Research Article

Study of Human Motion Recognition Algorithm Based on Multichannel 3D Convolutional Neural Network

Yang Ju

Department of Physical Education, Tianjin University of Commerce, Beichen, Tianjin 300134, China

Correspondence should be addressed to Yang Ju; juyang@tjcu.edu.cn

Received 9 April 2021; Revised 10 May 2021; Accepted 20 May 2021; Published 27 May 2021

Academic Editor: Zhihan Lv

Copyright © 2021 Yang Ju. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Aiming at the problem that it is difficult to balance the speed and accuracy of human behaviour recognition, this paper proposes a method of motion recognition based on random projection. Firstly, the optical flow picture and Red, Green, Blue (RGB) picture obtained by the Lucas-Kanade algorithm are used. Secondly, the data of optical flow pictures and RGB pictures are compressed based on a random projection matrix of compressed sensing, which effectively reduces power consumption. At the same time, based on random projection compression data, it can effectively find the optimal linear representation to reconstruct training samples and test samples. Thirdly, a multichannel 3D convolutional neural network is proposed, and the multiple information extracted by the network is fused to form an output recognizer. Experimental results show that the algorithm in this paper significantly improves the recognition rate of multicategory actions and effectively reduces the computational complexity and running time of the recognition algorithm.

1. Introduction

Human motion recognition technology based on computer vision is widely used in robotics, video surveillance, virtual reality, and other fields [1, 2]. The methods to solve the problem of human action recognition are mainly divided into traditional algorithms and recognition algorithms based on deep learning [3, 4]. Traditional algorithms use the method of “feature extraction and expression + feature matching” to recognize human behaviour, and recognition algorithms based on deep learning learn object characteristics through neural networks and directly output the final recognition results [5].

Traditional algorithms recognize human behaviour by analysing the inherent characteristics of human behaviour, including motion information features, temporal and spatial interest points, and geometric features. Among them, spatial-temporal interest points have strong robustness to illumination changes, background differences, and environmental noise, feature expression is more adequate, and the recognition rate is the highest. Commonly used time and space points of interest include 3D-SIFT [6], HOG3D

[7], and ESURF [8]. However, this type of feature extraction is complicated and time-consuming to match, and it is difficult to meet real-time requirements in practical applications. In contrast, MHI [9], MEI [10], and so forth only need simple processing methods such as image difference to achieve the extraction of geometric features. However, the conventional way of using this kind of descriptors for feature expression often results in the compression of feature dimensions. Although the speed of feature extraction has been improved, the recognition rate will be greatly reduced due to the loss of temporal and spatial information. In recent years, deep learning methods have emerged, and the characteristics obtained through neural networks have more abstract and comprehensive descriptions of behaviour characteristics [11, 12]. The convolutional neural network can effectively reduce the amount of parameters to be trained due to its shared weight characteristics of the convolutional layer and is the most widely used.

In view of the excellent feature expression capabilities of deep neural networks, this paper uses convolutional neural networks instead of traditional feature descriptors as the feature expression method to ensure the accuracy of the

algorithm, and directly using the original video as the input of the neural network will bring about the problem of algorithm speed reduction. In order to take into account the speed of the algorithm, this paper compresses the data of optical flow pictures and RGB pictures based on the compressed sensing random projection matrix, which effectively reduces the complexity of the model, and the compression method is simple and efficient, ensuring the real-time performance of the algorithm. Input the compressed data into the network, obtain more time dimension features through six-layer 3D convolution and 4-layer 3D pooling, and increase the adaptability of the model through the dropout layer. Then input all the information to the fully connected layer and softmax classifier to obtain the final output result. The experimental verification on the public dataset shows that the accuracy of the algorithm in this paper has been improved.

2. Related Knowledge

With the production and development of various smart devices, in addition to paying more attention to smart machines and equipment, human beings also pay more and more attention to the recognition of human movements. At present, the study of human action behaviour has gradually become one of the important research issues in the field of machine vision.

