SoccerNet-v2: A Dataset and Benchmarks for Holistic Understanding of Broadcast Soccer Videos

Adrien Deliège* University of Liège

Anthony Cioppa* University of Liège

Silvio Giancola* KAUST

Meisam J. Seikavandi* Aalborg University

Jacob V. Dueholm* Aalborg University

Kamal Nasrollahi Aalborg University, Milestone Systems

Bernard Ghanem KAUST

Thomas B. Moeslund Aalborg University

Marc Van Droogenbroeck University of Liège

Abstract

Understanding broadcast videos is a challenging task in computer vision, as it requires generic reasoning capabilities to appreciate the content offered by the video editing. In this work, we propose SoccerNet-v2, a novel large-scale corpus of manual annotations for the SoccerNet [24] video dataset, along with open challenges to encourage more research in soccer understanding and broadcast production. Specifically, we release around 300k annotations within SoccerNet's 500 untrimmed broadcast soccer videos. We extend current tasks in the realm of soccer to include action spotting, camera shot segmentation with boundary detection, and we define a novel replay grounding task. For each task, we provide and discuss benchmark results, reproducible with our open-source adapted implementations of the most relevant works in the field. SoccerNet-v2 is presented to the broader research community to help push computer vision closer to automatic solutions for more general video understanding and production purposes.

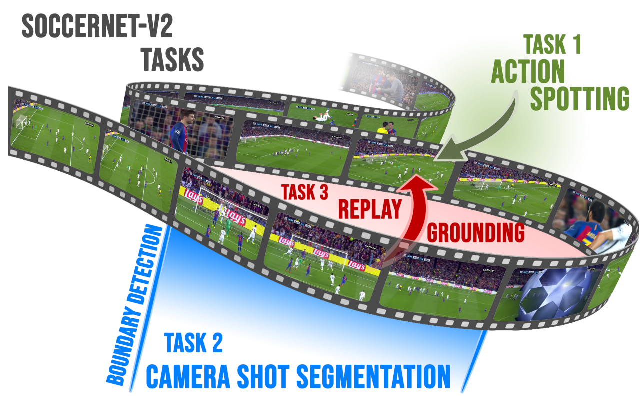

Figure 1. **SoccerNet-v2** constitutes the most inclusive dataset for soccer video understanding and production, with ~300k annotations, 3 computer vision tasks and multiple benchmark results.

1. Introduction

Sports is a profitable entertainment sector, capping $91 billion of annual market revenue over the last decade [15]. $15.6 billion alone came from the Big Five European Soccer Leagues (EPL, La Liga, Ligue 1, Bundesliga and Serie A) [16, 17, 18], with broadcasting and commercial activities being the main source of revenue for clubs [19]. TV broadcasters seek to attract the attention and indulge the curiosity of an audience, as they understand the game and edit the broadcasts accordingly. In particular, they select the best camera shots focusing on actions or players, allowing for semantic game analysis, talent scouting and advertisement placement. With almost 10,000 games a year for the Big Five alone, and an estimated audience of 500M+ people at each World Cup [60], automating the video editing process would have a broad impact on the other millions of games played in lower leagues across the world. Yet, it requires an understanding of the game and the broadcast production.

Recent computer vision works on soccer broadcasts focused on low-level video understanding [50], *e.g.* localizing a field and its lines [13, 20, 32], detecting players [12, 82], their motion [21, 47], their pose [6, 87], their team [34], the ball [67, 72], or pass feasibility [3]. Understanding frame-wise information is useful to enhance the visual experience of sports viewers [59] and to gather player statistics [73], but it falls short of higher-level game understanding needed for automatic editing purposes (*e.g.* camera shot selection, replay selection, and advertisement placement).

In this work, we propose a large-scale collection of manual annotations for holistic soccer video understanding and

(*) Equal contributions. More at https://soccer-net.org/.

Contacts: adrien.deliege@uliege.be, anthony.cioppa@uliege.be, silvio.giancola@kaust.edu.sa, meisamjam@gmail.com, jvdu@create.aau.dk.

several benchmarks addressing automatic broadcast production tasks. In particular, we extend the previous SoccerNet [24] dataset with further tasks and annotations, and propose open challenges with public leaderboards. Specifically, we propose three tasks represented in Figure 1: **(i)** *Action Spotting*, an extension from 3 to 17 action classes of SoccerNet's main task, **(ii)** *Camera Shot Understanding*, a temporal segmentation task for camera shots and a camera shot boundary detection task, and **(iii)** *Replay Grounding*, a task of retrieving the replayed actions in the game. These tasks tackle three major aspects of broadcast soccer videos: action spotting addresses the understanding of the content of the game, camera shot segmentation and boundary detection deal with the video editing process, and replay grounding bridges those tasks by emphasizing salient actions, allowing for prominent moments retrieval.

Contributions. We summarize our contributions as follows. **(i) Dataset.** We publicly release SoccerNet-v2, the largest corpus of manual annotations for broadcast soccer video understanding and production, comprising ~300k annotations temporally anchored within SoccerNet's 764 hours of video. **(ii) Tasks.** We define the novel task of replay grounding and further expand the tasks of action spotting, camera shot segmentation and boundary detection, for a holistic understanding of content, editing, and production of broadcast soccer videos. **(iii) Benchmarks.** We release reproducible benchmark results along with our code and public leaderboards to drive further research in the field.

2. Related Work

Video understanding datasets. Many video datasets propose challenging tasks around human action understanding [25, 68], with applications in movies [45, 48, 70], sports [40, 53, 61], cooking [14, 44, 62], and large-scale generic video classification [2, 41, 71]. While early efforts focused on trimmed video classification, more recent datasets provide fine-grained annotations of longer videos at a temporal [30, 38, 70, 84, 88] or spatio-temporal level [27, 49, 61, 80]. THUMOS14 [38] is the first benchmark for temporal activity localization, introducing 413 untrimmed videos, totalling 24 hours and 6k temporally anchored activities split into 20 classes, then extended to 65 classes in MultiTHUMOS [84]. ActivityNet [30] gathers the first large-scale dataset for activity understanding, with 849 hours of untrimmed videos, temporally annotated with 30k anchored activities split into 200 classes. A yearly ActivityNet competition highlights a variety of tasks with hundreds of submissions [22, 23]. Some datasets consider videos at an atomic level, with fine-grained temporal annotations from short snippets of longer videos [27, 51, 88]. Multi-Moments in Time [52] provides 2M action labels for 1M short clips of 3s, classified into 313 classes. Something-Something [26] collects 100k videos annotated with 147 classes of daily human-object interactions. Breakfast [44] and MPII-Cooking 2 [63] provide annotations for individual steps of cooking activities. EPIC-KITCHENS [14] scales up those approaches with 55 hours of cooking footage, annotated with around 40k action clips of 147 classes.

