arXiv:2309.12067v1 [cs.CV] 21 Sep 2023

SURVEY OF ACTION RECOGNITION, SPOTTING AND SPATIO-TEMPORAL LOCALIZATION IN SOCCER - CURRENT TRENDS AND RESEARCH PERSPECTIVES

A PREPRINT

Karolina Seweryn
NASK - National Research Institute,
Warsaw, Poland
karolina.seweryn@nask.pl

Anna Wróblewska
Faculty of Mathematics and Information Science,
Warsaw University of Technology,
Warsaw, Poland
anna.wroblewska1@pw.edu.pl

Szymon Łukasik
Faculty of Physics and Applied Computer Science,
AGH University of Science and Technology,
Cracow, Poland
slukasik@agh.edu.pl

September 16, 2025

ABSTRACT

Action scene understanding in soccer is a challenging task due to the complex and dynamic nature of the game, as well as the interactions between players. This article provides a comprehensive overview of this task divided into action recognition, spotting, and spatio-temporal action localization, with a particular emphasis on the modalities used and multimodal methods. We explore the publicly available data sources and metrics used to evaluate models' performance. The article reviews recent state-of-the-art methods that leverage deep learning techniques and traditional methods. We focus on multimodal methods, which integrate information from multiple sources, such as video and audio data, and also those that represent one source in various ways. The advantages and limitations of methods are discussed, along with their potential for improving the accuracy and robustness of models. Finally, the article highlights some of the open research questions and future directions in the field of soccer action recognition, including the potential for multimodal methods to advance this field. Overall, this survey provides a valuable resource for researchers interested in the field of action scene understanding in soccer.

Keywords action recognition, action spotting, soccer datasets, spatio-temporal action localization, modality fusion, multimodal learning

1 Introduction

Soccer is one of the most popular and lucrative sports worldwide, with billions of fans and many players. In recent years, there has been an increasing interest in using computer vision and machine learning techniques to automatically extract information from match recordings to get valuable insights about the strengths and weaknesses of teams. Understanding the actions that occur during a match is essential for both coaches and players to improve performance and gain a competitive edge. Similarly, scouts visit sports clubs to evaluate the performance and actions of young players to identify those with the most talent that could later be transferred to higher leagues. Automatic retrieval of such information could support scouts' decisions, saving money and time. There are many possible applications of this process in the television industry. For example, the ability to recognize game actions can enable producers to optimize

and automate the broadcast production process, emphasizing key aspects of the game to enhance spectator engagement. It is particularly valuable for real-time camera selection, post-game studio soccer analytics, and automatic highlights generation.

Action scene understanding has become an increasingly important area of research in the context of soccer [Giancola et al., 2018, Deliège et al., 2021, Li et al., 2021]. It poses unique challenges due to the complex and dynamic nature of the game. Players move quickly and often obscure each other, making it difficult to accurately track their movements. Moreover, soccer matches involve a wide range of actions, from simple passes to complex passages of play, tackles and shots on goal, which require different levels of analysis. With recent advancements in machine learning and computer vision, researchers have been exploring various approaches to improve the accuracy of action recognition in soccer. In particular, multimodal methods, which combine data from different sources such as video, audio, and other data, have shown promise in improving the accuracy and robustness of action recognition systems. These methods mirror how humans understand the world by utilizing multiple senses to process data. Employing multiple heterogeneous sources to train models presents both challenges and opportunities. The potential advantage is the improvement of model performance compared to unimodal representation, as incorporating additional modalities provides new information and reveals previously unseen relationships. However, multimodal learning also presents certain difficulties. Information from various sources can be redundant, and this should be filtered out in data representation as one vector. Some solutions build models for all modalities and then create a model combining the predictions. Another approach is to prepare an appropriate joint feature representation. The video data recorded during soccer games often includes information about fans' reactions and audio commentary. Also, Internet websites provide games and player statistics, live comments, and textual data with team analysis. Thus, soccer data can be valuable for researchers experimenting with multimodal learning.

This survey provides a comprehensive overview of recent research on action scene understanding, including action recognition (classification of actions in the trimmed video), spotting (detection and classification of actions in the untrimmed video) and spatio-temporal action localization (classification and tracking of specific actions and objects) in soccer, with a particular focus on available modalities and multimodal approaches. We explore the different data sources used in these methods, the various feature extraction and fusion techniques, and the evaluation metrics used to assess their performance. Also, we discuss the challenges and opportunities in this field, as well as the limitations and future research directions. By analyzing the state-of-the-art methods and identifying their strengths and weaknesses, this survey aims to provide a clear and up-to-date overview of the progress made in action recognition in soccer and to provide insights for researchers in this rapidly evolving area.

The following are the main contributions of this comprehensive literature review:

- We define three tasks in the area of soccer action understanding: action recognition, spotting, and spatio-temporal action localization, along with metrics used to assess the performance of these models.
- We prepare a list of soccer datasets for action understanding, highlighting the potential of applying multimodal methods.
- We examine a variety of existing state-of-the-art models in action recognition, spotting, and spatio-temporal action localization used in soccer as described in the literature.
- Based on the thorough assessment and in-depth analysis, we outline a number of key unresolved issues and future research areas that could be helpful for researchers.

The article is organized as follows. Subsection 1.1 highlights why this survey is important and what distinguishes it from others, while a discussion on the potential of using multimodal sources during training soccer models can be found in Section 1.2. Section 2 describes the research strategy. Tasks related to soccer action scene understanding are described in Section 3, and associated metrics and datasets are introduced in Section 4 and Section 5 respectively. Section 6 presents methods addressing the three analysed tasks. Future research directions, including the potential for multimodal methods, are discussed in the last section.

1.1 Motivation

Several publications listed in Table 1 have appeared in recent years reviewing machine learning systems that address the needs of the sports industry. Surveys [Thomas et al., 2017, Rahmad et al., 2018, Naik et al., 2022, Wu et al., 2022] show applications of computer vision to automatic analysis of various sports, two of which focus on the action recognition task. However, these surveys do not provide a comprehensive analysis of action detection in soccer, as different sports have different game motions and action types. Therefore, a detailed analysis of dedicated datasets and methods specific to soccer is necessary. While there are articles that focus on soccer [D'Orazio and Leo, 2010, Oskouie et al., 2012, Patil et al., 2014, Akan and Varlı, 2022], three of them were published before the release of relevant datasets

Table 1: List of previous similar reviews. ● denotes articles related to soccer, while ■ indicated articles about sport in general.

Review	Topic
A review of vision-based systems for soccer video analysis [D'Orazio and Leo, 2010]	●
Multimodal feature extraction and fusion for semantic mining of soccer video: A survey [Oskouie et al., 2012]	●
A survey on event recognition and summarization in football Videos [Patil et al., 2014]	●
Computer vision for sports: current applications and research topics [Thomas et al., 2017]	■
A Survey of Video Based Action Recognition in Sports [Rahmad et al., 2018]	■
A comprehensive review of computer vision in sports: open issues, future trends and research directions [Naik et al., 2022]	■
A survey on video action recognition in sports: datasets, methods and applications [Wu et al., 2022]	■
Use of deep learning in soccer videos analysis: survey [Akan and Varlı, 2022]	●

SoccerNet, SoccerNet-v2, and SoccerNet-v3 [Giancola et al., 2018, Deliège et al., 2021, Cioppa et al., 2022], which caused significant development of this field, including transformer-based solutions. Regarding action recognition, these articles describe binary models that classify only certain actions, such as goals or offside. A comprehensive review of various tasks was published in article [Wu et al., 2022] in 2022. Only one reported action recognition solution is evaluated on SoccerNet-v2 [Deliège et al., 2021], more specifically benchmark result is reported. Also, spatio-temporal action localization is not described in this publication. The potential of using multimodal inputs is only briefly described in previous surveys. Only one publication [Oskouie et al., 2012] addresses this topic; however, the methods outlined therein are not considered state-of-the-art nowadays. Mentioned publications focus mainly on action classification (recognition), while action localization in time or spatio-temporal space is more relevant in real-life scenarios.

1.2 Potential of Using Multimodality

Soccer is a sport that generates a vast amount of data, including videos, information on players, teams, matches, and events. Numerous matches are documented, and these recordings provide significant insights into the game. They vary in terms of video quality, used device and camera perspective (drone, single-camera, multiple cameras). Beyond the raw video feed, many characteristics can be extracted from video streams, including optical flow, players and ball trajectories, players' coordinates. When audio is available, it could be an additional data source capturing people's reactions, which can be represented using techniques like Mel Spectrogram [Xu et al., 2005]. Also, commentary data can be transcribed using automatic speech recognition systems such as Whisper [Radford et al., 2022] or vosk[1]. Unstructured texts containing match reports can be scrapped from various websites. Soccer clubs gather and analyse a vast amount of information about players to gain an advantage and choose the optimal tactics. Data from GPS systems, accelerometers, gyroscopes, or magnetometers can also be useful in soccer data analysis.

Combining many modalities is proven to achieve better results than using unimodal representations [Nagrani et al., 2021]. Soccer match recordings already contain many data types with predictive potential: audio, video, and textual transcriptions. Additional information about the outcomes of games or the time of individual events is also regularly reported on many websites. To sum up, investigating multimodal approaches is a logical step in soccer data analysis due to the accessibility of diverse data sources. As far as we know, there is no survey on the application of modality fusion in action recognition, spotting, and spatio-temporal action localization in sports videos. This survey provides a comprehensive overview of the current state of research and helps advance the field of action understanding in soccer by identifying areas for future research.

[1] `https://alphacephei.com/vosk/`

2 Definition of Research Strategy

The articles published between 2000 and 2022 are included in this survey. In order to find related articles, we used online databases such as Scopus [2], ScienceDirect [3], ACM [4], IEEE Xplore [5], SpringerLink [6] and SemanticScholar [7] and keyword search. The primary search keys were: *multimodal, multimodality, action recognition, sport, activity recognition, event recognition, event classification, action spotting, event detection, modality fusion, video, audio, text, activity localization, spatio-temporal action recognition, football, soccer, action localization, soccer dataset, soccernet*. Each query produced several articles. Additionally, we manually added many papers to our list by analyzing the references of the papers we identified. According to their relevance and year of publication, some of them were excluded from this analysis.

3 Problem Description

3.1 Actions

According to [Oskouie et al., 2012], soccer actions can be divided based on their relevance into primary and secondary events as depicted in Figure 1. The primary events directly affect the match's outcome and can directly cause goal opportunities, while the secondary actions are less important and do not affect the result of the match as much.

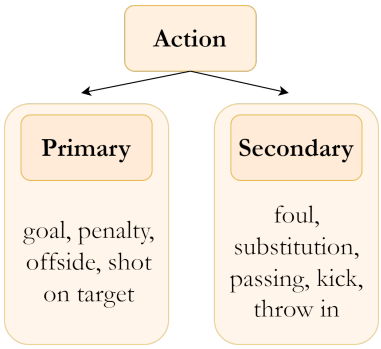


Figure 1: Difference between primary and secondary actions.

Action analysis can also be divided into three tasks that differ in their use of information about the time and localization of the event: action recognition, action spotting, and spatio-temporal action detection. Figure 2 highlights differences between mentioned tasks.

Action recognition, also known as **action/event classification**, is the classification of an event in a trimmed video. The model receives as input a series of frames; for the entire series, it has to predict which class the video refers to. In contrast, action spotting is a slightly different task which involves identifying the segment of the untrimmed video in which the action occurs, and then classifying it into predefined categories. An action can be temporally localized by defining a boundary that contains a start and end timestamp or a single frame timestamp, such as the start of the action. **Action spotting** is also referred to as **temporal action detection** or **temporal action localization**. In addition, we can distinguish an extension of this problem that incorporates actor location data called **spatio-temporal action detection (localization)**. This task aims to detect temporal and spatial information about action as moving bounding boxes of players. This involves detecting the location, timing, and type of actions performed by players, such as passing, shooting, tackling, dribbling, and goalkeeping. This task is particularly relevant in the analysis of an individual's behaviour and performance.



Figure 2: Comparison of tasks related to action analysis. Frames used in this visualization are from SoccerNet [Deliège et al., 2021] and MultiSports [Li et al., 2021] datasets.

3.2 Multimodality

Although we intuitively know what multimodality is, the formal definitions differ. One perspective is more human-centred and refers to the way people perceive the world, and other definitions are more related to machine-readable representations. According to Baltrušaitis et al. [2018], *"Our experience of the world is multimodal – we see objects, hear sounds, feel the texture, smell odours, and taste flavours. Modality refers to the way in which something happens or is experienced and a research problem is characterized as multimodal when it includes multiple such modalities."* (definition 1: human-centered) while Guo et al. [2019] define modality as *"a particular way or mechanism of encoding information"* (definition 2: machine-centered). Therefore, different encodings for one source (e.g. BERT embeddings and TFIDF encoding for text) would be considered multimodal according to definition 2, but not according to definition 1. This inconsistency was noted in [Parcalabescu et al., 2021], where authors proposed a new task-relative definition of multimodality: *"A machine learning task is multimodal when inputs or outputs are represented differently or are composed of distinct types of atomic units of information"*. Figure 3 highlights that inputs can be processed differently by people and machines. For instance, while people may perceive identical meaning in textual content presented in either text or image form, machine-extracted data can indicate significant dissimilarities. In this work, we analyse both human-centred and machine-centred multimodal models.

4 Metrics

The most common metric measuring classification performance among all analysed articles is mAP (mean average precision). In the case of action spotting avg-mAP is used, and for spatio-temporal action detection, Video-mAP@δ and Frame-mAP@δ are used. This section explains and summarizes various metrics used to measure the accuracy of the model's predictions.