Literature [13] proposed a detection method based on geometric constraints for the extraction of action targets. The background corner points corresponding to successive frames in the video sequence are used to meet the principle of geometric constraints. This method first extracts the Harris corner points of the previous frame and then uses the pyramid-layered Lucas-Kanade optical flow method to obtain the corresponding points in the next frame. The basic matrix is predicted by random sampling algorithm, and the background Angle and foreground Angle are identified. Finally, the moving target region is identified. This method can complete the detection of multiple moving targets with high precision in a dynamic background, but the number of image frames processed per second is relatively low. Literature [14] proposed a new spatiotemporal interest point detector. This method takes the detected spatiotemporal interest points as the centre and establishes a spatiotemporal gradient descriptor based on a polyhedron model to further characterize the visual characteristics of human actions in time and space. Then use the bag-of-words method to build a more effective codebook for the seen action features. Finally, the feature descriptor is combined with the artificially defined high-level action properties. Implicit support vector machine and coordinate descent method are used to solve the local optimal solution of the final recognition model. This method has a high recognition rate, but the feature description is too complicated.

The good results of deep learning in the field of image recognition in computer vision have made deep learning methods, especially convolutional neural networks, widely researched and applied in the field of computer vision. Compared with static images, video sequences have not only

appearance information but also motion information [15, 16]. Therefore, some recent studies have begun to try to design an action recognition model based on convolutional neural networks that can effectively utilize the appearance and motion information of video sequences. Amin et al. [17] studied and compared three widely used methods in a variety of CNN connection methods, namely, late fusion, early fusion, and slow fusion. The experimental results show that none of these methods can make full use of motion information and can only make moderate improvements to a single frame. The dual-stream convolutional neural network proposed by Liao et al. [18] trains the second CNN stream on the optical flow of the video frame to compensate for the defect that the superimposed RGB stream cannot make full use of time information. Here comes a certain performance gain. Al-Hammadi et al. [19] built a three-dimensional convolutional neural network (3DCNN), took the video sequence as the input of the network, and successfully extracted the spatiotemporal information of the behaviour sequence. Saleem [20] obtained time information and spatial information separately through the two-stream network and then merged them into behaviour feature descriptors. The premise of the above method is that the input video sequence needs to cover the entire behaviour sequence. The increase of the action duration will deepen the network structure and reduce the real-time performance. Improved dense trajectories (IDT) [21] uses dense trajectories algorithm. IDT with higher-dimensional encoding (IDTHDE) [22] is an improvement of the visual bag-of-words model and fusion of multidimensional features to achieve higher-dimensional feature encoding. These two action recognition methods are more traditional methods of manually designing features. Two-stream (TS) [23] recognizes actions by constructing a two-stream spatiotemporal network model, but the network is relatively shallow. Very deep two-stream (DTS) [24] has a very deep network architecture and improves the network in depth. The KVMF algorithm [25] intercepts multiple 3D volumes from the video segment as the input of the network and uses the prediction vector obtained from each volume to represent the category probability of the action.

In recent years, the emerging sparse representation theory has received widespread attention in the fields of machine learning, machine vision, and pattern recognition, providing a theoretical basis for signal classification based on sparse representation [26, 27]. Gan et al. [28] proposed a sparse representation based classifier (SRC) based on the sparse representation theory. The basic idea is to transform the pattern recognition problem into a signal sparse representation problem without considering other types of sample data. Yan et al. [29] proposed a sparse representation classification with random projection (SRC-RP), which tried to perform sparse representation classification on compressed sample data from random projections to reduce the energy cost of sensor nodes to improve the action recognition rate. Díaz et al. [30] proposed a fast sparse representation algorithm based on prototype representation, trying to use the K-SVD algorithm to construct a small overcomplete dictionary that satisfies the sparse representation condition and converts the sparse representation

problem into an l1-norm minimization problem. He et al. [31] proposed a fast sparse representation classification algorithm, which aims to solve the problem of high computational complexity of the traditional SRC algorithm. The basic idea is to use the K-nearest neighbour (KNN) algorithm to find a smaller training sample set used to sparsely represent the test samples, and an attempt is made to improve the recognition rate of the algorithm on the basis of reducing the complexity of the algorithm.

Therefore, this paper combines random projection algorithm and deep neural network and proposes an action recognition algorithm based on random projection. The experimental verification on the public dataset shows that the accuracy of the algorithm in this paper has been improved.