Soccer-related datasets. SoccerNet [24] is the first large-scale soccer video dataset, with 500 games from major European leagues and 6k annotations. It provides complete games with a distribution faithful to official TV broadcasts, but it only focuses on 3 action classes, making the task too simplistic and of moderate interest. SoccerDB [37] adds 7 classes and player bounding boxes for half of SoccerNet's videos and 76 extra games. However, it misses a complete set of possible actions and editing annotations to allow for a full understanding of the production of TV broadcasts. Yu *et al*. [85] released a dataset with 222 halves of soccer matches with annotations of actions, shot transitions, and player bounding boxes. They have few annotations and do not carry out any experiment nor propose any task. Pappalardo *et al*. [58] released a large-scale dataset of soccer events, localized in time and space. However, they focus on player and team statistics rather than video understanding, as they do not release any video. We address the limitations of these datasets by annotating all the interesting actions of the 500 SoccerNet games. Also, we provide valuable annotations for video editing, and we connect camera shots with actions to allow for salient moments retrieval.

Action spotting. Giancola *et al*. [24] define the task of action spotting in SoccerNet as finding the anchors of soccer events in a video and provide baselines based on temporal pooling. Rongved *et al*. [64] focus on applying a 3D ResNet directly to the video frames in a 5-second sliding window fashion. Vanderplaetse *et al*. [77] combine visual and audio features in a multimodal approach. Cioppa *et al*. [11] introduce a context-aware loss function to model the temporal context surrounding the actions. Similarly, Vats *et al*. [78] use a multi-tower CNN that accounts for the uncertainty of the action locations. Tomei *et al*. [74] fine-tune a feature extractor and use a masking strategy to focus on the frames after the actions. We build upon those works to provide benchmark results on our extended action spotting task.

Camera shot segmentation and boundary detection. Camera shot boundaries are typically detected by differences between frames, using pixels [5], histograms [56], motion [86] or deep features [1]. In soccer, Hu *et al*. [33] combine motion vectors and a filtration scheme to improve color-based methods. Lefèvre *et al*. [46] consider adaptive thresholds and features from a hue-saturation color space. Jackman [35] uses popular 2D and 3D CNNs but detects many false positives, as it appears difficult to efficiently process the temporal domain. Yet, these works are fine-tuned for only a few games. Regarding camera classification, Tong *et al.* [75] first detect logos to select non-replay camera shots, further classified as long, medium, close-up or out-of-field views based on color and texture features. Conversely, Wang *et al.* [79] classify camera shots for the task of replay detection. Sarkar *et al.* [66] classify each frame in the classes of [75] based on field features and player dimensions. Kolekar *et al.* [43] use audio features to detect exciting moments, further classified in camera shot classes for highlight generation. In this paper, we offer a unified and enlarged corpus of annotations that allows for a thorough understanding of the video editing process.

Replay grounding. In soccer, multiple works focus on detecting replays [66, 79, 81, 83, 89], using either logo transitions or slow-motion detection, but grounding the replays with their action in the broadcast has been mostly overlooked. Babaguchi *et al.* [4] tackle replay linking in American football but use a heuristic approach that can hardly generalize to other sports. Ouyng *et al.* [57] introduce a video abstraction task to find similarities between multiple cameras in various sports videos, yet their method requires camera parameters and is tested on a restricted dataset. Replay grounding can be likened to action similarity retrieval, as in [28, 39] for action recognition. Jain *et al.* [36] use a Siamese structure to compare the features of two actions, and Roy *et al.* [65] also quantify their similarity. We propose a task of replay grounding to connect replay shots with salient moments of broadcast videos, which could find further uses in action retrieval and highlight production.

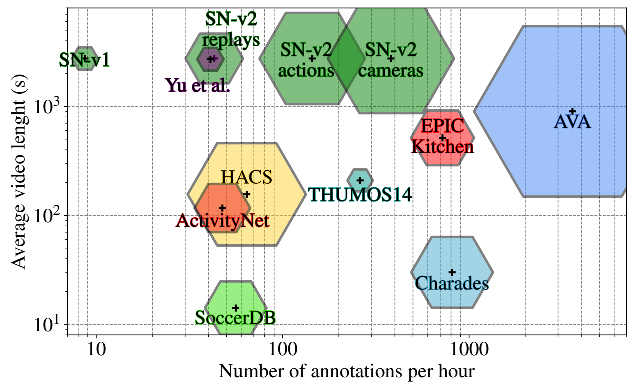

Figure 2. **Datasets comparison.** The areas of the tiles represent the number of annotations per dataset. SoccerNet-v2 (SN-v2) extends the initial SoccerNet [24] (SN-v1) with more annotations and tasks, and it focuses on untrimmed broadcast soccer videos.

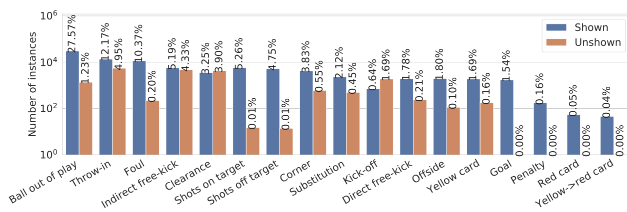

Figure 3. **SoccerNet-v2 actions.** Log-scale distribution of our **shown** and **unshown** actions among the 17 classes, and **proportion** that each class represents. The dataset is unbalanced, with some of the most important actions in the less abundant classes.

3. SoccerNet-v2 Dataset

Overview. Table 1 compares SoccerNet-v2 with the relevant video understanding datasets proposing localization tasks. SoccerNet-v2 stands out as one of the largest overall, and the largest for soccer videos by far. In particular, we manually annotated ~300k timestamps, temporally anchored in the 764 hours of the 500 games of SoccerNet [24]. We center the vocabulary of our classes on the soccer game and soccer broadcast domains, hence it is well-defined and consistent across games. Such regularity makes SoccerNet-v2 the largest dataset in term of events instances per class, thus enabling deep supervised learning at scale. As shown in Figure 2, SoccerNet-v2 provides the most dense annotations w.r.t. its soccer counterparts, and flirts with the largest fine-grained generic datasets in density and size.

We hired 33 annotators for the annotation process, all frequent observers of soccer, for a total of ~1600 hours of annotations. The quality of the annotations was validated by observing a large consensus between our annotators on identical games at the start and at the end of their annotation process. More details are provided in supplementary material. The annotations are divided in 3 categories: actions, camera shots, and replays, discussed hereafter.

Actions. We identify 17 types of actions from the most important in soccer, listed in Figure 3. Following [24], we annotate each action of the 500 games of SoccerNet with a single timestamp, defined by well-established soccer rules. For instance, for a corner, we annotate the last frame of the shot, *i.e.* showing the last contact between the player's foot and the ball. We provide the annotation guidelines in supplementary material. In total, we annotate 110,458 actions, on average 221 actions per game, or 1 action every 25 seconds. SoccerNet-v2 is a significant extension of the actions of SoccerNet [24], with 16x more timestamps and 14 extra classes. We represent the distribution of the actions in Figure 3. The natural imbalance of the data corresponds to the distribution of real-life broadcasts, making SoccerNet-v2 valuable for generalization and industrial deployment.

