Computer Vision in Sports

Jack T. Mahoney
jmahoney0419@gmail.com
Division of Science and Mathematics
University of Minnesota, Morris
Morris, Minnesota, USA

Abstract

The work presented in this paper is an overview of ball tracking in sports. The specific sport that will be primarily used is the sport of basketball. This paper looks at the components of a camera and the impact that those components have on determining the ball position. The paper also covers different sports and how ball tracking technology is used in each of them. The paper also introduces the process of finding the location of balls from images or frames of a camera.

1 Introduction

The sports of tennis, volleyball, and basketball involve a lot of movement and are fast paced games. Tennis, volleyball, and basketball also have a viewership together of over four billion worldwide. Tennis, volleyball, and basketball are all interesting sports and can be hard to understand if the people watching are not familiar with them. Automated ball tracking can help newcomers (and fans) better understand these sports. This paper will first cover tennis when it comes to ball tracking. Then transition into the sport of volleyball, which is similar to that of tennis in that it involves similar lines and has a similar play structure to tennis. Then transition into basketball and the complexities that introduces.

2 Background

The problem at hand is that the sports of tennis, volleyball, and basketball are all extremely fast playing sports which complicates both fan understanding and the ability to make the right call by referees. The use of computer vision can inform referee decisions, help automate the tracking of statistics, and provide information that makes the game easier to understand. These are all especially important to the games themselves and to the people watching them. The more someone understands the games they watch, the better and sometimes more accurate games lead to more people watching. There is constant research going on into what is the right way to use it and what are the drawbacks of using computer vision in specific sports. For example, the data competition known as Deep Sport Radar (see Section 8.1 and 2.1).

2.1 Kaggle

Kaggle is a platform for data scientists. The team at Kaggle held a competition that ran from 2020 to 2024 where the competition had different sport-related challenges. This paper focuses on the challenge of ball identification in a 3-dimensional scene using information gathered from a camera of known location and orientation. Kaggle's competition compiled the Deep Sport Radar data set. Details of the winning entries were published in Deep Sport Radar Version 1 [5]. One of their challenges involved identifying which pixels in a camera image belong to a ball. Once identified, a technique known as 3D Ball Localization [4] allow the center of the ball to be calculated relative to the court. This calculation requires knowing the details about the camera's internal structure, it's orientation, and location. This paper describes the details and complications of performing this calculation.

3 Camera

For this calculation, a camera is viewed as a lens which bends light into a single focused point (the optical center, See Figure 1) and then captures scene information on a sensor (Figure 1).

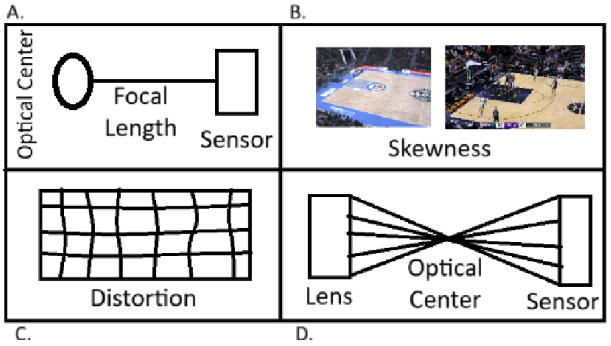

Figure 1. Image of Camera Components [5]

The focal length is the distance between the lens of the camera and the camera sensor. This can be different depending on the type of camera that is used. This is also taken into consideration when figuring out the location of the ball (Figure 1).

Figure 1 shows different lens distortions. Distortion is the process of light going into the camera and making objects or people look rounder or flatter depending on the lens.

3.1 Arena

An arena is where teams play games. Figure 3 contains examples of arenas from three different sports.

4 Coordinate Systems

Coordinate systems allow a set of values to be associated with positions in space. One common coordinate system uses three perpendicular lines known as axes locate a position in space. Values known as coordinates indicate how far to travel along those axes. Traditionally, these three values are known as coordinates and represented as x, y, and z.

Different coordinates systems produce different mappings between values and locations. We'll consider three. Two of them are two-dimensional and are different ways of representing locations in the "real world". The third is three-dimensional and represents locations on the sensor, and hence on an image (see Section 4.3 and Figure 2). Homogeneous coordinates are a representation of a two-dimensional coordinate system that uses 3 values (see Section 4.4).