3. Adaptive Spectral Clustering Algorithm

In this paper, firstly, the convolutional neural network is used instead of the traditional feature descriptor to ensure the accuracy of the algorithm. Secondly, the data of optical flow images and RGB images are compressed based on the compressed sensing random projection matrix, which effectively reduces the complexity of the model. Then, the compressed data are input into the network, and more time dimension features are obtained through six layers of 3D convolution and four layers of 3D pooling, and the adaptability of the model is increased through the dropout layer. Finally, all the information is entered into the full connection layer and the SoftMax classifier to get the final output.

3.1. Algorithm Steps. The basic process of behaviour recognition in this paper is shown in Figure 1. First, the random projection matrix is used to reconstruct the sample set used for training to reduce the interference of training samples of irrelevant categories, improve the accuracy of action recognition, and reduce the computational cost of classification at the same time. Then, use the training set to train the improved convolutional neural network in this paper, and get the model with parameters. Finally, use the model to make predictions on the test set.

Step 1: initialize training sample set $x = [x_1, x_2, \dots, x_n]$, test sample y , and random projection matrix m of each sensor node

Step 2: construct training sample set and test sample set

Step 3: normalize each column of vectors in the training sample set

Step 4: send the RGB picture compressed by the random projection matrix and its corresponding optical flow picture information into the 3D network, and obtain the action category

3.2. Random Projection. Random projection (RP) is a data compression method that uses a random projection matrix to reduce the dimensionality of high-dimensional data and has the advantage of low computational complexity [32–34].

Assuming that the human motion signal collected by the sensor node is $s \in Sn$, using a random projection matrix $m \in M$, the original signal s can be projected into a d -dimensional subspace; namely,

$$cm_n = m \times s_n. \quad (1)$$

In the formula, $cm \in Sn$ is the projected compressed data, and its computational complexity is only $O(dn)$.

In the practical application of projection matrix, the selection of appropriate human motion signals is very important [35, 36]. In order to accurately reconstruct and restore the original signal s from the d -dimensional measurement value cm with high probability, the projection matrix m must meet the restricted isometry property (RIP). That is, for any signal s , there is a RIP constant $a \in (0, 1)$, so that the following formula holds:

$$(1 - a) \times s^2 \leq (m's)^2 \leq (1 + a) \times s^2. \quad (2)$$

In the formula, m' is the submatrix formed by the relevant columns indicated by the index in m .

In order to effectively and reasonably use the limited computing resources of the system, a sparse binary matrix that satisfies the properties of RIP is selected as a random projection matrix for data compression, which provides reliable data for the subsequent construction of action recognition and classification algorithms.

The key to compressing data based on random projections is to find as few training samples as possible to optimally reconstruct the test samples linearly. First, on the sensor node i , use the random projection matrix m_i to randomly project the action vector v_{ij} to be recognized. Variable m_i is from the random projection matrix m on the sensor node i .

$$v'_{ij} = m_i \times v_{ij}. \quad (3)$$

In the remote data processing center, the motion vector to be recognized composed of k sensor nodes is

$$v'_{k,ij} = \{v'_{1,ij}, \dots, v'_{k,ij}\}^T. \quad (4)$$

Denote matrix m as a diagonal matrix composed of diagonal elements of random projection matrix m_i ; then formula (4) can be expressed as

$$v'_{ij} = \begin{Bmatrix} m_1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & m_k \end{Bmatrix} \times \begin{Bmatrix} v'_{1,ij} \\ \dots \\ v'_{k,ij} \end{Bmatrix} = m \times v'_{k,ij}. \quad (5)$$

At the same time, using the same random projection matrix to compare the training sample set,

$$mv'_{ij} = mv\beta. \quad (6)$$

If the sparse coefficient β is kept sparse enough, based on the compressed sensing theory, the l1-norm minimization problem can be used to solve the problem. Reconstruct the training sample set for the sparse representation of test samples, reduce the interference of training samples of

FIGURE 1: Flow chart of human motion recognition.

irrelevant categories, improve the accuracy of action recognition, and reduce the computational overhead of sparse representation classification.

3.3. Improved Action Recognition Algorithm of 3D Convolutional Neural Network. The core of the 3D convolutional neural network constructed in this paper uses a dual-stream 3D convolutional neural network, and its structure is shown in Figure 2. The algorithm is divided into three stages: data processing stage, feature extraction stage, and model fusion stage.