Additionally, we enrich each timestamp with a novel binary visibility tag that states whether the associated action is *shown* in the broadcast video or *unshown*, in which case the action must be inferred by the viewer. For example, this happens when the producer shows a replay of a shot off target that lasts past the clearance shot of the goalkeeper: the viewer knows that the clearance has been made despite it was not shown on the TV broadcast. Spotting unshown actions is challenging because it requires a fine understanding

Table 1. **Datasets.** Comparative overview of relevant datasets for action localization or spotting in videos. SoccerNet-v2 provides the second largest number of annotations and the largest in soccer. *computed with the 116k annotations of the 200 fully annotated games.

Dataset	Context	Duration (hrs)	#Actions	Classes	Density (act./hr)	Avg. events per class	Avg. video length (sec)
THUMOS14 [38]	General	24	6,363	20	260.4	318	209.2
ActivityNet [30]	General	648	30,791	200	47.5	154	116.7
Charades [70]	General	82	66,500	157	811	424	30
AVA [27]	Movies	107.5	385,446	80	3,585	4,818	900
HACS [88]	Human	~2,166	139,000	200	64.2	695	156
EPIC-Kitchen [14]	Cooking	55	39,596	149	720	266	514.3
SoccerNet [24]	Soccer	764	6,637	3	8.7	2,212	2750.4
SoccerDB [37]	Soccer	669	37,715	11	56	3,428	14.1
Yu *et al.* [85]	Soccer	167	6,850	11	41	623	2708.1
SoccerNet-v2 (actions)	Soccer	764	110,458	17	144	6,498	2750.4
SoccerNet-v2 (cameras)	Soccer	306	158,493	13	381*	8,976*	2750.4
SoccerNet-v2 (replays)	Soccer	764	32,932	–	43	–	2750.4

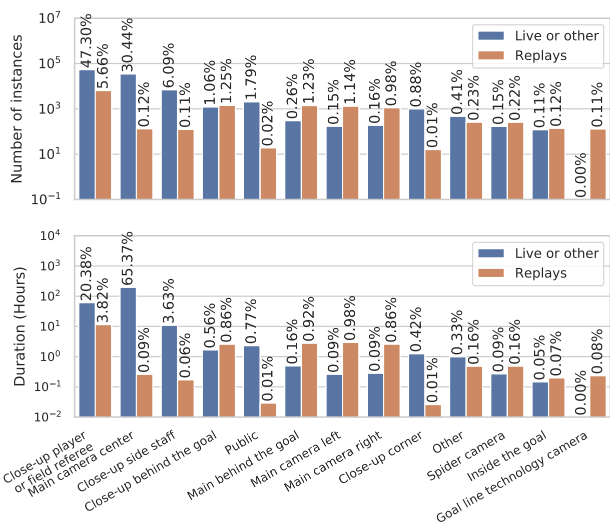

Figure 4. **Camera shots.** Log-scale distribution of our camera shot timestamps among the classes in terms of instances (top) and video duration (bottom), separated in **replays** and **live or other** sequences, and **percentage** of timestamps that each bar represents.

of the game, beyond frame-based analysis, as it forces to consider the temporal context around the actions. We annotate the timestamps of unshown actions with the best possible temporal interpolation. They represent 18% of the actions (see Figure 3), hence form a large set of actions whose spotting requires a sharp understanding of soccer. Finally, to remain consistent with SoccerNet [24], we annotate the team that performs each action as either *home* or *away*, but leave further analysis on that regard for future work.

Cameras. We annotate a total of 158,493 camera change timestamps, 116,687 of which are comprehensive for a subset of 200 games, the others delimiting replay shots in the remaining games (see hereafter). For the fully annotated games, this represents an average of 583 camera transitions per game, or 1 transition every 9 seconds. Those timestamps contain the type of camera shot that has been shown, among the most common 13 possibilities listed in Figure 4. We display their distribution in terms of number of occurrences and total duration. The class imbalance underpins a difficulty of this dataset, yet it represents a distribution consistent with broadcasts used in practical applications.

Besides, different types of transitions occur from one camera shot to the next, which we append to each timestamp. These can be abrupt changes between two cameras (71.4%), fading transitions between the frames (14.2%), or logo transitions (14.2%). Logos constitute an unusual type of transition compared with abrupt or fading, which are common in videos in the wild or in movies, yet they are widely used in sports broadcasts. They pose an interesting camera shot detection challenge, as each logo is different and algorithms must adapt to a wide variety thereof. For logo and fading camera changes, we locate the timestamps as precisely as possible at the middle of the transition, while we annotate the last frame before an abrupt change.

Eventually, we indicate whether the camera shot happens live (86.7%) with respect to the game, or shows a replay of an action (10.9%), or another type of replay (2.4%). The distribution in Figure 4 provides per-class proportions of replay camera shots and groups other replays and live shots.

Replays. For the 500 games of SoccerNet [24], we bound each video shot showing a replay of an action with two timestamps, annotated in the same way as for the camera shot changes. For each replay shot, we refer the timestamp of the action replayed. When several replays of the same action are shown consecutively with different views, we annotate all the replay shots separately. This gives one replay shot per type of view, all of which are linked to the same action. In total, 32,932 replay shots are associated with their corresponding action, which represents an average of 66 replay shots per game, for an average replay shot duration of

6.8 seconds. Retrieving a replayed action is challenging because typically, 1 to 3 replays of the action are shown from different viewpoints hardly ever found in the original live broadcast video. This encourages a more general video understanding rather than an exact frame comparison.

4. Broadcast Video Understanding Tasks

We propose a comprehensive set of tasks to move computer vision towards a better understanding of broadcast soccer videos and alleviate the editing burden of video producers. More importantly, these tasks have broader implications as they can easily be transposed to other domains. This makes SoccerNet-v2 an ideal playground for developing novel ideas and implementing innovative solutions in the general field of video understanding.

In this work, we define three main tasks on SoccerNet-v2: action spotting, camera shot segmentation with boundary detection, and replay grounding, which are illustrated in Figure 5. They are further motivated and detailed hereafter.

Action spotting. In order to understand the salient actions of a broadcast soccer game, SoccerNet [24] introduces the task of action spotting, which consists in finding all the actions occurring in the videos. Beyond soccer understanding, this task addresses the more general problem of retrieving moments with a specific semantic meaning in long untrimmed videos. As such, we foresee moment spotting applications in *e.g.* video surveillance or video indexing.

In this task, the actions are anchored with a single timestamp, contrary to the task of activity localization [30], where activities are delimited with start and stop timestamps. We assess the action spotting performance of an algorithm with the Average-mAP metric, defined as follows. A predicted action spot is positive if it falls within a given tolerance δ of a ground-truth timestamp from the same class. The Average Precision (AP) based on PR curves is computed then averaged over the classes (mAP), after what the Average-mAP is the AUC of the mAP computed at different tolerances δ ranging from 5 to 60 seconds.

Camera shot segmentation and boundary detection. Selecting the proper camera at the right moment is the crucial task of the broadcast producer to trigger the strongest emotions on the viewer during a live game. Hence, identifying camera shots not only provides a better understanding of the editing process but is also a major step towards automating the broadcast production. This task naturally generalizes to any sports broadcasts but could also prove interesting for *e.g.* cultural events or movies summarization.

