Hindawi
Scientific Programming
Volume 2021, Article ID 1016574, 8 pages
https://doi.org/10.1155/2021/1016574

Research Article

Key Frame Extraction for Sports Training Based on Improved Deep Learning

Changhai Lv,[1] Junfeng Li,[1] and Jian Tian [2]

[1]*Ministry of Sports, Shandong Technology and Business University, Yantai 264005, China*
[2]*School of Physical Education, Henan University, Kaifeng 475001, Henan, China*

Correspondence should be addressed to Jian Tian; 201513466@sdtbu.edu.cn

Received 20 July 2021; Revised 18 August 2021; Accepted 24 August 2021; Published 2 September 2021

Academic Editor: Muhammad Usman

With the rapid technological advances in sports, the number of athletics increases gradually. For sports professionals, it is obligatory to oversee and explore the athletics pose in athletes' training. Key frame extraction of training videos plays a significant role to ease the analysis of sport training videos. This paper develops a sports actions' classification system for accurately classifying athlete's actions. The key video frames are extracted from the sports training video to highlight the distinct actions in sports training. Subsequently, a fully convolutional network (FCN) is used to extract the region of interest (ROI) pose detection of frames followed by the application of a convolution neural network (CNN) to estimate the pose probability of each frame. Moreover, a distinct key frame extraction approach is established to extract the key frames considering neighboring frames' probability differences. The experimental results determine that the proposed method showed better performance and can recognize the athlete's posture with an average classification rate of 98%. The experimental results and analysis validate that the proposed key frame extraction method outperforms its counterparts in key pose probability estimation and key pose extraction.

1. Introduction

With the advent of artificial intelligence, performance analysis in sport has undergone significant changes in recent years. In general, manual analysis performed by trained sports analysts has some drawbacks such as being time-consuming, subjective in nature, and prone to human errors. Objective measurement and assessment for sports actions are indispensable to understand the physical and technical demands related to sports performance [1]. Intelligent sports action recognition methods are developed to provide objective analysis and evaluation in sport and improve the accuracy of sports performance analysis and validate the efficiency of training programs. Common sports action recognition systems can be developed using advanced machine learning methods to process the data collected via computer vision systems and wearable sensors [2]. Sports activities recorded through a computer vision system can be used for athlete action detection, movement analysis, and pose estimation [3]. The vision-based sports action recognition can provide real-time feedback for athletes and coaches. However, the player's actions in sports videos are more complex and skillful. Compared with daily activities, the analysis of sports videos is more challenging. This is because, the players while playing perform rapid and consistent actions within the camera view, thus degrading the action recognition performance [4].

In sports video analysis and processing for action recognition, pertinent and basic information extraction is a mandatory task. If the video is large, then it is hard to process the whole video in a short time while preserving its semantics [5]. The extraction of the key frame is a prime step of video analysis. The key frame provides eloquent information and is a summary of the entire video sequence [6]. A video is normally recorded 30 frames per second and contains additional information for the recognition of a particular computer vision task. Key frame detection is mainly applied in video summarization and visual localization in videos. To use all the frames of a video, more computational resources and memory are required. In many computer vision

applications, one or few key frames may be enough to accomplish the desired recognition results [3].

The key frames are applied in many applications such as searching, information retrieval, and scene analysis in videos [7]. The video represents a composite structure and is made of several scenes, shots, and several frames. Figure 1 shows the division of video into shots and frames. In many video and image processing tasks, such as scene analysis and sequence summarization, it is essential to perform an analysis of the complete video. During the analysis of videos, the major steps are scene segmentation, detection of shot margin, and key frame extraction [8, 9]. The shot is a contiguous, adjacent combination of frames recorded by a camera. The key objective of extracting key frames is to extract unique frames in a video and prepare the video sequences for quick processing [10]. In this paper, we propose an effective method for the extraction of a key frame from athlete sports video, which is accurate, fast, and efficient. The proposed key frame extraction model uses a long sports action video as input and extracts the key frames, which can better represent the sports action for recognition. We introduced an improved convolution neural network method to detect key frames in athletes' videos. We performed experiments on athletes' training video dataset to show the triumph of our method for key frames' detection.

We structured the rest of the paper as follows. In Section 2, related work is presented. Section 3 provides the detail of the proposed method. Sections 4 and 5 are about the experimental results and conclusion, respectively.

