Research Article

Improved Siamese Network-Based 3D Motion Tracking Algorithm for Athletes

Tao Lan

Hebei Sport University, Shijiazhuang 050041, China

Correspondence should be addressed to Tao Lan; lantao0575@hepec.edu.cn

Received 19 April 2022; Accepted 28 May 2022; Published 15 June 2022

Academic Editor: Wen Zhou

Copyright © 2022 Tao Lan. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

For the last few years, the application of Siamese network in athletes’ three-dimensional motion tracking has greatly improved the efficiency of sports training. However, the accuracy of the Siamese network tracking algorithm is limited to a large extent. To solve the above problems, based on the channel attention mechanism, the key feature information perception module is innovatively proposed to promote the discriminant ability of the network model and make the network focus on the convolution feature changes of the target. On this basis, an online adaptive mask strategy is proposed, which adapts the subsequent frames according to the output state of the cross-correlation layer learned online to highlight the foreground object. Compared with other algorithms in annotated data set and MOT17 data set, this algorithm has more stable initial tracking performance, significantly improved accuracy compared with the benchmark, and high robustness tracking effect in complex scenes.

1. Introduction

Human motion recognition and motion evaluation is a research hotspot at present. Action recognition analyses and processes the input video or 3D action data to determine which category different actions belong to. Motion recognition technology has practical application value in various industries such as human-computer interaction scene [1], surveillance video [2], gesture recognition [3], rehabilitation training [4], robot [5], and behaviour understanding [6]. Action evaluation is to judge the completion quality of specific actions. It is generally used in sports, dance, Tai Chi [7], and other professional fields. It can assist referees and coaches in scoring and help people with movement analysis and training.

As early as the 1970s, Johansson et al’s motion perception experiment of moving light spot confirmed that the three-dimensional human motion information can be analysed with the help of the two-dimensional model, which aroused many researchers’ interest in human motion recognition. A large number of subsequent research work on motion recognition emerged and achieved remarkable results. On the other hand, the research on movement evaluation is still in its infancy. Although there are some successful cases, such as golf swing [8] and badminton swing [9], what can be handled is mainly single and highly repetitive movements. For more complex movements, such as competitive aerobics [10], dance, 24 style Tai Chi [11], and opera [12], etc., this becomes completely different. For these complex actions, we should not only compare the “appearance similarity” but also make a breakthrough in the deeper “professional similarity.”

After full and in-depth investigation, this paper discusses the differences and relations between action recognition and action evaluation and summarizes the technical framework of action recognition and action evaluation from the perspective of complete data processing flow as shown in Figure 1.

In order to improve the accuracy of Siamese network tracker [13], on the basis of maintaining real-time performance, this paper innovatively proposed a new key feature information perception module to improve the discriminant ability of the Siamese network model, which includes multiscale feature extraction and attention mechanism [14]. This paper is used to remove all connection layer AlexNet [15] as a network of feature extraction, presents a multiscale sampling method to extract the target under the multiscale feature information, and uses the attention mechanism as the key to
enhance the target information, to capture the most distinct abstract semantic features [16], and then the similarity discriminant features are captured; the experimental results show that the tracking accuracy is significantly improved. In addition, in order to enhance the ability of Siamese network tracker to deal with complex scenes, a low-time consumption online adaptivemask strategy is proposed. By learning the complexity of background noise in the search image through cross-correlation output, we can mask the search image adaptively according to the complex situation and suppress a lot of background noise interference [17]. The tracker can maintain robust robustly performance in complex scenes.

2. Methodology

2.1. Siamese Neural Network. Figure 2 shows the structure of Siamese neural network. It has two or more neural networks with the same substructure. Each sub network has the same structure and shares parameters and weights [18]. The idea of Siamese network is to let two inputs pass through two subnetworks, respectively, and then extract the characteristics to get the characteristic vectors of the two inputs. Then, by constructing distance measurement functions, such as cosine distance [19] and Euclidean distance [20], it can be used to calculate the matching degree of the two inputs. During training, learn a similarity matching function according to the results for subsequent matching.

