Hindawi
Mobile Information Systems
Volume 2022, Article ID 8455164, 8 pages
https://doi.org/10.1155/2022/8455164

Research Article

A Novel Deep Learning-Enabled Physical Education Mechanism

Weiqi Wang[1] **and Jianan Jiang**[2]

[1] *Fuyang Normal University, Fuyang 236041, China*
[2] *University of Science and Technology LiaoNing, Anshan 114051, China*

Correspondence should be addressed to Jianan Jiang; lnkdjjn@ustl.edu.cn

Received 23 December 2021; Revised 14 January 2022; Accepted 17 January 2022; Published 8 March 2022

Academic Editor: Jianhui Lv

Race walking is one of the key events in the Tokyo Olympic Games, and also one of the strengths of China in athletics events. In recent years, China has made remarkable achievements in various race-walking competitions. However, with the improvement of the performance of race walkers, more and more technical problems have emerged, and the number of fouls due to nonstandard movements has increased significantly. It is a pity that athletes are disqualified for technical fouls in long-distance race-walking competitions. Therefore, it is necessary to introduce scientific training methods to help coaches strictly monitor the training process of athletes and accurately detect their standard degree of action in real-time. This paper mainly proposes a novel mechanism for foul recognition in race walking based on deep learning. Firstly, the image frames in the video are preprocessed by the Yolo algorithm to obtain the athletes' separated images. The U-Net network mixed with the attention mechanism is used to detect the athletes' actions to identify fouls and nonstandard actions, so as to assist the coach to identify the athletes' nonstandard actions in training and adjust them in time. Experiments show that the above method can identify the foul actions and nonstandard actions of multiple athletes in training at the same time quickly, and the recognition accuracy is higher than human eyes. It is more conducive to assist the coach to monitor and standardize the athletes' actions in the long-term training process, so as to reduce the error rate and improve the performance.

1. Introduction

Race walking originated in Britain in the 19th century and was developed on the basis of daily walking. The rules stipulate that the supporting legs must be straight, the two legs move forward alternately, keep uninterrupted contact with the ground, and do not leave the ground at the same time at any time, so as to ensure that there is no "flying" phenomenon, which is also the main difference between race walking and running. It is judged by the time it takes to complete the race. The normal walking speed is about 5 kilometers per hour, and the race walking is much faster. Men's walking race became an official event of the Olympic Games in 1908 and the women's walking race began in the Czech Republic in 1932. Currently, the longest distance in the Olympic event is 50 kilometers for men and 20 kilometers for women. The walking race is also the traditional advantage of the Chinese team in track and field events. From the 1984 Los Angeles Olympic Games to the 2016 Rio Olympic Games, China has won a total of 8 gold medals in track and field events of the Olympic Games, including 5 from walking race. By the end of the 2019 Doha track and field world championships, Chinese women's walking athletes have achieved four championships in five years in the 20 km walking event and made remarkable achievements.

Since race walking competition is usually time-consuming, there are very strict requirements for athletes' physical quality, psychological quality, tactics, and technical actions, especially for long-distance race walking, such as 20 km and 50 km, it is difficult for ordinary people to complete the whole process normally, athletes have to complete the game as soon as possible on the premise of maintaining the characteristic posture, the consumption of athletes in the competition is very large, which is not only testing the endurance of athletes but also testing the technical level of athletes. However, it is inseparable from the usual hard training to keep the standard actions in such a long competition. With the improvement of the performance of race walkers, the phenomenon of fouls in the competition is also gradually increasing. The influence of fouls on athletes in the competition is very huge. When the race walkers' actions show signs of violating the race-walking technology, they will be given a yellow card warning and the walking referee will give a red card when he watches visible flying or knee bending. When the same athlete receives three red cards, he will be disqualified. At present, in the competition, the referees cannot judge by any equipment. They can only rely on their own eyes to judge whether the athlete violates the rules. According to the literature [1], we know that there will be a certain misjudgment rate if he totally depends on his eyes. The referee may be affected by the factors such as the judgment distance and angle in the process of judgment. Therefore, for the athlete, it is very important to strictly regulate their own actions in daily training, resume the training process timely, adjust nonstandard actions, and avoid foul or be a misjudgment.