[2] https://www.scopus.com/
[3] https://www.sciencedirect.com/
[4] https://dl.acm.org/
[5] https://ieeexplore.ieee.org/Xplore/home.jsp
[6] https://link.springer.com/
[7] https://www.semanticscholar.org/

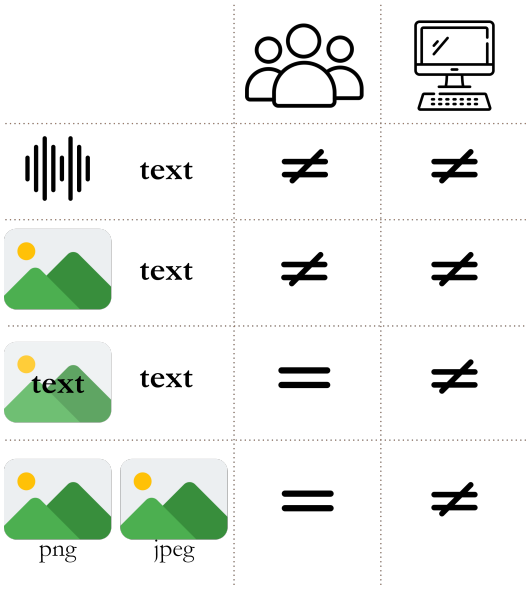


Figure 3: Difference between extracting information by people and machines. = denotes the same data, while $\neq$ means different. Visualization inspired by Parcalabescu et al. [2021].

4.1 Action recognition

Precision measures the model's ability to detect only relevant actions in a video. It is defined as the fraction of correctly classified samples (TP) out of all positive predictions (TP + FP):

$$Precision = \frac{TP}{TP + FP}.$$

Recall describes how well the model detects all relevant samples. It is defined as the fraction of correctly classified samples (TP) out of all positive ground truth (TP + FN):

$$Recall = \frac{TP}{TP + FN}.$$

F1-score is the harmonic mean of recall and precision:

$$F1\text{-}score = 2 \cdot \frac{precision \cdot recall}{precision + recall}.$$

The objective of the model is to achieve both high precision and recall. However, in practice, a trade-off between these metrics is chosen. The relationship between them can be depicted through **Precision-Recall (PR) curve**, which illustrates the values of precision on the y-axis and recall on the x-axis for different thresholds of model's confidence score. The term **AP** is an abbreviation for average precision and refers to the area under the precision-recall curve computed across all recall values. It is worth mentioning that a high AP value indicates a balance of both high precision and recall. Given that the PR curve often exhibits a zigzag pattern, calculating the area under the curve (AUC) accurately can be challenging. Therefore, various interpolation techniques, such as 11-points interpolation [Zhang and Zhang, 2009] or interpolation based on all points, are commonly employed. 11-points precision-recall curve interpolates precision at 11 recall levels (0.0, 0.1, ..., 1.0). The interpolation of precision ($P_{interp}(R)$ at recall level R is defined as the maximum precision with a recall value greater or equal than R and is defined in the equation below.

$$P_{interp}(R) = max_{R' \geq R} P(R').$$

Thus, the estimation of AP value can be defined as the mean of interpolated precisions over 11 recall values:

$$AP_{11} = \frac{1}{11} \sum_{R \in \{0.0, 0.1, \ldots, 1.0\}} P_{interp}(R)$$

Interpolation based on all points takes all points into account. AP is computed as the weighted mean of precisions at each recall (R) level. The weights are differences between current and previous recall values.

$$AP_{all} = \sum_{n=1}^{N} (R_n - R_{n-1}) P_{interp}(R),$$

where

$$P_{interp}(R_n) = max_{R' \geq R_n} P(R').$$

mean AP (mAP) measures the accuracy of a classifier over all classes and is defined as

$$mAP = \frac{1}{|\Gamma|} \sum_{\gamma \in \Gamma} AP(\gamma),$$

where Γ is a set of all possible classes and $AP(\gamma)$ denotes AP of class γ.

Top-N Accuracy is a fraction of correct predictions, where prediction is assigned as correct when any of the model's N highest probability scores match the expected answer (ground truth). In other words, it measures the proportion of correctly classified instances where the predicted class is among the N most probable classes. Top-N accuracy is a useful metric in situations where the exact class prediction may not be critical, but rather the identification of a set of probable classes that may be relevant.

4.2 Action spotting

Unlike the classic classification, action spotting also considers the time in which the action takes place. This aspect is also reflected in the metrics used to assess the models.

avg-mAP is the average of mAP for different tolerance values and can be defined as

$$avg\text{-}mAP = \frac{1}{|\Theta|} \sum_{\theta \in \Theta} mAP(\theta),$$

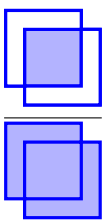
where Θ is a set of different tolerances. Authors of [Deliège et al., 2021] divided this metric into two subgroups: **loose avg - mAP** for tolerances 5s, 10s,..., 60s, and tight-avg-mAP with tolerance 1s, 2s, ..., 5s. If the high accuracy of action localization is required then **tight-avg-mAP** will be more appropriate.

4.3 Spatio-temporal action detection

Not only the time aspect but also the localization (bounding box coordinates) is important in spatio-temporal action detection. This section begins by introducing an assessment of the accuracy of the location of the action in a frame and then describes the associated metrics.

Intersaction over Union (IoU), also called Jaccard Index, measures the overlap of the predicted bounding box and the ground truth bounding box. IoU is invariant to the scale which means that the similarity between two shapes is not affected by the scale of their space.

$$IoU = \frac{\text{Area of Overlap}}{\text{Area of Union}} = \frac{P \cap GT}{P \cup GT} =$$

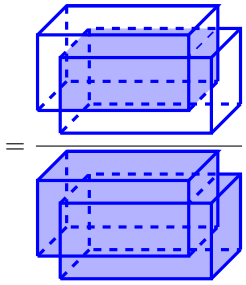


,

where P denotes prediction and GT means ground truth. An action is considered correctly classified and localized if the IoU between the predicted and the ground-truth bounding boxes is above a threshold δ.

3D Intersection over Union (3D IoU) The intersection over union in 3D is computed as an overlap between two cuboids.

$$3D\ IoU = \frac{\text{Volume of Overlap}}{\text{Volume of Union}} = \frac{P \cap GT}{P \cup GT} =$$



In the study introducing MultiSports dataset [Li et al., 2021] they defined 3D IoU (spatio-temporal-IoU) as IoU over the temporal domain multiplied by the average of the IoU between the overlapped frames (also used in [Singh et al., 2022, Weinzaepfel et al., 2015]).

MABO (Mean Average Best Overlap) **ABO** (Average Best Overlap) [Uijlings et al., 2013, Kalogeiton et al., 2017] for class c is defined as

$$ABO(c) = \frac{1}{|G^c|} \sum_{g_i^c \in G^c} max_{l_j \in L} IoU(g_i^c, l_j),$$

where G^c is a set of all ground truths for class c and L is predicted bounding box. The intersection over union (IoU) between every ground truth and predicted boxes (or tubes) is computed. Then, for each ground-truth box or tube, the overlap of the detection with the highest IoU value is retained (best-overlapping detection BO). Next, for every class, an average of all maximum intersections is calculated. The mean of **ABO** over all classes is called **MABO**.

Video-mAP@δ and Frame-mAP@δ Two groups of metrics can be considered to evaluate spatio-temporal action detectors: frame and video level. In the case of frame-level, metrics such as AP are computed for defined IoU threshold δ. Prediction is considered correct if its IoU with a ground truth box is greater than a given threshold and the predicted label matches the ground truth one. Then, mean AP is computed by averaging over all classes. In the literature, it is often referred to as **frame-mAP@δ**, **f@δ** or **f-mAP@δ**. 3D IoU between predicted tubes and ground truth tubes is often used to report video-level metrics. A model returns the correct tube if its 3D IoU with ground truth tube is above δ and correctly classified. Similarly, averaging the metrics (e.g. AP) over classes gives an overall metric such as **video-mAP@δ** (also denoted as **v-mAP@δ** and **v@δ**). By analogy, **Precision@k** and **Recall@k** can be defined.

Motion mAP and Motion AP In article [Singh et al., 2022], new metrics considering motion have been introduced. Actions are divided into three categories based on their motion size (large, medium, and small). With these labels, Average Precision (AP) is computed for each motion category. Computing the AP for each action class and then averaging the results for motion categories is referred to as the **Motion-mAP** while computing the AP for the motion categories regardless of action class is called the **MotionAP**. They are calculated at video and frame levels.

5 Datasets

The potential to use machine learning to analyse tactics and statistics in sports has automatically resulted in a significant increase in publicly available datasets [Giancola et al., 2018, Deliège et al., 2021, Tsunoda et al., 2017, Karimi et al., 2021, Li et al., 2021, Pappalardo et al., 2019]. Television broadcasters record sports games along with commentary, while online platforms offer detailed information and player statistics pertaining to the game. Thus, creating a multi-modal database should be relatively easy and intuitive. Table 2 examines the availability of different modalities in published soccer datasets for action recognition, spotting and spatio-temporal action localization.

Table 2: Modalities in soccer datasets with annotations of action detection task. Circles mean different tasks: ● - action classification, ■ - action spotting, ◀ - spatio-temporal action localization, ▶ - tracking, ▼ - camera shot segmentation, ◆ - replay grounding, ⬢ - highlight detection, ▲ - object detection.

	Dataset	Task	Video/ Photo	Audio	Publicly Available
soccer	SoccerNet [Giancola et al., 2018]	● ■	✓	✓	Yes
	SoccerNet-v2 [Deliège et al., 2021]	● ■ ▼ ◆	✓	✓	Yes
	SoccerNet-v3 [Cioppa et al., 2022]	● ■ ▼ ◆ ▲	✓	✓	Yes
	Football Action [Tsunoda et al., 2017]	● ▶	✓	?	No
	Comprehensive Soccer [Yu et al., 2018]	● ■ ▼ ▶	✓	✗	Yes
	SSET [Feng et al., 2020]	● ■ ▼ ▶	✓	✗	Yes
	SoccerDB [Jiang et al., 2020]	● ■ ▲ ⬢	✓	✗	Yes
	Soccer-logs [Pappalardo et al., 2019]	● ■	✗	✗	Yes
	VisAudSoccer [Gao et al., 2020]	● ◆	✓	✓	No
	SEV [Karimi et al., 2021]	●	✓	✗	Yes
	EIGD-S [Biermann et al., 2021]	● ■	✓	✓	Yes
multi sports	MultiSports [Li et al., 2021]	◀ ▶	✓	✗	Yes

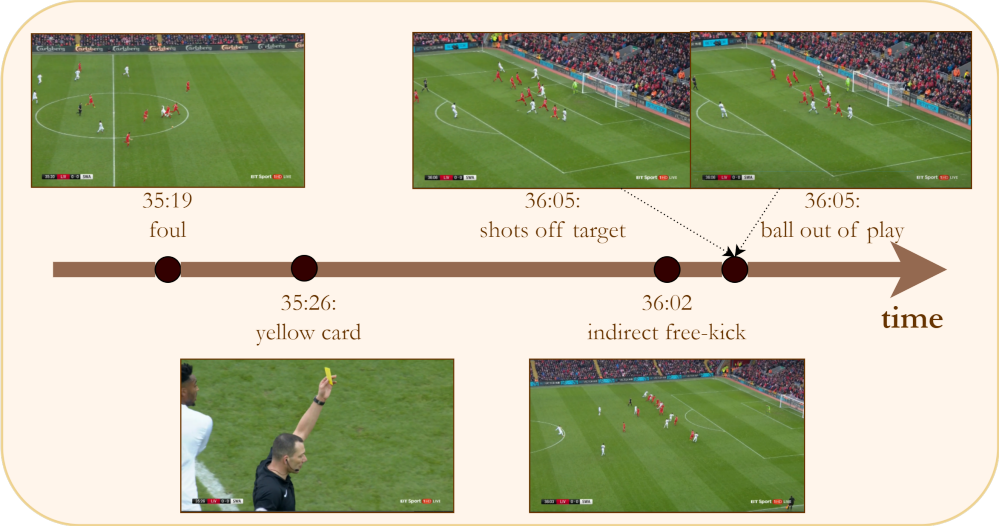


Figure 4: Examples of actions from SoccerNet-v2 dataset [Deliège et al., 2021]. Frames come from the match between Liverpool and Swansea (2017-01-21 - 15:30).

5.1 Soccer Datasets

SoccerNet SoccerNet[8] [Giancola et al., 2018] was introduced as a benchmark dataset for action spotting in soccer. The dataset consists of 500 soccer matches from the main European Championships (764 hours in total) and annotations of three action types: goal, card, and substitution. Each match is divided into two videos: one for each half of the match. The matches are split into a train (300 observations), test (100 observations), and validation datasets (100 observations). Also, the authors published extra 50 observations without labels for the challenge task.

SoccerNet-v2 SoccerNet-v2 [Deliège et al., 2021] enriched the SoccerNet dataset by manually annotating 17 action categories. In contrast to its predecessor, actions occur around 17 times more frequently (8.7 events per hour in SoccerNet, 144 actions per hour in SoccerNet-v2). Table 3 shows action types and their frequency. Each action has an assigned visibility category (shown and unshown) to indicate whether an action was shown in the broadcast video. Detecting unshown actions is very difficult and requires understanding the temporal context of actions. Figure 4

[8] `https://www.soccer-net.org/data`

Table 3: Number of actions in SoccerNet-v2 dataset [Deliège et al., 2021].

Action	#Train	#Test	#Valid	Total
Ball out of play	19097	6460	6253	31810
Clearance	4749	1631	1516	7896
Corner	2884	999	953	4836
Direct free-kick	1379	382	439	2200
Foul	7084	2414	2176	11674
Goal	995	337	371	1703
Indirect free-kick	6331	2283	1907	10521
Kick-off	1516	514	536	2566
Offside	1265	416	417	2098
Penalty	96	41	36	173
Red card	34	8	13	55
Shots off target	3214	1058	984	5256
Shots on target	3463	1175	1182	5820
Substitution	1700	579	560	2839
Throw-in	11391	3809	3718	18918
Yellow card	1238	431	378	2047
Yellow->red card	24	14	8	46

presents examples of actions for a selected match. Moreover, this dataset includes manual annotations for camera shot segmentation with boundary detection and replay grounding task. It is worth mentioning that SoccerNet-v2 recordings include audio commentary.