2. Related Work

With the advancement of sports, competition in sports is becoming a base to develop people's social life and emotions. In order to enhance the competitive skills of athletes, active investigation of sports training is one of the central issues. Many previous analysis methods in this field depend on using a segmentation-based approach [11]. These methods usually extract visual features from videos. One of the first attempts discovered local minimum changes within videos concerning similarity between the consecutive frames. Later on, other works augmented this approach by using the key points' detection method for local feature extraction and combining the key points to find the key frames [12]. All of these methods have a common shortcoming of extracting redundant frames rather than fully covering the video contents.

Another group of traditional methods is based on feature clusters and detects the key video frames with prediction of a prominent frame in individual clusters. Zhuang et al. [13] employed the joint entropy (JE) and mutual information (MI) between successive video frames and detected key frames. Tang et al. [14] developed a clustering method for recognizing the key frame using visual content and motion analysis. A frame extraction method for hand gesture images' recognition using image entropy and density clustering was presented in [15]. Cun et al. [16] developed a method for the extraction of key frames using spectral clustering. The feature locality in the video sequence was extracted using a graph as an alternative to relying on a similarity measure shared between two images.

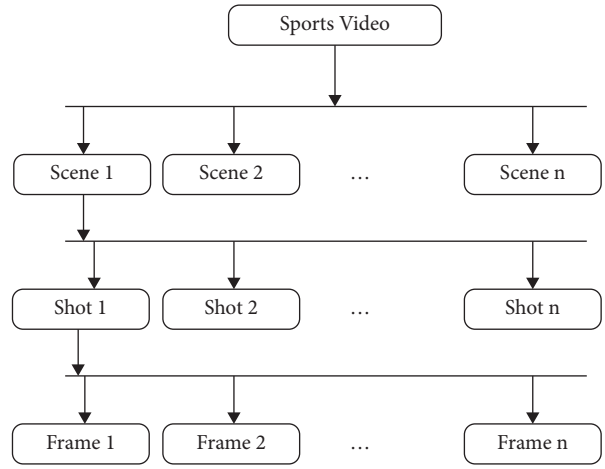


Figure 1: Structure of sport video.

To overcome the shortcomings of traditional frame detection methods, recent works focused on deep learning to perform key frame recognition in videos [17]. Deep learning has made a great breakthrough in the application of speech recognition, vision-based systems, human activity recognition, and image classification [18]. The deep learning models simulate human neurons and form the combination of low- and high-level features, to describe and understand objects [19]. Deep learning is relative to "shallow learning." The major difference between deep learning and "shallow learning" is that the deep model contains several nonlinear operations and more layers of a neural network [20]. "Shallow learning" relies on manual feature extraction and ultimately obtains single-layer features. Deep learning extracts different levels of features from the original signal from shallow to deep. In addition, deep learning can describe learning deeper and more complex features, to better express the image, which is conducive to classification and other tasks. The structure of deep learning is comprised of a large number of neurons, each of which is connected with other neurons. The process of deep learning is to update the weights through continuous iteration. Deep neural networks (DNN) are a deep network structure. The network structure of a deep neural network includes multiple single-layer nonlinear networks. At present, the more common networks can be categorized into feedback deep networks (FBDN), bidirectional deep networks (BDDN) [21], and feedforward deep networks (FFDN).

Different supervised and unsupervised deep learning methods have been suggested for key frame detection in sports videos which considerably enhance the performance of action recognition systems. Yang et al. [22] employed the method of generative adversarial networks for the detection of key frames in videos. For key features' extraction, CNNs were employed to extract the discriminant features which were encoded using long short-term memory (LSTM) networks. Another approach using bidirectional long short-term memory (Bi-LSTM) was introduced in [23]. The method was effective for extracting the highlighting the key

video's frames automatically. Huang and Wang [24] proposed a two-stream CNNs' approach to detect the key frames for action recognition. Likewise, Jian et al. [25] devised a unique key frame and shot selection model for summarization of video. Wen et al. [26] employed a frame extraction system through estimation of the pose probability of each neighboring frame in a sports video. Moreover, Wu et al. [27] presented a video generation approach based on key frames. In this study, we propose an improved key frame extraction technique for sports action recognition using a convolutional neural network. FCN is applied to get the ROI for a more accurate pose detection of frames followed by the application of a CNN to estimate the pose probability of individual frames.

