Sports Object Recognition and Tracking*

Shreyansh Sharma[1 [S3772241]]

Leiden University, Netherlands
S3772241@vuw.leidenuniv.nl

Abstract. Sports object recognition and tracking is a challenging task in computer vision. This paper presents and compares different methods for object recognition and tracking in sports videos. Pretrained models are used to reduce the training time and improve the accuracy of the models. Along with the comparison, a framework is proposed to evaluate the performance of the models. The framework can be accessed at https://github.com/shreyansh05s/SPORT.

Keywords: DETR · DeepSort · SportsMot.

1 Introduction

In recent years, the field of computer vision has seen a lot of progress in object recognition and tracking. Improved object recognition and tracking can be used in many applications such as autonomous driving, surveillance, and sports analysis. In sports, object recognition and tracking can be used to analyze the performance of the players and the team. It can also be used to analyze the performance of the referee and the umpire. The data generated from object recognition and tracking can be used for downstream tasks such as player tracking, player action recognition, and player pose estimation. This makes it vital to improve the accuracy of object recognition and tracking in sports.

Object recognition and tracking in sports is a challenging task due to the fast movement of the players and the ball. The players and the ball can be occluded by other players or the referee. The players can also be occluded by the audience. This makes it difficult to track the players and the ball.

In this paper, we present and compare different methods for object recognition and tracking in sports videos. The focus lies on utilizing pre-trained models to reduce the training time and improve the accuracy of the model. Along with the comparison, a framework is proposed to allow faster setup and evaluation of the models.

For the comparison, we use the SportsMot dataset [2]. The SportsMot dataset is a large-scale dataset for multi-object tracking in sports. It contains videos of different sports such as basketball, football, and volleyball. This makes it suitable for comparing different methods for object recognition and tracking in sports. MOT-16 and MOT-17 datasets [8] are also popular datasets for multi-object tracking.

* Supported by organization Leiden University.

2 Related Work

Usually, the task of object recognition and tracking is tackled in two steps. First, the objects are detected in each frame of the video. Second, the detected objects are tracked across the frames of the video. This is also commonly known as the tracking-by-detection approach. One such example for tracking-by-detection approach is the Deep SORT algorithm [14].

In recent years, there has also been a lot of progress in end-to-end object recognition and tracking. But for the purpose of this paper, we will focus on the tracking-by-detection approach.

2.1 Object Detection

Yolo [10] made a breakthrough in object detection by introducing a single neural network for object detection. Before Yolo, object detection was done using two neural networks. The first neural network was used to generate the region proposals. The second neural network was used to classify the region proposals. This made the object detection pipeline slow and inefficient.

After Yolo, many other single neural network object detection models were introduced. One such example is the DETR model [1]. The DETR model uses a transformer [11] to detect the objects in each frame of the video. The transformer is a neural network architecture that uses attention to process the input data. The DETR model uses a pre-trained ResNet-50 [3] as the backbone. The ResNet-50 is used to extract the features from the input image after which the extracted features are passed to the transformer. The transformer outputs the bounding boxes and the class labels of the detected objects.

2.2 Object Tracking

The Deep SORT algorithm uses a pre-trained object detection model to detect the objects in each frame of the video. The detected objects are then tracked across the frames of the video using the Kalman filter [4]. The Kalman filter is used to predict the position of the objects in the next frame of the video. The objects are then assigned to the predicted position based on the similarity between the predicted position and the detected position. The similarity is calculated using the Hungarian algorithm [5] for which features are extracted from the detected objects using a pre-trained model. This process is repeated for each frame of the video.

3 Dataset

SportsMot dataset [2] is a large-scale dataset for multi-object tracking in sports. The objective of the dataset is to detect and track the players exclusively. Both Train and Validation sets contain 45 videos, whereas Test set contains 150 videos.

3.1 Data Processing

The dataset is provided in the form of frames which are stored as images. The dataset needs to be loaded sequentially for evaluation and inference. This can be a bottleneck for the evaluation and inference of the models. But for training object detection models, the dataset can be loaded in parallel. For this an optimized data loader was implemented. With a higher number of workers and a larger batch size, the training time can be reduced significantly.