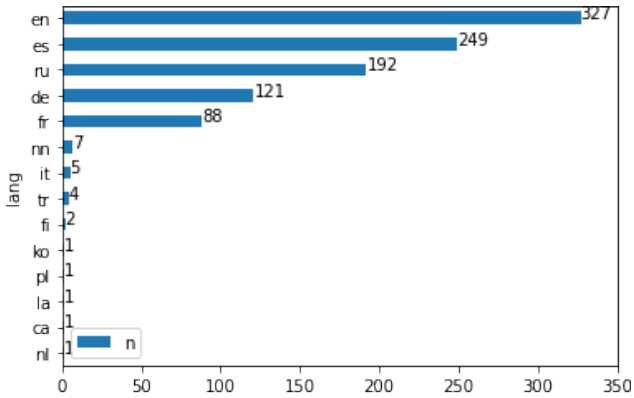


Figure 5: Distribution of commentary languages detected by Whisper [Radford et al., 2022] in SoccerNet-v2 [Deliège et al., 2021] dataset.

We analysed audio commentary in the SoccerNet-v2 with ASR (Automatic Speech Recognition) model called whisper [Radford et al., 2022]. We noticed that 37 out of 1000 observations (3.7%) do not have audio commentary. Figure 5 shows the distribution of detected languages. Each half of the match is analyzed as a separate observation. Sometimes, the first half of the game's commentary is in a different language than the second.

SoccerNet-v3 SoccerNet-v3 [Cioppa et al., 2022] is an extension of SoccerNet-v2 [Deliège et al., 2021] containing spatial annotations and associations between different view perspectives. Action annotations have been enriched with associated frames from the replay clips ($21,222$ of instances have been added). Therefore, it enables the exploration of multi-view action analysis. Also, they added lines and goals annotations, bounding boxes of players and referees ($344,660$ of instances), bounding boxes of objects including ball, flag, red/yellow card ($26,939$ of instances), multi-view player correspondences ($172,622$ of instances) and jersey numbers ($106,592$ of instances).

Football actions The dataset [Tsunoda et al., 2017] consists of two match recordings (each lasting 10 minutes) from 14 cameras located in different places. Five actions are manually annotated: pass, dribble, shoot, clearance, and loose

Table 4: Comparison of action types in action spotting datasets. *means that the background class (a category that does not belong to the main classes of interest) is not counted as a separate class.

Action		SoccerNet	SoccerNet-v2	ComprSoccer	SSET	SoccerDB
overhead kick		✗	✗	✓	✓	✗
solo drive		✗	✗	✓	✓	✗
goal		✓	✓	✓	✓	✓
shot	on target	✗	✓	✓	✓	✓
	off target		✓			
corner		✗	✓	✓	✓	✓
free kick	direct	✗	✓	✓	✓	✓
	indirect		✓			
penalty kick		✗	✗	✓	✓	✓
card	red	✓	✓	✓	✓	✓
	yellow		✓	✓	✓	
foul		✗	✓	✓	✓	✓
offside		✗	✓	✓	✓	✗
substitution		✓	✓	✗	✗	✓
ball out of play		✗	✓	✗	✗	✗
throw-in		✗	✓	✗	✗	✗
clearance		✗	✓	✗	✗	✗
kick off		✗	✓	✗	✗	✗
penalty		✗	✓	✗	✗	✗
yellow->red card		✗	✓	✗	✗	✗
injured		✗	✗	✗	✗	✓
saves		✗	✗	✗	✗	✓
corner&goal		✗	✗	✓	✓	✗
corner&shot		✗	✗	✓	✓	✗
free kick&goal		✗	✗	✓	✓	✗
free kick&shot		✗	✗	✓	✓	✗
#classes		3	17	11(+4)	11(+4)	10*
Duration [hours]		764	764	170	282	669
#events		6,637	110,458	6,850	10,619	37,715
Freq [#events/hour]		8.7	144.6	40.3	30.3	56.4

ball. Additionally, the authors provided annotations of the ball and players' 3D positions. Examples of images from this dataset are presented in Figure 6.



Figure 6: Example from Football actions dataset. Source: [Tsunoda et al., 2017].

Comprehensive Soccer Comprehensive Soccer dataset[9] [Yu et al., 2018, 2019] is a dataset containing 222 broadcast soccer videos (170 hours in total) in HD 720p (40% of observations) and 360p (60% of observations), 25 fps. They notice that most datasets focus on a single task, while a multi-task approach is necessary to analyse sports videos. Their dataset covers three tasks: shot boundary detection (far-view, medium-view, close-view, out-of-field view, playback shot), event detection and player tracking. They divided event annotation into two levels of granularity: event and story proposing 11 action classes: overhead kick, solo drive, goal, shot, corner, free kick, penalty kick, red card, yellow card,

[9] `http://media.hust.edu.cn/dataset/Datasets.htm`

foul, offside, and extra four-story labels: corner&goal, corner&shot, free kick&goal, free kick&shot. While action describes a single activity, the story provides a comprehensive narrative with contextual background (see Figure 7 for more details). They suggest that shot analysis can be essential to action analysis because various views can show different perspectives. For instance, a close-view shot can capture players, coaches and audiences when an event is happening, while far-view present tactics and the arrangement of players in the attacking and defensive team.

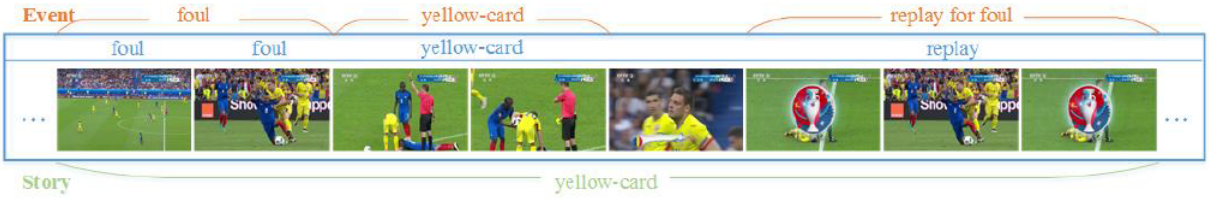


Figure 7: Difference between event and story from Comprehensive Soccer dataset. Source: [Yu et al., 2018].

SSET SSET dataset[10] [Feng et al., 2020] is an extension of **Comprehensive Soccer dataset [Yu et al., 2018]**. The authors have enriched the previous dataset with 128 recordings and increased the number of event annotations. Finally, the introduced dataset consists of 350 videos lasting 282h in total.

SoccerDB SoccerDB[11] [Jiang et al., 2020] is a dataset with annotations of four tasks: object detection, action recognition, action spotting, and video highlight detection for 343 soccer matches divided into 171, 191 video segments (some of them are from the SoccerNet dataset). It is worth mentioning that bounding boxes of players and the ball are available but not assigned to player numbers, so this dataset cannot be used for player tracking.

Although SoccerNet, SoccerNet-v2, Comprehensive Soccer, SSET, and SoccerDB are designed for the same task, the defined action labels differ. A comparison of available classes and their statistics can be found in Table 4.

Soccer-logs Soccer-logs[12] [13] [Pappalardo et al., 2019] is a large-scale dataset of temporal and spacial soccer events provided by Wyscout. Although they released a huge dataset with over 3 million events (100 times more than the largest open-source dataset SoccerDB), video analysis is hindered because the video files have not been available. Besides events data, authors provide files including annotations of competitions, matches, teams, players, referees and coaches.

SEV dataset It [14] [Karimi et al., 2021] consists of 42000 event-related images split into train, test, and validation data. The dataset includes annotations for 7 soccer events: corner kick, penalty kick, free kick, red card, yellow card, tackle, and substitute.

EIGD-S EIGD-S [15] [Biermann et al., 2021] is a dataset consisting of five soccer matches recordings with gold standard annotations for 125 minutes. The dataset includes multiple annotations for two matches from 4 experts and one inexperienced annotator. URLs of videos link to YouTube, where videos with audio paths are available. Annotation was prepared according to the proposed taxonomy assuming the hierarchical structure of events [Biermann et al., 2021]. Unlike the other datasets, EIGD-S contains annotations of high-level events, such as passes along with low-level events, including goals or cards.

VisAudSoccer Gao et al. [2020] proposed a new dataset (here denoted as **VisAudSoccer**), which contained data from 460 soccer game broadcasts, including 300 videos downloaded from SoccerNet [Giancola et al., 2018], lasting about 700 hours in total in video format. Audio data is available for 160 games with commentator voices categorized as "excited" and "not-excited". Events are divided into four classes: celebrate (1320 events), goal/shoot (1885 events), card (2355 events), and pass (2036 events). The dataset is not publicly available.

[10] `http://media.hust.edu.cn/dataset.htm`
[11] `https://github.com/newsdata/SoccerDB`
[12] `https://sobigdata-soccerchallenge.it/`
[13] `https://figshare.com/collections/Soccer_match_event_dataset/4415000`
[14] `https://github.com/FootballAnalysis/footballanalysis`
[15] `https://github.com/mm4spa/eigd`

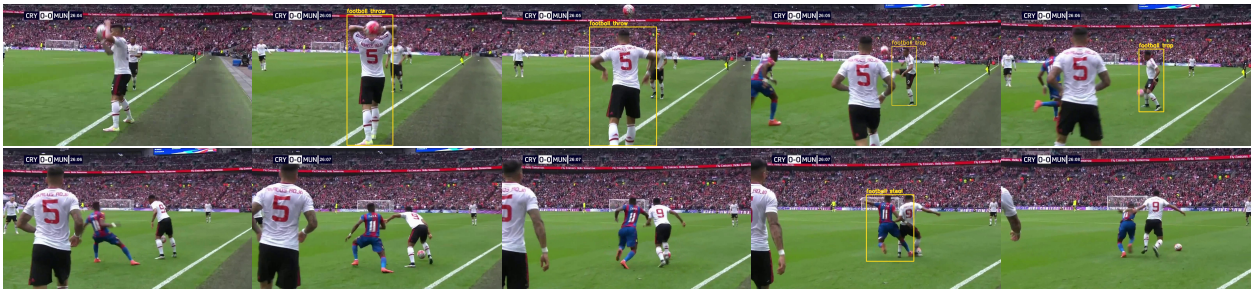


Figure 8: Examples of annotations from MultiSports [Li et al., 2021] dataset. Players participating in a given action are annotated with bounding boxes.

SoccerSummarization Gautam et al. [2022] [16] extended SoccerNet-v2 [Deliège et al., 2021] with news, commentaries and lineups from BBC.

5.2 Multi-Sports Datasets

MultiSports MultiSports[17] [Li et al., 2021] released spatio-temporal multi-person action detection dataset for basketball, volleyball, soccer, and aerobic gymnastics. After consulting with athletes, they proposed 66 action labels, e.g. soccer pass, trap, defence, tackle, and long ball. Additionally, a handbook was created to define actions and their temporal boundaries. Videos were downloaded from YouTube and then trimmed into shorter clips. In the end, 800 clips are available for each sport, which amounts to around 5 hours of recordings for soccer. The number of relevant action categories for soccer equals 15, and 12,254 instances were annotated. The authors emphasize that their dataset differs from other datasets due to its complexity, high quality, and diversity of videos. In the case of soccer, it is the first publically available dataset that contains spatio-temporal action annotations. Examples of MultiSports annotations can be found in Figure 8.

5.3 Other Datasets

UCF Sports UCF Sports[18] [Rodriguez et al., 2008, Soomro and Zamir, 2014] is a dataset for spatio-temporal action recognition with 10 possible classes: diving, golf swing, kicking, lifting, riding a horse, running, skateboarding, swinging-bench, swinging-side, and walking. 150 videos lasting about 6.39 seconds were gathered from broadcast television networks like ESPN and BBC. UCF Sports is known as one of the first datasets that published not only action class annotations but also bounding boxes of areas associated with the action. It was broadly used to conduct action classification experiments [Lui and Beveridge, 2011, Bregonzio et al., 2010, Wu et al., 2011, O'Hara and Draper, 2012] and spatio-temporal action recognition. It differs a lot from the task described in MultiSports dataset [Li et al., 2021] where the video is not temporally trimmed, multiple players and actions can be detected, and single action occurs only in a small subset of time. However, UTF-Sports initiated the advancement of this domain and inspired authors to develop interesting solutions. Due to the fact that this dataset contains only a few events related to soccer (kicking and running), the results and methods have not been widely described in this article.

UCF-101 The UCF-101 [Soomro et al., 2012] dataset was introduced as an action recognition dataset of realistic action videos from YouTube. 2 out of 101 action categories are related to soccer: soccer juggling and soccer penalty. Other action categories include diving, typing, bowling, applying lipstick, knitting etc.

Goal Tsagkatakis et al. [2017] proposed a dataset to classify a single *goal* class in soccer. The dataset consists of videos from YouTube: 200 2-3 second-long videos for *goal* class and 200 videos for *no-goal* class.

Offside Dataset A dataset[19] [Panse and Mahabaleshwarkar, 2020] that can be used to assess the effectiveness of a methodology for offside detection. It consists of about 500 frames that are publicly available. The authors highlight that this dataset has been carefully curated to include a diverse range of soccer match scenes demonstrating the different challenges such a system may encounter.

[16] `https://github.com/simula/soccer-summarization`

[17] `https://deeperaction.github.io/datasets/multisports.html`

[18] `https://www.crcv.ucf.edu/data/UCF_Sports_Action.php`

[19] `https://github.com/Neerajj9/Computer-Vision-based-Offside-Detection-in-Soccer`



Figure 9: Examples from artificially generated SoccER dataset [Morra et al., 2020].

