Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 2141194, 9 pages
https://doi.org/10.1155/2022/2141194

Research Article

Research on Intelligent Estimation Method of Human Moving Target Pose Based on Adaptive Attention Mechanism

Meishuang Ding[1] **and Jing Zhao**[2]

[1]*Employment and Entrepreneurship Guidance Center, Hefei Gongda Vocational and Technical College, Hefei, China*
[2]*School of Innovation and Entrepreneurship, Anhui Vocational and Technical College of Mechatronics, Wuhu, China*

Correspondence should be addressed to Meishuang Ding; 617541131@qq.com

Received 4 November 2021; Revised 19 January 2022; Accepted 2 February 2022; Published 23 February 2022

Academic Editor: Xiaohui Yuan

In daily physical education, posture performance is an important basis for making excellent results. This paper explores an intelligent method to estimate the target pose based on adaptive attention mechanism. First, the regional attention is iteratively generated from a global level to a local level based on the attention mechanism. Human decision-making patterns are imitated to evaluate the effectiveness of regional attention in real time. The level of attention mechanism is adaptively adjusted and focused layer by layer to achieve precise target detection and tracking. Second, with the target frame obtained from each frame, the pose estimation algorithm finds the key points of human body, enabling the human body pose optimization strategy to solve the crossover problem of the key points. Results of experiments on sports video images show that the proposed method has a higher accuracy in pose estimation than other algorithms and can help sportsmen adjust their training methods scientifically.

1. Introduction

Nowadays, sports results account for an increasing proportion of a student's total academic results. In daily physical education, posture performance is an important basis for giving excellent results. As physical teachers have always acquired the information about pose of students via field observation, there are many problems such as one-sidedness of information extraction and low teaching efficiency. With the development of the video technology, it is likely to intelligently analyze the pose using deep learning, which provides technical support for daily physical education.

Pose analysis includes target tracking and pose estimation, between which the former, as the prerequisite for analyzing and detecting the region of interest in the video, has been widely applied to a variety of scenarios. Target tracking algorithms can be roughly divided into the following two: (1) the algorithms that glean the information about target position, size, etc., by predicting the score map of the candidate area, but cannot perceive the aspect ratio of the target [1, 2]; (2) the algorithms that position the target accurately by carrying out bounding box regression, and making use of deep learning to predict the aspect ratio and adjust the predictive box [3, 4]. Among those deep learning-based target tracking methods, MDNet [5] is a masterpiece of early CNN-based tracking algorithms, which uses multidomain network branches to fit different target objects and then calculates the final result by external frame regression. SiamRPN [6] constructs a Region Proposal Network (RPN) structure based on twin network, where the template image and the search region use the exact same convolutional network to extract the feature map, the classification and regression are carried out through two independent network branches to determine the target location and size, and then the external frames are optimized by regression results. On the other hand, Fully Convolutional Network (FCN) [7] is the classical network structure in semantic segmentation, where all network layers use convolutional layers, and the original image size is recovered by upsampling layers, so that the output segmented image size is exactly the same as the input image size. Those target tracking algorithms are still restricted by appearance deformation, light fluctuation, fast

motion, and other problems. According to findings of neuroscience researches, human brain cannot receive all image information visually input. It responds to special areas and ignores the remainder until the response is completed. Such rapid response mechanism is called the attention mechanism [8, 9], which is completely identical to target tracking in essence. The attention mechanism has become an important part of deep web designing and has been widely applied to speech recognition, language translation, and target tracking.

Accurate pose estimation is a prerequisite for detection of target behaviors in video images. In the early pose estimation algorithms, complex structured prediction was performed, and global and local features were fused to predict the pose in a globally uniform manner [10–12]. With the advent of deep learning, recent algorithms for pose estimation generally use ConvNets as the main module to predict the key points of human pose either from the top down or from the bottom up [13–16]. Top-down approaches usually use advanced human detectors to first detect all people from the image and then scale the image block containing a single person to a fixed size to send to a single person pose estimator for prediction. Bottom-up approaches, on the other hand, do not rely on human detectors and directly infer the location information of all human key points in the image and group the key points to obtain the pose of all people [17–19].

At present, artificial intelligence, represented by deep learning, has widespread applications [20]. Multilevel higher-order abstract features of the target object can be obtained from shallow to deep via operations of convolutional and pooling layers, which is consistent with the human perception mode running from globally to locally. However, the trial-and-error pattern of the target tracking network structure leads to a shortage of universal samples in invariant eigenspace and classification criteria generated by fixed architecture network. Faced with similar and dissimilar samples with different scales, the reliability of the target detection results of the fixed-scale learning model deteriorates. The traditional deep learning-based target tracking system can be boiled down to an open-loop system with uncertain image inputs and unassured target outputs [21–23]. Due to invariant eigenspace and the posterior statistics of the target tracking results, coupled with the absence of adaptive attention mechanisms, this system greatly differs from the human decision-making pattern that can adaptively adjust the multilevel eigenspace and validate the reliability of the target tracking results in real time.

