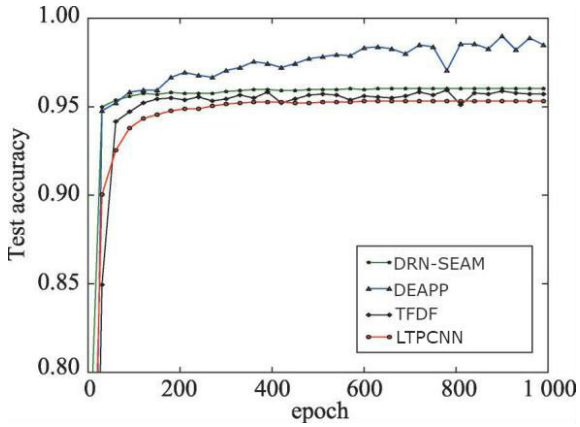

Fig. 13. Accuracy curve on UCI

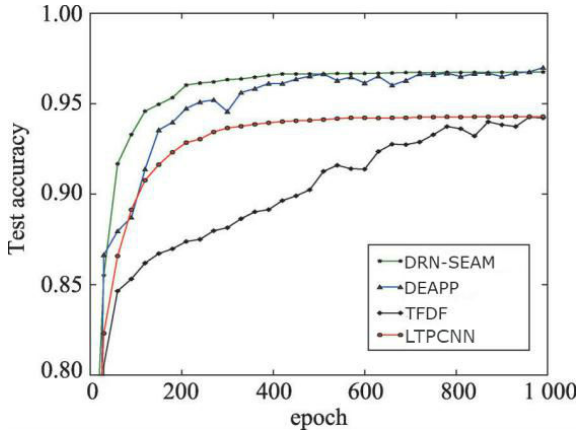

Fig. 14. Accuracy curve on WISDM

rately classified after 200 epochs. In this paper, the precision-recall (PR) curve is used to identify the accuracy index to evaluate the performance of action recognition algorithms. First, the PR value is calculated by predicting the contour processing of the action, and then the PR curve is drawn. Figure 14 is the basketball action PR curve. As shown in figure 15, the recognition PR curve on the basketball action dataset with proposed method can achieve better performance.

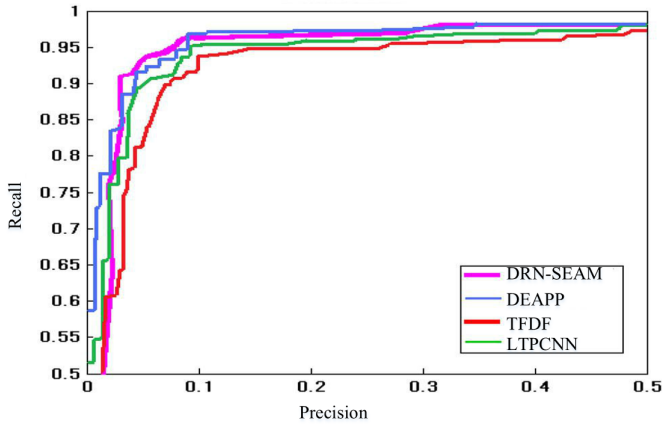

Fig. 15. PR curve on the basketball action dataset

4.4. Time consuming test

This experiment verifies the time-consuming advantages of the proposed DRN-SEAM structure in this paper. The experiment counts the time when the four networks are trained for 500 iterations simultaneously on the UCI, WISDM and basketball data set. The advantages of the DRN-SEAM are obtained by the time comparison. The experimental results are shown in tables 11-13. It can be seen from table 11 that the real-time performance of DRN-SEAM network has great advantages than other three networks. It shows that the proposed method has been greatly improved in efficiency as well as accuracy.

Table 11. Time consuming on UCI/s

Method	Epoch=100	Epoch=200	Epoch=300	Epoch=400	Epoch=500
DRN-SEAM	85	90	95	97	100
DEAPP	120	140	160	180	220
TFDF	150	180	210	240	270
LTPCNN	500	900	1300	1700	2100

Table 12. Time consuming on WISDM/s

Method	Epoch=100	Epoch=200	Epoch=300	Epoch=400	Epoch=500
DRN-SEAM	90	100	110	120	130
DEAPP	140	160	180	200	220
TFDF	160	200	240	280	320
LTPCNN	450	850	1250	1650	2050

Table 13. Time consuming on basketball/s

Method	Epoch=100	Epoch=200	Epoch=300	Epoch=400	Epoch=500
DRN-SEAM	74	84	89	97	105
DEAPP	120	145	170	195	215
TFDF	155	195	235	275	315
LTPCNN	480	891	1400	1871	2390

5. Conclusions

Action recognition is an important research direction in computer vision, which has worldwide applications, such as video surveillance, human-robot interaction and so on. Due to the influence of complex background and multi-angle changes, accurate recognition and analysis of human action in real-life scenarios is still a challenging problem. In order to improve the accuracy of action detection and recognition, this paper modifies the residual network by improving the order of "BN+ReLU+convolutional layer" in the residual block. And we introduce the attention mechanism and adjust the structure of the network convolution kernel to improve the recognition and classification effect of the model. The experimental results show that the proposed network model is better than the traditional deep residual network in terms of classification accuracy and convergence speed. In this study, we confine our technique to static images of human action.