In the traditional training process, coaches use naked eyes to observe or through video resumes to help athletes correct their actions. However, due to the limitations of human eyes, it is not efficient enough to judge by naked eyes completely and there are many misjudgments and omissions. With the increasing development of artificial intelligence technology, video image capture and wearable device technology are constantly promoting the development of sports behavior recognition. There are many methods to sample and judge athletes' behavior information based on wearable technology [2], but this method is less intuitive than the video capture method, which is not conducive to playback analysis and the recognition accuracy of actions is also not high enough. With the rapid development of the neural network in the field of computer vision, more and more methods to assist pedestrian detection and behavior recognition through image segmentation appear, and u-net is undoubtedly the most commonly used and simplest segmentation model to complete image segmentation. It is simple, efficient, easy to build, and can be trained from small data sets, It can help us quickly identify the athletes in training in the video, but if we want to quickly obtain the most effective information and identify the athletes' foul actions, we need to introduce the attention mechanism. The attention mechanism is being more and more widely used in various fields. Considering that the race-walking competition and training are usually carried out by multiple people at the same time, it is also a problem to be solved to identify whether the actions of multiple athletes are all standardized. Yolo algorithm can well preprocess the image frames and help us solve the problem of separation of human and scene for multi athletes, U-Net network combined with attention mechanism is used to identify the foul actions of all athletes in the video stream containing multiple athletes.

Accordingly, the contributions of this paper are summarized as follows: (i) The video stream of race walking training is obtained by recording video, the image frames in the video stream are preprocessed by the Yolo algorithm to obtain the image of athletes separate with a scene, and the u-net network combined with attention mechanism is used to identify the foul actions and nonstandard actions in the video stream of multiple athletes; (ii) test the effectiveness and accuracy of the method in (i) through experiments to see whether it can effectively detect the foul actions and nonstandard actions in a single athlete or multi athletes race walking, and compare the results with the detection results from naked eyes to judge whether the accuracy of foul recognition has been improved, so as to determine whether the method has a positive impact on auxiliary training.

The rest of this paper is organized as follows: The related work is in Section 2. The proposed method is introduced in Section 3. Section 4 shows the implementation method of foul actions recognition. The experimental design and result analysis are introduced in Section 5 and Section 6, respectively. Finally, Section 7 concludes this paper and gives future research directions.

2. Related Work

In paper [1], the author studies the relationship between the observation state of race walking judges and CFF and shows that the CFF of eyes should be regarded as the physiological indexes in choosing the qualified race walking judges. In article [2], the author analyzed the video data recorded of 30 men during the World Cup walking competition and came to the conclusion that step length and stride length are the key areas that must be coordinated in long-distance walking competition.

In recent years, technology plays an important role to help training and judgment in sport; in paper [3], the author proposes the use of a wearable inertial system to derive novel biomechanical indices for the assessment of performance and infringements in race-walking, where the result shows that these indices can be implemented on a wearable inertial system to assist training and judgment in race-walking. Paper [4] shows the preliminary result on the use of a wearable inertial system for the assessment of performances and infringements in race-walking in 2019. Current judging of race walking in international competitions relies on subjective human observation to detect illegal gait, which naturally has inherent problems, the research in [5] aims to determine whether an inertial sensor could improve accuracy based on monitoring every step of seven races walkers in training and competition. In paper [6], Tabori et al. have proposed machine-learning algorithms for automatic detection of infringements (both LOGC and bent knee). Paper [7] presents a new motion analysis protocol for race-walking, through setting up a motion capture system and a force platform to record both kinematic and dynamic aspects of the athletes' action to detect infringement of the rules based on the measure of knee flexion-extension and the loss of ground contact. Paper [8] placed an inertial sensor at L5/S1 of the vertebral column of an Italian national team athlete to acquire timing measurements of the LOGC to validate an inertial system able to detect the loss of ground contact (LOGC) in race-walking in real training conditions, results show that the inertial system can improve the accuracy in detecting the visible LOGC. Paper [9] also aims to develop an innovative approach based on a wearable inertial system, which enables objective evaluations on the loss of ground contact in race-walking.

Paper [10] proposes an attention mechanism LSTM framework for human action recognition in videos. In [11], a coattention model-based recurrent neural network (CAM-RNN) is proposed, where the CAM is utilized to encode the visual and text features, and the RNN works as the decoder to generate the video caption. Paper [12] introduces a regularized attention mechanism for graph attention networks. Paper [13] proposes an attention mechanism LSTM Framework for Human Action Recognition in Videos.

In the paper [14], the author presents temporal deformable convolutional encoder-decoder networks that fully employ convolutions in both encoder and decoder networks for video captioning. In paper [15], the author proposes a two-stream framework based on combinational deep neural networks to extract both temporal and spatial features by exploring the usage of 3D convolutional networks on both raw RGB frames and motion history images, and tune the weights of different feature channels since the network is trained end-to-end from learning combinational encoding of multiple features to LSTM-based language model. Paper [16] proposes a Multimodal Memory Model (M3) which builds a visual and textual shared memory to model the long-term visual-textual dependency and furthermore guides visual attention on described visual targets to solve visual-textual alignments. In [17], a video summarization technique that uses motion descriptors computed in the compressed domain is described. It can either speed up conventional color-based video summarization techniques or rapidly generate a key-frame-based summary by itself. A key-frames extraction method is proposed with Kekere's Proportionate Error (KPE) codebook generation techniques of vector quantization with ten different codebook sizes and two color-spaces (RGB and KLUV) in [18].