Dimension is the number of independent values needed in a coordinate system to represent a location. The origin is the point in the coordinate system at which the axes intersect.

4.1 Camera Coordinate System

The camera coordinate system functions as the frame of reference established by the camera's spatial position. It's origin is defined by the camera's optical center, designated as the null vector $(0, 0, 0)$. This system is essential for quantitatively determining the three-dimensional spatial relationship, specifically the distance, between an object (e.g., a basketball) and the camera (see Figure 2). One axis known as the optical axis extends outward from the optical center and intersects the lens perpendicularly (see Figure 2).

4.2 Real World Coordinate System

The real-world coordinate system (RWCS) is a fixed, three-dimensional frame of reference used to define the absolute spatial location of objects. The selection of the origin for this system is context-dependent, typically corresponding to a defined geometric center within the physical environment. For sports applications, the origin is commonly set at the center of the playing area, such as the center circle of a basketball court or the midpoint of a tennis or volleyball court. This system is crucial for accurately determining the absolute position of the object in physical space (see Figure 2).

4.3 Switching Between

Determining the absolute three-dimensional position of an object, such as a basketball, necessitates the use of both the camera coordinate system and the real-world coordinate system. Consequently, coordinate system transformation is required to map data between these two frames of reference. This essential conversion is mathematically governed by the camera matrix, which is detailed in the Section 8.4.1 discussing the ball center and projection models (see Figure 2).

4.4 Homogeneous Coordinate System

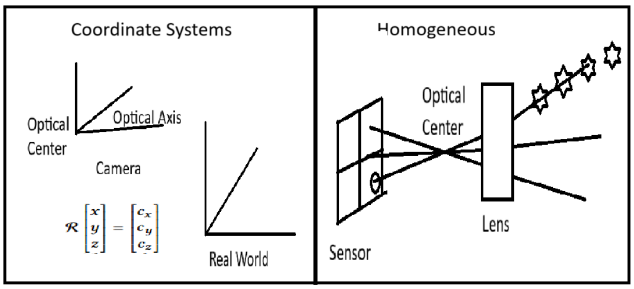

Figure 2. Coordinate Systems and Homogeneous

Homogeneous coordinates represent the projection of the optical axis through the camera lens. When looking at Figure 2 the stars on the left image represent homogeneous coordinates. Also the circle on the camera sensor insicates where all of those stars point to. This is why since all 3D points along a single optical axis map to the same 2D pixel coordinate, a single camera system suffers from depth ambiguity. As a result, this projection loses the ability to determine the object's distance (depth) from the camera shown in Figure 2.

5 Computer Vision

Computer vision is a field dedicated to enabling computers to extract meaningful information from digital media, such as images, video streams, or still frames. In this context, the process involves analyzing the pixel data within the input media to detect and track specific objects, such as a basketball, thereby determining its position and identity.

6 Tennis

Tennis presents a unique challenge for computer vision due to the critical need for precise call accuracy. Since a point is scored if the ball is not returned or lands out of bounds, accurate determination of the ball's impact location is paramount. This demand for high accuracy makes computer vision an essential technology for ensuring fairness and correctness in officiating (see Figure 3).

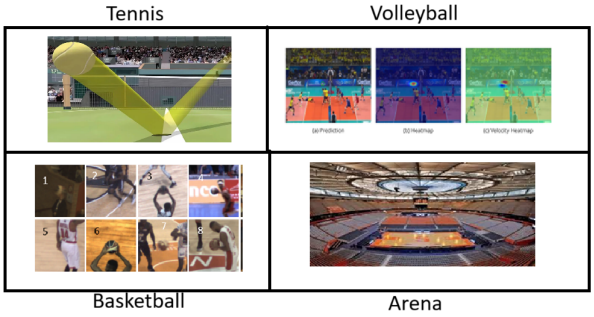

Figure 3. Images of Different Sports and an Arena

6.1 The Positives

The primary benefit of integrating these systems is the enhancement of game accuracy. Artificial Intelligence (AI) and computer vision collectively achieve this precision by processing visual data at a level of detail and speed inaccessible to human perception. Specifically, in tennis, this involves fusing data from multiple synchronized cameras positioned around the court, which we will not get into in this paper. This high-level accuracy can directly influence match outcomes by eliminating some human errors [6].