3. Methods

3.1. Overview of CNN. CNN is an artificial neural network that mimics the human brain and can grip the training and learning of layered network structures. CNN uses the local receptive field to acquire autonomous learning capability and handle huge data images for processing. CNN is a specific type of FFDN. It is extensively used for recognition of images. CNN represents image data in the form of multidimensional arrays or matrices. CNN extracts each slice of an input image and assigns weights to each neuron based on the important role of the receptive field. Simultaneous interpretation of weight points and pooling functions reduces the dimension of image features, reduces the complexity of parameter adjustment, and improves the stability of network structure. Lastly, prominent features are generated for classification, so they are broadly used for object detection and classification of images.

CNN is primarily comprised of the input layer, convolution layer, pooling layer, full connection layer, and output layer. The input image is given to the input layer for processing. The convolution layer performs convolution operation over the input matrix between the input layer and convolution layer, and the input image is processed for feature extraction. The function of the pooling layer is to take the maximum value of the pixels in the target area of the input image, to condense the resolution of the feature image and avoid overfitting. The full connection layer is composed of zero or more neurons. Each neuron is linked with all the neurons in the preceding layer. The obtained feature vector is mapped to the output layer to facilitate classification. The function of the output layer is to classify feature vectors mapped from the full connection layer and create a one-dimensional output vector, with dimensions equal to the number of classes.

3.2. Deep Key Frame Extraction

3.2.1. Proposed Algorithm. In this section, we provide the details of the proposed deep key frame extraction method for sports training. The method is based on athlete skeleton extraction. As illustrated in Figure 2, the proposed frame extraction technique consists of four steps: preprocessing of the athlete training video, ROI extraction based on FCN, skeleton and feature extraction, and CNN-based key frame extraction. The proposed deep frame extraction method examines the poses of athletes in training videos. It first divides input videos into frame sequences followed by exploring ROI. FCN is applied for the extraction of foreground features of the athlete. Next, all the video frames are cropped according to the extracted ROI in the first frame.

3.2.2. Extracting Athletes' Skeletons. We used the ROI image extracted by the FCN network and the previously labeled ground truth to make the training data of the deep skeleton network. The original training image and the labeled ground truth are shown in Figure 3.

The Matlab (R2015a) software was used to extract the athletes' skeleton information of ground truth. The Matlab 'bwmorph' function was applied to perform the morphological operation on all images. The general syntax of Matlab bwmorp function is as follows:

BW2 = bwmorph (BW, operation, n), which applies morphological operation n times and n can be inf; in this case, the operation is repeated until the image no longer changes. Table 1 lists some of the different morphological operations that can be performed on images.

The different morphological operations can be selected to generate the athlete's skeleton information. The athlete's skeleton information of the four key postures is shown in Figure 4.

It can be seen from the athletes' skeleton information map that the four key postures have different athletes' skeleton information. The 373 labeled images were used to extract their athletes' skeleton information as the label of training deep skeleton network.

3.2.3. Generation of Athlete Skeleton Information. We prepared the training and test files, as shown in Figure 5. The left side represents the original image, whereas the right side is the ground truth.

Because the CNN network is changed from VGG (visual geometry group) network, some parameters of the VGG network are selected. The VGG is the conventional CNN architecture and consists of blocks, where each block consists of 2D convolution and max pooling layers. Similar to the FCN training method, the deep skeleton is different from the traditional single-label classification network but uses the image of athletes' skeleton information as the label.

After 20000 iterations, the trained model is obtained. The test set was randomly selected to test the recognition performance. According to the predicted value of each pixel, after normalization, the predicted gray image is drawn. The original and predicted images are shown in Figure 6.

