EAI Endorsed Transactions
on Scalable Information Systems
Research Article EAI.EU

Feature extraction of dance movement based on deep learning and deformable part model

Shuang Gao[1] and Xiaowei Wang[2,*]

[1]College of Music, Anhui Normal University, Wuhu, 241000, Anhui, China
[2]Software College, Shenyang Normal University, Shenyang 110034, China
Email:407077764qq.com;zxcvfdsa5024@foxmail.com

Abstract

In complex scenes, the accuracy of dance movement recognition is not high. Therefore, this paper proposes a deep learning and deformable part model (DPM) for dance movement feature extraction. Firstly, the number of filters in DPM is increased, and the branch and bound algorithm is combined to improve the accuracy. Secondly, deep neural network model is used to sample points of interest according to human dance movements. The features extracted from the DPM and deep neural network are fused. It achieves a large reduction in the number of model parameters and avoids the network being too deep. Finally, dance movement recognition is performed on the input data through the full connection layer. Experimental results show that the proposed method in this paper can get the recognition result more quickly and accurately on the dance movement data set.

Keywords: DPM, dance movement feature extraction, deep neural network model.

Received on 23 December 2021, accepted on 31 December 2021, published on 05 January 2022

doi: 10.4108/eai.5-1-2022.172783

*Corresponding author. Email: zxcvfdsa5024@foxmail.com

1. Introduction

Human motion recognition is a promising research field in computer vision. It has important research value in monitoring system, intelligent home and virtual reality. Different from simple image recognition, human motion recognition is affected by many factors, such as chaotic background, different image acquisition equipment, insufficient categories of human motion database and so on [1,2]. At present, there are mainly traditional machine learning and deep learning methods for human motion recognition.

In traditional human action recognition methods, manual design features are widely used for recognition, such as HOG/HOF et al, [3]. In reference [4], a new method of motion recognition was designed, and the obtained behavior features were fed into a vector machine to train the model. However, the recognition accuracy of traditional methods in real scene data sets is not too high, and the recognition results are not good [5].

In recent years, deep learning has attracted much attention in the field of human motion recognition. Convolutional neural network (CNN) is a commonly used model in deep learning [6,7], which realizes image recognition through convolutional pooling of input images. The AlexNet network in 2015 made the CNN structure deeper [8]. Inception V2 network [9] made the CNN wider. Reference [10] sends both the time of video data and image information into CNN, but the effect of dual-stream network was not ideal, and consumed a lot of memory and training time. Reference [11] proposed a new

projection strategy in order to improve the recognition accuracy of tiny actions, which projected images onto multiple Cartesian planes to retain more behavior information. In reference [12], CNN convolution kernel was transformed into 3D, and time dimension was added into the convolution kernel, which improved accuracy and increases training difficulty. Reference [13] combined CNN with LSTM to improve the accuracy of recognition. Reference [14] proposed Deformable Convolurtional Networks (DCN), which had more flexible position deformation ability compared with ordinary convolutional networks. Reference [15] combined Deformable Part Model (DPM) with CNN to improve the accuracy of pedestrian detection. In reference [16], pooling layer was improved in order to learn the unstable convolution features in images, and the experimental verification results were good. Although convolutional neural network improved the accuracy of image recognition, it was easy to lose the early feature information in the process of human action recognition.

To sum up, this paper integrates the improved DPM for dance movement recognition on the basis of deformable convolutional neural network (DCN). It combines traditional machine learning with deep learning to improve the accuracy and detection speed of dance movement recognition.

2. DCN

Deformable convolutional neural network (DCN) is proposed to improve the adaptability of convolutional neural network to geometric deformation through deformable convolution and deformable region of interest pooling [17]. These two methods are based on the idea of further irregular migration of the convolution kernel sampling position information in the process of convolution pooling. Offsets are obtained by adding new modules without additional supervision. The deformable convolution is the ordinary convolution. The deformable convolution is to add an offset to the sampling position of the ordinary convolution. Deformable pooling is the addition of an offset for each bin location of the candidate region of interest. The offset is obtained by an additional layer of full connection. After the addition of modules, the deformable convolutional neural network can still be trained using the standard back propagation algorithm.

The comparison between ordinary convolution and deformable convolution is shown in figure 1. Where, the size of the convolution kernel is 3×3. Figure 1(a) shows ordinary convolution. The 3×3 convolution kernel is a rectangle of 9 points. Figure 1(b) and figure 1(c) represent deformable convolution. It adds an offset (arrow) to figure 1(a). Figure 1(b) shows the addition of offsets (arrows pointing to points) to the deformable convolution. Figure 1(c) shows the special case where a deformable convolution can reverse an angle.