Paper [19] is talking about the usage of the LSTM encoder-decoder algorithm for detecting anomalous ADS-B messages and [20] presents an encoder-decoder model for automatic video captioning based on Yolo Algorithm. Paper [21] proposes an improved YOLOv3-tiny for object detection based on the idea of feature fusion, compared with YOLOv3-tiny, the accuracy of the improved network structure is increased by 6.3%, and the detection speed is 31.8fps in 2019. Paper [22] proposes a faster detection method for real-time object detection based on a convolution neural network model called Single Shot MultiBox Detection (SSD), increase the accuracy in identifying objects. The research of paper [23, 24] is based on real-time multiple object detection through YOLO.

3. The Proposed Method

In recent years, with the continuous development of computer hardware, the computing power required by deep learning has been satisfied, so it has been widely applied in the field of computer vision research, surpassing the results of traditional machine learning. CNN convolutional neural network is a kind of feed-forward neural network with deep structure and convolution computation. As one of the representative algorithms of deep learning, it has achieved good results in large-scale recognition tasks, breaking through traditional classification methods and outperforming recognition naked eyes. However, at present U-Net is the most widely used convolutional neural network architecture in the field of image segmentation. The U-Net architecture combined with attention mechanism can focus limited attention on the effective features of the recognized athletes, thus saving resources rapid access to pertinent information to distinguish whether the athletes' actions are standardized. The overall architecture of this method is shown in Figure 1.

As it can be seen from Figure 1, there are four modules in our pipeline.

In the image acquisition module, the images during race walking competition or training through video are sampled to construct the training and testing dataset. It can be achieved by modern image processing tools, such as the OpenCV.

In the image matting module, the extracted images are preprocessed by YOLO-V3, such that the background can be filtered and only the 2D bounding box of the human is left.

The effective features are provided by the attention based on the U-Net structure, so as to output a three-dimensional probability vector composed of normal, foul, and nonstandard action to identify the three kinds of actions.

3.1. The Yolo Algorithm.

Object detection is a key task in the field of computer vision, which can be regarded as the combination of image classification and location. Since Ross Girshic K proposed R-CNN in 2013, Fast R-CNN, Faster R-CNN, YOLO, and other algorithms for object detection have been proposed one after another. The full name of Yolo is you only look once, which means that you can identify the category and location of objects (including pedestrians) in the picture only by browsing once. Compared with R-CNN series algorithms, Yolo does not need to generate a large number of candidate boxes first and then use a convolutional neural network to classify and regress the candidate boxes. It takes the whole image as the input of the network, and directly regresses the location and category of the bounding box at the output layer, so as to predict the object category and location in one step, saving time and hardware cost; however, it also can detect from real-time video and has stronger generalization ability.

In terms of implementation method, Yolo$*$s convolutional neural network divides the input image into $S \times S$ grid, Then, uses each cell to detect the targets whose center point falls in the grid. Each cell will predict N bounding boxes and confidence score which includes the possibility $P(\text{object})$ of bounding boxes containing targets and the accuracy $A(\text{object})$ of bounding boxes. When the bounding box is a background, $P(\text{object}) = 0$, When the bounding box contains a target, $P(\text{object}) = 1$. Therefore, the confidence score can be defined as C, $C = P(\text{object}) \ast A(\text{object})$. The size and position of the bounding box are represented by four values: (X, Y, W, H), where (x, y) is the central coordinate of the

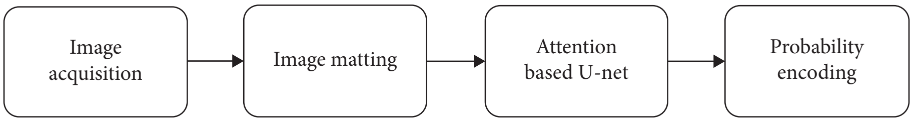

Figure 1: The proposed pipeline for action recognition.

bounding box, and W and H are the width and height of the bounding box. Thus, the predicted value of each bounding box contains five elements: (X, Y, W, H, C), the first four represent the size and position of the bounding box, and the last one is the confidence score. Border category confidence score is actually the conditional probability under each border-box confidence score, which can be expressed as $P(\text{class}|\text{object})$, It represents the possibility that the target object in the bounding box belongs to each category and whether the bounding box matches the target object. In general, the prediction frame of the network is filtered according to the confidence score of the category to identify the target object (including people).