The white portion in the figure indicates the skeleton information of athletes. The higher the value is, the more likely it is to be the skeleton information of athletes. Next, the nonmaximum suppression (NMS) algorithm is used to find the athletes' skeleton information. The NMS technique is used in several image processing tasks. It is a group of algorithms that chooses one entity out of many entities. We

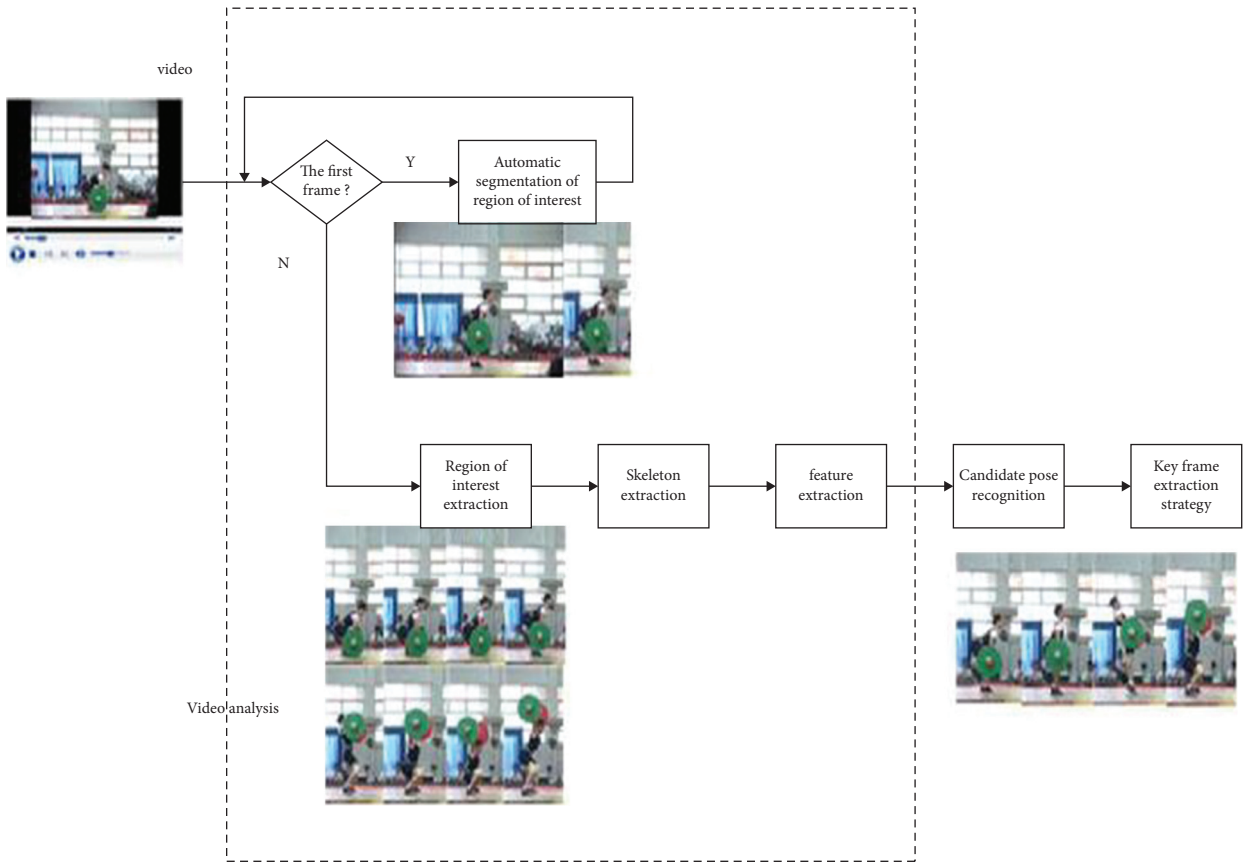


Figure 2: Algorithm framework.

