Hindawi
Scientific Programming
Volume 2021, Article ID 9956710, 10 pages
https://doi.org/10.1155/2021/9956710

Research Article

Machine Vision and Big Data-Driven Sports Athletes Action Training Intervention Model

Hui Jiang,[1] Ping wang,[2] Lei Peng,[3] and Xiaofeng Wang [4]

[1]*Institute of Physical Education, Dezhou University, Dezhou 253023, Shandong, China*
[2]*Baoding Vocational and Technical College, Baoding, 071000, China*
[3]*College of Physical Education, Hengshui University, Hengshui 053000, Hebei, China*
[4]*Sports Department of Hebei Vocational College of Rail Transportation, Tianjin, Hebei, China*

Correspondence should be addressed to Xiaofeng Wang; wangxiaofengdeqq@163.com

Received 8 March 2021; Revised 8 April 2021; Accepted 26 April 2021; Published 17 May 2021

Academic Editor: Shah Nazir

In recent years, athlete action recognition has become an important research field for showing and recognition of athlete actions. Generally speaking, movement recognition of athletes can be performed through a variety of modes, such as motion sensors, machine vision, and big data analysis. Among them, machine vision and big data analysis usually contain significant information which can be used for various purposes. Machine vision can be expressed as the recognition of the time sequence of a series of athlete actions captured through camera, so that it can intervene in the training of athletes by visual methods and approaches. Big data contains a large number of athletes' historical training and competition data which need exploration. In-depth analysis and feature mining of big data will help coach teams to develop training plans and devise new suggestions. On the basis of the above observations, this paper proposes a novel spatiotemporal attention map convolutional network to identify athletes' actions, and through the auxiliary analysis of big data, gives reasonable action intervention suggestions, and provides coaches and decision-making teams to formulate scientific training programs. Results of the study show the effectiveness of the proposed research.

1. Introduction

Estimation of sports athletes' action pose [1–3] is an emerging frontier research direction. At present, most of them are based on traditional image processing methods. The research steps of human body posture estimation based on neural network are firstly to detect the human body on the input sports athlete images, secondly to estimate the athlete's action posture, and finally evaluate the estimation results and compare the correct posture to give action intervention [4] suggestions. The human body detection of sports athletes is an important step in constructing a neural network for motion posture estimation. It usually uses related algorithms to determine whether there is a target research object in the input image. If there is, mark its specific location in the picture and use a border or red box subject circled. The research of motion pose estimation first needs to detect the athlete's human body, and the input picture is used to extract the bounding box of the athlete's portrait through a specific human body detector.

Since the single-person pose estimation target has only a single object, the recognition technology is relatively mature and the recognition speed is faster. TOSHEV et al. [5] first proposed the DeepPose method based on neural networks, which proposed a cascaded DNN-like regression, which can achieve high-precision estimation. The advantage of this method is to estimate the attitude in a holistic manner and has good generalization performance; TOMPSON et al. [6] proposed a new hybrid architecture, which is composed of deep convolutional neural networks and Markov random fields. Including an additional "torso joint heat map" to merge data to select the correct feature activation in the chaotic scene, YANG et al. [7] presented a new end-to-end human pose estimation framework, combining DCNN with parts. Expressible deformations are mixed together, DCNN is used to return the heat map of each body part, and the

structured output of deep learning of neural network is used to further simulate the relationship between body joints.

The above work proves the powerful performance of machine vision in the recognition of athletes' action gestures, so it is effective to use machine vision neural networks to participate in the intervention of sports athletes' action recognition. However, the existing machine vision-based methods rarely consider the combination of sports athletes' historical training and competition big data for analysis. Therefore, based on this shortcoming, we propose a novel combination of machine vision and big data analysis of athletes' postures. The estimation method can accurately estimate the motion posture. In addition, we also calculate the loss of the estimation result and the correct action posture. Following is the main innovative points of this paper:

This paper recommends a novel machine vision method to recognize sports athletes' actions. Among them, the machine vision method is realized by spatiotemporal attention map convolutional network.

This paper combines machine vision methods with big data technology. Because machine vision methods are used alone, even if they can accurately recognize action gestures, it is difficult to give intervention suggestions. Therefore, this paper considers athletes' historical training and competition big data. Auxiliary analysis is helpful for the formulation of training action intervention programs.

On the basis of the machine vision method, this paper proposes to use big data analysis to analyze athletes' historical training and competition data and to recognize losses in computer actions in order to develop scientific intervention strategies.

We have conducted sufficient comparative experiments and ablation studies. The athlete's movement intervention method based on machine data and big data analysis can provide a scientific basis for sports coaches to develop a reasonable training plan.