6.2 Drawbacks

Despite the technical benefits, the accuracy of computer vision systems in tennis remains a subject of debate [6]. Furthermore, a significant drawback, particularly noted at major tournaments like Wimbledon, is the impact on spectator experience. The system's immediate, exclusive use diminishes the suspense that was intrinsic to the older, challenge-based approach. Under the previous system, a player-initiated challenge would trigger a review process, culminating in a visual confirmation or rejection of the call displayed publicly on the stadium screen, a moment that actively engaged the audience [6].

6.3 3D Imaging in Tennis

Obtaining the 3D image of the tennis court is a little easier than that of basketball which is mentioned in this paper. The way the 3-dimensional space of the court is obtained is by taking several images of the court or frames of the court in the case of live video and then looking for the corners of the court. An example of this is shown in Figure 4.

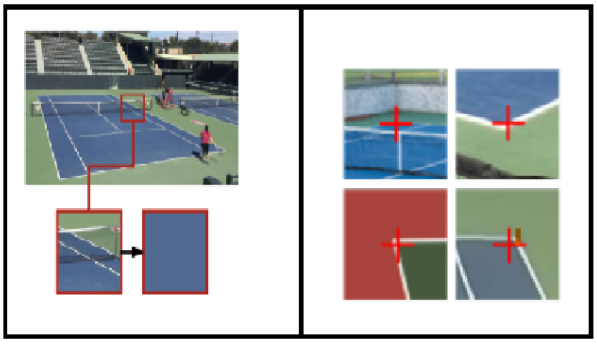

Figure 4. Tennis Court Edges [3] and In and Out of Bounds

After identifying the corner points, the computer uses these feature values to calculate distances between them [3], then connects the points to form a 3D wireframe of the court. The positions of internal features, such as the net, are then obtained from their fixed distances relative to the court's center point [3].

A special simplifying case in court analysis arises when the playable area and the out-of-bounds region are differentiated by color. [Image showing a tennis court with a contrasting color between the areas inbound and outbound is shown in Figure 4] This color contrast significantly aids the vision system's ability to segment the image, making it easier to accurately distinguish between valid and invalid playing areas [3]. This demonstrates that the reconstruction and tracking models can be simplified in certain contexts based on inherent visual features like color.

7 Volleyball

Volleyball is another interesting sport. Volleyball adds another level of complexity to the model because there is more than one player on each side of the net and there are more lines to consider. The ball also moves in several different directions at a high speed. This means there are more factors for the model to consider when tracking the ball.

7.1 Player Positioning

In the case of volleyball, the model generated in 3D Ball Localization would need to keep track of the player's position on the court eventually because of the ball being hidden by player sometimes which we will not cover this calculation. Doing this involves looking at the court and then using the video to show where the volleyball is in terms of a virtual arena or court. This is done several times to calculate all the positions of the players that are on the court. This gives a full image of what is going on at any point in the game [2]. This concept is also crucial to ball identification because players can obstruct the player there is a calculation to handle this but we are not going to get into this calculation.

7.2 Ball Prediction

In volleyball, player positioning is leveraged to create a predictive model for shot trajectory which we will not cover the model involved but it could use the calculation to help with prediction. This model is trained using data from numerous games to anticipate the ball's likely destination. Prediction is based on analyzing the players' momentum during the play and their apparent intention or movement vectors [2].

The ball prediction software operates by first processing an input image (or frame) and layering it with a heatmap that indicates the ball's current location [2]. This heatmap is then combined with the live video feed. By synthesizing these two inputs, the model predicts the ball's subsequent position and path, as visually represented in Figure 3 [2].

7.3 Challenge

One challenge of volleyball is that there are not as many camera angles available to use. This can make some of the prediction harder because the camera might lose the ball more often than if multiple different cameras were used. The

lesser amount of camera angles also is a challenge because of all of the players in such a condensed area.

8 Basketball

Basketball is the most complicated of the three. It involves many different actions by many different players. The movements of those players and the way that the player interacts with other players is important to. There are also several different lines on the court that are important to the process. This all makes basketball a much more challenging sport to model, which is why the challenge is based on this sport.

