SoccerNet: A Scalable Dataset for Action Spotting in Soccer Videos

Silvio Giancola, Mohieddine Amine, Tarek Dghaily, Bernard Ghanem
King Abdullah University of Science and Technology (KAUST), Saudi Arabia

`silvio.giancola@kaust.edu.sa, maa249@mail.aub.edu, tad05@mail.aub.edu, bernard.ghanem@kaust.edu.sa`

Abstract

In this paper, we introduce SoccerNet, *a benchmark for action spotting in soccer videos. The dataset is composed of 500 complete soccer games from six main European leagues, covering three seasons from 2014 to 2017 and a total duration of 764 hours. A total of 6,637 temporal annotations are automatically parsed from online match reports at a one minute resolution for three main classes of events (Goal, Yellow/Red Card, and Substitution). As such, the dataset is easily scalable. These annotations are manually refined to a one second resolution by anchoring them at a single timestamp following well-defined soccer rules. With an average of one event every 6.9 minutes, this dataset focuses on the problem of localizing very sparse events within long videos. We define the task of* spotting *as finding the anchors of soccer events in a video. Making use of recent developments in the realm of generic action recognition and detection in video, we provide strong baselines for detecting soccer events. We show that our best model for classifying temporal segments of length one minute reaches a mean Average Precision (mAP) of 67.8%. For the spotting task, our baseline reaches an Average-mAP of 49.7% for tolerances δ ranging from 5 to 60 seconds. Our dataset and models are available at* https://silviogiancola.github.io/SoccerNet.

1. Introduction

Sports is a lucrative sector, with large amounts of money being invested on players and teams. The global sports market is estimated to generate an annual revenue of $91 billion [17], whereby the European soccer market contributes about $28.7 billion [18], from which $15.6 billion alone come from the Big Five European soccer leagues (EPL, La Liga, Ligue 1, Bundesliga and Serie A) [19, 20]. After merchandising, TV broadcast rights are the second major revenue stream for a soccer club [21]. Even though the main scope of soccer broadcast is entertainment, such videos are also used by professionals to generate statistics, analyze strategies, and scout new players. Platforms such as Wyscout [83], Reely [59], and Stats SPORTVU [72] have made sports analytics their core business and already pro-

Figure 1. Example of events defined in the context of soccer. From top to bottom: **Goal**: the instant the ball crosses the goal line. **Substitution**: the instant a players enters the field to substitute an other player. **Card**: the instant the referee shows a card to a player.

vide various products for advanced statistics and highlights.

In order to get such statistics, professional analysts watch a lot of broadcasts and identify all the events that occur within a game. According to Matteo Campodonico, CEO of Wyscout, a 400 employee company focusing on soccer data analytics [83], it takes over 8 hours to provide up to 2000 annotations per game. With more than 30 soccer leagues in Europe, the number of games is very large and requires an army of annotators. Even though Amazon Mechanical Turk (AMT) can provide such workforce, building an annotated dataset of soccer games comes at a significant cost.

Automated methods for sports video understanding can help in the localization of the salient actions of a game. Several companies such as Reely [59] are trying to build automated methods to understand sports broadcasts and would benefit from a large-scale annotated dataset for training and evaluation. Many recent methods exist to solve generic human activity localization in video focusing on sports [7, 10, 26, 42]. However, detecting soccer actions is a difficult task due to the sparsity of the events within a video. Soccer broadcast understanding can thus be seen as a sub-problem of video understanding, focusing on a vocabulary of sparse events defined within its own context.

Contributions. **(i)** We propose the task of *event spotting* within a soccer context. We define *events* as actions anchored to a single timestamp in a video and, thus, proceed to define and study the task of *spotting* events within soccer videos (Section 3). **(ii)** We propose *SoccerNet*, a scalable dataset for soccer video understanding. It contains 764 hours of video and 6,637 instances split in three classes (Goal, Yellow/Red Card, and Substitution), which makes it the largest localization dataset in term of total duration and number of instances per class (Section 4). **(iii)** We provide baselines for our dataset in the tasks of video chunk classification and event spotting. Our *minute classifier* reaches a performance of 67.8% (mAP) and our *event spotter* an Average-mAP of 49.7% (Section 5).

2. Related Work

This paper relates to the topics of Sports Analytics, Activity Recognition and Action Localization Datasets. We give a brief overview of work relevant to each of these topics and highlight how our paper contributes to each of them.

