FootyVision: Multi-Object Tracking, Localisation, and Augmentation of Players and Ball in Football Video

Peter Andrews
MediaFutures and t2i Lab,
University of Bergen
Bergen, Norway
peter.andrews@uib.no

Njål Borch
Schibsted
Oslo, Norway
njaal.borch@gmail.com

Morten Fjeld
MediaFutures and t2i Lab,
University of Bergen
Bergen, Norway
morten.fjeld@uib.no

Figure 1: FootyVision multi-object tracking and top-down localisation in viewpoints with limited visual information.

Abstract

Football video content analysis is a rapidly evolving field aiming to enrich the viewing experience of football matches. Current research often focuses on specific tasks like player and/or ball detection, tracking, and localisation in top-down views. Our study strives to integrate these efforts into a comprehensive Multi-Object Tracking (MOT) model capable of handling perspective transformations. Our framework, FootyVision, employs a YOLOv7 backbone trained on an extended player and ball dataset. The MOT module builds a gallery and assigns identities via the Hungarian algorithm based on feature embeddings, bounding box intersection over union, distance, and velocity. A novel component of our model is the perspective transformation module that leverages activation maps from the YOLOv7 backbone to compute homographies using lines, intersection points, and ellipses. This method effectively adapts to dynamic and uncalibrated video data, even in viewpoints with limited visual information. In terms of performance, FootyVision sets new benchmarks. The model achieves a mean average precision (mAP) of 95.7% and an F1-score of 95.5% in object detection. For MOT, it demonstrates robust capabilities, with an IDF1 score of approximately 93% on both ISSIA and SoccerNet datasets. For SoccerNet, it reaches a MOTA of 94.04% and shows competitive results for ISSIA. Additionally, FootyVision scores a HOTA(0) of 93.1% and an overall HOTA of 72.16% for the SoccerNet dataset. Our ablation study confirms the effectiveness of the selected tracking features and identifies key attributes for further improvement. While the model excels in maintaining track accuracy throughout the testing dataset, we recognise the potential to enhance spatial-location accuracy.

CCS Concepts

• **Computing methodologies** → **Tracking**; **Object detection**; *Image processing*.

Keywords

Multi-Object Tracking, Object Detection, Homography, Deep Learning, Computer Vision, Football Analytics, Sports Analytics

ACM Reference Format:
Peter Andrews, Njål Borch, and Morten Fjeld. 2024. FootyVision: Multi-Object Tracking, Localisation, and Augmentation of Players and Ball in

ICMIP '24, April 20–22, 2024, Osaka, Japan

ACM ISBN 979-8-4007-1616-4…$xx.xx
https://doi.org/XXXXXXX.XXXXXXX

Football Video. In *Proceedings of International Conference on Multimedia and Image Processing (ICMIP '24).* ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Football commands an immense global audience, exceeding five billion enthusiasts, with its fan base spanning Europe, Latin America, the Middle East, and Africa [11]. This number is increasing with growing enthusiasm for leagues that once received little attention. For example, the FIFA Women's World Cup has become evermore popular, seeing its viewership numbers climb from 1.12 billion in 2019 to a staggering 2 billion in 2023 [21]. As viewership grows, so does the demand for more in-depth analysis of strategic components and player performance monitoring.

Innovations in Computer Vision (CV), Deep Learning (DL), and Graphical Processing Units (GPUs) have significantly advanced sports video content analysis, enhancing object detection, tracking, and reidentification. Despite these technological strides, applying DL to sports videos remains challenging due to the fast-paced nature of team sports, where distortions and occlusions frequently obscure player and ball detection and tracking. Football, in particular, presents additional challenges with its rapid player movements, uniform appearance, and frequent occlusions, thus making the task of Multi-Object Tracking (MOT) especially complex. Moreover, the intricacy of DL models for discrete tasks often compromises their performance and real-time processing capabilities. Consequently, an integrated system that can accurately track football players and the ball in near real-time, directly from raw video feeds, is not yet available. This gap hinders the potential for real-time analysis and event detection, which is essential for coaches, players, and avid fans.

FootyVision aims to bridge the existing research gap by delivering a comprehensive all-in-one solution for tracking and localising players and the ball in a top-down perspective, as demonstrated in Figure 1. This figure exemplifies FootyVision's application, showcasing its ability to perform well under scenarios with limited visual information. FootyVision utilises a custom-trained YOLOv7 network to detect players and the ball while accommodating football broadcast footage's dynamic and fast-paced nature. The perspective transformation module repurposes activation maps from the YOLOv7 to localise players in a top-down view and augment visualisations in the image space. This innovative method allows for extracting geometric information relevant to computing homographies even in viewpoints that otherwise lack relevant information for the computations.

FootyVision's tracking algorithm aims to overcome the limitations of previous research by providing an in-depth analysis of tracking features suitable to track identities which are similar in appearance. Our comprehensive ablation study dissects the tracking algorithm to pinpoint the features critical for maintaining player identities over successive frames. FootyVision's analysis is further distinguished by its application of CLEAR-MOT and Higher Order Tracking Accuracy (HOTA) metrics. We reference the videos used, which are publicly available, establishing a robust baseline for future research endeavours. Our dual contribution of methodological innovation and detailed performance evaluation addresses current challenges and sets a new standard for tracking accuracy in football analytics. For player and ball object detection, FootyVision achieves a mean average precision (mAP) of 95.7% and an F1-score of 95.5%. Moreover, for tracking on the SoccerNet dataset, FootyVision attains a Multi-Object Tracking Accuracy (MOTA) of 94.04%, a HOTA of 72.16%, and HOTA(0) of 93.1%. Consequently, FootyVision establishes a competitive benchmark for subsequent research and contributes new state-of-the-art results for football MOT.

2 Related Work

This section explores previous work related to player and ball detection and tracking in football, the rectification of sports videos, and all-in-one models for football video processing.

2.1 Player and Ball Detection and Tracking in Football

Early work in this area consisted of detecting and tracking players only due to difficulties in tracking the ball caused by motion blur and shape distortions. Many researchers approached the detection of players via segmentation of the foreground from the background via colour histograms [2, 7, 20, 25, 39, 49]. Segmentation of the players would then be undertaken via either blob detection [2, 12] or a particle filter [7, 27]. Representing players as dense particle collections, which are shape-invariant, assists in mitigating track loss due to occlusion, as described in [7]. Another method to overcome occlusions was to track players through multiple camera viewpoints [12, 27, 49]. However, multi-viewpoint tracking is only as reliable as the feature descriptor for the region, and early efforts still suffered from densely populated regions [49]. Other methods of tracking involve graph-based methods, where blobs are represented as nodes and edges showing the distance between the blobs [12]. Liu et al. [25] implemented an unsupervised approach to tracking, where two scans were made over the video to learn colour histograms for masking the background and sampling players using a Boosted Cascade of Haar features. Baysal et al. [1] employed Histogram of Oriented Gradients (HOG) descriptors with a Support Vector Machine (SVM) for tracking. This method involved sampling dense particles on the field and analysing colour and motion for player localisation.