8.1 Deep Sport Radar

This is a competition that is sponsored by Kaggle. The competition involves several challenges. The specific challenge that is discussed here is 3D ball localization in a calibrated scene [5]. The paper with the equations to calculate the balls position is 3D Ball Localization from A Single Calibrated Image [4]. Three-dimensional ball localization is the process of figuring out the precise location of an object. A calibrated image is taking camera positions and converting them into real world coordinates.

8.2 The Camera Setup

The camera setup is a crucial factor in the overall tracking process. The distance between the camera and the center of the court significantly influences the visual perception and scaling of the scene.

This variability directly affects the model's perception of spatial metrics, particularly the critical distance thresholds related to play validation, such as determining if a shot is a three-pointer or if a ball is out of bounds. For instance, the NBA three-point line distance ranges from 22 to 23.75 feet (the shot closest to the middle vs. the corner), and the model's calculation of this distance is sensitive to the camera's relative proximity to the basket. Therefore, the setup needs precise calibration to compensate for the perspective changes caused by arena-specific camera placements.

8.3 The Location of the Shot

In basketball, the three-point line defines the boundary for scoring three points. The precise distance of this line from the hoop is variable and depends on the specific level of play (e.g., high school, collegiate, professional) and league or in the case of college, the NCAA division you are in.

This geometric variability presents a critical challenge for computer vision models, as the system must be tuned to recognize and apply the correct location for each specific venue and league. Consequently, the development of robust tracking algorithms requires a comprehensive training dataset that accounts for diverse arena geometries. Examples of stadiums included in such training sets are Sportica in Gravelines, PdS Jean Weille de Gentilly in Nancy, Le Jeu de Paume in Blois, and PdS de Caen in Caen [5].

8.3.1 Helping with the Location. To simplify and precisely determine the exact locations of players or objects, the complex three-dimensional (3D) scene data is transformed into a simplified two-dimensional (2D) representation. This process begins by analyzing the broadcast image alongside the 3D court model to determine the players' coordinates (x, y, z). The z (depth) variable is then discarded by projecting the player's position onto the court plane, effectively marking their location with a reference point [1]. Subsequently, the 3D court model is virtually rotated around the z-axis to align the perspective with a top-down view of the court. Finally, the projected 2D coordinates are mapped onto a standardized $x - y$ grid, correlating them back to their respective positions in the original planar representation of the court [1].

Figure 5. Top down image of a basketball court

Figure 6. Shows an augmentation of the court in a 3D representation

8.4 Localizing the Ball

When finding the position of the ball in the 3D space of the arena we would use a Convolutional Neural Network (CNN).

This is taking a model and then run it for several iterations to find the best way to classify the objects. The camera matrix is what is used to calculate the the position of the camera. Which is shown in 8.4 [4].

8.4.1 Camera Matrix. $k = \begin{bmatrix} f * \mu_x & \gamma & u_x \\ 0 & f * \mu_y & u_y \\ 0 & 0 & 1 \end{bmatrix}$

The top left argument, $f * \mu_x$, in this matrix shows the focal length represented by f which is then multiplied by the scaling factor of μ_x which gives the scaling factor for the x coordinate. The top middle argument, γ, which is the relative skewness of the camera. Which skewness is when parallel lines appear to not be parallel because of the camera. The top right argument, u_x, helps figure out the x coordinate of the camera. The center argument, $f * \mu_y$, is the same as the focal length of top left argument the only difference is that it is the y coordinate instead of x and μ_y is the scaling factor for the y coordinate. The middle right, u_y, helps figure out the y coordinate of the camera. The bottom right argument, 1, helps to preserves the homogeneous nature of the matrix. Then send the result of the camera matrix which is denoted k to the ball center function [4].

The Figure 7, shows how this matrix works. The matrix takes real world coordinates and then turns them into camera coordinates for the rest of the model. The light from the real world goes into the lens. The lens then distorts the light to meet at a point that is the optical center. This center is where the camera coordinate system starts. After it passes through the optical center, it goes to the optical sensor which is the exact point in the camera in which the camera sees it. The point can also depend on the focal length of the camera which can be different depending on the type of camera that is being used in the arena.