Therefore, in order to solve the problems of the traditional target tracking network construction model and imitate the human cognitive model, this paper explores an intelligent target pose estimation method based on the adaptive attention mechanism. First, the regional attention is iteratively generated from a global level to a local level based on the attention mechanism. Human decision-making patterns are imitated to evaluate the effectiveness of regional attention in real time, and the level of attention mechanism is adaptively adjusted and focused layer by layer to achieve precise target detection and tracking. Second, combined with the target frame obtained from each frame, the pose estimation algorithm finds the key points of human body, enabling the human body pose optimization strategy to solve the crossover problem of the key points. Results of experiments on sports video images show that the proposed method has a higher accuracy in pose estimation than other algorithms and can help sportsmen adjust their training methods scientifically.

The innovations of this paper can be summarized as follows: (1) an iterative generation mechanism of regional attention from global to local is proposed; (2) a real-time evaluation method of regional attention effectiveness is constructed; (3) an adaptive adjustment mechanism of regional attention is established.

2. Target Detection Model Based on Adaptive Attention Mechanism

This paper proposes a target detection method based on the adaptive attention mechanism, which is composed of convolutional layers, target positioning network, target position feature evaluation and cropping module, etc. See Figure 1 for its structure and functions.

First, the target detection model completes the first round of reliability test, by sending the input image to the deep neural network M_0 for training and getting the deep feature map F_1. According to the entropy theory, a target evaluation function can be established to validate the reliability of F_1. If the threshold is satisfied, positioning of the target to be detected in the background image is considered reliable, and the trained deep network is stored as the first model into the detection model set of the attention mechanism; otherwise, the order of attention is adaptively adjusted to provide heuristic information for the positioning network. F_1 is then input into the positioning network weighted by the level of adaptive attention, and the target area is clipped according to the positioning characteristics of the output target. Adjust the clipped image to the size consistent with that of the input image, and start the second round of test; enter the deep neural network M_1 again to extract the deep feature map; repeat the above process to establish the second attention mechanism detection model, or generate a new image to be detected to initiate the next round of test until the test requirement is met or the number of focusing reaches the threshold. Finally, the image of the target precisely positioned is obtained.

2.1. Reliability Test of Target Detection Results and Adjustment Mechanism of Attention. Facing uncertain detection/recognition results, people will adaptively adjust their cognitive strategies to further perceive micro-information and get more reliable results. For the feature maps extracted by the Resnet-18 deep neural network, a reliability test is required to provide a quantitative basis for generation of the attention detection model and construction of the focus relocation model.

Define the train set as U, and let $U = \{U_1, U_2, \ldots, U_h\}$. Using Resnet-18 model M_0, U_j is given the fully connected