More recent developments in Convolution Neural Networks (CNNs) have allowed for learning feature maps at various scales, meaning smaller objects can be detected, enabling detection of both player and ball [13, 23, 28, 29]. You Only Look Once v3 (YOLOv3) [31] has provided state-of-the-art results for both player and ball detection in Naik et al. [28, 29]. Naik et al. [29] combined YOLOv3 with the Simple Online Realtime Tracking (SORT) [4] algorithm to achieve high-accuracy tracking, which coped well with partial occlusions. Garnier et al. [13] implemented a Single Shot MultiBox Detector (SSD) with feature embeddings and reinforcement learning to provide robust tracking. Linke et al.'s [24] exploration into the validity of TRACAB's optical tracking systems underscores the evolution of tracking accuracy, particularly highlighting the Gen5 model's superiority in providing reliable spatio-temporal data.

Other literature in football player tracking and analytics reflects on utility. Vidal-Codina et al.'s [44] study on automatic event detection and Sanford et al.'s [33] research on group activity detection

demonstrate the advanced use of tracking data. These works highlight the application of tracking systems for in-depth analysis of in-game events and team dynamics, respectively.

2.2 Rectification of Sports Images

Image rectification is the projection of images onto a common plane. Traditionally, the relationship between two image planes is mapped by an isometric transformation performed by the homography matrix H. To achieve this, at least four image correspondences must be matched. Hayet et al. [17, 18] mapped H for both image-to-template and image-to-image, where lines and ellipses were used to compute the homography. In Hayet et al. [18], search areas were minimised by re-projecting the last estimate into the current frame, leading to real-time inference. Dubrofsky et al. [10] extended the Direct Linear Transform (DLT) algorithm to accommodate for both points and lines, verifying the results on ice-hockey images. For ice hockey rink rectification, Gupta et al. [14] implemented Dubrofsky et al.'s extended DLT algorithm. Here, ellipses were detected to define new keypoints via intersections and polars, offering a more robust method of rectifying ice-hockey images.

Recent applications of sports pitch rectification are mainly based around DL solutions with multitask learning [30, 42], self-supervised learning [37] and spatial transformers [36]. These networks can learn more complex patterns within data but at the cost of processing power. Some of these works have aimed to provide an end-to-end solution for camera pose estimation, estimating the camera intrinsic $K[R|T]$ via homographies [6, 36]. CNNs, in particular, U-Net, have proved to be reliable feature extractors for lines and area-based segmenters in Citraro et al. [6] and Sha el al.[36], respectively. Nie et al. [30] estimated keypoints via an encoder network to compute H. In the scenario where there were not enough keypoints available to satisfy the computation of H, a two-channel dense feature regressor increased keypoints.

Other recent proposals have opted for computationally less costly methods of localising players in a top-down view. For instance, Stein et al. [39] preprocessed a panoramic view of the football pitch and mapped frames and homographies. While Scott el al. [34] aligned drone images with Iterative Closest Point (ICP) fitting, bypassing the need for computing homographies.

2.3 All-in-one Models

Most research commits to solving independent tasks without encapsulating each task into an all-in-one model. Theiner et al. [43] addressed the problem by combining CenterTrack [50] to detect and track players, a two Generative Adversarial Network (GAN) solution for camera pose estimation, TransNet [38] for viewpoint classification, and colour histogram and DBScan for team assignment. Indeed, while the solution provided state-of-the-art results utilising some of the most successful DL models for each task, performance regarding inference times and hardware was not recorded. Garner and Gregoir's [13] model utilised a Single Shot Multibox Detector (SSD) for player and ball detection with a reidentification feature embedding for tracking. The homography transformation was computed by keypoint masks from Effeicient-Netb3 on a feature pyramid network ready for processing by the DLT algorithm once matched to templates. In a situation when keypoints were sparse, a Deep Homography CNN based on ResNet-18 computed H. Stein et al. [39] also processed football footage to provide in-game analysis with impressive speeds on consumer-level hardware. Instead of using ML approaches to lay the groundwork for their analysis, they used background subtraction with edge and colour-based detection for player tracking. Moreover, a two-step homography transformation registered the current view into a pre-computed panoramic image of the pitch, which was then mapped to a pitch template.

Other research has developed all-in-one models for different use-case scenarios. For instance, Honda et al. [19] created a model to predict pass receivers using LSTM and transformer encoders, with YOLOv5 for player detection and ICP registration for frame mapping. In a different approach, Scott et al. [34] developed a drone and GNSS-based model for an aerial perspective, using YOLOv5 for detection. Though it lowers image processing needs, it struggles with adaptability and cost, especially for small-scale use and diverse camera movements. Our work contributes to the all-in-one models, encompassing MOT and perspective transformations.

3 Theoretical Framework

FootyVision is an all-in-one model for MOT and perspective transformations for uncalibrated broadcast video. FootyVision consists of three main modules:

- YOLOv7 network for player and ball detection
- Tracking module to assign identities to the detections
- Perspective transformation module to compute homographies for top-down localisation of tracks

Figure 2 provides an overview of the system architecture. Firstly, an uncalibrated video source is streamed into the YOLOv7 network, where detections undergo Non-Maxima Suppression (NMS) before being forwarded to the tracking module. Feature maps are extracted from activations in early layers of the YOLOv7 network. These feature maps provide substantial information regarding the geometric layout of the football pitch, which is relevant for computing the homographies through lines, intersections, and ellipses. The output from the model is tracked player and ball identities with their respective top-down view localisations suitable for strategic analysis.

3.1 MOT

MOT of football players and the ball is a challenging task due to the similar appearance of players, motion blur, and shape distortions. Our innovative all-in-one model achieves state-of-the-art MOT while repurposing information from the object detection network to remove bystander detections and compute perspective transformations. Our tracking algorithm consists of YOLOv7 object detection for players and the ball, team detection, feature embedding, and identity assignment with the Hungarian algorithm. The following subsections will explain each task in detail.

3.1.1 YOLOv7 Object Detection When the FootyVision was developed, YOLOv7 was the most recent addition to the YOLO family. In this research, we trained YOLOv7, from Wang et al. [46], on a custom dataset. YOLOv7 is designed for fast and accurate detection, particularly of smaller objects, making it well-suited for tasks like identifying a ball in football broadcast videos. The implementation of focal loss in its cost function enables YOLOv7 to detect small,

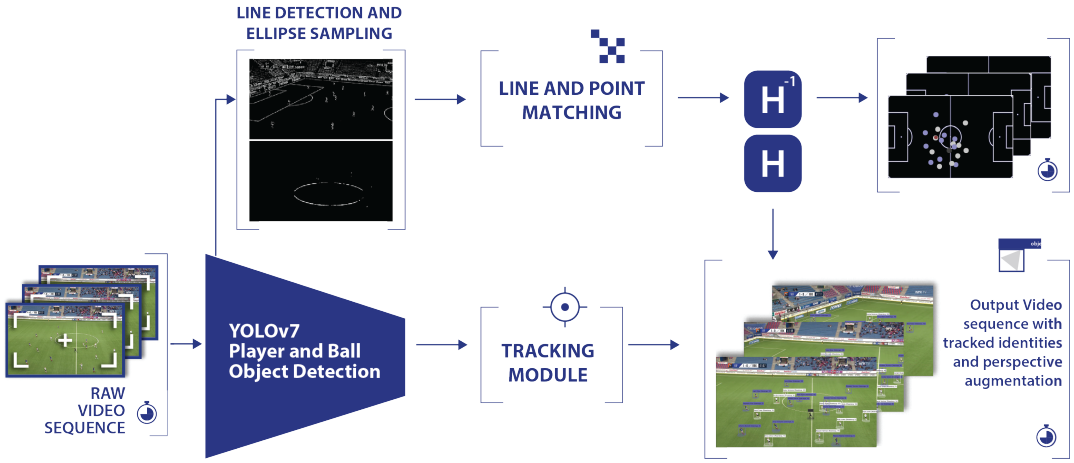


Figure 2: FootyVision, based on a YOLOv7 backbone, outputs bounding boxes to a tracking module for identity assignment from an incrementally constructed gallery. It employs a unique approach using intermittent activation maps for detecting lines, intersections, and ellipses, facilitating homography computation to localise players and the ball in a top-down viewpoint.

hard-to-see objects effectively. Additionally, the updated input resolution of 640x640 enhances its capability to identify smaller objects, increasing detection probability.

