Design of action detection system in wrestling match video based on 3D convolutional neural network

Yang Liu*

Faculty of Table Tennis, Badminton and Tennis,
Chengdu Sport University,
Wuhou District, Sichuan, Chengdu, China
Email: liuyang@cdsu.edu.cn
*Corresponding author

Qinyu Mei

School of Football,
Chengdu Sport University,
Wuhou District, Sichuan, Chengdu, China
Email: 1003229801@qq.com

Xin Gan

Faculty of Table Tennis, Badminton and Tennis,
Chengdu Sport University,
Wuhou District, Sichuan, Chengdu, China
Email: 87497590@qq.com

Ya Zhu

School of Wushu,
Chengdu Sport University,
Wuhou District, Sichuan, Chengdu, China
Email: 393211558@qq.com

Yongjie Wang

Faculty of Table Tennis, Badminton and Tennis,
Chengdu Sport University,
Wuhou District, Sichuan, Chengdu, China
Email: 100901@cdsu.edu.cn

Abstract: At present, there are some problems in motion detection in wrestling video at home and abroad, such as low detection accuracy and poor robustness. A motion detection system combining three-dimensional convolution neural network and recursive neural network is studied and designed. It uses three-dimensional convolution to obtain low-level feature code, then uses recursive memory module to obtain timing features, and finally completes motion detection according to timing features. Under the ratio of these three parameters, the accuracy of 3D-CNN convolutional neural network structure is higher than that of 2D-CNN. When the ratio of the influence factor of circular memory module P to that of circular memory module C is 1, the accuracy of 3D-CNN improves the fastest and the accuracy is close to 20%. The research results provide a new idea for the development of human motion detection and recognition technology.

Keywords: 3D-CNN; three-dimensional convolutional neural network; action detection; wrestling video; spatio-temporal features.

Reference to this paper should be made as follows: Liu, Y., Mei, Q., Gan, X., Zhu, Y. and Wang, Y. (2022) 'Design of action detection system in wrestling match video based on 3D convolutional neural network', *Int. J. Wireless and Mobile Computing*, Vol. 22, No. 1, pp.29–37.

Biographical notes: Yang Liu is a member, lecturer, with master's degree and Bachelor of Arts from China. He is a national second-class athlete of international wrestling and a national first-level referee of international wrestling. He is now the deputy director of the Graduate Management Office of the Department of Ball, the Secretary of the Party branch of 2021 and the Secretary of the General Branch of the Youth League.

Qinyu Mei is a Member of the Communist Party of China, Bachelor of education, Master student, International Wrestling National First-Class Athlete, National Second-Class Referee. As a professional Athlete in Ningbo Sports School, in 2009 to 2014, in 2014 to 2021, in September 2016 to June 2020 and in September 2020.

Xin Gan is a Member of the Communist Youth League, Bachelor of Education, Chengdu University of Physical Education Master student. September 2013–June 2016, Xinxian No. 1 Senior High School; September 2016–June 2020, Henan Agricultural University and September 2020–June 2022, Chengdu University of Physical Education.

Ya Zhu is an Active Party Membership, Teaching Assistant; Graduation University: University Beijing Sports University, Graduate Student Beijing Sports University, Major: Sports Training.

Yongjie Wang is a Member of the Communist Party of China, and received Bachelor of Education degree from Chengdu University of Physical Education. September 2012–June 2015, Master student; September 2015–June 2019, Lijin County No. 2 Middle School and September 2019–June 2022, School of Physical Education, Liaocheng University, Chengdu School of Physical Education.

1 Introduction

Human behaviour detection systems have made important progress in the fields of smart home, military security, and smart cities. The sports world is keen to analyse the performance of athletes during the game through videos, in order to provide a criterion for the final performance of the athletes. Traditional action recognition mainly includes detecting moving targets, feature extraction and feature understanding (Tateno et al., 2020; Moriwaki et al., 2019; Funke et al., 2019). With the continuous enhancement of computer technology, deep learning is a new research hotspot in machine learning at this stage, and it has a wide range of applications in computer vision, speech recognition, motion detection, image classification, etc. The three-dimensional Convolutional Neural Network (3D-CNN) is a network that uses the time dimension to expand the two-dimensional Convolutional Neural Network (2D-CNN), which is able to realise the effective extraction of video features. Compared with traditional methods, depth learning methods such as 2D-CNN, 3D-CNN and Recurrent Neural Network (RNN) have ideal results in different data sets. They not only have good generalisation performance, but also can greatly reduce labour cost and improve detection accuracy through automatic learning. Each depth method has different advantages. CNN is good at processing grid data, while RNN has a very powerful function in sequence data extraction (Fu and Aldrich, 2019; Fan, 2021; Wu et al., 2021). In view of this, this study proposes a motion detection system for 3D-CNN and Recurrent Neural Network (RNN) in wrestling match videos, which aims to make a huge contribution to the automatic detection of human motion. The research consists of three parts. The second part focuses on the motion detection system in 3D-CNN and RNN wrestling video. The third part analyses the performance of the motion detection system in wrestling video. The fourth part summarises the results and points out the direction and shortcomings of follow-up research.

