Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-20390-w

Automated soccer event detection and highlight generation for short and long views

Maira Afzal[1] · Jamal Hussain Shah[1] · Saeed ur Rehman[1] · Fahad Ahmed Khokhar[2] · Mussarat Yasmin[1] · Seifedine Kadry[3,4]

Received: 1 March 2024 / Revised: 14 September 2024 / Accepted: 10 October 2024

Abstract

The computer vision field has wide applications in various areas, including sports. Almost all sports events have been exploiting the best features. Sports videos are structure-based, and due to this characteristic, these videos can be categorized into interesting and non-interesting events. Identifying the view of the video and separating important events from non-interesting events is challenging. However, correct view detection can lead to correct event detection. Various researchers have proposed many strategies for detecting events in sports videos. Significant research shows some gaps while generating highlights due to limited work available in the long or short view; there is still a need for some automated methods. The main purpose of this research work is to detect key events by normalizing the data, extracting features, fusing those features, and classifying them. In this research, events are detected by classification. A new dataset is created for research purposes. The benchmark dataset is divided into two subsets, which can be used separately or as part of a larger dataset. A proposed novel approach for event highlight generation in long and short view is presented using a fusion of AlexNet and VGGNet architectures to explore the model's efficiency in the context of accurate highlight generation deep learning models fused with handcrafted. AlexNet and VGG16 pre-trained deep CNN models are applied to extract deep prior features, which are then combined with HOG to improve the results. The proposed methodology undergoes evaluation on a dataset that we created, resulting in an accuracy of 99.6%.

Keywords Event detection (ED) · Soccer · Classification · LOMO · Automatic highlight generation (AHG) · Feature fusion · Feature extraction · Deep prior features · Deep HOG features

✉ Fahad Ahmed Khokhar
fahadahmed.khokhar@unifi.it

[1] Department of Computer Science, COMSATS University Islamabad, Wah Campus, Rawalpindi, Pakistan

[2] Department of Mathematics and Information, University of Florence, Florence, Italy

[3] Department of Computer Science and Mathematics, Lebanese American University, Beirut, Lebanon

[4] Department of Applied Data Science, Noroff University College, Kristiansand, Norway

Published online: 31 October 2024

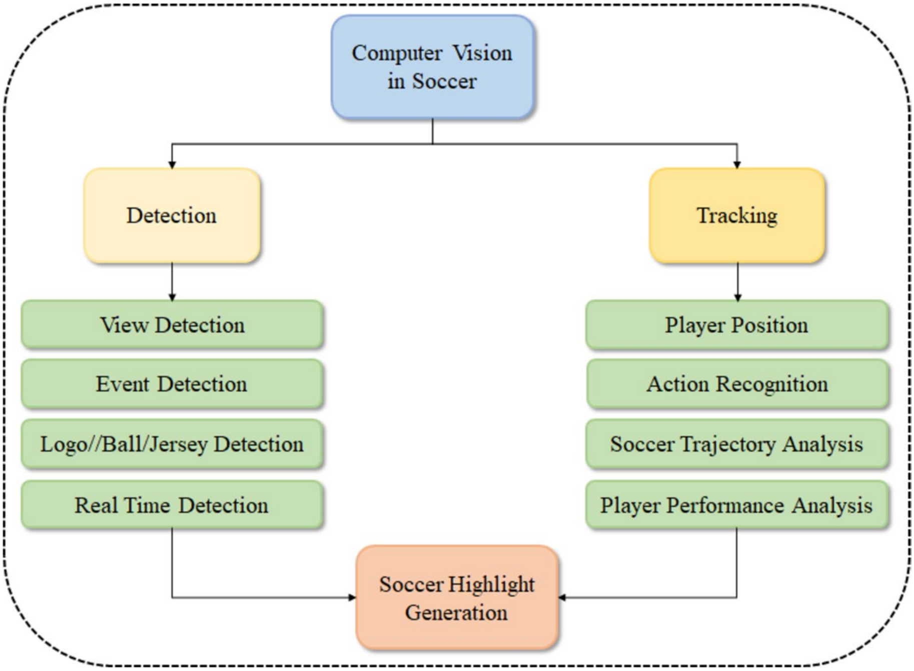

Computer Vision in Soccer
Detection
Tracking
View Detection
Event Detection
Logo//Ball/Jersey Detection
Real Time Detection
Player Position
Action Recognition
Soccer Trajectory Analysis
Player Performance Analysis
Soccer Highlight Generation

Fig. 1 Techniques for soccer highlight generation

1 Introduction

Soccer (Football) ranks as one of the most popular and frequently watched sports worldwide [1]. Soccer is played in over 200 countries with around 265 million players [2] which makes it a successful enterprise that generates billions of dollars [3]. 5.4 billion people watched the 2022 FIFA World Cup, and soccer's worldwide market is expected to reach around $37127.7 million in the sports business by 2027. Sports activities are becoming popular all over the world and this popularity has drawn a large audience [4]. Soccer videos also rank among the most popular social media content. A significant percentage of people like to see only the important events in a game. A Soccer game lasts about 90 min and contains some exciting events that can capture viewers' attention. Analysis of soccer videos has attracted considerable interest in the last decade due to its popularity. The primary difficulty in event highlight generation lies in identifying and selecting which events should be recognized as significant and deserving of inclusion in the highlights. The sequence or series of actions is also quite unexpected in every soccer match. Computer science is experiencing rapid growth and exerting a significant influence on soccer. As a result of tremendous growth in sports activity throughout the world, the need for automatic highlight production from a sports film has skyrocketed. Highlight of a match includes the most essential incident from the full-length video. The creation of video recaps and highlights from sporting events is of great importance to broadcasters. Most people find shorter videos more appealing than the whole match which can easily be transmitted over any network [5]. The creation of highlights of matches is one of the most demanding requirements among the broadcasting sectors because of its viewing and streaming or channel transmission [6]. Smart and concise videos are easy to stream, take up less storage space, and are easy to handle for smartphone users and so on [7]. A variety of techniques for highlight generation are illustrated in Fig. 1.

The majority of key moments of the game are identified manually. However, manual annotation and event clipping, where traditional operators are utilized to determine the start and end of an event is a time-consuming, costly, and tiresome job [8]. There are many techniques such as logo, replay, scoreboard detection, etc. [9], but event detection is still a difficult task, despite all these methods [10]. In these techniques, manual approaches are used for event detection, such as replay detection and then identifying the event [11]. Numerous potential applications have been proposed, which include a number of approaches such as soccer ball detection [10] and player monitoring [12] are used for collecting match statistics [13], or event detection (e.g., scoreboard, goal, outside back, etc.) [14]. Verification tactical analysis highlights automation recognition, referee judgments, visual annotation, automatic game summaries, and so on, but there is still a need for some automated methods. Machine learning-based [15] soccer highlight generation has the potential to address these issues by automating the process. The importance of past soccer matches reduces after some time, so highlights must be generated automatically [16]. Focus is on developing a model based on machine learning that is capable of analyzing soccer videos, and identifying key moments. Significant research shows that there are some gaps while generating highlights in the long view. Detecting events from a video and maintaining the rate of accuracy regardless of the dataset is a critical factor in most research around there. The major contributions of this paper are.

- Proposed a combined approach of deep learning models with traditional machine learning techniques for analyzing soccer events and views. This integration enables the gathering of detailed visual information, which may improve the understanding and accuracy of the data.
- Developed a model that can extract high-resolution strong visual features called "deep prior features" from the first convolutional layer of pre-trained CNN AlexNet and VGG-16 architecture which can generate results for both long and short views.
- A newly created dataset.

The rest of the paper is organized as follws. Section 2 discusses the literature review. The proposed model is described in Section 3. Section 4 presents the experimental results and analysis. The conclusion is summarized in Section 5.

