3D-CNN-based Action Recognition Algorithm for Basketball Players

Zhilei Cui
College of Physical Education, Taiyuan University of Technology, Taiyuan 030024, China
E-mail: cuizhilei@tyut.edu.cn

Keywords: Basketball technical action; Target detection; Action recognition; Single shot multibox detector; 3D-CNN

Received: April 21, 2024

The development of artificial intelligence has led to numerous methods for human action recognition. In basketball, its technical action features are obvious, so the feasibility of recognizing and classifying its technical actions is high. However, the existing action recognition methods are difficult to effectively utilize continuous frames, resulting in poor accuracy of basketball technical action recognition. Thus, the study suggests a continuous frame action identification approach based on the single-shot multi-edge detection algorithm and 3D convolutional neural network in order to enhance the performance of technical action recognition. The experimental results revealed that single shot multibox detector algorithm accurately recognizes the human body in the image and labels its confidence level. In addition, in basketball action recognition, the loss value of original frame was 6.0 and 6.8 on the training set and validation set, respectively, and the loss value of crop frame was 5.1 and 5.9 on the training set and validation set, respectively. 3D convolutional neural network achieved the highest classification accuracy of about 88.3% for the stop-and-go jump shot action in the original frame and its crop frame with an average recognition rate of about 90.3%. The recognition accuracy of original frame and crop frame increased with the increase of epoch, and reached a stable state when the epoch was 30. The presence of variable features in the European step, change of direction, and Sam Gould's action led to misjudgment of both original frame and crop frame. The accuracy of the original frame training set and test set were 0.91 and 0.81, respectively, and the accuracy of the crop frame training set and test set were about 0.92 and 0.81, respectively. After the fusion of the original frame features and the crop frame features, the average recognition rate was about 94.6%, which was significantly higher than that of the single-resolution recognition. In addition, with the increase of frame input, the F1-score gradually increased, while the false positive rate gradually decreased. When the frame input was 7, the F1-score and the misjudgment rate were 0.79 and 0.19, respectively. When the frame input was 16, the F1-score and the misjudgment rate were 0.92 and 0.05, respectively. The above results show that the continuous frame action recognition method based on single-shot multi-frame detection algorithm and three-dimensional convolutional neural network can realize the accurate recognition of the technical action in basketball video.

Povzetek: Predstavljen je algoritem za prepoznavanje košarkarskih akcij, ki temelji na 3D-CNN in algoritmu za zaznavanje večrobnih okvirjev. Rezultati kažejo, da metoda bistveno izboljša točnost prepoznavanja tehničnih akcij.

1 Introduction

Basketball, the second most popular sport globally, boasts over 2.7 billion fans. With the rise of short-video platforms, all kinds of high-quality videos provide people with rich spiritual food. Basketball technical videos and game highlights are popular among basketball fans. Recognizing the technical moves in the videos can promote the sport of basketball. In the field of video action recognition (AR), among the traditional recognition methods, improved dense trajectorie (iDT) has the best recognition performance, but its recognition speed is too slow. Moreover, the emergence of deep learning provides new ideas for video AR, common deep learning AR methods are single-frame video image-based recognition methods, two stream convolutional neural network (TSCNN), long short-term memory (LSTM) network and 3D convolutional neural network (3D-CNN) [1-2]. Among the recognition methods based on single-frame video images the selected frames are difficult to represent the whole video and it is difficult to match the extracted features with the action features. The spatial convolution of TSCNN is only performed on a single frame and cannot learn pixel-level correspondence between spatial and temporal features. Although LSTM can handle the timing problem better, it suffers from long training time and high consumption of computational resources. 3D-CNN, on the other hand, can effectively capture temporal and spatial information and can handle volumetric data, and thus performs well in video classification and AR [3-5]. However, due to the complexity of basketball technical action (BTA) and 3D-CNN also suffers from the problem of long training time. Therefore, the study suggests a basketball video target detection (TD) and augmented reality (AR) model based on single shot multibox detector (SSD) algorithm and 3D-CNN in order to increase training speed and identification accuracy and decrease the consumption of computer resources. The innovation of the study is to combine the TD of SSD to generate AR of dual-resolution 3D-CNN with low-resolution image input, which provides a new idea for BTA recognition.

The study is divided into six chapters, in which the introduction first provides background information and explanations of the importance of the study. Chapter two is a literature review, which will describe the SSD algorithm (SSDA) and related research on 3D-CNN. Chapter 3 is the study methodology, which will investigate the SSDA and dual-resolution 3D-CNN architectures. Chapter 4 is the experimental results, which will analyze the performance of the research methods, TD and AR. Chapter five is a discussion, which will analyze the advantages of the proposed method. Chapter 6 is the conclusion of the research results, and will summarize the research results of this paper.

2 Related works

The two-stage algorithm's candidate region generation and following pixel or feature resampling phases are removed by the SSDA, which combines the regression concept and the anchor frame mechanism. All calculations are contained in a single network. As a result, it is extensively employed in many different industries and offers the benefits of simple training and quick speed. For the problem of video face detection, Liu et al. suggested a face detection approach based on SSDA and Res-Net. This method uses kernel correlation filtering to track successive n frames and Res-Net as the core network of SSDA. The approach can achieve real-time detection and increase the precision of video face detection, according to experimental data [6]. For the purpose of solving the vehicle recognition problem in intelligent transportation, Zhao et al. suggested a detection approach based on the SSDA and feature pyramid augmentation strategy. By using a feature pyramid augmentation strategy, the method enhanced SSD's feature extraction (FE) capabilities, and it enhanced its localization capabilities by cascading the detection mechanism. According to experimental results, this method's detection time is less than that of previous approaches [7]. Liu et al. proposed a detection model based on SSDA and depth separable fusion hierarchical feature model for the pedestrian detection problem. The model effectively reduced the complexity of the model by depth separable convolution and realized feature fusion enhancement by using hierarchical structure. On the INRIA dataset, experimental results showed that the improved model's leakage detection rate was only 9.68% [8]. Li et al. proposed a detection model based on multi-block SSDA for the small TD problem of on-site monitoring of railroad drones. Experimental results were obtained from this model. After segmenting the image into overlapping blocks, the model transfers the blocks individually to the SSD and utilizes the concept of non-maximum suppression sub-layer suppression and filtering algorithms were used to remove overlapping frames from the sub-layers. The multi-block SSD model's overall accuracy was shown to be 96.6% in experimental findings, 9.2% better than the conventional SSDA [9]. For the problem of electromagnetic luminous surface defect detection, Xu et al. suggested an enhanced SSD technique based on feature fusion, which employs the concept of feature pyramid network. The enhanced SSDA outperforms previous algorithms in terms of electromagnetic luminescence defect detection, according to experimental data [10].

3-dimensional convolution neural network (3D-CNN) is a popular choice in multi-channel image processing because, in contrast to 2D-CNN, it is capable of capturing discriminative features along spatial and temporal dimensions, generating multiple information channels from adjacent video frames, and performing convolution and downsampling in each channel independently to obtain the final feature representations by combining the information from the video channels. Xu and Zhang proposed a 3D-CNN gesture estimation method. The method used the depth image to reconstruct the 3D spatial structure of the hand, and converted the hand model to voxel grid by 3D-CNN, which in turn realized the hand gesture estimation. Experimental results revealed that the improved 3D-CNN model can achieve an average accuracy of 87.98% with an average absolute error of only 8.82 mm [11]. Rehman et al. proposed a 3D-CNN-based tumor classification model for the problem of brain tumor detection and automated classification. This model extracted the brain tumor features by 3D-CNN and realized the verification and classification of the features by feed-forward neural network. The outcomes revealed that this 3D-CNN model could classify brain tumors with an accuracy of up to 98.32% [12]. For the problem of concentration estimate of gas mixtures, Pareek et al. suggested a concentration estimation network based on 3D-CNN and constrained Boltzmann machine. The network processed sensor array data using 3D-CNN, and for end-to-end gas concentration estimation, it used a constrained Boltzmann machine. The results indicated that the network was able to accurately predict the concentration of the gas mixture [13]. Chaddad et al. proposed a prediction model based on Gaussian mixture model and 3D-CNN for the problem of predicting the survivability of pancreatic ductal adenocarcinoma patients. The model utilized a Gaussian mixture model to model the distribution of learned features obtained from preoperative computed tomography, followed by FE and learning via 3D-CNN, and finally a robust classifier based on random forest to predict survival outcomes. The experimental results revealed that the ROC of this model was 0.72, which was much higher than other models [14]. For the purpose of classifying hyperspectral images, Roy et al. presented a hybrid spectral convolutional neural network-based classification model. The model utilized 3D-CNN for joint spatial-spectral feature representation and further learned more advanced spatial representation through 2D-CNN. According to experimental findings, the hybrid spectral convolutional neural network successfully decreased model complexity while increasing the classification accuracy of hyperspectral images [15]. A summary of the related literature is shown in Table 1.