Siamese neural network can be divided into the following two parts from the network structure.

2.1.1. Feature Extraction. The main part of the Siamese network is used to extract the characteristics of the two inputs, respectively, to obtain the feature information that can effectively describe the input. It is usually realized by convolutional neural networks (CNNs) [21]. CNN usually includes convolution layer, pooling layer, full-connection layer, nonlinear activation function, and other parts, which can
extract the high-dimensional semantic information of the input image. Compared with hog feature, LBP feature, and CN colour feature, CNN feature has better robustness under the influence of complex background change and target appearance change.

2.1.2. Decision Network. After extracting the characteristics of the two inputs, it is necessary to determine their similarity according to the matching function. During training, the model parameters are continuously adjusted in combination with the network output and the real output to acquire a decision function with good performance. In different tasks, decision networks have different forms. Some are measured directly by loss function [22] and measurement function [23], and some continue to use neural networks [24] to verify the confidence of results, which improve decision reliability.

2.2. Target Tracking Algorithm for Siamese Network. In the target tracking algorithm of Siamese networks, the input is target image Z and candidate image X. Through template feature and correlation matching, a correlation response graph based on candidate image x is obtained, and then the location of the tracking target is determined by the maximum response. The problem of target tracking is transformed into a problem of learning similarity function.

The similarity function of fully connected Siamese network is defined as follows:

\[ f(k, i) = \varphi(k)^\ast \varphi(i) + h, \]  (1)

where \( \varphi \) is the CNN feature extraction function, \( \ast \) represents the convolution operation, and the correlation between them can be calculated. \( H \) is the deviation value, and \( f(k, i) \) represents the similarity score of \( k \) and \( i \). The more similar the reference target \( k \) is to the candidate image \( i \), the higher the return score of the function is, and vice versa. Several candidate images are selected through a certain strategy in the frame to be detected, and a fractional response can be obtained after calculating the similarity score, respectively. The position with the highest score in the response graph is the target position of the tracking object in the frame predicted by the algorithm.

In the tracking phase, in the current frame, a large search area \( i \) is obtained centred on the target centre detected in the previous frame, which can be obtained by padding the bounding box of the previous frame. Then, feature extraction function is used \( \varphi \) The features of the reference target \( i \) in the region and the first frame are extracted, respectively, such as \( \varphi(k) \) and \( \varphi(i) \); then these features are obtained by convolution operations \( \varphi(k) \) and \( \varphi(i) \). The similarity score vector \( f(k, i) \) is used to obtain the similarity score map, in which the position with the largest score is the position corresponding to the target in the frame.

Taking the Siamese FC algorithm as an example, in model training and tracking, the network input is the real marker box of the first frame and the search area of the current frame, respectively. The two input images are \( 127 \times 127 \) and \( 255 \times 255 \), respectively, and the final network output fractional image size is \( 17 \times 17 \). Each position in the score graph represents the confidence value of a candidate region as the target region. Later, using appropriate processing methods can get more accurate target centre position.

2.3. Key Feature Information Perception Module. The Siamese network tracker can be modelled by the following formula as follows:

\[ R(i, k) = \text{CORR}(\varphi(i), \varphi(k)), \]  (2)

where \( i \) and \( k \) are the input template image and search image, respectively, \( \varphi \) is the feature extraction network, \( \text{CORR} \) is the cross-correlation operation, \( R \) is the matrix response graph, and the target centre is located through the maximum value in \( R \). In the formula, \( \varphi \) parameter is shared in template image and search image branch. Simply using AlexNet as the feature extraction network cannot fully tap the potential of the Siamese network structure, so this paper proposes a key feature information sensing module embedded in AlexNet. In order to prove the universality and effectiveness of the module, SiamFC-DW is also used as a benchmark for comparative experiments.

