The Application of Object Detection in Sports

Dylan Runjia Li
Tsinghua International School, Zhongguancun North St. Haidian District, Beijing, China

Keywords: Object Detection, Sports, Computer Vision.

Abstract: Object detection, which aims to both detect where multiple objects are in an image and classify them, is a significant task in computer vision and graphics due to its broad applications in robotics, autonomous driving, and sports analytics. In sports, it is increasingly used to track player movements and develop strategies for team-based games. Despite the fast development of object detection, a comprehensive study of its application in sports remains lacking. To address this, this paper analyzes state-of-the-art methods such as YOLO, SSD, RetinaNet, and R-CNN variants, highlighting their strengths and limitations and improvements over previous methods. Furthermore, the application of these methods in sports such as soccer, basketball, and baseball are introduced along with the many challenges of implementing object detection in sports such as occlusion and fast motion. These applications that have been developed with object detection solve a variety of issues in these sports, mostly dealing with the collection of different types of sports data. Advancements like these can greatly increase the accessibility of sports data by reducing the cost required to collect it, opening up its many use cases such as aid in coaching and strategy to a wider audience beyond major professional leagues.

1 INTRODUCTION

Artificial intelligence and machine learning have grown tremendously in recent years, revolutionizing the field of computer vision. Nowadays, machine learning is not just applied to classify what an image is through image classification, but also to detect and label multiple objects within one image. This detection and labelling ability, known as object detection, represents an improvement in how machines understand and interact with the world.

This task of object detection is often regarded as a subtask of image classification, but its ability to identify multiple objects and their positions within images makes it much more valuable and powerful for use in society. Unlike image classification, object detection's ability to locate and identify multiple targets has found applications for it across diverse fields, ranging from security systems to autonomous vehicles. Particularly, object detection has emerged as a transformative technology in sports, enabling real-time player tracking, ball detection, etc. This paper aims to analyze mainstream methods and introduce specific applications of object detection in various sports.

Current state-of-the-art methods for object detection can be categorized into two main types: one-stage methods and two-stage methods. One-stage methods include models such as YOLO (You Look Only Once) (Redmon et al., 2016), SSD (Single Shot Detector) (Liu et al., 2016), and RetinaNet (Lin et al., 2018). YOLO divides an image into a grid and predicts bounding boxes and class probabilities for objects in each grid cell simultaneously. SSD uses a neural network that processes feature maps to predict object locations and class probabilities across various scales and aspect ratios in a single pass, improving performance on smaller objects. RetinaNet similarly uses anchor boxes of different scales and aspect ratios, but it also uses the focal loss function to address class imbalance. These one-stage models prioritize inference speed and are often used in real-time applications. In contrast, two-stage methods prioritize detection accuracy and are often slower than one-stage methods. Two-stage methods include R-CNN (Region-based Convolutional Neural Network) (Girshick et al., 2014) and its evolutions like Fast R-CNN (Girshick, 2015) and Faster R-CNN (Ren et al., 2016). R-CNN, one of the first CNN object detection models, uses multiple separate stages in its detection process. It starts with a selective search algorithm to generate region proposals that are evaluated by a CNN for features, and then a support vector machine classifies the features. The Fast R-

Li, D. R.
The Application of Object Detection in Sports.
DOI: 10.5220/0013678300004670
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 2nd International Conference on Data Science and Engineering (ICDSE 2025)*, pages 72-76
ISBN: 978-989-758-765-8
Proceedings Copyright © 2025 by SCITEPRESS – Science and Technology Publications, Lda.

CNN iteration on R-CNN uses a single CNN to process all the region proposals at once, creating and then classifying a shared feature map. Finally, the Faster R-CNN improvement replaces the external region proposal generation with an integrated convolutional region proposal network, allowing end-to-end training for better performance.

The distinct characteristics of each detection algorithm naturally lead themselves to different sports applications. By detecting the positions of players and other objects of interest like balls, object detection systems can be used to analyze player performance and also team strategies in sports such as soccer and basketball. For example, heatmaps of a soccer player's movement and shot positions can created by detecting and tracking the player on the field and mapping their locations. This can be used by opposing teams to better learn the player's play style and the best way to defend against them. Combined with data for other players on the team, strategies, formations, and plays can also be detected by a computer, aiding a coach in preparing for upcoming matchups in sports like soccer, basketball, and American football. Additionally, object detection data can also be used for purposes in the sports entertainment industry with live television annotations. Strategy analyses can be shown on live television broadcasts automatically, and highlights can also be automatically presented by detecting key events like goals and successful plays.

