Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 5509095, 13 pages
https://doi.org/10.1155/2022/5509095

Research Article

Intelligent Optimization Algorithm of 3D Tracking Technology in Football Player Moving Image Analysis

Peng Sun,[1] Xiang Zhao,[2] Yu Zhao,[3] Ni Jia,[3] and Dawei Cao[2]

[1]*China Football College of Beijing Sport University, 100084 Beijing, China*
[2]*College of Physical Education, Huaibei Normal University, Huaibei, 235000 Anhui, China*
[3]*College of Physical Education, Minzu University of China, 100081 Beijing, China*

Correspondence should be addressed to Xiang Zhao; zhaox@chnu.edu.cn

Received 13 May 2022; Revised 17 June 2022; Accepted 7 July 2022; Published 27 July 2022

Academic Editor: Akshi Kumar

Computer vision technology began to affect the development of football. There is increasingly high-tech in football broadcast technology, and many application tools have emerged in the field of football broadcast video analysis. The purpose of this paper is to study the improvement of target tracking algorithm for football broadcast video and to study the intelligent optimization algorithm of 3D tracking technology in football player moving image analysis. This paper proposes to select four models of YOLOv5 to perform target detection experiments in football broadcast videos and analyzes the principle of the Deep SORT multitarget tracking algorithm. At the same time, it is based on the 3D tracking and 3D pose estimation of players based on cross-view correlation matching, and to measure the comprehensive performance of the tracker in the football scene, experiments are carried out on the accuracy and speed of the tracker under the football datasets of four different scenes. The experimental results in this paper show that the MOTA values corresponding to the 3D tracking results and 2D projection results obtained in the campus dataset are only 50 and 56.2. This is much lower than the tracking performance when based on other similarity matrices. The MOTA value of the obtained tracking result (92.6) is very close and significantly outperforms other methods. CCOT performs better on datasets 28, 29, 31, and 32, ECO stands out on dataset 38, and Siamese also performs well on datasets 22 and 36.

1. Introduction

The rapid development of the Internet continues to impact everything, and the television broadcasting industry is no exception. With the promotion and dissemination of football events and the rapid development of big data, football broadcasting is becoming increasingly technological in technology. The broadcast of mainstream football leagues is trying to integrate various high-tech, such as the use of 3D vision to display the starting players, VAR playback technology, TrueView technology, and Hawkeye technology. Taking the Spanish football league as an example, in 2018, it used a series of industry-leading technologies such as Skycam, True View 360°, and Mediacoach in event broadcast and event analysis. However, at present, to realize these technologies requires large investment, and the use of manual labor is time-consuming and labor-intensive. Small- and medium-sized competitions simply cannot support such expenses.

It needs to meet the high-level semantic analysis requirements of related players' technical action playback, video summary generation, and so on. These contents need to be completed through target tracking. At this point, the help of some auxiliary technologies is needed, such as offside penalty, football goal line technical analysis, and foul behavior recognition. Therefore, target tracking is the basic task for most practical applications in football at present, which has research value in both theory and practice.

The innovations of this paper are as follows: (1) it analyzes the principle of the Deep SORT multitarget tracking algorithm and finds that the ID increment of the algorithm is serious in the football scene. It reduces the ID increment phenomenon caused by intraclass occlusion by introducing a trajectory scoring mechanism to the matching stage. The

experimental results show that the improved algorithm has an obvious inhibitory effect on the ID increment speed. (2) This paper introduces and analyzes the theoretical basis of the correlation filtering algorithm and the Siamese network algorithm. It decomposes the correlation filter into multiple simple formulas and forms a network structure, which is fused into the Siamese network model as a correlation filter layer. It deduces the back-propagation formula of the relevant filter layer for the training of the network model. (3) It designs the experimental comparative analysis. To measure the overall performance of the tracker in football scenarios, it experiments the tracker in terms of accuracy and speed under four different scenarios of football datasets. It selects six representative tracking algorithms for horizontal comparison and conducts detailed thinking and analysis of the experimental results.

2. Related Work

The use of 3D eye-tracking systems to measure on-screen gaze is gaining traction. Stapleton and Koo found the effectiveness of a biokinetic visibility aid for night cyclists compared to other configurations in an intersubject blind experiment using 3D eye-tracking technology [1]. Vision-based technologies have received increasing attention due to their label-free and inexpensive configurations. Lee and Park researched various sensing technologies to locate workers and equipment on construction sites. They also proposed an efficient camera calibration method for locating entities tens of meters away from the camera [2]. In real-time 3D ball tracking for motion analysis in computer vision technology, complex algorithms to ensure accuracy can be time-consuming. On the CPU-GPU platform, Hou et al. proposed dual-stream system flow thread allocation based on view priority and reweighting for binary search [3]. The dual-stream system process allocates tasks that do not have data dependencies to different streams to process each frame, realizing parallelism at the system structure level. Predictive visual attention facilitates adaptation to virtual museum environments and provides context-aware and interactive user experiences. Zhou et al. designed a deep learning model. They also tested using EDVAM to predict the user's subsequent visual attention based on previous eye movements [4]. Human skeleton tracking systems often have difficulty handling lost tracking. Nguyen et al. proposed a multiview system for 3D human skeleton tracking based on multicue fusion [5]. Mendicino et al. aimed to develop and implement a complete integrated tracking system with very high accuracy both spatially and temporally per pixel [6]. In computer vision, tracking humans across camera views remains challenging. Especially to address these challenges, Liu et al. proposed a stochastic attribute grammar model for leveraging complementary and discriminative human attributes to enhance cross-view tracking [7]. For complex scenes, there are frequent occlusion, obvious lighting changes and other difficulties. In this case, most existing appearance and geometric cues are not reliable enough to distinguish humans in the camera view.

3D motion capture systems have been used to validate commercial electronic performance and tracking systems. Aughey et al. aimed to determine the effectiveness of the VisionKit computer vision system for 3D motion capture in a stadium environment. Experiments show strong agreement between VisionKit and 3D motion capture in every activity performed [8]. In professional football, almost every team today uses tracking technology to monitor performance during training and games. By tracking data, Goes et al. can gain valuable insights into how and why tactics perform in football matches. Each team has about 500 pass interactions in a game [9]. Dai and Lu studied an improved bioimage tracking algorithm for athletes' cervical spine health under color feedback. Their aim was to propose a new algorithm to improve the detection and tracking accuracy [10]. To assist football training in colleges and universities, Zhu proposed an edge computing-based football robot path planning algorithm and an improved PSO (particle swarm optimization) algorithm [11]. However, after conducting event detection and player tracking experiments on the dataset, the results show that the existing content cannot fully solve the task of video analysis and needs to be improved. However, 3D motion capture cannot be used for large capture areas, such as full football fields, because many fragile cameras need to be placed around the capture space and these cameras lack the proper depth of field.

3. Three-Dimensional Tracking Technology of Football Images

3.1. Football Video Analysis. A complete football broadcast video usually consists of various types of shots such as close-up shots, medium shots, long shots, and off-field shots. During the game, with the constant switching of the camera, it can display the game situation on the field, the cheering situation outside the field, the personal skills of the players, and the game atmosphere to the audience in an all-round way [12]. However, these shot switches will greatly affect the effectiveness of the target tracking algorithm and will frequently lose the detected player information, resulting in the continuous growth of IDs in the video. This is not conducive to accurate player tracking. The fixed three-shot dataset is shown in Figure 1.