Apart from loading the images they also need to be processed before they can be used for training. Preprocessing includes resizing, normalization, augmentation and converting to tensors. These operations are specific for each model and are accessed from the pretrained model coonfigurations. Huggingface's transformers library [13] is used for this purpose.

4 Framework

The purpose of the framework is to allow faster setup and evaluation of the models on SportsMot dataset. The framework is implemented using PyTorch [9] and Huggingface's transformers library [13]. It is designed to be modular and extensible. This makes it easier to compare different models on the same dataset.

The framework consists of three main components. The first component is the data loader. The data loader is used to load the dataset optimally. The second component are the models. The models are used to detect and track the objects in the videos. Currently there are two types of models: object detection models and object tracking models. For Object Detection, the DETR model [1] and Conditional DETR model [7] are implemented which are pre-trained on the COCO dataset [6]. And for Object Tracking, the Deep SORT algorithm [14] is implemented. The third component is the evaluation and inference component. The evaluation component is used to evaluate the performance of the models. The inference is done on the validation or test set and the results are generated in the format required by the SportsMot dataset. The results can then be uploaded to the SportsMot website for evaluation.

The framework also includes a demo website for viewing the results. The demo website is implemented using Streamlit. This makes it easier to view the results of the models. An example of the demo website can be seen in Figure 1.

5 Challenges

The main challenge faced during the project was the training time and computational resources. The time required to train the models was very high. And adding the time needed to do hyperparameter tuning made it even worse. This made it difficult to achieve the desired results and come up with a new model.

Reprocucibility is also a challenge in deep learning. Trying to compare the results of different models is difficult as each needs a different environment and dataset loading pipeline. So most time was spent on creating a framework that can be used to compare different models on the same dataset.

