A Sport Athlete Object Tracking Based on Deep Sort and Yolo V4 in Case of Camera Movement

Yao Zhang[1], Zhiyong Chen*[1], Bohan Wei[2]
Department of Physical[1],Department of Science[2],Tongji Zhejiang College, Jiaxing, 314001, China
e-mail: 75351981@qq.com

2020 IEEE 6th International Conference on Computer and Communications (ICCC) | 978-1-7281-8635-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICCC51575.2020.9345010

***Abstract*—Object tracking task has always been a major problem in the CV field. It is different from object detection. Object detection only needs to identify the type of object, while tracking task needs to identify its unique identity when a specific object is detected, such as REID problem. In sports-related fields, object tracking technology also has huge applications. For example, in football matches, camera tracking of footballs and tracking of athletes require this technology. This paper takes NBA and World Cup related scenes as the identification object, and aims to establish a tracking system for all players in the game, complete the real-time tracking of each athlete, so as to obtain relevant track information. In addition, the system can also help teachers review the competition in related education links and find the shortcomings of each student. Unlike most of filtering algorithms in past, this paper used the more cutting-edge deep learning technology in recent years, the YoloV4 and Sort's advanced version Deep Sort.**

Keywords-Object tracking; REID; sports; Deep Learning; Yolo V4; Deep Sort

I. INTRODUCTION

Object tracking is an important research direction in computer vision with a wide range of applications, such as: video surveillance, human-computer interaction, unmanned driving, etc. In the past two to three decades, the visual object tracking technology has made great progress, especially the deep learning be used in object tracking in the last two years has achieved satisfactory results, making the object tracking technology a breakthrough.

In sports, the application field of object tracking technology is also very extensive. In some sports competetion, such as NBA, when the team makes the corresponding offensive and defensive plan, it usually focuses on the opposing star players, such as the Golden State Warriors Curry's three-pointers and the Los Angeles Lakers James' fast break. In the analysis of opposing players, offensive and defensive lines need special attention. If the defensive side can grasp the opposing player's offensive line, it can well formulate the corresponding defense. And many players have a preferred offensive line, the defender can find their offensive line by analyzing multiple games.

This paper proposed a tricking method based on deep learning, it can circumvent many shortcomings of traditional algorithms. For example, the traditional model is influenced by appearance distortions, lighting changes, fast motion and motion blur, and background similarities, resulting in poor algorithm results. By using this technology, we can use a fixed camera to record each player's movement trajectory throughout the game for data analysis and related strategy formulation. Similarly, this technology can also be applied to other scenes that need to track trajectories and objects.

II. RELATED WORK

Before the deep learning and correlation filtering tracking algorithm appeared, the field of object tracking has been using classic algorithms. At this stage, the algorithms mostly use probability density and image edge features as tracking standards, make the object search direction has been along the direction of the rising probability gradient, such as the Meanshift[1], Particle Filter[2] and Kalman Filter[3].

Meanshift is a tracking method based on the probability density distribution, which makes the search of the object always follow the direction of the rising probability gradient, and iteratively converges to the local peak of the probability density distribution. First, Meanshift models the object, such as using the color distribution of the object to describe the object, and then calculates the probability distribution of the object on the next frame of image, so as to iteratively obtain the local densest area. Meanshift is suitable for situations where the color model of the object and the background are quite different, and it was also used for face tracking in the early days. Due to the fast calculation of the Meanshift method, many of its improved methods are still applied so far.

Particle Filter is based on particle distribution statistics. Taking tracking as an example, the tracking object is modeled first, and a similarity metric is defined to determine the matching degree between the particles and the object. In the process of object search, it will sprinkle some particles according to a certain distribution (such as uniform distribution or Gaussian distribution), count the similarity of these particles, and determine the possible position of the object. At these positions, more new particles are added in the next frame to ensure that the object is tracked with a greater probability.

Kalman filter is often used to describe the motion model of the object. It does not model the characteristics of the object, but models the motion model of the object. It is often used to estimate the position of the object in the next frame. In addition, the another classic tracking method is optical flow tracking based on feature points. Some feature points are extracted on the object, and then the optical flow matching points of these feature points are calculated in the next frame, and the position of the object is obtained by statistics. In the process of tracking, new feature points need to be constantly added, and feature points with poor

confidence are deleted to adapt to the changes in the shape of the object in motion. In essence, it can be considered that optical flow tracking is a method that uses a set of feature points to characterize a object model.