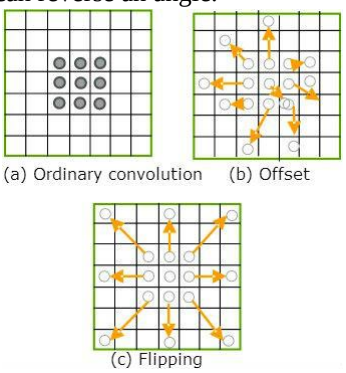

Figure 1. Traditional convolution and deformable convolution

3. DPM

Deformable part model (DPM) is an identification and detection method based on human body parts [18]. The model DPM for recognizing human motion is composed of root filter, component filter and relative position between filters. In general, the root filter has a lower resolution and the component filter has a higher resolution. Assuming that each target specifies the position of each filter in the model in the feature pyramid, $z=(t_0,\cdots,t_m)$ represents the DPM recognition model. Where, m is the number of filters, which can be set by itself. $t_i=(x_i,y_i,l_i)$ represents the coordinates (x_i,y_i) and the number of layers l of the position of the i-th filter.

The score of the total test window is shown in equation (1):

$$score(t_0,\cdots,t_m)=\sum_{i=1}^{m}F_i'\times\phi(H,t_i)-\sum_{i=1}^{m}d_i\times\phi_d(dx_i,dy_i)+b \quad (1)$$

Where

$$(dx_i,dy_i)=(x_i,y_i)-(2(x_0,y_0)+v_i) \quad (2)$$

$\sum_{i=1}^{n}F_i'$ represents the fractional sum of all the filters. $\phi(H,t_i)$ is the position of each component filter relative to the feature pyramid. d_i is the offset loss coefficient. $\phi_d(dx_i,dy_i)$ represents the offset loss score of the i-th component filter. b and v_i represent different offset real

numbers. (x_i, y_i) is the spatial position of the filter response maximum value of the i-th component. (x_0, y_0) represents the coordinates of the layer where the root filter is located. $2(x_0, y_0)$ means that the resolution of the component filter is twice that of the root filter.

4. Proposed feature extraction

Traditional convolution operations use convolution kernels with fixed shapes, generally rectangular. Deformable convolution adds an offset for each convolution sampling point so that the convolution window convolves according to the region of interest. The window shape of standard convolution for human body recognition is regular rectangle, while the window shape of deformable convolution network can be changed according to the object instance. Taking the two-dimensional convolution with a convolution kernel of 3×3 as an example, M represents the effective receptive field:

$$M = \{(-1,-1),(-1,0),\cdots,(0,-1),(1,1)\} \quad (3)$$

After sampling on the input feature graph x, the offset is added to each sampling point and multiplied by the weight w. For the position P_0 on the output feature graph, the deformable convolution is expressed as:

$$y(P_0) = \sum_{P_n \in M} (w(P_n) \cdot x(P_0 + P_n + \Delta P_n)) \quad (4)$$

Where, the offset parameter $\{\Delta P_n \mid n = 1, \cdots, N\}$, $N = |M|$ represents any position in M. ΔP_n only has a certain influence on the input layer pixels, but does not affect the weight w, so both w and ΔP_n need to be trained.

On the basis of ordinary convolution, a convolution layer is added to obtain the *offset*, which is ΔP_n in equation (4). The *offset* generated has two directions of horizontal and vertical coordinates.

First, an increased convolution layer (conv) is used to obtain the offset of deformable convolution. Then the *offset* is shifted in the convolution kernel to complete the deformable convolution. Where x is the input feature graph, and y is the output feature graph. Interest pooling divides feature graphs into k×k bins (k=7 is a freely set parameter).

In bin, deformable pooling of interest of row i and column j is expressed as:

$$y(i,j) = \sum_{Q \in bin(i,j)} \frac{x(Q_0 + Q_n + \Delta Q_{ij})}{n_{ij}} \quad (5)$$

Where Q_0 is the point in the upper-left corner of bin. Q_n represents any position in bin. n_{ij} indicates the number of pixels in bin. ΔQ_{ij} is the offset. The structure of deformable pooling network is shown in figure 2.