The organization of the paper is given as follows: Section 2 of the paper represents the related research in the area. Section 3 describes the methodology of the paper with the details of the research done. Section 4 shows the experiments and results of the study conducted for the proposed study. The paper is concluded in Section 5.

2. Related Research

Song et al. [8] proposed a deeply structured model to predict human pose sequences in unconstrained videos. The model can be effectively trained in an end-to-end manner and can simultaneously represent the appearance of body joints and their temporal and spatial relationships. Hossain et al. [9] used the time information on the 2D joint position sequence to estimate the 3D pose sequence. A sequence-to-sequence network composed of layered standardized LSTM units is also designed. The network has shortcut connections to connect the input to the output on the decoder side and imposes time smoothing constraints during the training process. Pavllo et al. [10] proved that a complete convolution model based on temporal convolution on 2D key points can effectively estimate the 3D pose in the video. It also introduces back projection, which can start from the 2D key point prediction of the unmarked video, then estimate the 3D pose, and finally back-project it to the input 2D key point.

CAO et al. [11] proposed an effective method for detecting multiperson poses in images. This method has high accuracy on multiple public benchmarks and greatly exceeds the multiperson detection level of MPII dataset in performance and efficiency. This method expresses bottom-up association scores through partial affinity fields (PAF). PAF is a set of two-dimensional vector fields that can encode the position and direction of limbs in the image domain, allowing bottom-up analysis, and can maintain high precision in real time at the same time and is not affected by the number of people in the image. Chen et al. [12] used a top-down approach to propose a new neural network structure called cascaded pyramid network (CPN), which includes two stages: GlobalNet and RefineNet. GlobalNet is a functional pyramid network that can accurately locate "simple" joint points (such as eyes and hands) but cannot accurately identify occluded or invisible joint points; RefineNet integrates all the feature representations of GlobalNet and mines the loss of online joint points and explicitly handles hard joint points.

In view of the above-mentioned research, the literature proves that machine vision has excellent performance in human body posture, which is superior to traditional image processing algorithms. Therefore, this also proves the rationality of the machine vision method proposed in this paper.

3. Methodology

Estimation of sports athletes' action pose [1–3] is an emerging frontier research direction. At present, most of them are based on traditional image processing methods. As shown in Figure 1, the general procedure is to first obtain the front and side images of the person and then extract the outline of the human body through image processing. Identify the key size points, then establish a function model of the human body dimension curve through statistical analysis and curve fitting, and import the complete athlete action data record table into the large sports action database after measurement by related auxiliary tools. With the fast development of computer vision technology [13–18], human body posture estimation has begun to be researched with neural network models [19–23], which has significantly improved the accuracy and robustness of human body posture estimation, has expanded the scope of application, and has been deeply integrated into sports competition and sports training.

Figure 2 is a flowchart of the overall architecture of our MVBD-Net algorithm. First, we extract the human pose from the motion images of sports players. Secondly, the adjacency matrix of the graph is constructed to feed the graph convolutional neural network to obtain the posture feature, and then it is input into the fully connected layer and the Softmax function to obtain the output of the action posture estimation. Secondly, this article also uses big data analysis technology to obtain the weight of the pose estimation and perform feature

Figure 1: Example of human pose estimation.

fusion with the output of the graph convolution network and finally calculate the loss with the label.

3.1. Human Detection. For football player training image data, this article first conducts human body detection. Human body detection is an important step in constructing a neural network for human posture estimation. It usually uses related algorithms to determine whether there is a target research object in the input image. If there is, mark its specific location in the picture, and circle the target object with a frame or red box stand up. Human body pose estimation research first needs human body detection. The input picture is used to extract the bounding box of the person through a specific human body detector. Common human body detectors include Yolo and R–CNN.

Since multiperson pose estimation cannot determine the specific location and total number of people in the image, it is much more difficult to implement than single-person pose estimation. There are usually two methods to achieve multiperson pose estimation: (1) first provide a human detector, then estimate each component separately, and finally get the pose of each person. This method is a top-down method; detect all the parts in the image, and then associate and group different parts. This method is a bottom-up method. Therefore, this article uses the YOLO algorithm for human detection (as shown in Figure 3).