Camera shot temporal segmentation consists in classifying each video frame among our 13 camera types and is evaluated with the mIoU metric. Concurrently, we define a task of camera shot boundary detection, where the objective is to find the timestamps of the transitions between the camera shots. For the evaluation, we use the spotting mAP metric with a single tolerance δ of 1 second as transitions are precisely localized and happen within short durations.

Replay grounding. Our novel replay grounding task consists in retrieving the timestamp of the action shown in a given replay shot within the whole game. Grounding a replay with its action confers it an estimation of importance, which is otherwise difficult to assess. Derived applications may be further built on top of this task, *e.g.* automatic highlight production, as the most replayed actions are usually the most relevant. Linking broadcast editing to meaningful content within the video not only bridges our previous tasks, but it can also be applied to any domain focusing on salient moments retrieval. We use the Average-AP to assess performances on this task, computed as described for the spotting task but without averaging over the classes. We choose this metric as replay grounding can be seen as class-independent action spotting conditioned by the replay sequence.

5. Benchmark Results

General comments. SoccerNet [24] provides high and low quality videos of the 500 games. For easier experimentation, it also provides features from ResNet [29], I3D [8] and C3D [76] computed at 2 fps, further reduced with PCA to 512 dimensions. Following [11, 24], in our experiments, we use the ResNet 512-dimensional frame features acting as compressed video representations as they yielded better results in early experiments. We adapt the most relevant existing methods to provide benchmark results on the SoccerNet [24] test set. We release our codes to reproduce them, and we will host leaderboards on dedicated servers.

5.1. Action Spotting

Methods. We adapt or re-implement efficiently all the methods that released public code on SoccerNet [24].

1. *MaxPool* and *NetVLAD* [24]. Those models pool temporally the ResNet features before passing them through a classification layer. Non-overlapping segments of 20 seconds are classified as to whether they contain any action class. In testing, a sliding window of 20 seconds with a stride of 1 frame is used to infer an actionness score in time, reduced to an action spot using NMS. We consider the basic yet lightweight max pooling and a learnable NetVLAD pooling with 64 clusters. We re-implement the method based on the original code for a better scaling to 17 classes.

2. *AudioVid* [77]. The network uses NetVLAD to pool temporally 20-second chunks of ResNet features, as well as VGGish [31] synchronized audio features, subsampled at 2 fps. The two sets of features are temporally pooled, concatenated and fed to a classification module, as in [24]. Similarly, the spotting prediction is at the center of the video chunk. We scaled the classification module to 17 classes.

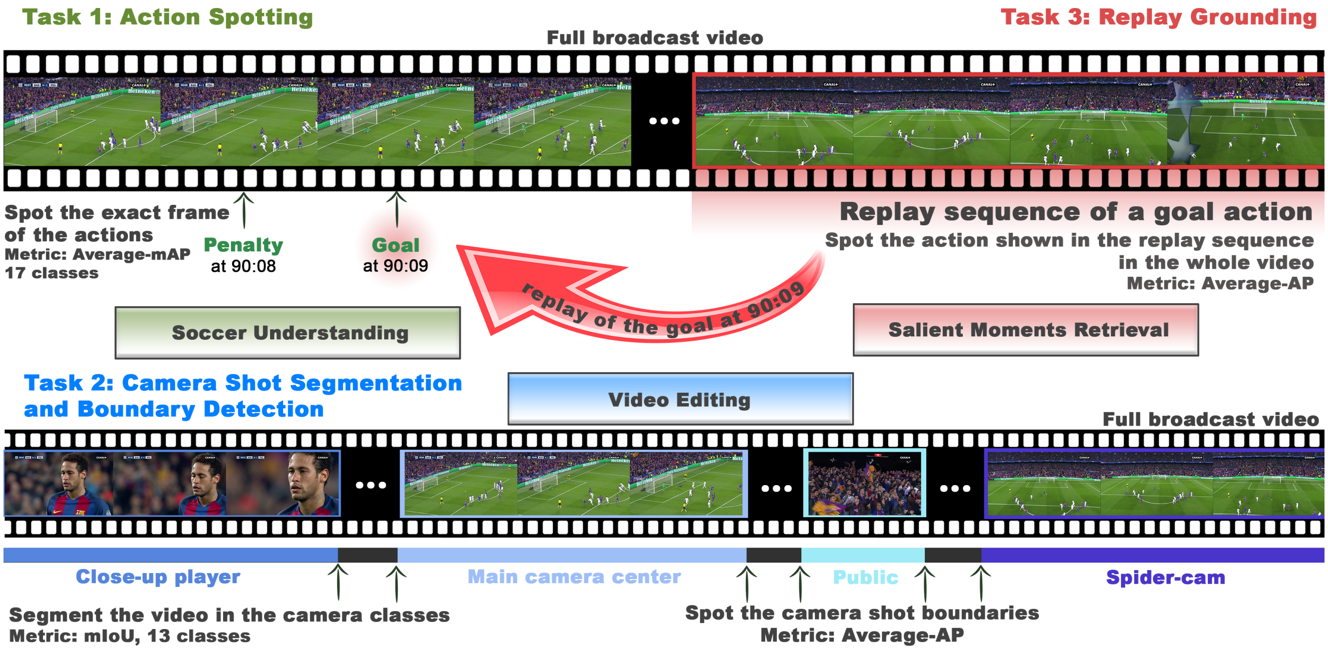

Figure 5. **Tasks overview.** We define a 17-class action spotting task, a 13-class camera shot segmentation and boundary detection tasks, and a novel replay grounding task, with their associated performance metrics. They respectively focus on understanding the content of broadcast soccer games, addressing broadcast video editing tasks, and retrieving salient moments of the game.

3. *CALF* [11]. This network handles 2-minute chunks of ResNet features and is composed of a spatio-temporal features extractor, kept as is, a temporal segmentation module, which we adapt for 17 classes, and an action spotting module, adapted to output at most 15 predictions per chunk, classified in 17 classes. The segmentation module is trained with a context-aware loss having four context slicing hyperparameters per class. Following [11], we determine optimal values for them with a Bayesian optimization [54]. We re-implement the method and optimize the training strategy based on the existing code to achieve a decent training time.

Results. We provide the leaderboard of our benchmark results for action spotting in Table 2. We further compute the performances on shown/unshown actions as the Average-mAP for predicted spots whose closest ground truth timestamp is a shown/unshown action. We show qualitative results obtained with CALF in Figure 6.

The pooling approaches MaxPool and NetVLAD are not on par with the other methods on SoccerNet-v2. We believe the hard pruning with MaxPool has a restricted learning capacity, limited to a single fully connected layer. Similarly, NetVLAD may lag behind because of a non-optimal choice in the design of the spotting module, in particular the Non-Maximum Suppression that discards the results with confidence score below 0.5. AudioVid prevails on the shown instances and on 5/17 actions classes. Injecting audio features appears to help with visible actions, as the sound is usually synchronized with the image. Also, it performs best on ac-

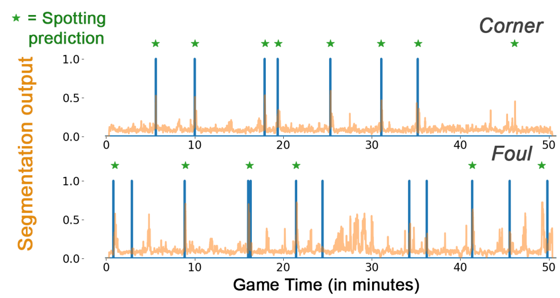

Figure 6. **Action spotting result** obtained from CALF adapted: temporal segmentation, ground truth, and spotting predictions. The network performs well on corners with only one false positive, and moderately on fouls with a few false negatives.

tions preceded or followed by the whistle of the referee, underlining the importance of audio features. Yet, the audio features appear less useful on unshown instances. CALF performs best globally, on the unshown instances and on most action classes. The context-aware loss focuses on the temporal context to spot the actions, which is useful for this task. This emphasizes the benefits of the temporal context surrounding the actions, that contains valuable information.