Table 1: Summary of related literature

Author	Method	Index
Liu et al. [6]	A face detection method based on the SSD algorithm and Res-Net	Accuracy is over 92%
Zhao et al. [7]	Vehicle detection method based on SSD algorithm and feature pyramid enhancement strategy	80.6% accuracy and 14 ms detection time
Liu et al. [8]	Pedestrian detection model based on SSD algorithm and deep separable fusion hierarchical feature model	The missed detection rate is 9.68%
Li et al. [9]	A small object detection model based on the multi-block SSD algorithm	The accuracy rate is 96.6%
Xu et al. [10]	Improved SSD algorithm based on feature fusion	Accuracy is over 90%
Xu and Zhang [11]	A 3D-CNN gesture estimation method based on the end-to-end hierarchical model and physical constraints	The average accuracy is 87.98%, and the mean absolute error is only 8.82 mm
Rehman et al. [12]	Tumor classification model based on 3D-CNN	The highest accuracy rate is up to 98.32%
Pareek et al. [13]	Concentration estimation network based on a 3D-CNN and a confined Boltzmann machine	Accuracy is over 91%
Chaddad et al. [14]	Pretive model of viability in patients with pancreatic ductal adenocarcinoma based on a Gaussian mixed model and 3D-CNN	The ROC is 0.72
Roy et al. [15]	Hyperspectral image classification model based on a hybrid spectral convolutional neural network	The accuracy rate is nearly 99%

In summary, both SSDA and 3D-CNN are widely used in image processing. However, facing the problem of complex BTA and strong correlation of consecutive frames of basketball video, it is difficult for existing image processing techniques to accurately recognize and classify BTA. Based on this, the study proposes a BTA recognition algorithm based on SSDA and 3D-CNN with dual-resolution 3D-CNN, in order to realize the accurate recognition and classification of BTA.

3 Action recognition algorithm for basketball video based on SSDA and 3D-CNN

Basketball sports videos can accurately reflect the technical movements of basketball and have certain advantages in teaching. However, due to the wide variety of BTAs, it is difficult to classify them by traditional AR techniques, and there is a lack of methods to utilize continuous image frames. Therefore, in order to better recognize and classify BTAs in sports videos, the study proposes a video clipping and AR method based on SSD and 3D-CNN.

3.1 Video cropping algorithm based on SSDA

Before the recognition of video actions, the video needs to be cropped to select the motion region and reduce the size of image frames. As a lightweight TD algorithm, the SSD method addresses the drawback of the YOLO algorithm, which is its inability to detect small targets, with its advantages of high detection accuracy and quick operation time. Although both SSDA and YOLO algorithm perform detection through CNN network, SSDA performs detection in the intermediate layer instead of after the fully connected layer. The SSDA first generates prediction frames (PFs) on the input image (IM), then labels the locations based on the PFs and feature maps (FMs), and outputs the classification categories. Lastly, the non-maximum suppression technique is used to eliminate the redundant and PFs that do not match the confidence expectation [16-17]. Figure 1 depicts the SSDA's framework.

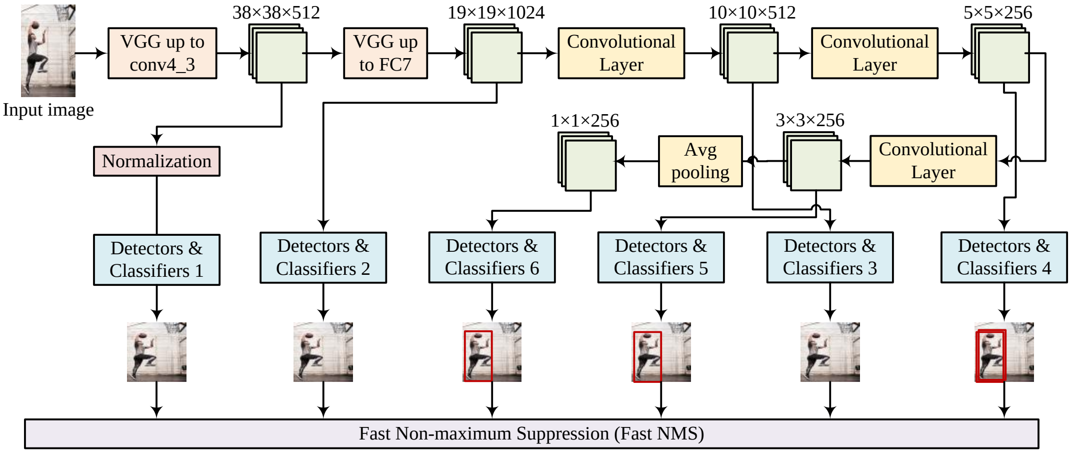
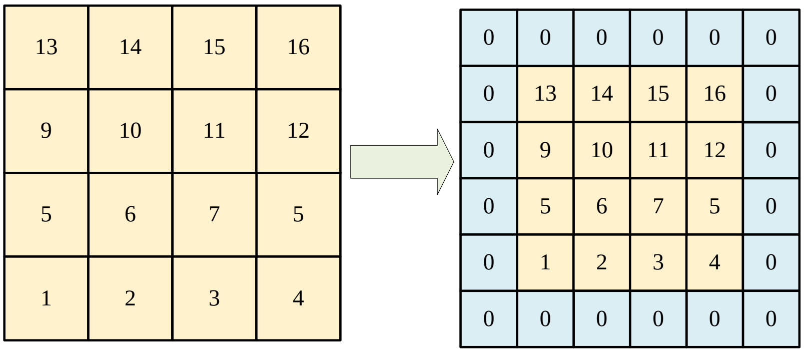

Input image
VGG up to conv4_3
38×38×512
VGG up to FC7
19×19×1024
Convolutional Layer
10×10×512
Convolutional Layer
5×5×256
1×1×256
Avg pooling
3×3×256
Convolutional Layer
Normalization
Detectors & Classifiers 1
Detectors & Classifiers 2
Detectors & Classifiers 6
Detectors & Classifiers 5
Detectors & Classifiers 3
Detectors & Classifiers 4
Fast Non-maximum Suppression (Fast NMS)

Figure 1: The framework of the SSDA

In Figure 1, the base network of the SSDA is VGG 16, and it is modified. Firstly, the fully connected layers FC6 and FC7 of VGG 16 are replaced with convolutional layers (CLs) of 3×3 and 1×1, then the Dropout and FC8 layers are deleted, and then the Pool 5 of the pooling layer is changed to 3×3 with stride=5. At the same time, in order to obtain a denser score mapping, the experiment also includes the Atrous algorithm, and more CLs are also added to VGG 16 in order to increase the FMs. The additional network is then a CNN network with gradually decreasing scales, which can be used to detect targets of different scales [18-19]. In the SSDA, the VGG16 and the additional CLs are responsible for the FE: first, the size of the IM is $300\times300\times3$, then the convolutional kernel (CK) of Conv 1 is initialized and two convolutional operations are performed, and the CK and activation function for the convolutional operations are the $[3,3]$ and the ReLU functions, respectively. Since the convolution operation results in the loss of image boundary information, in order to preserve the boundary features, it is necessary to set the Padding parameter to same, which indicates that the image boundary is 0 [20]. The resulting features are subjected to the maximum pooling process following the convolution. Then, Conv 2 to 5 performs the identical action. It is important to note that Convolution 3 to 5 require three convolution computations total. Taking padding=1 as an example, the padding operation is shown in Figure 2.