3.1.2 Removing Bystanders Naik et al. [29] proposed removing bystanders by classifying the background. We approach the problem by extracting an activation mask from the YOLOv7 network to remove bystanders classified as players. This methodology repurposes already processed information while simplifying the YOLOv7 training to detect just players and the ball. For our model, we do not regard linesmen as bystanders because they are integral members of the game.

3.1.3 Team Assignment For each frame, we extract a mask in Hue Saturation and Value (HSV) space and sample colours within a range that corresponds to the player's shirt colours. We use the bounding boxes of the player class from YOLOv7 to extract the Region of Interest (ROIs) of the three colour masks and assign a team based on which mask contains the maximum count of white pixels. The outcome is each bounding box being assigned to a team or classified as the referee.

3.1.4 Feature Embeddings Reidentification aims to assign identities to objects to be identified through consecutive frames and multiple camera angles. Currently, to the author's knowledge, there are no readily available datasets for reidentification of football players, therefore, we use Wieczorek et al.'s [47] model trained on the Market-1501 dataset. We first instantiate a gallery based on the feature embeddings, and then we take the weighted average of the feature vectors from a preset number of frames to provide historical context to the feature embedding. When a track is lost, we keep it alive before "flushing" tracks that have been inactive for a certain period of time.

3.1.5 Identity Assignment The Hungarian algorithm solves the linear assignment problem by yielding the minimal cost from the cost matrix C. In our research, the cost matrix is a product of the cosine similarity of the feature embedding and gallery, bounding box IoU, bounding box distance, and velocity.

$$CosineSimilarity = \cos(\theta) = \frac{A \cdot B}{||AB||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

Equation 1 outputs a similarity matrix of size m x n, where m is the number of current detections and n is the number of identities stored in the gallery.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (2)$$

Equation 2 is the Jaccard Index, otherwise known as IoU, which is computed between the current bounding boxes and those of the previous iteration collected from the respective instances of the player class.

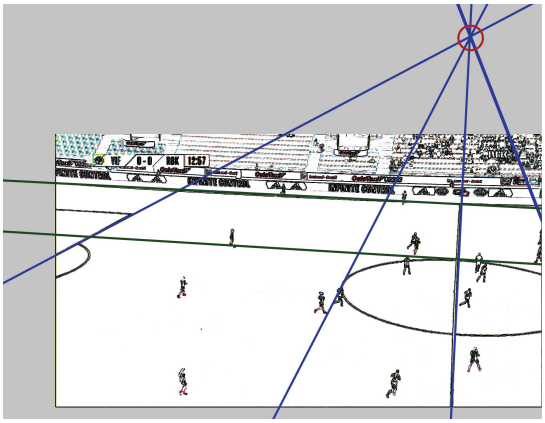
$$C = \lambda_{\text{feat}}(1 - J) + \lambda_{\text{iou}} \cos(\theta) + \lambda_{\text{dist}}(|c_x - c_y|)^2 + \lambda_{\text{vel}} V \quad (3)$$

We combine the two together while inverting the J matrix to meet the expectations of the Hungarian algorithm. Both distance, $(|c_x - c_y|)^2$, and velocity, V, are normalised before multiplying with their respective lambda weights. The lambda weights were identified during an ablation study outlined in Section 4.6. The Hungarian algorithm yields the corresponding identities that can be assigned to each bounding box, and the corresponding class is updated.

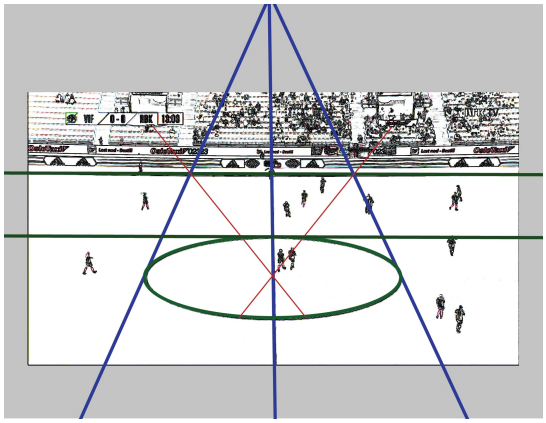
3.2 From Activation Maps to Homographies

An all-in-one model needs to localise player positions in a top-down view to assist in strategically analysing team formations. We compute the homography matrix by extracting lines and ellipses from the activation maps. Computing homographies is a well-documented procedure, so we refer the readers to Hartley & Zisserman [16] for comprehensive explanations. Our work uses the extended DLT algorithm documented in Dubrofsky et al. [10].

3.2.1 Line and Ellipse Detection To suit the dynamic nature of broadcast football videos we have employed two different processing routes based on the viewpoint of the image. We first denormalise a feature map from layer one before removing background noise



(a) Additional lines with two detected vertical lines.



(b) Additional lines with one detected vertical line.

Figure 3: Extending points and lines via ellipse detection and sampling. When two vertical lines are detected, a line is constructed connecting these lines' vanishing point to the ellipse's furthest horizontal point (3a). When no other vertical line is present, the ellipse is sampled, and a line is constructed from the intersection of the back line to the outermost horizontal ellipse point (3b).

from another activation map that masks the playing field. Depending on the angle and location of the center line, the viewpoint is classified and processed accordingly:

- **Side Viewpoint**: we split lines into horizontal and verticle clusters via RANSAC with a cost function measuring the distance from computed vanishing points (VP). Lines are then clustered using DBSCAN before fitting a line to each cluster for the best approximation.
- **Central Viewpoint**: First, an activation map which clearly depicts the central ellipse in the field is extracted and fitted with an ellipse using least squares, documented in Halir & Flusser [15]. When two verticle lines are present, the vanishing points are computed to be the first component of a line segment (see Figure 3a). While the outermost point of the ellipse is set as the second component. In the case of one vertical line, points are sampled around the ellipse circumference and an intersection is computed with the back horizontal line. This intersection, along with the outermost point of the ellipse, defines the line segment. A further two horizontal lines are added perpendicular to the verticle center line, which cross through the center top and center bottom points of the ellipse (see Figure3b).

3.2.2 Line and Intersection Matching The detected lines and intersections are matched to a database of templates with ground truth labels of the correspondences. A similarity matrix using Euclidean distance is computed for templates, selecting the one with the least distance. The Hungarian Algorithm assigns identities to each line using this similarity matrix as the cost matrix C. We stack the line and intersection matches into matrix A (as defined in Dubrofsky et al. [10]) and solve for matrix H via Singular Value Decomposition (SVD). With the detections and correspondences, it is possible to compute the H matrix with intersection points and lines.