5.2. Camera Segmentation and Boundary Detection

Methods. 1. *Basic model*. For our first baseline for the segmentation part, we train a basic model composed of 3 layers of 1D CNN with a kernel of 21 frames, hence aggregated in time, on top of ResNet features, and a MSE loss.

Table 2. **Leaderboard for action spotting** (Average-mAP %). Methods with codes publicly available were tested on SoccerNet-v2.

	SoccerNet-v1	SoccerNet-v2	shown	unshown	Ball out	Throw-in	Foul	Ind. free-kick	Clearance	Shots on tar.	Shots off tar.	Corner	Substitution	Kick-off	Yellow card	Offside	Dir. free-kick	Goal	Penalty	Yel.→Red	Red card
Counts (test set)	1369	22551	18641	3910	6460	3809	2414	2283	1631	1175	1058	999	579	514	431	416	382	337	41	14	8
MaxPool [24]	-	18.6	21.5	15.0	38.7	34.7	26.8	17.9	14.9	14.0	13.1	26.5	40.0	30.3	11.8	2.6	13.5	24.2	6.2	0.0	**0.9**
NetVLAD [24]	49.7	31.4	34.3	23.3	47.4	42.4	32.0	16.7	32.7	21.3	19.7	55.1	51.7	45.7	33.2	14.6	33.6	54.9	32.3	0.0	0.0
AudioVid [77]	56.0	39.9	**43.0**	23.3	54.3	50.0	**55.5**	22.7	46.7	26.5	21.4	66.0	**54.0**	**52.9**	35.2	24.3	**46.7**	69.7	**52.1**	0.0	0.0
CALF [11]	62.5	**40.7**	42.1	**29.0**	**63.9**	**56.4**	53.0	**41.5**	**51.6**	**26.6**	**27.3**	**71.8**	47.3	37.2	**41.7**	**25.7**	43.5	**72.2**	30.6	**0.7**	0.7

Other SoccerNet-v1 results but with no public code available: Rongved *et al*. [64]: 32.0 ; Vats *et al*. [78]: 60.1 ; Tomei *et al*. [74]: **75.1**.

2. *CALF (seg.)* [11]. We adapt CALF as it provides a segmentation module on top of a spatio-temporal features extractor. We replace its loss with the cross-entropy for easier experimentation and we focus on the segmentation by removing the spotting module. The number of parameters is reduced by a factor of 5 compared with the original model.

3. *Content* [9]. For the boundary detection task, we test the popular scene detection library PySceneDetect. We use the Content option, that triggers a camera change when the difference between two consecutive frames exceeds a particular threshold value. This method is tested directly on the broadcast videos provided in SoccerNet [24].

4. *Histogram, Intensity* [69]. We test two scene detection methods of the Scikit-Video library. The Histogram method reports a camera change when the intensity histogram difference between subsequent frames exceeds a given threshold [55]. The Intensity method reports a camera change when variations in color and intensity between frames exceed a given threshold. Those methods are tested directly on the broadcast videos provided in SoccerNet [24].

5. *CALF (det.)* [11]. Since we can see the camera shot boundary detection as a spotting task, we recondition the best spotting method CALF by removing the segmentation module to focus on detection. Following a grid search optimization, we use 24-second input chunks of ResNet features and allow at most 9 detections per chunk.

Results. We provide a leaderboard of our benchmark results for these task in Table 3. We further compute the performances per transition type as the mAP for predicted spots grouped by the transition of their closest ground truth.

Regarding the segmentation, even with 5x more parameters, the basic model trails behind CALF. Hence, simplistic architectures may not suffice for this task, and more sophisticated designs can rapidly boost performances. For the boundary detection, Histogram prevails, yet it ranks only third on fading transitions where the deep learning-based CALF is the best. The learning capacity of CALF may explain its performance consistency across transition types. Intensity, Content, and Histogram are intrinsically tailored for abrupt transitions. Intensity and Content are particularly bad on logos, while Histogram still performs well.

Table 3. **Leaderboard for Camera Shot Segmentation** (mIoU %) and **Boundary Detection** (mAP %).

Method	Camera Seg.	Bound. Det.	Transition: Abrupt	Transition: Fading	Transition: Logo
Basic model	35.8	-	-	-	-
CALF [11] (seg.)	47.3	-	-	-	-
CALF [11] (det.)	-	59.6	59.0	**58.0**	61.8
Intensity [69]	-	64.0	74.3	57.2	28.5
Content [9]	-	62.2	68.2	49.7	35.5
Histogram [69]	-	**78.5**	**83.2**	54.1	**82.2**

5.3. Replay Grounding

Methods. Given the novelty of this task, there is no off-the-shelf method available. We choose to adapt our optimized implementations of NetVLAD [24] and CALF [11] within a Siamese neural networks approach [7, 10, 42].

As input for the networks, we provide the ResNet features representations of a fixed-size video chunk and a replay shot. We either repeat or shorten the latter at both sides so that it has the same duration as the former. Ideally, for a chunk containing the action replayed (positive sample), the networks should output a high confidence score along with a localization prediction for spotting the action within the chunk. Otherwise (negative sample), they should only provide a low confidence score, and spotting predictions will be ignored. Negative samples are sampled either among chunks containing an action of the same class as the action replayed (hard negative), or among chunks randomly located within the whole video (random negative). The hard negatives ensure that the network learns to spot the correct actions without simply identifying their class, while the random negatives bring some diversity in the negative mining.

We test two sampling strategies. At each epoch, for each replay shot, we select: *(S1)* only 1 sample: a positive with probability 0.5, or a hard or random negative each with probability 0.25; *(S2)* 5 samples: 1 positive, 2 hard and 2

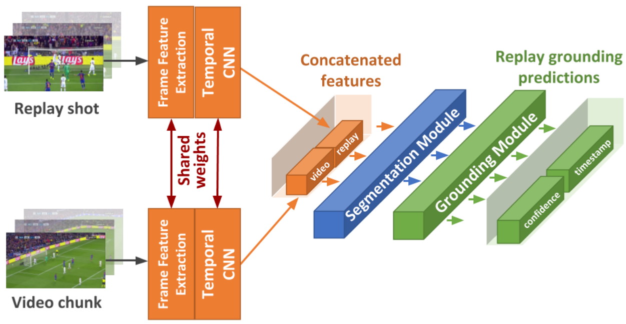

Figure 7. **Replay grounding pipeline** of our adaptation of CALF. In a Siamese network approach, the replay shot and the video chunk share a **frame features extractor**. Their features are concatenated and fed to the **segmentation module**. The **grounding module** outputs a confidence score on the presence or absence of the action in the replay shot, and an action spotting prediction.

random negatives. For both S1 and S2, the positive is a chunk randomly shifted around the action timestamp. The adaptations specific to each method are the following.