13	14	15	16
9	10	11	12
5	6	7	5
1	2	3	4

0	0	0	0	0	0
0	13	14	15	16	0
0	9	10	11	12	0
0	5	6	7	5	0
0	1	2	3	4	0
0	0	0	0	0	0

Figure 2: Padding operation diagram

In Figure 2, after many convolutions, the size of the output picture will keep decreasing. In order to avoid the size of the picture becoming smaller after convolution, padding is applied to the periphery of the picture. When padding=1, the size of padding is 1, the value of padding is 0, and the output size changes from the original 4×4 to 6×6. Since the SSDA changes FC6 and FC7 of VGG 16 into CLs and changes the stride of the pooling layer, Pool 5, in order to cope with this change, Atrous algorithm, which performs the null convolutional operation, is introduced in VGG 16, and its The two CKs for convolution computation are $[1,1]$ and $[3,3]$, respectively. Figure 3 displays the schematic diagram of the void convolution.

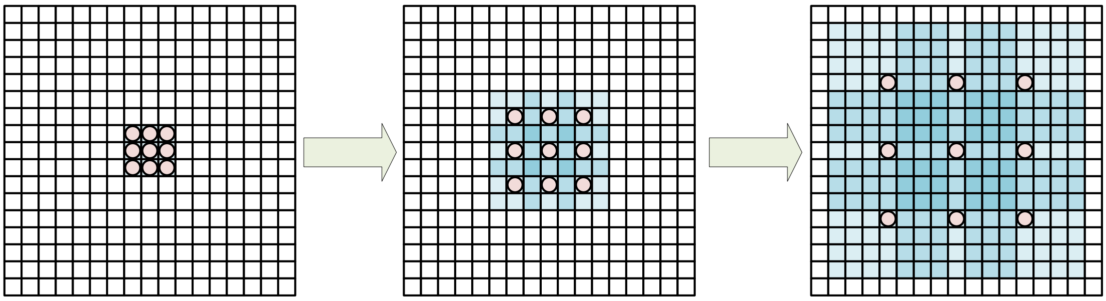


Figure 3: A Schematic representation of the atrous convolution

From Figure 3, it can be observed that for the conventional convolution after the feeling field is 3×3, the feeling field can be enlarged by performing the cavity convolution operation on its FM, and the size of the feeling field at this time is 7×7. Continuing to carry out the cavity convolution, the size of the feeling field is enlarged to 15×15. After cavity convolution, the two convolution calculations are carried out by Conv 6, and the CKs are $\left[1,1\right]$ and $\left[3,3\right]$, respectively, and the second convolution calculation is of a step size of two. Then the same operation is carried out by Conv 7 to 9. It is worth noting that Conv 7 to 9 all step to carry out the filling operation. After the above operations, six FMs can be obtained, and the detection results can be obtained according to the FMs. The bounding box's position and the FMs' confidence are included in the detection value, which is obtained using the $\left[3,3\right]$ convolution calculation. While the number of CKs needed for the bounding box location is four times the number of a priori frames, the CKs needed for classification in the prediction stage are calculated as the product of the number of a priori frames needed for the FM and the classification category [21-22]. Since each target of the FM corresponds to a number of default frames, whereas for the targets in the PFs a convolution operation is performed to obtain confidence and bounding box location information, and the true values are matched with the default frames to obtain negative and positive samples. The intersection over union (IoU) formula on the concatenation of predicted and true frames is given in Equation (1).

$$IoU=\frac{\left|A\cap B\right|}{\left|A\cup B\right|} \tag{1}$$

In Equation (1), A and B denote the PF and the true frame, respectively. As the number of CNN layers increases, the abstraction level of the extracted image features increases gradually. Whereas, since neurons are locally aware and locally connected, the underlying FM retains a large amount of detail information. When the multi-scale FMs of the SSDA are matched with the PFs, they become equal to distinct perceptual fields, which improves the SSDA's recognition capability. The PF scale calculation formula is shown in Equation (2).

$$s_k=s_{\min}+\frac{s_{\max}-s_{\min}}{m-1}\left(k-1\right),k\in\left[1,m\right] \tag{2}$$

In Equation (2), s_k denotes the ratio of the a priori frame size relative to the image, $s_{\max}$ and $s_{\min}$ denote the maximum and minimum values of the ratio, respectively, which are generally 0.2 and 0.9. m denotes the number of FMs and k denotes the number of a priori frames. It is worth noting that the a priori box scale obeys the linear increment rule, i.e., as the size of the feature image decreases, the a priori box scale increases linearly. Equation (3) displays the default box (DB) height calculating formula.

$$h_k^{a_r}=\frac{s_k}{\sqrt{a_r}} \tag{3}$$

In Equation (3), $h_k^{a_r}$ is the height of the DB, a_r represents the aspect ratio of the prediction box (PB), and its value set is $\left\{1,2,3,\frac{1}{2},\frac{1}{3}\right\}$. The formula for calculating the width of the DB is shown in Equation (4)

$$w_k^{a_r}=s_k\sqrt{a_r} \tag{4}$$

In Equation (4), $w_k^{a_r}$ represents the width of the DB. The formula for calculating the center of the PB is shown in Equation (5).

$$\left[\frac{i+0.5}{\left|f_k\right|},\frac{j+0.5}{\left|f_k\right|}\right] \tag{5}$$

In Equation (5), $\left|f_k\right|$ denotes the size of the FM. Since more DBs increase the time required for training and computational complexity, some of the CLs remove the DBs with aspect ratios of 3 and 1/3. Moreover, the SSDA also eliminates the DBs where the recognized object is located outside the box and the PBs where the confidence is less than the set prediction, this operation is realized by the non-maximal value suppression algorithm.

3.2 3D-CNN-based video action recognition algorithm

After cropping the video by the SSDA, the motion region can be selected, which reduces the difficulty for subsequent AR. When performing BTA recognition, since the input is a sequence of video frames. Therefore, not only the spatial representation of the motion but also the temporal order of the motion needs to be considered. The 3D-CNN increases the temporal dimension compared to the traditional 2D-CNN, which makes the 3D-CNN effectively deal with the temporal order of actions. The 3D convolution is shown in Figure 4.

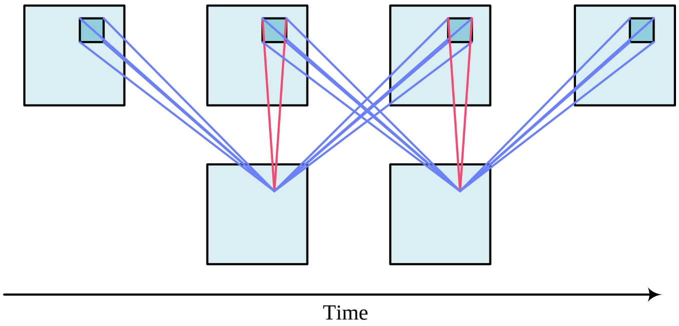


Figure 4: 3D convolution schematic diagram

In Figure 4, 3D convolution can produce multiple FMs by sharing a CK, and each FM contains time dimension information between them. The CKs in the time dimension are encoded with different colors and the same colors share weights. FE is achieved by applying the same 3D CK to the overlapping 3D cubes in the input video. The 3D convolution formula is given in Equation (6).

$$v_{lij}^{xyz} = f\left(\sum_{h=0}^{H_i-1} \sum_{w=0}^{W_i-1} \sum_{s=0}^{S_i-1} k_{ijm}^{hws} V_{(i-1)j}^{(x+h)(y+w)(z+s)} + b_{ij} \right) \quad (6)$$

In Equation (6), i and j denote the feature blocks in the previous layer and the CKs in the current layer, respectively. l denotes the time and v_{lji}^{xyz} denotes the 3D convolution result. H_i and W_i denote the height and width of the CK, respectively, and m denotes the index of the FM connected to the current layer. k_{ijm}^{hws} denotes the value of the CK at (h,w), and S_i denotes the size of the CK in the spectral dimension. b_{ij} denotes the offset, and $V_{(i-1)j}^{(x+h)(y+w)(z+s)}$ denotes the convolution result of the previous layer. Since the performance of CNN is affected by factors such as hyperparameters and algorithmic architecture, the study improves the dual-resolution 3D-CNN in order to reduce the training time while ensuring the training quality. The so-called dual-resolution 3D-CNN refers to processing the same image at different resolutions and then fusing its features to improve the AR capability. However, since the image is composed of numerous pixel points, the amount of data will be larger when it is directly input, resulting in longer training time [23-24]. The color information of the image is not very useful in AR, so the image needs to be grayscaled to reduce the amount of data, and also to avoid the interference of the background, people's clothing and light. The 3D-CNN architecture is shown in Figure 5.

