Multimedia Systems (2023) 29:897–915
https://doi.org/10.1007/s00530-022-01027-0

REGULAR PAPER

Use of deep learning in soccer videos analysis: survey

Sara Akan[1] · **Songül Varlı**[1]

Received: 2 July 2022 / Accepted: 15 November 2022 / Published online: 3 December 2022

Abstract
The demand for video analysis has been rapidly increasing in the last decade. Video analysis plays a critical role in various technologies, including medical diagnosis, security surveillance, robotics, and sport. Soccer is the most popular sport in our culture, with millions of fans. Many video analysis approaches have been developed in recent years to assist and provide important information to spectators, referees, coaches, and players. Most of these approaches are aimed towards detecting and tracking players or the ball, event detection, and analysis of the game. For this purpose, various classical or deep learning-based strategies have been used. This study investigates deep learning-based techniques that have been proposed over the last few years to analyze football videos. The purpose of this study is not to compare current methodologies, but to show the most recent research in the field. This paper investigates the challenges of soccer video analysis and its application groups, e.g., player/ball detection and tracking, event detection, and game analysis. This paper also reviews the used deep learning-based methods, their performance, advantages, and disadvantages in soccer videos, and finally, concludes with future potential in the analysis of soccer videos.

Keywords Soccer · Football · Survey · Review · Overview · Deep learning · Event detection · Player detection · Ball detection · Player tracking · Ball tracking · Game analysis · Team performance

1 Introduction

Sport is a physical and mental activity that can be carried out both individually or as a team with predetermined rules. Coaches and sports experts have an essential effect in evaluating and training players to achieve high performance in sports. Due to the growing number of fans in sports games such as soccer, the need for further analysis of the game and the provision of extensive information to sports fans is needed. Different classes of people need different types of information, e.g., the public may be required to see the events from a variety of perspectives; to fine-tune their plans or actions, amateurs need to learn the motions of professional athletes; to design their methods, athletes and coaches study the tactics and moves of competitors and broadcasting organizations must identify, store, organize, and summarize videos of sporting events, as well as explain the games. Athletes and coaches analyze the tactics and movements of opposing players and teams to design their strategies, and broadcasting agencies are required to detect, store, categorize and summarize the videos of sports competitions and explain the games. Performance analysis is the process of determining an athlete's level of performance. Two types of this kind of analysis are technical analysis and tactical analysis. The former addresses the question of how the game is played by the player, while the latter analyses which activity is carried out.

In the last few years, different approaches have been applied to analyze football videos in the field of computer vision. Initially, many researchers proposed classical approaches for soccer video analysis. Deep learning has been a popular subject of current attention in sports video analysis as a result of its success in various computer vision fields. In general, sports video analysis can be divided into several aspects such as ball, trajectory detection, and motion recognition, as well as tracking players, especially in soccer games.

✉ Sara Akan
sara.behjatjamal@gmail.com;
sara.behjatjamal@std.yildiz.edu.tr

Songül Varlı
svarli@yildiz.edu.tr

[1] Department of Computer Engineering, Yildiz Technical University, Istanbul, Turkey

Fig. 1 Some instances of changes in ball size, shape, and color at various speeds [11]

In general, the two available sources of soccer video streams are TV broadcast cameras and fixed cameras placed around the playing area. Broadcast image analysis to detect important events in soccer games has been the subject of many research [1]. On the other hand, some studies have used custom cameras carefully located around the playing field for even more particular tasks that broadcast cameras cannot solve, such as modeling player actions, capturing and analysis of team techniques, and evaluating player performances.

Due to the increasing demand for soccer video analysis, research in this area has been increased. Some studies summarize and compare proposed soccer video analysis schemes in recent years. The first review of the soccer video analysis that summarizes existing techniques was presented in 2010 [1]. Later, some other review papers have been presented to summarize the soccer video analytics works [2–8]. However, they did not provide a complete summary of all available application types and focused only on the analysis of specific subsets of approaches [7] reviews approach for implementing text resources (e.g., social networks) for event detection and video summary based on video streams.

Techniques for tracking players or the ball were the topic of research in [2, 5]. The most relevant available tracking techniques are mentioned and compared based on the specific features employed, such as motion, color, and edges. In [5], state-of-the-art technology corresponding to preprocessing and tracking algorithms applied in strategies to track players is analyzed. Moreover, [8] provides an overview of the approaches and methodologies proposed over the last 2 decades to analyze soccer video sequences. According to the literature, survey/review studies on this subject in the past generally focused on classical methods. However, the increase in the performance of deep learning-based methods and the fact that they are amenable to real-time implementation have given importance to learning-based methods. In the same way, the use of these methods in video analysis has also increased. The purpose of this article is not to cover the existing methods but to comprehensively explore the purposes of using deep learning-based methods in soccer video analysis. In addition, the classical methods have been briefly described.

The rest of the paper is organized as follows: the second section discusses the challenges in the ball and player analysis. In the third section, player and ball detection/tracking, event detection, and game analysis in football videos are explained. Section 4 carries out an overview of the current techniques for the analysis of soccer videos. In Sect. 5, a discussion of observations in the literature review and developments so far is presented. Section 6 presents the conclusion and possible future work.

2 Challenges

Algorithms such as player tracking, player detection, and activity analysis have been faced with several difficulties. For example, overlapping players wearing the same jersey, unpredictable trajectories, poor lighting conditions, the ball not being visible due to a wide camera view, and interactions between the ball and the players can cause complex problems. As a result, the analysis of soccer videos has become a very challenging subject for scientists.

2.1 Players

In order to properly assess the game, it is crucial to detect and track players. However, to achieve successful results, difficulties need to be overcome such as overlapping of players, unexpected camera movements, light changes, poor resolution in faraway players, blurring of moving players, or players staying static for long periods [9].

2.2 Ball

Tracking ball movement is extremely important for any information extraction from ball-based sports videos. Due to the nature of the game, the view of the ball, direction, and background changes dynamically. In addition, there are many challenges and issues such as occlusion, false detection, lighting variation, and different video quality with varying frame rates. Therefore, detection and monitoring processes become difficult.

2.3 Speed

The speed of the ball plays an important role in detecting and tracking the ball [10]. In soccer games, the ball is constantly

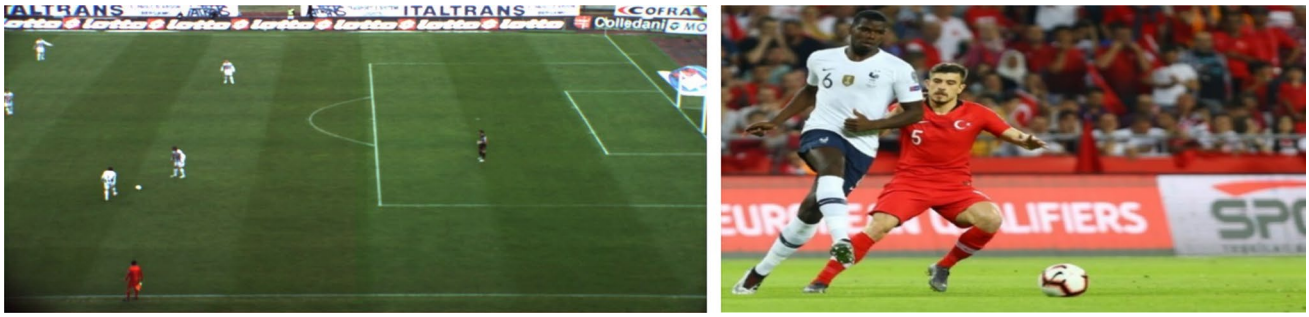


Fig. 2 Examples of ball images with different distance

Fig. 3 Examples of of partial or complete occlude with players

moving and can sometimes have high speed. The speed of the ball changes throughout the game. As seen in Fig. 1, if the speed of the ball increases, the shape, color, and size of the ball appear to be different. Moreover, sometimes the shape of the ball looks more oval because of the blur caused by the speed. Furthermore, balls sometimes consist of more than one color and the acceleration of the ball causes the colors to be mixed. Consequently, features extracted from the fixed ball may not be very useful for detecting the moving ball.

2.4 Quality

Detecting the ball from long-range video footage of a football match is a challenge. First, the ball is very small compared to other objects appearing in the observer scene. The size of the ball varies considerably depending on the position. In long-range video footage of football matches, the ball may appear as small as 8 pixels when directly in front of the camera on the far side of the field, and as large as 20 pixels when near the field. In Fig. 2, examples of ball images with different distances are given.

2.5 Occlusion

As seen in Fig. 3, when the ball is under the control of the player, it can be completely or partially covered by the legs or bodies of the players because of its small size. In this case, it is very difficult to detect and track the ball. In such cases, simple ball detection methods with motion-based background subtraction may fail.

Fig. 4 Examples of false detection

2.6 False detection

Looking at the examples in Fig. 4, the small-sized ball may not be easily distinguished from parts of the player's body (head or socks) or background clutter (small litter on the field or parts of stadium advertisements). These situations lead to false detection. Several methods have been developed to overcome difficulties such as speed, occlusion, and different appearances in the studies carried out so far. However, it

Fig. 5 Examples of player detection and tracking in two different frames

was generally not successful in false detection. Taking this issue into account, detection and tracking is still a subject that remains open to study.

3 Application groups

In general, soccer video analysis studies can be divided into four groups, according to their purposes and needs: player detection and tracking; ball detection and tracking; event detection; and game analysis. The first and second groups include applications that focus on spotting and tracking the players and the ball. These applications often employ several strategies to resolve tricky analytical conditions, such as players' occlusions, unexpected camera movements, or sudden illumination changes. The purpose of event detection applications is to identify the most important events that occurred during the broadcast of the match, such as goals, fouls, goal kicks, corner kicks, penalties, and so on. Eventually, applications in the game analysis are focused on analyzing what happens during matches. Typically, analysis methods such as low-level detection and tracking are applied first. The results are then used to provide various high-level analyses of the match. The high-level analysis includes the distance traveled by players from each team [12], possession statistics [13], offside detection [14], and team tactics [15]. Usually, apps in event detection are offline and used when the match is over. On the other hand, detection, tracking, and game analysis applications are frequently employed during the game or in real time. For example, it is used to calculate statistics or analyze goals or offsides after a few minutes [1].