This paper chooses to embed this module after the third layer of AlexNet because the features extracted from the first three layers are relatively shallow image features, while the last two layers are more abstract semantic features. In this paper, the key of design feature information perception module is as shown in Figure 3. Firstly, feature information of different scales is obtained through the maximum pooling sampling layer in various scales and fused. It enables the receiving field of each pixel to have rich convolution characteristics, thus providing the tracker with more prior knowledge about the target position. This paper adopts the sampling pooling under nuclear size of \( 3 \times 3 \) and \( 5 \times 5 \). However, in the process of feature information fusion at different scales, a large amount of interference information is brought, resulting in unstable performance of the tracker. As shown in Table 1, on the basis of SiamFC, the multiscale feature extraction strategy was used to train the models for many times in the same way, and their performance was tested. The results showed that the performance fluctuation was large and the improvement was not obvious.

In order to obtain stable and more robust performance, the network should selectively enhance the key feature information of the target and suppress useless feature information, that is, only capture the most significant image attributes of the target. A simple and effective way is to assign different weights to different channels of convolution features, which can be expressed by the formula as follows:

\[ P = Ps = \sum_{z=1}^{C} p_z s_z, \]  (3)

where \( P = [p_1, p_2, \ldots, p_C] \) said untreated convolution, \( C \) is the characteristics of the channel number, \( P = [p_1, p_2, \ldots, p_C] \) said each channel gives different weights to produce new features, and \( s = [s_1, s_2, \ldots, s_C] \) as
weights. This paper adopts channel attention mechanism to generate channel information weight value $s$ from the origin.

Figure 4 shows the channel attention module. By explicitly modelling the interdependence between channels, it adaptively recalibrates the characteristic response of channels and applies it to Figure 3, which only slightly increases the model complexity and computation burden but significantly improves the accuracy. The process for generating the weight $s$ through the channel attention mechanism is split into "squeeze" and "incentive." Firstly, Adaptive avgpool is used to compress the global features into a channel descriptor. Formally, the spatial dimension of $P, B \times M \times C,$ is compressed into a statistical vector $k \in \mathbb{R}^C,$ and the $k$ element of $z$ is calculated by formula as follows:

$$k_z = F_{sv}(u_z) = \frac{1}{B \times M} \sum_{x=1}^{B} \sum_{y=1}^{M} u_x(x, y) .$$  \hfill (4)

The rest of the scheduling operation follows the Squeeze operation. The scheduling aims to learn the nonlinear dependencies and non-murexes between the channels because you want to make sure that you allow the model to emphasize multiple channels and not just perform one-hot activation. Choose to use the activation function Sigmoid with the threshold mechanism. The procedure is illustrated by the following formula:

$$s = F_{sv}(k, M) = \sigma(a(k, M)) = \sigma(M_2 \delta(M_1 k)) ,$$  \hfill (5)

where $\delta$ represents the parameters of ReLU, $M_1$ and $M_2,$ respectively, which enhance the generalization ability of attentional mechanism through continuous full-connection layer structure. $\sigma$ represents Sigmoid function. Sigmoid function is often used as the activation function of neural network. It can map a real number to the interval of $(0, 1)$ and can be used for binary classification. The effect is better when the feature difference is complex or the difference is not particularly large.

![Figure 3: Key feature information perception module structure 1.](image)

Table 1: AUC performance of the training model on OTB100.