3.2. The U-Net Network. Semantic segmentation is an important branch in the field of image processing and machine vision. It needs to judge the category of each pixel of the image and segment it accurately. At present, it is widely used in the field of automatic matting. U-Net appeared in 2015, which is the most commonly used segmentation model now, it can be regarded as a simplified structure based on the FCN model. It is efficient and simple, can be trained with small data sets, and has higher accuracy than the FCN network. It mainly uses a U-shaped network structure to obtain context information and location information. Figure 2 shows the network model of U-Net [25].

U-Net is actual a U-shaped structure of encoder-decoder. As shown in Figure 2, the left side is the convolution layer for feature extraction, and the right side is the upper sampling layer. The u-net structure contains 4 revolutionary layers and 4 corresponding upsampling layers. Therefore, we can initialize the weight first and then train the model, or use the existing convolution layer mechanism and the trained weight value through the following up sampling layer to train. The feature map obtained from each convolution layer of the U-Net network will be concatenated to the corresponding upper sampling layer so that the feature map of each layer can be effectively used in subsequent calculations, that is, skip connection. Thus, the final feature map contains both high-level features and low-level features, which improves the accuracy of the model.

3.3. Attention Mechanism. Attention mechanism was first applied in computer vision and later developed in the field of NLP. This mechanism focuses limited attention on key information, so as to save resources and quickly obtain the most effective information.

In fact, attention mechanism is a process of filtering out a small amount of important information from a large amount of information, ignoring unimportant information, and better-weighted fusion of information. Focus on its corresponding value according to its weight, the larger the weight, the more aggregation. And the weight indicates the importance of information. Value is its corresponding information, as shown in Figure 3.

The specific calculation process of attention mechanism can be roughly summarized into three stages: (i) calculate the correlation between query and key; (ii) normalize the value obtained from (iii) to obtain the weight coefficient; (iv) assign the weight and summarize the value according to the weighting coefficient.

In stage (i), different functions and calculation mechanisms can be introduced to calculate the correlation by calculating the vector dot product or vector cosine similarity of Query and Key, or by introducing an additional neural network; In stage (ii), a calculation method similar to SoftMax is introduced to convert the values in the first stage. While normalizing, the original calculated values can be sorted into a probability distribution in which the sum of the weights of all elements is 1, highlighting the weights of important elements; In stage (iii), the weight coefficient corresponding to value (i) is calculated in stage (ii) is introduced for weighted summation to obtain the Attention Value for Query.

4. Proposed Method

4.1. Athletes Recognition Based on Yolov3. In this paper, we mainly detect athletes in multi-person race walking competitions or training by pedestrian detection based on yolov3. We use a convolutional neural network to detect athletes in the video. After the video frame is input, it first enters the target detection network based on Yolov3, extracts the features through darknet-53, then performs up sampling and feature fusion, finally performs regression analysis to obtain the prediction frame information, as shown in Figure 4.

As a single-stage detector, Yolo directly classifies and predicts the objects at each position of the feature map without generating candidate regions. In this paper, we use the labelme pedestrian data set, and then through the built Yolo algorithm to generate the model and train, we can accurately identify multiple race walkers and separate the race walkers from the image. Finally, several bounding boxes are output , resize output to $256 * 256$ for actions' recognition of next step.

4.2. Attention Based U-Net Network. Based on the discussion of the above methods, we propose the attention-based U-Net network structure to extract the action state of athletes. For the specific network structure, please refer to Figure 5 below.

According to Figure 5, the network input is 256×256 image frames extracted by YOLOv3, and the output is a