2 Motion detection system for 3D-CNN and RNN wrestling videos

The research creatively adopts the network structure of 3D-CNN and circular memory module to design two parallel modules P and C and refined loss function. Modules P and C implement candidate video segment proposal and classification tasks, respectively. Wrestling match videos contain many action clips of the athletes, but the light wrestling action only occupies part of the clips. This requires an action detection system for wrestling match videos to omit irrelevant backgrounds and videos, and only focus on actions related to the athlete's match score, so as to accurately identify and classify the types of human movements. The motion detection system for wrestling match videos first needs to effectively compress the original video data and obtain low-level video features, then extract the timing information of the compressed features and analyse the content included in each part of the video, and finally use the acquired timing information to complete the athlete's motion detection (Pan et al., 2021; Brumann et al., 2021; Ding et al., 2019). The system architecture designed by this study includes three parts: encoding low-level video feature ϕ, extracting timing information, and detecting timing actions. The specific schematic diagram is shown in Figure 1. The coded low-level video reference video is the original video frame, and the extraction method is 3D-CNN. At this stage, the time step is used as the video level feature extraction unit, which divides the video into segments

composed of video frames, and obtains low-level video features through 3D-CNN. Then, use the recurrent memory module to obtain high-level features and use the recurrent memory module to obtain the timing information of the video, so as to obtain high-level timing features, and finally complete the athlete's timing action detection, accurately obtain the athlete's motion segment and complete the action classification (Hong et al., 2019; An et al., 2021; Liu et al., 2018).

Set the input continuous frame sequence video that including L frames to $X = \{x_l\}_{l=1}^{L}$, each time step includes δ frames, and the total included time steps are referring to $T = L/\delta$. The video data is extracted through the 3D-Convolutional (C3D) neural network, and the principal component analysis method is used to reduce the dimensionality of the features, and the Softmax classification layer is no longer needed. The principal component analysis method for the optimisation of description vector is the process of principal component extraction by taking the feature vector as a sample set. In the implementation steps of principal component extraction, firstly, the description vector is selected as the sample set, and secondly, the average value of all description vectors is calculated. Thirdly, the mean value obtained in the previous step is subtracted from each description vector to form a matrix, and then the matrix is multiplied by its inverse to form a covariance matrix and the subsequent matrix is diagonalised. The remaining elements after diagonalisation are eigenvalues, which are standardised at the same time. Each eigenvalue corresponds to an eigenvector. Fourth, select the largest N eigenvalue as the optimised eigenvalue, and retain the number of eigenvalues according to a certain proportion. The higher the proportion, the more information is relatively retained. On the contrary, the smaller the proportion, the less information is retained. Finally, the eigenvectors corresponding to the two eigenvalues are extracted to form a new matrix, which is transposed and multiplied by the original sample vector to obtain the denoised data. These data maximise the amount of information compared with the original data.

C3d is an efficient temporal feature extractor, which is very suitable for feature extraction of video data, with simple network structure and fast running speed. The entire network includes 8 convolutional layers, 3 fully connected layers and 5 maximum pooling layers. This network structure is a very standard convolution structure, which is different from the two-dimensional convolution structure in the pooling pool and the convolution layer. The schematic diagram is shown in Figure 2. As shown in this figure, 8 convolutional layers are divided into 5 groups, and each group needs to be connected with a maximum pooling layer. The first group contains only one convolutional layer, and there are 64 convolutional kernels. The pooling layer is to compress the feature map, which can enhance the robustness of the feature (Wang et al., 2018; Luo et al., 2021; Rongved et al., 2021). Owing to the important value of time sequence information extraction, only the spatial dimension is considered in the first group of convolutional pooling processing, and the time sequence is not considered. During the pooling process, the step size is 1×2×2, and the pooling core is 1×2×2. The second group of convolution is also a convolution layer. The number of convolution kernels is increased to 128 to resist the reduction of spatial dimensions, combining low-level features and forming high-level features, and then completing three-dimensional pooling processing to increase the robustness of time and space. During the pooling process, the step size is 2S×2×2, and the pooling core is 2×2×2. The third group of convolution consists of two convolution layers, increase the number of convolution kernels to 256 and complete three-dimensional pooling. The fourth and fifth groups of convolutions are both convolutional layers, the fourth group increases the number of convolution kernels to 512, the fifth group does not increase the number of convolution kernels, which is followed by two fully connected layers, and the number of output units is 4096. The fully connected layer is a common dimensionality reduction network layer, and its main function is to receive commonly used linear and neural network activation functions. Finally, the C3D network is coded through time, and feature coding is referring to $f = \left(\{x_i\}_{i=(t-1)\times\delta+1}^{t=\delta}\right)$.