<table>
<thead>
<tr>
<th>SiamFC (%)</th>
<th>Model 1 (%)</th>
<th>Model 2 (%)</th>
<th>Model 3 (%)</th>
<th>Model 4 (%)</th>
<th>Model 5 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>58.3</td>
<td>58.6</td>
<td>60.8</td>
<td>60.2</td>
<td>57.5</td>
<td>60.4</td>
</tr>
</tbody>
</table>
This paper also proposes another structure that can replace Figure 3, which is shown in Figure 5. Different from Figure 3, the features obtained by sampling under multiscale maximum pooling are not directly fused. Instead, the features of these different scales are input into channel attention for weight allocation, and finally the calibrated features are fused. This paper uses SiamFC and SiamFC-DW as reference algorithms to compare these two structures. The comparison results are shown in Table 2. SiamFC with structure 1 improved its accuracy by 6.6% compared with the benchmark, exceeding the maximum increase of 4.1% in Table 1, and its speed decreased by 16fps. With structure 2, the accuracy improved by 7.3%, and the speed drop increased to 30fps, but it was still well above the real-time requirements. SiamFC-DW improved by 2.7% and 3.5%, respectively, under two different structures, indicating that the key feature information sensing module proposed in the paper has enhanced on the basis of strong enough network discrimination ability.

In this paper, structure 1 of Figure 3 is embedded into SiamFC feature extraction network AlexNet. After end-to-end training, the model is applied to Bolt and Board, two video sequences of OTB100, and the feature information output from the cross-correlation layer learned by the model is visualized.

### 2.4. Online Adaptive Mask

After embedding key feature information sensing module in feature extraction network, the discriminant ability of the model has been improved qualitatively. However, the performance of the tracker is still not robust enough in complex scenes to resist the interference of some seriously similar objects. Therefore, this paper also proposes an online adaptive mask strategy to suppress the interference information and highlight the foreground object to deal with the complex scene. The strategy achieves adaptive effect through online learning of mask parameters. Compared with the traditional image mask, the adaptive mask in this paper can capture the dynamic information of the target in the video stream. However, the traditional method cannot adapt to the change of the target, and the suppression process will bring loss to the foreground information of the image.

The form of the online adaptive mask is as follows:

\[
A(i_n, j_n) = \frac{1}{\alpha \sqrt{2\pi \sigma_i \sigma_j}} \exp \left(-\frac{i_n^2 + j_n^2}{2\sigma_i^2 + 2\sigma_j^2}\right),
\]

where \(f^*\) and \(f\) represent the search image before and after the mask, respectively, \(n\) represents the current frame sequence, \(i_n\) and \(j_n\) represent each pixel of the search image, \(A\) represents the Gaussian mask function, and the parameters \(\sigma_i\) and \(\sigma_j\) represent the horizontal mask degree and vertical mask degree of the search image. The following three steps can complete the detailed online adaptive mask.

**Step 1.** According to the aspect ratio of the target frame, the degree of mask \(\sigma_i\) in the horizontal axis and \(\sigma_j\) in the vertical axis of the search image in frame 1 can be determined by formula (7). The adaptive process first needs to know the confidence information of the historical frame, so the Gaussian mask function parameter of the previous x frame is the same as that of the first frame.

\[
\left[\sigma_i, \sigma_j\right] = \begin{cases} 
\left[\beta, \beta(1 - \theta)\right], & m \geq nr, \\
\left[\beta(1 - \theta), \beta\right], & m \leq nr, \\
\left[\beta, \beta\right], & \theta \leq 0.1
\end{cases}
\]

where \(nr\) and \(\theta\) are set to 1.8 and 0.1, respectively, and \(\beta\) is the initial value of \(\sigma\) and is set to 95.

**Step 2.** The confidence factor adopted in this paper is the average peak correlation energy (APCE). The APCE is obtained from the response graph obtained by searching the image of the current frame and then compared with the average APCE of several historical frames. Thus, the magnitude of \(\sigma_i\) and \(\sigma_j\) can be determined by the APCE formula as follows:
APCE = \frac{\left( \frac{r_{\text{max}} - r_{\text{min}}}{\text{mean} \left( \sum_{m,b} (r_{m,b} - r_{\text{min}})^2 \right)} \right)^2}{\text{mean} \left( \sum_{m,b} (r_{m,b} - r_{\text{min}})^2 \right)}.
\tag{8}

where \( r_{\text{max}}, r_{\text{min}}, \) and \( r_{m,b} \) represent the maximum value and minimum value of the response graph fraction obtained by cross-correlation between the search image features and the target template image features and each value of row \( W \) and column \( H \), respectively.