After the advent of deep learning, the classic tracking methods have been abandoned, because these classic methods cannot handle and adapt to complex tracking changes, and their robustness and accuracy are surpassed by cutting-edge algorithms.

In this paper, we proposed the tracking algorithm based on deep learning, use Yolo V4[4] and Deep Sort[5]. After comparing various models, we found that when the classic algorithm processes the human body tracking of athletes in sports competitions, the classic algorithm will lose the object due to the overlap of people and the blurred image of the lens during high-speed movement.

Therefore, in order to solve the complex situation of fast motion and overlapping objects in sports scenes, this article combines two leading algorithms in the field of object detection and object tracking to realize the motion tracking of each athlete in the video.

Figure 1. The figure we test by KCF+DSST[6], the test video is BAR vs CEL in World Cup, Messi dribbling the ball solo. At the beginning of the video,we chose the Messi(as the left figure), and then CEL member overlapped with Messi, the model tracked to the wrong object.

III. DeepSort & Yolo V4

A. Object Tracking & Object Detection

In many situations, unprofessionals will often confuse these two concepts. From the visual image, the effect of the two algorithms seems to use a rectangular frame to label the target object in the image or video, and then identify the type of the object.

However, compared to Object detection, Object tracking has an additional function, which is to re-identify the characteristics of the target. For example, the video scene is a shopping mall, and the Object detection algorithm will label everyone in the video, but it does not know who is who. For it, the characters in each frame of the image are different people, or simply "people". The Object tracking algorithm will number each person and recognize the person based on characteristics such as appearance and clothing. Even if the lens is deflected or the angle is changed, it can still recognize everyone in the picture like an old friend.

Figure 2. The difference between object detection and obeject tracking.

B. Sort

Sort[7] is a fast online multi-target tracking (MOT) algorithm, based on the TBD (Traking-by-Detection) strategy. These features determine the practicability of SORT. SORT played a benchmark role in the MOT field at that time.

In Sort, the author uses the CNN-based network Faster RCNN and the traditional pedestrian detection ACF two detection models. In addition, in order to solve the action prediction and data association, the very efficient Kalman filter and Hungarian algorithm are used.

As the first-generation in Sort series algorithm, it pays more attention to the speed and efficiency of the algorithm. Compared with the SOTA algorithm in 2016, its speed can reach 260Hz, which is 20 times faster than the former. However, the defect of this algorithm is that when the target is lost, it cannot be retrieved, and the ID can only be re-updated through detection. This does not conform to the common sense of the tracking algorithm, so the Deep Sort be proposed to solve this problem.

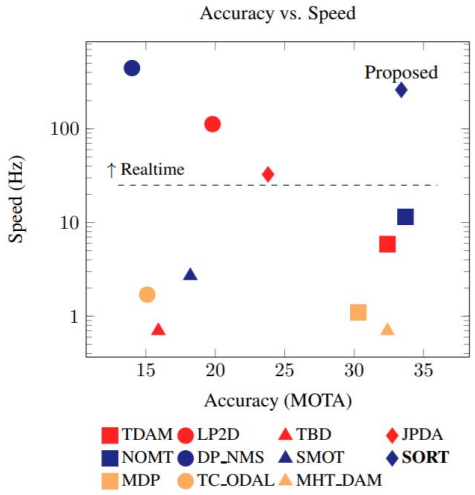

Figure 3. Each marker indicates a trackers accuracy and speed measured in frames per second.

C. Deep Sort

The current mainstream object tracking algorithms are based on the Tracking-by-Detection (detection & tracking to make the effect more stable) strategy, it means object tracking is based on the result of object detection. DeepSORT also uses this strategy.

DeepSort is an improvement based on Sort target tracking. It's a deep learning model trained offline on the REID data set. In the real-time target tracking process, the apparent feature of the target is extracted for nearest neighbor matching, which can improve the object tracking effect in the presence of occlusion. At the same time, the problem of target ID loss is also reduced.

The core idea of the algorithm is to use a traditional single-hypothesis tracking method, which uses a recursive Kalman filter to associate frame-by-frame data.