Table 5: Methods used for action recognition in analysed articles. Features are represented as ● - image, ■ - audio.

Article	Dataset	Method	Features	mAP	Top-1 Acc
[Giancola et al., 2018]	SoccerNet	AvgPool	●	40.7	-
[Giancola et al., 2018]	SoccerNet	MaxPool	●	52.4	-
[Giancola et al., 2018]	SoccerNet	NetVLAD	●	67.8	-
[Giancola et al., 2018]	SoccerNet	NetRVLAD	●	67.4	-
[Giancola et al., 2018]	SoccerNet	NetFV	●	64.4	-
[Giancola et al., 2018]	SoccerNet	SoftBOW	●	62.0	-
[Vanderplaetse and Dupont, 2020]	SoccerNet	AudioVid	● ■	73.7	-
[Gan et al., 2022]	SoccerNet-v2	PM	● ■	-	62.4
[Gao et al., 2020]	VisAudSoccer	I3D [Carreira and Zisserman, 2017]	●	95.2	90.1
[Gao et al., 2020]	VisAudSoccer	I3D-NL [Wang et al., 2018a]	●	96.9	92.5
[Gao et al., 2020]	VisAudSoccer	ECO [Zolfaghari et al., 2018]	●	96.3	92.2
[Gao et al., 2020]	VisAudSoccer	SlowFast [Feichtenhofer et al., 2018]	●	95.1	88.1

SoccER Soccer Event Recognition [20] [Morra et al., 2020] is a synthetic dataset consisting of 500 minutes of game recordings gathered from the open source Gameplay Football engine that can be an approximation of real game. 1.6 million atomic events and $9,000$ complex events are annotated. Atomic events (kicking the ball, ball possession, tackle, ball deflection, ball out, goal, foul, penalty) are spatio-temporally annotated. Complex events occur over a wide area, involve multiple participants, or can be composed of multiple other events (pass, pass then goal, filtering pass, filter pass then goal, cross, cross then goal, tackle, shot, shot then goal, saved shot). Examples can be found in Figure 9.

GOAL GrOunded footbAlL commentaries (**GOAL**), dataset [Suglia et al., 2022] contains 1107 game recordings transcribed to text. Although this dataset is dedicated to tasks such as commentary retrieval or commentary generation, its contents can also be valuable as an additional modality in action recognition.

6 Methods

6.1 Action Recognition and Spotting

Action analysis in soccer has been an important task and has attracted many researchers. The first articles extracted video features and, based on that, classified clips into predefined categories using rule-based algorithms or classical machine learning models.

[20] `https://gitlab.com/grains2/slicing-and-dicing-soccer`

Table 6: Methods used for action spotting in analysed articles. Features are represented as ● - image, ■ - audio, ● - graph. * denotes that model was evaluated on the challenge dataset.

Article	Dataset	Method	Features	Avg mAP	Tight Avg mAP	Acc
[Deliège et al., 2021]	SoccerNet	NetVLAD	●	49.7	-	-
[Vanderplaetse and Dupont, 2020]	SoccerNet	AudioVid	● ■	56.0	-	-
[Cioppa et al., 2020]	SoccerNet	CALF	●	62.5	-	-
[Vats et al., 2020]	SoccerNet	MTTCNN	●	60.1	-	-
[Norgård Rongved et al., 2020]	SoccerNet	3dCNN	●	32.0	-	-
[Ma et al., 2020]	SoccerNet	NetVLAD + self-attention	●	-	-	74.3
[Tomei et al., 2021]	SoccerNet	RMS-Net	●	65.5	-	-
[Rongved et al., 2021]	SoccerNet	2D-CNN AudVid	● ■	-	-	90.85
[Mahaseni et al., 2021]	SoccerNet	CNN + Dilated RNN	●	63.3	-	-
[Shi et al., 2022]	SoccerNet	Multiple Scene Encoder	●	66.8	-	-
[Karimi et al., 2022]	SoccerNet	CNN-GRU metric learning	●	64.9	-	-
[Deliège et al., 2021]	SoccerNet-v2	MaxPool	●	18.6	-	-
[Deliège et al., 2021]	SoccerNet-v2	NetVLAD	●	31.4	-	-
[Deliège et al., 2021]	SoccerNet-v2	AudioVid	● ■	40.7	-	-
[Deliège et al., 2021]	SoccerNet-v2	CALF	●	41.6	-	-
[Zhou et al., 2021]	SoccerNet-v2	Vidpress Sports	●	74.1	-	-
[Giancola and Ghanem, 2021]	SoccerNet-v2	NetVLAD++	●	53.4	-	-
[Zhu et al., 2022]	SoccerNet-v2	transformer	●	52.04*	-	-
[Cioppa et al., 2021]	SoccerNet-v2	CC+RN+FCL	● ▲	46.8	-	-
[Cartas et al., 2022]	SoccerNet-v2	RGB+Audio+ Graph	● ■ ▲	57.8	-	-
[Darwish and El-Shabrway, 2022]	SoccerNet-v2	STE	●	74.1	58.5	-
[Shi et al., 2022]	SoccerNet-v2	Multiple Scene Encoder	●	75.3	-	-
[Cao et al., 2022]	SoccerNet-v2	SpotFormer	●	76.1	60.9	-
[Hong et al., 2022]	SoccerNet-v2	E2E-Spot 800MF	●	74.1	61.8	-
[Chen et al., 2022]	SoccerNet-v2	Faster-TAD	●	-	54.1	-
[Soares et al., 2022]	SoccerNet-v2	DU+SAM+mixup	●	77.3	60.7	-
[Soares and Shah, 2022]	SoccerNet-v2	DU+SAM+mixup +Soft-NMS	●	78.5	65.1	-

Khan et al. [2018a] experimented with a short 5-minute long video, where events (ball possession and kicking) were classified with a rule-based system. The event detector took as an input bounding boxes of ball with associated confidence scores. Similarly, a rule-based system consulted with soccer experts was proposed in [Khaustov and Mozgovoy, 2020] to classify events such as ball possession, successful and unsuccessful passes, and shots on goal. It was evaluated on two datasets from Data Stadium and Stats Perform.

Initially, models relied mainly on feature engineering extracting semantic concepts in clips [Ye et al., 2005, Hosseini and Eftekhari-Moghadam, 2013, Kolekar and Sengupta, 2015, Raventós et al., 2015, Tavassolipour et al., 2014, Xie and Tong, 2011]. Colour, texture and motion are represented. Also, representation is enriched with mid-level features, including camera view labels, camera motion, shot boundary descriptions, object detections, counting players, grass ratio, play-break segmentation, dominant colour, or penalty area. Audio descriptors such as whistle information or MPEG audio features are also used [Hosseini and Eftekhari-Moghadam, 2013, Kolekar and Sengupta, 2015, Raventós et al., 2015, Kapela et al., 2015, Li et al., 2003, Xiong et al., 2003]. Particularly, audio keywords such as long-whistling, double-whistling (indicating foul), multi-whistling, excited commentator speech, and excited audience sounds can assist in detecting events such as a free kick, penalty kick, foul, and goal in soccer [Xu et al., 2003]. These features are fed to classifiers, such as SVM [Ye et al., 2005, Zhao et al., 2015, Sadlier and O'Connor, 2005], Hidden Markov models (HMM) [Qian et al., 2011, Itoh et al., 2013, Wang et al., 2004, Xiong, 2005, Pixi et al., 2010, Leonardi et al., 2004, Qian et al., 2010], bayesian networks [Tavassolipour et al., 2014, Huang et al., 2006], hierarchical Conditional Random Field [Nisha et al., 2009], or fuzzy logic [Song and Hagras, 2017]. Along with the development of science and access to better computing machines, video representation improved (VGG-16 backbone [Yu et al., 2019]), and classifiers became more complex, e.g. Long-short Term Memory (LSTM) [Fakhar et al., 2019, Tsunoda et al., 2017, Jiang et al., 2016, Yu et al., 2019], CNN [Hong et al., 2018, Khan et al., 2018b, Jiang et al., 2016], or GRU [Jiang et al., 2016].

A very interesting approach was investigated in [Xu et al., 2008, Lanagan and Smeaton, 2011], where data published on the Internet, including Twitter posts, were used to identify events in various games (like soccer and rugby). In [Tang et al., 2018], the authors use the live text of soccer matches as additional input to the model. Text model composed of TextCNN [Kim, 2014], LSTM with attention [Yang et al., 2016] and VDCNN [Conneau et al., 2017] detect events in time and classify them. Then, Optical Character Recognition (OCR) links video time to associated texts. If necessary, a video-model is employed to detect events. Another noteworthy method was proposed by Vidal-Codina et al. [2022], who utilised tracking data and a tree-based algorithm to detect events. A similar solution was suggested in [Richly et al., 2016] where positional data was employed to feed event classifiers such as SVM, K-Nearest Neighbors and Random Forest.

In [Giancola et al., 2018], authors introduced SoccerNet dataset together with benchmarks for action classification and spotting tasks. They achieved an average-mAP of 49.7% for a threshold ranging from 5 to 60 seconds in the spotting task. They compared different pooling layers (Average Pool, Max Pool, SoftDBOW [Philbin et al., 2008], NetFV [Lev et al., 2015, Perronnin and Larlus, 2015, Sydorov et al., 2014], NetVLAD [Arandjelović et al., 2015] and NetRVLAD [Miech et al., 2017]) and video representation (I3D[Carreira and Zisserman, 2017], C3D [Tran et al., 2015], and ResNet [He et al., 2015] features) with a sliding window approach at 0.5s stride (see Table 5). The same pooling methods were investigated in [Vanderplaetse and Dupont, 2020], where input was enriched with audio variables. The video was represented as ResNet features, and audio stream feature extractions were done with VGGish architecture (VGG [Simonyan and Zisserman, 2014] pretrained on AudioSet [Gemmeke et al., 2017]). It is worth mentioning that using modality fusion improved mAP of 7.43% for the action classification task and 4.19% for the action spotting task on the SoccerNet dataset. Experiments showed that mid-fusion was the most effective method (73.7%), while early fusion achieved the worst performance (64%). The result for late fusion is 68.4%. Similarly, the authors of [Rongved et al., 2021] conducted experiments with multimodal models combining video and audio features in various settings. These experiments prove that combing modalities can lead to an improvement in model performance. They also acknowledged that the highest gain was observed in the classification of goal class, which can be associated with the audio reaction of supporters. According to the authors of [Mahaseni et al., 2021], enhancing event spotting may be achieved significantly by including short-range to long-range frame dependencies within an architecture. They have introduced a novel approach based on a two-stream convolutional neural network and Dilated Recurrent Neural Network (DilatedRNN) with Long Short-Term Memory (LSTM) [Hochreiter and Schmidhuber, 1997]. The Two-Stream CNN captures local spatiotemporal features required for precise details, while the DilatedRNN allows the classifier and spotting algorithms to access information from distant frames. Ma et al. [2020] used the self-attention mechanism to extract key frames and the NetVLAD network to obtain the temporal window-level (60s) features. The results of the classifier trained on the SoccerNet [Giancola et al., 2018] have improved from 67.2% to 74.3% accuracy by adding an attention mechanism. ResNet3d pretrained on Kinetics-400 [Norgård Rongved et al., 2020] were found to be inferior to state-of-the-art models. However, the authors stated that this architecture is competitive in real-time settings or when the precise temporal localization of events is crucial. A novel loss function CALF (Context-Aware Loss Function) [Cioppa et al., 2020] for the segmentation model was introduced to improve action spotting model

training. Instead of focusing on a single timestamp, CALF analyses the temporal context around the action. Frames are grouped into categories: far before, just before, just after, far after an action, and transition zones with associated parameters. Outputs of segmentation block feed spotting layer with YOLO-like loss. This architecture has significantly outperformed the baseline model ($+12.8\%$ of avg-mAP).

RMS Net [Tomei et al., 2021] is a solution inspired by regression methods in object detection, combining classification and regression loss during the training. The model produces outputs that comprise both the probability of an action class and its corresponding temporal boundaries. The authors of this solution have also proposed masking procedures and implemented strategies to address the issue of data imbalance, which has led to an improvement in the mAP metric. They noticed that certain indicators of events tend to appear shortly after the event itself. By analyzing reactions, it is possible to infer whether an action has occurred or not. The masking procedure was constructed to focus only on frames occurring after an event which allows for a more targeted analysis of relevant video segments.

Karimi et al. [2022] implemented a Siamese neural network to conduct experiments of metric learning to detect soccer events. The most promising combination was Siamese Network with contrastive loss [Koch et al., 2015], Efficient-NetB0 [Tan and Le, 2021], and gated recurrent units (GRU) [Chung et al., 2014]. In [Vats et al., 2020], they introduced a multi-tower temporal convolutional neural network (MTTCNN) which considers that particular events occur with different frequencies in sports datasets.

Gao et al. [2020] proposed new action classification dataset including SoccetNet [Giancola et al., 2018] videos. This article presents benchmarks for the classification of four actions (goal/shoot, yellow/red card, celebrate and pass) with I3D [Carreira and Zisserman, 2017], I3D-NL [Wang et al., 2018a], ECO [Zolfaghari et al., 2018] and SlowFast [Feichtenhofer et al., 2018].

After the release of SoccerNet-v2 [Deliège et al., 2021], the action spotting task gained even more scientific interest [Cartas et al., 2022, Zhou et al., 2021, Darwish and El-Shabrway, 2022, Shi et al., 2022, Cao et al., 2022, Gan et al., 2022, Giancola and Ghanem, 2021, Hong et al., 2022, Soares et al., 2022]. Published benchmarks reached the average mAP for tolerances ranging from 5s to 60s metric of 41.6, and two years later, the result increased to 78.5 [Soares and Shah, 2022]. Table 6 presents results reported in analysed articles for action spotting task. It is worth noting that together with the increase in the number of classes (from 3 in SoccerNet to 17 in SoccerNet-v2), the task has been made more difficult. For instance, the performance of CALF decreased from 62.5 on SoccerNet to 41.6 on SoccerNet-v2.