Video Title	Time Stamp
2016 Leicester 0-0 Arsenal	07:42-09:32
2016 Leicester 3-0 Watford	46:22-47:24
2016 Leicester 0-1 West Bromwich	11:49-12:39
2016 Liverpool 0-0 Manchester United	45:52-47:26
2016 Arsenal 3-0 Chelsea	48:59-49:21
2016 Hull City 1-0 Arsenal	12:57-13:21
2016 Leicester 4-2 Manchester City	51:07-59:57

Table 1: Training Data from SoccerNet Dataset.

4 Experiments

This section details the training of YOLOv7 and evaluation methods for the tracking algorithm, with Section 4.6 offering a comprehensive ablation study on the individual components' impact on overall performance.

4.1 Datasets

We now outline the datasets used to train the YOLOv7 network and specify test datasets used for evaluating the MOT algorithm.

4.1.1 Object Detection The training of our player and ball YOLOv7 object detector is based on two datasets: ISSIA [9] and SoccerNet [8]. The ISSIA dataset consists of 15707 annotated images of players and balls. Frames containing just the goalkeeper were reduced, as there were already biases in the training data due to the player-ball ratio on the football pitch at any given time. The ISSIA dataset was re-synchronised with its labels due to a noticeable delay in tracks. We extended the dataset further with additional labelled images from the SoccerNet dataset annotated using Computer Vision Annotation Tool (CVAT) [35] (see Table 1 for details).

4.1.2 Tracking The tracking algorithm was tested on three datasets consisting of 1000 frames each (Table 3). All testing data consisted of previously unseen footage to evaluate how the model generalises to varying football video data. The chosen SoccerNet sequence offers closer player views similar in size to those in the ISSIA dataset but also includes varied viewpoints and camera movements, leading to more challenging sequences with motion blur and scale variations. In contrast, the ISSIA sequence uses a static camera, simplifying the

Method	Dataset	Pr(%)↑	Re(%)↑	F1-Score(%)↑	mAP(%)↑	Acc(%)↑	FPS ↑
DLBT [22]	ISSIA - Ball	93.25	73.25	-	-	**87.45**	10
FootAndBall [23]	ISSIA - Player	-	-	-	92.1	-	-
Small-Soccer Player [32]	ISSIA - Player	-	-	-	**97.3**	-	-
YOLOv3-SORT [29]	ISSIA - Player and Ball	96.57	93.3	91.6	93.47	-	23.7
FootyVision (Proposed)	ISSIA, SoccerNet - Player and Ball	**97.2**	**94.0**	**95.5**	95.7	-	**30.1**

Table 2: Object Detection Metrics Compared with State-of-the-Art Models. Table adapted from from Naik et al. [29].

task but offering lower video quality compared to the other datasets. We also tested on our industry partner, NRK's, dataset featuring a Norwegian football team with a wider angle and dynamic camera. The video clips were selected based on the inherent challenges they presented. For example, SoccerNet and NRK clips contained full and partial occlusions, varying camera angles and perspectives, a degree of motion blur, and distortions in players and the ball. The ISSIA clip was selected based on deducing information from Naik et al. [28], as the specific clip was not explicitly stated.

Dataset	Video Title	Time Stamp	Frames
SoccerNet	2016 Manchester United 4-1 Leicester	59:57	1000
ISSIA	Film Roll 1	00:00	1000
NRK	Valerenga vs Rosenborg	12:41	1000

Table 3: Datasets used with respective times and frames.

4.2 Training YOLOv7

YOLOv7 was trained on our extended dataset with a learning rate of 0.001. The model used PyTorch on a desktop computer with an NVIDIA A5000 Graphics card, AMD Ryzen 9 5950x 16-Core Processor and 64GB RAM. We employed several preprocessing techniques to enhance the training process and mitigate overfitting on the training dataset, including HSV augmentation, random translations, random scaling, and random horizontal flipping. These techniques enable the model to learn more robust features and generalise better to new data by introducing diversity and variability into the training samples.

4.3 Metrics

We use established metrics such as precision, recall, F1 score, and IoU to evaluate object detection performance. Our MOT evaluation utilises established metrics, such as CLEAR-MOT [3] and HOTA [26]. We refer readers to these papers for an in-depth description of each metric.

4.4 Implementation Details

Inference was performed on an MSI laptop with a 12th Gen Intel I9-12900HK 2.90Ghz CPU, NVIDIA GeForce RTX 3080 Ti GPU, and 32GB Ram. The DNNs were frozen before editing the network outputs with the ONNX library to access intermediate activation maps. Inference was carried out using a TensorRT Engine with Cuda 11.2 and CUDNN 8.2 acceleration. The model was written in Python with libraries Tensorflow 2.8, Numpy, Sklearn, SkImage, and OpenCV 4.5.2 with GPU support.

4.5 Evaluation

Table 2 compares our FootyVision model against other state-of-the-art football object detection models. Specifically, FootyVision achieves a precision (Pr) of 97.2%, recall (Re) of 94%, and F1-Score of

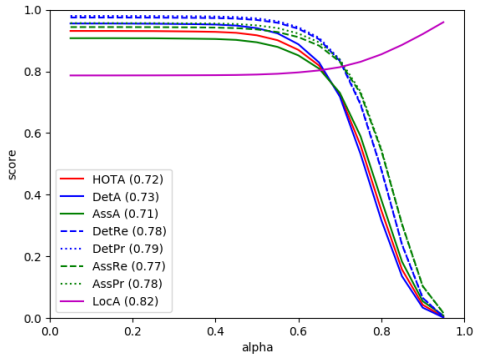


Figure 4: HOTA Evaluation.

95.5%, indicating superior performance in identifying players and the ball. While our mean Average Precision (mAP) score of 95.7% is highly competitive, it is important to note that Hurault & Haro's [32] model, which does not detect balls, achieves a slightly higher mAP of 97.3%. However, we should contextualise this comparison by recognising that their model does not include ball detection, a critical aspect of our evaluation. In terms of computational efficiency, our optimisation of the YOLOv7 model through TensorRT and GPU acceleration has significantly improved processing speed, achieving 30.1 frames per second (fps) for player and ball detection.

Regarding the performance of the tracking algorithm, FootyVision's performance is comparable and, in some cases, optimal compared to previous research. Considering Table 4, FootyVision generally outperforms other models in precision, recall, and IDF1. Considering MOTA, FootyVision outperforms on the SoccerNet dataset, achieves good results for the NRK dataset, and slightly underperforms with respect to ISSIA dataset compared to Naik et al. [28]. Concerning Multi-Object Tracking Precision (MOTP), FootyVision generally underperforms compared with documented results from Naik et al. The results further indicate that FootyVisions major drawback is the accuracy of the localisation of the bounding boxes when tracked across frames. While FootyVision demonstrates confidence in tracking and identity retention across frames, as indicated by its IDF1 score, its performance on the ISSIA dataset is notably lower regarding MOTA. This could be attributed to ISSIA's lower video quality, affecting spatial localisation and overall accuracy. However, despite these limitations, the high IDF1 score for ISSIA implies effective tracking. In contrast, the model shows improved results in the SoccerNet dataset, suggesting better generalisation and object detection in higher-quality videos.