1. *NetVLAD* [24]. We use NetVLAD to pool temporally the replay shot and the video chunk separately, but with shared weights. We compare the features obtained for the shot with those of the chunk through a cosine similarity loss, zeroed out when smaller than 0.4 to help the networks focus on improving their worst scores. In parallel, we feed the video features to a 2-layer MLP acting as spotting module to regress the spotting prediction within the chunk.

2. *CALF* [11]. We feed the replay shot and a video chunk to the shared frame feature extractor. Then, we concatenate the feature vectors along the temporal dimension, and give the resulting tensor to the remaining modules of the network. We set the number of classes to 1 in the segmentation module to provide per-frame insight. The spotting module outputs the confidence score on the presence of the replayed action in the chunk. We further set its number of detections to 1 as one action is replayed and might be spotted in the chunk. This architecture is represented in Figure 7.

For these methods, at test time, we slice the video associated with the replay in chunks. We obtain at most one grounding prediction per chunk, all of which are kept when computing the Average-AP metric.

Results. The leaderboard providing our benchmark results for action spotting is given in Table 4 for video chunks of different sizes. NetVLAD with S1 performs poorly, so no result is reported. Our adaptation of CALF achieves the best performance, with a chunk size of 60 seconds and with S2 as sampling strategy. Its demonstrated ability to aggregate the temporal context may explain this success. All the methods yield their best results with chunk sizes around 60 seconds, which presumably provides the most appropriate compromise between not enough and too much temporal

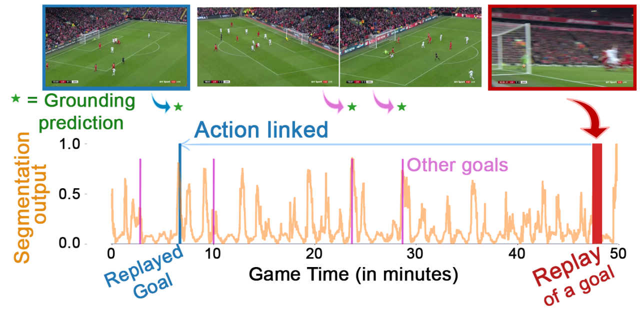

Figure 8. **Replay grounding result** of CALF adapted. We display the **replay shot** of a goal, its **ground truth** spot, the **other goals**, the temporal **segmentation output**, and the **grounding predictions**. The replayed goal is correctly spotted, two goals are rightly avoided, but two false positive predictions are also spotted, incidentally when other goals occurred. An insightful visualization can be appreciated in our **video in supplementary material**.

Table 4. **Leaderboard for replay grounding** (Average-AP %), along with sampling strategy during training.

Method	Video chunk size (seconds)						
	30	40	50	60	120	180	240
NetV. [24]+S2	23.9	22.9	24.3	22.4	7.5	–	–
CALF[11]+S1	16.7	19.6	28.0	32.3	32.0	26.9	22.0
CALF[11]+S2	8.2	14.7	28.9	**41.8**	40.3	27.2	14.4

context for an efficient replay grounding. An example of result from CALF is given in Figure 8, showing that it can correctly learn to link a replay with its action without necessarily spotting all the actions of the same class. This underlines both the feasibility and the difficulty of our novel task. For a more relevant visualization experience, we invite the reader to consult our **video in supplementary material**.

6. Conclusion

We release SoccerNet-v2, the largest soccer-related set of annotations, anchored on top of the original SoccerNet [24]'s 500 untrimmed broadcast games. With our ~300k annotations, we further extend the tasks of action spotting, camera shot segmentation and boundary detection, and we define the novel task of replay grounding. We propose and discuss several benchmark results for all of them. In addition, we provide codes to reproduce our experiments, and we will host public leaderboards to drive research in this field. With SoccerNet-v2, we aim at pushing computer vision closer to automatic solutions for holistic broadcast soccer video understanding, and believe that it is the ideal dataset to explore new tasks and methods for more generic video understanding and production tasks.

Acknowledgments. This work is supported by the DeepSport project of the Walloon Region, the FRIA, KAUST Office of Sponsored Research (award OSR-CRG2017-3405), and the Milestone Research Program at Aalborg University.

References

[1] Sadiq H. Abdulhussain, Abd R. Ramli, M. I. Saripan, Basheera M. Mahmmod, Syed Al-Haddad, and Wissam A. Jassim. Methods and challenges in shot boundary detection: A review. *Entropy*, 20(4):214, March 2018. 2

[2] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. YouTube-8M: A large-scale video classification benchmark. *CoRR*, September 2016. 2

[3] Adrià Arbués Sangüesa, Adriàn Martín, Javier Fernández, Coloma Ballester, and Gloria Haro. Using player's body-orientation to model pass feasibility in soccer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3875–3884, June 2020. 1

[4] Noboru Babaguchi, Yoshihiko Kawai, Yukinobu Yasugi, and Tadahiro Kitahashi. Linking live and replay scenes in broadcasted sports video. In *ACM workshops on Multimedia*, pages 205–208, November 2000. 3

[5] John S. Boreczky and Lawrence A. Rowe. Comparison of video shot boundary detection techniques. In *Storage and Retrieval for Still Image and Video Databases IV*, pages 170–179, March 1996. 2

[6] Lewis Bridgeman, Marco Volino, Jean-Yves Guillemaut, and Adrian Hilton. Multi-Person 3D Pose Estimation and Tracking in Sports. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2487–2496, June 2019. 1

[7] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a "siamese" time delay neural network. In *International Conference on Neural Information Processing Systems (NIPS)*, pages 737–744, November 1993. 7

[8] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017. 5

[9] Brandon Castellano. Pyscenedetect: Video scene cut detection and analysis tool, 2014. `https://github.com/Breakthrough/PySceneDetect`. 7

[10] Davide Chicco. Siamese neural networks: An overview. *Artificial Neural Networks*, 2190:73–94, 2021. 7

[11] Anthony Cioppa, Adrien Deliège, Silvio Giancola, Bernard Ghanem, Marc Van Droogenbroeck, Rikke Gade, and Thomas B. Moeslund. A context-aware loss function for action spotting in soccer videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13126–13136, 2020. 2, 5, 6, 7, 8

[12] Anthony Cioppa, Adrien Deliège, Maxime Istasse, Christophe De Vleeschouwer, and Marc Van Droogenbroeck. ARTHuS: Adaptive Real-Time Human Segmentation in Sports Through Online Distillation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2505–2514, June 2019. 1

[13] Anthony Cioppa, Adrien Deliège, and Marc Van Droogenbroeck. A bottom-up approach based on semantics for the interpretation of the main camera stream in soccer games. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1846–1855, June 2018. 1

[14] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. The EPIC-KITCHENS dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, May 2020. 2, 4