Giancola and Ghanem [2021] proposes a novel architecture known as NetVLAD++, which is based on NetVLAD pooling method. Similarly to [Cioppa et al., 2020], they take into consideration both frames prior to the action (past) and frames subsequent to the action (future). The authors noted that certain actions share the same characteristics prior to the event but differ in what happens after the action. They provided the goal and shot as an example, highlighting that both actions have the same pre-event characteristics, but can be differentiated based on the events that follow the action. Future and past clips are processed independently as two levels of temporal context utilizing NetVLAD pooling layers. This approach surpasses the previous state of the art, notably outperforming traditional NetVLAD [Deliège et al., 2021] by 22 percentage points.

Darwish and El-Shabrway [2022] [21] presented two versions of Spatio-Temporal Encoder (STE) build by convolution layers, max pooling and fully connected layers. The architecture is distinguished by the speed of model training and low complexity while maintaining good performance. The first version of the solution achieves 74.1% avg-mAP on the test SoccerNet-v2 dataset and 40.1% of tight avg-mAP. The modified model increased tight avg-mAP to 58.5%. Input frames are represented with Baidu features [Zhou et al., 2021]. Model's layers are divided into spatial and temporal representations. For the temporal encoder, they proposed three different time scales to capture events: [T, T/2, 1]. STE-v2 differs from STE-v1 in the last layer. The first version output predicted action class while STE-v2 returns a prediction of frame index in addition to the label.

After the development of transformer [Vaswani et al., 2017] neural networks in computer vision [Neimark et al., 2021, Dosovitskiy et al., 2020], a large and growing body of literature has investigated these architectures in action spotting task [Zhou et al., 2021, Soares et al., 2022, Chen et al., 2022, Gan et al., 2022, Cao et al., 2022, Shi et al., 2022, Zhu et al., 2022].

Article [Zhou et al., 2021] suggested that video representation is crucial in automatic match analysis. Contrary to other solutions using ResNet and ImageNet features, they fine-tuned action recognition models such as TPN [Yang et al., 2020], GTA [He et al., 2020], VTN [Neimark et al., 2021], irCSN [Tran et al., 2019], and I3D-Slow [Feichtenhofer et al., 2018] on SoccerNet-v2 [22] in semantic feature extraction block (this set of features will be referred as Baidu soccer features). Then, features from all extractors are concatenated into one vector. NetVLAD++ and transformer-based

[21] `https://github.com/amdarwish/SoccerEventSpotting/tree/main/STE-v2`

[22] Features can be downloaded from `https://github.com/baidu-research/vidpress-sports`.

models are used to detect actions. Experiments showed that clip representation using all five feature types is much better than single-backbone. Transformer architecture with all features achieves 73.7% of avg-mAP, while NetVLAD++ trained on the same representation achieves 74.1% on test SoccerNet-v2. Furthermore, they proved that the proposed video representation is better than previously used ResNet features.

Inspired by work of [Zhou et al., 2021], the study of [Cao et al., 2022] also represented video features as a fusion of multiple backbones. In addition to Baidu features, they chose action recognition models such as VideoMAE [Tong et al., 2022] and Video Swin Transformer [Liu et al., 2021b] assuming that these models can extract task-specific features. On the top of each feature extractor is a multilayer perceptron that reduces dimensionality and retains main semantic information. Then, features are concatenated. The results of experiments show that applying dimensionality reduction improved the accuracy of the action spotter. The model consists of multiple stacked transformer encoder blocks that process each frame individually, followed by a feed-forward network to perform temporal segmentation. Finally, Soft-NMS [Bodla et al., 2017] is used to filter predictions. The network of [Zhou et al., 2021] has been improved with the following adjustments: an additional input feature normalization, a learnable position embedding is used instead of sin-cos position encoding, changes of some hyperparameters values, and focal loss is applied to address data imbalance. Summing up the results, using only Baidu features achieved 56.9% of tight avg-mAP, while enriching representation with VideoMAE and Video Swin Transformer increases the performance to 58.3% of tight avg-mAP. The final model has an average-mAP equal to 76.1% and a tight average-mAP equal to 60.9% on the test set of SoccerNet-v2.

Zhu et al. [2022] used transformer [Fan et al., 2021] to extract features from soccer videos. The architecture generated action proposals with a sliding window strategy. Then, the proposed clips are fed to a transformer-based action recognition module. Transformer representation is then processed by NetVLAD++[Giancola and Ghanem, 2021].

Multiple Scene Encoders architecture [Shi et al., 2022] uses a representation of multiple frames to spot action because some actions consist of subactions (e.g. goal can be represented as running, shooting and cheering). The paper reports 55.2% Average-mAP using ResNet features, and 75.3% with Baidu embedding features [Zhou et al., 2021], once again showing a significant advantage thanks to the appropriate video representation.

Multimodal transformer-based model using both visual and audio information through a late fusion was introduced in [Gan et al., 2022] for action recognition. The transformer model is designed to capture the action's spatial information at a given moment and the temporal context between actions in the video. The input to the model consists of the raw video frames and audio spectrogram from the soccer videos. Video streams are processed with ViViT transformer [Arnab et al., 2021] and audio with a model based on Audio Spectrogram Transformer [Gong et al., 2021]. Then, modalities representations are connected with the late fusion method as a weighted average of encoder results. Unlike the other articles modelling action spotting on the SoccerNet dataset, the authors of this article report results with the Top-1 Accuracy metric, so it is difficult to compare their results with other papers. However, the article has a broad analysis with reference to different architectures. The best analysed model trained exclusively on visual input achieved 60.4%, and the multimodal transformer proposed by [Gan et al., 2022] achieved Top-1 Accuracy equal to 62.4%.

Action-spotting models commonly rely on using pretrained features as input due to the computation difficulties of end-to-end solutions. Hong et al. [2022] offers a more efficient, end-to-end architecture called E2E-Spot for accurate action spotting. Each video frame is represented by RegNet-Y [Radosavovic et al., 2020] with Gate Shift Modules (GSM) [Sudhakaran et al., 2019]. Then, similarly to [Tsunoda et al., 2017], a recurrent network is used – the resulting data sequence is modelled through a GRU network [Chung et al., 2014], which creates a temporal context and generates frame-level class predictions.

As the development of action spotting solutions has progressed, the importance of accurate models capable of precisely localizing actions within untrimmed videos has become increasingly acknowledged [Chen et al., 2022, Soares et al., 2022, Cao et al., 2022]. The performance of these models is commonly evaluated using the tight-avg-mAP metric, which measures their effectiveness within a specific small tolerance range (1, . . . , 5 seconds).

Inspired by Faster-RCNN [Ren et al., 2015], authors of [Chen et al., 2022] built architecture called Faster-TAD. Features are extracted using SwinTransformer [Liu et al., 2021a]. Then, similarly to the Faster-RCNN approach, 120 best proposals of action boundaries are generated. The boundary-based model was implemented to take into account the variance in action duration, as some actions may last only a few seconds while others may extend for several minutes. Then, the modules for correcting the proposals of action location and their classification work in parallel. The authors proposed an advanced context module consisting of three blocks (a Proximity-Category Proposal Block, a Self-Attention Block, and a Cross-Attention Block) to get semantic information for classification purposes. Proximity-Category Proposal Block gathers contextual information, a Self-Attention Block establishes relationships among proposals, and finally, a Cross-Attention Block gathers relevant context from raw videos concerning proposals. Architecture is enriched with Fake-Proposal Block for action boundary refinement and atomic features for better clip representation. They report 54.1% of tight-avg-mAP, which is a 7.04% gain compared to [Zhou et al., 2021].

Soares et al. [Soares et al., 2022] also propose a solution that tackles the problem of imprecise temporal localization. The model returns detection confidence and temporal displacement for each anchor. The architecture consists of a feature extractor (ResNet-152 with PCA and Baidu soccer embeddings fine-tuned on SoccetNet-v2 [Zhou et al., 2021]) followed by MLP. Then, features are processed by u-net [Ronneberger et al., 2015] and transformer encoder. Additionally, they experimented with Sharpness-Aware Minimization (SAM) [Foret et al., 2021] and mixup data augmentation [Zhang et al., 2017]. Their solution significantly boosted performance with tight avg-mAP: from 54.1 achieved by [Chen et al., 2022] to 60.7. Then, they introduced improvements to this solution [Soares and Shah, 2022] and won SoccerNet Challenge 2022. First, they modified preprocessing by resampling the Baidu embeddings [Zhou et al., 2021] to get greater frame frequency (2 FPS) and applying late fusion to combine them with the ResNet features. Also, soft non-maximum suppression (Soft-NMS) [Bodla et al., 2017] was applied in the postprocessing step. These modifications resulted in a 4.4 percentage point improvement over [Soares et al., 2022] measured by tight avg-mAP.

Cioppa et al. [Cioppa et al., 2021] conducted experiments exploring the utilization of camera calibration data in action spotting. The first phase involved the implementation of an algorithm based on Camera Calibration for Broadcast Videos (CCBV) of [Sha et al., 2020]. Results of Mask R-CNN [He et al., 2017] model for object detection combined with camera calibration module allow preparing diverse feature sets, including top view representations with a 3D convolutional network, feature vectors representations (ResNet-34 [He et al., 2015] and EfficientNet-B4 [Tan and Le, 2019]), and a player graph representation with graph convolutional network DeeperGCN [Li et al., 2020a]. In the graph, players are represented as nodes with edges connecting two players if the distance between them is less than 25 meters. SoccerNet-v2 labels were divided into *patterned* and *fuzzy* groups based on prior knowledge of the predictive potential of player localization data for classifying these labels. Player localization plays a crucial role in the classification of *patterned* classes (e.g. penalty, throw-in, cards) but is not relevant for *fuzzy* labels (substitution, ball out of play, foul). Two separate CALF [Cioppa et al., 2020] networks were trained for each class group: one using calibration data to improve the classification of *patterned* classes, and the other using only ResNet features for *fuzzy* labels. They reported an avg-mAP of 46.8%, outperforming the unimodal CALF by 6.1 percentage points.

Although the use of graphs had already found its application in sports analysis [Qi et al., 2020, Passos et al., 2011, Stöckl et al., 2021, Buldú et al., 2018], Cioppa et al. [Cioppa et al., 2021] were the first to use graph-based architecture to spot actions in untrimmed soccer videos. Then, Cartas et al. [Cartas et al., 2022] developed another graph based-architecture that resulted in a substantial improvement over its predecessor, achieving an average mAP of 57.8%. Similarly to the previous solution, players are represented as nodes with attributes such as location, motion vector and label (player team 1/2, goalkeeper 1/2, referee), and are connected to another player by edge based on proximity (less than 5 meters). Players and referees are detected by the segmentation module PointRend [Kirillov et al., 2019], and their position is projected onto a 2D pitch template through a homography matrix of camera calibration [Cioppa et al., 2021]. After data cleaning, referees and players are classified using a convolutional network supported by a rule-based algorithm and clustering. Player information is enriched with motion vector represented by preprocessed optical flow extracted with FlowNet 2.0 CSS [Ilg et al., 2016, Reda et al., 2017]. The model architecture consists of four dynamic edge graph CNN [Wang et al., 2018b] blocks followed by NetVLAD [Arandjelović et al., 2015] pooling layer. The authors experimented with multiple modalities and found that a graph-only model achieved 43.3% of Average-mAP while adding video features increased the metric to 51.5%. Furthermore, incorporating audio features from VGGish network [Simonyan and Zisserman, 2014] with both video and graph streams resulted in an average mAP of 57.8%, surpassing both unimodal and bimodal methods.

6.2 Spatio-Temporal Action Localization

The release of the MultiSports dataset [Li et al., 2021] for spatio-temporal localization of multiple sportsmen may contribute towards a better understanding of actions performed by individual players. Approaches to addressing this challenge can be categorized into frame-level and clip-level models [Li et al., 2021]. The frame-level models predict the bounding box and action type for each frame and then integrate these predictions. Conversely, clip-level methods, also called action tubelet detectors, endeavour to model both temporal context and action localization. Authors of the MultiSports dataset published benchmarks for the proposed task training frame-level models (ROAD [Singh et al., 2016], YOWO [Köpüklü et al., 2019]) and clip-level models (MOC [Li et al., 2020b], SlowOnly [Feichtenhofer et al., 2018] and SlowFast [Feichtenhofer et al., 2018]). Results of experiments are summarized in Table 7.

ROAD [Singh et al., 2016] is an algorithm for real-time action localization and classification which uses the Single Shot Multibox Detector (SSD) [Liu et al., 2015] method to independently detect and classify action boxes in each frame, without taking into account temporal information. Afterwards, the predictions from each frame are combined into action tubes through a novel algorithm. Similarly, You Only Watch Once (YOWO) [Köpüklü et al., 2019] method for identifying actions in real-time video streams links results from individual frames into action tubes through a dynamic programming algorithm. It uses two concurrent networks: a 2D-CNN to extract spatial features from key frames and

Table 7: Methods used for spatio-temporal action localization in analysed articles. ● means image.