2 Related work

Speech recognition, logo, replay, and slow motion detection are the most frequently used approaches by researchers for event detection [17]. Most of them rely on dominant frame color [18] to classify their view type. S. Tseng et al. [19] utilize the dominant color to determine the frames in view type, implying that the dominating color of long views represents grass, so their ratio is comparatively large. According to their research, image's dominating colors are not green in other view types (medium, short, and outer view). Their strategy is not applicable in some scenarios. The foundation of this approach depends on major color detection. It does not work in all cases. Figure 2 depicts different medium, short, and outer view frames with green as the primary color.

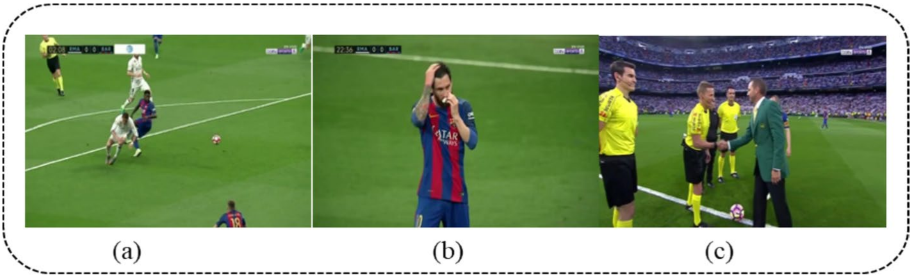
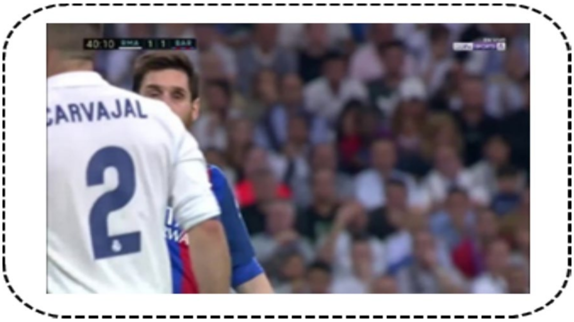


Fig. 2 View classifications (**a**) Short View, (**b**) Medium View, and (**c**) Outer View

Fig. 3 Short View in soccer [21]



DA Sadlier et al. [20] present a technique for short views where the player's face will be exactly at the center of the image which is not true in many cases as mentioned in Fig. 3.

In [22], E. R. M. Faizal et al. developed a framework for event localization in long view. To extract spatiotemporal features [23], Bi-stream CNN architecture is employed by Dilated-Recurrent-Neural-Networks, which is subsequently used for event detection. DRNNs perform better than RNNs and accurately utilize information from previous frames. Using the SoccerNet dataset model, 63.3% mAP was obtained. In [24], X. He et al. employ deep learning [25] technology to extract a player's trajectory in soccer, i.e., to follow the player's goal course. The target identification approach uses deep learning to create new multi-scale characteristics and adjust the creation rules of anchor points within acquired movies, making it more suited for tiny target recognition jobs in football match scenes. The generating rules are based on a complicated decision support system that tracks targets. The Hungarian method is applied, and this approach aims to convert the video's multi-target tracking issue into a data connection problem utilizing a decision support system. Internally, this system uses a detection-based tracking technique. In [26], H. Bai et al. develop an application that can track players' positions and movements in real-time and use it for live broadcasts. To achieve this purpose, they created a system utilizing a deep learning approach. In this approach, a convolutional neural network is developed to extract the rich visual attributes of players in soccer game videos. The network is trained on a huge number of data sets containing comparable objects, which increases the algorithm's capacity to recognize the same team members. In [27], P. Modi et al. present a combined approach that utilizes computer vision techniques with deep learning methods to improve object tracking in videos. They extract the position of objects by predicting the placement of bounding boxes in video frames using deep learning and YOLOv8 architecture. In the next step, blurring and optical flow are used to achieve accurate object tracking. This comprehensive technique guarantees that the item is consistently identified throughout the video frames. Their research provides encouraging results in real-time monitoring of football in football match films. In [28], G. Jin et al. propose an FRCN algorithm relying on deep learning technology that can perform player target tracking in soccer game footage while remaining resilient against player occlusion.

In [29] Zhu, He, et al. propose an effective model based on a transformer for soccer video action detection. The video features are first extracted using the multi-scale ViT. Next, for better temporal information and improved temporal comprehension, a sliding window method is implemented. To get the final results, the features are finally entered into the NetVLAD + +. This model can complete the SoccerNet Challenge 2022 Action Spotting Task and learn a hierarchy of resilient representations. In [30], M. Cao et al. present a SpotFormer technique which is a straightforward yet powerful framework for accurate activity spotting. In particular, effective backbone networks are used to extract supplementary features. These networks allow us to simply and effectively reduce feature dimensionality. A spotting network based on the transformer is designed. Farm-wise features are given to this designed network which uses spatiotemporal data. This technique shows remarkable performance and acquires a 0.609 tight mAP score. In [31], Zhang, Yixiao, et al. present an innovative technique based on an attention multi-modal neural network. This network implements a multiple-phase fusion training approach for event classification. Results demonstrate that the suggested approach excels over the transformer-based video approach by 4.43% on top-1 accuracy upon the Soccernet-V2 dataset. The suggested multi-modal combines three modalities: an image sequence, an audio, and a recently suggested sports formation modality. K. Tasaka et al. [32] suggest a model that recognizes the ball and players before kick motions motion detection [33] in the 4k Multiview soccer video using YOLO-v3 and open posture, which obtain 85% F-score and 73% Precision. The limitation of existing technology is that it cannot tackle the occlusion problem. If the cameras [34] cannot catch the ball, then ball detection will remain undetected. Pose estimation [35] will fail partly if body portions are obscured. Akhaee et al. [36] propose a model for the detection of nine different events in soccer; two CNN and Variational Autoencoder (VAE) are used to distinguish soccer images and other images. To categorize the yellow cards and red cards, Fine-Grained Image Classification (FGIC) was used, which raised the accuracy to 93%. G. S. Choi et al. [37] suggested a system based on detection for evaluating soccer events, which detects five events: free and corner kicks, centre-line, goals, and (soccer ball) passing. The 3DResNet34 architecture convolutional network framework is used in this investigation, which achieves an accuracy of 96.81%. Their proposed model is the restricted number of events, and highlights generation is not considered. In [38], Sifan Ma et al. introduce a novel soccer video event recognition [39] system applying the "self-attention mechanism," which grabs significant frames. NetVLAD architecture is employed to get time window-level features. The video clips [40] are then classified into 4 classes, which are Players'- substitution, Red cards, Goal, and Yellow cards. The self-attention approach helps to achieve classification [41] accuracy from 67.2% to 74.3% on the SoccerNet dataset. OA Nergård Rongved et al. [42] develope a CNN-based Audio-Visual combination model. Audial information is used to evaluate the consequences of detection. Visual and audio components are more useful for goal event detection [43] than cards and player substitution. On soccerNet-v2, this model obtained average mAP values of 72.2% for CALF, 69.7% for AudioVid, and 54.9% for NetVLAD, respectively. A. Sen et al. [44] propose a model for action detection in soccer. A combined CNN and RNN model is used to classify ten different soccer events. They extract spatial features using VGG Net for event detection in video frames. Although these spatial features are not enough to solve the gradient vanishing problem, VGG is updated to hybrid GRU, which can achieve an accuracy of up to 94%. BT Naik et al. [45] introduce a real-time soccer and player tracking