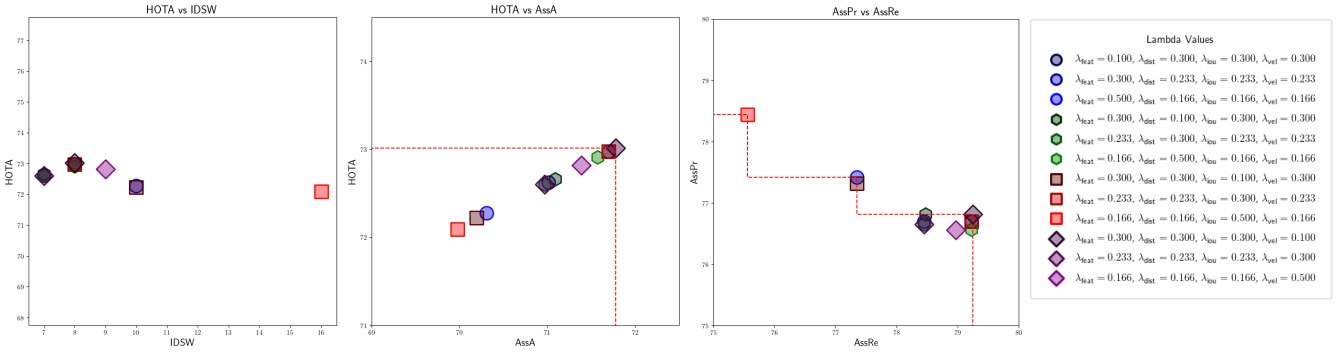
The analysis of FootyVision's tracking using HOTA metrics, as shown in Table 5, indicates strong performance on the SoccerNet dataset. The high HOTA(0) across tests suggests effective track

Method	Dataset	Pr(%)↑	Re(%)↑	IDF1(%)↑	IDR(%)↑	IDP(%)↑	MOTA(%)↑	MOTP(%)↑	MODA(%)↑	FPS ↑
DLBT [22]	ISSIA	93.25	73.25	-	-	-	-	-	-	**10**
Small-Soccer Player [32]	ISSIA	-	-	-	-	-	**97.3**	-	-	-
YOLOv3-SORT [29]	SoccerNet	90.1	89.2	85.8	-	-	87.2	**83.1**	-	**11.3**
	ISSIA	93.2	**91.7**	87.3	-	-	93.7	**88.6**	-	8.7
FootyVision (Proposed)	SoccerNet	**97.287**	**96.784**	**93.45**	**93.208**	**93.693**	**94.04**	78.972	**94.085**	9.7
	NRK	95.9	96.143	89.335	89.448	89.222	91.952	70.2	92.033	9.65
	ISSIA	**97.857**	90.301	**93.927**	**90.301**	**97.857**	88.323	69.84	**88.323**	9.52

Table 4: Multi-Object Tracking Metrics Compared with State-of-the-Art Models. Table adapted from Naik et al. [29].

Dataset	HOTA(%)↑	DetA(%)↑	AssA(%)↑	DetRe(%)↑	DetPr(%)↑	AssRe(%)↑	AssPr(%)↑	LocA(%)↑	HOTA(0)(%)↑	LocA(0)(%)↑	IDSW ↓
SoccerNet	72.168	73.169	71.27	78.137	78.543	77.225	78.171	81.749	93.108	78.677	6
ISSIA	64.096	61.294	67.361	64.908	70.339	71.419	73.417	75.328	94.671	69.397	0
NRK	62.078	63.602	60.703	68.634	68.461	63.691	71.553	75.318	89.909	70.048	14

Table 5: HOTA Metrics with configuration $\{\lambda_{\text{feat}} = 0.25, \lambda_{\text{dist}} = 0.2, \lambda_{\text{iou}} = 0.3, \lambda_{\text{vel}} = 0.25\}$ for SoccerNet $\{\lambda_{\text{feat}} = 0.233, \lambda_{\text{dist}} = 0.233, \lambda_{\text{iou}} = 0.3, \lambda_{\text{vel}} = 0.233\}$ for ISSIA, and $\{\lambda_{\text{feat}} = 0.266, \lambda_{\text{dist}} = 0.266, \lambda_{\text{iou}} = 0.266, \lambda_{\text{vel}} = 0.2\}$ for NRK.



(a) HOTA vs IDSW. **(b) HOTA vs AssA.** **(c) AssPr vs AssRe.**

Figure 5: Scatter plots considering performance with respect to tracking lambdas.

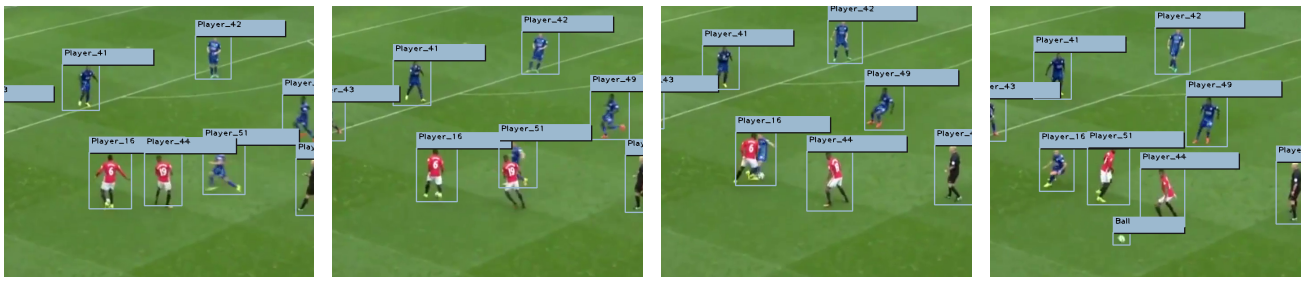
maintenance over frames. While localisation isn't pixel-perfect, as evidenced by the decent but improvable localisation accuracy LocA(0), the tracker consistently identifies correct object regions. Figure 4 illustrates performance metrics, such as Detection Accuracy (DetA, 0.73), Association Accuracy (AssA, 0.71), Detection Recall (DetRe, 0.78), Detection Precision (DetPr, 0.79), Association Recall (AssRe, 0.77), Association Precision (AssPr, 0.78), and Localisation Accuracy (LocA, 0.82) over varying alpha thresholds. As alpha becomes more stringent, a notable decline in performance scores indicates a significant trade-off between the precision of localisation and the overall tracking accuracy. Specifically, the decrease in the HOTA score to lower values as alpha tightens reflects the challenge of maintaining high accuracy when stricter localisation criteria are applied. This trend is consistent across evaluation metrics, confirming the impact of localisation precision on the tracking performance.

Considering Table 5 SoccerNet and NRK results, DetA surpasses AssA, indicating better detection accuracy than association. Conversely, ISSIA excels in association, suggesting effective track maintenance over time despite spatial localisation issues. The NRK dataset's lower AssRe and higher identity switch (IDSW) imply challenges in maintaining track consistency, unlike ISSIA, which shows no identity switches due to fewer tracks and static camera footage. SoccerNet, with six IDSW in an occlusion-heavy sequence, reveals the algorithm's partial efficiency during occlusions while highlighting limitations under complex conditions.

Both CLEAR-MOT and HOTA metics have indicated that while FootyVision's tracking algorithm is successful in maintaining tracks, it struggles with spatial accuracy of the bounding boxes. The HOTA metrics give further insights into the nuances of the tracking algorithm, revealing a high-level capability to maintain track identity even when the localisation precision is not stringent. The high HOTA scores across the datasets imply robustness in track identity maintenance, which is essential for applications where understanding the flow and pattern of play is more critical than the exact spatial accuracy. However, the lower LocA scores highlight the need to refine the algorithm's spatial precision.