3.1 Player detection and tracking

The initial stage in player tracking is detecting the position of players at any given time. To enhance team performance in sports such as soccer, examining both individual player movement trajectories and the general composition of the team can give invaluable information to the team coach. Preferably, the coach would like to have access to a comprehensive record of all players' positions numerous times during the workout or game. There is a lot of research available for the detection and tracking of players. For example, pedestrian tracking or player tracking for other sports can also be used for football players. Therefore, this field is sufficiently developed. However, its fully automatic tracking and tagging is still a subject of study. Examples of player detection and tracking are shown in Fig. 5**.**

3.2 Ball detection and tracking

The game of football is shaped around a ball. The potential of ball tracking in real time is essential for assessing TV broadcasts as well as helping the referee. Therefore, the automatic detection of a ball is important for statistics such as possession, as well as the classification of goals and other events. Some existing computer vision-based target detection systems have now been developed for commercial uses even in world cup games. For instance, 108 soccer grounds[1] in 2017 have goal-line technology (GLT) installations that FIFA certified. These systems use multi-view high-speed cameras that cover each target area. It is not feasible to implement a ball tracking system as it would require a large number of calibrated cameras to roll out a ball tracking system across the field. Due to the rapid movement and small size of the ball, it is very difficult to detect and track the ball according to the size of the field. The majority of ball detection algorithms produce acceptable results by selecting objects in motion as ball candidates and classifying them by color and shape [16]. However, the ball is usually partially or completely covered by the players. In this case, only the detection algorithm may not be useful. Instead, combining detection and tracking can provide more accurate results to predict ball position. Whoever possesses the ball may be represented and utilized as a part of the ball trajectory by combining information about the locations of the players and the ball [17]. Unlike the players, the ball cannot be assumed to move in the ground plane. Instead, it is assumed to move in a 3-dimensional space. More than one calibrated

[1] https://www.fifa.com/technical/football-technology.

Fig. 6 Example of ball detection and tracking

camera is usually installed to resolve such uncertainties. In [18], a series of states that the ball can take, such as flying, rolling, or being possessed, are defined. Various ball tracking approaches are available to track the ball, such as the Kalman filter, Particle filter, trajectories, Data association, and Graf. In Fig. 6**,** examples of ball detection are given.

be able to see the most attractive aspects of the game in this manner. Some apps are intended to automatically create video recaps of the most intriguing games of a match.

However, one of the fundamental disadvantages of event detection applications is their high processing costs, since the whole video of the game needs to be analyzed. Furthermore, all of the tasks carried out by these apps are often completed after the game, i.e., offline. As a result, this research often has no runtime or computational constraints.

3.4 Game analysis

Some studies concentrate on team strategies and gathering individual and global statistics. These data are important not just for viewers, but also for referees, coaches, and players because it provides both a greater comprehension of the game and an evaluation of the teams' strategies. It may also be used to design training sessions to help players perform better [20]. The majority of these applications try to analyze the players and the ball positions during the game to provide various indicators that can make the game easier to understand [6], such as the distance traveled by each team's players, the distance from the ball to the goal in a free-kick, positions taken by players throughout the game, and scoring opportunities.

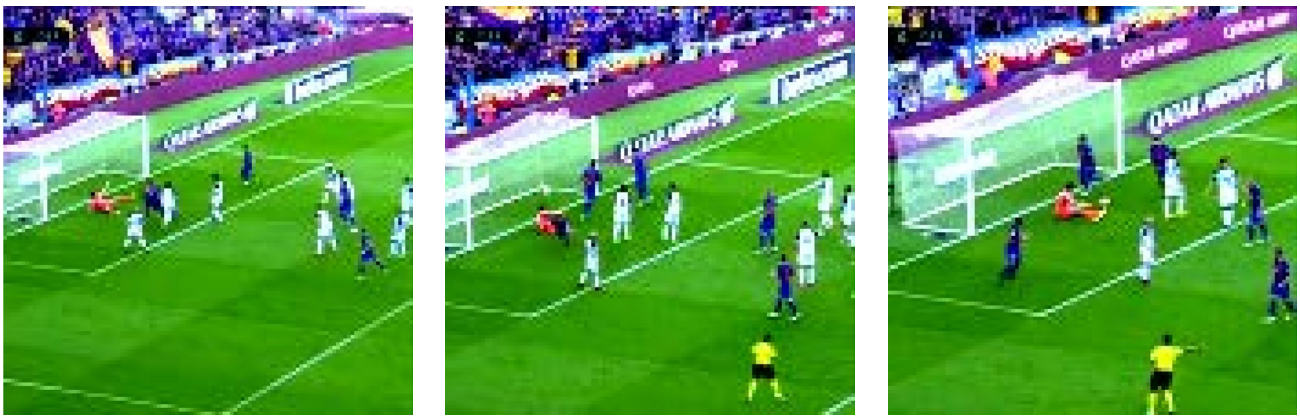
Table 1 Comparison between available techniques

	Model	Advantage	Drawback
Classical methods	Blob-based	Detect candidate object regions and reduce the number of false alarms	It is very difficult to detect and track the small size thing
			Useful just for detection
	Feature-based	Allow data collection or data production based on prior experience	Manually extract important information
		Can optimize performance criteria as well	Consider the dependence
		Computational complexity	Useful just for detection and tracking
Deep learning methods	CNN	Capture spatial information automatically in the image patches	Could not able to capture the temporal information in video data
	RNN	Automatic temporal information collection for sequential data	Has short memory ability, making them inapplicable in a real situation
			Gradient explosion
	LSTM	Automatic temporal information collection for sequential data	Take longer to train
			More memory training is needed
			Easy to overfit
			Dropout is significantly difficult to implement
			Sensitive to different initializations of random weights

3.3 Event detection

These applications aim to detect the most essential events that occur during a football match broadcast, such as goal-scoring, fouls, shots on goal, corners, penalties, and offside positions automatically (see Fig. 7). These applications are the result of a desire to automate the process of detecting events that could be of interest to the audience. Users will

(a) Card.



(b) Goal.

Fig. 7 Samples of frames taken from the SoccerNet dataset [19] represent three different types of events

4 Available techniques

As seen in Fig. 8 in this paper, the methods used so far are divided into two groups: classical methods and methods based on deep learning. In order to analyze a football video, it is often necessary to detect the players. As a result, traditional approaches such as blob-based and feature-based methods are preferred. On the other hand, deep learning approaches provide more effective analysis in many cases. Due to the inadequacy of annotated datasets created from sports games, many athlete detection algorithms use deep learning models that have been trained before fine-tuning for sports branches. In the following sections, both classical methods and deep learning methods are discussed in more detail. Table 1 shows a comparison between classical and deep learning models.

4.1 Classical approaches

In this section, broad research of classical methodologies and their applications in soccer video analysis has been conducted. Many non-deep learning approaches for detecting players and the ball have been developed [5, 21]. Classical approaches can be divided into two groups: blob-based approaches [22, 23] and feature-based methods [24, 25]. In the following section, an overview of both approaches, as well as algorithms that use them, is provided.

4.1.1 Overview of the blob base methods

The main purpose of blob-based methods is to detect candidate object regions and reduce the number of false alarms. These methods are mostly based on background subtraction or region-based segmentation (see Fig. 9) [23]. The foreground is a series of pixels that drastically change their position in the current frame relative to their position in the previous frame. Such a set of pixels constitutes an object that moves in each subsequent frame. The Background is a series of pixels that do not change their position or change their position by a few pixels per frame. Background subtraction achieves low-level features [26]. It is a popular method for separating the moving parts of a scene into background and foreground. Gaussian Mixture Models (GMMs) [27], Frame differencing [28], Temporal Median Filtering (TMF) [29],