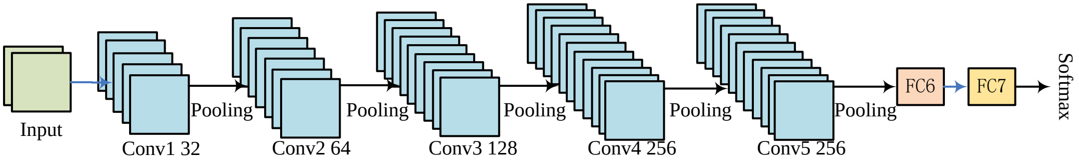


Figure 5: Architecture of the 3D-CNN

In Figure 5, the original frame (OF) 3D-CNN in AR consists of five CLs, FC6 and FC7. The IM can get the final result by Softmax after five consecutive convolution and pooling operations and then two FCs. The CK size and step size of the CLs are $[3,3,3]$ and $[1,1,1]$, respectively, the pooling window and step size of the first layer are $[2,2,1]$, and the rest of the pooling windows and steps are $[2,2,2]$. In addition, crop frame (CF)-3D-CNN differs from OF-3D-CNN in that it consists of 4 CLs, FC5 and FC6, and all pooling windows and steps are $[2,2,2]$. In addition, to further improve the training efficiency, the 2D weight parameters need to be utilized to initialize the 3D convolutional weight parameters [25-26]. Due to the high background similarity of consecutive frame images, the 2D weight matrix needs to be utilized to initialize the 3D weight matrix, which is calculated in Equation (7).

$$W_t^{3D} = \frac{W_t^{2D}}{T} \tag{7}$$

In Equation (7), W_t^{3D} and W^{2D} denote the 3D weight matrix and 2D weight matrix, respectively, and T denotes the timing information. In addition, in order to get different 3D weight matrices, it is necessary to initialize their scaling, which is calculated in Equation (8).

$$W_t^{3D} = \alpha_t W_t^{2D}, (\alpha_t > 0, \sum_{t=1}^{T} \alpha_t = 1) \tag{8}$$

In Equation (8), α_t denotes a random constant. Negative weights initialization is also required to set the values of the sub-matrices of the 3D weight matrix. The negative weights initialization formula is shown in Equation (9).

$$\begin{cases} W_t^{3D} = \alpha W_t^{2D} \\ \alpha_t = \begin{cases} \dfrac{2T-1}{T} & t=1 \\ \dfrac{1}{T} & 2 \le t \le T \end{cases} \end{cases} \tag{9}$$

After the OF image and CF image are processed by the dual-resolution 3D-CNN, the corresponding weight files will be obtained, which will be used as the model parameters for frame sequence prediction to obtain the feature vectors. At this time, the feature vectors need to be fused to facilitate the classification and recognition of features. The feature fusion calculation formula is shown in Equation (10).

$$Y = X_{p,q,d}^{a} + X_{p,q,d}^{b} \tag{10}$$

In Equation (10), Y denotes the final feature representation. $X_{p,q,d}^{a}$ denotes the feature of OF image and $X_{p,q,d}^{b}$ denotes the feature of CF image. p denotes the feature image height and q denotes the feature image width. The final feature representation is used as an input to SVM to perform action classification recognition. SVM is chosen for AR because it can effectively deal with high-dimensional data and linear indivisibility, and it is based on the Vapnik-Chervonenkis dimension principle and resultant risk minimization [27-28]. The schematic diagram of SVM to find the risk minimization cutoff is shown in Figure 6.

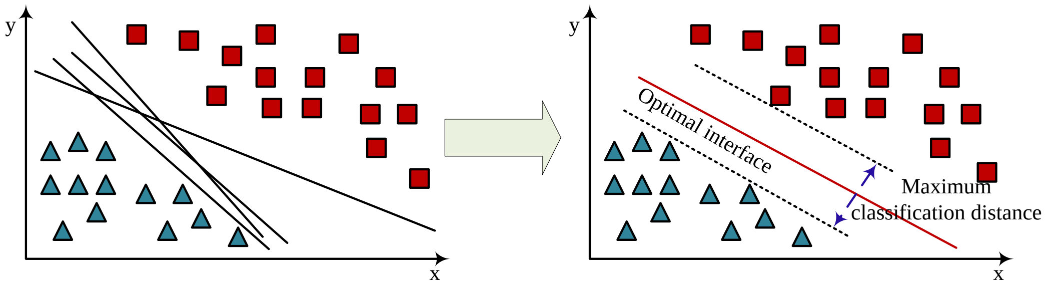


Figure 6: Schematic diagram of SVM searching for the minimum risk boundary

In Figure 6, in the two-dimensional data space, there exist numerous straight lines that can classify the data. However, when the classification straight line is too close to a certain sample, its sensitivity to the noise signal is high, resulting in a weak generalization ability, so the classification effect of the straight line is not optimal. At this time, SVM can find the optimal dividing line by minimizing the distance between the training samples and the classification straight line. Figure 7 displays the schematic diagram used by SVM to classify linearly indivisible data.

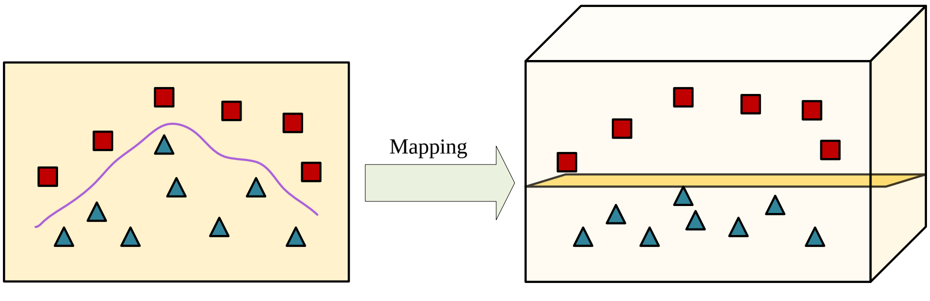


Figure 7: Schematic diagram of linear indivisible data classification

When faced with linearly indivisible data, as in Figure 7, SVM will use the kernel function to map the low-dimensional linearly indivisible data into the high-dimensional space, creating high-dimensional linearly divisible data. Then the optimal classification hypersurface is obtained in the high-dimensional space by linearly differentiable method [29-30]. In addition, when training the 3D-CNN model, a suitable loss function is needed to realize the weight update. The update strategy is mini-batch and the loss function is shown in Equation (11).

$$loss = -\frac{1}{u}\sum_{i=1}^{u}\left[x_i \log\left(z_i\right)+\left(1-x_i\right)\log\left(1-z_i\right)\right] \tag{11}$$

In Equation (11), u denotes the size of the mini-batch, and x_i and z_i denote the predicted and true values of the i th sample in each batch, respectively. The gradient descent formula is shown in Equation (12).

$$w_t = w_{t-1} - \eta\nabla_w J\left(w\right) \tag{12}$$

In Equation (12), w_t denotes the weights, η denotes the learning rate, and $J\left(w\right)$ denotes the cost function. Additionally, the study uses the Adam optimizer to shorten the training period in order to guarantee the training rate. At this time, the weight calculation formula is shown in Equation (13).

$$\begin{cases} w_t = w_{t-1} - \eta\hat{c}\varnothing\sqrt{\hat{s}+\text{э}} \\ c = \beta_1 c + \left(1-\beta_1\right)\nabla_w J\left(w\right) \end{cases} \tag{13}$$

In Equation (13), $\hat{c}$ denotes the correction of the momentum vector and $\hat{e}$ denotes the correction of the gradient-squared accumulation vector. c denotes the momentum vector and β_1 denotes the decaying momentum hyperparameter. The gradient-squared cumulative vector formula is shown in Equation (14).

$$e = \beta_2 e + \left(1-\beta_2\right)\nabla_w J\left(w\right)\otimes\nabla_w J\left(w\right) \tag{14}$$

In Equation (14), e denotes the gradient squared accumulation vector and β_2 denotes the scaling decay hyperparameter. The formulas for the correction of the momentum vector and the gradient-squared accumulation vector are given in Equation (15).

$$\begin{cases} \hat{c} = \dfrac{c}{1-\beta_1^t} \\ \hat{e} = \dfrac{e}{1-\beta_2^t} \end{cases} \tag{15}$$

Since the initial values of the momentum vector and the gradient-squared accumulation vector are both 0, the correction facilitates the improvement of both during the initial phase of training.