Figure 1 Framework of action detection system for wrestling video

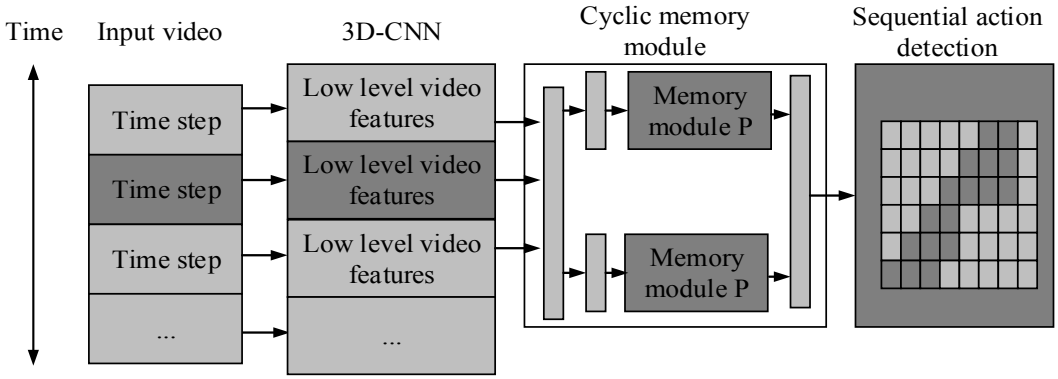
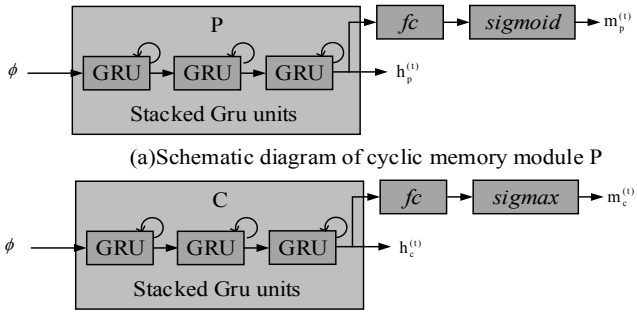


Figure 2 Schematic diagram of C3d network structure

Conv1a 64
Pool 1
Conv2a 128
Pool 2
Conv3a 256
Conv3b 256
Pool 3
Conv4a 512
Conv4b 512
Pool 4
Conv5a 512
Conv5b 512
Pool 5
Fc6 4096
Fc7 4096
Fc8 500

The most common cyclic memory modules are the Long Short-Term Memory (LSTM) and the Gate Recurrent Unit (GRU), which are both optimised for RNN and can solve long-term memory problems and prevent gradients. The LSTM training process first calculates the error of the last layer, then updates the parameters through the gradient descent algorithm, and then passes it layer by layer until all the parameters are updated. In the long and short-term memory network, there are eight groups of parameters that need to be learned, which are forgetting gates, input gates, output gates, and the weight matrix and bias term of the unit state. The weight matrix is different in calculating the two directions of back propagation. GRU is a simplified structure of LSTM, whose calculation is simpler and can create larger network. The basic unit of the cyclic memory module is GRU (Hung and Su, 2021; Wang et al., 2021; Jaa et al., 2020). In order to enhance the learning ability and expressive ability of the module, the research created a module composed of multi-layer GRU units, and designed two parallel recurrent memory modules P and C. The structures are shown in Figures 3(a) and 3(b), respectively. The inputs of the network structure are all low-level coding features ϕ, and the final hidden layer output results of the stacked GRU units are $h_p^{(t)}$ and $h_c^{(t)}$, respectively, and the output results of the two parallel modules P and C are combined to complete the action behaviour detection. Module P completes the classification of the video action or background, and module C completes the classification of the previous video. $m_p^{(t)}$ and $m_c^{(t)}$ are K-dimensional vector and C+1-dimensional vector, respectively. C is the sum of the category number of the action behaviour and the background. The training of the module is completed by the idea of a Single Shot multiple box Detector (SSD) in object detection, and the module P obtains K default right-aligned video segments at each time step. The video segment obtained at the time step is referred to as $\left[b_s^{(t)}, b_e^{(t)}\right]$, $b_s^{(t)}$ and $b_e^{(t)}$ are the beginning and the right boundary of the video segment, respectively. The specific implementation process of the semantic constraints of module P is as follows. A fully connected layer is added to the hidden layer output result of the GRU stacking unit, and the final output result is obtained through the Sigmoid function. This value represents the confidence score of each default video segment, and determines whether the feature is background or action behaviour by comparing with the real value. Similar to module P, the specific implementation process of the semantic constraints of module C is as follows. A fully connected layer is added to the hidden layer output result of the GRU stack unit, and the final output result is obtained through the Sigmax multi-classification layer. This value indicates whether the past video belongs to an action category by comparing with the real value.

Figure 3 Network structure of cyclic memory modules P and C



(a)Schematic diagram of cyclic memory module P

(b)Schematic diagram of cyclic memory module C

The timing behaviour detection basis is the hidden layer output of the cyclic memory modules P and C, $m_p^{(t)}$ and $m_c^{(t)}$ are to better train the recurrent memory modules P and C, so as to set reasonable semantic constraints. After generating K default video segments of different lengths at each time step, the actions contained in the video will be recognised and detected (tang et al., 2020; Rahman et al., 2021; Pratiwi et al., 2021). Before that, it is necessary to serially fuse the hidden layer output $h_p^{(t)}$ and $h_c^{(t)}$ of the recurrent memory module, and then use the final feature $h_{det}^{(t)} = h_h^{(t)} \| h_c^{(t)}$ to obtain the detection output result, which is referred to as $D^{(t)} = f_{out}\left(h_{det}^{(t)}\right)$. As a K×(C+1)-dimensional matrix, each row of $D^{(t)}\hat{U}\left\{\left(b_s^{(t)}, b_e^{(t)}, v_k^{(t)}\right)\right\}_{k=1}^{K}$ is a $v_k^{(t)}$ vector of (C+1)-dimension, which refers to the default category confidence score of the video segment whose number of rows is K, which means K default right-aligned video segments of each time step, $b_s^{(t)}$ and $b_e^{(t)}$, respectively represent the start and end positions of the default video segment.