YOLO loss calculation during training is shown in the following equation:

$$\begin{aligned}
&\lambda_{\text{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \Pi_{ij}^{\text{obj}} \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \right] \\
&+ \lambda_{\text{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \Pi_{ij}^{\text{obj}} \left[\left(\sqrt{w_i} - \sqrt{\hat{w}_i} \right) + \left(\sqrt{h_i} - \sqrt{\hat{h}_i} \right)^2 \right] \\
&+ \sum_{i=0}^{S^2} \sum_{j=0}^{B} \Pi_{ij}^{\text{obj}} \left(C_i - \hat{C}_i \right)^2 \\
&+ \lambda_{\text{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \Pi_{ij}^{\text{noobj}} \left(C_i - \hat{C}_i \right)^2 \\
&+ \sum_{i=0}^{S^2} \Pi_{ij}^{obj} \sum_{c \in \text{classes}} \left(p_i(c) - \hat{p}_i(c) \right)^2,
\end{aligned} \tag{1}$$

where x, y, w, C, p is the network predicted value, $\hat{x}, \hat{y}, \hat{w}, \hat{C}, \hat{p}$ is the labeled value, Π_i^{obj} indicates that the object falls into the grid i, and Π_{ij}^{obj} and Π_{ij}^{noobj}, respectively, indicate that the object falls and does not fall into the jth bounding box of the grid i.

3.2. Spatiotemporal Attention Graph Convolutional Network

3.2.1. Graph Convolutional Network. The human body motion posture is in the form of graphics, rather than two-dimensional or three-dimensional grids, which makes it difficult to simply use convolutional networks. GCN (as shown in Figure 4) is a general and effective framework for learning to represent graph structure data. Various GCN variants have achieved the most advanced results on many tasks. For skeleton-based human behavior recognition, skeleton-based data can be obtained from motion capture devices or pose estimation algorithms in videos. Usually the data is a frame sequence, and each frame will have a set of joint coordinates. On the basis of a given sequence of human joints in the form of two-dimensional or three-dimensional coordinates, a spatiotemporal graph with joints as graph nodes and the natural connectivity of human body structure and time as graph edges is constructed. Graph convolution can be defined as follows:

$$f_{\text{out}}(v_{ti}) = \sum_{v_{ti} \in N} \frac{1}{z_{ti}(v_{ti})} f_{in}(v_{ti}) W(l_{ti}(v_{ti})), \tag{2}$$

where f_{in} is the feature vector input of node v_{ti} and W is a weight function and is mapped from the graph label l_{ti}: V_{t} from K, which can be used to assign a label to each graph node v_{ti}.

3.2.2. Spatiotemporal Attention. In the algorithm model of this paper, a new spatiotemporal attention mechanism is proposed, which captures the dynamic spatiotemporal correlation on the athlete's action network. This mechanism contains two kinds of attention, namely, spatial attention and temporal attention.

Space Attention. In the spatial dimension, the conditions of nodes at different locations influence each other, and the mutual influence is very dynamic.

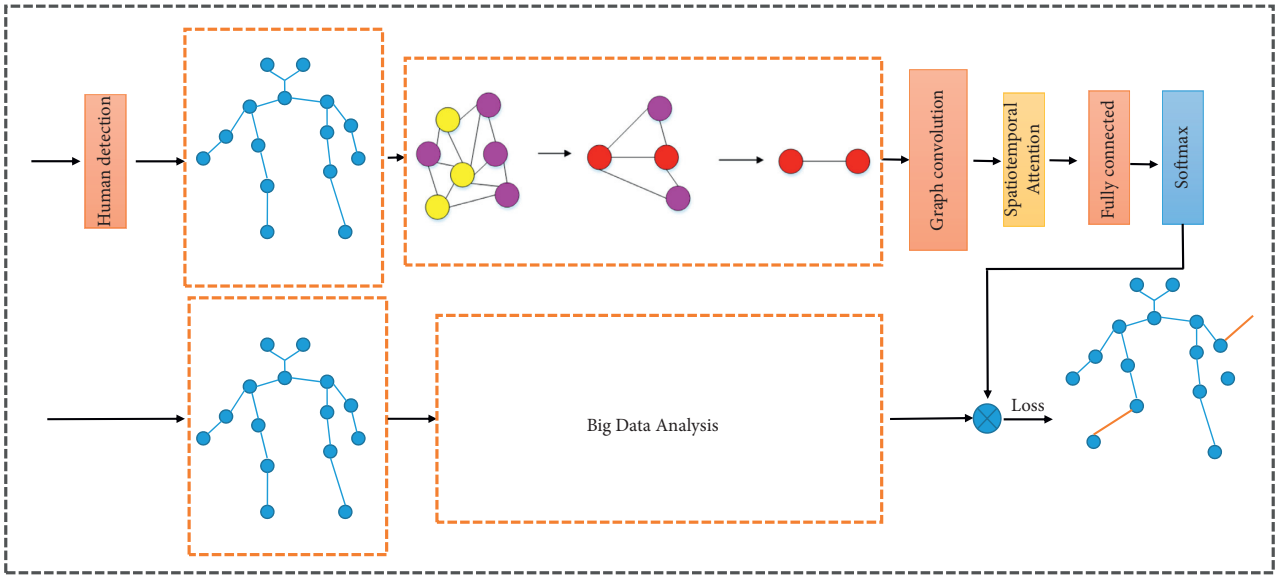

Figure 2: The flowchart of the overall architecture of our MVBD-Net algorithm.