4.6 Ablation Study

Our ablation study aimed to analyse and validate the impact of varying the lambda weights in the cost matrix construction on identity assignments by the Hungarian algorithm. By systematically adjusting the weights, we employed the leave-one-out method to better understand each component's role in the algorithm's performance. Considering Figure 5a, a balance of all components tends to reduce IDSW except for λ_{dist}, which consistently yields a low IDSW even when $\lambda_{feat} = 0.5$. However, when λ_{dist} is deactivated,



(a) Frame 922. (b) Frame 928. (c) Frame 940. (d) Frame 952.

Figure 6: SoccerNet Dataset: Tracking through full and partial occlusions. "Player_44" retains his identity through a partial occlusion in Figure 6b. Whereas, "Player_16" and "Player_51" switch identities after a full occlusion in Figure 6c.

there are eight identity switches, showing the other metrics collectively compensate to maintain a low incidence of ID switches. Similarly, λ_{vel} follows the same pattern, although it displays an optimal window in the lower ranges.

We observed a positive correlation between HOTA and AssA, as depicted in Figure 5b. The most effective weight combination for performance was $\{\lambda_{\text{feat}} = 0.3, \lambda_{\text{dist}} = 0.3, \lambda_{\text{iou}} = 0.3, \lambda_{\text{vel}} = 0.1\}$, which balanced the attributes with minimal emphasis on velocity. This balance was crucial as high-velocity weight reduced the IDSW but impacted HOTA negatively. Other notable combinations with lower IDSWs supported the importance of balanced weights across all components for optimal tracking performance. The stepped appearance in AssPr vs AssRe (Figure 5c) displays a trade-off concerning both these measures. The trade-off between precision and recall is often the case in tracking algorithms. The outlier, $\{\lambda_{\text{feat}} = 0.166, \lambda_{\text{dist}} = 0.166, \lambda_{\text{iou}} = 0.5, \lambda_{\text{vel}} = 0.166\}$, is consistent over all metrics in previous visualisations, indicating the lowest ranking configuration has a higher precision at the cost of recall. In contrast, the lambda configurations that perform well in HOTA, AssA, and IDFW tend to have higher recall than precision.

In conclusion, our ablation study confirmed the importance of each tracking component and the balance of lambda weights in the cost matrix for identity assignment with the Hungarian algorithm. Optimal adjustment of these weights is crucial for balanced component interaction and improved tracking.

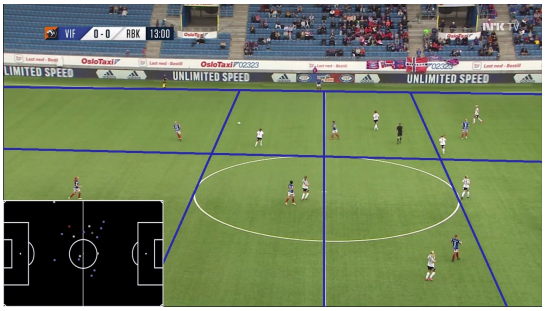
5 Discussion

To derive deeper insights, we delve into the evaluation and ablation study results of FootyVision, uncovering its strengths and weaknesses. By cross-validating these findings with qualitative data, we aim to enrich our discussion and present a more nuanced understanding. Additionally, We examine the perspective transformation results, suggesting improvements for future research.

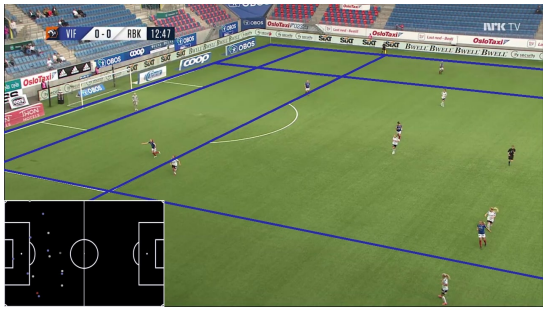
Our evaluation shows that training on SoccerNet and ISSIA datasets yielded improved object detection results compared to previous research. Considering Naik et al.'s [28] evaluation, their model underperformed on the SoccerNet dataset, whereas it had state-of-the-art performance on the ISSIA dataset. In contrast, FootyVisions tracking algorithm outperformed current state-of-the-art for SoccerNet datasets and retrieved competitive results for the ISSIA dataset. This shows that training YOLOv7 on a dataset with more variety than ISSIA helps it generalise to more varied data while maintaining satisfactory results. FootyVisions main limitation was maintaining spatial localisation throughout the lifespan of the track. Considering tracking of football players, maintaining good association and localisation is imperative to identify and localise players accurately in a top-down viewpoint. Therefore, it is necessary to improve these points before considering using such a model for real-time applications. The results indicate that performance varies across different datasets, particularly occlusion scenarios. For instance, the NRK dataset exhibited a higher incidence of IDSW than SoccerNet. This phenomenon suggests that factors such as camera proximity and the resulting angle relative to the field impact identity maintenance; the wider camera angle in NRK footage, which presents players on a smaller scale, likely contributes to these challenges. Our ablation study revealed that the feature parameter λ_{feat} significantly influences tracking effectiveness, especially when integrated with additional features. Tighter camera shots provide more prominent player segmentations, yielding richer visual data for the feature vectors. The diversity of features embedded in these vectors is crucial, as smaller player representations in the NRK dataset limit feature differentiation, leading to potential confusion during occlusions. Our algorithm's reliance on a combination of IOU, centroid distance, feature embeddings, and velocity attributes underscores the complexity of player tracking.

The ablation study highlights the importance of integrating attributes like IOU and feature embeddings for effective tracking. Over-reliance on isolated features leads to errors. Figure 6 illustrates an identity switch during full occlusion and identity recovery in partial occlusion. The algorithm's struggle with full occlusions is linked to high IoU and minimal spatial or feature distance between detections. To enhance identity tracking, especially through occlusions, we suggest using transfer learning with a reidentification network and football-specific datasets for more distinctive feature embeddings. Alternatively, implement a tracking system across multiple camera angles. Both approaches would improve robustness and reliability in MOT under challenging conditions.

The effectiveness of utilising activation maps from various layers to provide effective masks for lines and ellipses needed to compute the homography matrix is evident in the homography transformations shown in Figure 7. While the proposed methodology works well in challenging viewpoints, such as central views with limited information, there are still discrepancies in transformations due to lens barrel distortion, which can be seen in wider viewpoints



(a) Central Field Close Homography with Extended Line and Intersections from Ellipse Detection.



(b) Left Field Wide View Homography from Lines and Intersection.

Figure 7: Frames with Corresponding Inverse Homography Transformations from Lines, Intersections, and Ellipses.

(Figure 7b). Correcting lens distortion before applying the Hough transform would have improved perspective accuracy.

Output from FootyVision can provide rich insights from the football match. In particular, spatial-temporal data can provide information regarding the strategic elements of the game. For example, Wu et al. [48] used spatio-temporal data to gain deeper insights into team formations, assisting the user in their understanding of current game states. Moreover, rule-based systems built upon spatio-temporal data can infer in-game events. Tanaka-Ishii et al. [40, 41], Voelz et al. [45], and Binsted et al. [5] all showed how using these event systems can lay the groundwork for automated commentary. In our own research, FootyVision's data is the basis for interactive commentary in AiCommentator (in press, 2024).