The loss function of the network includes the loss function of the timing information extraction stage, the loss function of the timing behaviour detection stage, and the total loss function. The loss of the time sequence information extraction

stage is caused by the back propagation of the loss function, and the purpose is to achieve the semantic constraints of the training process (Zhu et al., 2021; Kandel and Castelli, 2020; Lukic et al., 2019; Yang et al., 2018). The semantic constraint of the recurrent memory module P is to compare the semantic constraint output $r^{(t)} = m_p^{(t)}$ of the module with the real label $y^{(t)}$ of the corresponding video segment. $r^{(t)}$ is the K-dimensional vector, which refers to the proportion of K right-aligned default videos that needed to detect the action behaviour. $y^{(t)}$ refers to the detection of whether the type of the video is an action video segment or a background, whose value is 0 and 1. When the time overlap between the detection video and the action behaviour video segment exceeds the threshold 1, the value of $y^{(t)}$ is 1, otherwise, the value is 0. $y^{(t)}$ is also a *K*-dimensional vector, which represents the true labels of the *K* default video segments within time step *t*. When the time step is set to *t*, the semantic constraint damage function of module P is formula (1).

$$L_p = -\sum_{k=1}^{K}\left\{w_0^k y_k^{(t)} \log r_k^{(t)} + w_1^k \left(1 - y_k^{(t)}\right) \log\left(1 - r_k^{(t)}\right)\right\} \tag{1}$$

This function represents the multi-label loss of weighted binary cross entropy. The weighting coefficients are w_0^k and w_1^k, respectively and, and their values refer to the ratio of positive and negative samples of each default video segment of a specific length. The semantic constraint of recurrent memory module C is to compare the semantic constraint output $m_c^{(t)}$ of this module with the real category $c^{(t)}$ of the video. When $m_c^{(t)}$ is the background of the video segment, $w^{(t)}$ can be regarded as a constant ρ, otherwise, the value is 1. $c^{(t)}$ is sparse (*C*+1) dimensional 0–1 vector. ρ indicates the proportion of detected background categories. When the time step is set to *t*, the semantic constraint damage function L_c of module C is shown in formula (2).

$$L_c = -w^{(t)} \log\left(m_c^{(t)}\left[c_k^{(t)}\right]\right) \tag{2}$$

The purpose of the semantic constraint module is to obtain the required learning information through the recurrent memory unit, which is an additional loss in the training phase. The real loss is the loss of sequential action behaviour detection, which can be divided into positioning loss and classification loss. The classification loss L_{detc} of the classification output $v_k^{(t)}$ when the time step is set to *t* is shown in formula (3).

$$L_{detc} = -\sum_{k=1}^{K} w_k^{(t)} \log\left(v_k^{(t)}\left[z_k^{(t)}\right]\right) \tag{3}$$

The ratio of the default video of each specific length is referring to $w_k^{(t)}$, the real category of the first default video segment in the *k*-th time step is $z_k^{(t)}$. Set the time step to *t*, the locate loss is as shown in formula (4).

$$L_{detl} = \frac{1}{2}\sum_{k=1}^{K}\left\{\left(\frac{\left(v_k^{(t)}\left[z_k^{(t)}\right]\right)^2}{\left(O_k^{(t)}\right)^a} - 1\right) \times \left[z_k^{(t)} > 0\right]\right\} \tag{4}$$

$O_k^{(t)}$ refers to the overlap ratio *k* between the default time boundary and the real time boundary of the first default video segment at the time step *t*. When the video segment represents the background, the value of $z_k^{(t)}$ is 0, otherwise it is 1. Comprehensive analysis shows that the total loss *L* of samples in a batch of data is formula (5).

$$L = \sum_{(X,Y)\hat{I}\chi}\sum_{t}\left(\lambda_p \times L_p + \lambda_c \times L_c + \lambda_{det} \times L_{det}\right) \tag{5}$$

λ_p, λ_c and λ_{det} refer to the weighted parameters of losses in each stage.

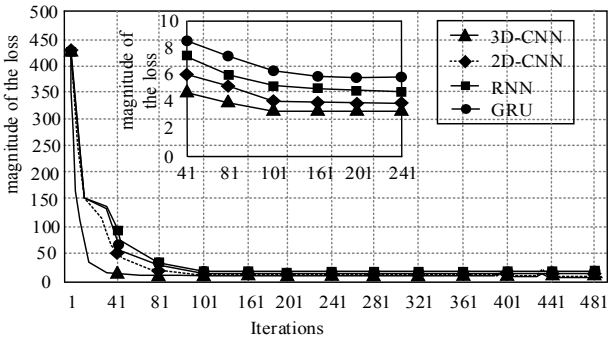
The study training sample is a sliding window with time sequence length of T_w, obtained by the dense sampling and the long video division, and set $T_w >> K$ to ensure the effectiveness of the training sample. Dense sampling can not only make the data have the characteristics of diversity, but also improve the learning ability of the network, thereby avoiding repeated operations. In the process of network training, the total loss function is used to obtain the error and complete the back propagation. It is necessary to set different loss proportions at different training times. In the initial stage of training, appropriately increase the proportion of semantic constraint loss to obtain a more efficient cyclic memory module. In the later stage of training, appropriately adjust the proportions of positioning loss and classification loss, so as to accomplish goals better (Moriwaki et al., 2019). In the stage of completing the network action behaviour test, considering that the output of the network is the proportion of the categories related to the default video segment, the research introduces the non-maximum consistent method to remove part of the data. This method has extremely high processing speed, and the operation is very simple. The performance indicators of motion detection and recognition are mean Averary Precision (mAP) of different categories, including video-level mAP and frame-level mAP. These two indicators can respectively detect the detection performance of the algorithm in time and space.