DeepSort uses (u,v,γ,h,$\dot{x}$,$\dot{y}$,$\dot{\gamma}$,$\dot{h}$) parameters to describe the state of motion. The (u,v) is the center coordinates of the bounding box, r is the aspect ratio, and h is the height. The remaining four variables represent speed information in the image coordinate system. The algorithm uses a standard Kalman filter based on a constant velocity model and a linear observation model to predict the target motion state, and the predicted result is (u, v, r, h).

The author of DeepSort also considered the association of movement information and the association of target appearance information.

$$d^{(1)}(i,j) = (d_i - y_i)^T S_i^{-1}(d_i - y_i), \qquad (1)$$

d_j is position of the j-th detection frame, y_i is the predicted position of the i-th tracker to the target, S_i is the covariance matrix between the detection position and the average tracking position. The Mahalanobis distance takes into account the uncertainty of the state measurement by calculating the standard deviation between the detection position and the average tracking position. If the Mahalanobis distance of a certain association is less than the specified threshold t, then the association to set the motion state is successful. The formulated in Eq.2.

$$b_{i,j}^{(1)} = \mathbb{I}\left[d^{(1)}(i.j) \le t^{(1)}\right], t^{(1)} = 9.4877, \qquad (2)$$

when the uncertainty of motion is very low, Mahalanobis distance matching is a suitable method, but when kalman filtering is used for motion state estimation in image space, it is only a rough prediction. Especially when the camera is in motion, the Mahalanobis distance association method will be invalid and causing an ID switch error. Therefore, DeepSort introduced the second association method, for each detection block d_j, find a feature vector r_i, and $\| r_i\|$=1. For each tracking target, DeepSort construceted gallery, and saved last 100 frames successfully associated with each tracking target, then calculate the minimum cosine distance between the latest 100 successfully associated feature sets of the i-th tracker and the feature vector of the j-th detection result of the current frame. The formulated in Eq.3.

$$d^{(2)}(i,j) = \min\{1 - r_j{}^T r_k^i | r_k^i \in R_i\}, \qquad (3)$$

then use the linear weighting method to obtain the final formula:

$$c_{i,j} = \lambda d^{(1)}(i,j) + (1-\lambda) d^{(2)}(i,j). \qquad (4)$$

When $c_{i,j}$ between the threshold, it can be considered that the target is the same one.

The Table 1 shows the result on MOT16 [8].

TABLE I. TRACKING RESULT ON THE MOT16[8]

		MOTA	*MOTP*	*MT*	*ML*	*ID*	*FM*	*FP*	*FN*	*RUNTIME*
KDNT[9]	BATCH	68.2	79.4	41.0%	19.0%	933	1093	11479	45605	0.7HZ
LMP[10]	BATCH	71	80.2	46.9%	21.9%	434	587	7880	44564	0.5HZ
MCMOT_HDM[11]	BATCH	62.4	78.3	31.5%	23.2%	1394	1318	9855	57257	35HZ
NOMTwSDP16[12]	BATCH	62.2	79.6	32.5%	31.1%	406	642	6119	63352	3HZ
EAMTT	ONLINE	52.5	78.8	19.0%	34.9%	910	1321	4407	81223	12HZ
POI[9]	ONLINE	66.1	79.5	34.0%	20.8%	805	3093	5061	55914	10HZ
SORT[7]	ONLINE	59.8	79.6	25.4%	22.7%	1423	1835	8968	63245	60HZ
DEEP SORT[5]*	ONLINE	61.4	79.1	32.8%	18.2%	781	2008	12852	56668	40HZ

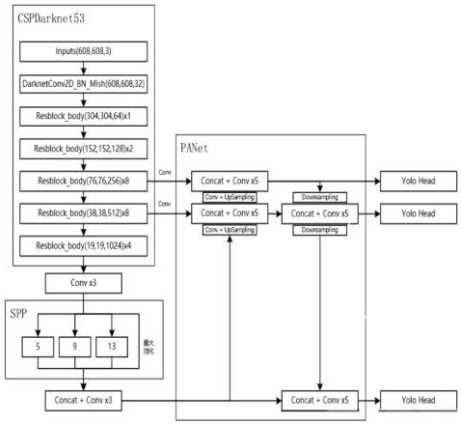

Figure 4. YoloV4 network structure.

D. Yolo V4

As the technical-grade object detection algorithm, Yolo v4 took over the "Best object detection algorithm" of YoloV3, it improved MAP, reduced training cost and improved FPS.