6 Limitations

Comparing our work with others is challenging due to the inconsistent documentation in previous studies. We address this by detailing specific open-source video segments while implementing CLEAR-MOT and HOTA metrics for clearer benchmarking. While FootyVision provides a rich dataset that can infer events and strategic elements, its main drawback is its inability to process in real time and track through prolonged periods of full occlusion reliably. Therefore, post-processing is necessary to correct identity switches, connect fragmented tracks, and label the tracks. Future research should look to optimise tracking algorithms through multiple video streams to retain track identities while improving the speed of computation. We suggest using DL models and optical character recognition to detect shirt numbers and link the associated identities automatically. We plan to correct barrel distortion in future iterations to enhance the accuracy of perspective transformations.

7 Conclusion and Future Work

We have introduced FootyVision, an innovative approach to MOT, localisation, and augmentation in football videos. Utilising a YOLOv7 backbone for object detection, FootyVision excels in detecting players and the ball in football broadcast video while surpassing accuracy with MOT. Qualitative results show that FootyVision's novel perspective transformation module copes well with viewpoints that contain limited visual information by extracting geometric features from YOLOv7 activation maps. Our extensive quantitative evaluation across ISSIA and SoccerNet datasets demonstrates the robustness and adaptability of FootyVision while setting a comprehensive baseline for future research in MOT of players and the ball in football video. An ablation study analyses the effect of each tracking feature with respect to the challenges of tracking identities with high similarity. We find a balanced contribution of all features achieves state-of-the-art MOT results. Despite these advancements, challenges in tracking accuracy and localisation still remain prevalent during periods of high occlusion. Addressing these challenges forms the primary avenue for future research. FootyVision represents a substantial advancement in sports analytics, offering insights and capabilities to analyse and enjoy football. The system not only sets a new standard in accuracy but also lays the groundwork for future innovations in the field. Our current work AiCommentator (in press, 2024), has shown the plausibility of utilising FootyVision for interactive platforms. We look forward to continuing to improve FootyVision while exploring the diverse applications and developments possible from this research.

Acknowledgments

This research was funded by SFI MediaFutures partners and the Research Council of Norway, grant number 309339. We thank Ayça Ünlüer for the illustrations and Philippa Beckman for proofreading.

References

[1] Sermetcan Baysal and Pınar Duygulu. 2015. Sentioscope: a soccer player tracking system using model field particles. *IEEE Transactions on Circuits and Systems for Video Technology* 26, 7 (2015), 1350–1362. https://doi.org/10.1109/TCSVT.2015.2455713

[2] Michael Beetz, Suat Gedikli, Jan Bandouch, Bernhard Kirchlechner, Nico von Hoyningen-Huene, and Alexander Perzylo. 2007. Visually tracking football games based on TV broadcasts. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*.

[3] Keni Bernardin and Rainer Stiefelhagen. 2008. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing* 2008 (2008), 1–10.

[4] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. 2016. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*. IEEE, 3464–3468.

[5] Kim Binsted and Sean Luke. 1999. Character Design for Soccer Commentary. *Lecture Notes in Computer Science* (1999). https://doi.org/10.1007/3-540-48422-1_2

[6] Leonardo Citraro, Pablo Márquez-Neila, Stefano Savare, Vivek Jayaram, Charles Dubout, Félix Renaut, Andres Hasfura, Horesh Ben Shitrit, and Pascal Fua. 2020. Real-time camera pose estimation for sports fields. *Machine Vision and Applications* 31, 3 (2020), 1–13. https://doi.org/10.1007/s00138-020-01064-7

[7] Anthony Dearden, Yiannis Demiris, and Oliver Grau. 2006. Tracking football player movement from a single moving camera using particle filters. (2006). https://doi.org/10.1049/cp:20061968

[8] Adrien Deliege, Anthony Cioppa, Silvio Giancola, Meisam J Seikavandi, Jacob V Dueholm, Kamal Nasrollahi, Bernard Ghanem, Thomas B Moeslund, and Marc Van Droogenbroeck. 2021. Soccernet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4508–4519. https://doi.org/10.1109/CVPRW53098.2021.00508

[9] Tiziana D'Orazio, Marco Leo, Nicola Mosca, Paolo Spagnolo, and Pier Luigi Mazzeo. 2009. A semi-automatic system for ground truth generation of soccer video sequences. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*. IEEE, 559–564. https://doi.org/10.1109/AVSS.2009.69

[10] Elan Dubrofsky and Robert J Woodham. 2008. Combining line and point correspondences for homography estimation. In *International symposium on visual computing*. Springer, 202–213. https://doi.org/10.1007/978-3-540-89646-3_20

[11] FIFA. 2021. The football landscape – the vision 2020-2023. https://publications.fifa.com/en/vision-report-2021/the-football-landscape/

[12] Pascual Figueroa, Neucimar Leite, Ricardo ML Barros, Isaac Cohen, and Gerard Medioni. 2004. Tracking soccer players using the graph representation. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, Vol. 4. IEEE, 787–790. https://doi.org/10.1109/ICPR.2004.1333890

[13] Paul Garnier and Théophane Gregoir. 2021. Evaluating soccer player: From live camera to deep reinforcement learning. *arXiv preprint arXiv:2101.05388* (2021). https://doi.org/10.48550/arXiv.2101.05388

[14] Ankur Gupta, James J Little, and Robert J Woodham. 2011. Using line and ellipse features for rectification of broadcast hockey video. In *2011 Canadian conference on computer and robot vision*. IEEE, 32–39. https://doi.org/10.1109/CRV.2011.12

[15] Radim Halır and Jan Flusser. 1998. Numerically stable direct least squares fitting of ellipses. In *Proc. 6th International Conference in Central Europe on Computer Graphics and Visualization. WSCG*, Vol. 98. Citeseer, 125–132. https://doi.org/10.1109/34.765658

[16] Richard Hartley and Andrew Zisserman. 2003. *Multiple view geometry in computer vision*. Cambridge university press. https://doi.org/10.1017/CBO9780511811685.008

[17] Jean-Bernard Hayet and Justus Piater. 2007. On-line rectification of sport sequences with moving cameras. In *Mexican international conference on artificial intelligence*. Springer, 736–746. https://doi.org/10.1007/978-3-540-76631-5_70

[18] Jean-Bernard Hayet, Justus Piater, and Jacques Verly. 2004. Robust incremental rectification of sports video sequences. In *British machine vision conference (BMVC'04)*. Citeseer, 687–696. https://doi.org/10.5244/C.18.71

[19] Yutaro Honda, Rei Kawakami, Ryota Yoshihashi, Kenta Kato, and Takeshi Naemura. 2022. Pass Receiver Prediction in Soccer Using Video and Players' Trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3503–3512. https://doi.org/10.1109/CVPRW56347.2022.00394

[20] P Huang and A Hilton. 2006. Football player tracking for video annotation. In *The 3rd European Conference on Visual Media Production (CVMP 2006)-Part of the 2nd Multimedia Conference 2006*. IET, 175–175. https://doi.org/10.1049/cp:20061941

[21] Euromonitor International. 2023. Women's World Cup 2023 viewership to cross 2 billion, double from 2019: Euromonitor International. https://www.euromonitor.com/press/press-releases/july-20232/

[22] Paresh R Kamble, Avinash G Keskar, and Kishor M Bhurchandi. 2019. A deep learning ball tracking system in soccer videos. *Opto-Electronics Review* 27, 1 (2019), 58–69. https://doi.org/10.1016/j.opelre.2019.02.003

[23] Jacek Komorowski, Grzegorz Kurzejamski, and Grzegorz Sarwas. 2019. Footandball: Integrated player and ball detector. *arXiv preprint arXiv:1912.05445* (2019). https://doi.org/10.5220/0008916000470056

[24] Daniel Linke, Daniel Link, and Martin Lames. 2020. Football-specific validity of TRACAB's optical video tracking systems. *PloS one* 15, 3 (2020), e0230179.