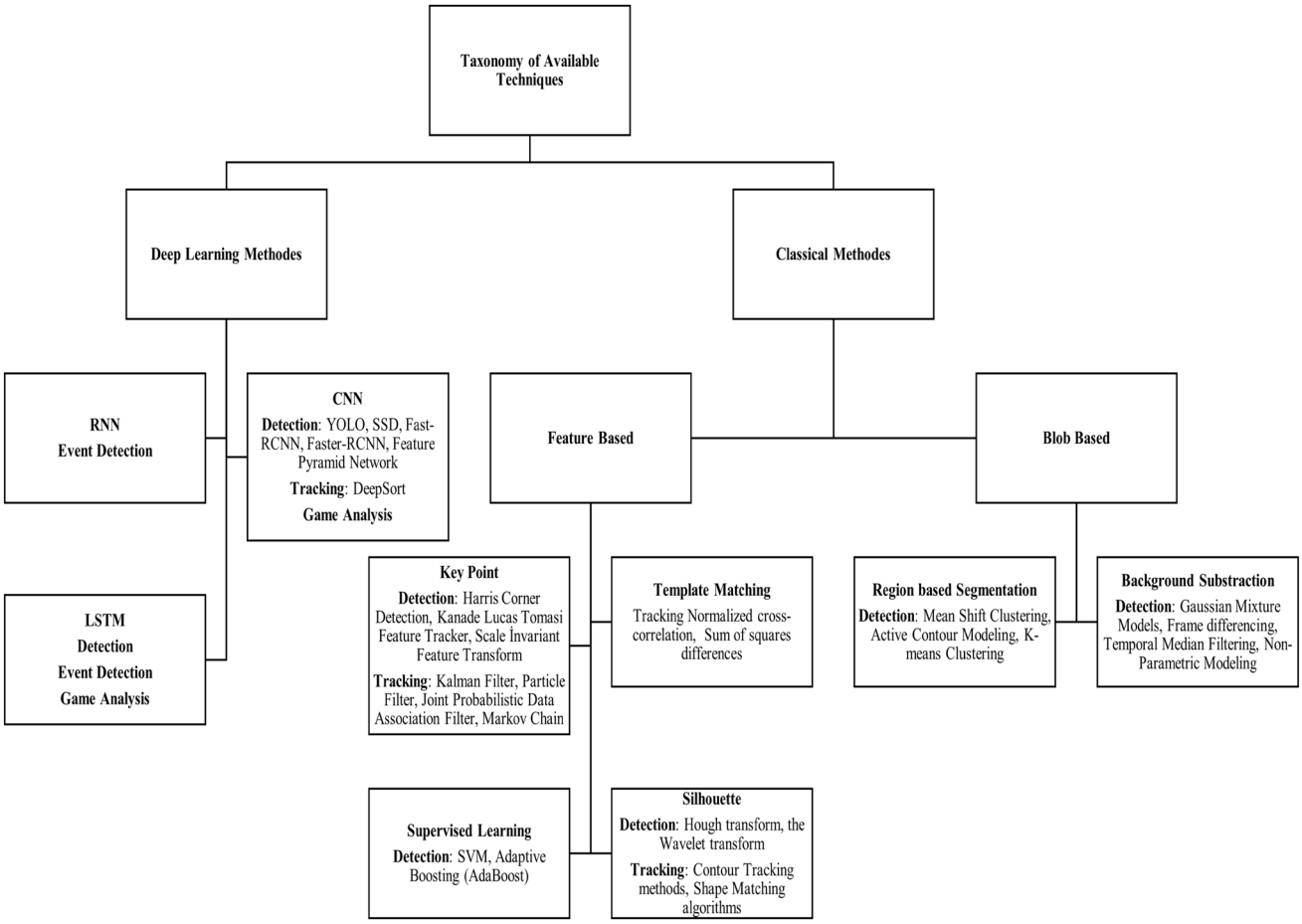


Fig. 8 Diagram of available techniques

and Non-Parametric Modeling (NPM) [30] are some of the most used background subtraction algorithms. On the other hand, region-based segmentation aims to divide video frames into different regions that contain uniform features (for example, color, texture, lighting, etc.) within each frame. Some of them are Mean Shift Clustering [31], Active Contour Modeling [32], and K-Means Clustering [33]. They are mostly used to detect players [22, 23] or balls [16, 34, 35] in soccer videos. The human outline shape plays a very important role in detecting players. However, the same is not true for the ball. Because of its small size, the ball can sometimes be entirely or partially occluded by the player's legs or body when it is under the player's possession. In this case, it is very difficult to detect and track the ball. In such cases, simple ball detection methods with motion-based background subtraction may fail.

4.1.2 Overview of the feature base methods

Feature-based methods make use of features to solve specific problems [36]. They consist of feature descriptors and feature extractors. They are used to manually extract important information from the sports video for further analysis. However, these methods can only extract low-level features [37].

Feature-based methods are generally used for object detection and tracking. It is usually used in soccer videos to indicate the locations of the players and the ball during the match. Based on the information obtained, it is possible to assess questions such as the strong and weak abilities of the players, as well as the role they perform on the playing field. Here, features such as the appearance of the object, its geometric shape, color, edge, and texture are used. These characteristics are employed in approaches such as key point, template matching, silhouette, and supervised learning, as illustrated in Fig. 8.

The search for key points in image texture is the foundation of key point algorithms for object detection. The most popular methods are Harris Corner Detection [38], Kanade Lucas Tomasi Feature Tracker [39], and Scale Invariant Feature Transform [40]. Methods such as Kalman Filter [41], Particle Filter [42], Joint Probabilistic Data Association

Table 2 An overview of the major aspects of the most important football strategy presented in the last decade

Year	References	Source	Tasks	Methods	Performance criteria
2020	[72]	Multi-view 4 K soccer videos	Detection of player and kick movements	Based on an object detection (YOLO)/tracking (MilTrack) technique and pose estimation (OpenPose)	*F*-measure = 0.85
2019	[73]	Broadcast videos	Player detection and tracking	CNN-based detection and histogram-based association for tracking	mAP = 87.43%
2018	[74]	Broadcast videos	possession of the ball	CNN-based detection, DeepSort-based tracking, and histogram-based matching for team identification	Acc = 0.84
2020	[75]	Broadcast videos	Player detection and tracking	Use of feature pyramid network (FPN) and faster-RCNN	MOTA = 92.9
2019	[58]	ISSIA dataset	Player and ball detection and tracking	CNN-based classification by detecting foreground objects using the median filter	Acc = 0.87
2019	[76]	ISSIA dataset	Ball detection and tracking	CNN-based detection and particle filter for tracking	Acc = 0.90
2019	[77]	ISSIA dataset	Ball detection	CNN base	Acc = 0.951
2019	[78]	ISSIA dataset	Player and ball detection	Based on CNN with feature pyramid network (FPN)	mAP = 0.833
2016	[79]	Robocopy humanoid league	Ball detection	CNN-based object detection and categorization (humanoid soccer ball detection)	Pr = 80%
2021	[80]	Broadcast videos	Player tracking	Machine learning and deep reinforcement learning	–
2022	[81]		Player and ball detection and tracking	YOLOv3 and SORT	Acc = 0.93
2022	[82]	ISSIA and SoccerNet	Player tracking	DeepPlayerTrack	Acc = 0.96
2018	[83]	SVSED (broadcast video)	Event detection and scene classification	An end-to-end approach using CNN with deep transfer learning	Acc = 0.89
2018	[84]	Broadcast videos	Event detection	Dividing the video into parts using pixel difference and edge change ratio and highlighting the event using 3D-CNN	Acc = 0.94
2021	[85]	Broadcast videos	Event detection	Use of variational autoencoder for separating football images and CNN for event detection	Pr = 0.90
2021	[86]	SoccerNet	Event detection	Combining visual and audio features based on CNN	Pr = 0.66
2020	[87]	SoccerNet	Event detection	Combining the self-attention mechanism, NetVLAD and innovative loss calculation methods	Acc = 0.74
2020	[88]	SoccerNet	Event detection	use a multi-tower CNN	mAP = 70.1%
2016	[89]	Broadcast videos	Event detection	CNN for fully exploiting features and RNN for dealing with the temporal relation	Pr = 0.92
2021	[90]	SoccerNet	Spotting events	Two-stream CNN Inspired by the model of DilateRNN and LSTM units	mAP = 63%
2019	[91]	RoboCup videos	Ball detection and tracking	Use of three spatio-temporal models: TCN, ConvLSTM and ConvGRU	Acc = 0.96
2019	[63]	UCF101	Event detection and video summarization	By leveraging the spatio-temporal learning capability of 3D-CNN and LSTM	Acc = 0.96

Table 2 (continued)

Year	References	Source	Tasks	Methods	Performance criteria
2019	[92]	Broadcast videos	Event detection	VGG and LSTM	Pr = 0.27
2019	[93]	Broadcast videos	Event detection	Histogram base feature extraction and LSTM	Pr = 0.92
2021	[94]	SoccerNet-V2	Event detection	Based on NetVLAD + +	Avg-mAP = 0.53
2019	[64]	Broadcast videos	Video summarization	Using LSTM based on event and audio features	Pr = 0.47
2020	[95]	F24	Game analysis	Using deep reinforcement learning and LSTM	–
2020	[96]	Broadcast videos	Game analysis	CNN base	–
2021	[97]	Broadcast videos	Game analysis	Convolutional autoencoder	Acc = 0.76
2019	[59]	Broadcast videos	Game analysis	DELM	*F*-measure = 0.87
2020	[98]	Broadcast videos	Scene classification	Pre-trained AlexNet convolutional neural network (CNN)	Acc = 0.99

Filter [43], and Markov Chain [44] are some common key point algorithms for tracking. Objects detected in sequential video frames are associated by establishing relationships between their key points.

Silhouette Detection analyzes video frames by looking for silhouettes and contours similar to a particular model. The relationship between models and silhouettes or contours is based on properties such as shape and density. Hough transform [45] (see Fig. 10) and the Wavelet transform (WT) [46] are some of these methods for object detection. Following the detection of a silhouette, the area covered by the object in each frame is estimated and tracking is performed. Contour Tracking methods [47] and Shape Matching algorithms [48] are the most popular tracking methods.

Supervised Learning focuses on detecting similar objects in each frame by learning the features during the training phase. These methods are used for both object detection and classification. SVM [49] and Adaptive Boosting (AdaBoost) [50] are types of this method.

Template Matching, on the other hand, tracks object by analyzing the changes in shape or appearance of objects caused by rotation or transformation. Normalized cross-correlation [51] and Sum of squares differences [52] are types of this method.

Fig. 9 Example of player detecting with blob-based methods [23]

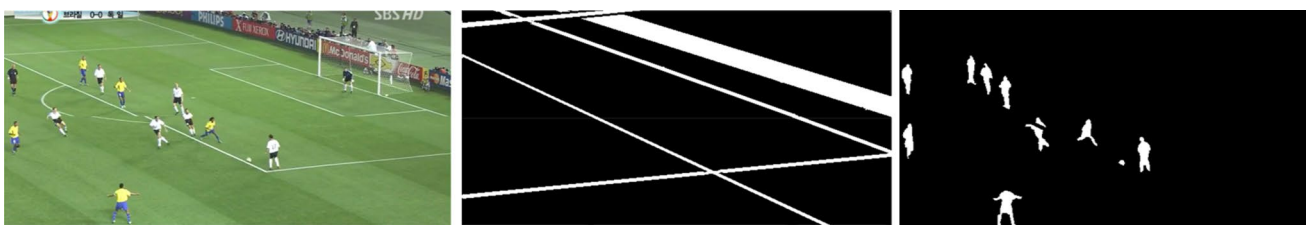


Fig. 10 Example of player detecting with feature-based methods [8]

Feature-based methods are not suitable for real-time processes [16]. Another disadvantage of this method is that the properties obtained can only be applied to certain problems, which is not useful for a real scenario or different datasets or problems. Such methods used for sports video analysis can only extract low-level features and are not very successful at capturing high-level semantic information [53].