4 Video action recognition result analysis

Among basketball enthusiasts, basketball instructional videos and basketball jams are popular and widely studied and imitated. To enhance the technical action learning method available to basketball enthusiasts, the study suggests a video AR approach that utilizes 3D-CNN and SSD. The study will evaluate the CF generation of SSDA and the AR performance of 3D-CNN in order to assess the recognition performance of the approach. The datasets for the COCO and VOC, which contain 328,000 images with 2,500,000 instances labeled and more than 5,000 examples classified in 82 out of 91 categories, respectively, will be used for the SSDA tests. There are twenty-one categories in the VOC dataset, including bicycles, dogs, and cats. Since the SSDA's, goal is to recognize people in images, irrelevant images will be ignored. The test for 3D-CNN will be performed on the homemade BTA dataset, which contains 1800 videos, 300 videos for each of the change of direction, tonbay, turn, Eurostep, Sam Gaudet, and sharp stop and jump shot movements, and the length of each video is from 0.5 s to 1.5 s. Fifty videos will be selected as the validation set for each category of movements in the BTA dataset, and the rest of the videos will be used as the training set. In the experiments, the deep learning framework is Keras, the programming language is Python 3.7, the batchsize and Epoch are 32 and 50, respectively, and the image input frame is 16 frames. The SSDA employs a confidence threshold of 0.9, a ratio of the prediction box minimum to original size of 0.2, a prediction box maximum to original size of 0.9, and an IoU threshold of 0.5. The original images have a resolution of 112112 pixels, the cropped frames have a resolution of 6464 pixels, and the number of output units in the fully connected layer is 1024. The recognition results of SSDA on COCO dataset are shown in Figure 8.

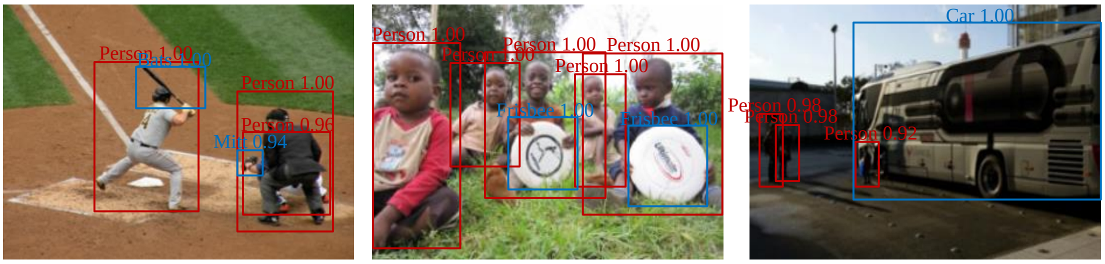


Figure 8: Identification results of the SSDA on the COCO dataset

In Figure 8, the SSDA accurately recognizes the people in the pictures in all COCO datasets and marks their confidence levels. Meanwhile, for non-human objects, the SSDA marks them with other colors and shows the category they belong to. It can be seen that the SSDA accurately recognizes human beings in the images and marks the non-human objects for the purpose of recognizing only human bodies. The result of CF generation based on SSDA is shown in Figure 9.

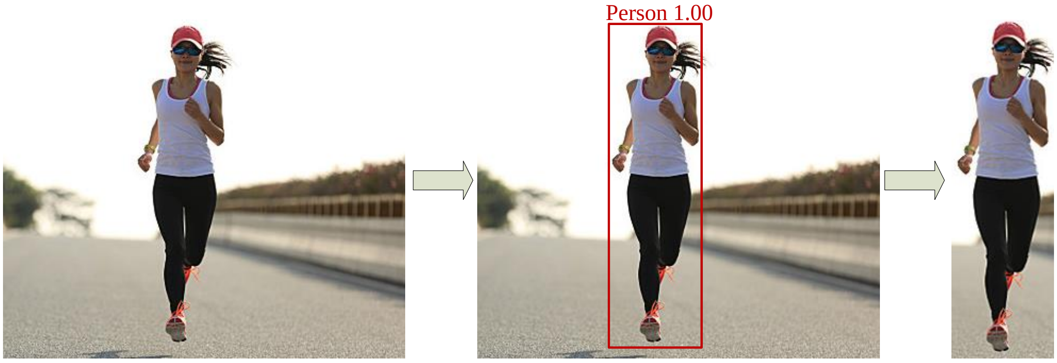


Figure 9: The crop frame generation results based on the SSDA

The human body in the picture is identified in Figure 9 following the SSDA's TD on the OF, and its confidence level is 1.00. Then after OpenCV obtains the four coordinates of the detection frame, it can be clipped on the OF in order to generate the CF. As an example, the CF generation result of the sharp stop jump shot in BTA is shown in Figure 10.

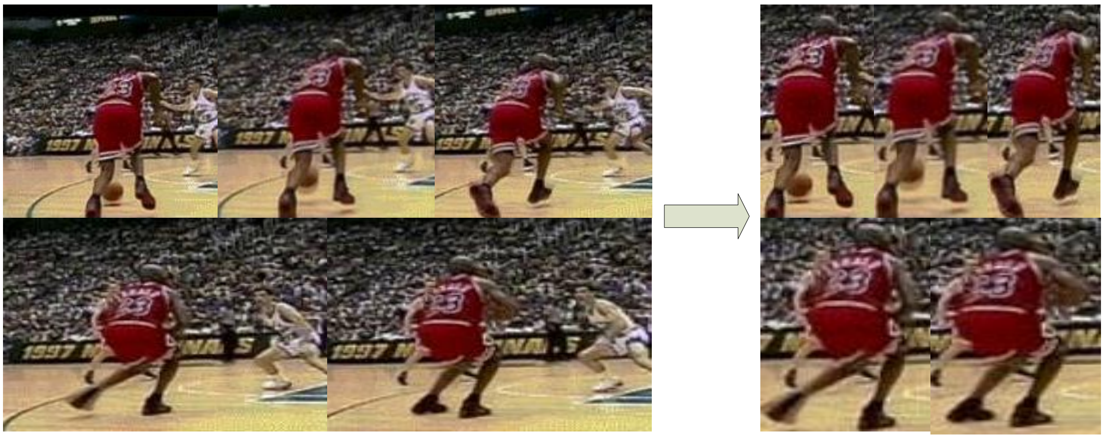
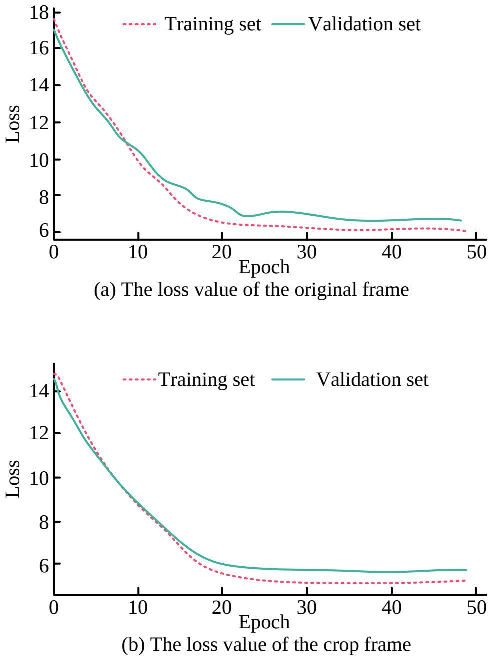


Figure 10: Cropped frame of stop jumper

In Figure 10, the OF is processed by the SSDA to obtain the human body in the detection frame of the continuous frame image of the video, and in order to avoid the interference of the spectators and other players, the confidence threshold is adjusted up to 0.95 to obtain the CF image of the target human body. Since the size of CF is inconsistent, it will be preprocessed first to ensure the consistent size when performing the subsequent AR with dual-resolution 3D-CNN. The OF and CF loss values of the dual-resolution 3D-CNN are shown in Figure 11.