[15] Deloitte. Global sports market - total revenue from 2005 to 2017 (in billion u.s. dollars). In *Statista - The Statistics Portal*, 2017. Retrieved October 30, 2017, from `https://www.statista.com/statistics/370560/worldwide-sports-market-revenue/`. 1

[16] Deloitte. Market size of the european football market from 2006/07 to 2015/16 (in billion euros). In *Statista - The Statistics Portal*, 2017. Retrieved October 30, 2017, from `https://www.statista.com/statistics/261223/european-soccer-market-total-revenue/`. 1

[17] Deloitte. Revenue of the biggest (big five*) european soccer leagues from 1996/97 to 2017/18 (in million euros). In *Statista - The Statistics Portal*, 2017. Retrieved October 30, 2017, from `https://www.statista.com/statistics/261218/big-five-european-soccer-leagues-revenue/`. 1

[18] Deloitte. Revenue of the top european soccer leagues (big five*) from 2006/07 to 2017/18 (in billion euros). In *Statista - The Statistics Portal*, 2017. Retrieved October 30, 2017, from `https://www.statista.com/statistics/261225/top-european-soccer-leagues-big-five-revenue/`. 1

[19] Deloitte. Top-20 european football clubs breakdown of revenues 2018/19 season (in million euros). In *Statista - The Statistics Portal*, 2020. Retrieved October 25, 2020, from `https://www.statista.com/statistics/271636/revenue-distribution-of-top-20-european-soccer-clubs/`. 1

[20] Dirk Farin, Susanne Krabbe, Peter de With, and Wolfgang Effelsberg. Robust camera calibration for sport videos using court models. In *Storage and Retrieval Methods and Applications for Multimedia*, pages 80–91, December 2003. 1

[21] Panna Felsen, Pulkit Agrawal, and Jitendra Malik. What will happen next? Forecasting player moves in sports videos. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3362–3371, October 2017. 1

[22] Bernard Ghanem, Juan Carlos Niebles, Cees Snoek, Fabian Caba Heilbron, Humam Alwassel, Victor Escorcia, Ranjay Krishna, Shyamal Buch, and Cuong Duc Dao. The ActivityNet large-scale activity recognition challenge 2018 summary. *CoRR*, August 2018. 2

[23] Bernard Ghanem, Juan Carlos Niebles, Cees Snoek, Fabian Caba Heilbron, Humam Alwassel, Ranjay Khrisna, Victor Escorcia, Kenji Hata, and Shyamal Buch. ActivityNet challenge 2017 summary. *CoRR*, October 2017. 2

[24] Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. SoccerNet: A Scalable Dataset for Action Spotting in Soccer Videos. In *IEEE Conference on Computer*

Vision and Pattern Recognition (CVPR) Workshops, pages 1711–1721, June 2018. 1, 2, 3, 4, 5, 7, 8

[25] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2247–2253, December 2007. 2

[26] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The "Something Something" video database for learning and evaluating visual common sense. In *IEEE International Conference on Computer Vision (ICCV)*, pages 5843–5851, October 2017. 2

[27] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. AVA: A video dataset of spatio-temporally localized atomic visual actions. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6047–6056, June 2018. 2, 4

[28] Seyed Mohammad Hashemi and Mohammad Rahmati. View-independent action recognition: A hybrid approach. *Multimedia Tools and Applications*, 75(12):6755–6775, June 2016. 3

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016. 5

[30] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970, June 2015. 2, 4, 5

[31] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron Weiss, and Kevin Wilson. CNN architectures for large-scale audio classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135, March 2017. 5

[32] Namdar Homayounfar, Sanja Fidler, and Raquel Urtasun. Sports field localization via deep structured models. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4012–4020, July 2017. 1

[33] Yichuan Hu, Bo Han, Guijin Wang, and Xianggang Lin. Enhanced shot change detection using motion features for soccer video analysis. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1555–1558, July 2007. 2

[34] Maxime Istasse, Julien Moreau, and Christophe De Vleeschouwer. Associative Embedding for Team Discrimination. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2477–2486, June 2019. 1

[35] Simeon Jackman. Football shot detection using convolutional neural networks. Master's thesis, Linköping University, June 2019. 2

[36] Hiteshi Jain, Gaurav Harit, and Avinash Sharma. Action quality assessment using siamese network-based deep metric learning. *CoRR*, February 2020. 3

[37] Yudong Jiang, Kaixu Cui, Leilei Chen, Canjin Wang, and Changliang Xu. Soccerdb: A large-scale database for comprehensive video understanding. In *International Workshop on Multimedia Content Analysis in Sports*, pages 1–8, October 2020. 2, 4

[38] Yu-Gang. Jiang, Jingen Liu, Amir Roshan Zamir, George Toderici, Ivan Laptev, Mubarak Shah, and Rahul Sukthankar. THUMOS Challenge: Action Recognition with a Large Number of Classes. http://crcv.ucf.edu/THUMOS14/, 2014. 2, 4

[39] Imran N Junejo, Emilie Dexter, Ivan Laptev, and Patrick Púrez. Cross-view action recognition from temporal self-similarities. In *European Conference on Computer Vision (ECCV)*, pages 293–306, October 2008. 3

[40] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale Video Classification with Convolutional Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 2

[41] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, May 2017. 2

[42] Gregory R. Koch. Siamese neural networks for one-shot image recognition. In *International Conference on Machine Learning (ICML) Deep Learning Workshop*, July 2015. 7

[43] Maheshkumar H. Kolekar and Somnath Sengupta. Bayesian network-based customized highlight generation for broadcast soccer videos. *IEEE Transactions on Broadcasting*, 61(2):195–209, June 2015. 3

[44] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 780–787, June 2014. 2

[45] Hilde Kuehne, Hueihan Jhuang, Estibaliz Garrote, Tomaso. Poggio, and Thomas Serre. HMDB: a large video database for human motion recognition. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2556–2563, November 2011. 2

[46] Sébastien Lefèvre and Nicole Vincent. Efficient and robust shot change detection. *Journal of Real-Time Image Processing*, 2:23–34, August 2007. 2

[47] Mehrtash Manafifard, Hamid Ebadi, and Hamid Abrishami Moghaddam. A survey on player tracking in soccer videos. *Computer Vision and Image Understanding*, 159:19–46, June 2017. 1

[48] Marcin Marszalek, Ivan Laptev, and Cordelia Schmid. Actions in context. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2929–2936, June 2009. 2

[49] Pascal Mettes, Jan C van Gemert, and Cees GM Snoek. Spot on: Action localization from pointly-supervised proposals.