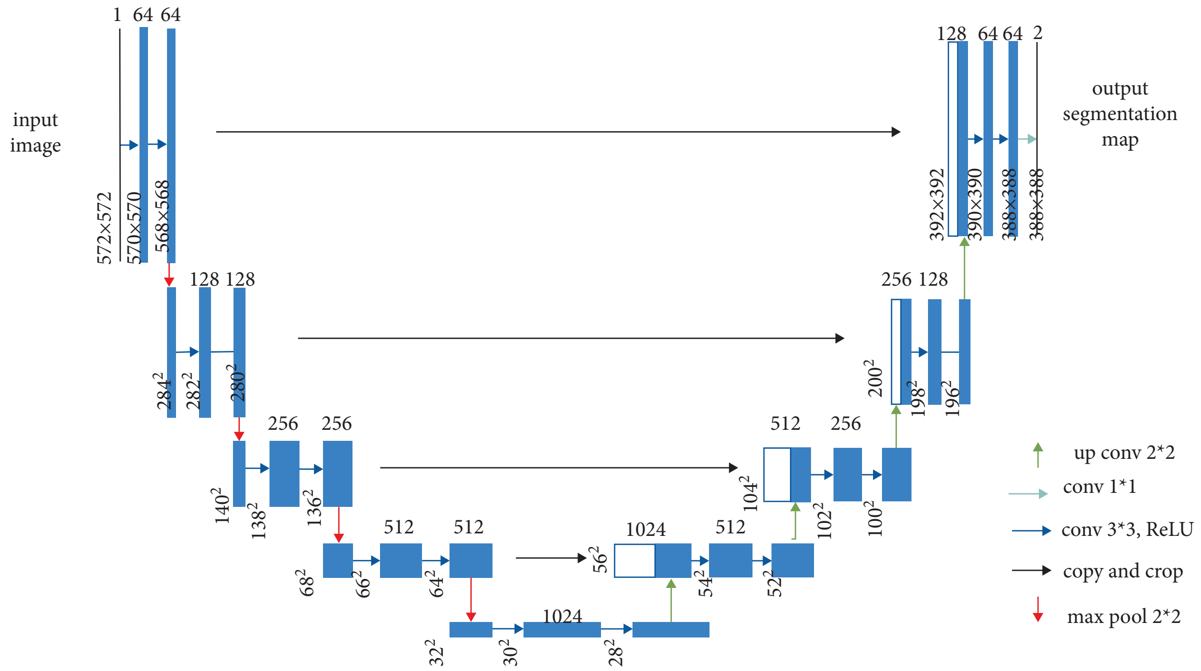

FIGURE 2: U-Net network model.

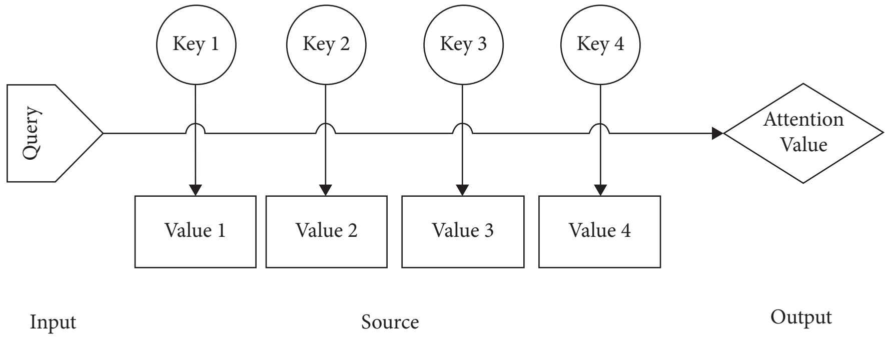

FIGURE 3: Attention mechanism.

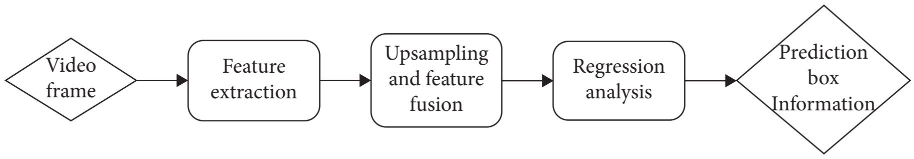

FIGURE 4: Pedestrian detection process based on Yolov3.

three-dimensional vector X, assuming $X = [X_1, X_2, X_3]\ T$, where X_1 represents the probability of normal action judged from this frame, X_2 represents the probability of nonstandard action judged from this frame, X_3 represents the probability of foul action judged from the frame, and they satisfy $X_1 + X_2 + X_3 = 1$.

The network structure combines the encode part of U-Net architecture and attention mechanism to centrally learn the lower limbs' actions of race-walking athletes and extract the corresponding features. The network front end consists of full convolution and max pooling. Among them, $3 \times 3 \times 16$ represents a 3×3 full convolution network, 16 represents the number of output channels, Relu is used as the activation function, the stripe is 1, zero padding is used, and the number of padding is 1. It can be seen from the figure that the space is down-sampled three times, and the features sampled in the first two times pass through the Attention module. The attention module here adopts CBAM in the document [26–29]. The aggregation layer is used to fuse the features of three scales. In order to ensure consistent resolution, 4x and 2x up samplings are used to align the features [26]. Max pooling channel represents maximum pooling in the channel layer,