[25] Jia Liu, Xiaofeng Tong, Wenlong Li, Tao Wang, Yimin Zhang, and Hongqi Wang. 2009. Automatic player detection, labeling and tracking in broadcast soccer video. *Pattern recognition letters* 30, 2 (2009), 103–113. https://doi.org/10.5244/C.21.3

[26] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. 2021. Hota: A higher order metric for evaluating multi-object tracking. *International journal of computer vision* 129 (2021), 548–578.

[27] Jesus Martinez-del Rincon, Elias Herrero-Jaraba, J Raul Gomez, Carlos Orrite-Urunuela, Carlos Medrano, and Miguel A Montanes-Laborda. 2009. Multicamera sport player tracking with Bayesian estimation of measurements. *Optical Engineering* 48, 4 (2009), 047201. https://doi.org/10.1117/1.3114605

[28] Banoth Thulasya Naik and Mohammad Farukh Hashmi. 2021. Ball and Player Detection & Tracking in Soccer Videos Using Improved YOLOV3 Model. (2021). https://doi.org/10.21203/rs.3.rs-438886/v1

[29] B Thulasya Naik and Md Farukh Hashmi. 2022. YOLOv3-SORT: detection and tracking player/ball in soccer sport. *Journal of Electronic Imaging* 32, 1 (2022), 011003. https://doi.org/10.1117/1.JEI.32.1.011003

[30] Xiaohan Nie, Shixing Chen, and Raffay Hamid. 2021. A robust and efficient framework for sports-field registration. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1936–1944. https://doi.org/10.1109/WACV48630.2021.00198

[31] Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018).

[32] Coloma Ballester Samuel Hurault and Gloria Haro. 2020. Self-Supervised Small Soccer Player Detection and Tracking. *CoRR* abs/2011.10336 (2020). https://doi.org/10.1145/3422844.3423054 arXiv:2011.10336

[33] Ryan Sanford, Siavash Gorji, Luiz G Hafemann, Bahareh Pourbabaee, and Mehrsan Javan. 2020. Group activity detection from trajectory and video data in soccer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 898–899.

[34] Atom Scott, Ikuma Uchida, Masaki Onishi, Yoshinari Kameda, Kazuhiro Fukui, and Keisuke Fujii. 2022. SoccerTrack: A Dataset and Tracking Algorithm for Soccer With Fish-Eye and Drone Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3569–3579. https://doi.org/10.1109/CVPRW56347.2022.00401

[35] Boris Sekachev, Nikita Manovich, Maxim Zhiltsov, Andrey Zhavoronkov, Dmitry Kalinin, Ben Hoff, TOsmanov, Dmitry Kruchinin, Artyom Zankevich, DmitriySidnev, Maksim Markelov, Johannes222, Mathis Chenuet, a andre, telenachos, Aleksandr Melnikov, Jijoong Kim, Liron Ilouz, Nikita Glazov, Priya4607, Rush Tehrani, Seungwon Jeong, Vladimir Skubriev, Sebastian Yonekura, vugia truong, zliang7, lizhming, and Tritin Truong. 2020. *opencv/cvat: v1.1.0*. https://doi.org/10.5281/zenodo.4009388

[36] Long Sha, Jennifer Hobbs, Panna Felsen, Xinyu Wei, Patrick Lucey, and Sujoy Ganguly. 2020. End-to-end camera calibration for broadcast videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 13627–13636. https://doi.org/10.1109/CVPR42600.2020.01364

[37] Feng Shi, Paul Marchwica, Juan Camilo Gamboa Higuera, Michael Jamieson, Mehrsan Javan, and Parthipan Siva. 2022. Self-Supervised Shape Alignment for Sports Field Registration. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 287–296. https://doi.org/10.1109/WACV51458.2022.00382

[38] Tomáš Souček and Jakub Lokoč. 2020. Transnet V2: an effective deep network architecture for fast shot transition detection. *arXiv preprint arXiv:2008.04838* (2020). https://doi.org/10.48550/arXiv.2008.04838

[39] Manuel Stein, Halldor Janetzko, Andreas Lamprecht, Thorsten Breitkreutz, Philipp Zimmermann, Bastian Goldlücke, Tobias Schreck, Gennady Andrienko, Michael Grossniklaus, and Daniel A Keim. 2017. Bring it to the pitch: Combining video and movement data to enhance team sport analysis. *IEEE transactions on visualization and computer graphics* 24, 1 (2017), 13–22. https://doi.org/10.1109/TVCG.2017.2745181

[40] Kumiko Tanaka-Ishii, K. Hasida, and I. Noda. 1998. Reactive Content Selection in the Generation of Real-time Soccer Commentary. *ACL* (1998). https://doi.org/10.3115/980691.980778

[41] Kumiko Tanaka-Ishii, I. Noda, I. Frank, H. Nakashima, K. Hasida, and H. Matsubara. 1998. MIKE: an automatic commentary system for soccer. *Proceedings International Conference on Multi Agent Systems (Cat. No.98EX160)* (1998). https://doi.org/10.1109/ICMAS.1998.699067

[42] Shuhei Tarashima. 2021. Sports Field Recognition Using Deep Multi-task Learning. *Journal of Information Processing* 29 (2021), 328–335. https://doi.org/10.2197/ipsjjip.29.328

[43] Jonas Theiner, Wolfgang Gritz, Eric Müller-Budack, Robert Rein, Daniel Memmert, and Ralph Ewerth. 2022. Extraction of positional player data from broadcast soccer videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 823–833. https://doi.org/10.1109/WACV51458.2022.00153

[44] Ferran Vidal-Codina, Nicolas Evans, Bahaeddine El Fakir, and Johsan Billingham. 2022. Automatic event detection in football using tracking data. *Sports Engineering* 25, 1 (2022), 18.

[45] Dirk Voelz, Elisabeth André, Gerd Herzog, and Thomas Rist. 1999. Rocco: A RoboCup Soccer Commentator System. *Lecture Notes in Computer Science* (1999). https://doi.org/10.1007/3-540-48422-1_4

[46] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. 2022. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696* (2022). https://doi.org/10.48550/arXiv.2207.02696

[47] Mikołaj Wieczorek, Barbara Rychalska, and Jacek Dąbrowski. 2021. On the unreasonable effectiveness of centroids in image retrieval. In *Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part IV 28*. Springer, 212–223. https://doi.org/10.48550/arXiv.2104.13643

[48] Yingcai Wu, Xiao Xie, Jiachen Wang, Dazhen Deng, Hongye Liang, Hui Zhang, Shoubin Cheng, and Wei Chen. 2018. Forvizor: Visualizing spatio-temporal team formations in soccer. *IEEE transactions on visualization and computer graphics* 25, 1 (2018), 65–75.

[49] Ming Xu, James Orwell, and Graeme Jones. 2004. Tracking football players with multiple cameras. In *2004 International Conference on Image Processing, 2004. ICIP'04.*, Vol. 5. IEEE, 2909–2912. https://doi.org/10.1109/ICIP.2004.1421721

[50] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. 2020. Tracking objects as points. In *European Conference on Computer Vision*. Springer, 474–490. https://doi.org/10.1007/978-3-030-58548-8_28