4.2 Deep learning

In this section, an overview of the deep learning architecture and its application in soccer video analysis is reviewed based on previous work. In [54], studies that used a deep learning approach in sports video analysis were examined. In soccer videos, deep learning algorithms are used to perform various tasks such as action detection [19, 55, 56], player detecting [55] and tracking [5, 57], ball tracking [58], tactical analysis [59], passing feasibility [60], talent scouting [61], game analysis [62] or highlighting [63, 64].

4.2.1 Overview of the deep learning architecture

There has been a considerable quantity of studies based on traditional approaches for soccer video analysis in recent years. However, these classical methods of feature descriptors and extractors are problem-based. In other words, the extracted features are only applicable to certain problems. In a real scenario, it is not useful for other unrelated datasets or problems. Other than that, classical methods for soccer video analysis can only extract low-level features. They are not particularly effective in capturing high-level semantic information.

The deep learning architecture classifies input images or video frames directly via multiple processing layers for automatically learning and representing data [65]. Unlike classical methods, it does not require any feature descriptors or feature extractors. For example, deep learning architecture uses methods such as local perception, down pooling, weight sharing, and a multi-convolution kernel to automatically learn local features from only a portion of an image instead of a whole image [66].

Deep learning techniques have high-level and complex action recognition and classification capabilities, which have attracted the attention of many researchers [67]. Convolutional neural networks (CNN), recurrent neural networks (RNN), and long short-term memory (LSTM) are examples of extensively used deep learning models.

In analyzing video data, since video sequences vary dynamically over time, researchers encounter several challenges in extracting temporal information. Therefore, RNN and LSTM are preferred. These models can extract temporal information from video data. Some research has been done on the combination of both CNN and LSTM models to extract spatio-temporal information.

Deep learning's superior performance has led to its use in video analysis. Initially, CNN was used independently to extract information from images [68]. However, 2D-CNN cannot extract temporal information from video sequences. To overcome this problem, 3D-CNN is created to extract both spatial and temporal information from video frames [69].

The RNN-based technique successfully captures temporal information, with current predictions based on both current and previous observations [70]. However, RNN architecture has only short-term memory, which makes RNN not applicable in real-world scenarios. The LSTM model has been presented as a solution to this problem. This approach can extract temporal information using sequential video data. The LSTM model includes a memory unit that determines when hidden states should be remembered and when they should be forgotten [71]. Because of its advantages, the LSTM model is frequently employed in computer vision applications.

4.2.2 Use of deep learning in soccer video analysis

In this section, applications in football video analysis based on deep learning architecture are reviewed. Table 2 summarizes the major aspects of the most typical deep learning algorithms published in recent years for different operations such as ball or player detection/tracking, event detection, and game analysis on football video. According to the data in this table, these strategies can be classified based on different criteria. According to the data in the sixth column of the table, the performance of the examined strategies was evaluated according to the following criteria:

- F-measure: is the harmonic mean of precision and recall. The fraction of true positive examples among those that the model classified as positive is known as precision. The percentage of examples out of all positive examples that are categorized as positive is called recall, also known as sensitivity:

$$\text{precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}} \tag{1}$$

$$\text{recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \tag{2}$$

$$F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{3}$$

- Accuracy (Acc): the fraction of correctly classified images to all processed images:

$$\text{Acc} = \frac{\text{true positives} + \text{true negatives}}{\text{true positives} + \text{true negatives} + \text{false positives} + \text{false negatives}} \tag{4}$$

- Precision (Pr): The ratio between the total number of correct positive predicts and all positive predictions.
- Mean average precision (mAP): a standard metric used in the assessment of object detection methods. Mean average precision is the mean of the precisions for each class. AP_i is the average precision of class i and N is the number of classes. The recall and precision at the n-th threshold are, respectively, R_n and P_n. Precision-recall is related to AP, which can be calculated as the area under the curve. AP is related to precision-recall and can be calculated as the area under the curve.

$$\text{AP} = \sum_n \left(R_n - R_{n-1}\right) P_n \tag{5}$$

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^{N} \text{AP}_i \tag{6}$$

- Multiple object tracking accuracy (MOTA): a cumulative metric that combines the number of false positives (FP), false negatives (FN), and ID switches (IDs). gtDet is ground truth detection:

$$\text{MOTA} = 1 - \frac{|\text{FP}| + |\text{FN}| + |\text{IDs}|}{|\text{gtDet}|} \tag{7}$$

- Avg-mAP: the area under the mAP curve.

4.2.2.1 CNN *Player/ball detection and tracking* This section describes methods recommended by CNN in recent years to detect and track players and the ball. These strategies serve as the foundation for the majority of game analysis studies. In order to comprehend each player's activities, it is critical for automated game analysis to recognize kicking motions in soccer videos. In multi-view 4 K football videos, Jianfeng et al. [72] employed YOLOv3 to detect players and the ball in the video frame before detecting kick motions for ball tracking, while they used the Open-Pose method to predict players' poses. They then detected the kick by calculating the shortest Euclidean distance between the perceived ball's center and the players' bodies. They used four separate cameras to examine it from various perspectives.

Another newly proposed approach is detailed in [73], which uses data association logic to track players by analyzing the location and size of bounding boxes generated from a CNN-based detection, as well as a set of probabilities associated with such detections and their color histograms.

In [74], a CNN-based strategy is proposed to determine the placement of players on the playing field and which player controls the ball during the match. Initially, The CNN-based YOLO9000 approach was used to detect football players, and for data augmentation, Deep Convolutional Generative Adversarial Networks (DCGAN) were used. The DeepSort tracking approach provided in [99] is then used to track the detected players, and each player's team is determined using a simple histogram-based matching method. Finally, to identify the player in control of the ball, a trained CNN is used to classify each player as either a "player with the ball" or a "player without the ball".

Detection algorithms in soccer videos sometimes meet challenges, since player sizes are small in comparison to the field and the looks of players from the same team are quite similar. In [75], a deep learning-based soccer player detection and tracking method is presented that can handle various conditions such as small players, different stadiums, teams, and lighting settings. Feature Pyramid Network (FPN) and Faster-RCNN are used to detect players [100]. Since the approach was trained unsupervised, manual explanations were not required.

Kambel et al. [58], proposed a new two-stage method that combined a median filter to automatically detect foreground objects in soccer videos with a CNN-based approach to classify objects in each frame as background, ball, and player. A probabilistic bounding box overlapping approach is also proposed for robust ball track validation.

In [76], a CNN-based deep network architecture is proposed to detect the ball in long shots. This architect's goal is to detect tiny objects and decrease processing time. It was later improved [77] to detect any size ball, as well as [78] to detect players. To detect the ball, [79] presents a CNN-based strategy that focuses on the center of the ball.

An open-source approach is proposed in [80], using only Deep Reinforcement Learning, without human training and guidance, to track and evaluate players. This method has failed to follow the ball at critical moments such as the initial pass and shot, and it has also failed to solve the problem of identity change.

In [81], YOLOv3 and real-time SORT-based ball and player tracking approaches are proposed to track the ball and players. The proposed approach is split into two parts.

YOLOv3 is first used for detecting and classifying the player, the ball, and the background since it removes the detected objects outside the playing field as background. The SORT method is then utilized for tracking, which employs Kalman filtering and bounding box overlap. This methodology effectively handles challenging situations such as partial occlusion for ball and players that reappear after a few frames. It fails, however, when players are significantly occluded.

In [82], a new approach (DeepPlayer-Track) is proposed to track players and referees by representing deep features to keep the tracking identity. A sophisticated multi-player tracker with deep features of player and referee identification is being developed further to give identity-coherent trajectories. The suggested technique is divided into two parts: (1) YOLOv4 detects and classifies players, soccer balls, referees, and backgrounds; (2) The SORT approach is used to connect distinct identities with the player identification (ID) coefficient, which improves tracking performance. The disadvantage of this method is that the identity of the player is changed when the player with the same jersey color is occluded.

Event detection Soccer video scene and event classification are two primary tasks for soccer video analysis. In general, these two tasks are handled separately. Hong et al. [83] proposed an approach for solving two problems at the same time that used deep transfer learning CNN and obtained an accuracy of more than 89%. Transfer learning was used to avoid the model from overfitting. The event classifications are corner, free kick, goal, and penalty, while the scene classifications are the long view and close-up view. They introduced a novel Football Video Scene and Event dataset (SVSED) in this work, consisting of 600 (unintelligible) video clips divided into six groups.

In [84], a 3D-CNN-based approach is proposed for detecting events like corners, shots, goals, and penalty kicks. In most prior studies, two-dimensional CNN was employed. Two-dimensional CNN, however, is only suited for spatial features. It is unable to acquire temporal characteristics. As a result, 3D-CNN is used in this work to include these temporal measurements. First, soccer video event markings are recognized using pixel differencing and edge change ratios. Following this, semantic properties of segmented frames are collected, and CNN was used to classify soccer event types such as Corner, Shoot, Goal Attempt, and Penalty Kick.

In [85], a deep learning approach is proposed to detect events in a football match by detecting the difference between red and yellow card images. The suggested technique for detecting football match events uses a variational autoencoder (VAE) to first determine whether a football square exists. The CNN architecture is then utilized to categorize the events in the next phase. The images of red and yellow cards were then classified using fine-grain image classification.