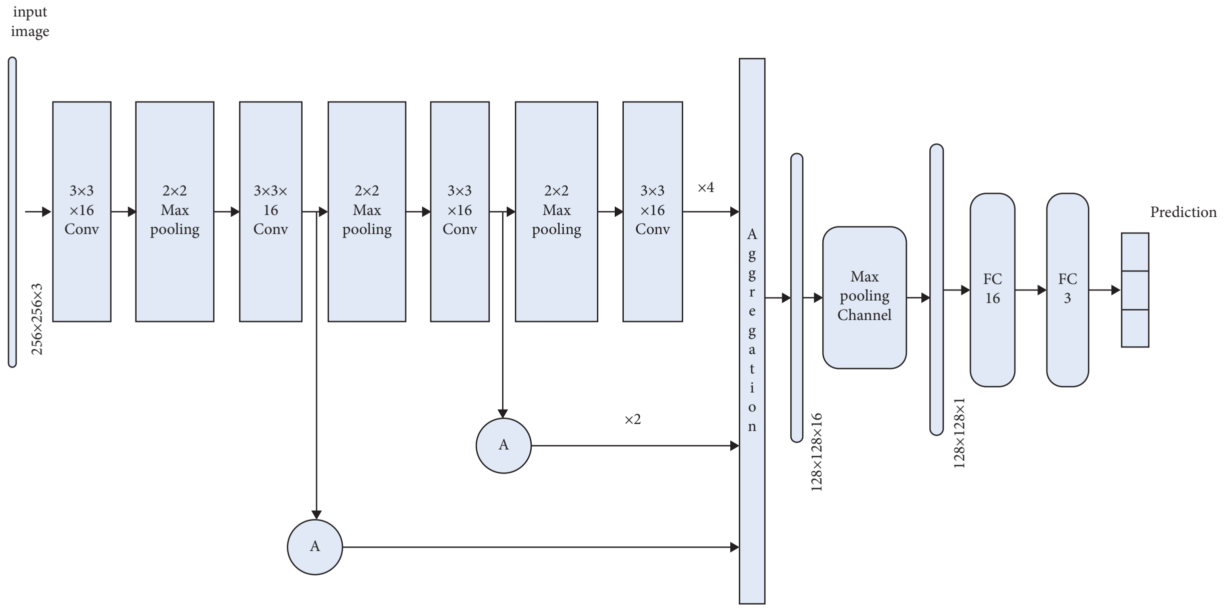

Figure 5: Attention-based U-Net network structure.

which can quickly integrate the information of features. The last two layers FC16 and FC3 represent the whole connection layer of 16 neurons and the whole connection layer of 3 neurons, respectively. Finally, FC3 neurons can obtain a probability value of 0~1 through sigmoid function.

4.3. Training Loss. Each task is associated with a loss function [27]. We select loss function for the task of action classification. Multi classification is actually an extension of two classifications:

$$L = \frac{1}{N}\sum_i L_i = -\frac{1}{N}\sum_i \sum_{c=1}^{M} w_{ic}\log\left(X_{ic}\right). \quad (1)$$

In formula (1): N is the length of encoding probability, M represents a number of classifications, $w\ ic$ represents Symbolic function (0 or 1), if the actual classification of sample i is equal to C, the value is 1, otherwise, the value is 0; X_{ic} represents prediction probability of observation sample i belonging to classification C.

Now, we use these expressions to calculate the value of the loss function in the above example.

Sample1 loss = −(0 + 0 × log 0.3 + 1 × log 0.4) = 0.91.

Sample2 loss = −(1 × log 0.1 + 0 × log 0.2 + 0 × log 0.7) = 2.30.

5. Experimental Design

5.1. Training Model

5.1.1. Experimental Dataset Sources. In this paper, the dataset is obtained by collecting the data from the Race Competition and daily training. Namely, 30% of the 500 groups come from the 20 km race walking competition, and the last comes from the data recorded in the process of athletes' daily training and simulated competition.

5.1.2. Network Training. As mentioned earlier, this paper will use Yolo as the input of the motion detection network, so there is no retraining. In network training, the batch size is 32 and the learning rate is 0.001. Adam optimizer is used to learn 200 epoch.

5.2. Validity Verification. We record the 20 km race walking training process of 5 groups of women and 5 groups of men by video. Six athletes participate in each training, and the race-walking athletes are required to make fouls and use nonstandard actions from time to time. The number of fouls and nonstandard actions in the training of each group of athletes is determined by our Yolov3 algorithm combined with an attention mechanism based on the U-Net model and naked eyes tracking by coaches, respectively.

6. Experimental Results and Analysis

6.1. Experimental Results. The experimental results of action classification of the 20 km race walking training process are shown in Tables 1–3, here group1-group5 are women groups and group6-group10 are men groups.

6.2. Result Analysis. Through the above experimental results, we can observe the following:

(i) Through the new method based on deep learning adopted in this paper we can successfully identify the foul and nonstandard actions of each athlete in the multi person race walking.

(ii) Using the new discrimination method proposed in this paper, more fouls and nonstandard actions are identified in each race, about 8% higher in an average form. This shows that in the process of 20 km race walking, the accuracy rate of identifying fouls and

Table 1: Foul and nonstandard action recognition.