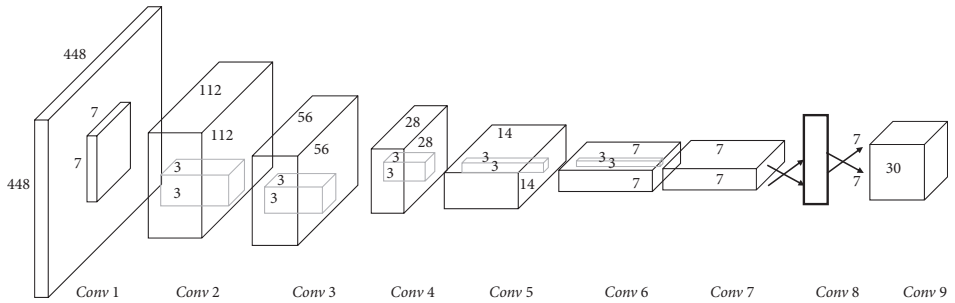

Figure 3: The flowchart of the overall architecture of Yolo.

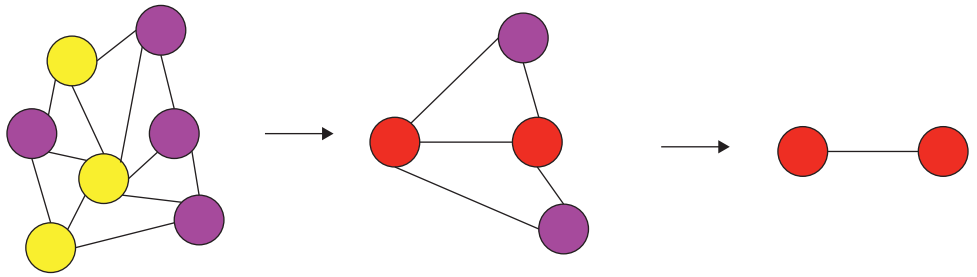

Figure 4: Graph convolutional neural network process.

$$S = V_s \times \sigma\left(\left(x_h^{r-1} w_1\right) w_2 \left(w_3 x_h^{r-1}\right)^T + b_s\right), \tag{3}$$

$$S_{ij}' = \frac{\exp\left(S_{ij}\right)}{\sum_{j=1}^{N} \exp\left(S_{ij}\right)}, \tag{4}$$

where x_h^{r-1} represents the input of the r th space-time block and $_C^{r-1}$ is the number of channels of input data in the r th layer. The attention matrix S is dynamically calculated based on the current input of this layer. T_{r-1} represents the length of the time dimension in the r th layer, V_s, b_s, w_1, and w_2, and w_3 represent the learning parameters, and σ represents the activation function.

Time Attention. In the time dimension, there is a correlation between the conditions of human joints in different time periods, and the correlation is also different in different situations. Similarly, this paper uses the attention mechanism to adaptively give different attention to the data.

$$E = V_e \times \sigma\left(\left(x_h^{r-1}\right)^T u_1\right) u_2\left(u_3 x_h^{r-1} + b_e\right), \tag{5}$$

$$E_{ij}' = \frac{\exp\left(E_{ij}\right)}{\sum_{j=1}^{T_{r-1}} \exp\left(E_{ij}\right)}, \tag{6}$$

where V_e, b_e, u_1, u_2, u_3 is the learning parameter and the time correlation matrix E is determined by the changing input. The value of an element E_{ij} j in E semantically represents the strength of dependence between time i and time j, and finally E is normalized by the Softmax function.

3.2.3. Spatiotemporal Attention Fusion. The spatial attention network generates the spatial attention heat map to guide the action classification network of football players to extract effective spatiotemporal features from the spatial region of interest. The time attention mechanism automatically mines discriminative time-domain video clips from the original training or competition videos and uses these video clips for network training, while eliminating the interference of other video clips on the classifier. The proposed spatiotemporal attention model separately trains two network models in RGB video frames and optical flow sequences, namely, spatial subnetwork (SN) and temporal subnetwork (Temporal Network, TN). Then, the weighted fusion of the Softmax prediction scores of the two networks is used as the basis for behavior classification, which can effectively improve the classification robustness of the network. It should be noted that when training on RGB video frame data, the parameters of the spatial attention network need to be pretrained on the optical flow prediction database in advance, and when training on optical flow sequence data, the spatial attention network and the behavior classification network share weights.