SORT-based technique using YOLOv3, which was initially utilized for identifying and categorizing the ball, player, and background. Kalman filtering is then used for tracking. This system handles difficult scenarios well, including partial occlusion for both the players and the ball. Nevertheless, it fails when players are heavily obscured. Model achieves a tracking accuracy of 93.7%. In [46], K Vats et al. propose a Multi-Tower 1D CNN architecture. A pre-trained 2D-CNN model is utilized to get the feature vector by passing video frames. Then, the feature vector is fed into the 1D CNN layers. Lastly, class probabilities are calculated by combining the inputs from the parallel layers. The SoccerNet dataset is used for model validation with a 70.4% mAP success rate for card detection, goal events, and player substitution. In [47], S Giancola et al. propose a new NetVLAD-based feature pooling approach that embeds spatial and temporal information for action detection in the soccer video stream. To pool information, the context is divided into two parts: the context after and before the event occurrence. Using a clustering algorithm-based method, this pooling technique can understand the past and future contexts in a feature space. The model learns on the vast SoccerNet-v2 dataset, and it achieved a 53.4% mAP for action detection A. M. Tekalp et al. [48] suggest a framework for the detection of three different activities of soccer: i) goal, ii) penalty boxes, and iii) referees in match footage. The color region that is prominent and shorter boundary lines are also detected with this proposed model that is resistant to changes in the dominant color. Overall, the algorithm obtained a recall rate of 97.3% and an accuracy rate of 91.7%. In [49], H.R. Pourreza et al. present a Fuzzy-Inference-System (FIS), which is dependent on replay, logo, audience interest, identification of view type, and player detection for highlight generation. Three different events are analyzed in this technique that are Free kicks, Penalties, and Corners.

3 Proposed methodology

Overall, the proposed model of soccer event detection is depicted in Fig. 4 Image frames are extracted from soccer video footage, and then further image scaling is performed on the images.

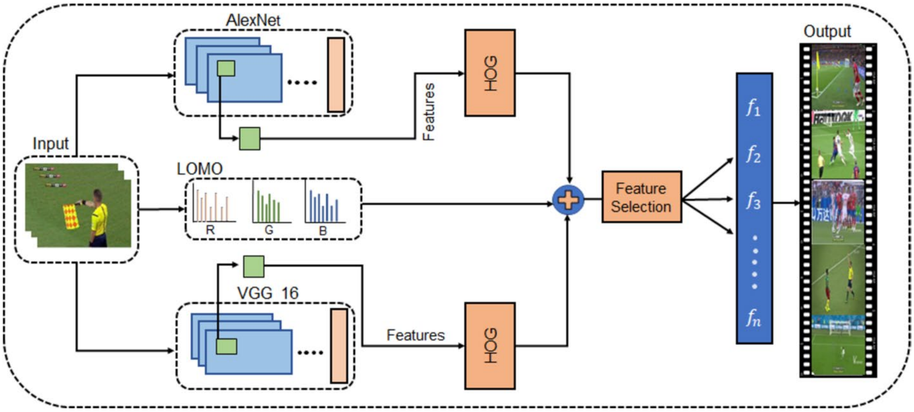


Fig. 4 Proposed model of soccer event detection and highlight generation

This proposed system is evaluated in two steps. Initially, high-resolution strong visual features are extracted from the very first convolution layer of deep learning models, and named those features as "deep prior features," which can reduce processing time and help achieve better results. These deep prior features are extracted using CNN-based AlexNet and VGG-16 models. Then, these deep prior features are fused with HOG and LOMO to get a single feature vector. The last step of this model is to select features for output.

3.1 Preprocessing

In this proposed work, image processing is applied to improve the visual representation of images for better classification results. Image resizing is done on the self created dataset because original images are 640×360, which can increase the processing time.

3.1.1 Image scaling

In the suggested technique, image scaling is used to resize an image without losing image quality, vector graphic images are resized utilizing geometric information. Let the original image is $I_{640X360}$. Image scaling is applied to convert it into $I_{227X227}$. Let S be the scaling factor of an image, W represents the width, and H is used to represent the height of the image. $S\prime(width)$ represents the width of the scaled image, Similarly $S\prime(Height)$ represents the height of the scaled image. Scaling is applied on both the width and height of the image as illustrated in Fig. 5, which can be performed as:

$$S\prime(Width) = S * W \tag{1}$$

$$S\prime(Height) = S * H \tag{2}$$

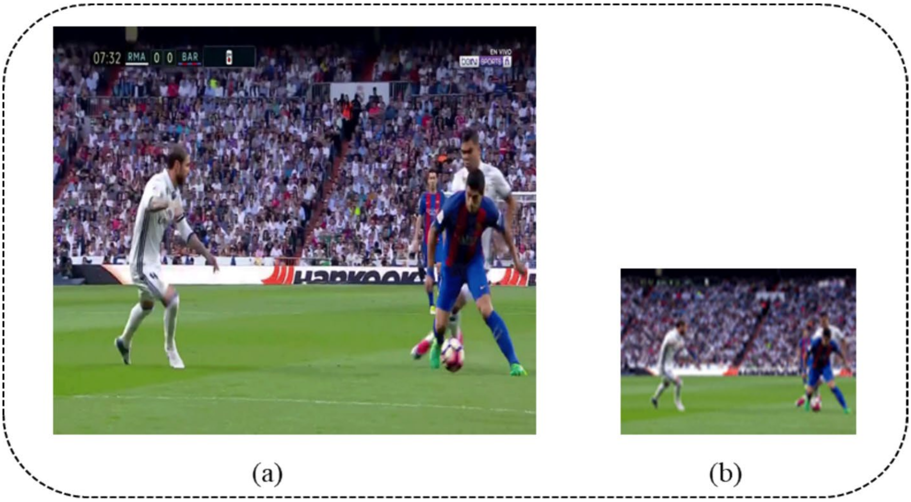


Fig. 5 (**a**) Original image (640, 360) (**b**) Scaled image (227, 227)

3.2 Feature extraction

A key step in machine learning and pattern identification is feature extraction, which is mainly used to minimize the dimensionality of given data while retaining its most relevant and discriminative properties. Features can indicate a variety of characteristics in image processing, including color, texture, form, and spatial connections between objects.

In the proposed model, handcrafted HOG and LOMO features are used, and deep learning models AlexNet and VGG-16 are applied to extract prior deep features. These models are used for feature extraction from the first convolutional layer by passing the dataset via the network.

3.3 Deep HOG feature extraction

In the proposed model, the input image $I_{227X227}$ is given to the Pre-trained CNN models AlexNet and VGG-16, which extract high-resolute deep prior features obtained from the relevant activation parameters of the first convolutional layer *(conv1),* which can reduce training with accurate results.

3.3.1 Feature extraction using LOMO

LOMO feature descriptors are mostly designed to extract colored features for image or video dataset analysis. Algorithms of LOMO descriptor identify local maxima of image intensity and gradient values to extract information. In the proposed work, LOMO descriptors [50] are used to extract useful characteristics from self created dataset, such as red and yellow cards, goal attempts, and other important ones in the context of view and event classification. Our dataset consists of colored images, since color image processing [51] has been gaining importance. The initial step of LOMO involves converting the image to grayscale. Following that, the intensity value differences between each pixel and its neighboring pixels are computed for the red, green, and blue channels individually. These computed differences are then subjected to a thresholding operation, resulting in binary patterns for each channel. Finally, the binary patterns obtained from the three channels are merged to create the ultimate LOMO descriptor. Let the histograms of RGB image are H_1, H_2, and H_3; its features can be extracted as:

$$f(LOMO) = \left[H_1, H_2, H_3 \ldots \ldots \ldots H_n\right] \tag{3}$$

By utilizing the RGB channels to include color information, LOMO descriptors effectively encompass the texture and color attributes of an image.

3.3.2 Feature extraction on AlexNet

AlexNet is a neural network model that recognizes images by using layers to detect different features like edges and objects. AlexNet constitutes an 8-layer CNN composed of 3 FC layers, 5 convolutional layers and maximum pooling. The construction of the AlexNet is illustrated in Fig. 6.