As shown in Figure 1, the ISSIA dataset, which is captured by a fixed static lens, has a certain degree of limitation when applied to full football videos. Generally speaking, there are three main aspects of football video analysis: detection of football events, detection and tracking of players and football, and game analysis [13]. Among them, football event detection is to detect and locate events such as goals, fouls, red and yellow cards, penalty kicks, and free kicks, also known as the generation of video summaries. Detecting and tracking players or the ball is an intuitive representation of football video. It has to deal with player occlusions, lighting changes, and sudden camera movements.

Then, it gets the coach's tactical changes and adjustments by analyzing the players' movements and positions [14]. The content of the game analysis includes players' running distance and speed, heat map, ball possession statistics,

Figure 1: Fixed three-shot dataset.

offside detection, team tactics, etc. And the tactical analysis requires a big-picture view, which can be obtained by looking at the game from the perspective of God. Often the content displayed in the video is limited by the lens, which requires drone shooting and dozens of cameras to shoot without blind spots [15]. In the establishment of the football dataset, the available datasets mainly focus on a single annotation or a single static scene shot.

3.2. Sports Video 3D Tracking Technology. The field of sports video analysis is one of the popular branches in the field of vision technology research. At present, people have used vision technology in many sports video analysis tasks, such as tracking of the ball and players in sports videos, action recognition for individual or team sports, performance scoring of players or teams, and eagle-eye techniques for judging whether the ball is out of bounds or whether a goal is scored [16]. Compared with ordinary surveillance video, the richness and complexity of sports video content bring greater challenges to the implementation of visual technology. Many vision-based solutions cannot meet the application demands of sports video analysis in terms of speed and accuracy. Up to now, many sports data analysis companies still rely on manual annotation. Based on the two-dimensional tracking of players under a single camera, we will further explore how to achieve three-dimensional tracking of players under multiple cameras. This helps to further implement more advanced semantic tasks such as behavior recognition, motion capture, virtual scene reproduction, player performance scoring, and event detection in sports video analysis.

3.3. Improvement of Single Target Tracker. In football games, players sometimes move faster, while the size of the search box is fixed in the single-target tracking algorithm. Tracking failure occurs if the player moves out of the search box [17]. It uses Kalman filtering technology to estimate the position of the target in the current frame and adjust the position of the search box, so that the player can be tracked when the player moves quickly. Next, it needs to update the covariance matrix P corresponding to $X(k \mid k-1)$:

$$P(k|k-1) = Ap(k-1|k-1)A' + Q, \tag{1}$$

where $P(k \mid k-1)$ is the covariance matrix corresponding to $X(k \mid k-1)$, $P(k-1 \mid k-1)$ is the covariance matrix corresponding to $X(k-1 \mid k-1)$, A' represents the transpose matrix of A, and Q is the covariance matrix of the system process.

After it has the predicted result, it needs to perform the optimal estimation in the next step. The optimal estimation should be performed by combining the predicted value and the measured value:

$$X(k|k) = X(k|k-1) + kg(k)(Z(k) - HX(k|k-1)), \tag{2}$$

where $Kg(k)$ is the Kalman gain:

$$Kg(k) = \frac{P(k|k-1)H'}{HP(k|k-1)H' + R}. \tag{3}$$

Now that the optimal estimate has been obtained, the covariance matrix needs to be updated, and the algorithm can run autoregressively:

$$P(k|k) = I - Kg(k)Hp(k|k-1), \tag{4}$$

where I is the identity matrix. When the system enters the $k+1$ state, $P(kl \mid k)$ is $P(k \mid k-1)$ by formula (1), and the algorithm can run autoregressively.

3.4. Feature Extraction of Players in Football Videos. To effectively distinguish football from other goals on the pitch, it is necessary to select appropriate features to describe the football goals. According to the observation, it can be found that the color of the football target is usually single, so the color feature can be used to describe the football target [18]. At the same time, considering that the Histogram of Oriented Gradient (HOG) feature describes the edge information of the image from the local image patch, it can effectively describe the shape and edge of the football target and distinguish the football from the target with similar color. This section first introduces the color feature and HOG feature extraction method of football.

3.4.1. Color Feature Extraction. Commonly used color models include RGB color model, HSI color model, etc. The color of the soccer goal is usually stable as a single color in the game video, and the brightness change is not obvious. To reduce computation, the color features of soccer goals are extracted in RGB space [19]. To reduce the amount of calculation, it quantizes the grayscales of the three color channels R, G, and B of the image to U level, respectively, and defines the quantization function:

$$b(l_i)\colon R^2 \longrightarrow \{1, 2, \cdots, U\}. \tag{5}$$

The grayscale values of the three color channels at I_i are, respectively, mapped to U quantization levels. Then, the single-color channel grayscale distribution of the target area is

$$P_N = \{P_N^M\}, u = 1, 2, \cdots, U, \tag{6}$$

where $N = \{R, G, B\}$ represents each color channel and the value of the u-th interval is calculated according to formula (7):

$$P_N^M = \sum_{i=1}^{M} k\left(\left\|\frac{l_c - l_i}{h}\right\|\right)\sigma[b(l_i) - u], \tag{7}$$

where $k(\cdot)$ is the kernel function, which adopts the Gaussian kernel function, so that the pixel gray level near the center position in the target area obtains a larger weight. h represents the nuclear window width. For objects with a rectangular area, h takes half the length of the area's diagonal. The role of the delta function is to determine whether the gray value of I in the target area falls within the uth interval. As shown in formula (8), it connects the gray distributions of the three color channels and normalizes them to obtain the target color feature:

$$P_{\text{color}} = C * \{P_R, P_G, P_B\} = \left\{P^{(M)}\right\}, m = 1, 2, \cdots, 3U, \tag{8}$$

where C represents the normalization coefficient, and the calculation formula is as follows:

$$C = \frac{1}{3 * \sum_{i=1}^{M} k((l - l_i)/h)}. \tag{9}$$

When only the color feature is used to characterize the target area, the part of the rectangular area beyond the edge of the target of interest is also considered. When the proportion of this part increases, it has a greater impact on the feature extraction results [20]. The color features are weighted by Gaussian with the position information, and the pixels close to the center of the target area are given large weights. The regions far from the center of the target region are assigned small weights, which weakens the influence of nontarget parts on feature extraction.