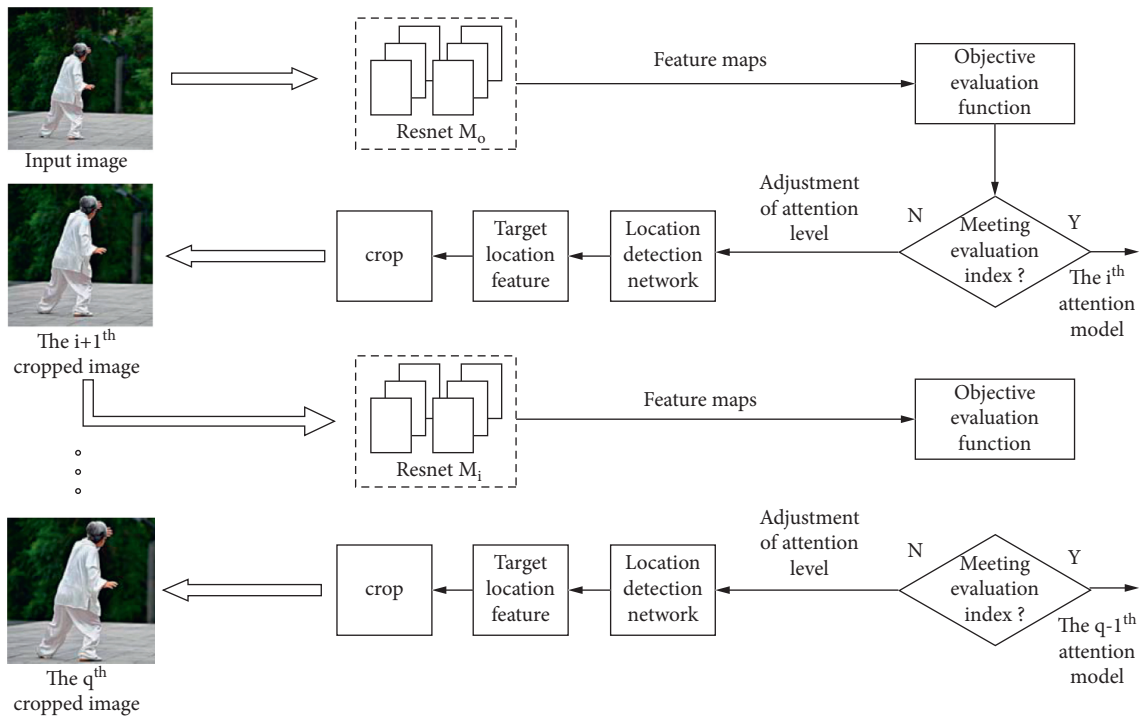

Figure 1: Target detection model based on adaptive attention mechanism.

feature vector $\mathbf{F}_j = [f_{j1}, f_{j2}, \ldots, f_{jd}]$ of the feature map, where d is the feature dimension. Take the fully connected feature vector set, given to U's real target area via the pretrained Resnet-18 model N, as $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_h]$, where $\mathbf{y}_j = [y_{j1}, y_{j2}, \ldots, y_{jd}]$, and define the target detection error entropy of U_j as

$$A_j = -\frac{\left\|1_{h\times 1} \times \mathbf{F}_j - \mathbf{Y}\right\|_2}{\sum_{j=1}^{h}\left\|1_{h\times 1} \times \mathbf{F}_j - \mathbf{Y}\right\|_2} \ln \frac{\left\|1_{h\times 1} \times \mathbf{F}_j - \mathbf{Y}\right\|_2}{\sum_{j=1}^{h}\left\|1_{h\times 1} \times \mathbf{F}_j - \mathbf{Y}\right\|_2}. \quad (1)$$

Larger entropy values v indicate less reliability of the feature information about the positioning area of the sample in the current deep network model. When the reliability threshold $\varrho = \lambda(\sum_{i=1}^{h} A_i)/h$ is exceeded, it is necessary to further focus and relocate the sample to perceive detailed information and save the current deep network model M_0 into the attention detection model set. Otherwise, the samples with smaller entropy values are deleted, and the train set is thus updated.

Let the proposed attention adjustment mechanism be $q \leftarrow q + \Delta q, 1 \le q \le q_{\max}, 1 \le \Delta q \le q_{\max}$, where Δq represents the increment to levels of attention adjusted according to feedback, and $q_{\max}$ represents the maximum level of attention.

2.2. Position Detection Network Model. The feature map generated by the deep neural network is used as the input of the position detection network. Under the constraint of the heuristic information of adaptive attention, the deep neural network is evolved synergistically with the position detection network through Alternating Training to get better detection results. Since the input item is the extracted deep feature map, it is possible to construct the position detection network with two fully connected layers without the need for carrying out complicated operations such as convolution, pooling, and activation.

Define the deep features of the input image U, which is extracted by the deep neural network, as $W_c * U$, where $*$ represents convolution, activation, pooling, and other operations, and W_c represents all parameters of the deep neural network. The position detection network model will more accurately deliver a target prediction area, expressed as follows:

$$\left[\frac{t_x, t_y, t_l}{\Delta q}\right] = g(W_c * U), \quad (2)$$

where t_x and t_y respectively represent the center coordinates of the square target prediction area, t_l half of the side length of the area, and $g()$ the position detection network. By analyzing the mechanism of human cognition, we can see that the granularity of cognitive information is changed in a nonuniform manner from macroscopically to microscopically. Therefore, in order to accomplish the nonuniform adjustment of the target prediction area, the transformation function is established for cognitive information granularity with decline characteristics, and the heuristic information of the increment to level of attention is used to weight the target prediction area.

In order to guarantee that an effective area is selected during the forward propagation, and the network can be optimized during the backward propagation, a two-dimensional rectangular function is proposed as a basis to perform the approximate clipping operation. Assuming that the upper left corner of the original image is the origin of

coordinates, the coordinates of the upper left and lower right corners of the target area are

$$t_{x1} = t_x - \frac{t_l}{\Delta q},\ t_{y1} = t_y - \frac{t_l}{\Delta q},$$
$$t_{x2} = t_x + \frac{t_l}{\Delta q},\ t_{y2} = t_y + \frac{t_l}{\Delta q}. \tag{3}$$