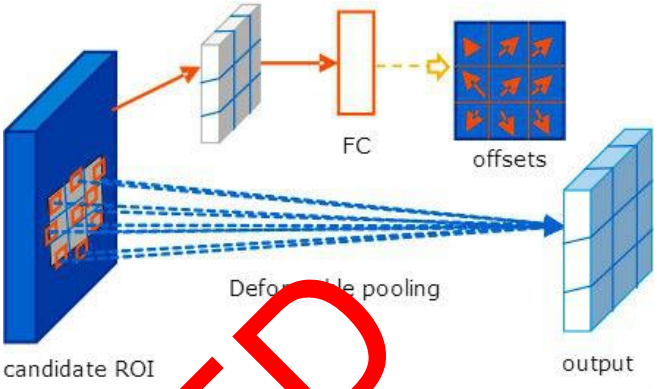

Figure 2. Structure of deformable pool of interest

Candidate regions of interest (ROI) are obtained by pooling the feature map [19,20]. Then the offset of the region of interest is obtained through the new full connection layer (FC). The input and output of the deformable convolutional network are the same as that of the ordinary convolutional network, except that during training, the newly added weights of the convolutional layer and the full connection layer for learning the offset are initialized to zero.

The deformable part is actually a model composed of a root filter and a component filter. The quality of DPM model is mainly determined by component filter. According to formula (1), $\sum_{i=1}^{n} F_i'$ represents the total fraction sum of filters. Therefore, with the increase of the number of component filters, the calculation of $\sum_{i=1}^{n} F_i'$ will become larger and the calculation speed will be correspondingly slower.

Experiments show that with the increase of component filters, although the detection accuracy is improved, the operation speed will be slowed down. In order to improve the experimental effect, we need to find a balance between the number of component filters and the operation speed. When the number of traditional DPM component models is 5, the accuracy of the collected human motion data set is 83.75%. When the number of component filters is increased to 8, the accuracy reaches 95.67%, and the accuracy of human motion recognition is improved by more than 11%. However, when the number of filters reaches more than 8, the improvement of accuracy is no longer obvious, which also brings some

calculation difficulty to the experiment. Therefore, this paper increases the component models in DPM to 8.

DPM usually divides the human body into five parts in human movement recognition, namely head, left upper body, right upper body, legs and feet. In this paper, three models are added into the head, left shoulder, right shoulder, left abdomen, right abdomen, leg and left and right feet. The improvement point is that the original left upper body is further divided into left shoulder and left abdomen. The right upper body is divided into the right shoulder and the right abdomen. The feet are divided into left and right feet. The subdivided model can more accurately identify similar actions, such as running and playing football.

In order to solve the problem of excessive workload of DPM artificial design features, this paper integrates Branch and Bound (BB) algorithm [21] into DPM model to locate human body. Branch represents that the whole image is divided into several small regions, and the score value of each branch in the small region is calculated respectively. Bound is to set the optimal solution function boundary for each region and automatically find the maximum value in this region. The maximum region obtained by BB algorithm can be used as the region of interest for human movement recognition.

Traditional DPM recognition of human movements takes about 11s. After the BB algorithm is fused and the number of filters is increased to 8, the detection speed of DPM is increased by about 3 times, and the result can be obtained in about 3.8s. According to the experimental data, traditional DPM takes a long time. Combined with BB algorithm, it can quickly get the maximum value of the function in the image and remove most impossible hypothetical target actions, thus effectively improving the detection speed. The steps for the improved DPM to recognize human movements are as follows:

Step 1. Extracting features.

DPM uses directional gradient histogram (HOG) for feature extraction.

Step 2. DPM model.

Formula (1) is the semantic model of DPM, and modeling is to establish the structural model through the semantic model of formula (1). The DPM structural model consists of a root filter and several component filters. In this paper, experiments show that increasing the number of component filters from 5 to 8 can effectively improve the accuracy of human movement detection.

Step 3. DPM model training.

The root filter is initialized first. The region of interest determined by BB algorithm is scanned by root model, and the highest score is determined as the location of the root filter. The component filter is then initialized. The position of the component filter is determined according to the position relationship between the component filter and the root filter. After determining one component, the position of the next component is continued to be searched until the position of all component filters is determined. Finally, DPM is continuously updated until the accuracy was less than 0.01.

Step 4. The trained model is used to identify and classify the actions in the data set.

The following will introduce the depth feature extraction.

The proposed lightweight deep learning network that combines shallow and deep networks in this paper is shown in figure 3. This network model mainly consists of two modules: one is shallow multi-scale network module, the other is deep network module. Specifically, the model in this paper consists of a convolution layer with a convolution kernel of 3×3, a maximum pooling layer, three densely connected shallow multi-scale modules, a convolution layer with a convolution kernel of 1×1×1, a deep network module, a full connected layer and softmax. Densely connected shallow multi-scale modules are mainly responsible for extracting and combining local features of video sequences to form longer and broader deep features. The main function of deep network module is to better integrate the features extracted from the previous module by virtue of its better abstraction ability and enhance the ability of the whole network model to express the features. The combination of shallow and deep networks enables the model in this paper to better represent the temporal and spatial features of video sequences and achieve better recognition results on the premise that the network model is not heavy enough for dance movement recognition tasks.