3.4.2. HOG Feature Extraction. For each pixel in the image area, it calculates the horizontal and vertical gradients at that point. The horizontal gradient and vertical gradient calculation methods of the image include orthogonal gradient operator, Roberts operator, Sobel operator, and Prewitt operator. Experiments show that large templates and smoothing operations will reduce the performance of feature description. In this paper, the central orthogonal gradient operator is used, and the horizontal gradient and vertical gradient are calculated as follows:

$$G_x(x, y) = I(x + 1, y) - I(x - 1, y), \tag{10}$$

$$G_y(x, y) = I(x, y + 1) - I(x, y - 1). \tag{11}$$

In formula (10) and formula (11), $I(x, y)$ represents the gray value at (x, y); then, the modulus value and direction of the directional gradient at the (x, y) point can be expressed as

$$\begin{aligned} p(x, y) &= \sqrt{G_x^2(x, y) + G_y^2(x, y)}, \\ \theta(x, y) &= \arctan\frac{G_y(x, y)}{G_x(x, y)}. \end{aligned} \tag{12}$$

3.5. Multitarget Tracking Algorithm. For the multitarget tracking problem, the current common tracking-by-detection is to associate the unreliable target detection results with the existing trackers. Deep SORT (Simple Online And Realtime Tracking) is a typical hybrid algorithm framework combining deep learning and traditional methods in the field of multitarget tracking and achieves a relatively stable tracking effect.

3.5.1. Deep SORT Multitarget Tracking Algorithm. Deep SORT starts from an input video stream and first achieves object detection by executing an object detection algorithm (YOLOv5). It converts the box obtained by the detector into detections and then uses Deep SORT to implement tracking based on the detection results, as shown in Figure 2.

As shown in Figure 2, this design can better optimize the detector or tracker to a certain extent for the tracking effect and can basically achieve the effect of real-time tracking. The speed can be adjusted according to the scale of Re-ID.

The algorithm input of the multitarget tracking stage is the target frame information detected by YOLOv5. It performs Kalman filter trajectory prediction based on the input detection frame and then uses the Hungarian algorithm to perform cascade matching or IOU matching between the predicted prediction frame and the detection frame detected in the current frame. The successful matching is the tracking success, and finally, it uses the Kalman filter to update.

For the association of motion information, it uses the square of the Mahalanobis distance between the prediction result of the Kalman filter and the new detection result for data association. Its calculation formula is as follows:

$$d^{(1)}(i, j) = \left(d_j - y_i\right)^T S_i^{-1}\left(d_j - y_i\right). \tag{13}$$

For the four-dimensional measurement space, the 0.95th quantile of the chi-square distribution was used as the corresponding Mahalanobis distance threshold.

It updates this list after each match, such as removing some target feature sets that have been moved out of the shot, keeping the newest features, and deleting the old ones. It computes all appearance description feature vectors tracked by the i-th object based on the Re-ID. Its calculation formula is as follows:

$$d^{(2)} = \min\left\{1 - r_j^T r_k^{(i)} \middle| r_k^{(i)} \in R_i\right\}. \tag{14}$$

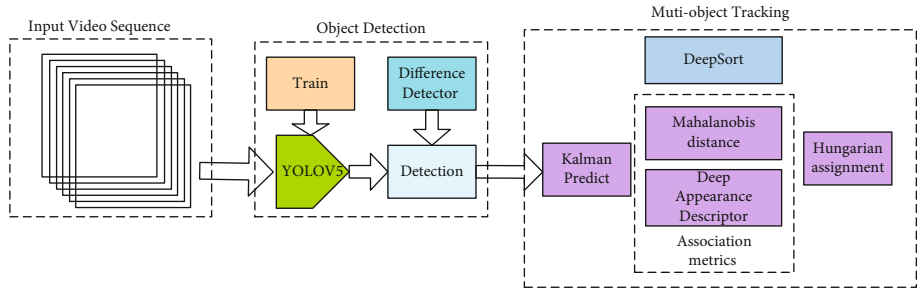

Figure 2: Principle of Deep SORT.

The cosine distance threshold is obtained in advance from the training set. If the cosine distance is less than the preset threshold, it is considered that the detection frame and the tracking frame are successfully associated. The final measure of the matching process is a linear weighting of the two measures, as shown in

$$= \lambda d^{(1)}(i,j) + (1-\lambda) d^{(2)}(i,j). \quad (15)$$

When parameter $C_{i,j}$ belongs to the intersection of the two metric thresholds, it is considered that the two have completed the association matching. In the experiment, we try to set λ to 0 and only use the appearance information because the camera motion of the football video is too large.

Finally, the update stage of Kalman filter updates the mean square error and saves the feature map of the detector. The update phase is accompanied by state transitions, as shown in Figure 3.

As shown in Figure 3, both the detection frame and the prediction frame are in an uncertain state when they first enter the matching module. Then, after several matches, if the number of successful matches is greater than n_init, the trajectory will be converted from the initial indeterminate state to the deterministic state. If the detection frame has not been matched, it will directly enter the deletion state. Deep-in-Deep SORT refers to the introduced Re-ID model, as shown in Figure 4.

As shown in Figure 4, this is a deep model that extracts the target appearance information, and the model finally outputs a 128-dimensional vector. In the current situation, it seems that the effect of appearance features in Deep SORT tracking is not obvious.

3.5.2. Improvement Strategy Based on Deep SORT. From a common intuition, detection and tracking are two complementary problems, but the detection results are not always reliable. In crowded scenes, pose changes and occlusions often lead to detection failures, such as false detection, missed detection, and inaccurate boundaries. However, Deep SORT does not consider the problem of unreliable detection, and it directly deletes the trajectories that have not been successfully matched. The target association process is shown in Figure 5.

As shown in Figure 5, to handle unreliable detection in online mode, it extends traditional detection tracking by collecting candidates from the outputs of detection and tracking. However, combining the outputs of detection and tracking leads to an excess of candidates. Therefore, it creates a Regionalization-based Fully Convolutional Neural Network (R-FCN) classifier that classifies candidate results in terms of space. Each classified candidate region is defined as a region of interest (RoI). To explicitly embed spatial information into the score map, it divides an RoI into $k * k$ units. Each unit represents the spatial location information of the object, and each score map corresponds to this unit. During training, the ground truth values are randomly sampled as positive samples, and the same number of RoIs is taken from the background as negative samples.

It only uses the information of the last track to formulate the confidence of the track. It defines $R_{\text{detection}}$ as the detection result value associated with the trajectory and R_{track} as the tracking prediction value after the last detection association was successful. The definition of tracking trajectory confidence is as follows:

$$\begin{aligned} \text{con}_{\text{trk}} &= \max\left(1 - \log\left(1 + \alpha R_{\text{track}}, 0\right) \cdot \left(R_{\text{detection}} \geq 2\right) \cdot \mu,\right. \\ P &= p(y|B,x) \cdot \left(\mu(x \in C_{\text{det}})\right) + \text{con}_{\text{track}} \mu(x \in C_{\text{track}}), \end{aligned} \quad (16)$$

where C_{det} represents the detected candidate, C_{track} represents the candidate from the tracking output, and $\text{con}_{\text{track}}$ ranges from 0 to 1, which represents the penalized candidate in the uncertain trajectory. It finally filters out reasonable candidate targets according to nonmaximum suppression and obtains updated confidence.