IV. CONCLUSION

In this paper,we used INRIA Person Dataset to train Yolo V4, the result as Fig.5, the max iou is 0.97673.

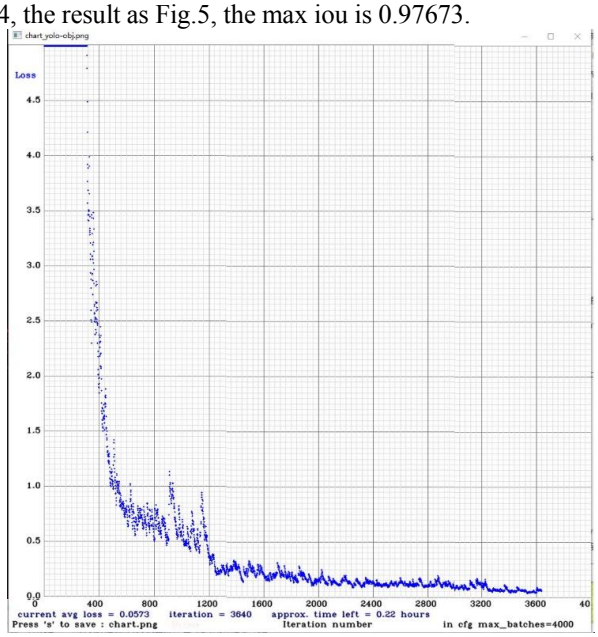

Figure 5. The loss curve figure of Yolo V4 in INRIA Person Dataset.

Then we used Market1501 & MARS datasets to train DeepSort. The result as Tabel 2.

TABLE II. THE RESULT OF DEEPSORT IN MARKET1501 & MARS

	Score
mAP	68.2%
CMC curve, Rank-1	82.3%
CMC curve, Rank-5	90.4%
CMC curve, Rank-10	93.7%

We tested the model in BAR vs CEL video, the initial detection effect figure is as follows.

Figure 6. The initial detection effect.

The player marked incorrectly above is marked as number nine by the model (pink rectangle).

Figure 7. Player 9 is the wrong tracked by KCF+DSST

In two seconds later, the player 9 and player 16 had a confrontation, which appears to be overlapping in the image

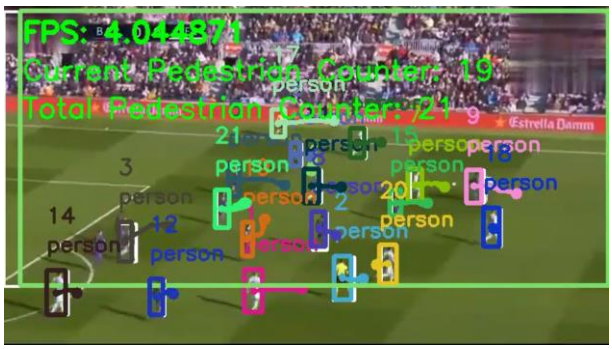

Figure 8. Player 9 and Player16 overlapped.

When player 9 grabs the ball and the two players separated, the model can still identify the player 9 and player 16(pink and red rectangle).

Figure 9. Player 9 and Player 16 separated.

It can be seen from the above results that the model can solve the object tracking problem well in this kind of scene

with small targets and slow lens shift. In order to verify the effect of the model under fast lens, we tested an NBA image.

Samely, we chosed the DEN vs LAL's video.

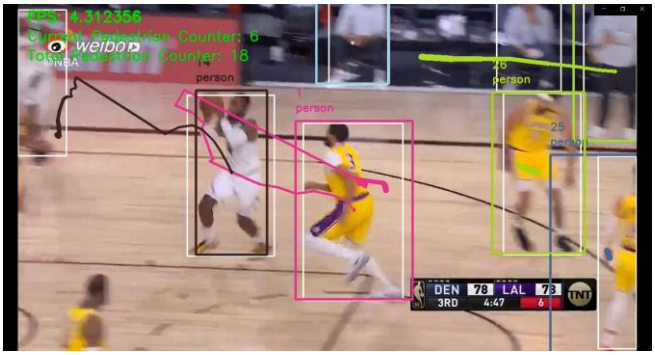

Figure 10. Player 14 is Lebron James.