However, these applications can run into challenges such as occlusion, fast motion, complex backgrounds, small objects, multi-object tracking, and diverse appearances when implemented. With occlusion, players or other objects can often be blocked from the view of the camera, which could lead the object detector to lose track of it, similar to how a fast movement from a player or a ball could. Additionally, complex backgrounds such as crowds and stadiums can confuse models, along with small objects like balls. Keeping track of multiple objects while maintaining their identity can also be challenging, and different appearances that objects can take such as players in different poses also pose a challenge for object detectors to recognize them.

In summary, object detection is a very powerful tool with many capabilities that allow computers to perform ever more complicated tasks that are utilized in many fields. These capabilities are achieved through different approaches such as YOLO, SSD, and R-CNN, each with their own advantages and disadvantages.

2 METHODS

2.1 R-CNN

The introduction of the original R-CNN (Region-based Convolutional Neural Network) (Girshick et al., 2014) was a very important landmark for the field of computer vision and the object detection task. It demonstrated the potential of CNNs for achieving high performance in object detection. R-CNN was one of the first models to effectively use deep learning on object detection, achieving performance that was greatly improved over previous methods. Essentially, R-CNN functions by creating region proposals on an image, then computing features for each region using a CNN, and finally classifying the features.

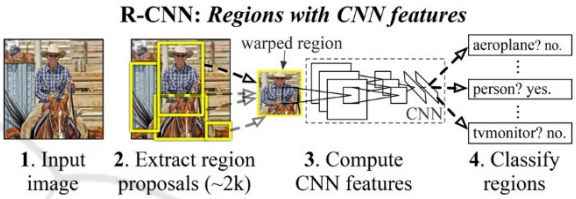

Figure 1: The stages of the R-CNN model (Girshick et al., 2014).

Figure 1 illustrates the main steps of R-CNN. First, it calculates around 2000 region proposals that can be generated using different external methods, but the original paper used selective search in its testing. Next, each region is warped into a 227 x 227 RGB image with 16 pixels of padding around the region proposal. This image is fed into the Caffe implementation of AlexNet (Krizhevsky et al., 2012) to compute a 4096-dimensional feature vector of the region. Finally, this feature vector is classified by a support vector machine.

2.2 YOLO

YOLO (You Look Only Once) (Redmon et al., 2016) was one of the first object detection models that could truly run in real-time. The big difference between YOLO and previous methods like R-CNN was that it unified the entire object detection process from bounding box detection to classification all into one step performed by a single convolutional neural network. The major advantage of performing object detection in one step is the massively improved inference time of only one step of processing. However, in its state in 2014, this method lacked the accuracy that slower models such as R-CNN had.

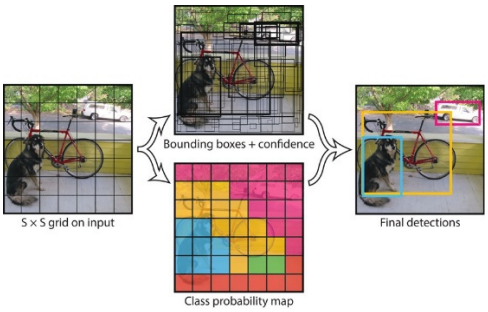

Figure 2: The process of the YOLO model (Redmon et al., 2016).

With YOLO, as illustrated by Figure 2, in contrast to R-CNN, there is not a separate step to generate regions before a main classification step. Instead, YOLOv1 resizes the input image to a 448 x 488 resolution and then divides it into 7 x 7 grid. Afterwards, each grid cell predicts 2 bounding boxes with confidence scores and one class probability score for the cell. These grid-based predictions are then combined for a final output of classified bounding boxes. YOLO is fast and efficient, but can have worse performance than two-stage methods and also have poor localization, especially for smaller objects that only occupy a small region of a grid cell. Additionally, detecting overlapping objects for YOLO can be challenging as it only predicts a limited number of bounding boxes for each cell.

2.3 SSD

SSD (Single Shot MultiBox Detector) (Liu et al., 2016) was another attempt to achieve the same goal as YOLO: real-time object detection. Conversely to YOLO, SSD uses an external CNN such as VGG16 to process the image and generate multiscale feature maps. These feature maps are at multiple scales, in contrast to YOLOv1's single scale feature maps, allowing SSD to perform much better on smaller objects. Figure 3 shows an example of these multiscale feature maps in play.

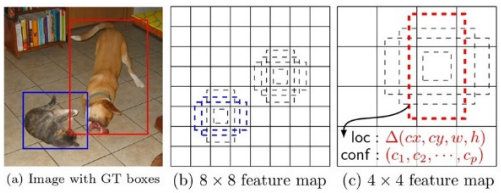

Figure 3: Multiscale feature maps in SSD (Liu et al., 2016).