Let $I_{227X227}$ is the input image; its deep prior features using AlexNet can be extracted as:

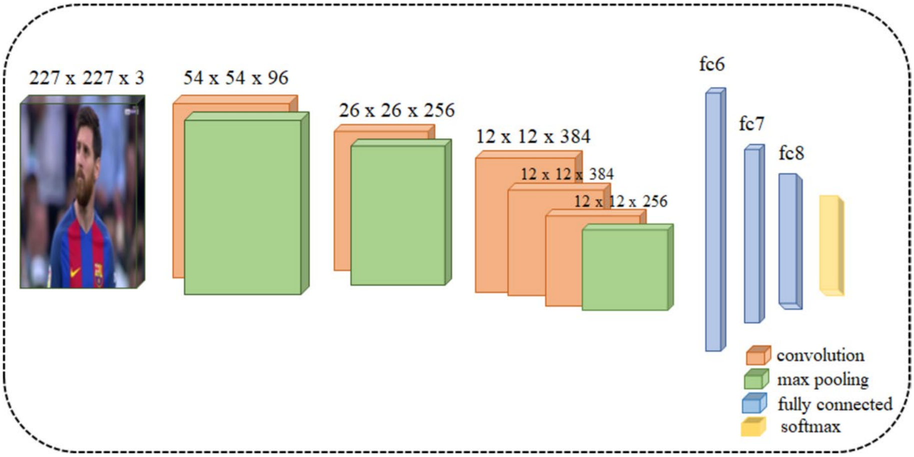


Fig. 6 Structure of AlexNet model

$$f(Alex) = Activation(Conv1(I_{227X227})) \tag{4}$$

The very first convolution layer is evaluated to extract deep prior features rather than the FC layer; because of this, the process of extracting features takes less training time and gives more accurate results than extracting features from the FC layer, hence attaining 99.4% accuracy. These deep prior features are then fused with VGG-16 and LOMO after which HOG is also applied on these features.

In the presented model, the activation function is performed on AlexNet. Figure 7 shows the activation function on an image where it is normalized.

The process of getting important visual details from convolutional layers shows how deep learning models can automatically understand and show complex visual information. Figure 8 shows the identification of deep prior features at the *Conv1* layer. White spots in the activation show where the channel is strongly located.

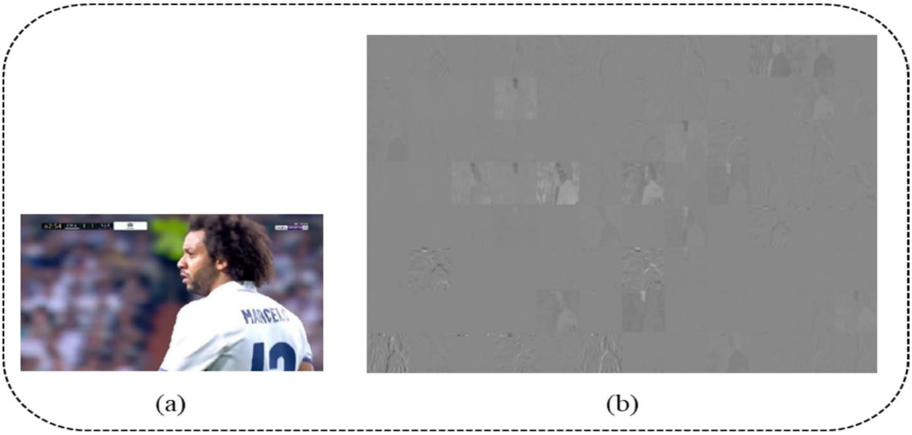


Fig. 7 Activation on AlexNet (**a**) Original image (**b**) Activation function on AlexNet

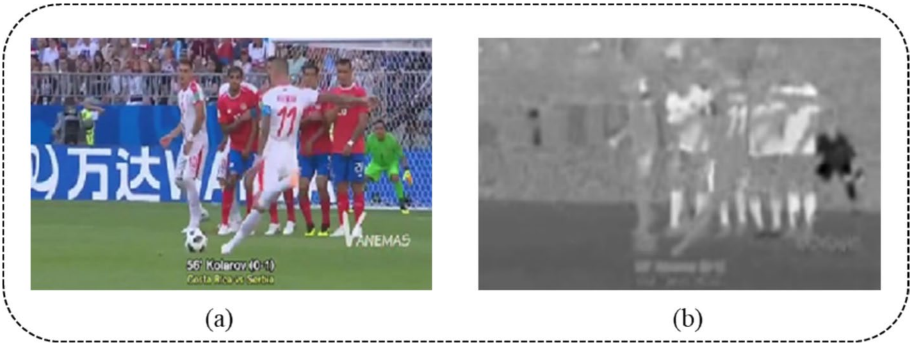


Fig. 8 (**a**) Original image (**b**) Deep prior features at *Conv1*

3.3.3 Feature extraction on VGG-16 net

VGG Net has an ability to extract strong visual features from photos [52]. Figure 9 depicts the fundamental construction of a VGG-16 Net. As more convolutional layers are added, VGG-16 Net progressively learns more complex and abstract features.

Let $I_{227X227}$ is the input image; its deep features using VGG-16 Net can be extracted as:

$$f(VGG16) = Activation(Conv1(I_{227X227})) \tag{5}$$

While extracting features using VGG-16, the very first convolution layer is evaluated to extract deep prior features rather than the FC layer, which will take less extracting time. VGG-16 is a dense model that extracts prior features from *Conv1* to give more accurate and fast results. Then, these deep prior features are fused with HOG and LOMO. Furthermore, pre-trained VGG-16 Net model shows remarkable results while training on a substantial dataset comprising of 38,728 images.

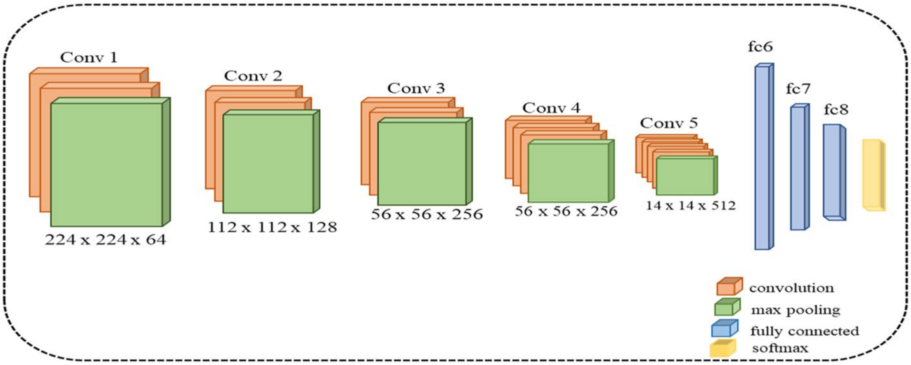


Fig. 9 Structure of VGG-16 model

3.4 Deep HOG feature fusion

A fusion approach is presented for soccer events, and view recognition is integrated with deep and handcrafted features. Figure 10 illustrates the whole feature fusion process. In this feature fusion step, deep prior features extracted by applying AlexNet and VGG-16 are fused with HOG as:

$$f(deepHOGfeatures) = \left[f(Alex)f(VGG16)f(HOG)\right] \tag{6}$$

3.5 Hand crafted feature extraction

Handcrafted features have been successfully implemented for decades and remain an effective technique when paired with ML classifiers. This research uses LOMO and HOG as handcrafted features to train the proposed model.

3.5.1 Feature extraction using HOG

The histogram of oriented gradients (HOG) is an approach for feature extraction that is commonly used for object detection as well as for action recognition [53]. The HOG descriptor is utilized for human identification, particularly in a soccer match scenario where numerous players are present.