To clip, the original image and the coordinates are masked. In order to prevent images from discontinuity caused by the traditional masking method based on 0/1 dot product, this paper adopts the logistic regression function $\sigma(x) = 1/1 + e^{-kx}$ to construct a similar step function and uses a step that is about zero to produce an effect similar to 0/1 masking effect. When k is large enough, $x \geq 0, \sigma(x) \approx 1$; $x < 0, \sigma(x) \approx 0$. At this point, $\sigma(x)$ approximates a first-order function. If $x_0 < x_1$, $\sigma(x - x_0) - \sigma(x - x_1)$ is a smooth step function; if $x < x_0$ or $x > x_1$, $\sigma(x - x_0) - \sigma(x - x_1) \approx 0$; if $x_0 < x < x_1$, $\sigma(x - x_0) - \sigma(x - x_1) \approx 1$. Therefore, the following function is constructed:

$$U^{at} = U \odot M\left(t_x, t_y, \frac{t_l}{\Delta q}\right), \tag{4}$$

where U^{at} represents the image area that is worthy of attention, and $M()$ the calculated masking results of coordinates, as shown in the following:

$$M\left(t_x, t_y, \frac{t_l}{\Delta q}\right) = \left[\sigma(x - t_{x1}) - \sigma(x - t_{x2})\right] \cdot \left[\sigma\left(y - t_{y1}\right) - \sigma\left(y - t_{y2}\right)\right]. \tag{5}$$

After the original image is clipped, the useless information is deleted. As the features extracted from the localized image are weakened due to the reduced resolution, the bilinear interpolation method is adopted to enlarge the clipped target image to the size that the original image has so as to extract finer target features.

2.3. Backpropagation Process of the Target Detection Network. The output values t_x, t_y and t_l of the position detection network should be backpropagated for iterative optimization. Since these three coordinate parameters are the same in location, their backpropagation process is the same. Here, take t_x as an example, and calculate its derivatives by the chain rule:

$$L'_{rank}(t_x) = \frac{\partial L_{rank}}{\partial t_x} \propto D_{top} \odot \frac{\partial M\left(t_x, t_y, t_l/\Delta q\right)}{\partial t_x}, \tag{6}$$

where $\odot$ represents element-Wise product, and D_{top} represents the derivative propagated by the previous network. If $L'_{rank}(t_x) < 0$, t_x increases, or it decreases. Further, calculate $-\|L_{rank}(t_x)\|_2^2$ to determine the parameter optimization direction. $M'(t_x)$ represents the derivative of the mask function to t_x, which can be defined by the following piecewise function:

$$M'(t_x) = \begin{cases} < 0, & t_x \longrightarrow t_{x1}, \\ > 0, & t_x \longrightarrow t_{x2}, \\ = 0, & \text{otherwise}. \end{cases} \tag{7}$$

$M'(t_y)$ can be deduced in the same way. Since $M'(t_l/\Delta q)$ is positive on both sides of the boundary and negative inside the boundary, it can be defined as follows:

$$M'\left(\frac{t_l}{\Delta q}\right) = \begin{cases} > 0, & t_x \longrightarrow t_{x1} \text{ or } t_x \longrightarrow t_{x2}, \\ t_y \longrightarrow t_{y1} \text{ or } t_y \longrightarrow t_{y2}, & \\ < 0, & \text{otherwise}. \end{cases} \tag{8}$$

In summary, since the interval of negative values in the derivative is consistent with *that of* $M'(t_x)$, it is easy to conclude that $L'_{rank}(t_x)$ is positive. Similarly, $L'_{rank}(t_y)$ and $L'_{rank}(t_l/\Delta q)$ are also positive. Therefore, t_x, t_y, t_l will all decrease in the next iteration, indicating focused areas of attention. This is consistent with the human perception model.

2.4. Multimode Selection Method for Target Detection. The target detection model based on the adaptive attention mechanism can establish an attention detection model set, in which models are similar in structure but different in parameters. When the sample set is tested against the multi-attention detection model and since a model cannot give full play to its advantages in the premise of averagely weighted outputs, the decision-making attention mechanism is used according to the defined target detection error entropy to detect the outputs of the model set.

To obtain the output feature map and the focused images that have been clipped, the test set images are, respectively, input to each attention detection model in the model set.

According to formula (1), the target detection error entropy $A_q, 1 \le q \le q_{max}$ of q_{max} models with respect to the input image is calculated. When $A_{q_1} \le A_{q_2} \le \cdots \le A_{q_m}, q_1, q_2, \ldots, q_m \in [1, q_{\max}]$, select the model M_{q_1} corresponding to A_{q_1} as the best model, of which the output image is the final adaptive positioning image.

3. Pose Estimation Model Based on Deep Learning

3.1. Pose Estimation Model Based on Hourglass Network. DeepPose utilizes deep learning to transform the problem of pose estimation into the one related to joint point regression and performs estimated regression of the entire image to position each human joint point. Another type of method integrates multiscale features to generate a heatmap of the key points of the human body. Based on the above two methods, the hourglass network adopts the convolutional layer architecture of upsampling first and then downsampling. It integrates multiscale features at the bottom layer and the top layer of the network, improving the prediction accuracy of key points. The structure and the parameters of each module of the hourglass network are shown in Figure 2. The image resolution of the network input in Figure 2 is 256×256, Max pool in the figure stands for Down Sample, Up Sample stands for Up Sample, and Res stands for Residual Module, as shown in Figure 3.