Article	Dataset	Method	Features	f@0.5	v@0.2	v@0.5
[Li et al., 2021]	MultiSports	ROAD [Singh et al., 2016]	●	3.90	0.00	0.00
[Li et al., 2021]	MultiSports	YOWO [Köpüklü et al., 2019]	●	9.28	10.78	0.87
[Li et al., 2021]	MultiSports	MOC [Li et al., 2020b] (K=7)	●	22.51	12.13	0.77
[Li et al., 2021]	MultiSports	MOC [Li et al., 2020b] (K=11)	●	25.22	12.88	0.62
[Li et al., 2021]	MultiSports	SlowOnly Det., 4 × 16 [Feichtenhofer et al., 2018] (K=11)	●	16.70	15.71	5.50
[Li et al., 2021]	MultiSports	SlowFast Det., 4 × 16 [Feichtenhofer et al., 2018] (K=11)	●	27.72	24.18	9.65
[Singh et al., 2022]	MultiSports	TAAD + TCN	●	**55.3**	-	**37.0**
[Faure et al., 2022]	MultiSports	HIT	●	33.3	**27.8**	8.8

a 3D-CNN to extract spatio-temporal features from key frames and preceding frames. Then, the features from these two networks are combined through a channel fusion and attention mechanism and fed into a convolution layer to predict bounding boxes and action probabilities directly from video clips. Another approach was proposed in article [Li et al., 2020b] introducing Moving Center Detector (MOC detector). It models an action instance as a series of moving points and leverages the movement information to simplify and enhance detection. The framework consists of three branches: (1) Center Branch for detecting the center of the action instance and action classification, (2) Movement Branch for estimating the movement between adjacent frames to form a trajectory of moving points, and (3) Box Branch for predicting the size of the bounding box at each estimated center. They return tubelets, which are then linked into video-level tubes through a matching process. SlowFast [Feichtenhofer et al., 2018] comprises of two parallel branches. The slow branch identifies spatial semantics that exhibits minimal fluctuations, thus allowing for a low frame rate. Conversely, the fast branch is responsible for detecting rapid changes in motion, requiring a high frame rate to operate effectively. During training, data from the fast branch is fed to a slow neural network and at the end, the results of the two networks are concatenated into one vector. Faster R-CNN [Ren et al., 2015] with a ResNeXt-101-FPN [Lin et al., 2016, Xie et al., 2016] backbone was used to detect people. As the name suggests, the SlowOnly model uses only the slow path of SlowFast. Table 7 summarizes the results and indicates that the SlowFast detector achieved the best performance within benchmark models. Metrics are computed for all sports, not only for soccer.

The results obtained by Gueter Josmy Faure1 et al. in [Faure et al., 2022] [23] suggest that including pose information can be valuable to predict actions. Authors motivate their architecture with the fact, that actions can be defined as interactions between people and objects. Their multimodal Holistic Interaction Transformer Network (HIT) fusing a video stream and a pose stream surpasses other models on the MultiSports dataset. Each stream composes of person interaction, object interaction and hand interaction to extract action patterns. For each modality, Intra-Modality Aggregator (IMA) facilitates learning valuable action representations. Then, an Attentive Fusion Mechanism (AFM) is utilized to merge the various modalities, retaining the most significant features from each modality. 3D CNN backbone [Feichtenhofer et al., 2018] processes video frames, Faster RCNN [Ren et al., 2015] with ResNet-50-FPN [Xie et al., 2016, Lin et al., 2016] backbone predict bounding boxes and spatio transformer [Zheng et al., 2021] is pose encoder. This method outperformed others in terms of f-mAP@0.5 and v-mAP@0.2.

In existing solutions to spatio-temporal action recognition, tube detection involves extending a bounding box proposal at a keyframe into a 3D temporal cuboid and pooling features from nearby frames. However, this approach is not effective when there is significant motion. The study [Singh et al., 2022] propose cuboid-aware feature aggregation to model spatio-temporal action recognition. Also, they improve actor feature representation through actor tracking data and temporal feature aggregation along the tracks. The experiments show that the proposed method called Track Aware Action Detector (TAAD) outperforms others, especially for large-motion actions.

[23] `https://github.com/joslefaure/HIT`

6.3 Summarizing Multimodal Action Scene Understanding

Multimodal machine learning is a powerful approach combining different pieces of information to understand complex phenomena comprehensively. Fusing information from multiple modalities leads to a deeper comprehension of the underlying processes, enabling superior predictive performance compared to unimodal models. In the realm of soccer, the adoption of multimodal approaches has successfully improved the accuracy of predictive models through the integration of diverse sources of data and the extraction of more meaningful insights [Vanderplaetse and Dupont, 2020, Rongved et al., 2021, Cartas et al., 2022, Cioppa et al., 2021, Zhou et al., 2021].

Classical methods of action recognition relied mostly on data preparation and feature engineering. Thus, authors extracted different features from video clips, including logo frame detection, audio MPEG descriptors, camera motion, zoom indicator, colour layout, dominant colour, referee's whistle indicator, view category, and Histogram of Oriented Gradients (HoG) [Dalal and Triggs, 2005]. It is worth noting that these methods widely used a combination of audio and visual features [Hosseini and Eftekhari-Moghadam, 2013, Kolekar and Sengupta, 2015, Raventós et al., 2015, Kapela et al., 2015, Li et al., 2003, Xiong et al., 2003, Xu et al., 2003]. The more recent methods have primarily relied on visual embedding alone [Deliège et al., 2021, Zhou et al., 2021, Cioppa et al., 2020, Tomei et al., 2021]. However, using a fusion of multiple representations of a single source (e.g. Baidu embeddings for video [Zhou et al., 2021]), that can also be considered as multimodality, has proven to be more effective than using a single model to represent video (e.g. ResNet features). The release of Baidu soccer embeddings has resulted in researchers favouring them over ResNet features originally presented by SoccerNet authors. Moreover, Baidu embeddings were further extended by two additional models [Cao et al., 2022]. Experiments showed that incorporating audio streams [Vanderplaetse and Dupont, 2020, Rongved et al., 2021, Cartas et al., 2022], graph networks [Cartas et al., 2022, Cioppa et al., 2021] and optical flow [Cartas et al., 2022] can also provide significant value to the model. In the case of spatio-temporal action localization models, the fusion of a video stream and a pose stream surpassed other solutions [Faure et al., 2022].

7 Discussion

Automatic action recognition and localization is relevant from the perspectives of many in the soccer industry: coaches, scouts, fans and broadcasters. The Internet provides various sources of information on match results, highlights, and non-structured data. Also, many matches are recorded via different cameras, and TV and radio provide audio commentary to some matches. Therefore, it seems that soccer can be an excellent source of multimodal data. However, collecting high-quality and realistic soccer videos is complex and challenging for several reasons.

Datasets One of the main challenges in gathering annotated soccer data is that it can be challenging to obtain the necessary data due to licences and limited public availability of broadcast content. An even bigger challenge is assessing data for smaller or less popular leagues. Preparing annotated data is a laborious, time-consuming and expensive task because it involves manual annotation of video footage. This process may require a team of trained analysts to watch and annotate every match. The quality of the annotations may depend on the level of expertise of the analysts, which can further affect the accuracy of the data. To avoid this, a match is sometimes annotated by several people, with the recorded data being cross-referenced. Nevertheless, the interpretation of any event is always subjective and open to numerous interpretations. For instance, two analysts may disagree on whether a particular incident should be classified as a foul or not. Also, obtaining accurate frame level annotations of events is difficult due to inaccuracies in defining the beginning and end of the action. This subjectivity can result in inconsistencies in the data and make it difficult to compare or analyze different datasets.

To be useful and informative, soccer data must meet certain requirements that ensure its quality, reliability and usefulness to the industry. Annotations should be prepared in a standardized manner to ensure comparability and consistency. Firstly, current sport datasets provide access to trimmed short clips [Rodriguez et al., 2008], but this assumption is unrealistic. To apply models to real scenarios, they should be trained on whole untrimmed videos, such as SoccerNet [Giancola et al., 2018, Deliège et al., 2021]. Secondly, broadcast-like videos with moving camera and replays are the most natural to collect. It could be difficult and costly to gather videos from multiple cameras for the same scene; however, middle camera recordings or drone recordings could be valuable and useful for everyday training, even being affordable for smaller clubs.

Although there are already quite a few soccer datasets, there is still room for improvement. Action recognition is essential in the analysis of matches in lower leagues to find talented players. Unfortunately, many games are not broadcasted, and the installation of many cameras on the pitch is too expensive. Thus, datasets consisting of drone recordings or middle camera-only recordings and models trained on them could supplement the work of scouts. Moreover, even though there are a lot of multimodal soccer data sources, there is a lack of publicly available datasets, including them. The development of spatio-temporal action localization methods in soccer can lead to the easy highlighting of actions

performed by individual players. Furthermore, combining results with homography allows statistics per action to be computed. For instance, *"Player [] covers a distance [] while dribbling with an average speed of []"*. Despite the wide range of applications, there is no dedicated soccer dataset for this task. In contrast, authors of [Ibrahim et al., 2016] proposed a widely used volleyball dataset consisting of sportsman position and action annotation along with group activity labels.

Methods and Potential of Multimodality Soccer action scene understanding, which can be divided into action classification, spotting and spatio-temporal action localization, is a crucial aspect of analyzing soccer matches. The mentioned tasks vary in difficulty, with action classification being the easiest and action spotting being a more complex task that involves classification and finding temporal localization of actions. Spatio-temporal action localization makes the task more difficult by adding the spatial aspect. Considering the temporal context of actions is essential for several reasons. Firstly, some actions follow each other and information that one occurred should increase the probability of the associated action. For instance, yellow cards occur after fouls, which are then followed by free kicks. Moreover, analysis of frames before and after an action can provide valuable information about the occurrence of actions. Before a goal-scoring situation, one team runs towards the opposite goal in the video, and the audio track can capture the reactions of reporters and fans, including cheers or boos. Also, the results of an action can be deduced from what happens after the event. If there is a goal, players and their fans celebrate it. It can be challenging to train models to localize actions when sudden position changes and fast movement occur during attacks and dribbling. The duration of events can also vary significantly, such as the difference between the time taken for a yellow card and the ball being out of play.

Model performance on benchmark datasets has fluctuated over the past few years. Action spotting on SoccerNet [Giancola et al., 2018] has gained about 17 percentage points of average mAP in comparison to the baseline (from 49.7 to 66.8). Similarly, after the release of SoccerNet-v2 [Deliège et al., 2021] in 2021, containing 17 action classes, the average mAP increased from 18.6 to 78.5. A similar trend is observed on the MultiSports [Li et al., 2021] dataset for spatio-temporal action localization, where the best model published by dataset's authors was outperformed in terms of frame-mAP@0.5 (almost twice as good) and in terms of video-mAP@0.5 (four times as good) [Singh et al., 2022]. Huge improvements were made thanks to enriching data representation with other data sources, such as audio [Vanderplaetse and Dupont, 2020, Rongved et al., 2021, Cartas et al., 2022], graphs [Cartas et al., 2022, Cioppa et al., 2021], and pose [Singh et al., 2016]. Although SoccerNet includes reporter's commentary track, text input has not yet been used in modelling. A textual input can provide a plethora of valuable information as sports commentators describe the unfolding events on the pitch. Thus, work on the enriching of representation with text data is promising; experiments will be needed to verify this conjecture. Multimodal models can also be defined as a different representation of the same input, e.g. concatenation of various text embeddings. This approach was used in many articles [Zhou et al., 2021, Shi et al., 2022, Cao et al., 2022, Soares et al., 2022, Soares and Shah, 2022]. The most groundbreaking is the article of Baidu Research [Zhou et al., 2021] which published multiple semantic features of action recognition models, which were then used in other articles.

To sum up, action scene understanding in soccer has attracted much attention from research teams in recent years. Many publicly available datasets have been released, and models have improved the accuracy of action spotting and recognition. Nevertheless, some interesting and relevant problems remain to be addressed, including spatio-temporal action localization datasets and models dedicated to soccer and experiments with multimodal data such as textual commentary.

Acknowledgments

We would like to thank Olaf Skrabacz and Filip Boratyn for his valuable insights, comments and recommendations. This research was supported by NASK - National Research Institute.

Declarations

Availability of data and materials

Not applicable

Code availability

Not applicable

References

S. Akan and S. Varlı. Use of deep learning in soccer videos analysis: survey. *Multimedia Systems*, pages 1–19, 12 2022. doi:10.1007/s00530-022-01027-0.

R. Arandjelović, P. Gronát, A. Torii, T. Pajdla, and J. Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:1437–1451, 2015.

A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic, and C. Schmid. Vivit: A video vision transformer. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6816–6826, 2021.

T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

H. Biermann, J. Theiner, M. Bassek, D. Raabe, D. Memmert, and R. Ewerth. A unified taxonomy and multimodal dataset for events in invasion games. In *Proceedings of the 4th International Workshop on Multimedia Content Analysis in Sports*, MMSports'21, page 1–10, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450386708. doi:10.1145/3475722.3482792. URL `https://doi.org/10.1145/3475722.3482792`.

N. Bodla, B. Singh, R. Chellappa, and L. S. Davis. Soft-nms — improving object detection with one line of code. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5562–5570, 2017.

M. Bregonzio, J. Li, S. Gong, and T. Xiang. Discriminative topics modelling for action feature selection and recognition. In *Proceedings of the British Machine Vision Conference*, pages 8.1–8.11. BMVA Press, 2010. ISBN 1-901725-40-5. doi:10.5244/C.24.8.

J. M. Buldú, J. Busquets, J. H. Martínez, J. L. Herrera-Diestra, I. Echegoyen, J. Galeano, and J. Luque. Using network science to analyse football passing networks: Dynamics, space, time, and the multilayer nature of the game. *Frontiers in Psychology*, 9, 2018.

M. Cao, M. Yang, G. Zhang, X. Li, Y. Wu, G. Wu, and L. Wang. Spotformer: A transformer-based framework for precise soccer action spotting. In *2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6, 2022. doi:10.1109/MMSP55362.2022.9948888.

J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

A. Cartas, C. Ballester, and G. Haro. A graph-based method for soccer action spotting using unsupervised player classification. In *Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports*, MMSports '22, page 93–102, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450394888. doi:10.1145/3552437.3555691. URL `https://doi.org/10.1145/3552437.3555691`.

S. Chen, C. Chen, W. Li, X. Tao, and Y. Guo. Faster-tad: Towards temporal action detection with proposal generation and classification in a unified network, 2022. URL `https://arxiv.org/abs/2204.02674`.