Let the coordinate be *(x,y)* where gradient orientation in the directions of x and y are $\varnothing(x, y)$ and the gradient magnitude represented as *G(x,y)* is computed for each of the pixels in each cell as:

$$f(HOG) = \sum(G(x, y), \varnothing(x, y)) \tag{7}$$

By employing HOG, most effective features are extracted. The output of applying HOG on the dataset is illustrated in Fig. 11. It works by characterizing a picture as a series of

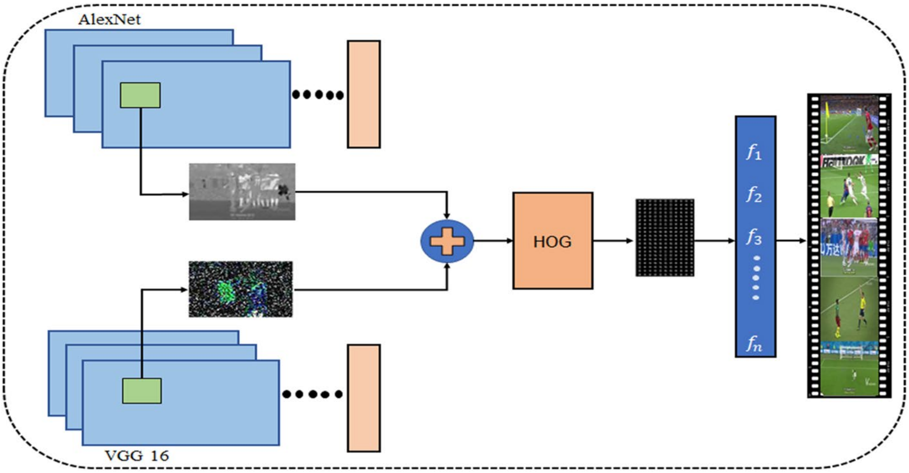


Fig. 10 Deep HOG fusion

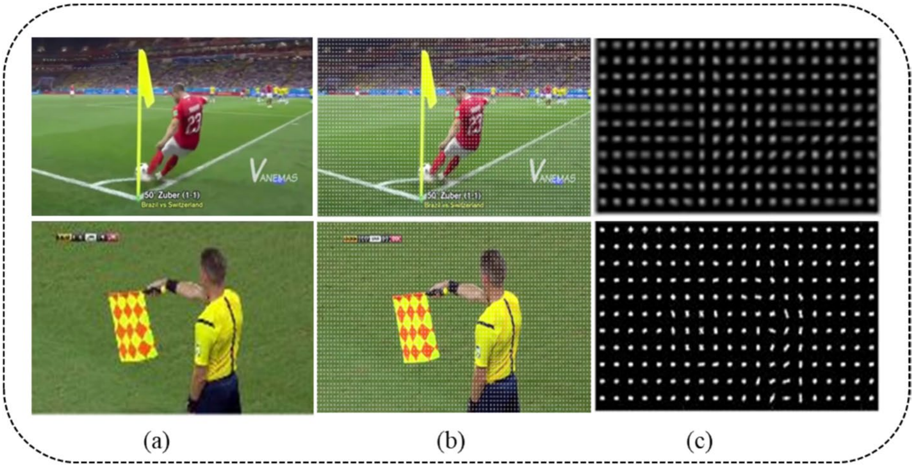


Fig. 11 HOG features extraction (**a**) Original image (**b**) Extract & plot HOG features (**c**) HOG features

local histograms, which identify gradient orientation patterns in a certain area of the picture. The image is split into 50% overlapped blocks, with each block further subdivided into cells.

3.6 Feature fusion

In this proposed model, best features are extracted from the first layer of VGG-16 and AlexNet. These features are then combined with HOG, which is the best local descriptor. It analyses the localized region of an image instead of the whole image at once. HOG can analyse local features more deeply such as edges and shapes that are required to detect objects like balls, and players in soccer events. Combining deep prior features with HOG provides the best features for classifiers that contain detailed and texture information from both deep model and HOG to give a more precise representation. Through this feature fusion technique, the presented model shows significant improvement in event classification. This leads to increased accuracy and reduced computation time.

Accuracy is improved by 99.6% illustrating the effect of combining deep features of AlexNet and VGG-16 with fine details of the HOG descriptor. Deep learning features excel at complex patterns, while HOG excels at detecting local patterns. This combination is helpful while classifying soccer events because it captures both the high-level semantic features and low-level fine-grained texture features to enhance the accuracy of proposed model. This improved model highlights the importance of exploring more alternative techniques of feature fusion and its implementation in image classification.

This fusion technique shows how the best parts of many deep learning models can be used to improve accuracy. Different soccer events can be identified and old games can be watched. In this last feature fusion step, all extracted features combined with HOG enhance the performance of presented model.

$$f(fussion) = [f(LOMO)f(DeepHOGfeatures)] \tag{8}$$

AlexNet is well-known for extra abstract high-level features, but due to VGG-Net's deep architecture, it extracts finer-grain details. When both deep prior features are combined, this leads to increased performance.

4 Results and analysis

In this accompanying section, the details of the selected datasets and the evaluation process are described. Detailed experimental evaluation is being carried out to examine the feasibility and efficacy of the proposed model in long and short views. The results are then compiled and compared to existing approaches. Finally, the results are summarized in the discussion section.

4.1 Dataset

We created our own dataset, which is split into two categories: view dataset and event dataset. YouTube soccer videos are used in this dataset. The view dataset consists of four different classes of 137,196 images; each class consists of 878,9 ~ 961,55 images approximately, and the event dataset contains 10 classes of 38,728 distinct images such that each class consists of 841 ~ 221,49 images approximately. This dataset consists of only RGB images. View dataset contains 4 classes: Long view, Medium view, Shorter view, and Outer view. Table 1 shows that the view dataset is distributed in test, train, and validation images. Also, the total number of classes, class names along with their labels, and the number of images in each class are summarized.

Event dataset contains ten different classes, which are: Red card, Spectator, Yellow card, Plenty stock, Corner, Player celebration, Offside, Free kick, Goal and Goal attempt. Table 2 shows how the event dataset is distributed in test, train, and validation images. Also, the total number of classes, class names along with their labels, and the number of images in each class are summarized.

A comprehensive comparison is made in Table 3 to evaluate our dataset's robustness with other existing datasets. This comparison is based on the number of classes, which include event classes, and view classes. It aims to highlight our dataset's unique qualities, and strengths.

Table 1 Description of view dataset

Class name	Class Label	Total images	No. of train images	No. of validation images
Long view	A	96,155	67,309	28,846
Medium View	B	16,645	11,652	4,993
Shorter View	D	15,607	10,925	4,682
Outer View	E	8,789	6,153	2,636
Total images in Dataset		137,196	96,039	41,157

Table 2 Description of event dataset

Class name	Class Label	Total images	No. of train images	No. of validation images
Red card	C	2,000	1,400	600
Corner	G	2,768	1,938	830
Yellow card	H	890	623	267
Plenty stock	I	3,307	2,315	992
Player celebration	J	10,219	7,154	3,065
Offside	K	1,491	1,044	447
Goal attempt	L	3,055	2,139	916
Goal	M	10,922	7,645	3,277
Free kick	R	3,235	2,265	970
Spectator	X	841	589	252
Total images in Dataset		38,728	27,110	11,618

Table 3 Comparision with existing datasets

Ref	Dataset	No. of Classes	No. of Events	No. of Views
[36]	SEV Dataset	10	8	Nil
[54]	IAUFD	10	10	Nil
[42]	SoccerNet	4	4	Nil
[55]	Novel Dataset	8	4	4
[56]	SoccerDB	11	11	Nil
	Proosed Dataset	14	10	4

4.2 Experiments and performance evaluation

First, the model proposed in this paper is evaluated on the soccer event dataset. Initially, to train this model, images from ten different events are selected. Events are classified using high-resolution deep prior features retrieved from the very first convolution layer. A distributed dataset that contains testing and validation images is used to evaluate the model. During training, the model is saved with the highest accuracy on the validation dataset and verified its performance on the test dataset. Nine different classifiers, including linear discriminant, linear SVM, Quadratic SVM, Cubic SVM, Fine KNN, Medium KNN, Cubic KNN, Weighted KNN, and Subspace Discriminant, are used to evaluate the performance of the proposed method. All simulations were performed on MATLAB 2018a using core i7 with 16GB RAM and an ITB hard disk.