Sports Analytics. Many automated sports analytics methods have been developed in the computer vision community to understand sports broadcasts [7, 23, 26, 41, 58]. They produce statistics of events within a game by either analyzing camera shots or semantic information. Ekin *et al*. [24] present a cornerstone work for game summarization based on camera shot segmentation and classification, followed by Ren *et al*. [60] who focus on identifying video production patterns. Huang *et al*. [35] analyze semantic information to automatically detect goals, penalties, corner kicks, and card events. Tavassolipour *et al*. [76] use Bayesian networks to summarize games by means of semantic analysis.

More recent work in this category focuses on deep learning pipelines to localize salient actions in soccer videos. Baccouche *et al*. [6] use a Bag-of-Words (BOW) approach with SIFT features to extract visual content within a frame. They use such representations to train a Long Short Term Memory (LSTM) network that temporally traverses the video to detect the main actions. Jiang *et al*. [38] propose a similar methodology using Convolution Neural Networks (CNN) to extract global video features rather than local descriptors. They also use a play-break structure to generate candidate actions. Tsagkatakis *et al*. [78] present a two-stream approach to detect goals, while Homayounfar *et al*. [33] recently present a deep method for sports field localization, which is crucial for video registration purposes.

The main impediment for all these works is the lack of reference datasets/benchmarks that can be used to evaluate their performance at large-scale and standardize their comparison. They all use small and custom-made datasets, which contain a few dozen soccer games at most. We argue that intelligent sports analytics solutions need to be scalable to the size of the problem at hand. Therefore, to serve and support the development of such scalable solutions, we propose a very large soccer-centric dataset that can be easily expanded and enriched with various types of annotations.

Activity Recognition. Activity recognition focuses on understanding videos by either detecting activities or classifying segments of video according to a predefined set of human-centric action classes. A common pipeline consists of proposing temporal segments [9, 11, 28, 64], which are in turn further pruned and classified [30, 80]. Common methods for activity classification and detection make use of dense trajectories [29, 79, 80, 81], actionness estimation [14, 28, 87], Recurrent Neural Networks (RNN) [8, 9, 25], tubelets [40, 63], and handcrafted features [11, 49, 86].

In order to recognize or detect activities within a video, a common practice consists of **aggregating** local features and **pooling** them, looking for a consensus of characteristics [42, 67]. While naive approaches use mean or maximum pooling, more elaborate techniques such as Bag-of-Words (BOW) [15, 68], Fisher Vector (FV) [36, 54, 56], and VLAD [4, 37] look for a structure in a set of features by clustering and learning to pool them in a manner that improves discrimination. Recent works extend those pooling techniques by incorporating them into Deep Neural Network (DNN) architectures, namely NetFV [46, 55, 74], SoftDBOW [57], and NetVLAD [3, 30]. By looking for correlations between a set of primitive action representations, ActionVLAD [30] has shown state-of-the-art performance in several activity recognition benchmarks.

To further improve activity recognition, recent works focused on exploiting **context** [10, 16, 50], which represent and harness information in both temporal and/or spatial neighborhood, or on **attention** [51], which learns an adaptive confidence score to leverage this surrounding information. In this realm, Caba Heilbron *et al*. [10] develop a semantic context encoder that exploits evidence of objects and scenes within video segments to improve activity detection effectiveness and efficiency. Miech *et al*. [50], winners of the first annual Youtube 8M challenge [2], show how learnable pooling can produce state-of-the-art recognition performance on a very large benchmark, when recognition is coupled with context gating. More recently, several works use temporal context to localize activities in videos [16] or to generate proposals [28]. Furthermore, Nguyen *et al*. [51] present a pooling method that uses spatio-temporal attention for enhanced action recognition, while Pei *et al*. [53] use temporal attention to gate neighboring observations in a RNN framework. Note that attention is also widely used in video captioning [34, 44, 48].

Activity recognition and detection methods are able to provide good results for these complicated tasks. However, those methods are based on DNNs and require large-scale and rich datasets to learn a model. By proposing a large-

scale dataset focusing on event spotting and soccer, we encourage algorithmic development in those directions.

Datasets. Multiple datasets are available for video understanding, especially for video classification. They include **Hollywood2** [47] and **HMDB** [45], both focusing on movies; **MPII Cooking** [62], focusing on cooking activities; **UCF101** [71], for classification in the wild; **UCF Sports** [61], **Olympics Sports** [52] and **Sports-1M** [42], focusing on sports; **Youtube-8M** [2] and **Kinetics** [43], both tackling large scale video classification in the wild. They are widely used in the community but serve the objective of video classification rather than activity localization.