3.6. Evaluation Criteria. The evaluation indicators of the tracker mainly include two aspects: speed and accuracy. In terms of speed, it uses frames per second (FPS) to evaluate the tracker, reflecting the real-time performance of the tracker. The evaluation of accuracy is more diversified, and the commonly used evaluation standards for image tracking are as follows: (1) center error, that is, the Euclidean distance between the tracking result position and the center of the standard position. However, this method has shortcomings, one is that it cannot measure the influence of the change of the target scale, and the other is that the definition of the center position of the target is not very accurate. (2) Regional overlap ratio, that is, the intersection ratio between the tracking result and the standard target frame, which generally uses the overlap of valid frames as an average. (3)

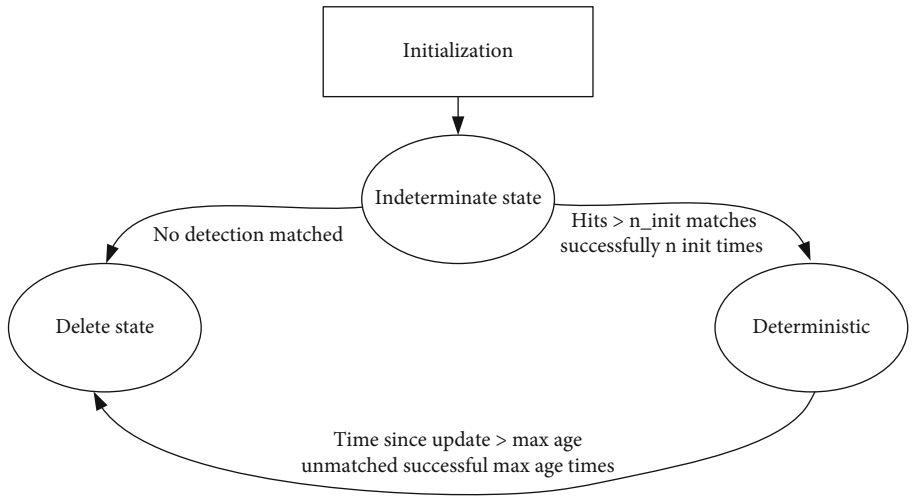

Figure 3: Three state transitions.

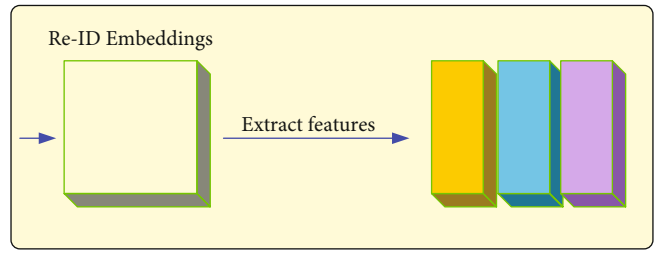

Figure 4: Re-ID model.

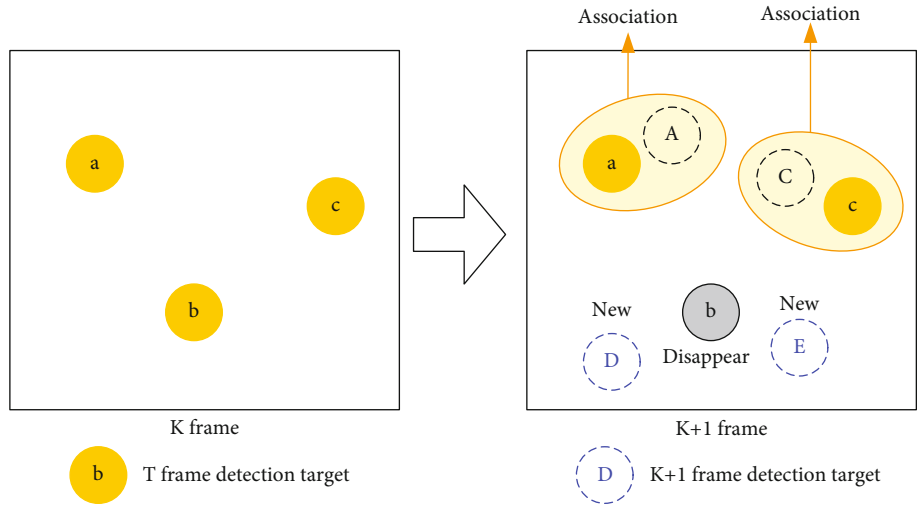

Figure 5: Target association process.

Tracking length, which is the number of frames from the start of tracking until the center error drops below the threshold. (4) Tracking failure rate, that is, the center error is greater than a certain threshold or the area overlap rate is less than a certain threshold, and the tracking failure rate is determined.

In the process of football video tracking, different trackers may produce two situations. One is that the tracking does not cause loss; that is, the target player can be located in each frame, but the area overlap is not high. The second is that the tracking loss is more serious, but the overlapping degree is high in the correct number of frames being tracked. To take into account the performance of the tracker in both cases, it adopts two indicators of accuracy and robustness to evaluate the tracker. Moreover, among several accuracy indicators of image tracking, the correlation between accuracy

and robustness is the weakest, which can comprehensively reflect the performance of a tracker. Accuracy is used to evaluate whether the tracking results are accurate. The higher the value, the higher the accuracy. The accuracy is borrowed from the definition of the area overlap ratio, as shown in

$$\text{Accuracy} = \frac{1}{N_{\text{valid}}} \sum_{i=1}^{N_{\text{valid}}} \varphi(i), \tag{17}$$

where N_{valid} represents the number of valid frames and (i) represents the tracker's accuracy on the i-th frame in the k -th repetition. Each tracker runs repeatedly in a sequence. $\varphi(i)$ is defined as

$$\varphi(i) = \frac{1}{N_{\text{rep}}} \sum_{k=1}^{N_{\text{rep}}} \varphi(i,k), \tag{18}$$

where N_{rep} is the number of repetitions and $\varphi(i,k)$ is the intersection ratio, as shown in

$$\varphi(t,k) = \frac{A_t(i,k) \cap A_{gt}(i,k)}{A_t(i,k) \cup A_{gt}(i,k)}, \tag{19}$$

where A_t represents the result box output by the tracking algorithm and A_{gt} represents the standard box in groundtruth. Robustness is used to evaluate the stability of the tracker, which represents the proportion of tracking failures in multiple tracking results. The larger the value, the worse the stability. Here, we use $F(k)$ to denote the number of times the tracker fails to track in the k-th repetition. Robustness is defined as

$$\text{Robustness} = \frac{1}{N_{\text{rep}}} \sum_{k=1}^{N_{\text{rep}}} F(k). \tag{20}$$

Tracking failure means that the overlap between the tracker output and the area of the standard box is 0.

4. Algorithm Research in 3D Tracking Image Analysis of Football Players

4.1. Player 3D Tracking and 3D Pose Estimation Based on Cross-View Correlation Matching

4.1.1. 2D Detection, Tracking, and 2D Pose Estimation of Players. Technologies such as visual object detection, tracking, and human pose estimation have developed rapidly in recent years. This chapter looks at the players in sports videos as the application of these techniques. First, it uses the proposed multitarget tracking method based on the context graph model to obtain the two-dimensional tracking trajectory of the player (the player detection result is provided by the ground truth). It then uses the CPN pose estimation method to estimate the 2D pose of the players in each tracking box on the trajectory. In this way, the two-dimensional trajectory information and two-dimensional attitude information of the players in each camera plane can be obtained. This facilitates further implementation of cross-view player matching, 3D pose estimation, and 3D tracking.