4. Experiments and Results

In this article, we collected 1000 training images of sports figures from the Leeds Sports Pose dataset and annotated 14 joints. Those images are challenging due to the different appearance and strong sharpness. The images in the Leeds Sports Pose dataset have been scaled so that the most prominent figures are approximately 150 pixels tall. Although each image in Leeds Sports Pose may contain multiple people, standard preprocessing for human detection has been performed to extract a single person. As in previous works, we use the subimages of these detected individuals as training and testing samples. In this way, the training and testing data contains only one person, and as mentioned earlier, in the testing phase, we only use the entire image (for the Leeds Sports Pose dataset, this means the entire subimage of a person) as a body patch.

4.1. Evaluation Methods. oks (object key point similarity) is the evaluation index of the commonly used human bone key point detection algorithm. This index is inspired by the IoU index in target detection. The purpose is to calculate the truth value and predict the similarity of the key points of the human body. The formula is as follows:

$$O_p = \begin{cases} KS = e^{\frac{\left\|\widehat{\theta}_i^P - \theta_i^P\right\|_2^2}{2s^2k_i^2}}, \\ OKS = \dfrac{\sum_i \exp\left\{-d_{pi}^2/2S_p^2\sigma_i^2\right\}\delta\left(v_{pi} > 0\right)}{\sum_i \delta\left(v_{pi} > 0\right)}, \end{cases} \tag{7}$$

where p represents the person with id p among all ground truth pedestrians in the current picture, $p \in (0, M)$, and M represents the number of pedestrians in the current picture. Since the training and test data contains only one person, M here is set to 1. i represents the key point with id i. d_{pi} represents the Euclidean distance between the key point with id i in the set of key points currently detected and the key point with id p in the ground truth pedestrian.

$$d_{pi} = \sqrt{\left(x_i^{\sim} - x_{pi}\right)\left(y_i^{\sim} - y_{pi}\right)}, \tag{8}$$

where $(x_i^{\sim}, y_i^{\sim})$ is the current key point detection result and (x_i, y_i) is the ground truth.

S_p represents the scale factor of the person with id p in the ground truth pedestrian, and its value is the square root of the area of the pedestrian detection frame:

$$S_p = \sqrt{wh}, \tag{9}$$

where w and h are the width and height of the detection frame.

σ_i represents the key point normalization factor of type i. This factor is the standard deviation between the ground truth key points in all the sample sets and the true value manually marked. v_{pi} represents the visibility of the i key points of the pedestrian with id p in the ground truth.

$\delta(*)$ means if the condition $*$ holds, then $\delta(*) = 1$; otherwise, $\delta(*) = 0$.

AP (average precision) is used to calculate the accuracy percentage of the test set. In single-person pose estimation, only one pedestrian is estimated at a time, that is, $M = 1$ in the *oks* indicator, so the ground truth in a picture is a pedestrian (GT), and a set of key points will be obtained after the key point detection of this pedestrian (DT), and finally calculate the similarity *oks* between GT and DT as a scalar, and then artificially give a threshold T, and then AP can be calculated from the *oks* of all pictures:

$$AP = \frac{\sum_p \delta\left(oks_p > T\right)}{\sum_p 1}. \tag{10}$$

PCP is the Percentage of Correct Parts. If the key distance between the positions of the two joint points and the real limb reaches at most half the length of the real limb, the joint point is considered to be correctly predicted.

4.2. Experimental Results of Different Methods. In order to verify the competitiveness of MVBD-NET in the task of human skeleton action recognition, the model trained on the

Table 1: PCP comparison on Leeds Sports Pose dataset using oks evaluation method.

Methods	Arm		Leg		Torso	Head
	Upper	Lower	Upper	Lower		
Dantone	0.53	0.35	0.74	0.71	0.82	0.78
Tian	0.45	0.38	0.52	0.69	0.81	0.65
Johnson	0.52	0.31	0.64	0.52	0.71	0.74
Wang	0.54	0.45	0.74	0.65	0.76	0.80
Pishchulin	0.43	0.55	0.68	0.71	0.89	0.81
Ours	**0.86**	**0.70**	**0.89**	**0.86**	**0.92**	**0.88**

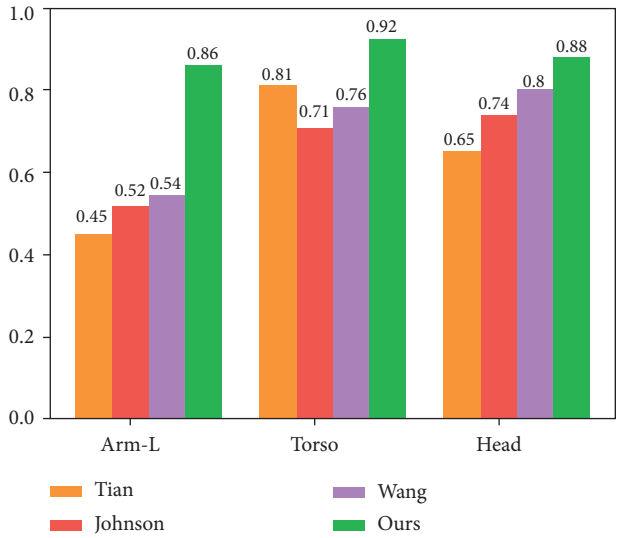

Figure 5: PCP comparison on LSP.