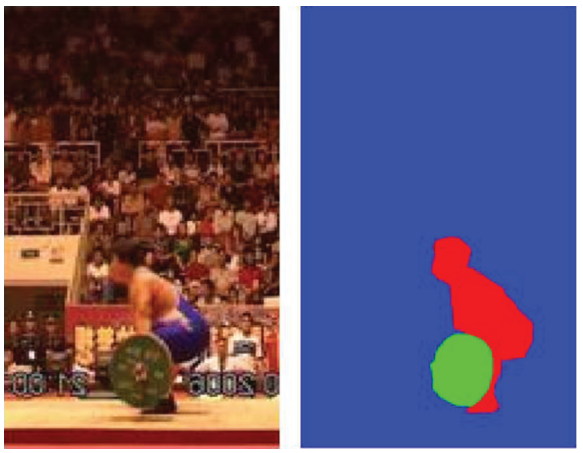
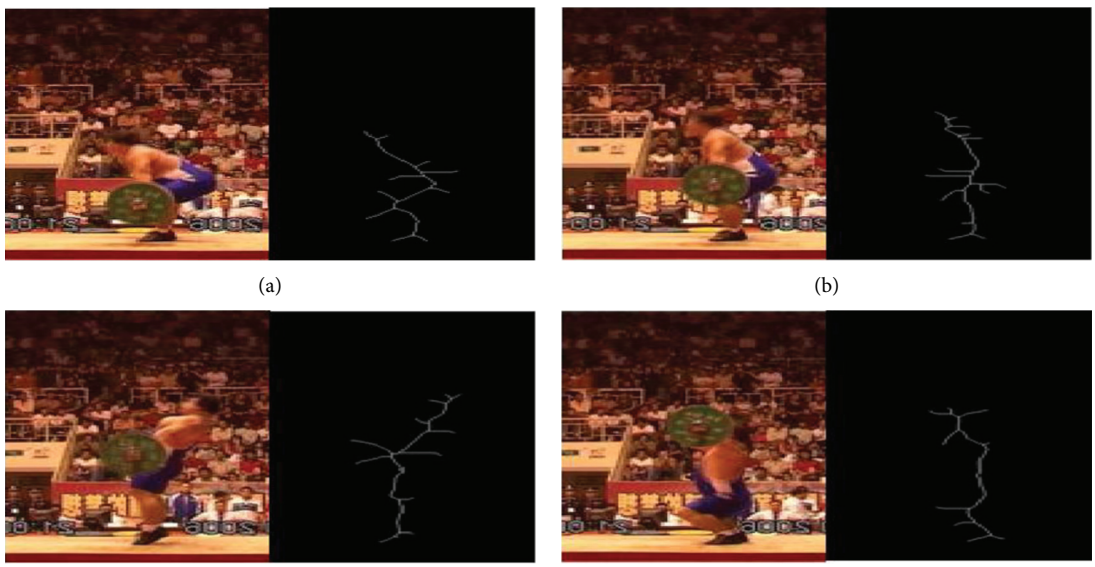


Figure 3: Original and ground truth.

Table 1: bwmorph morphological operations on images.

Operation	Description
Botha'	It is a morphological "bottom cap" transformation operation, and the returned image is the original image minus the morphological closing operation (closing operation: first expand and then corrode)
Bridge	Disconnected pixels: the value pixel is set to 1 if it has two nonzero unconnected (8 neighborhood) pixels
Clean	Remove isolated pixels (by O 1)
Close	Perform morphological closing operation (expansion before corrosion)
Diag	The diagonal filling is used to eliminate the 8 connected regions in the background
Dilate	The structure ones (3) are used to perform the expansion operation
Erode	The structure ones (3) are used to perform the corrosion operation
Fill	Fill in isolated internal pixels (0 surrounded by 1)



(a) (b) (c) (d)

Figure 4: (a) The original picture of athlete's and the ground truth of athletes' skeleton. (b) The original drawing of the knee lead and the ground truth of the athlete's skeleton. (c) The original drawing and the ground truth of the athlete's skeleton. (d) The original map of the highest point and the ground truth of the athlete's skeleton.

```
yuantu/shiyan795/257.jpg  label/shiyan795/257.png
yuantu/shiyan795/258.jpg  label/shiyan795/258.png
yuantu/shiyan795/259.jpg  label/shiyan795/259.png
yuantu/shiyan795/260.jpg  label/shiyan795/260.png
yuantu/shiyan795/261.jpg  label/shiyan795/261.png
yuantu/shiyan795/262.jpg  label/shiyan795/262.png
yuantu/shiyan795/263.jpg  label/shiyan795/263.png
yuantu/shiyan795/264.jpg  label/shiyan795/264.png
yuantu/shiyan795/265.jpg  label/shiyan795/265.png
yuantu/shiyan795/266.jpg  label/shiyan795/266.png
yuantu/shiyan795/267.jpg  label/shiyan795/267.png
yuantu/shiyan795/268.jpg  label/shiyan795/268.png
yuantu/shiyan795/269.jpg  label/shiyan795/269.png
yuantu/shiyan795/270.jpg  label/shiyan795/270.png
```

Figure 5: Training parameter.

Figure 6: Original and predicted results.

take 3 neighborhoods as an example to introduce the implementation of the NMS algorithm.