3 Performance analysis of motion detection system in wrestling video

In the experiment, the behaviour detection subset of the UCF101 data set is used to verify the performance of the model. The number of samples in the training set and the test set are 700 and 300, respectively, and the hyperparameter optimisation is completed by the five-fold cross-validation method. The learning rate is 0.001, the number of samples in each batch is 64, the constraint weight is reduced by 1/2 for every 5 K training of the cyclic memory module P, and the constraint weight is reduced by 1/2 for every 8.5K training of

the cyclic memory module P. Figure 4 is a common deep learning neural network training loss result. As a whole, it can be seen that the training loss values of the four network structures continue to decrease as the number of iterations increases. The loss values of the two structures of RNN and GRU are higher than the two convolutional network structures of 3D-CNN and 2D-CNN. Both the 3D-CNN and 2D-CNN convolutional neural network algorithms converge quickly when the number of iterations is about 20, and the gap between the two algorithms is not particularly obvious. However, when the number of iterations ranges from 20 to 100, compared with the 2D-CNN convolutional neural network algorithm, the 3D-CNN convolutional neural network algorithm converges faster and the loss value tends to be more stable.

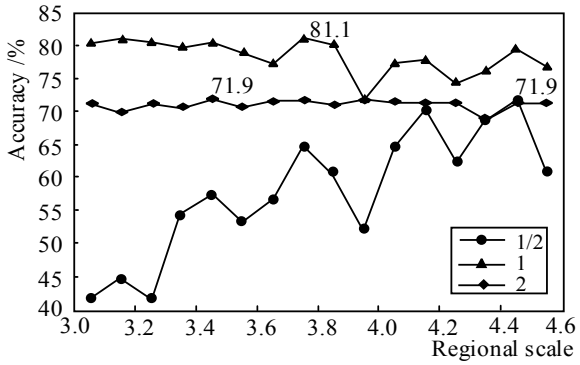
Figure 4 Training loss results of common deep learning algorithms



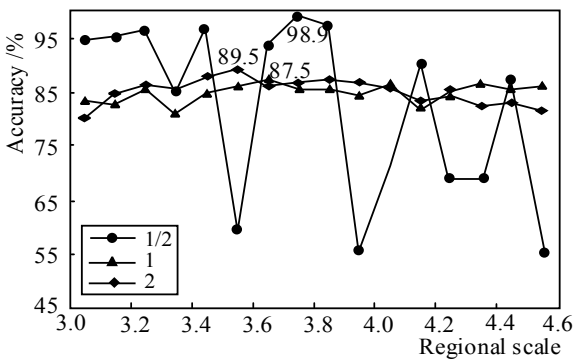
Experiments are performed to compare the accuracy at different regional scales. The accuracy results of the two convolutional neural network structures of 3D-CNN and 2D-CNN are shown in Figures 5(a) and 5(b), respectively. The experiment set the ratio of the impact factor of the recurrent memory module P and the impact factor of the cyclic memory module C to 1/2, 1, 2 respectively, and then use the 3D-CNN and 2D-CNN convolutional neural network structures to test the detection performance of the model. It can be seen that when the ratios of the 3D-CNN convolutional neural grid are 1/2, 1 and 2, the corresponding best regional scales are 4.4, 3.7 and 3.4, and the accuracy rates obtained are 71.9%, 81.1% and 71.9%, respectively. When the ratios of the two convolutional neural networks of 2D-CNN are 1/2, 1 2, respectively, the corresponding best regional scales are 3.7, 3.6, 3.5, and the accuracy rates obtained are 98.9%, 87.1% and 89.5%, respectively. Through data analysis, the accuracy of the two convolutional neural network structures is improved under the ratio of the three types of parameters, and when the ratio of the impact factor of the cyclic memory module P and the impact factor of the cyclic memory module C is 1/2, the accuracy rate of the 3D-CNN convolutional neural network structure has been improved the fastest, and the accuracy rate has increased by nearly 20%. When the ratio of the impact factor of the cyclic memory module P and the impact factor of the recurrent memory module C is 1, the accuracy of the two structures is not much different, and the maximum accuracy is only 6.0%.

The experiment uses the loss value judgment model to detect the effect in the wrestling match video, and uses Tensorboard to present the loss trend. The content image, style image, noise image and the entire loss result are shown in Figures 6 (a), 6(b), 6(c) and 6(d). The model loss value evaluates the effect of motion detection in the generated wrestling match from a quantitative perspective. Both the style image loss and the overall loss gradually decrease as the number of training increases, and the loss value quickly reaches the convergence value, which is 0 and 2.000e+6, respectively. The noise image loss curve shows a rapid increase first and then a slow convergence, and the value of the convergence is repeatable, and the loss peak value is 6.6e+4. The convergence speed of the content image loss is slow, and can reach an optimal convergence value, and the loss value has a certain repeatability.