In [87], a new soccer video event detection algorithm based on the self-attention mechanism is proposed. It retrieves important frames using the self-attention method and then gets time window-level characteristics using the NetVLAD network. Then, each video clip is divided into four categories (goals, red/yellow cards, substitutions, and others). The experimental findings demonstrate that the self-attention mechanism enhanced classification accuracy on the SoccerNet data set from 67.2 to 74.3%.

Since the videos are long and have a high data volume, it is difficult to explain the video frames in detail and determine the specific event in soccer video analysis. To solve this issue, [88] proposed using a multi-tower 1-dimensional CNN architecture for event detection in roughly annotated videos. To acquire the features vector, video frames are fed into a pre-trained 2D-CNN. The 1D convolutional towers are fed into the feature vectors. With the use of varied temporal receptive fields, each tower analyses input feature vectors on various temporal scales. Finally, the activations from the parallel towers are combined to get class probabilities. The approach is evaluated on the SoccerNet dataset. The average success for goal, card, and player substitution events is 70.4% mAP.

For automatic processing in soccer videos, the most important task is to obtain high-level semantic information during the game. For example, detecting the main event of the game allows focusing on the important details of the game. In [94] a new temporally aware learnable pooling module for event detection in a soccer broadcast video is presented. To pool on features, most deep learning networks employ max pooling or average pooling. Unlike previous pooling approaches, in [94], the context is separated into two parts to pool temporal information: the context before the event happens, and the context after the event occurs. They present CNN-based NetVLAD, a new pooling technique that can learn the context in a feature space utilizing a clustering algorithm-based approach. They trained and evaluated the new pooling method on the large-scale SoccerNet-v2 dataset.

Detecting events in videos is a complex process, and many approaches in the literature only consider visual information in videos. In [86], visual models based on CNN are used. Besides the visual features for cards, player substitutions, and goal events, additional audio information is added to evaluate the effect on detection accuracy. Existing state-of-the-art visual models based on the ResNet model were implemented and audited using the SoccerNet dataset. In addition, utilizing Log-Mel spectrograms, an audio-based model was developed, and experiments were conducted with this model in various configurations. Finally, a multimodal approach was used by combining video and audio models

using both feature fusion and averaging/maximizing softmax prediction values from independent models. This study aims to examine the effect on event detection performance when visual and auditory features are combined. Overall, improved performance has been observed, but it also shows differences in benefits from the multimodal approach according to different event types. More specifically, the combination of audio and visual features has proven to be more beneficial for Goal events than for Card and Substitution events.

Game analysis Fernandez and Bornn presented a CNN architecture that can predict the exact probability of potential soccer passes using low-level spatio-temporal soccer data to improve game state analysis in soccer [96].

The goal of the study in [97] is to use deep learning to assess the passing style of the player with higher precision. Pass2vec, a passing style descriptor that may define each player's passing style by combining precise information about Passes, is offered as a solution to this goal. The Convolutional Autoencoder was used to aggregate information on the position, length, and direction of the passes, and then pass vectors were created for each player. In this work, combining Autoencoder and extreme learning machine approaches increases event classification performance.

In [59], a new method based on Deep Extreme Learning Machine (DELM) is proposed to predict team tactics from player formations in soccer videos. The estimate of tactics is divided into two sections. First, the temporary tactic using DELM is estimated. The team tactics are modified in the second phase based on the relationship between the two teams' tactics and ball possession. According to the *F*-measure criteria, it has a success rate of 0.87. However, at the start of the game, he is unable to predict team configuration.

A pre-trained AlexNet Convolutional Neural Network (CNN) is used in [98] to classify sports video scenes (Bowling, Kicking, Boundary, Close-up, and Crowd scenes), especially to summarize sports videos.

4.2.2.2 RNN ***Event detection*** CNN techniques for soccer video event detection have two major drawbacks: a large number of training datasets are required, and temporal features are not considered. The RNN architecture was chosen to tackle this problem. Jiang et al. proposed a deep learning-based automatic event detection method by extracting important highlights from soccer videos via CNN and RNN [89]. They focused on 4 types of events: goal, goal attempt, corner, and card. In this method, first, the large video is divided into small video clips using scene classification and Play-Break rules. The CNN model was used to extract the image features and the RNN structure was used to capture the temporal relationship. However, three different RNN structures (traditional RNN model, LSTM model, and gated repeating unit (GRU) model) were used to determine the best model. Among them, LSTM shows the best performance. However, as the number of categories increases, the classification accuracy decreases.

Unlike a single image or a sequence of several frames, long videos contain rich temporal information. When used correctly, video analytics challenges can profit from this kind of data. As previously stated, RNNs have been effectively used for a variety of video analysis challenges in prior research. However, many of these videos are only a few minutes long. Training RNNs for long videos and modeling long-term dependencies using RNNs has become difficult due to gradient disappearance or explosion in general. To solve this issue, Mahaseni et al. proposed an approach for accurate event localization that considers long-term connections between video frames [90]. A bi-stream CNN extracts Spatio-temporal features which are then used by a Dilated Recurrent Neural Network (DRNN) for event detection. DRNN outperforms ordinary RNNs in event detection by properly using information gained from prior frames in the distant range. The model received a Mean-mAP score of 63.3% after being evaluated on the SoccerNet dataset [101].

4.2.2.3 LSTM ***Player/ball detection and tracking*** In [91], a new convolutional neural network approach is presented to detect the soccer ball in a sequence of images. Unlike current methods, which only use the existing frame or an image for detection, this method uses the past to track the ball efficiently when it is lost or partially occluded in some frames. This method uses Spatial–temporal correlation and detects the ball based on the trajectory of its movements using Temporal convolutional networks (TCN), ConvLSTM, and ConvGRU. TCN's feed-forward nature allows for quicker inference time, making it a great solution for real-time RoboCup soccer applications.

Event detection An event detection method based on 3D-CNN and LSTM network is proposed in [65]. The purpose of this method is to summarize a football match using event detection, which considers five events: corner kicks, free kicks, goal scenes, centerline, and throw-in. 3D-CNN based on ResNet was used to extract features from video clips. The evaluation was made on the UCF101 dataset, and it achieved an accuracy of around 96.81% as a success rate. However, the limited number of events is one of the biggest drawbacks of this study.

Yu et al. designed a deep learning-based soccer event detection involving event detection and story generation using the relationship between events and repetitions in soccer videos [92]. Shooting, corner, free-kick, yellow card, foul, offside, and goal were the seven categories of events. CNN and LSTM were used in this design. At first, reply clips were detected from input videos; then the events were discovered using VGG for extracting features from previous

Fig. 11 Example frames from three views from the ISSIA dataset

episodes from reply clips, and LSTM was used to determine the event type of the repeat clips using the same event detection model.

In [93], a strategy is proposed to automatically detect events such as goals, cards, fouls, and corners. To detect shot limitations, a histogram-based analysis and a linear SVM were utilized first. LSTM was then used to explore long-term temporal correlations in the array of features, and SoftMax was used to classify events.

To automatically summarize soccer video, a recurrent neural network (LSTM) that considers both event and auditory aspects was utilized. Thus, a multimodal deep learning-based algorithm is presented in [64]. Event features provide a shorter and better representation of the match, and audio helps to detect the excitement of the game. However, this is not an ideal approach for long videos such as football matches.

Game analysis Player evaluation, which provides a sense of a player's performance, is an essential task of sports statistics. Performance evaluation is important for team management and fan attendance. Soccer is the most difficult to analyze of all the major team sports. In [95], a novel method for evaluating all soccer movements using match data was proposed. Deep Reinforcement Learning is used to understand the spatio-temporal features. To capture the spatio-temporal complexity of the home team and the opponent team independently, two different LSTMs were employed. This article used the *F*24 soccer game dataset provided by Opta.[2]

4.3 Datasets

This section provides an overview of annotated soccer video datasets. Shared datasets make it possible to compare the performance of different algorithms directly on the same data and increase the transparency of research in the field. Moreover, the extremely time-consuming process of capturing and annotating large amounts of various videos can be reduced. However, the lack of soccer datasets hinders the rapid development of the soccer video analysis field.

The ISSIA Dataset [102] contains a video from a football match captured with 6 cameras, three on each side of the stadium, recording at 25 fps. Example frames from each side of the field are shown in Fig. 11. The positions of the players, referees, and the ball were manually noted from each camera. There are calibration images and measurements available to calibrate each camera to a common-world coordinate system. The ISSIA Soccer dataset includes six synchronized, long-shot views of the football pitch that were captured by six Full-HD DALSA 25-2M30 cameras. The dataset contains 20,000 annotated frames, each of which is tagged with information about the position of the ball and the bounding boxes of the players. This dataset is comprised of six different sequences in total. Many studies made use of four of them for training purposes, while the remaining two were employed for testing. They also use data augmentation so that they can increase the variety of instances used for training and reduce the likelihood of the model overfitting.

SSET dataset [103] was created to meet the shot segmentation, soccer event detection, and player tracking research needs. In total, 350 soccer videos were collected, including 282 h of various soccer games. The data set consists of three parts: (1) 5 shot types; (2) 11 different types of events; (3) the coordinates, width, and length of the players' bounding box and bounding boxes.

SoccerNet [19] is the most comprehensive publicly available soccer dataset for event detection. The dataset comprises 500 football matches from six major European leagues, spanning three seasons from 2014 to 2017, and totaling 764 h in length. For the three primary event classes—Goal, Yellow/Red Card, and Substitution—a total of 6,637 temporal assertions were automatically extracted from online match reports at one-minute resolution. Then, to better understand the broadcast videos, the SoccerNet dataset was developed. With 300 k annotations, three computer vision tasks (action identification, camera shot segmentation, and boundary detection), and several benchmark results, SoccerNet-v2 is the most comprehensive dataset for soccer video interpretation and production [56]. In several studies, the training set consisted of 300 matches, the

[2] https://www.statsperform.com/opta.

validation set consisted of 100 matches, and the testing set consisted of 100 matches.