J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014.

A. Cioppa, A. Deliège, S. Giancola, B. Ghanem, M. Van Droogenbroeck, R. Gade, and T. B. Moeslund. A context-aware loss function for action spotting in soccer videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

A. Cioppa, A. Deliège, F. Magera, S. Giancola, O. Barnich, B. Ghanem, and M. V. Droogenbroeck. Camera calibration and player localization in soccernet-v2 and investigation of their representations for action spotting. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4532–4541, 2021.

A. Cioppa, A. Deliège, S. Giancola, B. Ghanem, and M. Droogenbroeck. Scaling up soccernet with multi-view spatial localization and re-identification. *Scientific Data*, 9:355, 06 2022. doi:10.1038/s41597-022-01469-1.

A. Conneau, H. Schwenk, L. Barrault, and Y. Lecun. Very deep convolutional networks for text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1107–1116, Valencia, Spain, Apr. 2017. Association for Computational Linguistics. URL `https://aclanthology.org/E17-1104`.

N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1, 2005. doi:10.1109/CVPR.2005.177.

A. Darwish and T. El-Shabrway. Ste: Spatio-temporal encoder for action spotting in soccer videos. In *Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports*, MMSports '22, page 87–92, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450394888. doi:10.1145/3552437.3555704. URL `https://doi.org/10.1145/3552437.3555704`.

A. Deliège, A. Cioppa, S. Giancola, M. J. Seikavandi, J. V. Dueholm, K. Nasrollahi, B. Ghanem, T. B. Moeslund, and M. Van Droogenbroeck. Soccernet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4503–4514, 2021. doi:10.1109/CVPRW53098.2021.00508.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020.

T. D'Orazio and M. Leo. A review of vision-based systems for soccer video analysis. *Pattern Recognition*, 43 (8):2911–2926, 2010. ISSN 0031-3203. doi:https://doi.org/10.1016/j.patcog.2010.03.009. URL `https://www.sciencedirect.com/science/article/pii/S0031320310001299`.

B. Fakhar, H. Rashidy Kanan, and A. Behrad. 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, June 2019.

H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer. Multiscale vision transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6804–6815, 2021.

G. J. Faure, M.-H. Chen, and S.-H. Lai. Holistic interaction transformer network for action detection. *ArXiv*, abs/2210.12686, 2022.

C. Feichtenhofer, H. Fan, J. Malik, and K. He. Slowfast networks for video recognition. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6201–6210, 2018.

N. Feng, Z. Song, J. Yu, Y.-P. P. Chen, Y. Zhao, Y. He, and T. Guan. Sset: a dataset for shot segmentation, event detection, player tracking in soccer videos. *Multimedia Tools and Applications*, pages 1 – 22, 2020.

P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=6Tm1mposlrM`.

Y. Gan, R. Togo, T. Ogawa, and M. Haseyama. Transformer based multimodal scene recognition in soccer videos. In *2022 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–6, 2022. doi:10.1109/ICMEW56448.2022.9859304.

X. Gao, X. Liu, T. Yang, G. Deng, H. Peng, Q. Zhang, H. Li, and J. Liu. Automatic key moment extraction and highlights generation based on comprehensive soccer video understanding. In *2020 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–6, 2020. doi:10.1109/ICMEW46912.2020.9106051.

S. Gautam, C. Midoglu, S. Shafiee Sabet, D. B. Kshatri, and P. Halvorsen. Soccer Game Summarization using Audio Commentary, Metadata, and Captions. In *NarSUM '22: Proceedings of the 1st Workshop on User-centric Narrative Summarization of Long Videos*, pages 13–22. Association for Computing Machinery, New York, NY, USA, Oct. 2022. ISBN 978-1-45039493-2. doi:10.1145/3552463.3557019.

J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. Audio set: An ontology and human-labeled dataset for audio events. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, 2017.

S. Giancola and B. Ghanem. Temporally-aware feature pooling for action spotting in soccer broadcasts. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4485–4494, 2021.

S. Giancola, M. Amine, T. Dghaily, and B. Ghanem. Soccernet: A scalable dataset for action spotting in soccer videos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1792–179210, 2018. doi:10.1109/CVPRW.2018.00223.

Y. Gong, Y.-A. Chung, and J. R. Glass. Ast: Audio spectrogram transformer. *ArXiv*, abs/2104.01778, 2021.

W. Guo, J. Wang, and S. Wang. Deep multimodal representation learning: A survey. *IEEE Access*, 7:63373–63394, 2019.

B. He, X. Yang, Z. Wu, H. Chen, S.-N. Lim, and A. Shrivastava. Gta: Global temporal attention for video action understanding. In *British Machine Vision Conference*, 2020.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015.

K. He, G. Gkioxari, P. Dollár, and R. B. Girshick. Mask r-cnn. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:386–397, 2017.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–80, 12 1997. doi:10.1162/neco.1997.9.8.1735.

J. Hong, H. Zhang, M. Gharbi, M. Fisher, and K. Fatahalian. Spotting temporally precise, fine-grained events in video. In *European Conference on Computer Vision*, 2022.

Y. Hong, C. Ling, and Z. Ye. End-to-end soccer video scene and event classification with deep transfer learning. pages 1–4, 04 2018. doi:10.1109/ISACV.2018.8369043.

M.-S. Hosseini and A.-M. Eftekhari-Moghadam. Fuzzy rule-based reasoning approach for event detection and annotation of broadcast soccer video. *Applied Soft Computing*, 13(2):846–866, 2013. ISSN 1568-4946. doi:https://doi.org/10.1016/j.asoc.2012.10.007. URL https://www.sciencedirect.com/science/article/pii/S1568494612004565.

C.-L. Huang, H.-C. Shih, and C.-Y. Chao. Semantic analysis of soccer video using dynamic bayesian network. *IEEE Transactions on Multimedia*, 8(4):749–760, 2006. doi:10.1109/TMM.2006.876289.

M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori. A hierarchical deep temporal model for group activity recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1971–1980, 2016. doi:10.1109/CVPR.2016.217.

E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1647–1655, 2016.

H. Itoh, T. Takiguchi, and Y. Ariki. Event detection and recognition using hmm with whistle sounds. In *2013 International Conference on Signal-Image Technology & Internet-Based Systems*, pages 14–21, 2013. doi:10.1109/SITIS.2013.14.

H. Jiang, Y. Lu, and J. Xue. Automatic soccer video event detection based on a deep neural network combined cnn and rnn. pages 490–494, 11 2016. doi:10.1109/ICTAI.2016.0081.

Y. Jiang, K. Cui, L. Chen, C. Wang, and C. Xu. Soccerdb: A large-scale database for comprehensive video understanding. MMSports '20, page 1–8, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450381499. doi:10.1145/3422844.3423051. URL https://doi.org/10.1145/3422844.3423051.

V. S. Kalogeiton, P. Weinzaepfel, V. Ferrari, and C. Schmid. Action tubelet detector for spatio-temporal action localization. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4415–4423, 2017.

R. Kapela, K. McGuinness, A. Swietlicka, and N. E. O'Connor. Real-time event detection in field sport videos. In *Computer vision in Sports*, pages 293–316. Springer, 2015.

A. Karimi, R. Toosi, and M. A. Akhaee. Soccer event detection using deep learning. *arXiv preprint arXiv:2102.04331*, 2021.

A. Karimi, R. Toosi, and M. A. Akhaee. Soccer video event detection using metric learning. In *2022 12th International Conference on Computer and Knowledge Engineering (ICCKE)*, pages 048–052, 2022. doi:10.1109/ICCKE57176.2022.9959985.

A. Khan, B. Lazzerini, G. Calabrese, and L. Serafini. Soccer event detection. In *International Conference on Image Processing and Pattern Recognition (IPPR)*, pages 119–129, 04 2018a. doi:10.5121/csit.2018.80509.

M. Z. Khan, S. Saleem, M. A. Hassan, and M. U. G. Khan. Learning deep c3d features for soccer video event detection. *2018 14th International Conference on Emerging Technologies (ICET)*, pages 1–6, 2018b.

V. Khaustov and M. Mozgovoy. Recognizing events in spatiotemporal soccer data. *Applied Sciences*, 10(22), 2020. ISSN 2076-3417. doi:10.3390/app10228046. URL https://www.mdpi.com/2076-3417/10/22/8046.

Y. Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, Oct. 2014. Association for Computational Linguistics. doi:10.3115/v1/D14-1181. URL https://aclanthology.org/D14-1181.

A. Kirillov, Y. Wu, K. He, and R. B. Girshick. Pointrend: Image segmentation as rendering. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9796–9805, 2019.

G. Koch, R. Zemel, R. Salakhutdinov, et al. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2. Lille, 2015.

M. H. Kolekar and S. Sengupta. Bayesian network-based customized highlight generation for broadcast soccer videos. *IEEE Transactions on Broadcasting*, 61(2):195–209, 2015. doi:10.1109/TBC.2015.2424011.

O. Köpüklü, X. Wei, and G. Rigoll. You only watch once: A unified cnn architecture for real-time spatiotemporal action localization. *ArXiv*, abs/1911.06644, 2019.

J. Lanagan and A. Smeaton. Using twitter to detect and tag important events in sports media. 01 2011.

R. Leonardi, P. Migliorati, and M. Prandini. Semantic indexing of soccer audio-visual sequences: A multimodal approach based on controlled markov chains. *Circuits and Systems for Video Technology, IEEE Transactions on*, 14: 634 – 643, 06 2004. doi:10.1109/TCSVT.2004.826751.

G. Lev, G. Sadeh, B. Klein, and L. Wolf. Rnn fisher vectors for action recognition and image annotation. In *European Conference on Computer Vision*, 2015.

B. Li, H. Pan, and I. Sezan. A general framework for sports video summarization with its application to soccer. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, volume 3, pages III–169, 2003. doi:10.1109/ICASSP.2003.1199134.

G. Li, C. Xiong, A. K. Thabet, and B. Ghanem. Deepergcn: All you need to train deeper gcns. *ArXiv*, abs/2006.07739, 2020a.

Y. Li, Z. Wang, L. Wang, and G. Wu. *Actions as Moving Points*, pages 68–84. 10 2020b. ISBN 978-3-030-58516-7. doi:10.1007/978-3-030-58517-4_5.

Y. Li, L. Chen, R. He, Z. Wang, G. Wu, and L. Wang. Multisports: A multi-person video dataset of spatio-temporally localized sports actions. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13516–13525, 2021. doi:10.1109/ICCV48922.2021.01328.

T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. Feature pyramid networks for object detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2016.

W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*, 2015.

Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021a.

Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu. Video swin transformer. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3192–3201, 2021b.

Y. M. Lui and J. R. Beveridge. Tangent bundle for human action recognition. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 97–102, 2011. doi:10.1109/FG.2011.5771378.

S. Ma, E. Shao, X. Xie, and W. Liu. Event detection in soccer video based on self-attention. pages 1852–1856, 12 2020. doi:10.1109/ICCC51575.2020.9344896.

B. Mahaseni, E. R. M. Faizal, and R. G. Raj. Spotting football events using two-stream convolutional neural network and dilated recurrent neural network. *IEEE Access*, 9:61929–61942, 2021.

A. Miech, I. Laptev, and J. Sivic. Learnable pooling with context gating for video classification. *ArXiv*, abs/1706.06905, 2017.

L. Morra, F. Manigrasso, G. Canto, C. Gianfrate, E. Guarino, and F. Lamberti. Slicing and dicing soccer: Automatic detection of complex events from spatio-temporal data. In A. Campilho, F. Karray, and Z. Wang, editors, *Image Analysis and Recognition*, pages 107–121, Cham, 2020. Springer International Publishing.

A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun. Attention bottlenecks for multimodal fusion. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 14200–14213. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper/2021/file/76ba9f564ebbc35b1014ac498fafadd0-Paper.pdf`.

B. T. Naik, M. F. Hashmi, and N. D. Bokde. A comprehensive review of computer vision in sports: Open issues, future trends and research directions. *Applied Sciences*, 12(9), 2022. ISSN 2076-3417. doi:10.3390/app12094429. URL `https://www.mdpi.com/2076-3417/12/9/4429`.

D. Neimark, O. Bar, M. Zohar, and D. Asselmann. Video transformer network. *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3156–3165, 2021.

J. Nisha, S. Chaudhury, S. Dutta Roy, P. Mukherjee, K. Seal, and K. Talluri. A novel learning-based framework for detecting interesting events in soccer videos. pages 119 – 125, 01 2009. doi:10.1109/ICVGIP.2008.71.

O. A. Norgård Rongved, S. A. Hicks, V. Thambawita, H. K. Stensland, E. Zouganeli, D. Johansen, M. A. Riegler, and P. Halvorsen. Real-time detection of events in soccer videos using 3d convolutional neural networks. In *2020 IEEE International Symposium on Multimedia (ISM)*, pages 135–144, 2020. doi:10.1109/ISM.2020.00030.

S. O'Hara and B. A. Draper. Scalable action recognition with a subspace forest. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1210–1217, 2012. doi:10.1109/CVPR.2012.6247803.

P. Oskouie, S. Alipour, and A. M. Eftekhari Moghadam. Multimodal feature extraction and fusion for semantic mining of soccer video: A survey. *Artificial Intelligence Review*, 42, 08 2012. doi:10.1007/s10462-012-9332-4.

N. Panse and A. Mahabaleshwarkar. A dataset & methodology for computer vision based offside detection in soccer. MMSports '20, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450381499. doi:10.1145/3422844.3423055. URL https://doi.org/10.1145/3422844.3423055.

L. Pappalardo, P. Cintia, A. Rossi, E. Massucco, P. Ferragina, D. Pedreschi, and F. Giannotti. A public data set of spatio-temporal match events in soccer competitions. *Sci. Data*, 6(1):236, Oct. 2019.