Figure 5 Accuracy results of two network models



(a)Structure accuracy of 3D-CNN convolutional neural network



(b)Structure accuracy of 2D-CNN convolutional neural network

Figure 6 Content image, style image, noise image and the whole loss result

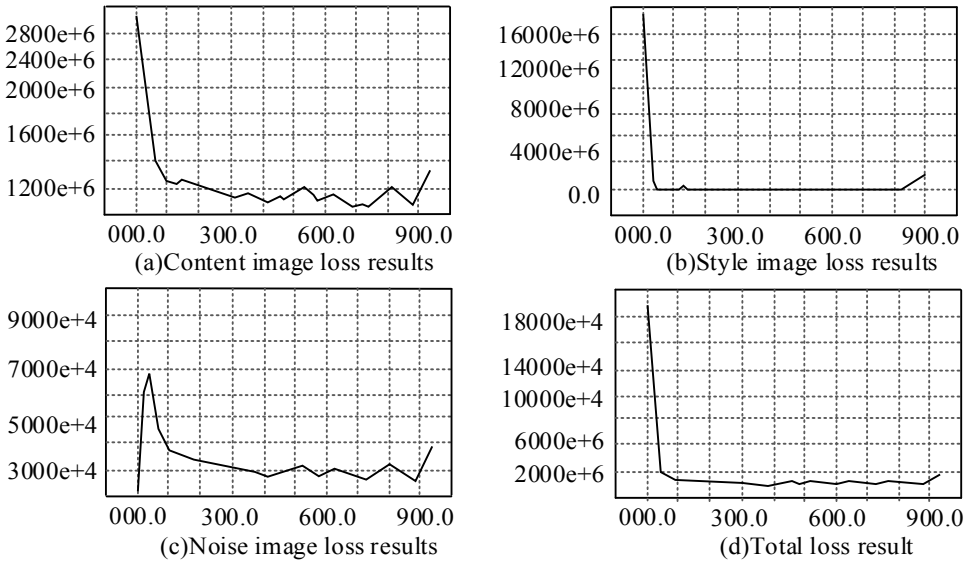
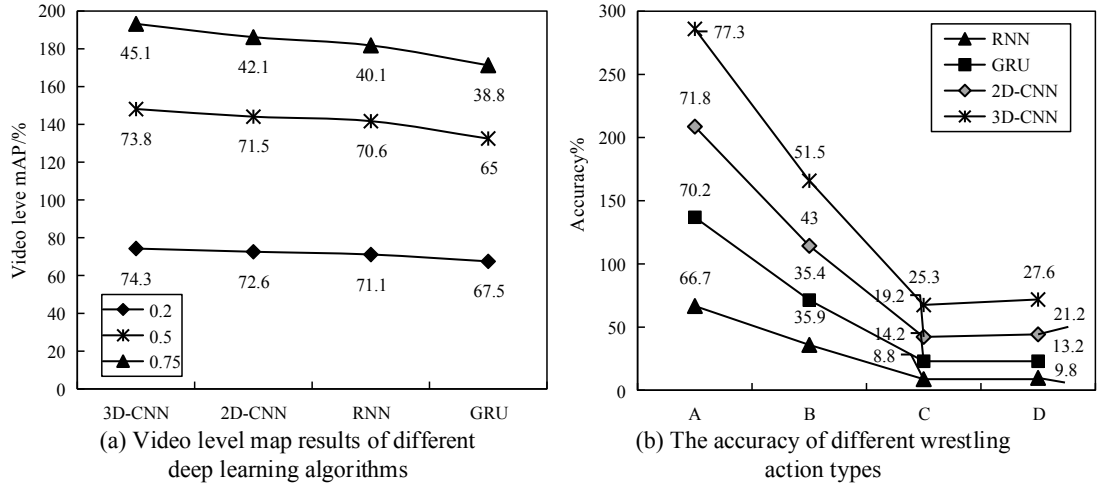


Figure 7 Map results of video level and accuracy of different wrestling action categories



Experiments set up different frame numbers to verify network performance through classification accuracy and frame-level mAP. As a result, as the number of consecutive frames increases, frame-level mAP and classification accuracy are improved to varying degrees, but when the number of consecutive frames is 6, frame level mAP and classification accuracy both reach their peaks. Therefore, when the number of consecutive frames is 6, the 3D-CNN convolutional neural network can obtain timing information very well. The experiment sets the overlap thresholds to 0.3, 0.5, 0.7, and the video level mAP results of different deep learning algorithms are shown in Figure 7(a). Compared with other deep learning algorithms, the 3D-CNN convolutional neural network algorithm has a better video level mAP under the three overlap thresholds.

When the overlap thresholds are 0.3, 0.5 and 0.7, the video level mAP of the 3D-CNN convolutional neural network is 74.3%, 73.8% and 45.1%, respectively, and the video level mAP of the 2D-CNN convolutional neural network is 72.6%, 71.5% and 42.1%, respectively. The video level mAP of the RNN neural network is 71.1%, 70.6% and 40.1% and the video level mAP of the GRU neural network is 67.5%, 65.0% and 38.8%, respectively. The accuracy of 3D-CNN convolutional neural network in different wrestling action categories is shown in Figure 7(b). The letter actions A–D are used to represent the side holding one shoulder backward, locking jaw over chest, lever holding bridge and side anti holding trunk backward in the wrestling competition video. It can be seen that the algorithm has very big differences in the detection accuracy of different actions. The 3D-CNN convolutional neural network has good detection performance for action A, but the detection performance for action C and action D is not ideal. This may be determined by the nature of the action itself, or it may be the reason for the small number of samples in the training set.