4.1.2. Cross-View Player Association Matching. The main task of this section is to obtain the two-dimensional trajectory and attitude information of players in each camera plane based on the previous section. It combines epipolar geometry constraints and depth appearance features for each player to match players across camera perspectives.

Inspired by this, this section intends to build a cross-view player relationship graph with players on each camera plane as nodes (player appearance features as node features) and connections of players on different camera planes as edges. It uses a multilayer graph convolutional neural network to supervise and learn the similarity relationship between each node (player) in the graph. The cross-view player appearance similarity learned based on the graphical model is more robust and discriminative than the similarity obtained directly by calculating the cosine distance. The cross-view player appearance similarity learning based on the graph model is shown in Figure 6.

As shown in Figure 6, it assumes that the multiview scene contains a total of 3 camera views and 3 target players. Among them, 3, 2, and 3 players can be seen in cameras 1, 2, and 3, respectively. After 2D detection and tracking of players in each camera plane, it renumbers all players. Then, it builds the graph with these 8 players as nodes. Each node is represented by the player's appearance feature vector (here, a pretrained Re-ID model is used to extract its appearance feature for each player). The edges in the graph represent connections between players. It should be pointed out that there is only connection between players located on different camera planes, and there is no connection between players on the same camera plane. Next, it introduces a graph convolutional neural network and uses it for similarity learning between players.

4.1.3. 3D Pose Estimation and 3D Tracking of Players. Two-dimensional tracking ID2D of players in each camera plane and cross-camera player matching results are obtained. In fact, the correlation matching of each player in time and space is completed, respectively. Next, this section will further implement 3D pose estimation and 3D tracking of players. On the one hand, for the cross-camera player matching group obtained by correlation matching in space, the three-dimensional pose information of each group of players can be obtained by using the triangulation algorithm or the 3DPS algorithm, respectively. On the other hand, combined with the two-dimensional tracking ID2D and cross-camera matching, according to the algorithm, the correlation matching of each player in the three-dimensional space can be realized naturally.

4.2. Experimental Results of Player 3D Tracking and 3D Pose Estimation.

Since the two core problems mainly solved in this chapter are 3D pose estimation and 3D multitarget tracking, the experiments in this chapter will also be carried out on these two tasks, respectively.

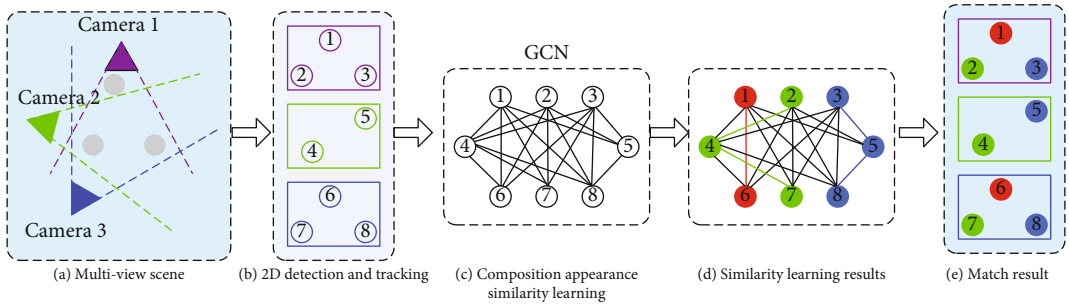

Figure 6: Cross-view player appearance similarity learning based on graph model.

4.2.1. Experimental Results of 3D Pose Estimation. The results obtained by each 3D pose estimation method in the campus dataset are shown in Table 1.

Table 1 shows the improvement effect of the method proposed in this chapter. Since the improvement of the method in this chapter is mainly reflected in the two cross-camera matching similarities (geometric similarity and appearance similarity), several sets of comparative experiments are mainly conducted for different similarities.

4.2.2. 3D Tracking Results. This chapter also projects the 3D tracking results back into each 2D camera plane to further verify the tracking effect. The comparison of the 3D tracking results of players in the APIDIS dataset is shown in Table 2.

As shown in Table 2, it can be seen from the comparison results that no matter in the 3D space or in the 2D camera plane, the tracking results obtained by the method based on the improved similarity matrix are clearly better than the tracking method based on the improved similarity. Especially for the appearance similarity, the original method based on the simple cosine similarity measure is almost difficult to complete the target matching across perspectives. The method based on graph model similarity metric learning proposed in this chapter can significantly improve the tracking effect. The comparison of pedestrian 3D tracking results in the campus dataset is shown in Table 3.

As shown in Table 3, when the cross-view similarity matrix is A, the MOTA values corresponding to the 3D tracking results and 2D projection results obtained by this method in the campus dataset are only 50 and 56.2. This is much lower than the tracking performance when based on other similarity matrices. The MOTA value of the obtained tracking result (92.6) is very close and significantly outperforms the other methods. Similarly, the improvement effects of the methods proposed in this chapter are similar to the campus dataset. It is worth mentioning that geometric similarity performs better on the campus dataset. The main reason for this is that the campus dataset contains only 3 people, so it is easy to distinguish. While the APIDIS dataset includes up to 10 individuals, the discriminativeness of geometric similarity decreases compared to the campus dataset, which includes only 3 individuals.

4.3. Motion Recognition Algorithm for Football Players. In the previous two chapters, the extraction and processing of features, that is, the tracking algorithm of Siamese-correlation filter fusion, were introduced. In this chapter, to test its comprehensive performance, after introducing the football tracking dataset and tracking evaluation indicators, it conducts experiments and compares it with other existing trackers.

The specific data are shown in Table 4.

As shown in Table 4, the dataset divides football videos into four scenes according to the different environments of the target players, namely, unoccluded scenes, occlusion scenes of players in the same team, occlusion scenes of players from different teams, and mixed dense scenes. The dataset contains a total of 80 video sequences and a total of 19908 video images. The resolution of each frame image is 624∗352. Player positions for each video sequence are represented in the text file groundtruth.txt. The first line in the text represents the start frame number and the end frame number of the tracking sequence, and each subsequent line represents the rectangular box position of the tracking target, which is represented by a quadruple (x, y, width, height). The number of occlusions in a single video sequence in the dataset ranges from 1 to 7 times. Occlusion can be divided into complete occlusion and partial occlusion according to the degree of occlusion.

4.4. Comparative Analysis of Tracking Accuracy

4.4.1. Experimental Environment. The hardware environment is as follows: Intel Core CPU i7 @ 2.8 GHz and MEM 16 G. The operating system used is macOS 10.14.4. The programming tools used are PyCharm 2017.1.4 and MATLAB R2017a.

4.4.2. Contrast Tracker. In the experimental part, in addition to the SiamCF (Siamese and correlation filter tracker) proposed in this paper, six tracking algorithms are selected for comparison. These six algorithms are all tracking algorithms for general fields. To reflect the performance differences of trackers of different methods, the algorithms selected in this comparative experiment include deep learning algorithm, correlation filtering algorithm, and Siamese network algorithm. The six tracking algorithms are DLT algorithm, MOSSE algorithm, KCF algorithm, CCOT algorithm, ECO algorithm, and Siamese algorithm. The DLT algorithm belongs to the deep learning algorithm; MOSSE, KCF, CCOT, and ECO belong to the correlation filtering algorithm; and Siamese belongs to the twin network algorithm. To avoid the influence of different experimental parameters,

Table 1: Comparison of player 3D pose estimation results in campus dataset.