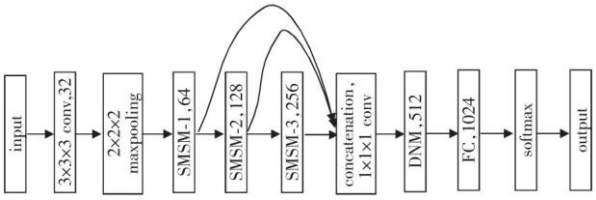

Figure 3. Lightweight deep learning network model combining shallow and deep networks, SDNet

Shallow multiscale modules

Generally speaking, motion feature extraction can be divided into global feature extraction and local feature extraction according to whether it is global feature extraction or not. However, for video-based human motion recognition tasks, most of the video background is complex environment or partial occlusion, and global feature extraction can not deal with this situation well. Multi-scale representation can be used to describe local features of video sequence in different scales in a simple way, which is convenient for analyzing local features of

video frames. Therefore, this model adopts multi-scale feature extraction to extract local feature of video sequence.

Inspired by the RFB module [22], this paper adopts the operation of expansive convolution [23] to expand the acceptance domain of the network model and proposes the shallow multi-scale module (SMSM) based on the shallow network, as shown in figure 4. Figure 4 takes the SMSM-1 module in figure 3 as an example. The 1×1×1 convolutional layer in this module does not contain nonlinear activation function, and the rest of the convolutional layer adopts ReLU as the nonlinear activation function. Specifically, this paper first uses 1×1×1 convolution layer to construct a bottleneck structure for each branch of the module. The main purpose of this structure is to reduce the number of channels for the feature mapping of the input to the next layer, so as to improve the calculation efficiency. Then, two superimposed 3×3×3 convolution layers are used to replace the 5×5×5 convolution layer, which can effectively reduce model parameters. Finally, the features extracted from convolution layers with different expansion coefficients are connected to the next structure of the model through the maximum pooling layer. Human action recognition based on video often contain a long-term dependence of space and time, and the expansion of convolution operation has better application of this kind of problem, at the same time, the operation can be pooling operation and without loss of information and keep the same parameter, through increasing receptive field, enables the convolution operation to the larger context of the output characteristic information. At the same time, it has been proved in reference [24] that this operation can improve the calculation speed and recognition accuracy. In this paper, two superposed 3×3×3 convolutional layers in each branch of the SMSM module are inflated, as shown in figure 4, and the expansion coefficients of each branch are respectively set as {1,2,3}. Meanwhile, in view of the expansion coefficients set as {1,3,5} in reference [25], this paper sets up two groups of comparative experiments with different expansion coefficients in the experimental stage.

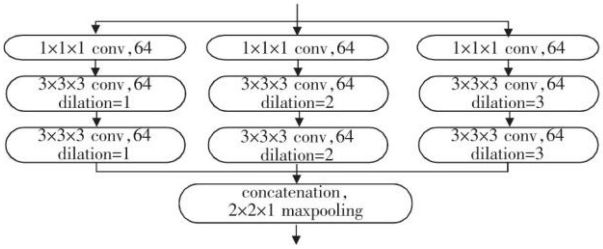

Figure 4. Shallow multi-scale module

Deep network module

Experiments in reference [26] had proved that densely connected networks not only did not bring redundancy problems, but also alleviated the problem of gradient disappearance and enhance the generalization ability of the network. The main reason is that the dense connection greatly reduces the computation amount of each layer, and the features extracted from each layer except the last layer of the dense connection can be reused. Therefore, it can be seen that dense connections can improve the performance of network models. Therefore, in this paper, features of three shallow multi-scale modules with different number of filters are densely connected to form longer and broader deep features. In order to better integrate the multi-scale features extracted from SMSM, this paper improved the structure module of deep network, and used the improved Network-In-Network (NIN) module to form the deep network module in this model.