NMS in three neighborhoods is to judge whether the element $I[x]$ ($2 <= I <= W$-1) of a dimension group $I[w]$ is greater than its left neighbor $I[I$-1] and right neighbor $I[x+1]$ (Algorithm 1).

Lines 3–5 of the algorithm flow check whether the current element is greater than its left and right neighbor elements. If the condition is met, the element is the maximum point. For the maximum point $I[x]$, it is known that $I[x] > I[x+1]$, so there is no need to further process the $I+1$ position element. Instead, it directly jumps to the $I+2$ position, corresponding to the 12th line of the algorithm flow. If the element $I[x]$ does not meet the judgment condition of the third line of the algorithm flow, its right neighbor $I[x+1]$ is taken as the maximum candidate, corresponding to the seventh line of the algorithm flow. A monotonically increasing method is used to search the right until the element satisfying $I[x] > I[x+1]$ is found. If $I <= W-1$, this point is the maximum point, corresponding to lines 10-11 of the algorithm flow.

We used the NMS method of MATLAB toolkit, according to the results of deep skeleton network output, and, finally, determined the information pixels that may be athletes' skeleton. The predicted results and NMS results are shown in Figure 7. The test effect picture including the athlete skeleton is shown in Figure 8.

```
(1) x = 2
(2) While I ≤ w−1 do
(3) If I [x]x >x I [x + 1] then
(4)     If I [x] >= I [x−1] then
(5)         Maximum At (x)
(6) Else
(7)     x = x + 1
(8)     While x ≤ w−1 AND I [x] <= I [x + 1] do
(9)     x = x + 1
(10)    If x ≤ w−1 then
(11)    Maximum At (x)
(12) I = x + 2
```

Algorithm 1: NMS for three neighborhoods.

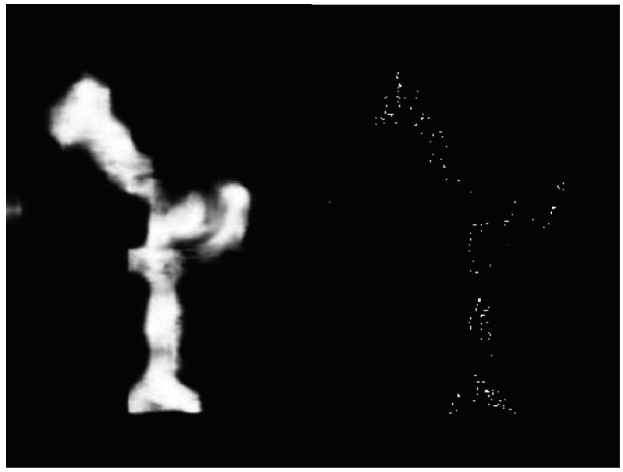


Figure 7: Prediction results and NMS results.

Figure 8: Test effect picture including athlete skeleton.

4. Results

In this section, we performed experimental analysis to confirm the performance of the proposed key frame extraction method. We performed experiments on sports videos collected from the Chinese Administration of Sports. All the videos contain four key athletes' poses.

4.1. CNN-Based Key Pose Estimation. The proposed key frame extraction method used CNN with ROI of the extracted video frames as input to predict probabilities of all poses. In all sports videos, there are four groups of key poses. The CNN model was used to calculate the probability of each frame for all frames estimated with accurate or inaccurate poses. Table 2 provides the classification results of 4 subjects corresponding to 4 poses of sport action videos. Firstly, 612 image frames are tested. Table 2 provides the number of correct and wrongly predicted frame number and the associated accuracy, sensitivity, and specificity for all poses predicted by CNN. It is evident that the accuracy, sensitivity, and specificity of pose probability estimated in the proposed model are higher than 90% on all the poses which provide a base for the ultimate key pose extraction in sports training.

4.2. Experimental Comparison. To ratify the superiority of the proposed key frame extraction method, we compared the obtained results with the existing pose estimation methods. The comparison results are shown in Table 3. Compared with the traditional deep learning method, the method in this paper has a great improvement. Because the athlete skeleton information is extracted from the key objects, the feature expression of human posture is enhanced, and the accuracy is improved. It can be observed that athletes'

Table 2: Test accuracy of four key frames.