The number of benchmark datasets focusing on action localization is much smaller. **THUMOS14** [39] is the first reasonably scaled benchmark for the localization task with a dataset of 413 untrimmed videos, totaling 24 hours and 6,363 activities, split into 20 classes. **MultiTHU-MOS** [85] is a subset of THUMOS, densely annotated for 65 classes over unconstrained internet videos. **ActivityNet** [12] tackles the issue of general video understanding using a semantic ontology, proposing challenges in trimmed and untrimmed video classification, activity localization, activity proposals and video captioning. ActivityNet 1.3 provides a dataset of 648 hours of untrimmed videos with 30,791 activity candidates split among 200 classes. It is so far the largest localization benchmark in terms of total duration. **Charades** [66] is a recently compiled benchmark for temporal activity segmentation that crowd-sources the video capturing process. After collecting a core set of videos from YouTube, they use AMT to augment their data by recording them at home. This dataset consists of a total of 9,848 videos and 66,500 activities. More recently, Google proposed **AVA** [31] as a dataset to tackle dense activity understanding. They provide 57,600 clips of 3 seconds duration taken from featured films, annotated with 210,000 dense spatio-temporal labels across 100 classes, for a total of 48 hours of video. While the main task of AVA is to classify these 3 seconds segments, such dense annotation can also be used for detection.

Within the multimedia community, **TRECVID** has been the reference benchmark for over a decade [5, 69]. They host a "Multimedia Event Detection" (MED) and a "Surveillance Event Detection" (SED) task every year, using the **HAVIC** dataset [73]. These tasks focus on finding all clips in a video collection that contain a given event, with a textual definition, in multimedia and surveillance settings. Also, Ye *et al.* [84] propose **EventNet**, a dataset for event retrieval based on a hierarchical ontology, similar to ActivityNet. We argue that these two datasets both focus on large-scale information retrieval rather than video understanding.

We propose *SoccerNet*, a scalable and soccer-focused dataset for event spotting. It contains 500 games, 764 hours of video and 6,637 instances split in three classes (Goal,
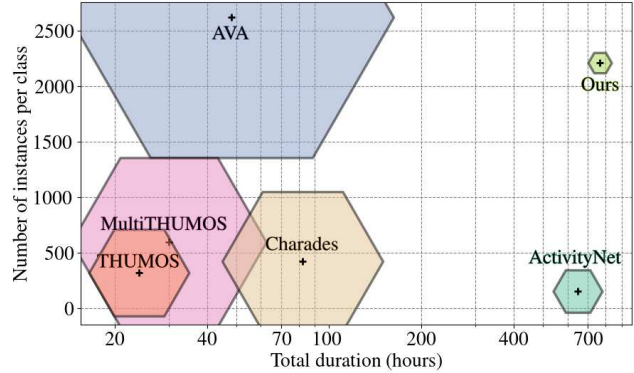

Figure 2. Dataset comparison in term of number of instance per class, and total duration. The size of the hexagon shows the density of the event within the video. Our dataset has the largest amount of instances per class and the largest total duration, despite being sparse, which makes the task of localization more difficult.

Yellow/Red Card, and Substitution), which makes it one of the largest dataset in term of total duration and number of instances per class. With an average of one event every 6.9 minutes, our dataset has a sparse distribution of events in long untrimmed videos, which makes the task of localization more difficult. The annotations are obtained within one minute at no cost by parsing sports websites, and further refined in house to one second resolution. We define our dataset as easily scalable since annotations are obtained for *free* from online match reports. Table 1 shows a breakdown description and comparison of the datasets available for the problem of action localization. Figure 2 shows a graphical comparison between these datasets in terms of the number of instances per class and the total duration of videos they contain.

3. Spotting Sparse Events in Soccer

In this context, we define the concept of *events* and the task of *spotting events* within soccer videos.

Events: Sigurdsson *et al.* [65] recently question the concept of temporal boundaries in activities. They re-annotated Charades [66] and MultiTHUMOS [85] (using AMT), and conclude that the average agreement with the ground truth is respectively 72.5% and 58.8% tIoU. This clearly indicates that temporal boundaries are ambiguous. However, Sigurdsson *et al.* [65] observe that central frames within an activity offer more consensus among annotators.