NIN module has a higher level of abstraction, in terms of human action recognition task, NIN module can be in the low-level features for the top of the model, can get as much as possible under the same action from different angles, and scale features remain unchanged, the ability to a higher level of abstraction can enhance human action recognition model in this paper the expression ability of local characteristics. NIN in reference [27] is shown in figure 5. The network can be approximated as consisting of three superposed NIN modules and a global average pooling layer. Each NIN module consists of a 3×3 convolution layer, one 1×1 convolution layer and a combination of multi-layer perceptrons (MLPConv), and one 3×3 convolution layer. Among them, MLPConv and other convolution layers all take ReLU as their activation function. MLPConv has a stronger ability to model the distribution of various potential concepts than the simple linear convolution layer, which can improve the feature expression of traditional CNN.

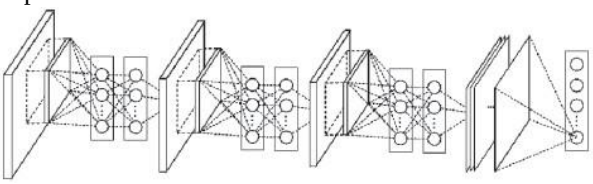

Figure 5. Overall structure of NIN

The deep network module of this article is shown in figure 6. NIN itself is superimposed by multiple CONV+MLPConv +CONV structures, and this paper adopts two such structures to superimpose to form deep network modules. Specifically, this paper firstly improves the original NIN in two aspects. On the one hand, considering the requirement of capturing spatio-temporal information as much as possible in the human motion recognition task in this paper, the convolution and pooling

of each layer of the original NIN are extended to 3D operation. On the other hand, the effective fusion of multi-scale features after intensive connection of shallow multi-scale modules is considered due to the extensive application of receptive field operation for a large range of shallow multi-scale modules in SDNet. The first CONV+MLPConv+CONV structure of deep network module is improved. The expansion convolution operation with expansion coefficient of 2 is added to all convolution operations of the first structure. The first improved structure is then connected to the second improved structure through a maximum pooling layer of 2×2×2. In the second improved structure, the convolution kernel expansion coefficient of all convolution is 1. Finally, considering that the global average pooling layer can reduce the number of parameters and make up for the shortcomings of the full connection layer, which is easy to over-fit, and meet the requirements of this paper to build a lightweight human action recognition model, the global average pooling method is still adopted in the classification layer as the feature mapping pooling method.

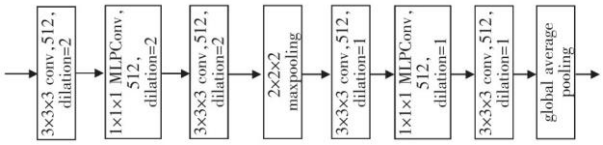

Figure 6. Deep networks module of SDNet

Feature fusion

In this paper, the feature graph of deformable convolutional neural network and DPM model is proposed to classify dance movements. Although convolutional neural network is superior to traditional machine methods in detection speed and accuracy, it has the problem of low accuracy of low-level feature extraction. The fusion of the improved DPM and the feature images extracted from the deformable convolutional network before the deformable pooling layer can effectively improve the accuracy of low-level feature extraction of the convolutional neural network and thus improve the recognition accuracy of the global network. The pre-processed data sets are input into the deformable convolutional network and the improved DPM model respectively. The candidate regions of interest are obtained by the deformable convolutional network and the feature images are obtained by the fusion of the DPM. The fused feature images are used as the input of the deformable pooling of interest.

The dance movement recognition algorithm in this paper has four steps.

Step 1. Body parts are obtained using improved DPM. After the data set is preprocessed and input into the network,DPM successively carries out the steps of feature extraction, modeling and training model.When DPM is used for human motion detection, BB algorithm can quickly get the optimal solution on the global graph, so that the DPM model can get the region of interest more quickly.

Step 2. Feature extraction from deformable convolutional networks. The same data set is used as input for the deformable parts model. First, three convolution blocks (including common convolution layer and pooling layer) are used to obtain low-level features. Then, the feature images are input into Region Proposal Network (RPN) [28,29] through two layers of deformable convolution blocks (including ordinary convolution layer, deformable convolution layer and pooling layer) to obtain candidate region of interest. Using back propagation algorithm to optimize parameters can also prevent over-fitting.

Step 3. Fusion of DPM features with candidate region of interest features. The fractions of root filter and component filter of the optimal DPM model were added, that is $\sum_{i=1}^{n} F_i'$ in equation (1) is obtained. Then feature graph F is obtained after a 3×3 convolution layer. Since deformable convolutional network is sampled layer by layer, up-sampling operation of feature graph extracted by CNN is required before feature fusion. F and up-sampled R fuse features using weighted summation. The fused feature graph is used as the input graph of the deformable pooling layer. Finally, through the full connection layer, Softmax is used to judge the action category.