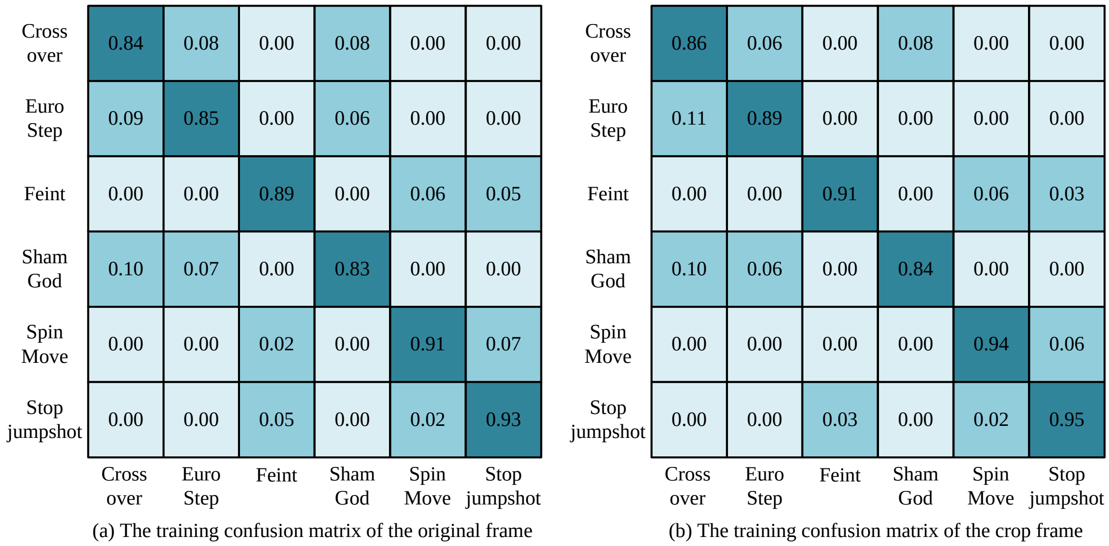


(a) The loss value of the original frame

(b) The loss value of the crop frame

Figure 11: Raw frame and trimmed frame loss values

From Figure 11(a), the loss value of OF on the training set starts to converge at epoch of 19, with a loss value of about 6.0. On the validation set, the loss value of OF starts to converge at epoch of 21, with a loss value of about 6.8. From Figure 11(a), the loss value of CF on both the training and validation sets starts to converge at epoch of 20, with loss values of 5.1 and 5.9. The above results show that 3D-CNN converges better for CF, but the convergence speeds of OF and CF are not much different. The training confusion matrices of OF and CF are shown in Figure 12.



	Cross over	Euro Step	Feint	Sham God	Spin Move	Stop jumpshot
Cross over	0.84	0.08	0.00	0.08	0.00	0.00
Euro Step	0.09	0.85	0.00	0.06	0.00	0.00
Feint	0.00	0.00	0.89	0.00	0.06	0.05
Sham God	0.10	0.07	0.00	0.83	0.00	0.00
Spin Move	0.00	0.00	0.02	0.00	0.91	0.07
Stop jumpshot	0.00	0.00	0.05	0.00	0.02	0.93

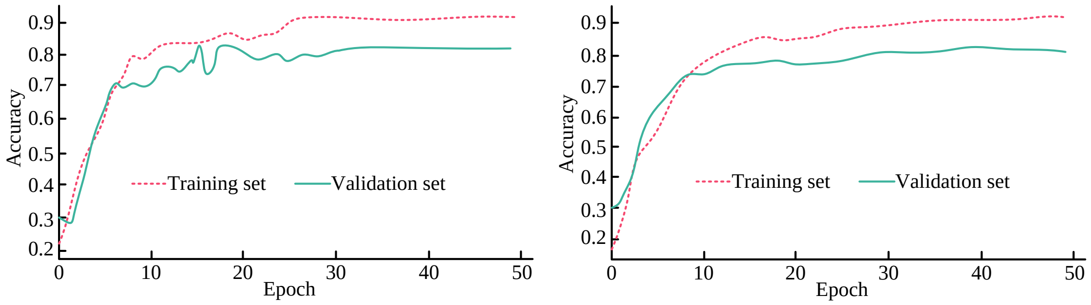
(a) The training confusion matrix of the original frame

	Cross over	Euro Step	Feint	Sham God	Spin Move	Stop jumpshot
Cross over	0.86	0.06	0.00	0.08	0.00	0.00
Euro Step	0.11	0.89	0.00	0.00	0.00	0.00
Feint	0.00	0.00	0.91	0.00	0.06	0.03
Sham God	0.10	0.06	0.00	0.84	0.00	0.00
Spin Move	0.00	0.00	0.00	0.00	0.94	0.06
Stop jumpshot	0.00	0.00	0.03	0.00	0.02	0.95

(b) The training confusion matrix of the crop frame

Figure 12: The training confusion matrix of raw frames and cropped frames

In Figure 12(a), among the ARs of OF, the highest classification accuracy is achieved for the hasty jump shot action, which has an average recognition rate of about 88.3%, but there is still the possibility of being misclassified as a turnaround or tonbay. In addition, there are cases of misclassification for the recognition of Eurostep, change of direction and Samgold movements, which may be due to the fact that all the above three movements are characterized by change of direction. In Figure 12(b), in the AR of CF, there are still misjudgment cases for the recognition of European step, change of direction and Samgold action, but their misjudgment probabilities have decreased. In addition, the classification accuracy for the sharp stop jump shot action is still the highest, and its average recognition rate is about 90.3%. It can be seen that the accuracy of AR has been improved after recognition by the SSDA and removal of interfering factors. The recognition accuracy of OF and CF are shown in Figure 13.



(a) The accuracy of the original frame

(b) The accuracy of the crop frame

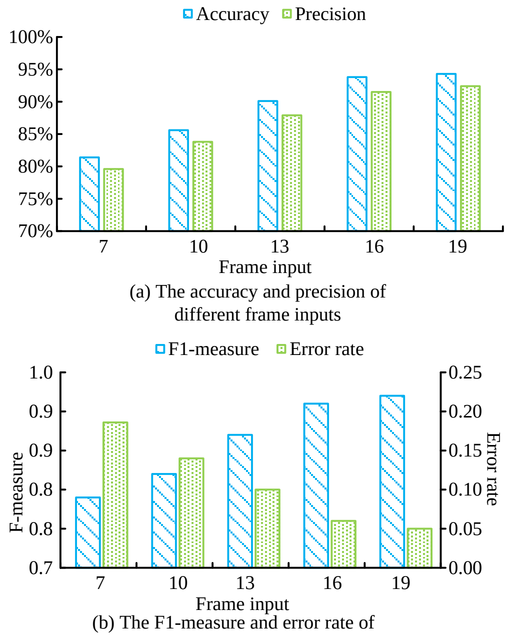
Figure 13: Recognition accuracy of raw frames and cropped frames

In Figure 13(a), the recognition accuracy (RA) of OF in both the training set and the validation set rises with the increase of epoch, and both reach a stable state when the epoch is 30, at which time its accuracy is about 0.91 and 0.81, respectively, but the RA of the test set fluctuates a lot in the interval of epoch 10 to 20. In Figure 13(b), the RA of the CF training set and the test set is basically the same with the trend of epoch, and also reaches a stable state when the epoch is 30, at which time the RA is about 0.92 and 0.81, respectively. After the training of OF and CF, the obtained weight file is used as a parameter for the recognition of consecutive frames, and the OF and CF features are fused and then classified by SVM to achieve continuous frame recognition. SVM classification can realize AR for continuous frames. The confusion matrix for feature fusion is shown in Figure 14.

	Cross over	Euro Step	Feint	Sham God	Spin Move	Stop jumpshot
Cross over	0.91	0.02	0.00	0.07	0.00	0.00
Euro Step	0.05	0.93	0.00	0.02	0.00	0.00
Feint	0.00	0.00	0.94	0.00	0.03	0.03
Sham God	0.05	0.02	0.00	0.93	0.00	0.00
Spin Move	0.00	0.00	0.02	0.00	0.95	0.03
Stop jumpshot	0.00	0.00	0.03	0.00	0.00	0.97

Figure 14: Confusion matrix of feature fusion

In Figure 14, compared with the single-resolution recognition of OF and CF, the RA of each BTA after feature fusion has increased, and the sharp jump shot is basically not misjudged as a turn, and the probability of misjudgment of the European step, change of direction and Sam Gaudet's action has decreased significantly. The average recognition rate after dual-resolution feature fusion is about 94.6%, which is significantly higher than single-resolution recognition. The above results show that dual-resolution recognition has obvious advantages over single-resolution recognition. Figure 15 displays the dual-resolution 3D-CNN's recognition performance under various frame inputs.