8.4.2 Ball Center. $b^c = K^{-1} * \mathcal{R} \begin{bmatrix} b_x \\ b_y \\ 1 \end{bmatrix}$

This function is designed to calculate the center of the ball in real-world coordinates. K^{-1} is the inverse of the camera matrix. It converts homogeneous coordinates into camera coordinates. The second argument, $\mathcal{R}$, represents a function that applies distortion correction to compensate for effects introduced by the camera lens.

The distortion function, $\mathcal{R}$, is applied to a coordinate matrix containing the following values [4]:

- The ball's x-coordinate: b_x
- The ball's y-coordinate: b_y
- A normalized z-coordinate (Homogeneous coordinate): 1

8.4.3 Edges. $e^c_{\pm} = K^{-1} * \mathcal{R} \begin{bmatrix} b_x \\ b_y \pm \frac{d}{2} \\ 1 \end{bmatrix}$

The function first employs the inverse of the camera matrix, denoted as K^{-1}. The second argument, R, is a distortion function applied to the ball's coordinates. This distortion correction takes an input matrix similar to that used for calculating the ball center, with the key difference being an offset introduced by dividing the y-coordinate by 2 and then adding and subtracting this value from the original ball center coordinates [4].

Using the corrected ball center and the detected ball edges, the system then calculates the three-dimensional (3D) location of the ball in the real-world coordinate system (i.e., its position within the arena). This spatial localization is governed by the formula presented in 8.4.4 [4].

8.4.4 Ball Location. $b^o = R^T * \frac{\phi * b^c}{e^c_{+y} - e^c_{-y}} + c^o$

This function calculates the exact location of the ball. This function takes three arguments.The matrix R is a rotation matrix that captures the orientation of the camera. R^T is the transpose of that rotation matrix. It will align the orientations of the world coordinate system and the camera coordinate system. The second argument is a fraction. The numerator of the fraction is ϕ which is the true ball size and then multiplies that by the ball's center. The denominator of the fraction is the top of the ball denoted by e^c_{+y} and then subtracts the bottom edge of the ball denoted by e^c_{-y}. This helps identify the true ball size. Then add c^o the exact position of the camera from the court.

9 Image Quality

The image quality is an especially key factor to consider regardless of the sport. If the image quality is not good, then it could lead to misinterpreted models and potentially lead to key issues in the prediction. An example of this can be seen in Deep Sport Radar Version 1 [5]. One of the examples is highlighted in Figure 7.

Figure 7. Image Quality [5]

These images show different qualities of an image. Figure 7.1 shows an image where the basketball is barely visible because of the darkness of the image which can really affect the model prediction. It might miss this and take it as part of

the background. Figure 7.2 shows an image that does a good job of brightness but the basketball is hard to distinguish from the court which makes it hard to figure out which pixels correspond to the ball, and which correspond to the court. Figure 7.3 is a great image because of the minimal distractions in the background and has good quality and resolution. Figure 7.4 shows an example of how the ball could blend into the rotating billboard behind the player, which sometimes might be a problem, but not always because the image changes. Figure 7.5 shows a misidentified object because it thinks that the hand of that player is the ball. Figure 7.6 is another good image. Figure 7.7 is an okay image. It has good brightness, but the computer might have trouble identifying the difference between the orange and white on the ball from the orange and white on the court. Figure 7.8 another good example [5].

10 Comparing and Results

Comparing the three sports and their implementations, there are some similarities but many differences. Tennis has fewer players and uses computer vision only to decide whether a ball is in or out, not for player tracking. In both volleyball and basketball, computer vision is used similarly, but in volleyball it focuses more on predicting the ball's trajectory. In basketball, it tracks the ball and predicts the location of it.

In Table 1 and Table 2 they show the results of the three different models used in testing. The first model shown is "the baseline that consists in applying Hough-circle transformation on the heatmap" in the table it is shown by BallSeg + HCT [4]. The second model is using the model covered in this paper. In this method they used the ball diameter to help with the estimation of the ball position [4]. The last model used in testing is a "method applied on oracle detections instead of using a detector" [4].