Step 4. Input the training set of the data set to train the model, and then input the data of the test set and validation set into the trained model to obtain the experimental results.

5. Experiments and analysis

5.1. Experimental environment and data set

In this paper, the hardware used to identify the system is NVIDIA GPU and the operating system is Windows 10. TensorFlow is combined with PyCharm to ensure the operation of the experiment. This paper selects two mainstream data sets in human motion recognition: MPII data set and MSCOCO Person Keypoints2017 data set (hereinafter referred to as MSCOCO data set). Both datasets are images collected in different scenarios that contain human movements in many complex situations, such as crowding, scale changes, occlusion and rotation. The annotated frames in the MPII dataset are divided into various action categories, such as running, skiing, and playing soccer, and the unannotated frames are used as the test set. A similar preprocessing method is also used

for MSCOCO data sets, which commonly include speech, playing tennis, surfing and cooking in complex scenes. Finally, the robustness of the proposed method is proved by experiments in actual dance movements.

5.2. Experimental training process and results

In the forward propagation stage of the network, the training set is first used as the sample input into the designed model. The image is sampled layer by layer from input layer to output layer. In the backward propagation stage of the network, the weight is updated by calculating the error between the actual output image and the original image, and the process is repeated until the error is less than the preset value. The loss function of this experiment is the cross entropy function.
Accuracy rate is defined as:

$$acc = \frac{\sum_{i=1}^{S} C_i}{S} \times 100\% \quad (6)$$

acc is the accuracy rate. The action category of the target in the dataset is C_i. S is the number of scenarios in the dataset

Table 1 and Table 2 are the experimental results on the MPII dataset and MSCOCO Person Keypoints2017 dataset respectively. As there are few methods combining traditional methods with CNN, the convolutional neural network-based method with better performance in action recognition is mainly used in the comparative experiment.

Table 1. Comparison of accuracy on MPII dataset

Method	acc/%
DeepCut[30]	44.2
DeeperCut[31]	59.6
OpenPose[32]	75.7
Proposed	78.9

Table 2. Comparison of accuracy on MSCOCO dataset

Method	acc/%
DeepCut[30]	61.9
DeeperCut[31]	64.1
OpenPose[32]	66.6
Proposed	70.2

The recognition methods in reference [30,31] all recognize actions after human body is detected on ordinary convolution. The accuracy of recognition system in reference [31] is improved because CNN network becomes deeper. The recognition system in reference [32] analyzes the position of the human body after identifying the human body's key nodes using the ordinary convolutional network. In this paper, the method of reference [30] is also adopted, but deformable convolution is combined with improved DPM to identify human movements. It can be seen from the data in Table 1 that the identification method in this paper has higher accuracy. In addition, due to the combination of BB algorithm, the calculation speed has also been greatly improved. Under the same experimental conditions, it takes 1.43s for convolution recognition system to process a single image, while the recognition system in this paper only takes about 0.2s.

On MPII data set, confusion matrices of some action categories are selected, as shown in figure 7. As can be seen from figure 7, the new method in this paper is relatively easy to identify actions with obvious characteristics, such as running and playing football. But there is a slight lack of recognition of complex movements, such as yoga and dancing, which are more confusing than running and playing football. Because there are so many kinds of yoga moves and they have a lot in common with dancing.

	run	walk	football	dance	yoga
run	0.91	0.04	0.05	0.01	0.02
walk	0.03	0.84	0.04	0.03	0.02
football	0.03	0.04	0.92	0.01	0
dance	0.03	0.05	0.01	0.85	0.05
yoga	0.02	0.03	0.02	0.09	0.81

Figure 7. Confusion matrix on MPII dataset

As can be seen from Table 2, the recognition system in this paper also has good experimental effect in complex scenes, with higher accuracy than the full convolutional network [32]. Although the recognition system in reference [31] adds tracking points on human body, the effect is not ideal.

Figure 8 is the confusion matrix of some action categories in the MSCOCO dataset. Most of the actions in

MSCOCO are complex, so the accuracy is lower than that in MPII data set. As shown in figure 8, taking a selfie and making a phone call are generally similar hand-held mobile device actions, which are easy to confuse. Eating and drinking are generally sedentary activities, which are also prone to false detection. However, for actions with different postures, such as selfies and cooking, the false detection rate of the recognition method in this paper can be lower than 1% or even reach 0.

	selfie	cooking	call	drinking	eating
selfie	0.84	0	0.14	0.01	0.02
cooking	0.01	0.74	0.02	0.08	0.14
call	0.11	0.02	0.81	0.03	0.01
drinking	0.01	0.05	0.03	0.76	0.11
eating	0.01	0.04	0.02	0.02	0.82