Similarly to YOLO, the input image is first resized to a standard resolution (either 300 x 300 or 512 x 512) before processing. Another difference between SSD and YOLO is that SSD makes use of predefined anchor boxes of different shapes, sizes, and aspect ratios that originate from the spatial locations in the feature maps. These anchor boxes are adjusted to fit the objects being detected in order to form the final output. This feature was also later used in YOLOv2 (Redmon & Farhadi, 2016) as they allowed SSD to detect objects of different shapes and sizes with better performance. With SSD, real-time object detection took another step forwards, resolving many of the limitations of YOLO.

2.4 RetinaNet

Like YOLO and SSD, RetinaNet (Lin et al., 2018) aims to detect objects in real-time. As SSD did, RetinaNet makes use of an external CNN to generate features at multiple scales. However, RetinaNet uses a Feature Pyramid Network (FPN) on top of ResNet (He et al., 2015) instead of VGG16 as SSD did. This FPN allows RetinaNet to outperform SSD on objects of all sizes, but adds complexity and computation. Furthermore, RetinaNet also makes use of anchor boxes at the locations in the feature maps, following in SSD's footsteps. Finally, RetinaNet uses the focal loss function to handle class imbalance, which, like the FPN, adds complexity to the model. While this added complexity improves performance when compared to SSD, it also increases computational requirements and slows down the model, making it less viable for time-critical applications. Additionally, the simplicity of SSD makes it easier to train and deploy as well. The structure of RetinaNet is shown in Figure 4.

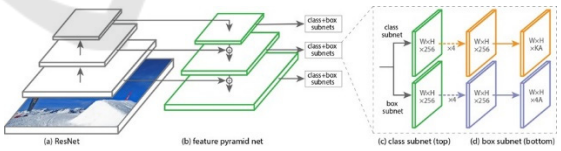

Figure 4: RetinaNet's ResNet-FPN structure (Lin et al., 2018).

3 APPLICATIONS

3.1 Soccer

In soccer and similar sports such as basketball, matches are often annotated for data in the form of recorded events such as passes, shots, and fouls with spatial and temporal information. This data has become very useful for many parties involved in

sports, including coaches, players, broadcasters, and fans. For example, the data collected from a game can be used by coaches to evaluate the performance and strategies of both their own teams and opposing teams. More specifically, data can reveal to a soccer coach that an upcoming opponent favours attacking from the right side, thus allowing the coach to make a strategic adjustment to have their team focus more on that right side on defense. Similarly, the coach could observe that their own team favours one side on either offense or defense, allowing them to adjust and balance out their strategy.

However, the current reliance on manual data collection by human operators makes the process costly and time-consuming. To address this, PassNet (Sorano et al., 2020), an AI model that combines YOLOv3 with image classification and LSTM to automatically annotate passes from a soccer video stream, was introduced. PassNet is made up of three separate modules, a feature extraction module with a CNN, an object detection module that makes use of YOLOv3 (shown in Figure 5), and a sequence classification module using LSTM. First, the feature extraction module and the object detection module are provided with the input sequence of frames from the video stream to generate features such as the positions of the ball and the players closest to it. Then, the sequence classification module takes these features as input and outputs a sequence of events. While not ready for mainstream use yet, PassNet is a great first step towards reducing the cost of sports analysis and making it more accessible to minor and non-professional sports leagues.

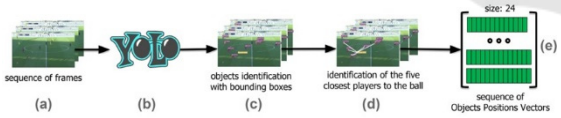

Figure 5: The architecture of PassNet's object detection module, including YOLOv3 at (b) (Sorano et al., 2020).

3.2 Basketball

A similar sport to soccer in terms of structure, basketball also lends itself to the application of object detection for the purpose of tracking players and the game for data analysis. With tracking data for basketball, coaches can make similar adjustments to strategy as soccer coaches would, reading different defensive and offensive strategies from automatically generated data. Additionally, analytics data could also be used for other purposes too. However, the major difference between basketball and soccer when viewed with object detection in mind is that basketball is played on a much smaller playing area. This means that basketball players more often occlude one another and the ball from the view of the camera, as soccer players are much more spread out on a large soccer field compared to basketball players on a small basketball court.