During specific pose estimation, the stacked hourglass network is repeatedly upsampled and downsampled to generate the heatmap. The point with the highest heatmap score is taken as the key prediction point for each corresponding joint, and the key points are then converged to obtain the distribution diagram of key points of the human body. In addition, the symmetrical topology of the hourglass network helps predict the pose in both forward and backward directions, which improves the accuracy of pose estimation.

3.2. Evaluation Criteria for Human Pose Estimation. After the key points of the human body are predicted, it is necessary to evaluate the similarity with the ground truth (GT). The common Object Keypoint Similarity (OKS) index is defined as follows:

$$OKS_p = \frac{\sum_i \exp\left\{-d_{pi}^2 / 2S_p^2 \sigma_i^2\right\} \delta\left(v_{pi} = 1\right)}{\sum_i \delta\left(v_{pi} = 1\right)}, \tag{9}$$

where p represents the ID number of the target in the GT, i the number of the key points of the human body, d_{pi} the Euclidean distance between the key points of each person in the GT and the predicted result, S_p the scale factor of the p-th target, σ_i the i-th normalization factor of the key point, v_{pi} the visibility of the i-th key point of the p-th target, and δ the function deduced by selecting the visible key point.

From formula (9), the similarity between two key points of the human body can be deduced. If there are M persons in an image among whom N persons are predicted, an $M \times N$ OKS matrix should be constructed, of which the maximum value of each row is used as the OKS value of the i-th person. For several images in the test set, the Average Precision (AP) is used to measure the human pose estimation. The calculation is as follows: if OKS is greater than the threshold t, detection of the key points succeeds; otherwise, it fails. Count the number of OKS values that are greater than t, and calculate the ratio to the total number of OKS values.

3.3. PoseFix-Based Pose Estimation Optimization Strategy. Incorrect pose estimation may usually be caused by jitter, switching, and loss. Learning from DeepPose and facing most of the key points with good pose input or small deviations, the PoseFix model [24] focuses on the trusted key points and corrects the postprediction results. It can effectively solve the crossover problem. Its structure is shown in Figure 4.

PoseFix uses Gaussian distribution to represent the input poses in the form of heatmap, which are then spliced with the input image and sent to PoseFix. Such practice is not only suitable for convolution operations, but helps estimate key points by making use of the information around the features. Another advantage of PoseFix is that it has no connection with the method of generating the key points of the pose. Thus, it can be used as a postprocessing optimization strategy for any pose estimation method.

4. Experimental Results and Analysis

4.1. Experimental Data. In order to verify the feasibility and effectiveness of the proposed method, solo sports videos such as long jump, high jump, and 100-metre sprint are selected. Each video is about 30 minutes in length.

The video is converted and clipped by OpenCV into images with $1080 * 1080$ resolution, among which 70% are sampled randomly to construct the train set, and the remaining 30% make up for the test set. In this experiment, the maximum attention level of the target detection model is $q_{\max} = 6$, and the training epoch is set as 2000. In order to guarantee smoothness of the human detection/recognition process, let $\Delta q = 1$ to construct a compact set model for attention detection. Extract similar training sample sets with different qualities to multilevelly position the features. All experiments run under the environment of CPU i7-8700, 32G memory, and GTX1080Ti.

4.2. Experimental Results and Analysis. Figure 5 displays the target positioning results of the method proposed in this paper. Take the three-level attention as an example.

We can see from the figure that, in the busy background of the training field, the direct target detection method may damage detection accuracy and comprehensiveness, because possible crowds and light changes may have an impact on extraction of the main target features. The adaptive attention mechanism, however, allows the input image to be continuously positioned and focused and avoids the interference from useless areas, which helps represent and detect the features.

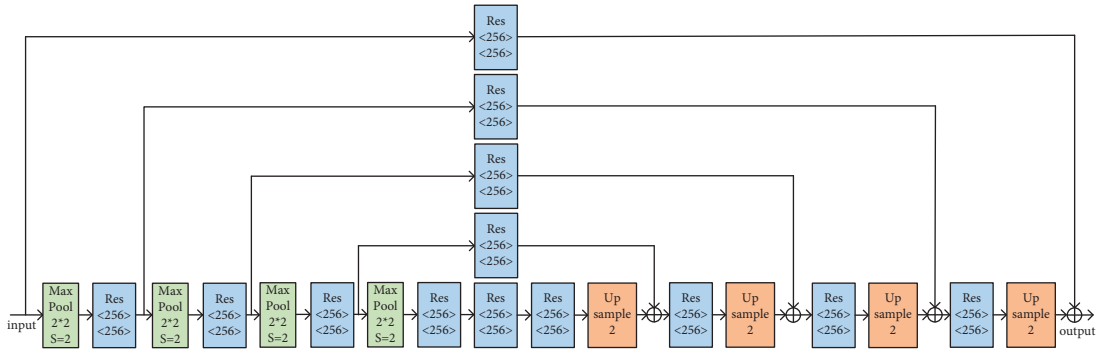

Figure 2: Hourglass network structure.