The input of each group of networks is divided into two inputs. The spatial stream CNN input is the continuous 16 frames of RGB picture information compressed by the random projection matrix, and the time stream CNN input is the optical flow corresponding to the RGB picture compressed by the random projection matrix. RGB pictures have more static characteristics in the spatial domain, and optical flow pictures have dynamic characteristics in the time domain. The RGB pictures and optical flow pictures are input to the improved convolutional neural network, and the spatial and temporal features are extracted at the same time, and the abstract high-dimensional features are obtained through the multilayer convolutional neural network to train and output the behaviour recognition model. In the test, the output of the 3DCNN SoftMax layer trained separately through the RGB picture, and the optical flow picture is weighted and fused, and the output result after the first model fusion is obtained as shown in the following formula:

$$O = \alpha \frac{I_s}{n} + (1 - \alpha) \frac{I_t}{n}. \quad (7)$$

In formula (7), I is the feature vector, α is a constant greater than 0 and less than 1, and n is the number of video frames.

Repeat the above experiment to obtain two weighted fusion models. When testing, the outputs of the SoftMax layer of the two models are averagely fused to calculate the final probability, as shown in the following equation:

$$p = \frac{n}{N}. \quad (8)$$

In formula (8), p is the final recognition rate, N is the number of videos tested, and n is the number of videos whose category label of top-1 is the same as the real label of the input video.

The improved neural network model in this paper is shown in Figure 3. In Figure 3, C represents the convolutional layer, and P represents the pooling layer. The main part of the AlexNet network classification effect is outstanding, and the model complexity is moderate, which can meet the real-time requirements of this paper. In order to further improve the speed of the algorithm, this article optimizes AlexNet partly. The model has 6 3D convolutional layers, 4 3D pooling layers, a dropout layer with a ratio of 0.5, two fully connected layers, and a SoftMax layer. The size of the 3D convolution kernel is $[3 * 3 * 3]$, where $3 * 3$ is the space dimension and 3 is the time dimension. The step size is 1, and padding is 1. From the first convolutional layer to the sixth convolutional layer, the numbers of convolution kernels are 64, 128, 256, 256, 512, and 512, respectively. Each convolutional layer is followed by a pooling layer. All the pooling types in this network structure are maximum pooling, which can effectively eliminate the estimated mean deviation caused by parameter errors of the convolutional

FIGURE 2: Schematic diagram of 3D convolutional neural network structure.

FIGURE 3: Improved fusion multichannel 3D convolutional neural network structure.

layer. The filter size of the first pooling layer is $[1 * 2 * 2]$, and the other filters are all $[2 * 2 * 2]$. After the dropout layer, the SoftMax layer is used for classification, and the behaviour category output is obtained. Because the feature enhancement effect of the LRN layer is random, it can improve the recognition rate for some tasks, but, for some tasks, it will

cause feature distortion and reduce the recognition rate. Therefore, this paper eliminates the LRN layer in the implementation process.

In order to further accelerate the speed of neural network convergence, this paper uses dropout technology for optimization. Dropout refers to discarding hidden nodes in the

neural network according to a certain probability during the training process, inactivating some neurons, reducing the complexity of the model, and effectively inhibiting the overfitting phenomenon of the neural network.

4. Results and Discussion

4.1. Experimental Data and Environment. The training of deep convolutional neural network has very high requirements for the configuration of the experimental platform. Therefore, the processor used in this article is Intel(R) Core(TM) i7-5930K CPU-3.50 GHz; installed memory is 16.0 GB; system type is a 64-bit Win7 operating system; the graphics card is GTX1080; the development tools are Café and VS2013.

The experimental data in this article are based on the public video action recognition dataset UCF101, WARD, and HMDB51.