graph convolutional network based on the spatiotemporal attention mechanism is compared with the current state of the art (SOA). The methods were compared experimentally. Compare the proposed MVBD-Net model with Dantone [24], Tian [25], Johnson [26], Wang [27], and Pishchulin [28]. The results of the different datasets are shown in the following Table 1 and Figure 5.

The experiment is mainly to verify the effectiveness of the time attention mechanism on the action recognition data set. This set of experiments designed two different convolution models based on time attention mechanism. The models are all based on the ST-GCN benchmark model with time attention mechanism. The temporal attention mechanism convolution model uses different video time-domain segmentation and selection methods to verify the influence of different time-domain segmentation methods and different segmentation parameters on the accuracy of behavior recognition. The training process of the benchmark model in this set of experiments is the same as the previous set of experiments. The experimental results are shown in Table 1. Among them, the average fusion mode (AFM) is to directly average the prediction results of each time-domain segment as the prediction result of the entire video; and the discrimination weighted (Discriminative Confidence Weighted, DCW) represents a time-domain fusion method based on weighted prediction credibility; that is, the most reliable video segment is selected according to the prediction credibility of each input segment, and the prediction results of these segments are weighted as the prediction result of the entire action.

4.3. Experimental Results of Ablation Studies on Spatiotemporal Attention Graph Convolutional Network. In this section, five sets of experiments are designed to verify in detail the effectiveness of the graph convolutional neural network based on the spatiotemporal attention mechanism in the task of skeletal action recognition. The first set of experiments shows the visualization results of the attention heat map of the corresponding frame learned by the attention model network [29, 30]. The second and third sets of experiments are slice experiments, which are used to independently verify the effectiveness of the spatial attention mechanism and the temporal attention mechanism. The fourth set of experiments is to integrate spatial attention mechanism and temporal attention mechanism into a graph convolutional neural network to achieve end-to-end training and apply it to action recognition tasks. This experiment is to verify the effect of introducing spatiotemporal attention mechanism into the graph convolutional neural network on the task of action recognition.

The experimental results are shown in Table 2. MVBD is a spatiotemporal attention network model established on the

Table 2: Average precision of joint detection on LSP.

LSP	Ankle	Knee	Hip	Wrist	Elbow	Shoulder	Neck	Head	mAP
P_{TF}	0.25	0.34	0.23	0.23	0.43	0.32	0.21	0.34	0.31
P_{HSV}	0.37	0.31	0.36	0.41	0.43	0.33	0.32	0.76	0.36
MVBD-net	**0.47**	**0.49**	**0.39**	**0.38**	**051**	**0.52**	**0.61**	**0.88**	**0.69**

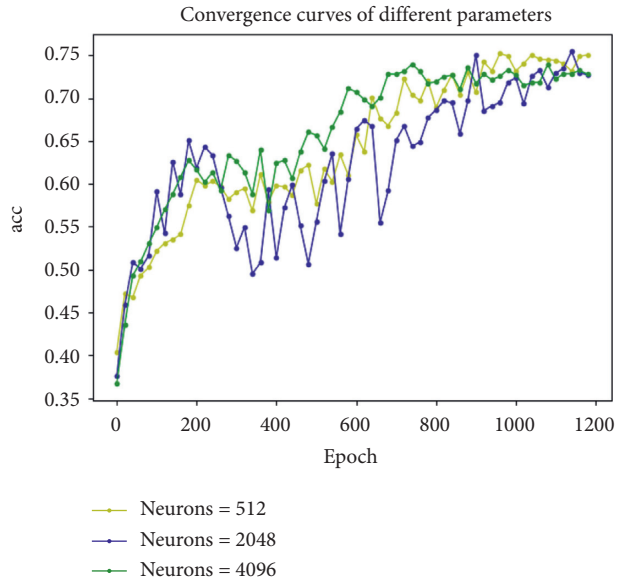

Figure 6: Convergence curves of different parameters.

basis of a recurrent neural network with long- and short-term memory. It can selectively focus on the joint differences of input frames and give different degrees of attention to the output of different frames, so it can extract distinguishable. The spatiotemporal features help action recognition. ST-GCN [21] breaks through the limitations of previous bone modeling methods and applies graph convolution to human skeleton action recognition, and the proposed model has strong generalization ability. AS-GCN [22] combines A-links and S-links into a generalized pose graph, further establishes a behavior structure graph convolutional network model, learns spatial and temporal characteristics, and can capture different action patterns more accurately and in detail.