In *European Conference on Computer Vision (ECCV)*, pages 437–453, October 2016. 2

[50] Thomas B. Moeslund, Graham Thomas, and Adrian Hilton. *Computer Vision in Sports*. Springer, 2014. 1

[51] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):502–508, February 2020. 2

[52] Mathew Monfort, Kandan Ramakrishnan, Alex Andonian, Barry A McNamara, Alex Lascelles, Bowen Pan, Quanfu Fan, Dan Gutfreund, Rogerio Feris, and Aude Oliva. Multi-moments in time: Learning and interpreting models for multi-action video understanding. *CoRR*, November 2019. 2

[53] Juan Carlos Niebles, Chih-Wei Chen, and Li Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *European Conference on Computer Vision (ECCV)*, pages 392–405, September 2010. 2

[54] Fernando Nogueira. Bayesian Optimization: Open source constrained global optimization tool for Python, 2014. `https://github.com/fmfn/BayesianOptimization`. 6

[55] Kiyotaka Otsuji and Yoshinobu Tonomura. Projection detecting filter for video cut detection. In *ACM International Conference on Multimedia*, pages 251—257, September 1993. 7

[56] Kiyotaka Otsuji and Yoshinobu Tonomura. Projection-detecting filter for video cut detection. *Multimedia Systems*, 1(5):205–210, March 1994. 2

[57] Jian-quan Ouyang, Jin-tao Li, and Yong-dong Zhang. Replay scene based sports video abstraction. In *International Conference on Fuzzy Systems and Knowledge Discovery*, pages 689–697, August 2005. 3

[58] Luca Pappalardo, Paolo Cintia, A. Rossi, Emanuele Massucco, P. Ferragina, D. Pedreschi, and F. Giannotti. A public data set of spatio-temporal match events in soccer competitions. *Scientific Data*, 6:236, October 2019. 2

[59] Konstantinos Rematas, Ira Kemelmacher-Shlizerman, Brian Curless, and Steve Seitz. Soccer on your tabletop. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4738–4747, June 2018. 1

[60] Felix Richter. Super bowl can't hold the candle to the biggest game in soccer. In *Statista - The Statistics Portal*, 2020. Retrieved October 25, 2020, from `https://www.statista.com/chart/16875/super-bowl-viewership-vs-world-cup-final/`. 1

[61] Mikel D Rodriguez, Javed Ahmed, and Mubarak Shah. Action MACH a spatio-temporal maximum average correlation height filter for action recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, June 2008. 2

[62] Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele. A database for fine grained activity detection of cooking activities. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1194–1201, June 2012. 2

[63] Marcus Rohrbach, Anna Rohrbach, Michaela Regneri, Sikandar Amin, Mykhaylo Andriluka, Manfred Pinkal, and Bernt Schiele. Recognizing fine-grained and composite activities using hand-centric features and script data. *International Journal of Computer Vision*, 119(3):346–373, September 2016. 2

[64] Olav A. Nergård Rongved, Steven A. Hicks, Vajira Thambawita, Håkon K. Stensland, Evi Zouganeli, Dag Johansen, Michael A. Riegler, and Pål Halvorsen. Real-time detection of events in soccer videos using 3D convolutional neural networks. In *IEEE International Symposium on Multimedia (ISM)*, December 2020. In press. 2, 7

[65] Debaditya Roy, C Krishna Mohan, and K Sri Rama Murty. Action recognition based on discriminative embedding of actions using siamese networks. In *IEEE International Conference on Image Processing (ICIP)*, pages 3473–3477, October 2018. 3

[66] Saikat Sarkar, Sazid Ali, and Amlan Chakrabarti. Shot classification and replay detection in broadcast soccer video. In *Advanced Computing and Systems for Security*, pages 57–66, February 2020. 3

[67] Saikat Sarkar, Amlan Chakrabarti, and Dipti Prasad Mukherjee. Generation of Ball Possession Statistics in Soccer Using Minimum-Cost Flow Network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2515–2523, June 2019. 1

[68] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local svm approach. In *International Conference on Pattern Recognition (ICPR)*, pages 32–36, August 2004. 2

[69] Scikit-Video Developers. Scikit-video: Video processing in python, 2015. `https://github.com/scikit-video/scikit-video`. 7

[70] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision (ECCV)*, pages 510–526, October 2016. 2, 4

[71] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, December 2012. 2

[72] Rajkumar Theagarajan, Federico Pala, Xiu Zhang, and Bir Bhanu. Soccer: Who has the ball? Generating visual analytics and player statistics. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1830–1838, June 2018. 1

[73] Graham Thomas, Rikke Gade, Thomas B. Moeslund, Peter Carr, and Adrian Hilton. Computer vision for sports: Current applications and research topics. *Computer Vision and Image Understanding*, 159:3–18, June 2017. 1

[74] Matteo Tomei, Lorenzo Baraldi, Simone Calderara, Simone Bronzin, and Rita Cucchiara. Rms-net: Regression and masking for soccer event spotting. In *International Conference on Pattern Recognition (ICPR)*, 2020. 2, 7

[75] Xiaofeng Tong, Qingshan Liu, and Hanqing Lu. Shot classification in broadcast soccer video. *Electronic Letters on Computer Vision and Image Analysis*, 7(1):16–25, November 2008. 3

[76] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497, 2015. 5

[77] Bastien Vanderplaetse and Stephane Dupont. Improved soccer action spotting using both audio and video streams. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3921–3931, June 2020. 2, 5, 7

[78] Kanav Vats, Mehrnaz Fani, Pascale Walters, David A Clausi, and John Zelek. Event detection in coarsely annotated sports videos via parallel multi-receptive field 1d convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 882–883, June 2020. 2, 7

[79] Jinjun Wang, EngSiong Chng, and Changsheng Xu. Soccer replay detection using scene transition structure analysis. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages ii/433–ii/436, March 2005. 3

[80] Philippe Weinzaepfel, Xavier Martin, and Cordelia Schmid. Human action localization with sparse spatial supervision. *CoRR*, May 2016. 2

[81] Wei Xu and Yang Yi. A robust replay detection algorithm for soccer video. *IEEE Signal Processing Letters*, 18(9):509–512, July 2011. 3

[82] Ying Yang and Danyang Li. Robust player detection and tracking in broadcast soccer video based on enhanced particle filter. *Journal of Visual Communication and Image Representation*, 46:81–94, July 2017. 1

[83] Ying Yang, Shouxun Lin, Yongdong Zhang, and Sheng Tang. A statistical framework for replay detection in soccer video. In *International Symposium on Circuits and Systems*, pages 3538–3541, May 2008. 3

[84] Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, and Li Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *International Journal of Computer Vision*, 126(2-4):375–389, April 2018. 2

[85] Junqing Yu, Aiping Lei, Zikai Song, Tingting Wang, Hengyou Cai, and Na Feng. Comprehensive dataset of broadcast soccer videos. In *IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 418–423, 2018. 2, 4

[86] Ramin Zabih, Justin Miller, and Kevin Mai. A feature-based algorithm for detecting and classifying scene breaks. In *ACM International Conference on Multimedia*, pages 189–200, November 1995. 2

[87] Dan Zecha, Moritz Einfalt, and Rainer Lienhart. Refining joint locations for human pose tracking in sports videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2524–2532, June 2019. 1

[88] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. Hacs: Human action clips and segments dataset for recognition and temporal localization. In *IEEE International Conference on Computer Vision (ICCV)*, pages 8668–8678, October-November 2019. 2, 4

[89] Zhao Zhao, Shuqiang Jiang, Qingming Huang, and Guangyu Zhu. Highlight summarization in sports video based on replay detection. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1613–1616, July 2006. 3