	Total	Correct	Wrong	Accuracy (%)	Sensitivity (%)	Specificity (%)
Pose 1	169	166	3	98.2	92.4	94.7
Pose 2	130	125	5	96.1	90.4	95.3
Pose 3	155	152	3	98.1	96.3	98.3
Pose 4	158	155	3	98.1	97.6	97.7

Table 3: Experimental comparison.

Method	Accuracy (%)
Wu et al. [27]	90.6
Jian et al. [28]	97.4
Proposed skeleton-based method	**97.7**

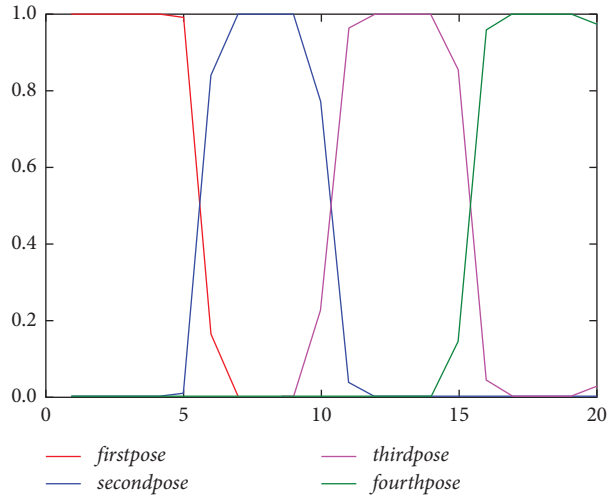


Figure 9: Test the results of the video.

skeleton extraction of key objects can improve the accuracy of classification. Wu et al. [27] achieved the highest accuracy of 90.6%, whereas Jian et al. [28] reported 97.4% accuracy. Compared with the aforementioned two methods, the proposed key frame extraction method achieved the highest average accuracy of 97.7% for all the pose categories.

4.3. Key Frame Extraction. Figure 9 shows the probability distribution of proposed skeleton-based key frame extraction for four groups of poses from training videos. It can be seen that the unique characteristics of each pose are properly captured, and the estimation of all four poses is good. In addition, the method in this paper has a very obvious performance in performance and effect and has a strong expression in each type of key posture. It combines FCN with CNN to extract ROI and distinct features and lays down the foundation for key frame extraction from sports videos. It further confirms that the proposed skeleton-based method conquers other key frame extraction methods. Test the results of the video is shown in Figure 9.

5. Conclusion

Object detection and behavior understanding in the video has become a point of contention in the field of machine vision. Sports video contains a lot of information related to the human movement which is complex and highly skilled. Compared with the analysis of human daily movements, the analysis and recognition movements in sports video are more challenging. In this study, we have presented a deep key frame extraction method for sport video analysis based on CNN. This paper extracts the key posture of athletes' training, to assist the coach to carry out more professional training for athletes, and puts forward a method to extract the skeleton information of human athletes. The human body in sports training action video is extracted through the skeleton of athletes to enhance the expression of features and the accuracy of key frame extraction. The experimental results and analysis validate that the proposed skeleton-based key frame extraction method outperforms its counterparts in key pose probability estimation and key pose extraction.

Data Availability

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

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This paper was supported by the 2019 Shandong Social Science Planning Project (item no. 19CTYJ16).

References

[1] L. Liang Li, S. Shuqiang Jiang, and Q. Qingming Huang, "Learning hierarchical semantic description via mixed-norm regularization for image understanding," *IEEE Transactions on Multimedia*, vol. 14, no. 5, pp. 1401–1413, 2012.

[2] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, Boston, MA, USA, June 2015.

[3] G. Lili and D. Shifei, "Research progress of deep learning," *Computer science*, vol. 42, no. 5, pp. 28–33, 2015.

[4] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proceedings Of IEEE Conference On Computer Vision AndPattern Recognition*, Anchorage, AK, USA, June 2008.

[5] G. Cheron, I. Laptev, and C. Schmid, "P-CNN pose-based CNN features for actionrecognition," in *Proceedings of IEEE International Conference on Computer Vision*, pp. 3218–3226, Santiago, Chile, December 2015.

[6] R. Girshick, J. Donahue, and T. Darrell, "Rich feature hierarchies for accurate objectdetection and semantic

segmentation," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 580–587, Columbus, OH, USA, June 2014.