The difficulty of action recognition lies in the huge changes in the specific actions. So the models often have poor accuracy. It performs better on one data set, but it is inferior to other models on the other data sets. Meanwhile, error data exists in the real-time situation of mobile phone sensor, this requires that the model needs to have high accuracy, adaptability, robustness and better data fault tolerance ability. Next work, RNN-based models will be researched with its better ability of handling the time sequence dependence information by combining the SE attention mechanism. Also, future studies will dynamically test our system on humans action with a robot system in real-world settings.

Availability of data and materials. The data used to support the findings of this study are available from the corresponding author upon request.

Competing interests. The authors declare that they have no conflicts of interest.

References

1. Peng L, Chen Z, Yang L T, et al. "Deep Convolutional Computation Model for Feature Learning on Big Data in Internet of Things," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 2, pp. 790-798, Feb. 2018.

2. Rajini A R, Abishek E, Ramesh S, et al. "Compact Printed Planar Eye Shaped Dipole Antenna for Ultra-Wideband Wireless Applications," *Journal of Applied Science and Engineering*, vol. 25, no. 5, pp. 761-766, 2021.
3. Yeh, J., Tsai, C. "A Graph-based Feature Selection Method for Learning to Rank Using Spectral Clustering for Redundancy Minimization and Biased PageRank for Relevance Analysis," *Computer Science and Information Systems*, Vol. 19, No. 1, pp. 141-164. (2022).
4. Zhong X, Huang W, Luo R, et al. "Video Human Behavior Recognition Based on ISA Deep Network Model," *International Journal of Pattern Recognition and Artificial Intelligence*, 2020. doi: 10.1142/S0218001420560121
5. S. Yin and H. Li. "Hot Region Selection Based on Selective Search and Modified Fuzzy C-Means in Remote Sensing Images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 5862-5871, 2020, doi: 10.1109/JSTARS.2020.3025582
6. Li M, Chen T, Du H. "Human Behavior Recognition Using Range-Velocity-Time Points," *IEEE Access*, vol. 8, pp. 37914-37925, 2020. doi: 10.1109/ACCESS.2020.2975676
7. Mandić, M. "Semantic Web Based Platform for the Harmonization of Teacher Education Curricula," *Computer Science and Information Systems*, Vol. 19, No. 1, pp. 229-250. (2022).
8. L. Jiao and J. Zhao. "A Survey on the New Generation of Deep Learning in Image Processing," *IEEE Access*, vol. 7, pp. 172231-172263, 2019, doi: 10.1109/ACCESS.2019.2956508.
9. Chen W. "A Novel Long Short-Term Memory Network Model For Multimodal Music Emotion Analysis In Affective Computing," *Journal of Applied Science and Engineering*, vol. 26, no. 3, pp. 367-376, 2022.
10. Ding S, Sun Y, An Y, et al. "Multiple birth support vector machine based on recurrent neural networks," *Applied Intelligence*, vol. 50, no. 7, pp. 2280-2292, 2020.
11. R. Jiao, T. Zhang, Y. Jiang and H. He, "Short-Term Non-Residential Load Forecasting Based on Multiple Sequences LSTM Recurrent Neural Network," *IEEE Access*, vol. 6, pp. 59438-59448, 2018.
12. Jiang F, Yuen K K R, Lee E W M. "A long short-term memory-based framework for crash detection on freeways with traffic data of different temporal resolutions," *Accident Analysis & Prevention*, vol. 141:105520, 2020.
13. Ronao C.A., Cho SB. "Deep Convolutional Neural Networks for Human Activity Recognition with Smartphone Sensors," *Neural Information Processing. ICONIP 2015. Lecture Notes in Computer Science*, vol. 9492. Springer, Cham.
14. M. Zeng et al., "Convolutional Neural Networks for human activity recognition using mobile sensors," *6th International Conference on Mobile Computing, Applications and Services*, Austin, TX, pp. 197-205, 2014.
15. Moya Rueda F, Grzeszick, René, Fink G, et al. "Convolutional Neural Networks for Human Activity Recognition Using Body-Worn Sensors," *Informatics*, vol. 5, no. 2, 2018.
16. Jain S, Rustagi A, Saurav S, et al. "Three-dimensional CNN-inspired deep learning architecture for Yoga pose recognition in the real-world environment," *Neural Computing and Applications*, pp. 1-15, 2020.
17. Yadav S K, Singh A, Gupta A, et al. "Real-time Yoga recognition using deep learning," *Neural Computing and Applications*, vol. 31, no. 12, pp. 9349-9361, 2019.
18. Alghyaline S. "Real-time Jordanian license plate recognition using deep learning," *Journal of King Saud University-Computer and Information Sciences*, 2020.
19. Lei, Zhang, Yang, et al. "RFR-DLVT: a hybrid method for real-time face recognition using deep learning and visual tracking," *Enterprise Information Systems*, 2020.
20. Qamar S, Jin H, Zheng R, et al. "3D Hyper-Dense Connected Convolutional Neural Network for Brain Tumor Segmentation," *IEEE, 14th International Conference on Semantics, Knowledge and Grids (SKG)* 2018. IEEE, 2019.
21. A. P. Tafti, F. S. Bashiri, E. LaRose and P. Peissig, "Diagnostic Classification of Lung CT Images Using Deep 3D Multi-Scale Convolutional Neural Network," *2018 IEEE Interna-*