Figure 8. Confusion matrix on MSCOCO dataset

6. Conclusion

In order to improve the accuracy of dance movement recognition in complex scenes, this paper proposes an improved recognition system combining DPM and deep learning. It adds the BB algorithm and the DPM to increase the dance movement recognition accuracy. But the amount of calculation is too large, which makes the detection speed slow down. The deformable convolution network model in the low-level feature extraction accuracy is low, so this article combines traditional recognition algorithm, to improve the CNN feature extraction accuracy. However, the recognition method proposed in this paper still has some problems in practical application, such as occlusion, low recognition accuracy, missing detection and false detection of complex actions. In the future, the hourglass network model will be integrated into the deformable convolutional network, and the stacked hourglass network model will be used to improve the recognition accuracy of a single joint and further improve the accuracy of dance movement recognition.

Acknowledgements.

This work was supported by: Anhui Province Federation of Social Sciences, Researching Project “Overseas Broadcast of Dancing Art of Anhui Local Special Characteristics “(2121CX156).

References

[1] Liu, J., Rahmani, H., Akhtar, N. et al. Learning Human Pose Models from Synthesized Data for Robust RGB-D Action Recognition. Int J Comput Vis 127, 1545–1564 (2019). https://doi.org/10.1007/s11263-019-01192-2

[2] Dianhuai Shen, Xueying Jiang, Lin Teng. A novel Gauss-Laplace operator based on multi-scale convolution for dance motion image enhancement[J]. EAI Endorsed Transactions on Scalable Information Systems, 2021. http://dx.doi.org/10.4108/eai.17-12-2021.172439

[3] Nishimura K, Yanabe M, Nagatani T, et al. Proposal on Improvement of Defect Identification Performance by Dimensional Reduction Focusing on Localization of HOG Feature in Electronic Component Inspection[J]. Journal of the Japan Society for Precision Engineering, 2019, 85.

[4] S. Ramyar, A. Homaifar, A. Karimoddini and E. Tunstel, "Identification of anomalies in lane change behavior using one-class SVM," 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2016, pp. 004405-004410, doi: 10.1109/SMC.2016.7844924.

[5] Wang X, Gong G, Li N, et al. Decoding pilot behavior consciousness of EEG, ECG, eye movements via an SVM machine learning model[J]. International Journal of Modeling Simulation and Scientific Computing, 2020(15):2050028.

[6] Qingwu Shi, Shoulin Yin, Kun Wang, Lin Teng and Hang Li. Multichannel convolutional neural network-based fuzzy active contour model for medical image segmentation. Evolving Systems (2021). https://doi.org/10.1007/s12530-021-09392-3

[7] Jisi A and Shoulin Yin. A New Feature Fusion Network for Student Behavior Recognition in Education [J]. Journal of Applied Science and Engineering. vol. 24, no. 2, pp.133-140, 2021.

[8] S. Gu, L. Ding, Y. Yang and X. Chen, "A new deep learning method based on AlexNet model and SSD model for tennis ball recognition," 2017 IEEE 10th International Workshop on Computational Intelligence and Applications (IWCIA), 2017, pp. 159-164, doi: 10.1109/IWCIA.2017.8203578.

[9] Baldassarre F, Morín, Diego González, Rodés-Guirao, Lucas. Deep Koalarization: Image Colorization using CNNs and Inception-ResNet-v2[J]. 2017. arXiv:1712.03400

[10] Aslan M F, Durdu A, Sabanci K. Human action recognition with bag of visual words using different machine learning methods and hyperparameter optimization[J]. Neural Computing and Applications, 2020, 32(12):8585-8597.

[11] Afza F, Khan M A, Sharif M, et al. A framework of human action recognition using length control features fusion and Weighted Entropy-Variances based Feature Selection[J]. Image and Vision Computing, 2021, 106:1-20.

[12] Indhumathi C, Murugan V, Muthulakshmii G. Human Action Recognition Using Spatio-Temporal Multiplier Network and Attentive Correlated Temporal Feature[J]. International Journal of Image and Graphics, 2021.

[13] S. Kumawat, M. Verma, Y. Nakashima and S. Raman, "Depthwise Spatio-Temporal STFT Convolutional Neural Networks for Human Action Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, doi: 10.1109/TPAMI.2021.3076522.

[14] Suresh A J, Visumathi J. Inception ResNet deep transfer learning model for human action recognition using LSTM - ScienceDirect[J]. Materials Today: Proceedings, 2020.