The UCF101 dataset contains 101 types of actions, with a total of 13 320 video segments. The UCF101 dataset is composed of network videos shot in an unconstrained real environment. The video frame pixels are relatively low, contain different lighting information, and have partial occlusion and camera movement. The dataset divides actions into five types:

- (a) Human-human interaction, such as hair cutting, head massage, and 5 other categories
- (b) Playing musical instruments, such as flute, violin, and 10 other categories
- (c) Contains only human sports, such as blowing candles, playing Tai Chi, and 16 other categories
- (d) Human-object interaction, such as blowing hair, cutting vegetables, and 20 other categories
- (e) Sports, such as playing billiards, breaststroke, and 50 other categories

The HMDB51 dataset contains 51 types of actions and a total of 6 849 video segments. Most of the HMDB51 dataset comes from movie clips, and a small part comes from video sites such as YouTube. Similarly, HMDB51 is also divided into five types:

- (a) Facial expressions interacting with objects, such as smoking and drinking and 3 other categories
- (b) General facial motions, such as smiling, speaking, and 4 other categories
- (c) Human-to-human interaction body movements, such as hugs, kisses, and 7 other categories
- (d) Human-to-thing interaction body movements, such as drawing swords, horse riding, and 18 other categories
- (e) General body movements, such as applause and 19 other categories such as handstand

The WARD database places wearable multisensor nodes on the anatomical parts of the human waist, left wrist, right wrist, left ankle, and right ankle and collects 20 subjects and 13 kinds of movement data in the natural state, including

standing, sitting, lying, walking forward, walking counter-clockwise, walking clockwise, turning left, turning right, going upstairs, going downstairs, jogging, jumping, and pushing wheelchair.

As an evaluation indicator, we selected the recognition accuracy rate. The recognition accuracy rate is the most commonly used evaluation index, and the calculation formula is

$$\text{accuracy} = \frac{C_m}{N_{\text{all}}}. \quad (9)$$

In the above formula, C_m represents the number of correct recognitions. N_{all} represents all the numbers.

4.2. Analysis of the Impact of LRN on the Effect of Behaviour Recognition. In the training phase, the data volume of each batch of training images is 50. Through continuous attempts, the selected reference learning rate is 0.0001, the learning rate decay coefficient is 0.9, the learning rate decay period is 1,000, and the maximum number of iterations is 25,000. On the HMDB51 dataset and UCF101 dataset, this article trains according to the above-mentioned training strategy and compares the impact of using the LRN layer and not using the LRN layer on the recognition effect. Figures 4 and 5 show the training sample loss and the test sample loss, respectively, and the variation curve of the recognition accuracy with the number of iterations.

It can be seen from Figure 4 that whether the LRN layer is included does not have much impact on the loss of training samples and the loss of test samples, and the attenuation curve is basically consistent with the number of iterations. It can be seen from Figure 5 that, after deleting the LRN layer, the accuracy of the test sample has been slightly improved. This just verifies that the LRN layer has a certain randomness in the optimization of features. Therefore, this paper removes the LRN layer.

4.3. Analysis of Dropout's Influence on the Effect of Behaviour Recognition. Similarly, according to the above training strategy, on the HMDB51 dataset and the UCF101 dataset, this paper further compares and analyses the influence of dropout on the training results of neural networks, including training sample loss, test sample loss, and recognition accuracy, respectively, as shown in Figure 6. As shown in Figure 7, the dropout ratio is set to 0.5.

Figure 6 reflects that the loss of training samples and test samples with and without dropout network is basically consistent with the decreasing trend of the number of iterations. However, it can be seen from Figure 6 that, at the beginning of training, the network without dropout has a very obvious fluctuation, and after dropout is added, the loss of training samples shows a smooth downward trend. This is because dropout reduces the complexity of the model through the inactivation of some neurons and suppresses the degree of overfitting of the neural network to a certain extent. Figure 7 shows that dropout improves the accuracy of behaviour recognition by about 1%, and the effect is

FIGURE 4: The change curve of the loss function of the training sample and the test sample with the number of iterations. (a) Loss value with training in HMDB51 dataset. (b) Loss value with training in UCF-101 dataset. (c) Loss value with testing in HMDB51 dataset. (d) Loss value with testing in UCF-101 dataset.

relatively not obvious. This is because the recognition rate has reached more than 90% when dropout is not used, and there is not much room for improvement, so that the effect of dropout is not significant.

4.4. Performance Analysis of Different Optimization Functions. In this experiment, under the same other conditions, the model recognition results constructed using different optimization functions are compared. The optimization functions are SGD, Adam, and RMSPR.

The experimental results are shown in Figure 8 and Tables 1 and 2. The abscissa of Figure 8 and Tables 1 and 2 represent the number of iterations, and the ordinate represents the accuracy rate. The lines of different legends in the

figure represent different optimization functions. It can be seen from Tables 1 and 2 that the recognition rate of the Adam optimization function model is the highest, and we get the accuracy rates of 0.936 and 0.916.