tional Conference on Healthcare Informatics (ICHI), New York, NY, 2018, pp. 412-414, doi: 10.1109/ICHI.2018.00078
22. Lin K, Li C, Zhao H, et al. "Face Detection and Segmentation Based on Improved Mask R-CNN," *Discrete Dynamics in Nature and Society*, 2020. doi: 10.1155/2020/9242917
23. Guoli Yan, Huiyan Wang, et al. "Semantic annotation for complex video street views based on 2D-3D multi-feature fusion and aggregated boosting decision forests," *Pattern Recognition the Journal of the Pattern Recognition Society*, vol. 62, pp. 189-201, 2017.
24. Weng Z, Guan Y. "Action recognition using length-variable edge trajectory and spatio-temporal motion skeleton descriptor," *EURASIP Journal on Image and Video Processing*, vol. 2018, no. 1, 2018.
25. M. Zhou, "Feature Extraction of Human Motion Video Based on Virtual Reality Technology," *IEEE Access*, vol. 8, pp. 155563-155575, 2020, doi: 10.1109/ACCESS.2020.3019233.
26. Jahandad, Suriani Mohd Sam, Kamilia Kamardin, Nilam Nur Amir Sjarif, Norliza Mohamed. "Offline Signature Verification using Deep Learning Convolutional Neural Network (CNN) Architectures GoogLeNet Inception-v1 and Inception-v3," *Procedia Computer Science*, vol. 161, pp. 475-483, 2019.
27. Shoulin Yin, Ye Zhang, Shahid Karim. "Large Scale Remote Sensing Image Segmentation Based on Fuzzy Region Competition and Gaussian Mixture Model," *IEEE Access*, vol. 6, pp. 26069-26080. 2018.
28. loffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv: 1502.03167, 2015.
29. J. Hu, L. Shen, S. Albanie, G. Sun and E. Wu. "Squeeze-and-Excitation Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 2011-2023.
30. Jeffrey W. Lockhart, Gary M. Weiss. "Limitations with Activity Recognition Methodology & Data Sets," *Proceedings of the 2014 ACM Conference on Ubiquitous Computing (UBICOMP) Adjunct Publication (2nd International Workshop on Human Activity Sensing Corpus and its Application)*, Seattle, WA, 2014.
31. Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, Jorge L. Reyes-Ortiz. "Energy Efficient Smartphone-Based Activity Recognition using Fixed-Point Arithmetic," *Journal of Universal Computer Science*, vol. 19, no. 9. May 2013.
32. Z. Li, Z. Zheng, F. Lin, H. Leung, and Q. Li. "Action recognition from depth sequence using depth motion maps-based local ternary patterns and CNN," *Multimedia Tools Appl.*, vol. 78, no. 14, pp. 19587-19601, 2019.
33. Bai D, Chen S, Yang J. "Upper Arm Motion High-Density sEMG Recognition Optimization Based on Spatial and Time-Frequency Domain Features," *Journal of Healthcare Engineering*, 2019, 2019:1-16.
34. Xu L, Yan S, Chen X, et al. "Motion Recognition Algorithm Based on Deep Edge-Aware Pyramid Pooling Network in Human-Computer Interaction," *IEEE Access*, vol. 7, pp. 163806-163813, 2019.

Xinxiang Hua is with College of Marxism, Zhengzhou University of Science and Technology, Zhengzhou, 450015 China. His research interests include image processing and education.

Received: March 22, 2022; Accepted: August 20, 2022.