4.4. Experimental Results of Ablation Studies on Different Parameters. Considering that there are a large number of parameters that can be optimized in the network structure designed in this paper, the use of different parameter settings will have different effects on the accuracy and operating efficiency of the model, so this paper conducts ablation experiment analysis on different parameter configurations. Since the three-dimensional pose estimation in this paper is implemented using a fully connected network, and the number of neurons in the fully connected layer is different, the number of model parameters and the prediction effect are also different. So in Figure 6, a different number of neurons (that is, Linear_size, representing fully connected layers) are analyzed. By comparing the loss function (loss) curve of the number of neurons, it is found that the loss value shows a gradually decreasing trend as the number of neurons increases from 256 to 4,096, which indicates the accuracy of the number of neurons. There was a positive correlation between model training and the number of neurons. The positive correlation indicates that the convergence of the model training will be higher with the increase of the number of neurons. As shown in Figure 7, this paper randomly selects 50 batch test data sets from the test set of the data set Human3.6 M to verify the prediction effect of the model. The batch-size is 64, and a total of 3 200 test data sets are used in this experiment. The abscissa in the figure represents the number of neurons in each layer, the parameters are set to 512, 1024, 2 048, and 4 096, and the ordinate represents the average value (mm) of node errors in the calculation of the batch of test data. By comparing the number of neurons with different numbers, the prediction result found that the error value of the number of neurons of 4 096 is significantly lower than that of the number of neurons of 2 048. Therefore, this article sets the number of neurons in the fully connected layer to 4096. The convergence curves of different parameters are shown in Figure 6.

4.5. Experimental Results of Ablation Studies on Different Submodule. Since the MVBD-NET algorithm uses the Yolo algorithm to extract the LSP human body pose and uses big data technology to do auxiliary analysis; in this section, we

Figure 7: The visualization results of athlete training action recognition on the LSP dataset.

Table 3: Average precision of different submodule on LSP.

LSP	Ankle	Knee	Hip	Wrist	Elbow	Shoulder	Neck	Head	mAP
Yolo-submodule	0.26	0.39	0.22	0.27	0.41	0.33	0.28	0.36	0.33
BigData-Submodule	0.41	0.48	0.39	0.32	0.49	0.46	0.47	0.72	0.56
MVBD-net	**0.47**	**0.49**	**0.39**	**0.38**	**051**	**0.52**	**0.61**	**0.88**	**0.69**

will analyze the impact of these strategies on the experiment. The results are shown in Table 3.

4.6. Visualization of Results. This section shows the visualization results of the MVBD-NET algorithm on the LSP data set. Due to the introduction of an effective spatiotemporal attention model and training strategy, which can extract discriminative spatiotemporal features, the MVBD-Net proposed in this paper has obtained the best classification accuracy among current similar methods on this data set. Figure 7 visually shows the results of the algorithm.

5. Conclusion

Nowadays, the athlete action recognition has become a significant research area for showing and recognition of athlete actions. Movement recognition of athletes can be accomplished through a variety of modes, such as motion sensors, big data analysis, and machine vision. This paper proposes a graph convolutional neural athlete motion recognition algorithm based on big data analysis and machine vision, which is used for athlete's motion gesture recognition and intervention. Among them, this article first uses the Yolo model to extract the athlete's action posture features and uses the spatiotemporal attention map convolutional neural network to estimate the posture. Secondly, we analyzed the historical training and competition data through big data analysis technology, obtained the posture weight, and performed feature fusion with the output of the graph convolutional neural network. Experimental results show that our MVBD-Net algorithm is effective. It can formulate reasonable training plans for coaches and athletes of sports teams, provide scientific basis, and improve training efficiency and training effects. The results achieved from the proposed study show the effectiveness of the study.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] D. C. Luvizon, D. Picard, and H. Tabia, "2d/3d pose estimation and action recognition using multitask deep learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5137–5146, Salt Lake City, UT, USA, June 2018.

[2] M. Liu and J. Yuan, "Recognizing human actions as the evolution of pose estimation maps," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1159–1168, Salt Lake City, UT, USA, June 2018.

[3] H. H. Pham, H. Salmane, L. Khoudour, A. Crouzil, S. A. Velastin, and P. Zegers, "A unified deep framework for joint 3d pose estimation and action recognition from a single rgb camera," *Sensors*, vol. 20, no. 7, p. 1825, 2020.

[4] G. Yang, L. Wang, X. Xu, and J. Xia, "Footballer action tracking and intervention using deep learning algorithm," *Journal of Healthcare Engineering*, vol. 2021, 2021.