4.2.1 Experiment 1: event classification using deep prior features

In this experiment, extracted deep prior features from the proposed model are used to classify ten different events. An approach of 30:70 was adopted for training and testing of validation of the proposed method. All testing results were obtained using

Table 4 Event classification using deep prior features

Performance Measures					
Classifiers	Accuracy (%)	Precision (%)	Recall (%)	F1-Score	Sensitivity
Subspace KNN	**99.6**	**99.6**	**91.4**	**15.7**	**9.6**
Fine KNN	99.5	99.6	91.3	15.9	9.7
Medium KNN	97.4	97.6	86.9	22.9	14.1
Cubic KNN	96.9	97.3	85.5	24.9	15.5
Weighted KNN	98.7	98.9	90.6	17.0	10.4
Cubic SVM	98.6	98.7	89.6	18.7	11.4
Medium Guassian SVM	96.5	97.0	85.5	2.60	15.5
Quadratic SVM	87.4	86.0	74.9	38.2	26.1
Fine Gussian SVM	96.7	97.0	83.5	27.5	17.2

Table 5 Class-based results on deep prior features

Predicted Class												
		65	**66**	**67**	**69**	**70**	**71**	**72**	**73**	**74**	**100**	**107**
Classified	**65**	2755	000	000	005	000	005	002	000	000	001	000
	66	000	605	002	003	000	000	000	000	001	000	000
	67	000	000	3229	000	000	000	002	000	003	001	000
	69	000	000	001	3054	000	001	000	000	000	000	000
	70	000	000	000	001	1489	001	000	000	000	000	000
	71	000	001	000	002	000	1486	001	001	000	001	000
	72	000	000	003	003	000	002	1483	001	000	000	000
	73	000	000	000	001	000	000	000	1490	001	000	000
	74	000	000	000	000	000	000	000	000	841	000	000
	76	002	001	003	009	000	006	000	001	001	005	000
	100	000	000	000	001	000	003	000	000	000	3051	000
	107	000	000	000	000	000	000	000	000	000	000	029
Recall (%)		99.5	99.0	99.8	99.9	99.9	99.6	99.4	99.8	99.9	99.8	99.9
Precision (%)		99.9	99.7	99.7	99.2	99.9	98.6	99.7	99.9	99.3	99.7	99.9
F1-Score		25.9	12.0	12.1	4.00	4.00	12.0	17.9	4.00	0.00	7.99	0.00

tenfold cross-validation. During this experiment, the best testing classification accuracy was recorded at 99.6%, with a precision rate of 99.6%, sensitivity of 9.6, and an F1-Score of 15.7 on Subspace KNN. Classification results of Subspace KNN compared with eight other classifiers are shown in Table 4.

4.2.2 Experiment 2: Class-based results on deep prior features

In this experiment, a class-wise analysis is conducted using deep prior features to evaluate the accuracy, recall, and precision of the proposed model. The results of this analysis are further compared with the performance of other existing state-of-the-art methods.

Table 5 highlights the findings of analysis, showing major differences across various classes. The proposed model's accuracy varied from class to class, and these differences may be due to the occurrence of inter-class similarities. It is important to note that classes that had low inter-class similarities outperformed the rest of the classes in terms of accuracy. Furthermore, the differences in accuracy between classes highlight the difficulties of event identification in soccer. Some events have similar characteristics, which makes classification more difficult. This intricacy emphasizes the importance of fine-tuned proposed model that provides an understanding of classes having low accuracy and requires more focus to achieve more accuracy.

4.3 Event classification using handcrafted features

In this experiment, features are extracted from the HOG, LBP, and SFTA. An approach of 30:70 was adopted for training and testing of validation of the proposed method. All testing results are obtained using tenfold cross-validation. During this experiment, the best testing classification accuracy using HOG is 99.1%, FNR 0.9%, and precision rate 99.0%, with an execution time of 623 s on Subspace KNN. For LBP, the best testing classification accuracy is 99.1%, FNR 0.9%, and precision rate 98.8%, with an execution time of 529 s on Subspace KNN. Classification accuracy using SFTA is 99.01%, FNR is 100%, and the precision rate is 98.8%, with an execution time of 445 s on Subspace KNN. In Table 6, classification results of Subspace KNN are compared with ten other classifiers.

4.4 Comparative analysis

Various studies in soccer analytics have been reviewed and compared in this comparative study based on their classification performance among different classes. A comparative analysis is conducted in Table 7 using different baseline classifiers including different types

Table 6 Event classification using handcrafted features

	HOG			LBP			SFTA		
Classifiers	Acc (%)	Pre (%)	FNR (%)	Acc (%)	Pre (%)	FNR (%)	Acc (%)	Pre (%)	FNR (%)
Subspace KNN	**99.1**	**99**	**.9**	**99.1**	**98.8**	**0.9**	**99**	**98.8**	**1**
Linear SVM	78.6	83.5	21.4	71.8	76.5	28.2	62.5	64.6	37.5
Quadratic SVM	98.1	97.7	1.9	96.6	96.9	3.4	91.5	93.4	8.5
Cubic SVM	99.0	98.7	1	98.8	98.6	1.2	98.1	97.8	1.9
Fine KNN	99.0	98.8	1	99	98.7	1	99	98.8	1
Medium KNN	97.4	98.0	2.6	97.1	97.7	2.9	96.6	97.1	3.4
Cubic KNN	97.2	95.8	2.8	96.7	96	3.3	96.2	95.6	3.8
Weighted KNN	98.7	97.6	1.3	98.6	97.9	1.4	98.5	98.1	1.5
Subspace Discriminant	88.2	69.1	11.8	60	71.9	40	54	55.9	46
Linear Discriminant	67.5	66.3	32.5	62	60.4	38	54.8	47.8	45.2
Medium Gaussian SVM	98.6	98.7	1.4	97.8	97.9	2.2	95.9	96.2	4.1

Table 7 Comparison with baseline classifiers

Baseline Classifiers	Accuracy (%)	Precision (%)	Sensitivity (%)
Linear SVM	95.7	96	95.75
Quadratic SVM	98.1	98	97.25
Cubic SVM	99	98.75	99
Medium Gaussian SVM	98.6	98.75	98.89
Medium KNN	97.4	97.25	97.25
Cubic KNN	97.2	97	97
Weighted KNN	98.7	98.75	98.5
Subspace KNN	98.6	90.1	88.5
Proposed Model	**99.6**	**99.6**	**9.6**

of suppot vector machines (SVM) and K-Nearest Neighbors (KNN). This comparison summarize the performance of several classifiers with the proposed model.

A comparative analysis between the proposed model and state-of-art work is presented in Table 8. Different and most common soccer events, including mid-field activity, ball-related occurrences, and throw-ins, are all considered in this comparison table. Each row in a table is related to a single paper and indicates the existence or absence of specific features within the classes. It also includes overall accuracy, which is an important metric in determining the effectiveness of the model.

The comparison table demonstrates how different research evaluates soccer events and applies different methodologies. It is critical to understand potential restrictions such as dataset variances. These findings help improve soccer analytics by assisting researchers in improving their techniques and accuracy in various soccer events.