4.5. Performance Analysis of Action Recognition. Under the same experimental conditions, three different random projections are selected, and the action recognition rate changes under different compression ratios are shown in Figure 9. It can be seen that, for different random projection matrices, the action recognition rate based on this algorithm increases as the data compression ratio increases, and the three random matrices almost obtain the same action recognition rate. However, a sparse binary random matrix can

(a)

(b)

FIGURE 5: Accuracy change curve of test sample with number of iterations. (a) Accuracy in HMDB51 dataset. (b) Accuracy in UCF-101 dataset.

(a)

(b)

FIGURE 6: Continued.

FIGURE 6: The change curve of the loss function of the training sample and the test sample with the number of iterations. (a) Loss value with training in HMDB51 dataset. (b) Loss value with training in UCF-101 dataset. (c) Loss value with testing in HMDB51 dataset. (d) Loss value with testing in UCF-101 dataset.

FIGURE 7: Accuracy curve of test samples with the number of iterations. (a) Accuracy in HMDB51 dataset. (b) Accuracy in UCF-101 dataset.

further reduce computing resources and power consumption by reducing the number of nonzero elements in each column of the matrix. The above results show that the sparse binary random matrix is helpful to reduce the computational complexity of the algorithm and improve the action recognition rate. Therefore, this article chooses a sparse binary matrix.

Figure 10 shows the comparison of the recognition accuracy between the method in this paper and the more typical action recognition methods in action recognition on

the three datasets. On the HMDB51 dataset, the recognition rates of IDT, IDTHDE, TS, DTS, KVME, and 3DCNN were 0.859, 0.879, 0.881, 0.912, 0.932, and 0.935, respectively. The method in this paper is the highest, which is 0.943. IDT uses the dense trajectory algorithm. IDTHDE improves the visual bag-of-words model and integrates multidimensional features to achieve higher-dimensional feature coding. These two action recognition methods are more traditional methods of manually designing features. TS recognizes actions by constructing a dual-stream spatiotemporal

FIGURE 8: The relationship between optimizer and recognition rate. (a) Accuracy in HMDB51 dataset. (b) Accuracy in UCF-101 dataset.

TABLE 1: The relationship between optimizer and recognition rate in HMDB51 dataset.

Iterations	SGD	RMSPROP	Adam
10	0.356	0.298	0.399
20	0.389	0.356	0.475
30	0.432	0.467	0.538
40	0.502	0.567	0.603
50	0.524	0.628	0.699
60	0.589	0.698	0.767
70	0.614	0.756	0.838
80	0.628	0.789	0.901
90	0.723	0.805	0.923
100	0.798	0.867	0.936

TABLE 2: The relationship between optimizer and recognition rate in UCF-101 dataset.

Iterations	SGD	RMSPROP	Adam
10	0.316	0.198	0.379
20	0.389	0.256	0.445
30	0.412	0.367	0.518
40	0.532	0.467	0.593
50	0.564	0.528	0.639
60	0.589	0.698	0.717
70	0.614	0.756	0.778
80	0.698	0.779	0.801
90	0.763	0.835	0.863
100	0.838	0.867	0.916

network model, but the network is relatively shallow. DTS has a very deep network architecture and improves the network in depth. The KVMF algorithm intercepts multiple

3D volumes from the video segment as the input to the network and uses the prediction vector obtained from each volume to represent the category probability of the action. It

FIGURE 9: Action recognition rate under different random projection matrices under different compression ratios.

FIGURE 10: Comparison of recognition accuracy of different algorithms.

can be seen from Figure 10 that the improved 3D convolutional neural network proposed in this paper has better action recognition capabilities on the three datasets.

5. Conclusion

In this research, a method of human action recognition based on random projection is proposed. First, use the optical flow picture and RGB picture obtained by the Lucas-Kanade algorithm. Secondly, the data of optical flow pictures and RGB pictures are compressed based on the random projection matrix of compressed sensing, which effectively reduces power consumption; and, based on random projection compression data, it can effectively find the optimal linear representation to reconstruct training samples and test samples. A multichannel 3D convolutional neural network is proposed, and two model fusions are used in the

test phase of the network to obtain the final recognition result. Compared with the traditional 3DCNN algorithm, the network constructed in this paper continuously adds dropout layers in the high-level feature extraction stage of the 3D convolutional network to reduce the overfitting situation during model training and speed up the training. In the output stage of the model, two fusion techniques are used to improve the stability and accuracy of the network model. Experimental results show that the algorithm in this paper not only achieves better recognition performance but also can better adapt to human action recognition problems in complex scenes and has better application prospects.