In the next five seconds, he will complete a layup and tip-up action, although the final tip-up was blocked.

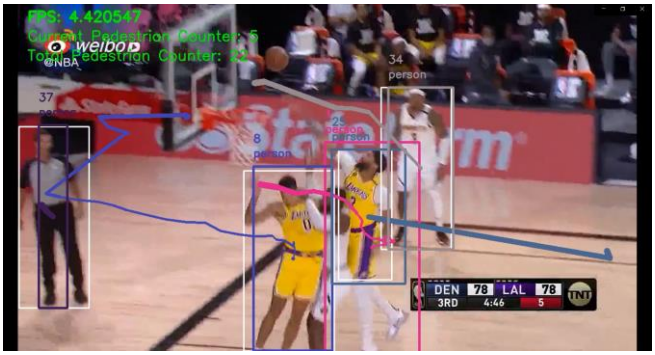

Figure 11. Lebron James complete the layup, and the detection loss the Lebron because of the overlap.

When the three people separated, the detection model re-identified the human body, and DeepSort also successfully identified the human body as Lebron(Player 14)

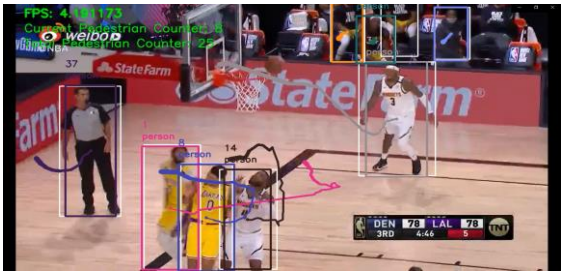

Figure 12. Lebron(Player14) been re-identified after three players separated.

ACKNOWLEDGMENT

This work was financially supported by Jia Xing Administration of Science & Technology(2020AY10031-Research on the development and application of edible fungus maturation and pathological analysis model construction and expert system based on machine vision). And supported by Zhejiang School Sports Association (zgtx202008 Construction and Application of Analysis and Recognition Model of Physical Education Classroom Teaching Behavior Based on Computer Vision)

REFERENCES

[1] Comaniciu D, Meer P. Mean shift: A robust approach toward feature space analysis[J]. IEEE Transactions on pattern analysis and machine intelligence, 2002, 24(5): 603-619.

[2] Van Der Merwe, R., Doucet, A., De Freitas, N., & Wan, E. A. (2001). The unscented particle filter. In Advances in neural information processing systems (pp. 584-590).

[3] Kalman, R. E. (1960). A new approach to linear filtering and prediction problems.

[4] Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv preprint arXiv:2004.10934.

[5] Wojke, N., Bewley, A., & Paulus, D. (2017, September). Simple online and realtime tracking with a deep association metric. In 2017 IEEE international conference on image processing (ICIP) (pp. 3645-3649). IEEE.

[6] J. F. Henriques, R. Caseiro, P. Martins, J. Batista,"High-Speed Tracking with Kernelized Correlation Filters", TPAMI 2015.

[7] Bewley, A., Ge, Z., Ott, L., Ramos, F., & Upcroft, B. (2016, September). Simple online and realtime tracking. In 2016 IEEE International Conference on Image Processing (ICIP) (pp. 3464-3468). IEEE.

[8] A. Milan, L. Leal-Taixe, I. Reid, S. Roth, and ´ K. Schindler, "Mot16: A benchmark for multi-object tracking," arXiv preprint arXiv:1603.00831, 2016.

[9] F. Yu, W. Li, Q. Li, Y. Liu, X. Shi, and J. Yan, "Poi:Multiple object tracking with high performance detection and appearance feature," in ECCV. Springer, 2016, pp. 36–42

[10] M. Keuper, S. Tang, Y. Zhongjie, B. Andres, T. Brox,and B. Schiele, "A multi-cut formulation for joint segmentation and tracking of multiple objects," arXiv preprint arXiv:1607.06317, 2016.

[11] B. Lee, E. Erdenee, S. Jin, M. Y. Nam, Y. G. Jung, and P. K. Rhee, "Multi-class multi-object tracking using changing point detection," in ECCV. Springer, 2016,pp. 68–83.

[12] W. Choi, "Near-online multi-target tracking with aggregated local flow descriptor," in ICCV, 2015, pp. 3029–3037.