(a) The accuracy and precision of different frame inputs

(b) The F1-measure and error rate of different frame inputs

Figure 15: Recognition accuracy and precision under different frame inputs

In Figure 15(a), the RA and precision rate gradually increase with the increase of frame input. When the frame input is 7, the accuracy and precision rate are 81.4% and 79.6%, respectively. When the frame input is 10, the accuracy and precision are 85.6% and 83.8%, respectively. In addition, when the frame input is 16, the accuracy and precision rate increase to 93.8% and 91.5%, respectively. In Figure 15(b), as the frame input increases, the F1-score gradually rises while the misclassification rate gradually decreases. When the frame input is 7, the F1-score and the misjudgment rate are 0.79 and 0.19, respectively. When the frame input is 10, the F1-score and the misjudgment rate are 0.82 and 0.14, respectively. In addition, when the frame input is 16, the F1-score and the misjudgment rate are 0.92 and 0.05, respectively. The reason why the larger the frame input is, the better the performance of AR is because it is difficult to recognize the AR when the frame number is small, it is difficult to recognize the corresponding timing information, so its recognition performance for continuous frames is poor. However, a high frame number also leads to an increase in training time and training volume, which shows that the appropriate frame input has a greater impact on the AR performance for continuous frames. The F1-score and recall of the dual-resolution 3D-CNN with different frame inputs are shown in Figure 16.

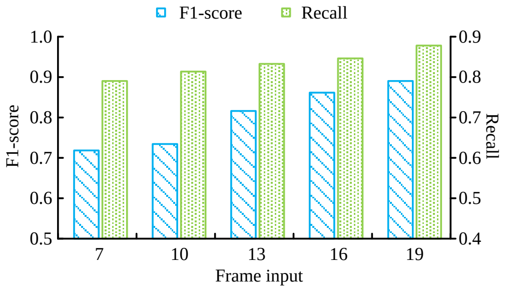


Figure 16: The F1-score and recall of the dual-resolution 3D-CNN with different frame inputs

Figure 16 illustrates that the F1-score and recall of the dual-resolution 3D-CNN increase with the number of input frames. The F1-score and recall are 0.70 and 0.80, respectively, when the input frame is 13. When the input frame is 14, the F1-score and recall are 0.81 and 0.85, respectively. The results of the ablation experiments of the BTA recognition model based on the SSDA and dual-resolution 3D-CNN are presented in Table 2.

Table 2: Results of ablation experiments of basketball technical action recognition model

SSD	Dual-resolution 3D-CNN	Accuracy
×	×	80.5%
×	√	84.3%
√	×	81.7%
√	√	89.2%

Table 2 indicates that in the absence of trimmed frames generated by the SSDA, the average RA of BTA is 80.5% for the ordinary 3D-CNN and 84.3% for the double-resolution 3D-CNN, respectively. In the case of cropped frames generated by the SSDA, the average RA of the ordinary 3D-CNN is 81.7% and 89.2% for the 3D-CNN, respectively. The preceding results demonstrate that the SSDA and dual-resolution 3D-CNN are effective in enhancing the RA of actions.

5 Discussion

In order to improve the efficiency of video analysis, a video cropping and action recognition method based on SSDA and 3D-CNN was proposed. The average recognition rate of this algorithm was about 94.6%, and the RA and precision gradually improved with the increase of frame input. A 3D-CNN pose estimation method based on an end-to-end hierarchical model and physical constraints was proposed by Xu and Zhang [11]. This method enabled gesture estimation by reconstructing the 3D spatial structure of hands using deep images and converting the hand model into a voxel mesh via 3 D-CNN. Nevertheless, although the method can effectively utilize deep images, it is challenging to accurately identify objects in complex environments. In comparison to the research of Xu and Zhang, the proposed method exhibited a higher degree of RA. The reason for the enhanced RA was that the research team opted to utilize the SSD object detection algorithm to generate a basketball technique action cropping frame dataset. The SSD detection algorithm is employed to process the original frame images of BTA videos, thereby enabling the identification of the human body within the detection box of each frame image. To prevent the inclusion of background elements that might otherwise interfere with the identification of the human subject, the confidence threshold is increased to 0.9, thereby allowing for the capture of the final human image. A hyperspectral image classification model based on a hybrid spectral convolutional neural network was proposed by Roy et al. [15]. The model employs 3D-CNN for joint spatial spectral feature representation, subsequently acquiring more advanced spatial representation through 2D-CNN. However, the model necessitates the resolution of the image, which is unable to process images with disparate resolutions. The dual-resolution 3D-CNN model is capable of handling continuous technical action frames at different input resolutions and fusing features at different resolutions to enhance recognition capabilities.

6 Conclusion

Compared with other sports, basketball has significant technical movement characteristics. Therefore, in basketball teaching, coaches often teach technical movements through video analysis. In addition, in basketball leagues, analyzing the technical characteristics of opposing athletes through game videos can also support the formulation of tactics. In view of this, in order to improve the efficiency of video analysis, the study proposes a video cropping and AR method based on SSDA and 3D-CNN. The experimental results revealed that the SSDA accurately recognizes the human body in the image and labels its confidence level. While for non-human objects other color criteria were used and that detection was abolished. While in basketball AR, the loss values of OF on the training and validation set start to converge at epochs of 19 and 21, with loss values of 6.0 and 6.8, respectively. The loss values of CF on both the training and validation set start to converge at epoch of 20, with loss values of 5.1 and 5.9, respectively. The 3D-CNN achieved the highest classification accuracy of the sharp stop jumper action in the OF of about 88.3%. In addition, due to the fact that the European step, change of direction and Samgold action all had change of direction features, which led to its misclassification situation. The recognition of CF by 3D-CNN was similar to that of OF, but its average recognition rate was 90.3%, which was higher than that of OF. The RA of OF and CF both epoch increased, and both reached a stable state when epoch was 30. The accuracy of the OF training set and the test set were 0.91 and 0.81, respectively, and the RA of the CF training set and the test set were about 0.92 and 0.81, respectively, with a small difference between the two. In addition, after fusing OF features and CF features, the average recognition rate was about 94.6%, which was significantly higher than the single resolution recognition. In addition, with the increase of frame input, the RA and precision rate gradually increased. When the frame input was 7, the accuracy and precision rate were 81.4% and 79.6%, respectively. When the frame input was 16, the accuracy and precision increased to 93.8% and 91.5%, respectively. The above results reveal that the BTA recognition performance for continuous frames is significantly improved after the SSDA recognizes and removes interfering factors and performs feature fusion. Although the study achieved some results in BTA recognition of continuous frames, the recognition performance of the proposed AR method in the case of occluded characters is doubtful because the basketball technology videos used are all unobscured videos. In view of this, future research will focus on how to improve the AR accuracy for occluded characters. In addition, how to apply the AR algorithm proposed by the study to short video platforms or web-side is also a worthwhile research direction.