L. Parcalabescu, N. Trost, and A. Frank. What is multimodality? In *Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR)*, pages 1–10, Groningen, Netherlands (Online), June 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.mmsr-1.1.

P. Passos, K. Davids, D. Araújo, N. S. C. Paz, J. Minguéns, and J. Mendes. Networks as a novel tool for studying team ball sports as complex social systems. *Journal of science and medicine in sport*, 14 2:170–6, 2011.

D. S. Patil, S. B. Waykar, International Journal of Science, Research (IJSR), and ijsr. A survey on event recognition and summarization in football videos. https://www.ijsr.net/get_abstract.php?paper_id=OCT14705, 2014. Accessed: 2023-2-23.

F. Perronnin and D. Larlus. Fisher vectors meet neural networks: A hybrid classification architecture. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3743–3752, 2015.

J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

Z. Pixi, L. Hongyan, and W. Wei. Research on event detection of soccer video based on hidden markov model. In *2010 International Conference on Computational and Information Sciences*, pages 865–868, 2010. doi:10.1109/ICCIS.2010.215.

M. Qi, Y. Wang, J. Qin, A. Li, J. Luo, and L. V. Gool. stagnet: An attentive semantic rnn for group activity and individual action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 30:549–565, 2020.

X. Qian, G. Liu, H. Wang, Z. Li, and Z. Wang. Soccer video event detection by fusing middle level visual semantics of an event clip. In G. Qiu, K. M. Lam, H. Kiya, X.-Y. Xue, C.-C. J. Kuo, and M. S. Lew, editors, *Advances in Multimedia Information Processing - PCM 2010*, pages 439–451, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. ISBN 978-3-642-15696-0.

X. Qian, H. Wang, G. Liu, and X. Hou. Hmm based soccer video event detection using enhanced mid-level semantic. *Multimedia Tools and Applications - MTA*, 60, 09 2011. doi:10.1007/s11042-011-0817-y.

A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL https://arxiv.org/abs/2212.04356.

I. Radosavovic, R. P. Kosaraju, R. B. Girshick, K. He, and P. Dollár. Designing network design spaces. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10425–10433, 2020.

N. Rahmad, M. A. As'ari, N. F. Ghazali, N. Shahar, and N. A. J. Sufri. A survey of video based action recognition in sports. *Indonesian Journal of Electrical Engineering and Computer Science*, 11:987–993, 09 2018. doi:10.11591/ijeecs.v11.i3.pp987-993.

A. Raventós, R. Quijada, L. Torres, and F. Tarrés. Automatic summarization of soccer highlights using audio-visual descriptors. *Springerplus*, 4(1):301, June 2015.

F. Reda, R. Pottorff, J. Barker, and B. Catanzaro. flownet2-pytorch: Pytorch implementation of flownet 2.0: Evolution of optical flow estimation with deep networks. https://github.com/NVIDIA/flownet2-pytorch, 2017.

S. Ren, K. He, R. B. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149, 2015.

K. Richly, M. Bothe, T. Rohloff, and C. Schwarz. Recognizing compound events in spatio-temporal football data. In *IoTBD*, pages 27–35, 2016.

M. D. Rodriguez, J. Ahmed, and M. Shah. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008. doi:10.1109/CVPR.2008.4587727.

O. A. N. Rongved, M. Stige, S. Hicks, V. L. Thambawita, C. Midoglu, E. Zouganeli, D. Johansen, M. Riegler, and P. Halvorsen. Automated event detection and classification in soccer: The potential of using multiple modalities. *Mach. Learn. Knowl. Extr.*, 3:1030–1054, 2021.

O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4.

D. Sadlier and N. O'Connor. Event detection in field sports video using audio-visual features and a support vector machine. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(10):1225–1233, 2005. doi:10.1109/TCSVT.2005.854237.

L. Sha, J. Hobbs, P. Felsen, X. Wei, P. Lucey, and S. Ganguly. End-to-end camera calibration for broadcast videos. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13624–13633, 2020. doi:10.1109/CVPR42600.2020.01364.

Y. Shi, H. Minoura, T. Yamashita, T. Hirakawa, H. Fujiyoshi, M. Nakazawa, Y. Chae, and B. Stenger. Action spotting in soccer videos using multiple scene encoders. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 3183–3189, 2022. doi:10.1109/ICPR56361.2022.9956667.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

G. Singh, S. Saha, M. Sapienza, P. H. S. Torr, and F. Cuzzolin. Online real-time multiple spatiotemporal action localisation and prediction. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3657–3666, 2016.

G. Singh, V. Choutas, S. Saha, F. Yu, and L. V. Gool. Spatio-temporal action detection under large motion. *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5998–6007, 2022.

J. C. V. Soares, A. Shah, and T. Biswas. Temporally precise action spotting in soccer videos using dense detection anchors. *ArXiv*, abs/2205.10450, 2022.

J. V. B. Soares and A. Shah. Action spotting using dense detection anchors revisited: Submission to the soccernet challenge 2022, 2022. URL `https://arxiv.org/abs/2206.07846`.

W. Song and H. Hagras. A type-2 fuzzy logic system for event detection in soccer videos. In *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6, 2017. doi:10.1109/FUZZ-IEEE.2017.8015426.

K. Soomro and A. Zamir. Action recognition in realistic sports videos. *Advances in Computer Vision and Pattern Recognition*, 71:181–208, 01 2014. doi:10.1007/978-3-319-09396-3_9.

K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *ArXiv*, abs/1212.0402, 2012.

M. Stöckl, T. Seidl, D. Marley, and P. Power. Making offensive play predictable-using a graph convolutional network to understand defensive performance in soccer. 2021.

S. Sudhakaran, S. Escalera, and O. Lanz. Gate-shift networks for video action recognition. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1099–1108, 2019.

A. Suglia, J. G. P. Lopes, E. Bastianelli, A. Vanzo, S. Agarwal, M. Nikandrou, L. Yu, I. Konstas, and V. Rieser. Going for goal: A resource for grounded football commentaries. *ArXiv*, abs/2211.04534, 2022.

V. Sydorov, M. Sakurada, and C. H. Lampert. Deep fisher kernels – end to end learning of the fisher kernel gmm parameters. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1402–1409, 2014.

M. Tan and Q. Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021.

M. Tan and Q. V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 2019. URL `http://proceedings.mlr.press/v97/tan19a.html`.

K. Tang, Y. Bao, Z. Zhao, L. Zhu, Y. Lin, and Y. Peng. Autohighlight : Automatic highlights detection and segmentation in soccer matches. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 4619–4624, 2018. doi:10.1109/BigData.2018.8621906.

M. Tavassolipour, M. Karimian, and S. Kasaei. Event detection and summarization in soccer videos using bayesian network and copula. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(2):291–304, 2014. doi:10.1109/TCSVT.2013.2243640.

G. Thomas, R. Gade, T. B. Moeslund, P. Carr, and A. Hilton. Computer vision for sports: Current applications and research topics. *Computer Vision and Image Understanding*, 159:3–18, 2017. ISSN 1077-3142. doi:https://doi.org/10.1016/j.cviu.2017.04.011. URL https://www.sciencedirect.com/science/article/pii/S1077314217300711. Computer Vision in Sports.

M. Tomei, L. Baraldi, S. Calderara, S. Bronzin, and R. Cucchiara. Rms-net: Regression and masking for soccer event spotting. *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 7699–7706, 2021.

Z. Tong, Y. Song, J. Wang, and L. Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *ArXiv*, abs/2203.12602, 2022.

D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. pages 4489–4497, 12 2015. doi:10.1109/ICCV.2015.510.

D. Tran, H. Wang, L. Torresani, and M. Feiszli. Video classification with channel-separated convolutional networks. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5551–5560, 2019.

G. Tsagkatakis, M. Jaber, and P. Tsakalides. Goal!! event detection in sports video. *Electronic Imaging*, 2017:15–20, 01 2017. doi:10.2352/ISSN.2470-1173.2017.16.CVAS-344.

T. Tsunoda, Y. Komori, M. Matsugu, and T. Harada. Football action recognition using hierarchical lstm. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 155–163, 2017. doi:10.1109/CVPRW.2017.25.

J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104:154–171, 2013.

B. Vanderplaetse and S. Dupont. Improved soccer action spotting using both audio and video streams. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3921–3931, 2020.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

K. Vats, M. Fani, P. Walters, D. A. Clausi, and J. S. Zelek. Event detection in coarsely annotated sports videos via parallel multi receptive field 1d convolutions. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3856–3865, 2020.

F. Vidal-Codina, N. Evans, B. Fakir, and J. Billingham. Automatic event detection in football using tracking data. *Sports Engineering*, 25, 09 2022. doi:10.1007/s12283-022-00381-6.

J. Wang, C. Xu, E. Chng, and Q. Tian. Sports highlight detection from keyword sequences using hmm. pages 599–602, 01 2004. doi:10.1109/ICME.2004.1394263.

X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018a.

Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (TOG)*, 38:1 – 12, 2018b.

P. Weinzaepfel, Z. Harchaoui, and C. Schmid. Learning to track for spatio-temporal action localization. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3164–3172, 2015.

F. Wu, Q. Wang, J. Bian, N. Ding, F. Lu, J. Cheng, D. Dou, and H. Xiong. A survey on video action recognition in sports: Datasets, methods and applications. *IEEE Transactions on Multimedia*, pages 1–25, 2022. doi:10.1109/TMM.2022.3232034.

X. Wu, D. Xu, L. Duan, and J. Luo. Action recognition using context and appearance distribution features. In *CVPR 2011*, pages 489–496, 2011. doi:10.1109/CVPR.2011.5995624.

S. Xie, R. B. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5987–5995, 2016.

W. Xie and M. Tong. A novel framework for soccer goal detection based on semantic rule. *Journal of Electronics (China)*, 28(4-6):670–674, 2011.

Z. Xiong. Audio-visual sports highlights extraction using coupled hidden markov models. *Pattern Anal. Appl.*, 8(1-2): 62–71, Sept. 2005.

Z. Xiong, R. Radhakrishnan, A. Divakaran, and T. Huang. Audio events detection based highlights extraction from baseball, golf and soccer games in a unified framework. In *2003 International Conference on Multimedia and Expo. ICME '03. Proceedings (Cat. No.03TH8698)*, volume 3, pages III–401, 2003. doi:10.1109/ICME.2003.1221333.

C. Xu, Y.-F. Zhang, G. Zhu, Y. Rui, H. Lu, and Q. Huang. Using webcast text for semantic event detection in broadcast sports video. *IEEE Transactions on Multimedia*, 10(7):1342–1355, 2008. doi:10.1109/TMM.2008.2004912.

M. Xu, N. Maddage, C. Xu, M. Kankanhalli, and Q. Tian. Creating audio keywords for event detection in soccer video. In *2003 International Conference on Multimedia and Expo. ICME '03. Proceedings (Cat. No.03TH8698)*, volume 2, pages II–281, 2003. doi:10.1109/ICME.2003.1221608.

M. Xu, L.-Y. Duan, J. Cai, L.-T. Chia, C. Xu, and Q. Tian. Hmm-based audio keyword generation. In *Advances in Multimedia Information Processing-PCM 2004: 5th Pacific Rim Conference on Multimedia, Tokyo, Japan, November 30-December 3, 2004. Proceedings, Part III 5*, pages 566–574. Springer, 2005.

C. Yang, Y. Xu, J. Shi, B. Dai, and B. Zhou. Temporal pyramid network for action recognition. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 588–597, 2020.

Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California, June 2016. Association for Computational Linguistics. doi:10.18653/v1/N16-1174. URL https://aclanthology.org/N16-1174.

Q. Ye, Q. Huang, W. Gao, and S. Jiang. Exciting event detection in broadcast soccer video with mid-level description and incremental learning. In *Proceedings of the 13th Annual ACM International Conference on Multimedia*, MULTIMEDIA '05, page 455–458, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595930442. doi:10.1145/1101149.1101250. URL https://doi.org/10.1145/1101149.1101250.

J. Yu, A. Lei, Z. Song, T. Wang, H. Cai, and N. Feng. Comprehensive dataset of broadcast soccer videos. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 418–423, 2018. doi:10.1109/MIPR.2018.00090.

J. Yu, A. Lei, and Y. Hu. Soccer video event detection based on deep learning. In I. Kompatsiaris, B. Huet, V. Mezaris, C. Gurrin, W.-H. Cheng, and S. Vrochidis, editors, *MultiMedia Modeling*, pages 377–389, Cham, 2019. Springer International Publishing. ISBN 978-3-030-05716-9.

E. Zhang and Y. Zhang. *Eleven Point Precision-recall Curve*, pages 981–982. Springer US, Boston, MA, 2009. ISBN 978-0-387-39940-9. doi:10.1007/978-0-387-39940-9_481. URL https://doi.org/10.1007/978-0-387-39940-9_481.

H. Zhang, M. Cissé, Y. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. *ArXiv*, abs/1710.09412, 2017.

W. Zhao, Y. Lu, H. Jiang, and W. Huang. Event detection in soccer videos using shot focus identification. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 341–345, 2015. doi:10.1109/ACPR.2015.7486522.

C. Zheng, S. Zhu, M. Mendieta, T. Yang, C. Chen, and Z. Ding. 3d human pose estimation with spatial and temporal transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11636–11645, 2021.

X. Zhou, L. Kang, Z. Cheng, B. He, and J. Xin. Feature combination meets attention: Baidu soccer embeddings and transformer based temporal detection. *ArXiv*, abs/2106.14447, 2021.

H. Zhu, J. Liang, C. Lin, J. Zhang, and J. Hu. A transformer-based system for action spotting in soccer videos. In *Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports*, MMSports '22, page 103–109, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450394888. doi:10.1145/3552437.3555693. URL https://doi.org/10.1145/3552437.3555693.

M. Zolfaghari, K. Singh, and T. Brox. Eco: Efficient convolutional network for online video understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 695–712, 2018.