5 Conclusion and future work

In this era of increasing demand for video content, extracting important events from lengthy soccer match recordings to generate highlights is of considerable importance. Previous research shows that there is a need for more efficient and precise approaches to this task. Current models are unable to achieve a balance between time efficiency and accuracy. This research provides a solution that reduces time while simultaneously improving

Table 8 Comparative analysis with existing state-of-art models

Ref	Baseline Models	No. of events	Overall results
[23]	CNN, Variational Auto Encoder (VAE)	8	Acc = 93.21%
[44]	VGG–GRU	8	Acc = 94%
[57]	3 D CNN	6	mAP = 77.2, mAR = 71.7
[54]	VGGNet 13, ResNet-18	10	-
[55]	PSPFSR	4	-
[58]	Fastest RCNN, YoloV5	4	Acc = 95%
Proposed Model	AlexNet, VGG-16	10	Acc = 99.6%

accuracy in detecting significant events. The fusion of traditional and deep features allows for the collection of precise visual information, which increases the interpretation of the model and improves accuracy.

In this research work, the effectiveness of using the deep learning models AlexNet and VGGNet with handcrafted HOG features is examined for recognizing ten different soccer events and four different views. To accomplish this, the dataset is created and a model is trained on this dataset to extract features. Robust visual features are extracted from soccer images. The fusion of deep prior features extracted from the first convolution layer of AlexNet and VGGNet models combines both high-level and low-level features effectively, while local properties and gradient-based information are improved by handcrafted features. Tables, graphs, and figures are used to present the experimental results. The classification results concerning execution time, model accuracy, recall, precision, F1-score, and sensitivity show the efficiency of the proposed method, which has performed outstandingly as compared to existing methods. The fusion of deep prior features extracted from the first convolution layer of AlexNet and VGGNet reduces the training time and achieves an accuracy of 99.6%. The proposed model is not restricted to soccer; it can be used in any other domain or to generate highlights of any sport, such as cricket or hockey. This can also be used in further soccer work, such as detecting player position, analysis of player performance, and so forth. The suggested work can be extended to real-time video analysis, and it would be the best direction for further research in this area. This strategy will work in any extent that involves feature extraction, fusion, and selection.

While this research focuses on developing a strong foundation, it is still open for future work to enhance these findings in real-time video analysis. In the future, while developing an efficient algorithm for real-time video analysis data handling, high computation power will be required. For real-time analysis, important models must be responsive to changes in input data and environmental conditions. Techniques such as transfer learning, self-supervised, and continuous learning for real-time analysis could be the future direction in this domain. Another domain for future research is dynamic scene understanding. There are still some challenges in this domain including breaking down the scenes into different frames, identifying unseen objects in real-time data, and semantic understanding. These challenges can be overcome by implementing a model that can handle occlusion, lighting conditions, and scene segmentation. To overcome with occlusions, future research should focus on detecting and tracking objects using models such as LSTM or GRU. These models can maintain temporal stability and can accurately predict the location of objects that are blocked. Handling lighting conditions can significantly improve the efficiency of models. To adjust the lighting condition among images, different models can be employed such as GANs or CycleGAN. Some data normalization techniques such as CLAHE before processing can also help to improve illumination consistency. For scenes, segmentation models like U-Net or DeepLab can perform best because of their remarkable segmentation capabilities. Future research can significantly improve the functionality of scene analysis systems, making them more robust more effective, and versatile across a wide range of practical applications. This future direction will not only handle occlusions and changing lighting conditions but will improve the real-time development.

Acknowledgements In this research article author received no funding from any profit or non-profit organization.

Funding The authors clarify that no funding is obtained for the research work.

Data availability Data for this research will be available upon request.

Declarations

Conflicts of interest Regarding this work authors have no conflicts of interest.

References

1. Krustrup P, Krustrup BR (2018) Football is medicine: it is time for patients to play!, vol 52. BMJ Publishing Group Ltd and British Association of Sport and Exercise Medicine, pp 1412–1414
2. Cuevas C, Quilón D, García N (2020) Techniques and applications for soccer video analysis: a survey. Multimed Tools Appl 79(39–40):29685–29721
3. Micelotta E, Washington M, Docekalova I (2018) Industry gender imprinting and new venture creation: the liabilities of women's leagues in the sports industry. Entrep Theory Pract 42(1):94–128
4. Tejero-de-Pablos A, Nakashima Y, Sato T, Yokoya N, Linna M, Rahtu E (2018) Summarization of user-generated sports video by using deep action recognition features. IEEE Trans Multimed 20(8):2000–2011
5. Jiang Y-G, Dai Q, Mei T, Rui Y, Chang S-F (2015) Super fast event recognition in internet videos. IEEE Trans Multimed 17(8):1174–1186
6. Darapaneni N, Kumar P, Malhotra N, Sundaramurthy V, Thakur A, Chauhan S, Thangeda KC, Paduri AR (2022) Detecting key soccer match events to create highlights using Computer Vision. arXiv preprint arXiv:2204.02573
7. D'Orazio T, Leo M (2010) A review of vision-based systems for soccer video analysis. Pattern Recogn 43(8):2911–2926
8. Valand JO, Kadragic H, Hicks SA, Thambawita VL, Midoglu C, Kupka T, Johansen D, Riegler MA, Halvorsen P (2021) AI-based video clipping of soccer events. Mach Learn Knowl Extract 3(4):990–1008
9. Pan H, Van Beek P, Sezan MI (2001) Detection of slow-motion replay segments in sports video for highlights generation. In 2001 IEEE international conference on acoustics, speech, and signal processing. proceedings (Cat. No. 01CH37221), vol 3. IEEE, pp 1649-1652
10. Huang C-L, Shih H-C, Chao C-Y (2006) Semantic analysis of soccer video using dynamic Bayesian network. IEEE Trans Multimed 8(4):749–760
11. Raghuram M, Chavan NR, Koolagudi SG, Ramteke PB (2016) Efficient audio segmentation in soccer videos. In 2016 IEEE Canadian conference on electrical and computer engineering (CCECE). IEEE, pp 1-4
12. Ali MN, Abdullah-Al-Wadud M, Lee S-L (2012) An efficient algorithm for detection of soccer ball and players. Proc 16th ASTL Control Netw 16:39-46
13. Barros RM, Misuta MS, Menezes RP, Figueroa PJ, Moura FA, Cunha SA, Anido R, Leite NJ (2007) Analysis of the distances covered by first division Brazilian soccer players obtained with an automatic tracking method. J Sports Sci Med 6(2):233
14. Hosseini M-S, Eftekhari-Moghadam A-M (2013) Fuzzy rule-based reasoning approach for event detection and annotation of broadcast soccer video. Appl Soft Comput 13(2):846–866
15. Ansari GJ, Shah JH, Yasmin M, Sharif M, Fernandes SL (2018) A novel machine learning approach for scene text extraction. Futur Gener Comput Syst 87:328–340
16. Chang S-F (2002) The holy grail of content-based media analysis. IEEE Multimed 9(2):6–10
17. Zhang D, Chang S-F (2002) Event detection in baseball video using superimposed caption recognition. In proceedings of the tenth ACM international conference on multimedia, pp 315-318. https://doi.org/10.1145/641007.641073
18. Murtaza M, Sharif M, Raza M, Shah JH (2013) Analysis of face recognition under varying facial expression: a survey. Int Arab J Inf Technol 10(4):378–388
19. Tseng VS, Su J-H, Huang J-H, Chen C-J (2008) Integrated mining of visual features, speech features, and frequent patterns for semantic video annotation. IEEE Trans Multimed 10(2):260–267
20. Sadlier DA, O'Connor NE (2005) Event detection in field sports video using audio-visual features and a support vector machine. IEEE Trans Circuits Syst Video Technol 15(10):1225–1233
21. Tang K, Bao Y, Zhao Z, Zhu L, Lin Y, Peng Y (2018) Autohighlight: Automatic highlights detection and segmentation in soccer matches. In 2018 IEEE international conference on big data (Big Data). IEEE, pp 4619-4624
22. Mahaseni B, Faizal ERM, Raj RG (2021) Spotting football events using two-stream convolutional neural network and dilated recurrent neural network. IEEE Access 9:61929–61942