[15] J Lee, Ahn B. Real-Time Human Action Recognition with a Low-Cost RGB Camera and Mobile Robot Platform[J]. Sensors (Basel, Switzerland), 2020, 20(10).

[16] Jia J G, Zhou Y F, Hao X W, et al. Two-Stream Temporal Convolutional Networks for Skeleton-Based Human Action Recognition[J]. Journal of Computer Science and Technology, 2020, 35(3):538-550.

[17] Fei L, Dan L C, Jie T, et al. Cascaded one-shot deformable convolutional neural networks: Developing a deep learning model for respiratory motion estimation in ultrasound sequences[J]. Medical Image Analysis, 2020, 65.

[18] Dou J, Li J. Robust object detection based on deformable part model and improved scale invariant feature transform[J]. Optik - International Journal for Light and Electron Optics, 2013, 124(24):6485-6492.

[19] Shoulin Yin, Hang Li, Desheng Liu and Sh nd Ka m. Active Contour Modal Based on Density-orie d R Clustering Method for Medical Image egment on [J]. Multimedia Tools and Applications ve 79, pp. 3 049-31068, 2020.

[20] S. Yin and H. Li. Hot Region Selection Based on Selective Search and Modified Fuzzy C-Means in Remote Sensing Images[J]. IEEE Journ l of S cted T pics in Applied Earth Observations nd Re ote Se ing, vol. 13, pp. 5862-5871, 2020, doi: 1 1109/ TARS.2020.3025582.

[21] Panja S, Mondal S K Analytics of an Imperfect Four-Layer Production Inven ry Model Under Two-Level Credit Period Using Branch-and-Bound Technique[J]. Journal of the Operations Research Society of China, 2020(1).

[22] S. Hou, Y. Li, Y. Pan, X. Yang and G. Yin, "A Face Detection Algorithm Based on Two Information Flow Block and Retinal Receptive Field Block," in IEEE Access, vol. 8, pp. 30682-30691, 2020, doi: 10.1109/ACCESS.2020.2973071.

[23] Yin, S., Li, H. & Teng, L. Airport Detection Based on Improved Faster RCNN in Large Scale Remote Sensing Images [J]. Sensing and Imaging, vol. 21, 2020. https://doi.org/10.1007/s11220-020-00314-2

[24] Xiaowei Wang, Shoulin Yin, Ke Sun, et al. GKFC-CNN: Modified Gaussian Kernel Fuzzy C-means and Convolutional Neural Network for Apple Segmentation and Recognition [J]. Journal of Applied Science and Engineering, vol. 23, no. 3, pp. 555-561, 2020.

[25] Li M, Sun Q. 3D Skeletal Human Action Recognition Using a CNN Fusion Model[J]. Mathematical Problems in Engineering, 2021, 2021.

[26] G. Huang, Z. Liu, L. Van Der Maaten and K. Q. Weinberger, "Densely Connected Convolutional Networks," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2261-2269, doi: 10.1109/CVPR.2017.243.

[27] Lin M, Chen Q, Yan S. Network In Network[J]. Computer Science, 2013.

[28] Shoulin Yin, Hang L Lin Teng et al. An optimised multi-scale fusion method f airp t detection in large-scale optical re ote sen ng ima s [J]. International Journal of Image nd Fusion ol. 11, no. 2, pp. 201-214, 2020. DO 10.1080/ 4798 2.2020.1727573

[29] Jing Y Hang Li, Shoulin Yin. Dynamic Gesture Recognitio Based on Deep Learning in Human-to-Computer Interfaces [J]. Journal of Applied Science and Engi eering, vol. 23, no. 1, pp.31-38, 2020.

[30] l U., Gall J. (2016) Multi-person Pose Estimation with Local Joint-to-Person Associations. In: Hua G., Jégou H. (eds) Computer Vision – ECCV 2016 Workshops. ECCV 2016. Lecture Notes in Computer Science, vol 9914, pp. 627-642. Springer, Cham. https://doi.org/10.1007/978-3-319-48881-3_44

[31] Insafutdinov E., Pishchulin L., Andres B., Andriluka M., Schiele B. (2016) DeeperCut: A Deeper, Stronger, and Faster Multi-person Pose Estimation Model. In: Leibe B., Matas J., Sebe N., Welling M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol 9910, pp. 34-50. Springer, Cham. https://doi.org/10.1007/978-3-319-46466-4_3

[32] Z. Cao, T. Simon, S. Wei and Y. Sheikh, "Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1302-1310, doi: 10.1109/CVPR.2017.143.