NO.	Group1	Group2	Group3	Group4	Group5	Group6	Group7	Group8	Group9	Group10
Foul by naked eyes	85	76	62	82	137	62	113	125	107	132
Foul by deep learning	97	86	98	90	190	73	131	200	183	194
Nonstandard by naked eyes	98	102	121	109	198	89	99	201	207	236
Nonstandard by deep learning	118	128	149	121	249	119	128	248	210	271

Table 2: Accurate rate comparison between different methods for Foul action estimation.

NO.	Group1	Group2	Group3	Group4	Group5	Group6	Group7	Group8	Group9	Group10
GT (Ground truth)	99	88	100	91	191	75	132	203	185	157
Eyes	86.7%	86.3%	88.6%	90.1%	84.6%	82.7%	85.0%	83.9%	88.4%	82.5%
AI	98.9%	97.7%	97.1%	98.9%	98.7%	97.3%	98.5%	97.3%	97.5%	97.5%

Table 3: Accurate rate comparison between different methods for Nonstandard estimation.

NO.	Group1	Group2	Group3	Group4	Group5	Group6	Group7	Group8	Group9	Group10
GT (Ground truth)	120	133	155	123	260	130	140	260	220	275
Eyes	80.8%	64.7%	63.2%	73.2%	190	73.1%	93.6%	76.9%	83.2%	79.2%
AI	98.3%	96.2%	96.1%	98.4%	249	91.2%	91.4%	95.4%	95.5%	98.5%

nonstandard actions by the new method is higher than that by the coach through naked eyes. Hence, it is more effective than human eyes to assist training.

7. Conclusions

In this paper, an attention-based network is proposed to identify the fouls and nonstandard actions of multiple athletes in the process of long-distance race walking, and the effectiveness of this method is verified by experiments. Through the experimental results, we know that the new method has less interference and higher accuracy than human eye recognition.

In the future research, we will continue to focus on improving the accuracy through training more data sets and reduce the complexity of the method so that we can not only use this method to identify actions to assist training, but also apply this method to the real race-walking competition to assist the referee to improve the accuracy of judgment and reduce misjudgment.

Data Availability

The data used to support the findings of the study are included in the paper.

Conflicts of Interest

The authors declare that there are no conflicts of interest in this paper.

Acknowledgments

This work was supported by 2020 Key Research Project of Humanities and Social Sciences in Anhui Universities Granted No. 2020jyxm1989.

References

[1] S. Xu and F. Jiao, "The observation state of race walking judge and CFF," *China Sport Science and Technology*, vol. 35, pp. 39–42, 1999.

[2] B. Hanley, A. Bissas, and A. Drake, "Kinematic characteristics of elite men's 50 km race walking," *European Journal of Sport Science*, vol. 13, no. 3, pp. 272–279, 2013.

[3] T. Caporaso, S. Grazioso, G. D. Gironimo, and A. Lanzotti, "Biomechanical indices represented on radar chart for assessment of performance and infringements in elite race-walkers," *Sports Engineering*, vol. 23, 2020.

[4] T. Caporaso, S. Grazioso, D. Panariello, G. Di Gironimo, and A. Lanzotti, "A wearable inertial device based on biomechanical parameters for sports performance analysis in race-walking: preliminary results," in *Proceedings of the II Workshop on Metrology for Industry 4.0 and IoT (MetroInd4.0&IoT)*, pp. 259–262, Naples, Italy, June 2019.

[5] J. B. Lee, R. B. Mellifont, B. J. Burkett, and D. A. James, "Detection of illegal race walking: a tool to assist coaching and judging," *Sensors*, vol. 13, p. 16, 2013.

[6] J. Taborri, E. Palermo, and S. Rossi, "Automatic detection of faults in race walking: a comparative analysis of machine-learning algorithms fed with inertial sensor data," *Sensors*, vol. 19, no. 6, 2019.

[7] G. Di Gironimo, T. Caporaso, D. M. Del Giudice, A. Tarallo, and A. Lanzotti, "Development of a new experimental protocol for analysing the race-walking technique based on kinematic and dynamic parameters," *Procedia Engineering*, vol. 147, pp. 741–746, 2016.

[8] G. D. Gironimo, T. Caporaso, G. Amodeo, and D. M. D. Maria, "Outdoor tests for the validation of an inertial system Able to detect illegal steps in race-walking," *Procedia Engineering*, vol. 147, pp. 544–549, 2016.

[9] G. D. Gironimo, T. Caporaso, D. M. D. Giudice, and A. Lanzotti, "Towards a new monitoring system to detect illegal steps in race-walking," *International Journal on Interactive Design and Manufacturing*, vol. 11, no. 2, pp. 1–13, 2016.

[10] C. Yan, Y. Tu, X. Wang, and Y. Zhang, "Corrections to "STAT: spatial-temporal attention mechanism for video captioning"," *IEEE Transactions on Multimedia*, vol. 22, no. 3, 830 pages, 2020.