Data Availability

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

Conflicts of Interest

The author declares no conflicts of interest

References

- [1] F. Zhu, L. Shao, J. Xie, and Y. Fang, "From handcrafted to learned representations for human action recognition: a survey," *Image and Vision Computing*, vol. 55, pp. 42–52, 2016.
- [2] L. Xia, C. C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *Proceedings of 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 20–27, IEEE, Washington, DC, USA, June 2012.
- [3] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, no. 6, pp. 976–990, 2010.
- [4] A. Akula, A. K. Shah, and R. Ghosh, "Deep learning approach for human action recognition in infrared images," *Cognitive Systems Research*, vol. 50, pp. 146–154, 2018.
- [5] E. P. Ijjina and K. M. Chalavadi, "Human action recognition in RGB-D videos using motion sequence information and deep learning," *Pattern Recognition*, vol. 72, pp. 504–516, 2017.
- [6] J. Yang, J. Huang, Z. Jiang et al., "3D SIFT aided path independent digital volume correlation and its GPU acceleration," *Optics and Lasers in Engineering*, vol. 136, Article ID 106323, 2021.
- [7] I. El-Henawy, K. Ahmed, and H. Mahmoud, "Action recognition using fast HOG3D of integral videos and Smith–Waterman partial matching," *IET Image Processing*, vol. 12, no. 6, pp. 896–908, 2018.
- [8] D. Mistry and A. Banerjee, "Comparison of feature detection and matching approaches: SIFT and SURF," *GRD Journals-Global Research and Development Journal for Engineering*, vol. 2, no. 4, pp. 7–13, 2017.
- [9] S.-J. Ahn, D.-M. Ko, and K.-S. Choi, "Cow behavior recognition using motion history image feature," in *Proceedings of International Conference Image Analysis and Recognition*, pp. 626–633, Springer, Cham, Switzerland, June 2017.
- [10] F. Yang, L. Zhang, S. Yu et al., "Feature pyramid and hierarchical boosting network for pavement crack detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 4, pp. 1525–1535, 2019.