[7] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[8] K. Hornik, M. Stinchcombe, and H. White, *Multilayer Feedforward Networks Are Universal Approximators*, Elsevier Science Ltd, Amsterdam, Netherlands, 1989.

[9] M. W. Gardner and S. R. Dorling, "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences," *Atmospheric Environment*, vol. 32, no. 14, pp. 2627–2636, 1998.

[10] S. Kulhare, S. Sah, S. Pillai, and R. Ptucha, "Key frame extraction for salient activity recognition," in *Proceedings Of the 2016 23rd International Conference On Pattern Recognition (ICPR)*, Cancun, Mexico, December 2016.

[11] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

[12] Z. Cernekova, I. Pitas, and C. Nikou, "Information theory-based shot cut/fade detection and video summarization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, pp. 82–91, 2006.

[13] Y. Zhuang, Y. Rui, T. S. Huang, and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering," in *Proceedings of The1998 International Conference on Image Processing. ICIP98*, vol. 1, pp. 866–870, Chicago, IL, USA, October 1998.

[14] H. Tang, H. Liu, W. Xiao, and N. Sebe, "Fast and robust dynamic hand gesture recognition via key frames extraction and feature fusion," *Neurocomputing*, vol. 331, pp. 424–433, 2019.

[15] R. VáZquez and A. Bandera, "Spatio-temporal feature-based key frame detection from video shots using spectral clustering," *Pattern Recognition Letters*, vol. 34, pp. 770–779, 2013.

[16] Y. L. Cun, B. Boser, and J. S. Denker, "Handwritten digit recognition with a back-propagation network," in *Proceedings of Advances in Neural Information Processing Systems*, pp. 396–404, Morgan Kaufmann Publishers Inc., Denver, CO, USA, 1990.

[17] K. Yu, Y. Lin, and J. Lafferty, "Learning image representations from the pixel level via hierarchical sparse coding," in *Proceedings of Computer Vision and Pattern Recognition*, pp. 1713–1720, IEEE Xplore, Colorado Springs, CO, USA, 2011.

[18] X. Yu, Q. Peng, L. Xu, F. Jiang, J. Du, and D. Gong, "A selective ensemble learning based two-sided cross-domain collaborative filtering algorithm," *Information Processing & Management*, vol. 58, Article ID 102691, 2021.

[19] L. Jianwei, L. Yuan, and L. xionglin, "Research progress of Boltzmann machine," *Computer Research and Development*, vol. 51, no. 1, pp. 1–16, 2014.

[20] X. Yu, F. Jiang, J. Du, and D. Gong, "A cross-domain collaborative filtering algorithm with expanding user and item features via the latent factor space of auxiliary domains," *Pattern Recognition*, vol. 94, pp. 96–109, 2019.

[21] B. Mahasseni, M. Lam, and S. Todorovic, "Unsupervised video summarization with adversarial LSTM networks," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–10, Honolulu, HI, USA, July 2017.

[22] H. Yang, B. Wang, S. Lin, D. Wipf, M. Guo, and B. Guo, "Unsupervised extraction of video highlights via robust recurrent auto-encoders," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4633–4641, Santiago, Chile, December 2015.

[23] A. Kar, N. Rai, S. Sikka, and G. Sharma, "adaptive scan pooling in deep convolutional neural networks for human action recognition in videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, July 2017.

[24] C. Huang and H. Wang, "Novel key-frames selection framework for comprehensive video summarization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, pp. 577–589, 2019.

[25] M. Jian, S. Zhang, L. Wu, S. Zhang, and X. Wang, "Deep key frame extraction for sport training," *Neurocomputing*, vol. 328, pp. 147–156, 2019.

[26] S. Wen, W. Liu, Y. Yang, T. Huang, and Z. Zeng, "Generating realistic videos from keyframes with concatenatedGANs," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, pp. 2337–2348, 2019.

[27] L. Wu, J. Zhang, and F. Yan, "A pose let based key frame searching approach in sports training videos," in *Proceedings of the Information Processing Association Annual Summit and Conference*, pp. 1–4, Hollywood, CA, USA, December 2012.

[28] M. Jian, S. Zhang, L. Wu, S. Zhang, X. Wang, and Y. He, "Deep key frame extraction for sport training," *Neurocomputing*, vol. 328, pp. 607–616, 2018.