[11] B. Zhao, Y. Li, and X. Lu, "Co-attention model based RNN for video captioning," *IEEE Transactions on Image Processing*, vol. 28, no. 99, pp. 5552–5565, 2019.

[12] U. S. Shanthamallu, J. J. Thiagarajan, and A. Spanias, "A regularized attention mechanism for graph attention networks," in *Proceedings of the ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 3372–3376, Barcelona, Spain, May 2020.

[13] C. I. Orozco, M. E. Buemi, and J. J. Berlles, "Towards an Attention Mechanism LSTM Framework for Human Action Recognition in Videos," in *Proceedings of the 2020 IEEE Congreso Bienal de Argentina (ARGENCON)*, pp. 105–109, Resistencia, Argentina, December 2020.

[14] J. Chen, Y. Pan, Y. Li, T. Yao, H. Chao, and T. Mei, "Temporal deformable convolutional encoder-decoder networks for video captioning," vol. 33, pp. 6–8, 2019, https://arxiv.org/abs/1905.01077.

[15] C. Zhang and Y. Tian, "Automatic video description generation via LSTM with joint two-stream encoding," in *Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR)*, pp. 753–756, IEEE, Cancun, December 2016.

[16] J. Wang, W. Wei, and H. Yan, "M3: multimodal memory modelling for video captioning," in *Proceedings of the 2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. pp.7512–7520, IEEE, Salt Lake City, UT, USA, June 2018.

[17] A. Divakaran, R. Radhakrishnan, and K. A. Peker, "Video summarization using descriptors of motion activity: a motion activity based approach to key-frame extraction from video shots," *Journal of Electronic Imaging*, vol. 10, no. 4, pp. 909–916, 2001.

[18] S. D. Thepade and P. H. Patil, "Novel video keyframe extraction using KPE vector quantization with assorted similarity measures in RGB and LUV color spaces," in *Proceedings of the 2015 International Conference on Industrial Instrumentation and Control (ICIC)*, pp. pp.2512–2520, IEEE, Pune, India, July 2015.

[19] E. Habler and A. Shabtai, "Using LSTM encoder-decoder algorithm for detecting anomalous ADS-B messages," *Computers & Security*, vol. 78, pp. 155–173, 2017.

[20] A. Shabtai and I. Habler, "Using LSTM encoder-decoder algorithm for detecting anomalous," *Ads-B Messages*, vol. 78, pp. pp1–8, 2019.

[21] H. Gong, H. Li, K. Xu, and Y. Zhang, "Object detection based on improved YOLOv3-tiny," in *Proceedings of the 2019 Chinese Automation Congress (CAC)*, pp. pp.1520–1528, IEEE, Hangzhou, China, February 2020.

[22] S. Kanimozhi, G. Gayathri, and T. Mala, "Multiple Real-time object identification using Single shot Multi-Box detection," in *Proceedings of the 2019 International Conference on Computational Intelligence in Data Science (ICCIDS)*, pp. pp.5512–5522, IEEE, Chennai, India, October 2019.

[23] B. Kumar, R. Punitha, and Mohana, "YOLOv3 and YOLOv4: multiple object detection for surveillance applications," in *Proceedings of the Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pp. 1316–1321, Tirunelveli, India, October 2020.

[24] M. Mahendru and S. K. Dubey, "Real time object detection with audio feedback using Yolo vs. Yolo_v3," in *Proceedings of the 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, p. pp.734, Noida, India, March 2021.

[25] P. Li, L. Zhang, J. Qiao, and X. Wang, "A semantic segmentation method based on improved U-net network," in *Proceedings of the 2021 4th international conference on advanced electronic materials, computers and software engineering (AEMCSE)*, pp. 600–603, Changsha, China, August 2021.

[26] S. Woo, J. Park, and J. Y. Lee, "CBAM: convolutional block Attention module," in *Proceedings of the European conference on computer vision*, pp. 541–549, Munich, Germany, July 2018.

[27] V. L. Tran and H. Y. Lin, "3D object detection and 6D pose estimation using RGB-D images and mask R-CNN," in *Proceedings of the 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1890–1898, IEEE, Glasgow, UK, December 2020.

[28] M. Braun, Q. Rao, Y. Wang, and F. Flohr, "Pose-RCNN: joint object detection and pose estimation using 3D object proposals," in *Proceedings of the IEEE International Conference on Intelligent Transportation Systems*, pp. 178–189, IEEE, Rio de Janeiro, Brazil, December 2016.

[29] Z. Zhao, K. Chen, and S. Yamane, "CBAM-Unet++: easier to find the target with the attention module CBAM," in *Proceedings of the 2021 IEEE 10th Global Conference on Consumer Electronics(GCCE)*, pp. 655–657, Kyoto, Japan, December 2021.