SoccerDB [104] is created by combining 270 games from the SoccerNet dataset and 76 soccer games from the Chinese Super League. It has been proposed for use in a variety of applications including object detection, action recognition, temporal action localization, and highlight detection.

5 Discussion

The worldwide popularity of football among team sports is precisely due to its simplicity. It is very easy to enjoy watching the game because grasping the basic rules is relatively simple compared to many other sports. For this reason, football video analysis has attracted the attention of many researchers in recent years. Available researches are generally based on classical methods or deep learning-based methods. This paper considers deep learning-based research on soccer videos.

Over the past few decades, various research has focused on sports video analysis based on blob-based and feature-based methods. The main purpose of blob-based methods is to detect candidate object regions and reduce the number of false alarms. These methods are mostly based on background subtraction or region-based segmentation. Therefore, it has an important role in player detection. However, the same is not true for the ball. Sometimes, when the ball is in the player's control, it can be completely or partially covered by the players. In such cases, blob-based ball detection methods may fail. Feature-based methods relate to features designed for specific problems. These consist of feature descriptors and feature extractors and are used to manually extract important information from football videos for further analysis. In general, feature-based methods are used for object detection and tracking. It is mostly applied to the players and the ball in football videos and is used to indicate their positions during the game. Methods based on features are not appropriate for real-time activities. Another drawback of this method is that the properties obtained can only be applied to certain problems, which does not make it useful in a real-world scenario, different datasets, or problems. Such methods used for soccer video analysis can only extract low-level features and are not very successful in capturing high-level semantic information.

The advancement in technology has led to the emergence of deep learning-based sports video analysis. Deep learning-based soccer video analysis is still a new topic that is open to research. The major drawback of using deep learning models is the need for many data to automatically learn features from inputs. Previously, the CNN model, which is one of the deep learning approaches, has shown successful performance in many areas. CNN has been used to extract information from images. However, it cannot extract temporal information in video sequences. Obtaining temporal information is quite challenging. As a result, the usage of RNN and LSTM algorithms was required. These models can extract temporal information from video data.

Although deep learning-based architecture works well in many domains, such as image classification, the predicted advancement in others, such as video classification or sports video analysis, has not been reached. As a result, it remains open for deep learning-based study, which many researchers are attempting to tackle.

6 Future directions

This section aims to give the reader some suggestions for potential future developments in the methods and uses of video soccer analysis. The integration of strategies to detect players, the ball, and events in the strategies for analyzing the game is becoming more and more prevalent, as can be seen throughout the article. The observed tendency, therefore, suggests that interest in applications that can identify events in real time will increase since many game analysis applications require real-time performance even though event detection is often performed offline. In addition, a clear trend has been noticed in the creation of applications utilizing deep learning techniques, as well as ball and player detection and tracking applications, event detection tactics, and in game analysis applications. Another area for research is the re-identification of players using deep learning techniques to track them more precisely. Due to recent advancements in increasing parallel processing capacity, deep learning cannot only provide good outcomes in many scenarios but also execute nearly in real time. The observed trends also point to an increasing interest in the development of game analysis-based applications and, more specifically, in tactical analysis-based strategies. Most of these tactics require input data consisting of players' positions at various points in matches. Because of the difficulty in overcoming common player detection and tracking issues such as occlusions with only the cameras used to broadcast the match, it may be preferable to gather location data using commercial tools based on some supplementary video cameras or wearable devices. However, the cost of these commercial tools is usually high. Therefore, it makes sense that broadcast signal-based detection and tracking techniques will continue to be proposed. Manufacturers are often hesitant to include sensors in ball detection and tracking systems. Therefore, vision-based approaches will still be required.

7 Conclusion

The most significant deep learning-based algorithms for analyzing soccer video sequences published in recent years are investigated in this paper. Existing methods are divided into two groups: classical methods and methods based on deep learning. This paper presents a comprehensive review of soccer video analysis by comparing both classical methods and the deep learning approach. The techniques used in general are grouped into three categories according to their purposes and needs, such as a ball or player detection/tracking, event detection, and game analysis. In summary, the deep learning approach overcome the drawbacks of classical approaches for analyzing soccer videos and performed better. Therefore, this approach will be used in the future by researchers, coaches, and sports professionals in game analysis and evaluation of players' performances to extract high-level semantic information.

Acknowledgements The authors declare no conflict of interest in this study.

Author contributions S.A wrote the main manuscript text.S.V Proposed the idea and reviewed the manuscrip.

Funding We confirm that we do not have a funding source.

Data availability The datasets analyzed during the current study are available in the following public domain resources: http://pspagnolo.jimdo.com/download/http://media.hust.edu.cn/dataset.htmhttps://www.soccer-net.org/data https://github.com/newsdata/SoccerDB.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. D'Orazio, T., Leo, M.: A review of vision-based systems for soccer video analysis. Pattern Recognit. **43**(8), 2911–2926 (2010). https://doi.org/10.1016/J.PATCOG.2010.03.009
2. Al-Ali, A., Almaadeed, S.: A review on soccer player tracking techniques based on extracted features. In: 2017 6th Int. Conf. Inf. Commun. Technol. Accessbility, ICTA 2017, vol. 2017(December), pp. 1–6 (Apr. 2018). https://doi.org/10.1109/ICTA.2017.8336015
3. Kamble, P.R., Keskar, A.G., Bhurchandi, K.M.: Ball tracking in sports: a survey. Artif. Intell. Rev. **52**(3), 1655–1705 (2019). https://doi.org/10.1007/S10462-017-9582-2/FIGURES/10
4. Khan, Y.S., Pawar, S.: Video summarization: survey on event detection and summarization in soccer videos. In: IJACSA) Int. J. Adv. Comput. Sci. Appl., vol. 6(11) (2015). Accessed: 26 Feb. 2022 (Online). Available: www.ijacsa.thesai.org
5. Manafifard, M., Ebadi, H., Abrishami Moghaddam, H.: A survey on player tracking in soccer videos. Comput. Vis. Image Underst. **159**, 19–46 (2017). https://doi.org/10.1016/J.CVIU.2017.02.002
6. Memmert, D., Koen, A., Lemmink, P.M., Sampaio, J.: Current approaches to tactical performance analyses in soccer using position data (2016). https://doi.org/10.1007/s40279-016-0562-5
7. Rehman, A., Saba, T.: Features extraction for soccer video semantic analysis: current achievements and remaining issues. Artif. Intell. Rev. 2012 413 **41**(3), 451–461 (2012). https://doi.org/10.1007/S10462-012-9319-1
8. Cuevas, C., Quilón, D., García, N.: Techniques and applications for soccer video analysis: a survey. Multimed. Tools Appl. **79**(39–40), 29685–29721 (2020). https://doi.org/10.1007/S11042-020-09409-0/FIGURES/16
9. Cuevas, C., Martínez, R., García, N.: Detection of stationary foreground objects: a survey. Comput. Vis. Image Underst. **152**, 41–57 (2016). https://doi.org/10.1016/J.CVIU.2016.07.001
10. Ishii, N., Kitahara, I., Kameda, Y., Ohta, Y.: 3D tracking of a soccer ball using two synchronized cameras. Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 4810 LNCS, pp. 196–205 (Dec. 2007). https://doi.org/10.1007/978-3-540-77255-2_22
11. Scheuer, C., et al.: Generating ball trajectory in soccer video sequences. In: Phys. Educ. Sport Child. Youth with Spec. Needs Res.—Best Pract.—Situat., pp. 343–354 (2006)
12. Barros, R.M.L., et al.: Analysis of the distances covered by first division brazilian soccer players obtained with an automatic tracking method. J. Sports Sci. Med. **6**(2), 233 (June 2007). Accessed: 17 Jan. 2022 (Online). Available: https://www.pmc/articles/PMC3786245/
13. Yu, X., Sen Hay, T., Yan, X., Chng, E.: A player-possession acquisition system for broadcast soccer video. IEEE Int. Conf. Multimed. Expo, ICME 2005, vol. 2005, pp. 522–525 (2005). https://doi.org/10.1109/ICME.2005.1521475
14. Hashimoto, S., Ozawa, S.: A system for automatic judgment of offsides in soccer games. In: 2006 IEEE Int. Conf. Multimed. Expo, ICME 2006—Proc., vol. 2006, pp. 1889–1892 (2006). https://doi.org/10.1109/ICME.2006.262924
15. Gerke, S., Linnemann, A., Müller, K.: Soccer player recognition using spatial constellation features and jersey number recognition. Comput. Vis. Image Underst. **159**, 105–115 (2017). https://doi.org/10.1016/J.CVIU.2017.04.010
16. Halbinger, J., Metzler, J.: Video-based soccer ball detection in difficult situations. Commun. Comput. Inf. Sci. **464**, 17–24 (2013). https://doi.org/10.1007/978-3-319-17548-5_2
17. Wang, X., Turetken, E., Fleuret, F., Fua, P.: Tracking interacting objects using intertwined flows. IEEE Trans. Pattern Anal. Mach. Intell. **38**(11), 2312–2326 (2016). https://doi.org/10.1109/TPAMI.2015.2513406
18. Maksai, A., Wang, X., Fua, P.: What players do with the ball: a physically constrained interaction modeling, pp. 972–981 (2016)
19. Giancola, S., Amine, M., Dghaily, T., Ghanem, B.: Soccernet: a scalable dataset for action spotting in soccer videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1711–1721 (2018)
20. Hossein-Khani, J., Soltanian-Zadeh, H., Kamarei, M., Staadt, O.: Ball detection with the aim of corner event detection in soccer video. In: Proc.—9th IEEE Int. Symp. Parallel Distrib. Process. with Appl. Work. ISPAW 2011—ICASE 2011, SGH 2011, GSDP 2011, pp. 147–152 (2011). https://doi.org/10.1109/ISPAW.2011.41
21. Shih, H.C.: A survey of content-aware video analysis for sports. IEEE Trans. Circuits Syst. Video Technol. **28**(5), 1212–1231 (2018). https://doi.org/10.1109/TCSVT.2017.2655624
22. D'Orazio, T., et al.: An investigation into the feasibility of real-time soccer offside detection from a multiple camera system. IEEE Trans. Circuits Syst. Video Technol. **19**(12), 1804–1818 (2009). https://doi.org/10.1109/TCSVT.2009.2026817