23. Yan L, Wang Q, Ma S, Wang J, Yu C (2022) Solve the puzzle of instance segmentation in videos: a weakly supervised framework with spatio-temporal collaboration. IEEE Trans Circ Syst Video Technol 33(1):393–406
24. He X (2022) Application of deep learning in video target tracking of soccer players. Soft Comput 26(20):10971–10979
25. Khokhar FA, Shah JH, Saleem R, Masood A (2024) Harnessing deep learning for faster water quality assessment: identifying bacterial contaminants in real time. Vis Comput 1-12. https://doi.org/10.1007/s00371-024-03382-7
26. Bai H, Yuanyuan C, Cheng Z (2023) Research on soccer player tracking algorithm based on deep learning. EAI international conference, BigIoT-EDU. Springer, pp 70–80
27. Modi P, Menon D, Verma A, Areeckal AS (2024) Real-time object tracking in videos using deep learning and optical flow. In 2024 2nd international conference on intelligent data communication technologies and internet of things (IDCIoT). IEEE, pp 1114-1119
28. Jin G (2022) Player target tracking and detection in football game video using edge computing and deep learning. J Supercomput 78(7):9475–9491
29. Zhu H, Liang J, Lin C, Zhang J, Hu J (2022) A transformer-based system for action spotting in soccer videos. In proceedings of the 5th international ACM workshop on multimedia content analysis in sports, pp 103-109. https://doi.org/10.1145/3552437.3555693
30. Cao M, Yang M, Zhang G, Li X, Wu Y, Wu G, Wang L (2022) Spotformer: a transformer-based framework for precise soccer action spotting. In 2022 IEEE 24th international workshop on multimedia signal processing (MMSP). IEEE pp 1-6
31. Zhang Y, Li B, Fang H, Meng Q (2023) A multi-modal transformer approach for football event classification. In 2023 IEEE international conference on image processing (ICIP). IEEE, pp 2220-2224
32. Xu J, Tasaka K (2020) 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
33. Cui Y, Yan L, Cao Z, Liu D (2021) Tf-blender: temporal feature blender for video object detection. In proceedings of the IEEE/CVF international conference on computer vision, pp 8138-8147. https://doi.org/10.1109/ICCV48922.2021.00803
34. Qadeer N, Shah JH, Sharif M, Dahan F, Khokhar FA, Ghazal R (2024) Multi-camera tracking of mechanically thrown objects for automated in-plant logistics by cognitive robots in Industry 4.0. Vis Comput pp 1-20. https://doi.org/10.1007/s00371-024-03296-4
35. Sharif M, Shah JH, Mohsin S, Raza M (2014) Facial Feature Detection and Recognition for Varying Poses. In: World Congress on Engineering and Computer Science pp 22–24
36. Karimi A, Toosi R, Akhaee MA (2021) Soccer event detection using deep learning. arXiv preprint arXiv:2102.04331
37. Agyeman R, Muhammad R, Choi GS (2019) Soccer video summarization using deep learning. In 2019 IEEE conference on multimedia information processing and retrieval (MIPR). IEEE, pp 270-273
38. Ma S, Shao E, Xie X, Liu W (2020) Event detection in soccer video based on self-attention. In 2020 IEEE 6th international conference on computer and communications (ICCC). IEEE, pp 1852-1856
39. Yan L, Han C, Xu Z, Liu D, Wang Q (2023) Prompt learns prompt: exploring knowledge-aware generative prompt collaboration for video captioning. In proceedings of international joint conference on artificial intelligence (IJCAI), pp 1622-1630. https://doi.org/10.24963/ijcai.2023/180
40. Yan L, Ma S, Wang Q, Chen Y, Zhang X, Savakis A, Liu D (2022) Video captioning using global-local representation. IEEE Trans Circ Syst Video Technol 32(10):6642–6656
41. Cheema Y, Cheema MN, Nazir A, Khokhar FA, Li P, Ahmed A (2024) A novel approach for improving open scene text translation with modified GAN. Vis Comput 1-13. https://doi.org/10.1007/s00371-024-03371-w
42. Nergård Rongved OA, Stige M, Hicks SA, Thambawita VL, Midoglu C, Zouganeli E, Johansen D, Riegler MA, Halvorsen P (2021) Automated event detection and classification in soccer: the potential of using multiple modalities. Mach Learn Knowl Extract 3(4):1030–1054
43. Liu D, Cui Y, Tan W, Chen Y (2021) Sg-net: Spatial granularity network for one-stage video instance segmentation. In proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9816-9825. https://doi.org/10.48550/arXiv.2103.10284
44. Sen A, Deb K (2022) Categorization of actions in soccer videos using a combination of transfer learning and gated recurrent unit. ICT Express 8(1):65–71
45. Naik BT, Hashmi MF (2023) YOLOv3-SORT: detection and tracking player/ball in soccer sport. J Electron Imaging 32(1):011003–011003
46. Vats K, Fani M, Walters P, Clausi DA, Zelek J (2020) Event detection in coarsely annotated sports videos via parallel multi-receptive field 1D convolutions. In proceedings of the IEEE/CVF

conference on computer vision and pattern recognition workshops, pp 882-883. https://doi.org/10.48550/arXiv.2004.06172

47. Giancola S, Ghanem B (2021) 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. https://doi.org/10.48550/arXiv.2104.06779
48. Ekin A, Tekalp AM, Mehrotra R (2003) Automatic soccer video analysis and summarization. IEEE Trans Image Process 12(7):796–807
49. Sigari M-H, Soltanian-Zadeh H, Pourreza H-R (2015) Fast highlight detection and scoring for broadcast soccer video summarization using on-demand feature extraction and fuzzy inference. Int J Comput Graph 6(1):13–36
50. Rani JSJ, Augasta MG (2020) A deep learning model for human re-identification with split LOMO and deep features using XQDA. Test Eng Manag 83:13776 (**ISSN 0193-4120**)
51. Hussain SJ, Chen Z, Mudassar R, Lin M (2016) Color based pre-rank categorization for person re-identification. In 2016 international conference on intelligent control and computer application (ICCA 2016). Atlantis Press, pp 293-96
52. Shu C, Hu X (2023) Improved image style transfer based on VGG-16 convolutional neural network model. J Phys: Conf Ser 2424(1):012021 (**IOP Publishing**)
53. Steffi DD, Mehta S, Venkatesh K, Dasari SK (2022) HOG-based object detection toward soccer playing robots. In computer vision and robotics: proceedings of CVR 2021. Springer, pp 155-163
54. Zanganeh A, Jampour M, Layeghi K (2022) IAUFD: A 100k images dataset for automatic football image/video analysis. IET Image Proc 16(12):3133–3142
55. Fakhar B, Rashidy Kanan H, Behrad A (2019) 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
56. Jiang Y, Cui K, Chen L, Wang C, Xu C (2020) Soccerdb: A large-scale database for comprehensive video understanding. In proceedings of the 3rd international workshop on multimedia content analysis in sports, pp 1-8. https://doi.org/10.48550/arXiv.1912.04465
57. Liu N, Liu L, Sun Z (2022) Football game video analysis method with deep learning. Comput Intell Neurosci 2022:1
58. Narayana Darapaneni PK, Malhotra N, Sundaramurthy V, Thakur A, Chauhan S, Thangeda KC, Paduri AR (2022) Detecting key soccer match events to create highlights using computer vision. https://doi.org/10.48550/arXiv.2204.02573

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.