	Player 1	Player 1	Player 1	Average value
Geometric similarity	69.18	87.57	94.06	83.60
	89.18	79.89	94.78	87.95
Appearance similarity	83.47	53.23	40.07	58.92
	89.18	80.53	87.10	85.60
True similarity	90.00	87.57	94.78	90.78
Geometry + appearance	90.82	79.89	94.79	88.50

Table 2: Comparison of player 3D tracking results in APIDIS dataset.

Similarity type	3D				2D			
	MOTA	FP	FN	FM	MOTA	FP	FN	FM
Geometric similarity	73.6	1952	1993	111	86.6	1456	2586	633
	74. l	1364	2512	426	87.9	2047	1798	251
Appearance similarity	-46.5	8171	13510	572	11.8	8443	15336	3845
	76.4	1576	1946	305	88.7	1579	1938	523
True similarity	79.1	1402	1729	228	89.0	1647	1828	391
Geometry + appearance	100	0	0	0	92.6	FP	1133	227

Table 3: Comparison of player 3D tracking results in campus dataset.

Similarity type	3D				2D			
	MOTA	FP	FN	FM	MOTA	FP	FN	FM
Geometric similarity	89.8	19	19	6	92.6	45	38	16
	96.5	1	12	8	96.2	8	34	22
Appearance similarity	50	31	156	24	56.2	52	420	73
	84.8	25	27	10	91.1	40	39	26
True similarity	89.6	19	20	7	92.1	45	41	19
Geometry + appearance	100	0	0	4	99.4	7	0	12

Table 4: Soccer dataset distribution.

Serial number	Scenes	Video sequence number	Number of video sequences	Frame number
1	Unobstructed scene	1-20	20	5385
2	Teammates block the scene	21-40	20	3954
3	Players from different teams block the scene	41-60	20	5060
4	Mixed dense scenes	61-80	20	5509
Total	Scenes	Video sequence number	80	19908

it tries to keep the related parameters in different algorithms consistent, and the parameters of the same algorithm in different scenarios remain unchanged.

4.4.3. Tracking in Four Different Scenarios

(1) Unobstructed Scene. An unobstructed scenario is one where there are no other players around the tracked target player. This scene is the simplest set of scenes in a football video. However, in an unobstructed scene, the target player usually has a faster speed and a larger shape change. Although the ability to discriminate between the target and the background is not high, the tracker needs to respond in time to the changes of the object. The accuracy of the tracker in this scenario is shown in Figure 7.

It can be seen from Figure 7 that the KCF, CCOT, ECO, Siamese, and SiamCF algorithms can handle unoccluded scenes well and have good performance in accuracy. The

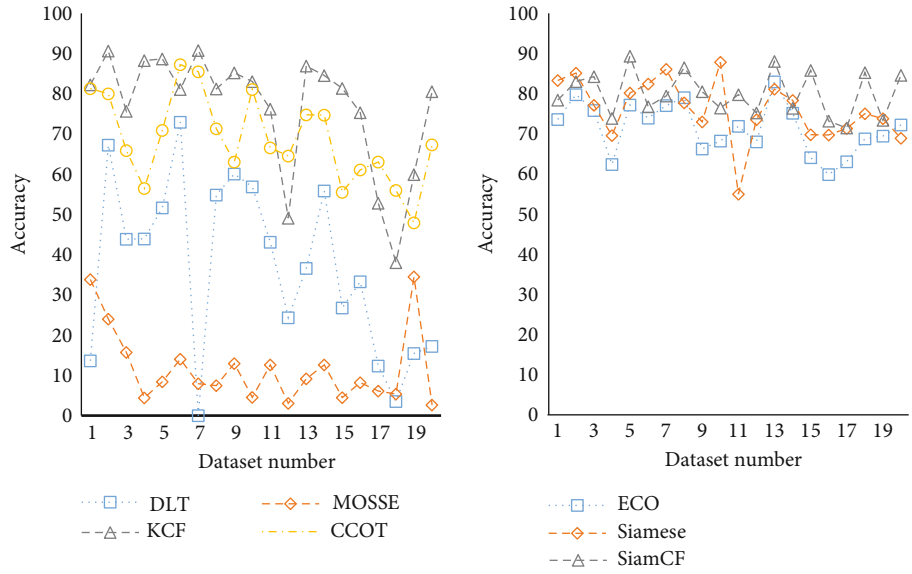

Figure 7: Tracking accuracy in unobstructed scene.

performance of DLT and MOSSE is less stable, and DLT completely loses the target in dataset 7 and dataset 18. SiamCF can achieve more than 80 accuracies on 9 datasets. Compared with the separate correlation filter ECO and the Siamese network Siamese algorithm, it shows high accuracy.

(2) Players on the Same Team Block the Scene. Teammate occlusion scenarios are very challenging scenarios. Players on the same team often have the same jersey and have very similar characteristics under long-range lenses. This is easy to cause a tracking drift, and once the tracking drifts to a player on the same team, it is often difficult to retrieve it. This scenario requires the tracker to have a strong ability to discriminate the subtle features of the tracked target and to have a strong error correction capability for the drifted target, so that the tracker can relocate after drifting. The accuracy of the tracker in this scenario is shown in Figure 8.

As can be seen from Figure 8, CCOT, ECO, Siamese, and SiamCF outperform the other three trackers in accuracy. Among them, CCOT performs better on datasets 28, 29, 31, and 32. ECO performed well on dataset 38, and Siamese also performed well on datasets 22 and 36. Overall, SiamCF outperforms other algorithms in comprehensive performance and in most dataset scenarios. For example, on dataset 37, other trackers have more or less offset.

(3) Players from Different Teams Block the Scene. The occlusion scene of different team players is the most common scene in football videos. Players from different teams are highly confrontational and prone to conflict. One or more players from different teams often appear around the target player. Players from different teams have different uniforms and have large differences in characteristics, which are easier to discriminate than players on the same team. However, due to the high probability of occlusion by players from different teams and many video sequences, the tracking accuracy in this scene has a greater impact on the final tracking accuracy. The accuracy of the tracker in this scenario is shown in Figure 9.

It can be seen from Figure 9 that the occlusion of players from different teams is easier to discriminate than the occlusion of players from the same team. However, on dataset 50, the accuracy of all trackers is not high due to the situation where the target player is completely occluded. After the occlusion is over, all trackers have tracking drift phenomenon. However, continue to track, it is found that SiamCF is able to relocate to the target player after tracking drift due to the introduction of the tracking result correction strategy. Overall, SiamCF has the best accuracy performance on 10 datasets, and its comprehensive performance is better than other trackers. Secondly, CCOT's performance in accuracy is also very good and better than the Siamese network algorithm. The performance of the KCF algorithm is very unstable, and the accuracy is high when the tracking is correct, but the drift is serious.

(4) Mixed Dense Scenes. Mixed-intensive scenarios usually occur when the confrontation is very intense, such as scoring a penalty area or a free kick in the front court. In this scene, the tracking target will change drastically whether it is occluded or deformed, such as alternative occlusion by players of the same team and different teams and players falling. Robust performance of the tracker in this scenario is a very big challenge. The accuracy of the tracker in this scenario is shown in Figure 10.