The research analyses the action detection effect of the designed 3D-CNN (scheme 1) and 3D-CNN (scheme 2) designed by other scholars in wrestling competition video[27]. The results are shown in Figure 8. Action 1-action 8, respectively represent eight actions: standing still, taking a step back, left arm flush with shoulder, right arm flush with

shoulder, left hand waving towards chest, recovering upward, waving forward and holding fist. The motion detection scheme designed by the Institute has better motion detection effect, and the accuracy of waving the left hand towards the chest is the highest, while the accuracy of clenching the fist is the lowest.

Figure 8 Motion detection effect of two schemes in wrestling video

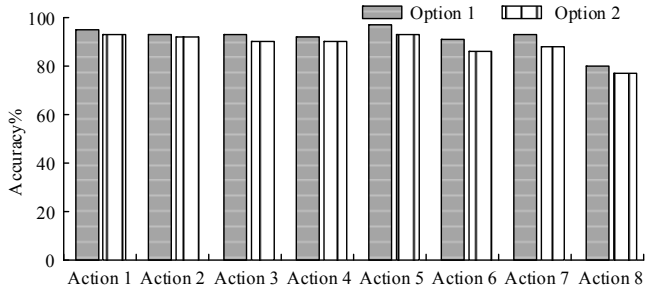


Figure 9 is a physical diagram of the action detection effect of the proposed method on wrestling competition video. It can be seen that the action types detected by the model are very clear and intuitive, and have strong practicability.

Figure 9 Physical picture of action detection effect in wrestling competition video

4 Conclusion

Motion detection in sports competitions has very important value and research significance. Aiming at the current status of human action behaviour detection, this study proposes an action video detection system in wrestling match video combining three-dimensional convolutional neural network and recurrent neural network. 3D-CNN and 2D-CNN and RNN and GRU two neural network algorithms have the same trend of loss value changes in the first 20 iterations, but in the interval of 20 to 100 iterations, the 3D-CNN convolutional neural network algorithm has a faster convergence speed and more stable loss value. The action detection results of the two convolutional neural network structures of 3D-CNN and 2D-CNN show that when the ratio of the impact factor of cyclic memory module P and the impact factor of recurrent memory module C is 1/2, 1, 2, the corresponding optimal area scale of 3D-CNN convolution neural network is 4.4, 3.7 and 3.4, and the accuracy rate is 71.9%, 81.1% and 71.9% in turn; while the optimal area scale corresponding to the 2D-CNN convolutional neural network structure is 3.7, 3.6 and 3.5, the accuracy rates obtained are 98.9%, 87.1% and 89.5%, respectively. Both the style image loss and the overall loss gradually decrease as the number of training increases, and the convergence values are 0 and 2.000e+6, respectively. The noise image loss curve shows a rapid increase first and then a slow convergence, and the value of the convergence is repeatable, and the loss peak value is 6.6e+4. When the overlap thresholds are 0.3, 0.5 and 0.7, the video level mAP of the 3D-CNN convolutional neural network is 74.3%, 73.8% and 45.1%, respectively, and the corresponding values are better than the other three deep learning algorithms. Limited by time and energy, there are still some problems in the research. The network structure needs to be further optimised in the follow-up to improve the detection accuracy of low-quality video images.

Acknowledgement

This work was supported by 2021 Chengdu philosophy and Social Sciences key research base – Chengdu world famous City Research Centre Project (Project No.: CDMC2021B09)

References

An, Q., Pan, Z. and You, H. (2021) 'Ship detection in Gaofen-3 SAR images based on sea clutter distribution analysis and deep convolutional neural network', *International Journal of Sensors*, Vol. 18, No. 2, pp.334–354.

Brumann, C., Kukuk, M. and Reinsberger, C. (2021) 'Evaluation of open-source and pre-trained deep convolutional neural networks suitable for player detection and motion analysis in squash', *International Journal of Sensors*, Vol. 21, No. 13, pp.4550–4574.

Ding, Y., Liu, Z. and Huang, M. et al. (2019) ''Depth-aware saliency detection using convolutional neural networks', *International Journal of Visual Communication and Image Representation*, Vol. 61, pp.1–9.

Fan, Y. (2021) 'Criminal psychology trend prediction based on deep learning algorithm and three-dimensional convolutional neural network', *International Journal of Psychology in Africa*, Vol. 31, No. 3, pp.292–297.

Fu, Y. and Aldrich, C. (2019) 'Flotation froth image recognition with convolutional neural networks', *International Journal of Minerals Engineering*, Vol. 132, pp.183–190.

Funke, I., Mees, S.T. and Weitz, J. et al. (2019) 'Video-based surgical skill assessment using 3D convolutional neural networks', *International Journal of Computer Assisted Radiology and Surgery*, Vol. 14, No. 7, pp.1217–1225.

Funke, I., Mees, S.T. and Weitz, J. et al. (2019) 'Video-based surgical skill assessment using 3D convolutional neural networks', *International Journal of International Journal of Computer Assisted Radiology and Surgery*, Vol. 14, No. 7, pp.1217–1225.