References

[1] Z. Guo, Y. Hou, R. Xiao, C. Li, and W. Li, “Motion saliency based hierarchical attention network for action recognition,” Multimedia Tools and Applications, vol. 82, no. 3, pp. 4533-4550, 2022. https://doi.org/10.1007/s11042-022-13441-7

[2] G. V. Reddy, K. Deepika, L. Malliga, D. Hemanand, C. Senthilkumar, and S. Gopalakrishnan, “Human action recognition using difference of gaussian and difference of wavelet,” Big Data Mining and Analytics, vol. 6, no. 3, pp. 336-346, 2023. https://doi.org/10.26599/BDMA.2022.9020040

[3] L. Liu, L. Yang, W. Chen, and X. Gao, “Dual-view 3D human pose estimation without camera parameters for action recognition,” IET Image Processing, vol. 15, no. 14, pp. 3433-3440, 2021. https://doi.org/10.1049/ipr2.12277

[4] G. Zhang, Y. Rao, C. Wang, W. Zhou, and X. Ji, “A deep learning method for video-based action recognition,” IET Image Processing, vol. 15, no. 14, pp. 3498-3511, 2021. https://doi.org/org/10.1049/ipr2.12303

[5] Q. Men, E. S. L. Ho, H. P. H. Shum, and H. Leung, “Focalized contrastive view-invariant learning for self-supervised skeleton-based action recognition,” Neurocomputing, vol. 537, no. 7, pp. 198-209, 2023. https://doi.org/10.1016/j.neucom.2023.03.070

[6] Y. Liu, R. Liu, S. Wang, D. Yan, B. Peng, and T. Zhang, “Video face detection based on improved ssd model and target tracking algorithm,” Journal of Web Engineering, vol. 21, no. 2, pp. 545-567, 2022. https://doi.org/10.13052/jwe1540-9589.21218

[7] M. Zhao, Y. Zhong, D. Sun, and Y. Chen, “Accurate and efficient vehicle detection framework based on ssd algorithm,” IET Image Processing, vol. 15, no. 13, pp. 3094-3104, 2021. https://doi.org/10.1049/ipr2.12297

[8] D. Liu, S. Gao, W. Chi, and D. Fan, “Pedestrian detection algorithm based on improved ssd,” International Journal of Computer Applications in Technology, vol. 65, no. 1, pp. 25-35, 2021. https://doi.org/10.1504/ij cat.2021.199965999996

[9] Y. Li, D. Han, H. Li, X. Zhang, B. Zhang, and Z. Xiao, “Multi-block SSD based on small object detection for UAV railway scene surveillance,” Chinese Journal of Aeronautics, vol. 33, no. 6, pp. 1747-1755, 2020. https://doi.org/10.1016/j.cja.2020.02.024

[10] Z. Xu, Z. Wu, and W. Fan, “Improved SSD-assisted algorithm for surface defect detection of electromagnetic luminescence,” Proceedings of the Institution of Mechanical Engineers, vol. 235, no. 5, pp. 761-768, 2021. https://doi.org/1748006X2199538

[11] Z. Xu, and W. Zhang, “3D CNN hand pose estimation with end-to-end hierarchical model and physical constraints from depth images,” Neural Network World, vol. 33, no. 1, pp. 35-48, 2023. https://doi.org/10.14311/NNW.2023.33.003

[12] A. Rehman, M. A. Khan, T. Saba, Z. Mehmood, and N. Ayesha, “Microscopic brain tumor detection and classification using 3D CNN and feature selection architecture,” Microscopy Research and Technique, vol. 84, no. 1, pp. 133-149, 2020. https://doi.org/10.1002/jemt.23597

[13] V. Pareek, S. Chaudhury, and S. Singh, “Hybrid 3DCNN-RBM network for gas mixture concentration estimation with sensor array,” IEEE Sensors Journal, vol. 21, no. 21, pp. 24263-24273, 2021. https://doi.org/10.1109/JSEN.2021.3105414

[14] A. Chaddad, P. Sargos, and C. Desrosiers, “Modeling texture in deep 3D CNN for survival analysis,” IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 7, pp. 2454-2452, 2020. https://doi.org/10.1109/JBHI.2020.3025901

[15] S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri, “HybridSN: Exploring 3D-2D CNN feature hierarchy for hyperspectral image classification,” IEEE Geoscience and Remote Sensing Letters, vol. 17, no. 2, pp. 277-281, 2020. https://doi.org/10.1109/LGRS.2019.2918719

[16] W. Chen, Y. Qiao, and Y. Li, “Inception-SSD: An improved single shot detector for vehicle detection,” Journal of Ambient Intelligence and Humanized Computing, vol. 13, no. 1, pp. 5047-5053, 2022. https://doi.org/10.1007/s12652-020-02085-w

[17] Z. Lyu, D. Zhang, and J. Luo, “A GPU-free real-time object detection method for apron surveillance video based on quantized MobileNet-SSD,” IET Image Processing, vol. 16, no. 8, pp. 2196-2209, 2022. https://doi.org/10.1049/ipr2.12483

[18] Q. Huang, Y. Zhang, Y. Huang, C. Mi, Z. Zhang, and W. Mi, “Two-stage container keyhole location algorithm based on optimized SSD and adaptive threshold,” Journal of Computational Methods in Sciences and Engineering, vol. 22, no. 5, pp. 1559-1571, 2022. https://doi.org/10.3233/JCM-226135

[19] W. A. Okaishi, A. Zaarane, I. Slimani, I. Atouf, and M. Benrabh, “A vehicular queue length measurement system in real-time based on SSD network,” Transport and Telecommunication, vol. 22, no. 1, pp. 29-38, 2021. https://doi.org/10.2478/ttj-2021-0003

[20] Y. Pan, M. Lin, Z. Wu, H. Zhang, and Z. Xu, “Caching-aware garbage collection to improve performance and lifetime for nand flash ssds,” IEEE Transactions on Consumer Electronics, vol. 67, no. 2, pp. 141-148, 2021. https://doi.org/10.1109/TCE.2021.3067604

[21] T. E. Trueman, A. K. Jayaraman, S. Jasmine, G. Ananthakrishnan, and P. Narayanasamy, “A Multi-channel convolutional neural network for multilabel sentiment classification using abilify oral user reviews,” Informatica, vol. 47, no. 1, pp. 109-113, 2023. https://doi.org/10.31449/inf.v47i1.3510

[22] L. Yao, and Z. Ge, “Cooperative deep dynamic feature extraction and variable time-delay estimation for industrial quality prediction,” IEEE Transactions on Industrial Informatics, vol. 17, no. 6, pp. 3782-3792, 2021. https://doi.org/10.1109/TII.2020.3021047

[23] K. Bhosle, and V. Musande, “Evaluation of deep learning CNN model for recognition of devanagari digit,” Artificial Intelligence and Applications, vol. 1, no. 2, pp. 114-118, 2023. https://doi.org/10.47852/bonviewAIA3202441

[24] R. A. A. Salvador, and P. C. Naval, “Towards a feasible hand gesture recognition system as sterile non-contact interface in the operating room with 3D convolutional neural networkm,” Informatica, vol. 46, no. 1, pp. 1-12, 2022. https://doi.org/10.31449/inf.v46i1.3442

[25] Y. Li, S. Yang, Y. Zheng, and H. Lu, “Improved point-voxel region convolutional neural network: 3d object detectors for autonomous driving,” IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 7, pp. 9311-9317, 2022. https://doi.org/10.1109/TITS.2021.3071790

[26] K. Zhu, W. Lu, J. Liu, X. Luo, and X. Zhao, “A lightweight 3d convolutional neural network for deepfake detection,” International Journal of Intelligent Systems, vol. 36, no. 9, pp. 4990-5004, 2021. https://doi.org/10.1002/int.22499

[27] D. Kim, M. E. Lipford, H. He, Q. Ding, V. Ivanovic, S. N. Lockhart, S. Craft, C. T. Whitlow, and Y. Jung, “Parametric cerebral blood flow and arterial transit time mapping using a 3D convolutional neural network,” Magnetic Resonance in Medicine, vol. 90, no. 2, pp. 583-595, 2023. https://doi.org/10.1002/mrm.29674

[28] R. Opfer, J. Krueger, L. Spies, A. C. Ostwaldt, H. H. Kitzler, and S. Schippling, and R. Buchert, “Automatic segmentation of the thalamus using a massively trained 3d convolutional neural network: higher sensitivity for the detection of reduced thalamus volume by improved inter-scanner stability,” European Radiology, vol. 33, no. 3, pp. 1852-1861, 2022. https://doi.org/10.1007/s00330-022-09170-y

[29] L. He, B. Ding, H. Wang, and T. Zhang, “An optimal 3d convolutional neural network based lipreading method,” IET Image Processing, vol. 16, no. 1, pp. 113-122, 2021. https://doi.org/10.1049/ipr2.12337

[30] B. Masoudi, S. Daneshvar, and S. N. Razavi, “Multi-modal neuroimaging feature fusion via 3d convolutional neural network architecture for schizophrenia diagnosis,” Intelligent Data Analysis, vol. 25, no. 3, pp. 527-540, 2021. https://doi.org/10.3233/IDA-205113