	$TP_{[\%]}$	$MAE_{[px]}$	$MAE_{[m]}$	$MAE_{[\%]}$
BallSeg+HCT	47 ±7	4.9 ±0.8	6.3 ±1.0	28 ±5
BallSeg+CNN	47 ±7	1.6 ±0.1	2.3 ±0.2	10 ±0.9
Oracal+CNN	100 ±0	1.9 ±0.1	2.8 ±0.2	12 ±0.6

Table 1. DeepSport testset [4]

	$TP_{[\%]}$	$MAE_{[px]}$	$MAE_{[m]}$	$MAE_{[\%]}$
BallSeg+HCT	83 ±2	4.6 ±0.5	5.1 ±0.5	24 ±4
BallSeg+CNN	83 ±2	1.6 ±0.2	1.8 ±0.2	10 ±0.7
Oracal+CNN	100 ±0	1.5 ±0.1	1.7 ±0.1	10 ±0.5

Table 2. 3D Ball Localization testset [4]

The BallSeg + HCT model is the baseline. BallSeg + CNN is the new model, which estimates ball diameter to improve position estimation. Oracle + CNN is an alternative method using oracle detection instead of diameter estimation [4].

$TP_{[\%]}$ is the percentage of pixels correctly identified as ball. $MAE_{[px]}$ is the mean absolute error in pixels. $MAE_{[m]}$ is the mean absolute error in meters of the ball center. and $MAE_{[\%]}$ is the mean absolute error as a percentage of model accuracy [4].

In the Deep Sport Radar competition, diameter estimation improved position accuracy compared with the baseline, even though $TP_{[\%]}$ was identical. The new model achieved a substantially lower MAE, showing that using diameter improves performance. However, Oracle detection outperformed the other methods on metrics other than MAE on the DeepSport dataset [4].

The 3D Ball Localization paper highlight that "having an automated estimation of ball 3D localization from a single camera can be extremely valuable, even with an imprecision of two meters" [4]. The error in meter approximation is why when used in a real life application sports organizations use multiple different camera angles to decrease the chance of this error.

11 Conclusion

Contemporary sports competitions move too quickly for many spectators to fully follow and interpret. Computer vision can help audiences better understand events and can support on-field officiating decisions that directly affect game outcomes.

The examples discussed show several practical uses of computer vision in sports. Tennis uses it systematically for line calls. Similar approaches were used in volleyball and then adapted to basketball for boundary detection and goal or basket validation.

Acknowledgments

Thank you to Professor Dolan and Professor Machasova for being my advisors in this process. I would also like to thank my Alumni reviewer Tom Harren.

References

[1] Code In a Jiffy. 2025. Build an AI/ML NBA Basketball Analysis system with YOLO, OpenCV, and Python. https://www.youtube.com/watch?v=QqVahw9tBfw&t=1085s

[2] Vanyi Chao, Hoang Quoc Nguyen, Ankhzaya Jamsrandorj, Yin May Oo, Kyung-Ryoul Mun, Hyowon Park, Sangwon Park, and Jinwook Kim. 2024. Tracking the Blur: Accurate Ball Trajectory Detection in Broadcast Sports Videos. In *Proceedings of the 7th ACM International Workshop on Multimedia Content Analysis in Sports (MM '24)*. ACM, 41–49. https://doi.org/10.1145/3689061.3689075

[3] Megan Fazio, Kyle Fisher, and Tori Fujinami. 2018. Tennis Ball Tracking: 3D Trajectory Estimation using Smartphone Videos. https://api.semanticscholar.org/CorpusID:4507065

[4] Gabriel Van Zandycke and Christophe De Vleeschouwer. 2022. 3D Ball Localization From A Single Calibrated Image. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 3471–3479. https://doi.org/10.1109/CVPRW56347.2022.00391

[5] Gabriel Van Zandycke, Vladimir Somers, Maxime Istasse, Carlo Del Don, and Davide Zambrano. 2022. DeepSportradar-v1: Computer

Vision Dataset for Sports Understanding with High Quality Annotations. In *Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports (MM '22)*. ACM, 1–8. https://doi.org/10.1145/3552437.3555699

[6] Zoya Yasmine. 2025. Calling it Out: What Wimbledon Can Teach Us About AI Automation. https://cjai.co.uk/usersubpost/calling-it-out-what-wimbledon-can-teach-us-about-ai-automation/. [Accessed 24-10-2025].