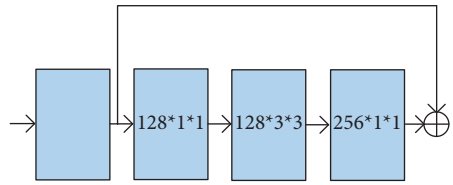

Figure 3: Residual module.

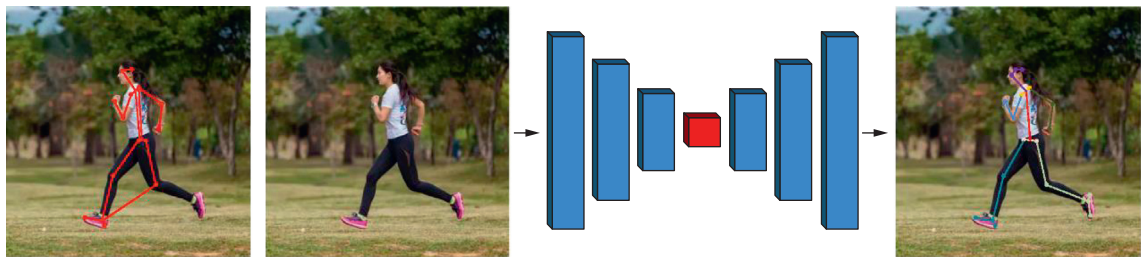

Figure 4: PoseFix structure.

Figure 6 shows the changing curve of target detection error entropy of the multilevel attention detection model under the condition of different maximum levels of attention q_{max}. Here, $q_{\max} = \{2, 3, 6\}$, where the Y- coordinate is the detection error entropy, and the X-coordinate is the number of iterations.

It can be seen from the figure that, with a decline in levels of attention focusing, the training efficiency of the model is increased. As the iteration process continues, the target detection error entropy of the model shows a rapid decline at $q_{\max} = 2$. When the number of iterations reaches 1000, the model presents the most reliable target detection results but works not stably. After the number of iterations exceeds 1000, the error entropy fluctuates, indicating that the attention mechanism malfunctions at $q_{\max} = 2$, and the target cannot be accurately positioned if some samples are focused only twice. Because of the lack of the ability in sample characterization, the feature extraction network should be further optimized. Similar problems occur at $q_{\max} = 3$. When the upper limit is raised ($q_{\max} = 6$), the training efficiency of the model decreases, because the model is no longer limited by the levels of attention and spends more time in seeking a more suitable focusing stage and extracting the marked features of the detected target. The reliability of target detection can be improved, as the iteration process goes on.

The precision and success rate of the proposed method with other target tracking algorithms are given in Table 1, where precision is defined as the distance between the predicted target position and the real target position $\text{dist} = \sqrt{(G_x - P_x)^2 + (G_y - P_y)^2}$. Tracking of this image is considered accurate if the distance is less than a threshold value. The success rate is calculated from the IoU score. If the IoU value is greater than a threshold, the tracking of this image is considered successful. Here, the precision threshold is set to 20 pixels, and the success rate threshold is set to 0.5.

As shown in Table 1, the proposed method is compared with seven mainstream tracking algorithms, including MDNet, SiamRPN, FCN, TADT, C-RPN, DAT, and SPM. It can be seen that the proposed method outperforms other

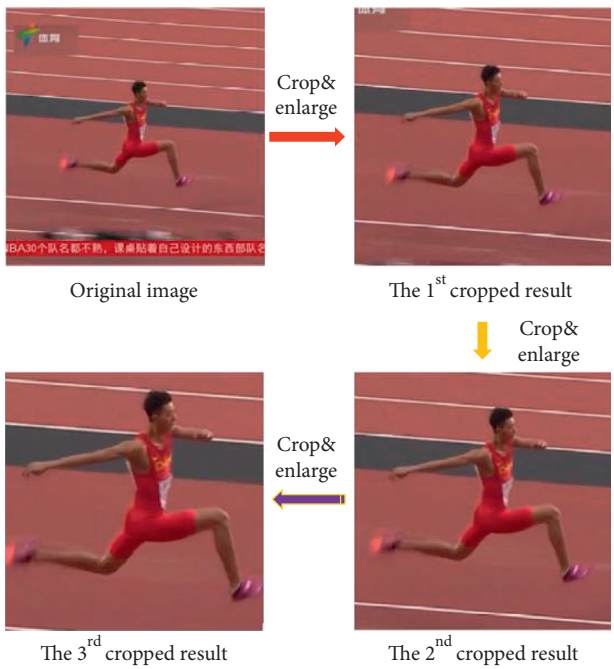

Figure 5: Schematic diagram of three-level attention focusing.