**Step 3.** Finally, the values of \( \sigma \) are adaptively updated according to formula (9) during the online tracking process of subsequent frames.

\[
\sigma_i = \begin{cases} 
\sigma_{i-1} (1 + \mu), & \varepsilon_{\text{div}} \geq \tau_1, \\
\sigma_{i-1} (1 - \mu), & \varepsilon_{\text{div}} \leq \tau_2, \\
\sigma_{i-1}, & \text{others,}
\end{cases}
\tag{9}
\]

where \( \mu \) is the factor of change of mask degree, set as 0.2, and the thresholds \( \tau_1 \) and \( \tau_2 \) are set as 1.175 and 0.825, respectively. The above formula only gets \( \sigma_{i-1} \), and the determination method of \( \sigma_{i-1} \) is the same. \( \epsilon_{\text{div}} \) represents the ratio of the current APCE to the average APCE of historical I frames, which can be calculated from formula (10), where \( x \) is set to 3.

\[
\epsilon_{\text{div}} = \frac{\text{APCE}_i}{\left( \sum_{z=0}^{n-1} \text{APCE}_z \right) / x} = \frac{x \cdot \text{APCE}_i}{\sum_{z=0}^{n-1} \text{APCE}_z}.
\tag{10}
\]

**3. Result Analysis and Discussion**

In the Literature [25] paper, experimental results are based on the MOT16 data set and nonstandard detectors. In this paper, the detection branch results of Literature [26] are used as the input of Literature [25] tracking stage, and the appearance features extracted by Literature [26] are used as the input of Literature [25] appearance similarity
calculation. Since MOT16 data set does not provide camera internal and external parameters, this paper carries out the comparison experiment of Literature [26] network on MOT17 data set and carries out the comparison experiment of overall tracking algorithm on annotation data set.

3.1. Data Correlation. MOT17 dataset is a public data set to evaluate the effectiveness of the multitarget pedestrian tracking algorithm and provides MOTA, MOTP, MT, and other evaluation indicators. The training of Literature [26] requires additional target orientation information and head bounding box, so this paper manually annotates the pedestrian orientation information and head bounding box based on MOT17 data set. Finally, this paper uses 4 training videos (3012 frames) and 3 test videos (2575 frames) in the data set and manually marks the orientation information and the head bounding head frame. The annotation data set is 3805 frames of pedestrian videos taken by ourselves. Like MOT17 data set, the information of pedestrian bounding frame, head bounding frame, and orientation are manually annotated, and the external and internal parameters of surveillance cameras are collected. In the training phase, the input data were randomly mirrored, and the padding processing in the lower right corner was used to ensure that the image length and width were equal. In this paper, the two branches of Literature [26] were trained in sections. Resnext-50 of the trunk network used the standard weights of ImageNet classification data set to train, cut the layers after Conv4, and then train the RPN and RCNN of the detection branch. After the training of the detection branch, the weight of the trunk network was fixed. The training is complete after network convergence. The Sigmoid activation function is used after Conv3 for the prediction branch and FC2&FC3 for the detection branch, and the ReLU activation function is used for the rest.

3.2. The Results. The test results of the proposed method on annotated data sets are shown in Table 3. The network detection branch achieves the same excellent detection performance as Faster RCNN under the structure of multitask branch and feature sharing. The accuracy of orientation prediction was evaluated by three value domains. About 72.23% of the predicted orientations were almost identical with the original labelled orientations, 91.21% of the predicted orientations were highly consistent with the