[5] A. Toshev and C. Szegedy, "Deeppose: human pose estimation via deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1653–1660, Ohio, Columbus, June 2014.

[6] J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," 2014, https://arxiv.org/abs/1406.2984.

[7] W. Yang, W. Ouyang, H. Li, and X. Wang, "End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3073–3082, Las Vegas, NV, USA, June 2016.

[8] J. Song, L. Wang, L. Van Gool, and O. Hilliges, "Thin-slicing network: a deep structured model for pose estimation in videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4220–4229, Honolulu, HI, USA, July 2017.

[9] M. R. I. Hossain and J. J. Little, "Exploiting temporal information for 3d human pose estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 68–84, Munich, Germany, September 2018.

[10] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3d human pose estimation in video with temporal convolutions and semi-supervised training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7753–7762, Long Beach, CA, USA, June 2019.

[11] Z. Cao, T. Simon, S. E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7291–7299, Honolulu, HI, USA, July 2017.

[12] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7103–7112, Salt Lake City, UT, USA, June 2018.

[13] X. Ning, S. Xu, W. Li, and S. Nie, "FEGAN: flexible and efficient face editing with pre-trained generator," *IEEE Access*, vol. 8, pp. 65340–65350, 2020.

[14] W. Cai and Z. Wei, "PiiGAN: generative adversarial networks for pluralistic image inpainting," *IEEE Access*, vol. 8, pp. 48451–48463, 2020.

[15] X. Ning, P. Duan, W. Li, Y. Shi, and S. Li, "A CPU real-time face alignment for mobile platform," *IEEE Access*, vol. 8, pp. 8834–8843, 2020.

[16] W. Cai and Z. Wei, "Remote sensing image classification based on a cross-attention mechanism and graph convolution," *IEEE Geoscience and Remote Sensing Letters*, vol. 7, 2020.

[17] X. Ning, F. Nan, S. Xu, L. Yu, and L. Zhang, "Multi-view frontal face image generation: a survey," *Concurrency and Computation: Practice and Experience*, Article ID e6147, 2020, in Press.

[18] W. Cai, B. Liu, Z. Wei, M. Li, and J. Kan, "TARDB-Net: triple-attention guided residual dense and BiLSTM networks for hyperspectral image classification," *Multimedia Tools and Applications*, vol. 80, pp. 1–22, 2021.

[19] X. Ning, W. Li, and J. Xu, "The principle of homology continuity and geometrical covering learning for pattern recognition," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 32, no. 12, Article ID 1850042, 2018.

[20] Z. Wang, C. Zou, and W. Cai, "Small sample classification of hyperspectral remote sensing images based on sequential joint deeping learning model," *IEEE Access*, vol. 8, pp. 71353–71363, 2020.

[21] X. Zhang, Y. Yang, Z. Li, X. Ning, Y. Qin, and W. Cai, "An improved encoder-decoder network based on strip pool method applied to segmentation of farmland vacancy field," *Entropy*, vol. 23, no. 4, p. 435, 2021.

[22] Z. L. Yang, S. Y. Zhang, Y. T. Hu, Z. W. Hu, and Y. F. Huang, "VAE-Stega: linguistic steganography based on variational auto-encoder," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 880–895, 2020.

[23] X. Ning, X. Wang, S Xu et al., "A review of research on co-training," *Concurrency and Computation: Practice and Experience*, vol. 24, 2021.

[24] Y. Kim and D. Kim, "A CNN-based 3D human pose estimation based on projection of depth and ridge data," *Pattern Recognition*, vol. 106, Article ID 107462, 2020.

[25] M. Dantone, J. Gall, C. Leistner, and L. Van Gool, "Human pose estimation using body parts dependent joint regressors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3041–3048, Portland, OR, USA, June 2013.

[26] Y. Tian, C. L. Zitnick, and S. G. Narasimhan, "Exploring the spatial hierarchy of mixture models for human pose estimation," in *Proceedings of the European Conference on Computer Vision*, pp. 256–269, Springer, Berlin, Heidelberg, October 2012.

[27] S. Johnson and M. Everingham, "Learning effective human pose estimation from inaccurate annotation," in *Proceedings of the IEEE Conference*, pp. 1465–1472, IEEE, Shanghai, China, June 2011.

[28] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, "Poselet conditioned pictorial structures," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 588–595, Portland, OR, USA, June 2013.

[29] 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.

[30] X. Yu, Y. Chu, F. Jiang, Y. Guo, and D. Gong, "SVMs classification based two-side cross domain collaborative filtering by inferring intrinsic user and item features," *Knowledge-Based Systems*, vol. 141, pp. 80–91, 2018.