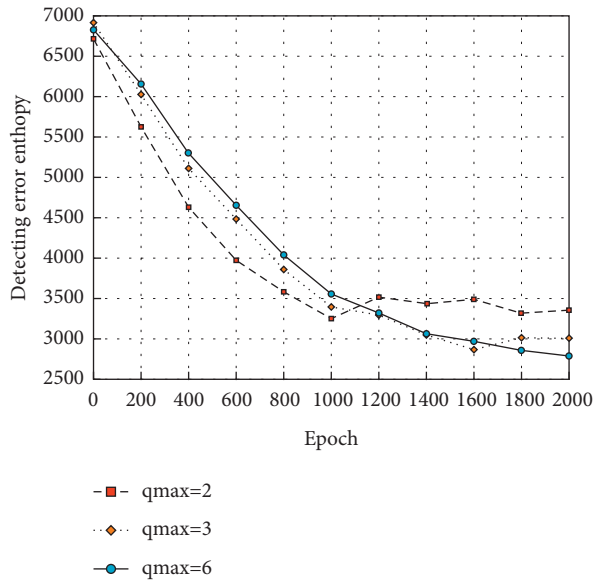

Figure 6: The changing curve of target detection error entropy under the condition of different levels of attention.

Table 1: Evaluation results of various trackers.

Recognition method	Precision	Success rate
Our method	0.910	0.685
MDNet [5]	0.861	0.641
SiamRPN [6]	0.891	0.666
FCN [7]	0.901	0.68
TADT	0.865	0.67
C-RPN	0.876	0.675
DAT	0.902	0.672
SPM	0.899	0.667

Table 2: Comparison in pose estimation between different methods.

Recognition method	AP	Posture estimation time (s)
Our method	76.2	0.27
Tompson et al. [25]	74.3	0.012
Feng et al. [26]	74.1	0.073
Cao et al. [27]	72.8	0.028
CPN [28]	75.8	0.18
Hourglass [29]	75.3	0.14

target tracking algorithms in terms of both precision and success rate of tracking, which is due to its imitation of human decision patterns, real-time evaluation of regional attention effectiveness, and iterative generation of regional attention from global to local to focus on samples at different scales. The above characteristics improve its tracking robustness in complex environments.

In addition, the algorithm proposed is also compared with other human pose estimation algorithms, such as Tompson et al. [25], Fang et al. [26], Cao et al. [27], CPN [28], and Hourglass [29]. Taking AP as an evaluation index, Table 2 gives the estimation results of the test sample set, and the average time consumed for pose estimation.

It can be seen from Table 1 that the proposed method performs better in pose estimating than other methods, because the multilevel focusing mechanism of the adaptive attention mechanism can get rid of the interference from useless areas in the complex image, which provides more accurate target areas for subsequent estimation. By integrating multiscale features at both the bottom level and the top level, the PoseFix pose optimization strategy enhances the pose feature representation effect. In addition, the pose estimation time allows to evaluate the overall complexity of the algorithm. Although the method proposed in this paper is inferior to other methods in time consumed, it in fact makes no difference with CPN and Hourglass under the same order of magnitudes. The long estimation time is due to the good real-time tracking and pose estimation performance as the model is exploring more suitable focus levels to obtain accurate target detection results.

5. Conclusions

Higher accuracy in pose estimation is a prerequisite for excellent sports results. In order to solve the problem related to low efficiency of labor-based physical education, this paper explores an intelligent method of estimating target posture based on adaptive attention mechanism, which provides technical support for daily physical education. First, according to the human decision-making model, regional attention is gradually focused and iteratively generated. By evaluating the performance of target features in real time, and adaptively adjusting levels of the attention, precise positioning of the detected target is achieved. Second, the PoseFix pose optimization strategy is adopted to solve the crossover problem of the key points of the human body posture estimated by the hourglass network. Experiments on

sports video images demonstrate the feasibility and effectiveness of this method.

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 research was supported in part by the Key Humanities and Social Science Research Project in Anhui Province (SK2020A0955) and Anhui School-Enterprise Cooperation Practice Education Base Project (2020sjjd102).

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[2] S. Ren, K. He, and R. Girshick, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[3] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "YOLOv4: optimal speed and accuracy of object detection," 2020, https://arxiv.org/abs/2004.10934.