23. Manafifard, M., Ebadi, H., Moghaddam, H.A.: Multi-player detection in soccer broadcast videos using a blob-guided particle swarm optimization method. Multimed. Tools Appl. **76**(10), 12251–12280 (2017). https://doi.org/10.1007/S11042-016-3625-6/TABLES/2
24. Ivankovic, Z., Rackovic, M., Ivkovic, M.: Automatic player position detection in basketball games. Multimed. Tools Appl. **72**(3), 2741–2767 (2014). https://doi.org/10.1007/S11042-013-1580-Z/TABLES/5
25. Ma'ckowiak, S.M.: Segmentation of football video broadcast. INTL J. Electron. Telecommun. **59**(1), 75–84 (2013). https://doi.org/10.2478/eletel-2013-0009
26. Turaga, P., Chellappa, R., Veeraraghavan, A.: Advances in video-based human activity analysis: challenges and approaches. Adv. Comput. **80**(C), 237–290 (2010). https://doi.org/10.1016/S0065-2458(10)80007-5
27. Cuevas, C., García, N., Salgado, L.: A new strategy based on adaptive mixture of Gaussians for real-time moving objects segmentation. Real Time Image Process. **6811**, 304–315 (2008). https://doi.org/10.1117/12.768139
28. Www, W., Patel, N.: International journal of emerging technology and advanced engineering motion detection based on multi frame video under surveillance system, vol. 2(1) (2012). Accessed: 16 Dec. 2021 (Online). Available: www.ijetae.com
29. Arce, G.R.: Nonlinear Signal Processing: A Statistical Approach. Wiley, London (2005)
30. Berjón, D., Cuevas, C., Morán, F., García, N.: Real-time nonparametric background subtraction with tracking-based foreground update. Pattern Recognit. **74**, 156–170 (2018). https://doi.org/10.1016/J.PATCOG.2017.09.009
31. Cheng, Y.: Mean shift, mode seeking, and clustering. IEEE Trans. Pattern Anal. Mach. Intell. **17**(8), 790–799 (1995). https://doi.org/10.1109/34.400568
32. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: active contour models. Int. J. Comput. Vis. 1987 14 **1**(4), 321–331 (1988). https://doi.org/10.1007/BF00133570
33. Hartigan, J.A., Wong, M.A.: Algorithm AS 136: a *K*-means clustering algorithm. Appl. Stat. **28**(1), 100 (1979). https://doi.org/10.2307/2346830
34. Rao, U.M., Pati, U.C.: A novel algorithm for detection of soccer ball and player. 2015 Int. Conf. Commun. Signal Process. ICCSP **2015**, 344–348 (2015). https://doi.org/10.1109/ICCSP.2015.7322903
35. Kia, M.: Ball automatic detection and tracking in long shot views. IJCSNS Int. J. Comput. Sci. Netw. Secur. **16**(6), 1 (2016)
36. Yang, H., et al.: Asymmetric 3D convolutional neural networks for action recognition. Pattern Recognit. **85**, 1–12 (2019). https://doi.org/10.1016/J.PATCOG.2018.07.028
37. Huiqun, Z., Hui, W., Xiaoling, W.: Application research of video annotation in sports video analysis. In: Proc. 2011 Int. Conf. Futur. Comput. Sci. Educ. ICFCSE 2011, pp. 62–66 (2011). https://doi.org/10.1109/ICFCSE.2011.24
38. Harris, C., Stephens, M.: A combined corner and edge detector
39. Lucas, B.D., Kanade, T., et al.: An iterative image registration technique with an application to stereo vision (1981)
40. Lowe, D.G.: Object recognition from local scale-invariant features. Proc. IEEE Int. Conf. Comput. Vis. **2**, 1150–1157 (1999). https://doi.org/10.1109/ICCV.1999.790410
41. Julier, S.J., Uhlmann, J.K.: New extension of the Kalman filter to nonlinear systems. Signal Process. Sens. Fusion Target Recogn. **3068**, 182–193 (1997). https://doi.org/10.1117/12.280797
42. Nieto, M., Cuevas, C., Salgado, L.: Measurement-based reclustering for multiple object tracking with particle filters. In: Proc.—Int. Conf. Image Process. ICIP, pp. 4097–4100 (2009). https://doi.org/10.1109/ICIP.2009.5413709
43. Habtemariam, B., Tharmarasa, R., Thayaparan, T., Mallick, M., Kirubarajan, T.: A multiple-detection joint probabilistic data association filter. IEEE J. Sel. Top. Signal Process. **7**(3), 461–471 (2013). https://doi.org/10.1109/JSTSP.2013.2256772
44. Oh, S., Russell, S., Sastry, S.: Markov chain Monte Carlo data association for multi-target tracking. IEEE Trans. Automat. Contr. **54**(3), 481–497 (2009). https://doi.org/10.1109/TAC.2009.2012975
45. Illingworth, J., Kittler, J.: A survey of the hough transform. Comput. Vis. Graph. Image Process. **44**(1), 87–116 (1988). https://doi.org/10.1016/S0734-189X(88)80033-1
46. Daubechies, I.: The wavelet transform, time–frequency localization and signal analysis. IEEE Trans. Inf. Theory **36**(5), 961–1005 (1990). https://doi.org/10.1109/18.57199
47. Athanesious, J., Suresh, P.: Implementation and comparison of kernel and silhouette based object tracking. Int. J. Adv. Res. Comput. Eng. Technol. **2**(3), 1298–1303 (2013)
48. Athanesious, J.J., Suresh, P.: Systematic survey on object tracking methods in video. Int. J. Adv. Res. Comput. Eng. Technol. **1**(8), 242–247 (2012)
49. Cortes, C., Vapnik, V., Saitta, L.: Support-vector networks. Mach. Learn. 1995 203 **20**(3), 273–297 (1995). https://doi.org/10.1007/BF00994018
50. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. **55**(1), 119–139 (1997). https://doi.org/10.1006/JCSS.1997.1504
51. Broder, J.: The Fourier Transform and Mis Applications. McGraw-Hill, New York (1992)
52. Powell, M.J.D.: A method for minimizing a sum of squares of non-linear functions without calculating derivatives. Comput. J. **7**(4), 303–307 (1965). https://doi.org/10.1093/COMJNL/7.4.303
53. Lee, J., Nam, D.W., Lee, J.S., Moon, S., Kim, K., Kim, H.: A study on composition of context-based soccer analysis system. Int. Conf. Adv. Commun. Technol. ICACT (2017). https://doi.org/10.23919/ICACT.2017.7890222
54. Rangasamy, K., As'ari, M.A., Rahmad, N.A., Ghazali, N.F., Ismail, S.: Deep learning in sport video analysis: a review. Telkomnika Telecommun. Comput. Electron. Control. **18**(4), 1926–1933 (2020). https://doi.org/10.12928/TELKOMNIKA.V18I4.14730
55. Cioppa, A., Deliege, A., Huda, N.U., Gade, R., Van Droogenbroeck, M., Moeslund, T.B.: Multimodal and multiview distillation for real-time player detection on a football field. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 880–881 (2020)
56. Deliege, A., et al.: 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, pp. 4508–4519 (2021)
57. Hurault, S., Ballester, C., Haro, G.: Self-supervised small soccer player detection and tracking. In: MMSports 2020—Proc. 3rd Int. Work. Multimed. Content Anal. Sport., pp. 9–18 (Nov. 2020). https://doi.org/10.1145/3422844.3423054
58. Kamble, P.R., Keskar, A.G., Bhurchandi, K.M.: A deep learning ball tracking system in soccer videos. Opto-Electronics Rev. **27**(1), 58–69 (2019). https://doi.org/10.1016/J.OPELRE.2019.02.003
59. Suzuki, G., Takahashi, S., Ogawa, T., Haseyama, M.: Team tactics estimation in soccer videos based on a deep extreme learning machine and characteristics of the tactics. IEEE Access **7**, 153238–153248 (2019). https://doi.org/10.1109/ACCESS.2019.2946378
60. Arbues-Sanguesa, A., Martin, A., Fernández, J., Ballester, C., Haro, G.: Using player's body-orientation to model pass feasibility in soccer. In: Proceedings of the IEEE/CVF Conference