Basketball-SORT (Hu et al., 2024) attempts to remedy this issue with a new motion-based approach to multi-object tracking that performs significantly better than previous methods. The main issue that this model solves is the problem of keeping track of different objects even when they are occluded and not getting them confused with other objects such as those that might be occluding them. Basketball-SORT achieves this by tracking the motions of objects from previous frames and projecting future positions to aid YOLOv8 in keeping track of 10 unique IDs for ten players on a basketball court. This is done with the BGR (Basketball Game Restriction) and RLLI (Reacquire Long-Lost Player IDs) modules in the algorithm. With more reliable object detection models such as Basketball-SORT, object detection can provide the sport of basketball the same benefits as object detection systems like PassNet could with soccer.

3.3 Baseball

As baseball is a vastly different sport when compared to soccer and basketball, there are also vastly different opportunities for object detection to be applied in baseball. One obvious application for object detection is on pitching, as that is the most important aspect of the sport, so much so that half of MLB team rosters can be just players dedicated to pitching. A baseball pitcher has an arsenal of different pitches that they can throw to hitters, and being able to distinguish between pitches to hit them is a baseball batter's hardest task, the most important of the sport. The trajectory of a baseball in a pitch is affected by many factors, most importantly the ball's speed and spin rate and the Magnus force that the surrounding air applies on it due to its speed and spin rate.

While professional leagues such as the MLB already have systems in place that measures the speed the spin rate of a pitch, these systems are very complex and expensive, involving multiple pieces of advanced hardware. Wen et al. (2022) proposes a system that uses YOLOv3-tiny (Adarsh et al., 2020) to track a baseball from television footage and uses this tracked data to calculate both its speed and spin rate using a model of the Magnus force developed in the same paper. Using this combination of AI object detection and physics modelling, this paper was able

to achieve results comparable to much more expensive and complex existing methods that are used in television broadcasts such as those of the MLB. These results are a major step towards making essential pitching data much more accessible to coaches, players, and fans alike.

4 CONCLUSIONS

In this paper, various mainstream methods of performing object detection using deep learning models were analyzed and applications of these methods in the sports industry were introduced.

The methods analyzed were YOLO, SSD, RetinaNet, and R-CNN (along with Fast R-CNN and Faster R-CNN). These methods can be placed into two categories, one-stage and two-stage, with YOLO, SSD, and RetinaNet falling into the one-stage category and R-CNN and its evolutions being categorized as two-stage. While one-stage models are much faster, they sacrifice accuracy and precision when compared to two-stage models that run slower but have a separate step to generate bounding box proposals.

With the fast nature of most sports, one-stage methods and, in particular, YOLO, are used the most in sports applications. These applications span a variety sports from soccer to basketball to baseball and fulfill many use cases. The greatest value in the application of AI object detection in sports is the possible reduction of cost over current methods of either manual labour or expensive and complex equipment and hardware. Object detection models are much cheaper and simpler to deploy, making sports data and analysis much more accessible and opening up the opportunity for a much wider audience to benefit from modern sports data.

REFERENCES

Adarsh, P., Rathi, P., Kumar, M. 2020. YOLO v3-Tiny: Object Detection and Recognition Using One Stage Improved Model. In *International Conference on Advanced Computing and Communication Systems*.

Girshick, R. 2015. Fast R-CNN. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Girshick, R., Donahue, J., Darrell, T., Malik, J. 2014. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

He, K., Zhang, X., Ren, S., Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Hu, Q., Scott, A., Yeung, C., Fujii, K. 2024. Basketball-SORT: An Association Method for Complex Multi-Object Occlusion Problems in Basketball Multi-Object Tracking. In *Multimedia Tools and Applications*.

Krizhevsky, A., Sutskever, I., Hinton, G. E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P. 2020. Focal Loss for Dense Object Detection. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A. C. 2016. SSD: Single Shot MultiBox Detector. In *Proceedings of the European Conference on Computer Vision*.

Redmon, J., Farhadi, A. 2017. YOLO9000: Better, Faster, Stronger. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A. 2016. You Only Look Once: Unified, Real-Time Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Ren, S., He, K., Girshick, R., Sun, J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems*.

Sorano, D., Carrara, F., Cintia, P., Falchi, F., Pappalardo, L. 2021. Automatic Pass Annotation from Soccer Video Streams Based on Object Detection and LSTM. In *Machine Learning and Knowledge Discovery in Databases.*

Wen, B.-J., Chang, C.-R., Lan, C.-W., Zheng, Y.-C. 2022. Magnus-Forces Analysis of Pitched-Baseball Trajectories Using YOLOv3-Tiny Deep Learning Algorithm. In *Applied Sciences*.