[4] C. Y. Wang, H. Y. M. Liao, and Y. H. Wu, "CSPNet: a new backbone that can enhance learning capability of CNN," in *Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1571–1580, Seattle, WA, United States, June 2020.

[5] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4293–4302, Las Vegas, NV, USA, June 2016.

[6] B. Li, J. Yan, and W. Wu, "High performance visual tracking with siamese region proposal network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8971–8980, Salt Lake City, UT, USA, June 2018.

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

[8] N. Chen, J. Zhu, J. Chen, and T. Chen, "Dropout training for SVMs with data augmentation," *Frontiers of Computer Science*, vol. 12, no. 4, pp. 694–713, 2018.

[9] J. Redmon, S. Divvala, and R. Girshick, "You Only Look Once: unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788, Las Vegas, NV, United States, June 2016.

[10] W. Ouyang, X. Chu, and X. Wang, "Multi-source deep learning for human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2337–2344, Columbus, OH, United States, June 2014.

[11] Q. Li, H. Tang, Z. Liu, J. Li, X. Xu, and W. Sun, "Optimal resource allocation of 5G machine-type communications for situation awareness in active distribution networks," *IEEE Systems Journal*, pp. 1–11, 2021.

[12] Q. Y. Li, T. Cao, and W. Sun, "An optimal uplink scheduling in heterogeneous PLC and LTE communication for delay-aware smart grid applications," *Mobile Networks and Applications*, vol. 26, no. 4, pp. 1–14, 2021.

[13] G. Hidalgo, Y. Raaj, and H. Idrees, "Single-network whole-body pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 6982–6991, Seoul, South Korea, June 2019.

[14] S. Kreiss, L. Bertoni, and A. A. Pifpaf, "Composite fields for human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11977–11986, Seoul, South Korea, June 2019.

[15] W. Sun, Q. Li, C. Zhao, and S.-K. Nguang, "Mode-dependent dynamic output feedback H∞ control of networked systems with Markovian jump delay via generalized integral inequalities," *Information Sciences*, vol. 520, pp. 105–116, 2020.

[16] W. Sun, L. Huang, Z. Liu, Q. Li, C. Zhao, and D. Mu, "Distributed controller design and stability criterion for microgrids with time-varying delay and rapid switching communication topology," *Sustainable Energy, Grids and Networks*, vol. 29, p. 100566, 2022.

[17] A. Maharjan, X. Yuan, Q. Lu, Y. Fan, and T. Chen, "Non-rigid registration of point clouds using landmarks and stochastic neighbor embedding," *Journal of Electronic Imaging*, vol. 30, no. 3, p. 031202, 2021.

[18] A. Maharjan and X. Yuan, "Registration of human point set using automatic key point detection and region-aware features," in *Proceedings of the IEEE winter conference on applications of computer vision (WACV)*, pp. 4–8, Waikoloa, HI, USA, June 2022.

[19] L. Kong, X. Yuan, and A. M. Maharjan, "A hybrid framework for automatic joint detection of human poses in depth frames," *Pattern Recognition*, vol. 77, pp. 216–225, 2018.

[20] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, "Geometric deep learning: going beyond euclidean data," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 18–42, 2017.

[21] W. T. Li, D. Jiao, and Q. Zhang, "Research on intelligent cognition method of self-exploding state of glass insulator based on deep migration learning," *Proceedings of the CSEE*, vol. 40, no. 11, pp. 3710–3720, 2020.

[22] Q. Zhang, W. Li, H. Li, and J. Wang, "Self-blast state detection of glass insulators based on stochastic configuration networks and a feedback transfer learning mechanism," *Information Sciences*, vol. 522, pp. 259–274, 2020.

[23] W. Li, H. Tao, H. Li, K. Chen, and J. Wang, "Greengage grading using stochastic configuration networks and a semi-supervised feedback mechanism," *Information Sciences*, vol. 488, pp. 1–12, 2019.

[24] M. R. Ronchi and P. Perona, "Benchmarking and error diagnosis in multi-instance pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 369–378, Venice, Italy, June 2017.

[25] J. Tompson, A. Jain, and Y. LeCun, "Joint training of a convolutional network and a graphical model for human pose estimation," in *Proceedings of the Conference and Workshop on Neural Information Processing Systems*, pp. 1799–1807, Montreal, Canada, December 2014.

[26] H. S. Fang, S. Xie, and Y. W. Tai, "RMPE: regional multi-person pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2353–2362, Venice, Italy, October 2017.

[27] Z. Cao, T. Simon, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pp. 1302–1310, Honolulu, HI, USA, July 2017.

[28] Y. Chen, Z. Wang, and Y. Peng, "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, USA, June 2018.

[29] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proceedings of the European Conference on Computer Vision*, pp. 483–499, Amsterdam, Netherlands, October 2016.