Hong, J., Luo, Y. and Mou, M. et al. (2019) 'Convolutional neural network-based annotation of bacterial type IV secretion system effectors with enhanced accuracy and reduced false discovery', *International Journal of Briefings in Bioinformatics*, Vol. 21, No. 5, pp.1–12.

Hung, P.D. and Su, N.T. (2021) 'Unsafe construction behavior classification using deep convolutional neural network', *International Journal of Pattern Recognition and Image Analysis*, Vol. 31, No. 2, pp.271–284.

Jaa, B., Ms, B. and Maa, C. et al. (2020) 'Convolutional neural network with batch normalization for Glioma and stroke lesion detection using MRI', *International Journal of Cognitive Systems Research*, Vol. 59, pp.304–311.

Kandel, I. and Castelli, M. (2020) 'How deeply to fine-tune a convolutional neural network: a case study using a histopathology dataset', *International Journal of Applied Sciences*, Vol. 10, No. 10, pp.3359–3378.

Liu, M., Jiang, J. and Wang, Z. (2018) 'Colonic polyp detection in endoscopic videos with single shot detection based deep convolutional neural network', *International Journal of IEEE Access*, Vol. 7, pp.334–3354.

Lukic, V., Gasperin F.D. and Brüggen, M. (2019) 'ConvoSource: radio-astronomical source-finding with convolutional neural networks', *International Journal of Galaxies*, Vol. 8, No. 1, pp.3–31.

Luo, Z., Ling, L. and Yin, J. et al. (2021) 'Deep learning of graphs with N-gram convolutional neural networks', *International Journal of IEEE Transactions on Knowledge and Data Engineering*, Vol. 29, No. 11, pp.2125–2139.

Moriwaki, H., Tian, Y-S., Kawashita, N. and Takagi, T. (2019) 'Privacy-preserved fall detection method with three-dimensional convolutional neural network using low-resolution infrared array sensor', *International Journal of J. Chemical and Pharmaceutical Bulletin*, Vol. 67, No. 5, pp.426–432.

Moriwaki, H., Tian, Y-S., Kawashita, N. and Takagi, T. (2019) 'Three-dimensional classification structure-activity relationship analysis using convolutional neural network', *International Journal of Chemical and Pharmaceutical Bulletin*, Vol. 67, No. 5, pp.426–432.

Pan, D., Zou, C. and Rong, H. et al. (2021) 'Early diagnosis of Alzheimer's disease based on three-dimensional convolutional neural networks ensemble model combined with genetic algorithm', *Journal of Biomedical Engineering*, Vol. 38, No. 1, pp.47–55.

Pratiwi, R.A., Nurmaini, S. and Rini, D.P. et al. (2021) 'Deep ensemble learning for skin lesions classification with convolutional neural network', *International Journal of IAES International Journal of Artificial Intelligence (IJ-AI)*, Vol. 10, No. 3, pp.563–570.

Rahman, T., Chowdhury, M. and Khandakar, A. et al. (2021) 'Transfer learning with deep convolutional neural network (CNN) for pneumonia detection using chest X-ray', *International Journal of Applied Sciences*, Vol. 10, No. 9, pp.3233–3251.

Rongved, O.A.N., Hicks, S.A. and Thambawita, V. et al. (2021) 'Using 3D convolutional neural networks for real-time detection of soccer events', *International Journal of Semantic Computing*, Vol. 15, No. 2, pp.161–187.

Tang, S., Yang, M. and Bai, J. (2020) 'Detection of pulmonary nodules based on a multiscale feature 3D U-Net convolutional neural network of transfer learning', *International Journal of PLoS ONE*, Vol. 15, No. 8. Doi: 10.1371/journal.pone.0235672.

Tateno, S., Meng, F. and Qian, R. et al. (2020) 'Privacy-preserved fall detection method with three-dimensional convolutional neural network using low-resolution infrared array sensor', *International Journal of Sensors*, Vol. 20, No. 20, pp.5957–5979.

Wang, S., Wu, K. and Chu, T. et al. (2021) 'SOSPCNN: structurally optimized stochastic pooling convolutional neural network for tetralogy of Fallot recognition', *International Journal of Wireless Communications and Mobile Computing*, No. 1, pp.5792975–5792993.

Wang, S.H., Lv, Y.D. and Sui, Y. et al. (2018) 'Alcoholism detection by data augmentation and convolutional neural network with stochastic pooling', *International Journal of Journal of Medical Systems*, Vol. 42, No. 1, pp.1–11.

Wu, X., Shi, Y. and Fomel, S. et al. (2021) 'FaultNet3D: predicting fault probabilities, strikes, and dips with a single convolutional neural network', *International Journal of IEEE Transactions on Geoscience and Remote Sensing*, Vol. 57, No. 11, pp.9138–9155.

Yang, W., Chen, Y. and Huang, C. et al. (2018) 'Video-based human action recognition using spatial pyramid pooling and 3D densely convolutional networks', *International Journal of Future Internet*, Vol. 10, No. 12, pp.115–125.

Zhu, W., Qu, J. and Wu, R. (2021) 'Straight convolutional neural networks algorithm based on batch normalization for image classification', *International Journal of Jisuanji Fuzhu Sheji Yu Tuxingxue Xuebao/Journal of Computer-Aided Design and Computer Graphics*, Vol. 29, No. 9, pp.1650–1657.