As can be seen from Figure 10, the tracker exhibits lower accuracy values on many datasets. In dataset 64, the accuracy values of CCOT and ECO are lower than 50, and the tracking accuracy is very low. SiamCF accuracy values perform better than the rest of the trackers. KCF performs best on datasets 61, 68, and 80 but is unstable. It is almost at the bottom of datasets 74, 79, etc. Siamese is more prominent on

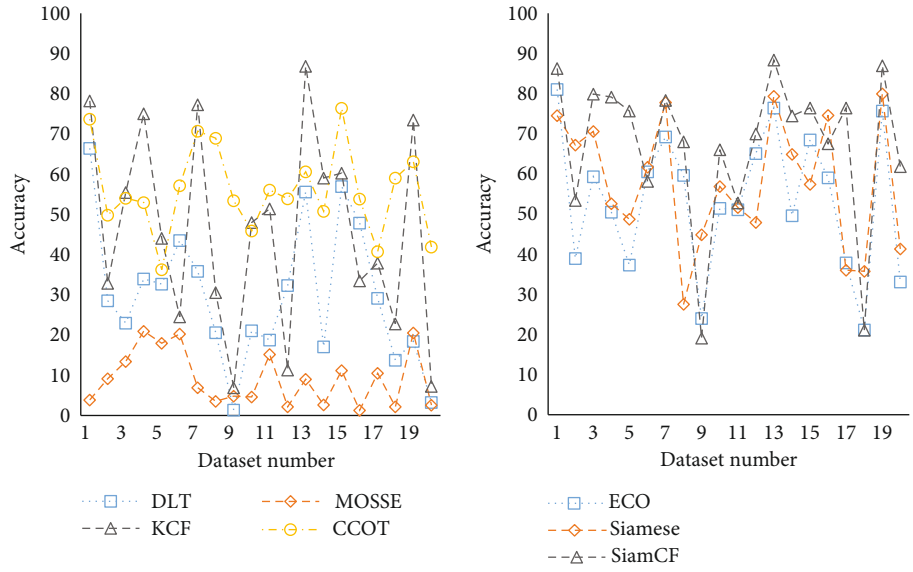

Figure 8: Tracking accuracy in the scene of teammate occlusion.

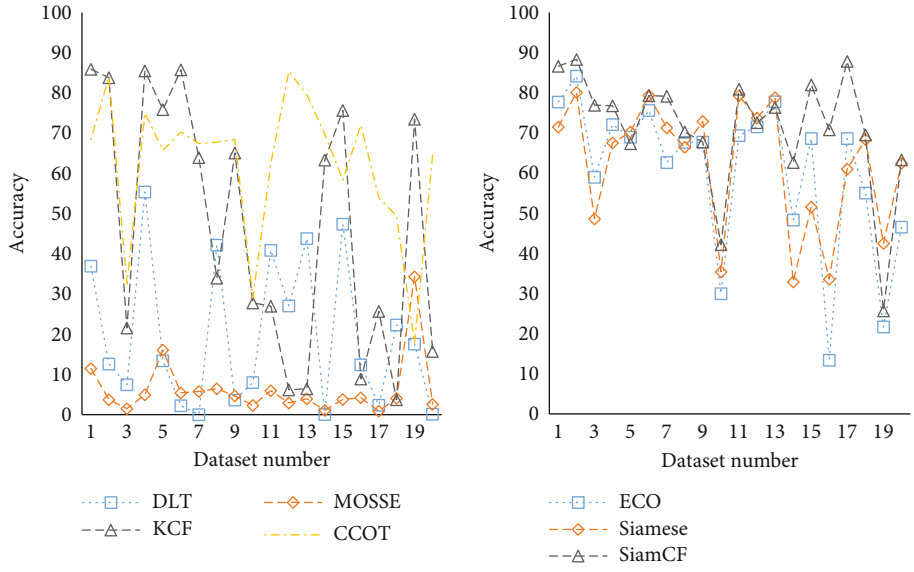

Figure 9: Tracking accuracy in the scene where players from different teams are occluded.

63, 65, 71, and 76. SiamCF has the best performance on 11 datasets and far outperforms other trackers in the accuracy of this scene. Especially on dataset 66, the deformation and occlusion of the target player are serious, and SiamCF still does not lose the target.

4.5. Comparative Analysis of Tracking Speed. In terms of temporal performance, the average frame rate (FPS) is used as the evaluation index, indicating the number of video frames that can be tracked per second. According to their performance in time performance, the order is MOSSE> KCF > DLT > Siamese>ECO > SiamCF>CCOT. When the FPS of the tracker exceeds 20 frames, the tracker can be considered to meet the real-time requirements. From this point of view, except MOSSE and KCF, the other trackers do not meet the real-time requirements. The average frame rate of each tracker is shown in Table 5.

As shown in Table 5, the MOSSE algorithm has the fastest tracking speed because it only extracts a traditional single feature, and the convolution operation in the correlation filter is converted to the frequency domain through Fourier transform to perform point multiplication, and no other redundant operations are caused. Although KCF introduces a variety of features and tracks in high-dimensional space, its method of constructing samples using circulant matrices greatly reduces the amount of sample computation. The calculation of all samples can be completed by only calculating one generated sample, which is also very prominent in frame rate performance. CCOT introduces multifeature fusion including deep features, and extending the features to the

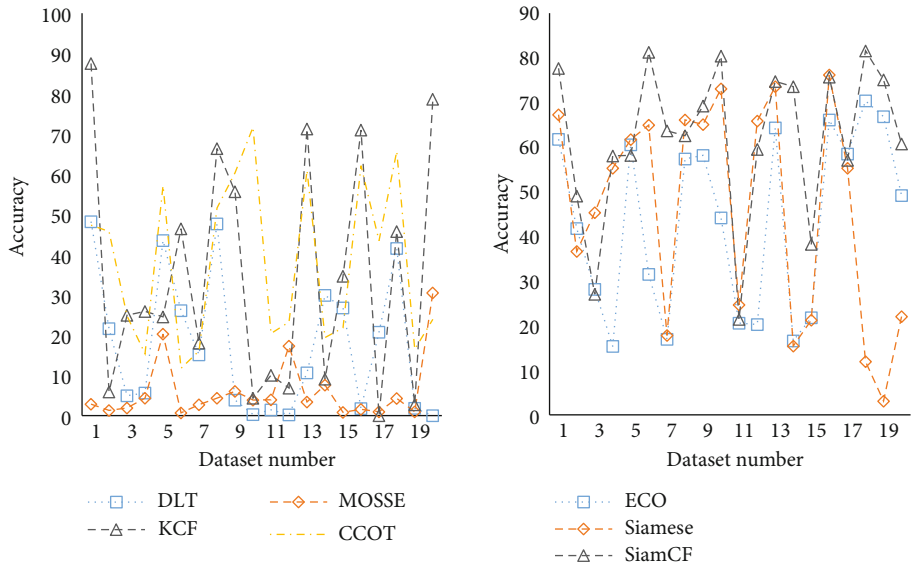

Figure 10: Tracking accuracy in mixed dense scenes.