<table>
<thead>
<tr>
<th>Output</th>
<th>Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human body bounding box detection</td>
<td>Accuracy 99.81%</td>
</tr>
<tr>
<td></td>
<td>Recall rate 98.63%</td>
</tr>
<tr>
<td></td>
<td>Mean IOU 76.35%</td>
</tr>
<tr>
<td>Head bounding regression box</td>
<td>Mean IOU 87.15%</td>
</tr>
<tr>
<td></td>
<td>CS_{loss} &lt; 0.001 72.23%</td>
</tr>
<tr>
<td>Orientation prediction</td>
<td>CS_{loss} &lt; 0.003 91.21%</td>
</tr>
<tr>
<td></td>
<td>CS_{loss} &lt; 0.010 99.67%</td>
</tr>
</tbody>
</table>

Table 3: Results of the proposed method.
original labelled orientations, and 99.67% of the predicted orientations were similar to the original labelled orientations.

In MOT17 data set, the Literature [26] and Literature [27] and Literature [25] algorithms are used to compare the tracking effect. Figure 6 shows the average accumulative number of ID Switch occurrences of the three algorithms with multiple frame inputs on MOT17 data set. It can be seen that the initial ID Switch error of the Literature [27] and Literature [25] algorithm increases rapidly and then tends to be stable. The initial tracking performance of this paper is obviously better than that of other algorithms.

This paper compress and tests the complete tracking algorithm on annotated data sets. Figure 7 is the result of the initial tracking error of the algorithm on the labelled data set. Different from Figure 6, the curve gap between Literature [26] and Literature [25] is small. After analysing the test data, it is found that compared with the labelled data set, MOT17 data set is more densely populated, which is prone to ID Switch error, while the labelled data are relatively sparse. This happens less frequently. Compared with Literature [26], the camera model projection proposed can effectively reduce ID Switch errors in another case. Table 4 shows the MOT evaluation indexes of each algorithm on the annotated data set. The algorithm proposed in this paper has higher MOTA and effectively reduces the occurrence of ID Switch.

4. Conclusion
This paper presents a three-dimensional motion tracking algorithm based on improved Siamese network. The tracking algorithm from two-dimensional plane is transferred to three-dimensional space, which reduces the occurrence of ID Switch and improves the initial stability of the tracking algorithm. In addition, a feature reuse network is proposed to reduce the calculation overhead of the algorithm.

To improve the accuracy of the Siamese network tracker, a general key information feature sensing module was proposed based on the channel attention mechanism to enhance useful information selectively. The module was embedded in the feature extraction network to effectively improve the network model’s discrimination ability. In this paper, a low-time consumption online adaptive mask strategy is proposed to highlight the foreground target, suppress the interference of background information to a large extent, and further improve the tracking accuracy while taking into account the tracking speed.

The proposed algorithm in this paper on the data from the routine surveillance cameras made excellent tracking performance, but real fisheye camera in the scene is also more common. Fisheye camera is spherical imaging, a wide-angle, image distortion, and spherical imaging characteristics different from general type straight surveillance cameras; how compatible fisheye camera is the research focus of next step.

Table 4: Multitarget tracking index result.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>MOTA (%)</th>
<th>MT (%)</th>
<th>ML (%)</th>
<th>ID Switch (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Literature [27]</td>
<td>60.5</td>
<td>24.9</td>
<td>27.5</td>
<td>24.4</td>
</tr>
<tr>
<td>Literature [25]</td>
<td>63.3</td>
<td>38.5</td>
<td>22.3</td>
<td>13.8</td>
</tr>
<tr>
<td>Literature [26]</td>
<td>65.7</td>
<td>55.4</td>
<td>17.2</td>
<td>12.5</td>
</tr>
<tr>
<td>Proposed</td>
<td>71.3</td>
<td>54.2</td>
<td>5.4</td>
<td>8.6</td>
</tr>
</tbody>
</table>

Data Availability
The labelled data set used to support the findings of this study is available from the corresponding author upon request.

Conflicts of Interest
The author declares that there are no conflicts of interest.

Acknowledgments
This work was supported by the Hebei Sport University.
References