- [11] Y. Ma, Y. Liu, Q. Xie, and L. Li, "CNN-feature based automatic image annotation method," *Multimedia Tools and Applications*, vol. 78, no. 3, pp. 3767–3780, 2019.
- [12] H. Wang, T. Zhao, L. C. Li et al., "A hybrid CNN feature model for pulmonary nodule malignancy risk differentiation," *Journal of X-Ray Science and Technology*, vol. 26, no. 2, pp. 171–187, 2018.
- [13] K. Yan, R. Zhao, E. Liu, and Y. Ma, "A robust fundamental matrix estimation method based on epipolar geometric error criterion," *IEEE Access*, vol. 7, pp. 147523–147533, 2019.
- [14] Y. Li, R. Xia, Q. Huang, W. Xie, and X. Li, "Survey of spatio-temporal interest point detection algorithms in video," *IEEE Access*, vol. 5, pp. 10323–10331, 2017.
- [15] A. Siarohin, S. Lathuilière, S. Tulyakov et al., "First order motion model for image animation," *Advances in Neural Information Processing Systems*, vol. 32, pp. 7137–7147, 2019.
- [16] X. Hu, Y. Huang, X. Gao et al., "Squirrel-cage local binary pattern and its application in video anomaly detection," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 4, pp. 1007–1022, 2018.
- [17] S. U. Amin, M. Alsulaiman, G. Muhammad, M. A. Mekhtiche, and M. Shamim Hossain, "Deep Learning for EEG motor imagery classification based on multi-layer CNNs feature fusion," *Future Generation Computer Systems*, vol. 101, pp. 542–554, 2019.
- [18] X. Liao, K. Li, X. Zhu, and K. J. R. Liu, "Robust detection of image operator chain with two-stream convolutional neural network," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 955–968, 2020.
- [19] M. Al-Hammadi, G. Muhammad, W. Abdul, M. Alsulaiman, M. A. Bencherif, and M. A. Mekhtiche, "Hand gesture recognition for sign language using 3DCNN," *IEEE Access*, vol. 8, pp. 79491–79509, 2020.
- [20] A. B. Saleem, "Two stream hypothesis of visual processing for navigation in mouse," *Current Opinion in Neurobiology*, vol. 64, pp. 70–78, 2020.
- [21] Y. Shi, Y. Tian, Y. Wang, and T. Huang, "Sequential deep trajectory descriptor for action recognition with three-stream CNN," *IEEE Transactions on Multimedia*, vol. 19, no. 7, pp. 1510–1520, 2017.
- [22] J. Song, Z. Yang, Q. Zhang et al., "Human action recognition with 3D convolution skip-connections and RNNs," in *Proceedings of International Conference on Neural Information Processing*, pp. 319–331, Springer, Cham, Switzerland, December 2018.
- [23] X. Wang, L. Gao, P. Wang et al., "Two-stream 3-d convnet fusion for action recognition in videos with arbitrary size and length," *IEEE Transactions on Multimedia*, vol. 20, no. 3, pp. 634–644, 2017.
- [24] X. Wei, J. Du, Z. Xue et al., "A very deep two-stream network for crowd type recognition," *Neurocomputing*, vol. 396, pp. 522–533, 2020.
- [25] S. Aslani, M. Dayan, L. Storelli et al., "Multi-branch convolutional neural network for multiple sclerosis lesion segmentation," *NeuroImage*, vol. 196, pp. 1–15, 2019.
- [26] J. He, S. Terashima, H. Yamada, and S. Kidera, "Diffraction signal-based human recognition in non-line-of-sight (NLOS) situation for millimeter wave radar," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 4370–4380, 2021.
- [27] R. Huang and M. Sun, "Network algorithm real-time depth image 3D human recognition for augmented reality," *Journal of Real-Time Image Processing*, vol. 18, no. 2, pp. 307–319, 2020.
- [28] L. Gan, P. Du, J. Xia, and Y. Meng, "Kernel fused representation-based classifier for hyperspectral imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 5, pp. 684–688, 2017.
- [29] K. Yan, Y. Xu, X. Fang, C. Zheng, and B. Liu, "Protein fold recognition based on sparse representation based classification," *Artificial Intelligence in Medicine*, vol. 79, pp. 1–8, 2017.
- [30] J. Díaz, F. Marton, and E. Gobbetti, "Interactive spatio-temporal exploration of massive time-Varying rectilinear scalar volumes based on a variable bit-rate sparse representation over learned dictionaries," *Computers & Graphics*, vol. 88, pp. 45–56, 2020.
- [31] Y. He, G. Li, Y. Liao et al., "Gesture recognition based on an improved local sparse representation classification algorithm," *Cluster Computing*, vol. 22, no. 5, pp. 10935–10946, 2019.
- [32] L. Ding, S. Li, H. Gao et al., "Adaptive neural network-based finite-time online optimal tracking control of the nonlinear system with dead zone," *IEEE Transactions on Cybernetics*, vol. 51, no. 1, pp. 382–392, 2021.
- [33] A. A. Mousavi, C. Zhang, S. F. Masri, and G. Gholipour, "Structural damage localization and quantification based on a CEEMDAN hilbert transform neural network approach: a model steel truss bridge case study," *Sensors*, vol. 20, no. 5, p. 1271, 2020.
- [34] F. Orujov, R. Maskeliūnas, R. Damaševičius, W. Wei, and Y. Li, "Smartphone based intelligent indoor positioning using fuzzy logic," *Future Generation Computer Systems*, vol. 89, pp. 335–348, 2018.
- [35] Y. Hu, Q. Chen, S. Feng, and C. Zuo, "Microscopic fringe projection profilometry: a review," *Optics and Lasers in Engineering*, vol. 135, Article ID 106192, 2020.
- [36] W. Wei, S. Liu, W. Li, and D. Du, "Fractal intelligent privacy protection in online social network using attribute-based encryption schemes," *IEEE Transactions on Computational Social Systems*, vol. 5, no. 3, pp. 736–747, 2018.