Table 5: Tracker average frame rate.

Tracker	Average frame rate (FPS)
DLT	7.31
MOSSE	169.9
KCF	134.72
CCOT	0.95
ECO	3.37
Siamese	6.28
SiamCF	1.69

continuous domain increases the computational load. Accuracy has improved, but framerate drops are noticeable. ECO performs acceleration operations such as dimension reduction and sample clustering on the basis of CCOT, and the frame rate performance is better. Siamese is a Siamese network tracker, and its two input branches perform two feature extraction operations on the template image and the detection image, respectively, which increases its computational load, and its speed decreases compared to the standard network model DLT tracker. SiamCF combines correlation filtering and Siamese network, adding filter layers to the Siamese network. Although it performs the best in terms of accuracy and robustness, the tracker does not perform as well as a single Siamese network or correlation filtering algorithm in terms of average frame rate due to the introduction of a new amount of computation.

5. Conclusions

This paper compares and analyzes the tracking results and time performance experiments of multiple trackers. In addition to the tracking algorithm proposed in this paper, it also selects six other comparison algorithms. Tracking experiments were conducted under four different scenarios in football videos. It explains the characteristics of these four scenarios, respectively, analyzes the various performances of different trackers in these four scenarios, and compares these trackers in terms of time performance. Experimental results show that different trackers perform differently in different scenarios. However, in general, SiamCF integrates correlation filtering and Siamese network structure and makes improvements to the characteristics of the football field. It has the highest average accuracy and average robustness and average frame rate performance under the current football dataset. The speed of SiamCF does not meet the requirements of real-time performance, and it is stretched in real-time tracking scenarios. Due to a series of operations such as extraction of multiple sets of features, feature continuity, dimensionality reduction, correlation, filtering, and fusion of twin networks, the tracking algorithm has a large amount of computation, which affects the tracking speed.

Data Availability

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by "The 13th 5-year Plan" of the Chinese National Education Science Program (ELA170479) which is "The construction of the dynamic mechanism model for the development of new campus football for Chinese teenagers."

References

[1] S. Timothy and H. S. Koo, "Bicyclist biomotion visibility aids: a 3D eye-tracking analysis," *International Journal of Clothing Science and Technology*, vol. 29, no. 2, pp. 262–269, 2017.

[2] Y. J. Lee and M. W. Park, "3D tracking of multiple onsite workers based on stereo vision," *Automation in Construction*, vol. 98, no. FEB., pp. 146–159, 2019.

[3] Y. Hou, Z. Deng, X. Cheng, and T. Ikenaga, "View priority based threads allocation and binary search oriented reweight for GPU accelerated real-time 3D ball tracking," *IEICE Transactions on Information and Systems*, vol. E101.D, no. 12, pp. 3190–3198, 2018.

[4] Y. Zhou, T. Feng, S. Shuai, X. Li, L. Sun, and B. L. Duh, "EDVAM: a 3D eye-tracking dataset for visual attention modeling in a virtual museum," *Frontiers of Information Technology & Electronic Engineering*, vol. 23, no. 1, pp. 101–112, 2022.

[5] M. H. Nguyen, C. C. Hsiao, W. H. Cheng, and C. C. Huang, "Practical 3D human skeleton tracking based on multi-view and multi-Kinect fusion," *Multimedia Systems*, vol. 28, no. 2, pp. 529–552, 2022.

[6] R. Mendicino, G. T. Forcolin, M. Boscardin et al., "3D trenched-electrode sensors for charged particle tracking and timing," *Nuclear Instruments and Methods in Physics Research Section A Accelerators Spectrometers Detectors and Associated Equipment*, vol. 927, pp. 24–30, 2019.

[7] X. Liu, Y. Xu, L. Zhu, and Y. Mu, "A stochastic attribute grammar for robust cross-view human tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 2884–2895, 2018.

[8] R. J. Aughey, K. Ball, S. J. Robertson et al., "Comparison of a computer vision system against three-dimensional motion capture for tracking football movements in a stadium environment," *Sports Engineering*, vol. 25, no. 1, pp. 1–7, 2022.

[9] F. R. Goes, M. Kempe, L. A. Meerhoff, and K. A. P. M. Lemmink, "Not every pass can be an assist: a data-driven model to measure pass effectiveness in professional soccer matches," *Big Data*, vol. 7, no. 1, pp. 57–70, 2019.

[10] C. Dai and Y. Lu, "Improved biological image tracking algorithm of athlete's cervical spine health," *Revista Brasileira de Medicina do Esporte*, vol. 27, no. 3, pp. 274–277, 2021.

[11] C. Zhu, "Applying edge computing to analyse path planning algorithm in college football training," *International Journal of Systems Assurance Engineering and Management*, vol. 12, no. 4, pp. 844–852, 2021.

[12] V. L. Goosey-Tolfrey, J. Zepetnek, M. Keil, K. Brooke-Wavell, and A. M. Batterham, "Tracking within-athlete changes in whole-body fat percentage in wheelchair athletes," *International Journal of Sports Physiology and Performance*, vol. 16, no. 1, pp. 1–6, 2020.

[13] A. Morozzi, S. Sciortino, L. Anderlini et al., "3D diamond tracking detectors: numerical analysis for timing applications with TCAD tools," *Journal of Instrumentation*, vol. 15, no. 1, pp. C01048–C01048, 2020.

[14] Q. Meng, Y. Xu, Z. Wu, S. Wang, and Z. Guo, "3D speckle tracking imaging in evaluation of left ventricular function in patients with right ventricular dual chamber septal pacing," *Chinese Journal of Medical Imaging Technology*, vol. 34, no. 7, pp. 1019–1023, 2018.

[15] S. Manmohan, A. Akram, and P. Clive, "Enhanced 3D localisation accuracy of body-mounted miniature antennas using ultra-wideband technology in line-of-sight scenarios," *IET Microwaves, Antennas and Propagation*, vol. 12, no. 1, pp. 1–8, 2018.

[16] Z. Guo, H. Bu, L. Song, J. Li, and Z. Feng, "Experimental test of a 3D parameterized vane cascade with non-axisymmetric endwall," *Aerospace Science and Technology*, vol. 85, no. FEB., pp. 429–442, 2019.

[17] W. Mahardika, S. Wibirama, R. Ferdiana, and S. S. Kusumawardani, "A novel user experience study of parallax scrolling using eye tracking and user experience questionnaire," *International Journal on Advanced Science Engineering and Information Technology*, vol. 8, no. 4, pp. 1226–1233, 2018.

[18] Z. Yanxing, L. Lei, and L. Beibei, "The discussion on interior design mode based on 3D virtual vision technology," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 23, no. 3, pp. 390–395, 2019.

[19] T. Long, "Research on application of athlete gesture tracking algorithms based on deep learning," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 9, pp. 3649–3657, 2020.

[20] L. Tang, C. Zhu, and H. Luo, "Training prediction and athlete heart rate measurement based on multi-channel PPG signal and SVM algorithm," *Journal of Intelligent Fuzzy Systems*, vol. 40, no. 4, pp. 1–12, 2020.