on Computer Vision and Pattern Recognition Workshops, pp. 886–887 (2020)
61. Decroos, T., Van Haaren, J., Bransen, L., Davis, J.: Actions speak louder than goals: valuing player actions in soccer. In: Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., pp. 1851–1861 (July 2019). https://doi.org/10.1145/3292500.3330758
62. Cioppa, A., Deliege, A., Van Droogenbroeck, M.: A bottom-up approach based on semantics for the interpretation of the main camera stream in soccer games, pp. 1765–1774 (2018)
63. Agyeman, R., Muhammad, R., Choi, G.S.: Soccer video summarization using deep learning. In: Proc.—2nd Int. Conf. Multimed. Inf. Process. Retrieval, MIPR 2019, pp. 270–273 (Apr. 2019). https://doi.org/10.1109/MIPR.2019.00055
64. Sanabria, M., Precioso, S.F., Menguy, T.: A deep architecture for multimodal summarization of soccer games, pp. 16–24 (2019). https://doi.org/10.1145/3347318.3355524
65. Voulodimos, A., Doulamis, N., Doulamis, A., Protopapadakis, E.: Deep learning for computer vision: a brief review. Comput. Intell. Neurosci. **2018**, 1 (2018). https://doi.org/10.1155/2018/7068349
66. Sargano, A.B., Angelov, P., Habib, Z.: A comprehensive review on handcrafted and learning-based action representation approaches for human activity recognition. Appl. Sci. **7**(1), 110 (2017). https://doi.org/10.3390/APP7010110
67. Elboushaki, A., Hannane, R., Afdel, K., Koutti, L.: MultiD-CNN: a multi-dimensional feature learning approach based on deep convolutional networks for gesture recognition in RGB-D image sequences. Expert Syst. Appl. **139**, 112829 (2020). https://doi.org/10.1016/J.ESWA.2019.112829
68. Meng, B., Liu, X.J., Wang, X.: Human action recognition based on quaternion spatial–temporal convolutional neural network and LSTM in RGB videos. Multimed. Tools Appl. **77**(20), 26901–26918 (2018). https://doi.org/10.1007/S11042-018-5893-9/TABLES/4
69. Asadi-Aghbolaghi, M., et al.: A survey on deep learning based approaches for action and gesture recognition in image sequences. In: 2017 12th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2017), pp. 476–483 (2017)
70. Yang, X., Molchanov, P., Kautz, J.: Multilayer and multimodal fusion of deep neural networks for video classification. In: MM 2016—Proc. 2016 ACM Multimed. Conf., pp. 978–987 (Oct. 2016). https://doi.org/10.1145/2964284.2964297
71. Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., Toderici, G.: Beyond short snippets: deep networks for video classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4694–4702 (2015)
72. Xu, J., Tasaka, K.: [Papers] Keep your eye on the ball: detection of kicking motions in multi-view 4K soccer videos. ITE Trans. Media Technol. Appl. **8**(2), 81–88 (2020). https://doi.org/10.3169/MTA.8.81
73. Sverrisson, S., Grancharov, V., Pobloth, H.: Real-time tracking-by-detection in broadcast sports videos. Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 11482 LNCS, pp. 399–411 (June 2019). https://doi.org/10.1007/978-3-030-20205-7_33
74. Theagarajan, R., Pala, F., Zhang, X., Bhanu, B.: Soccer: Who has the ball? Generating visual analytics and player statistics. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1749–1757 (2018)
75. Hurault, S., Ballester, C., Haro, G.: Self-supervised small soccer player detection and tracking. In: MMSports 2020—Proc. 3rd Int. Work. Multimed. Content Anal. Sport., pp. 9–18 (Oct. 2020). https://doi.org/10.1145/3422844.3423054
76. Komorowski, J., Kurzejamski, G., Sarwas, G.: BallTrack: football ball tracking for real-time CCTV systems. In: Proc. 16th Int. Conf. Mach. Vis. Appl. MVA 2019 (May 2019). https://doi.org/10.23919/MVA.2019.8757880
77. Komorowski, J., Kurzejamski, G., Sarwas, G.: DeepBall: deep neural-network ball detector. In: VISIGRAPP 2019—Proc. 14th Int. Jt. Conf. Comput. Vision, Imaging Comput. Graph. Theory Appl., vol. 5, pp. 297–304 (Feb. 2019). https://doi.org/10.5220/0007348902970304
78. Komorowski, J., Kurzejamski, G., Sarwas, G.: FootAndBall: integrated player and ball detector. In: VISIGRAPP 2020—Proc. 15th Int. Jt. Conf. Comput. Vision, Imaging Comput. Graph. Theory Appl., vol. 5, pp. 47–56 (Dec. 2019). https://doi.org/10.5220/0008916000470056
79. Speck, D., Barros, P., Weber, C., Wermter, S.: Ball localization for robocup soccer using convolutional neural networks. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 9776 LNAI, pp. 19–30 (2016). https://doi.org/10.1007/978-3-319-68792-6_2
80. Garnier, P., Gregoir, T.: Evaluating soccer player: from live camera to deep reinforcement learning. Preprint arXiv:2101.05388 (2021)
81. Naik, B.T., Hashmi, M.F.: YOLOv3-SORT: detection and tracking player/ball in soccer sport. J. Electron. Imaging **32**(1), 11003 (2022)
82. Naik, B.T., Hashmi, M.F., Geem, Z.W., Bokde, N.D.: DeepPlayer-track: player and referee tracking with jersey color recognition in soccer. IEEE Access **1**, 1 (2022)
83. Hong, Y., Ling, C., Ye, Z.: End-to-end soccer video scene and event classification with deep transfer learning. In: 2018 Int. Conf. Intell. Syst. Comput. Vision, ISCV 2018, vol. 2018(May), pp. 1–4 (May 2018). https://doi.org/10.1109/ISACV.2018.8369043
84. Khan, M.Z., Saleem, S., Hassan, M.A., Khan, M.U.G.: Learning deep C3D features for soccer video event detection. In: 2018 14th Int. Conf. Emerg. Technol. ICET 2018 (Jan. 2019). https://doi.org/10.1109/ICET.2018.8603644
85. Karimi, A., Toosi, R., Akhaee, M.A.: Soccer event detection using deep learning (Feb. 2021). Accessed: 26 Jan. 2022 (Online). Available: https://arxiv.org/abs/2102.04331v1
86. Andre Nergård Rongved, O., et al.: Automated event detection and classification in soccer: the potential of using multiple modalities. Mach. Learn. Knowl. Extr. **3**(4), 1030–1054 (2021). https://doi.org/10.3390/MAKE3040051
87. Ma, S., Shao, E., Xie, X., Liu, W.: Event detection in soccer video based on self-attention. 2020 IEEE 6th Int. Conf. Comput. Commun. ICCC **2020**, 1852–1856 (2020). https://doi.org/10.1109/ICCC51575.2020.9344896
88. Vats, K., Fani, M., Walters, P., Clausi, D.A., Zelek, J.: Event detection in coarsely annotated sports videos via parallel multi receptive field 1D convolutions
89. Jiang, H., Lu, Y., Xue, J.: Automatic soccer video event detection based on a deep neural network combined CNN and RNN, pp. 490–494 (Jan. 2017). https://doi.org/10.1109/ICTAI.2016.0081
90. Mahaseni, B., Faizal, E.R.M., Raj, R.G.: Spotting football events using two-stream convolutional neural network and dilated recurrent neural network. IEEE Access **9**, 61929–61942 (2021). https://doi.org/10.1109/ACCESS.2021.3074831
91. Kukleva, A., Khan, M.A., Farazi, H., Behnke, S.: Utilizing temporal information in deep convolutional network for efficient soccer ball detection and tracking. In: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 11531 LNAI, pp. 112–125 (July 2019). https://doi.org/10.1007/978-3-030-35699-6_9

92. Yu, J., Lei, A., Hu, Y.: Soccer video event detection based on deep learning. In: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 11296 LNCS, pp. 377–389 (Jan. 2019). https://doi.org/10.1007/978-3-030-05716-9_31
93. Fakhar, B., Rashidy Kanan, H., Behrad, A.: Event detection in soccer videos using unsupervised learning of spatio-temporal features based on pooled spatial pyramid model. Multimed. Tools Appl. **78**(12), 16995–17025 (2019). https://doi.org/10.1007/S11042-018-7083-1/TABLES/12
94. Giancola, S., Ghanem, B.: Temporally-aware feature pooling for action spotting in soccer broadcasts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4490–4499 (2021)
95. Liu, G., Luo, Y., Schulte, O., Kharrat, T.: Deep soccer analytics: learning an action-value function for evaluating soccer players. Data Min. Knowl. Discov. **34**(5), 1531–1559 (2020). https://doi.org/10.1007/S10618-020-00705-9/FIGURES/9
96. Fernández, J., Bornn, L.: SoccerMap: a deep learning architecture for visually-interpretable analysis in soccer. In: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 12461 LNAI, pp. 491–506 (Oct. 2020). https://doi.org/10.1007/978-3-030-67670-4_30
97. Cho, H., Ryu, H., Song, M.: Pass2vec: analyzing soccer players' passing style using deep learning. Int. J. Sport. Sci. Coach. **2021**, 17479541211033078 (2021)
98. Rafiq, M., Rafiq, G., Agyeman, R., Il Jin, S., Choi, G.S.: Scene classification for sports video summarization using transfer learning. Sensors **20**(6), 1702 (2020). https://doi.org/10.3390/S20061702
99. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: Proc.—Int. Conf. Image Process. ICIP, vol. 2017(September), pp. 3645–3649 (Feb. 2018). https://doi.org/10.1109/ICIP.2017.8296962
100. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. Adv. Neural Inf. Process. Syst. (2015). Accessed: 27 Sept. 2022 (Online). Available: https://github.com/
101. Giancola, S., Amine, M., Dghaily, T., Ghanem, B.: SoccerNet: a scalable dataset for action spotting in soccer videos, pp. 1711–1721 (2018). Accessed: 27 Sept. 2022 (Online). Available: https://silviogiancola.github.io/SoccerNet
102. D'Orazio, T., Leo, M., Mosca, N., Spagnolo, P., Mazzeo, P.L.: A semi-automatic system for ground truth generation of soccer video sequences. In: 6th IEEE Int. Conf. Adv. Video Signal Based Surveillance, AVSS 2009, pp. 559–564 (2009). https://doi.org/10.1109/AVSS.2009.69
103. Feng, N., et al.: SSET: a dataset for shot segmentation, event detection, player tracking in soccer videos. Multimed. Tools Appl. **79**(39–40), 28971–28992 (2020). https://doi.org/10.1007/S11042-020-09414-3/TABLES/13
104. Jiang, Y., Cui, K., Chen, L., Wang, C., Xu, C.: SoccerDB: a large-scale database for comprehensive video understanding. In: MMSports 2020—Proc. 3rd Int. Work. Multimed. Content Anal. Sport., pp. 1–8 (Oct. 2020). https://doi.org/10.1145/3422844.3423051