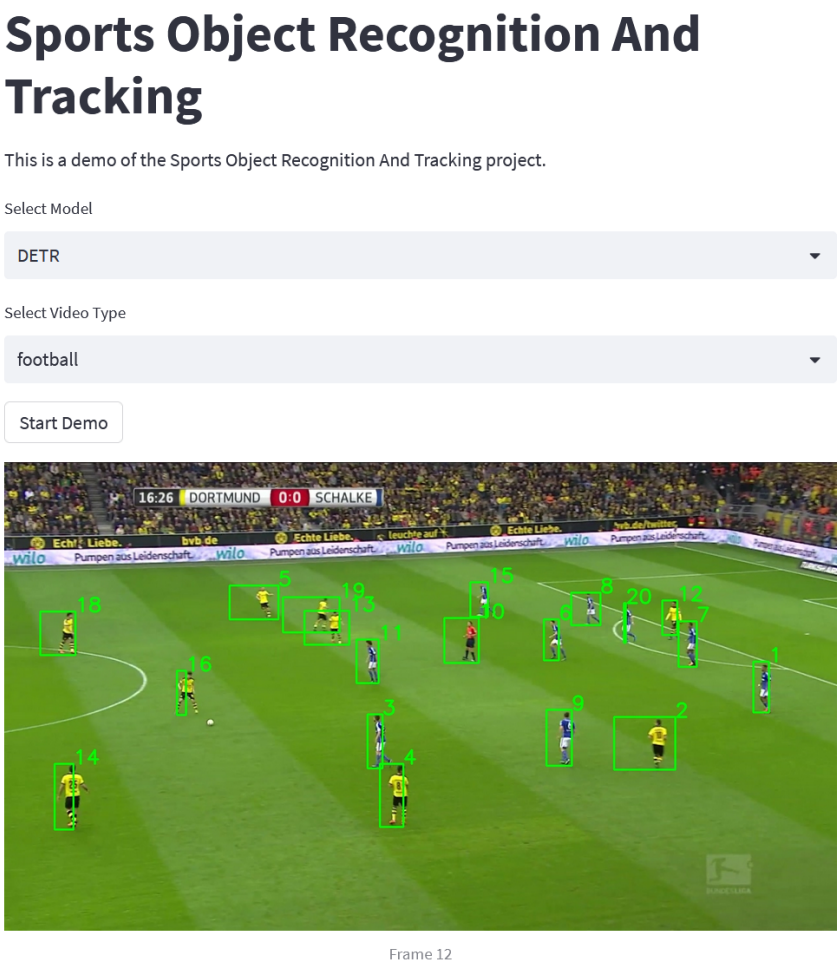

Fig. 1. Demo website for viewing the results

6 Conclusion

To conclude, the field of object detection and tracking has seen a lot of progress in the last few years. With the introduction of SportsMot dataset, it has even more practical applications it can be used for. Automatic annotation of sports videos can be used to generate highlights of the match. With even better hardware they can be used in real-time to generate live highlights of the match. This can be used to improve the experience of the viewers and also help the coaches in analyzing the performance of the players.

Accessing and setting up the dataset for benchmarking with a baseline model can be a challenging and time consuming task. The framework proposed in this paper can be used to fasten this process. With rapid improvements, this framework allows for easy integration of new models.

7 Future Work

For future work, the framework can be extended to include more models. Currently only two models are implemented for object detection and one for object tracking. More models can be added to the framework to allow for more comparisons. The framework can also be extended to include more datasets.

Additionally, only pretrained models were utlized for this project. Using the framework finetuning can be done in order to improve the performance of the models for the SportsMot dataset. This will acheive better results and can be used to publish results for competition. But this will require a lot of computational resources and time.

Another area of improvement is the hyperparameter tuning. Currently the hyperparameters are set to the default values. But with hyperparameter tuning the performance of the models can be improved. Either by using a grid search or by using a more advanced method like Bayesian optimization.

References

1. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. CoRR **abs/2005.12872** (2020), https://arxiv.org/abs/2005.12872
2. Cui, Y., Zeng, C., Zhao, X., Yang, Y., Wu, G., Wang, L.: Sportsmot: A large multi-object tracking dataset in multiple sports scenes. arXiv preprint arXiv:2304.05170 (2023)
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR **abs/1512.03385** (2015), http://arxiv.org/abs/1512.03385
4. Kalman, R.E.: A new approach to linear filtering and prediction problems. Transactions of the ASME–Journal of Basic Engineering **82**(Series D), 35–45 (1960)
5. Kuhn, H.W.: The hungarian method for the assignment problem. Naval Research Logistics Quarterly **2**(1-2), 83–97 (1955). https://doi.org/https://doi.org/10.1002/nav.3800020109, https://onlinelibrary.wiley.com/doi/abs/10.1002/nav.3800020109

6. Lin, T., Maire, M., Belongie, S.J., Bourdev, L.D., Girshick, R.B., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. CoRR **abs/1405.0312** (2014), http://arxiv.org/abs/1405.0312
7. Meng, D., Chen, X., Fan, Z., Zeng, G., Li, H., Yuan, Y., Sun, L., Wang, J.: Conditional DETR for fast training convergence. CoRR **abs/2108.06152** (2021), https://arxiv.org/abs/2108.06152
8. Milan, A., Leal-Taixe, L., Reid, I., Roth, S., Schindler, K.: Mot16: A benchmark for multi-object tracking (2016)
9. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E.Z., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. CoRR **abs/1912.01703** (2019), http://arxiv.org/abs/1912.01703
10. Redmon, J., Divvala, S.K., Girshick, R.B., Farhadi, A.: You only look once: Unified, real-time object detection. CoRR **abs/1506.02640** (2015), http://arxiv.org/abs/1506.02640
11. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. CoRR **abs/1706.03762** (2017), http://arxiv.org/abs/1706.03762
12. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. CoRR **abs/1703.07402** (2017), http://arxiv.org/abs/1703.07402
13. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 38–45. Association for Computational Linguistics, Online (Oct 2020). https://doi.org/10.18653/v1/2020.emnlp-demos.6, https://aclanthology.org/2020.emnlp-demos.6
14. Xu, Y., Wang, J.: A unified neural network for object detection, multiple object tracking and vehicle re-identification. CoRR **abs/1907.03465** (2019), http://arxiv.org/abs/1907.03465