Chen *et al.* [14] define the concept of *action* and *action-ness* by underlining 4 necessary aspects that define an action: an agent, an intention, a bodily movement, and a side-effect. Dai *et al.* [16] define an *activity* as a set of events or actions, with a beginning and an ending time. In our work, we define the concept of *event* as an action that is anchored in a single time instance, defined within a specific context

Table 1. Comparison of benchmark datasets currently tackling the task of action localization.

Dataset	Context	#Video	#Instance	Duration (hrs)	Sparsity (event/hr)	Classes	Instance per class
THUMOS'14 [39]	General	413	6363	24	260.4	20	318
MultiTHUMOS [85]	General	400	38690	30	1289.7	65	595
ActivityNet [12]	General	19994	30791	648	47.5	200	154
Charades [66]	General	9848	66500	82	811	157	424
AVA [31]	Movies	57600	210000	48	4375	80	2625
Ours	Soccer	1000	6637	764	8.7	3	2212

respecting a specific set of rules. We argue that defining every action with temporal boundaries is ambiguous for multiple reasons:

1. An action can occur in a glimpse, such as *"a man dropped his wallet"* or *"a man put a letter in the mail"*. While there are no well-defined boundaries for such actions, a sole instant can readily define these events.

2. An action can be continuous within a live video, hence it is unclear when it starts or stops. For instance, time boundaries in video for actions such as *"the night is falling"* or *"the ice is melting in my glass"*, rely on a subjective discrimination between measurable quantities such as the illumination level or visual changes in matter state.

3. An action can overlap and conflict with another. Consider a video of a man walking his dog, when he suddenly receives a phone call. It is not clear whether the activity *"taking a phone call"* actually cancels out the activity *"walking a dog"*, or the activity *"walking a dog"* should be split into two parts as opposed to one single segment overlapping the *"taking a phone call"* instance.

Current benchmarks such as THUMOS14 [39], ActivityNet [12], and Charades [66] only focus on activities with temporal boundaries and cope with ambiguities by anchoring an activity with a consensus between several annotators. This ambiguity motivates the recently developed AVA [31] dataset that attempts to tackle the atomic characteristic of actions by providing dense fine-scale annotations within a short time duration (3 seconds).

In the multimedia community, the concept of events is generally more vague and overlaps with the concept of actions and activities. In the MED task of the TRECVID benchmark [5], an event is defined as a kit which consists of a *mnemonic title* for the event, a *textual definition*, an *evidential description* that indicates a non-exhaustive list of textual attributes, and a set of *illustrative video examples*. They propose a specific set of events, providing a description and defining rules for the start and end times. Such work underlines our hypothesis that events need to be defined with a set of rules and within specific circumstances.

In the context of live soccer broadcasts, it is unclear when a given action such as *"scoring a goal"* or *"making a foul"* starts and stops. For similar reasons, the beginning and end of activities such as *"scoring a 3 points shot"* or a *"slam dunk"* in a basketball broadcast are ambiguous. We argue that sports respect well-established rules and define an action vocabulary anchored in a single time instance. In fact, soccer rules provide a strict definition of *"goal"*, *"foul"*, *"card"*, *"penalty kick"*, *"corner kick"*, *etc*. and also anchor them within a single time. Similarly, Ramanathan *et al*. [58] define the action *"basket-ball shot"* as a 3 seconds activity and its ending time as the moment the ball crosses the basket. Defining starting or stopping anchors around such events or fixing its duration would be considered as subjective and biased by the application.

Spotting: Rather than identifying the boundaries of an action within a video and looking for similarities within a given temporal Intersection-over-Union (tIoU), we introduce the task of *spotting*. Spotting consists of finding the anchor time (or *spot*) that identifies an *event*. Intuitively, the closer the candidate spot is from the target, the better the spotting is, and its capacity is measured by its distance from the target. Since perfectly spotting a target is intrinsically arduous, we introduce a tolerance δ within which a event is considered to be spotted (*hit*) by a candidate. We believe that event spotting is better defined and easier than detection since it focuses only on identifying the presence of an event within a given tolerance. An iterative process can refine such tolerance at will by using fine localization methods around candidate spots.

By introducing the task of spotting, we also define the metric to be used for evaluation. First of all, we define a candidate spot as positive if it lands within a tolerance δ around the anchor of an event. For each tolerance, we can recast the spotting problem as a general temporal detection problem, where the tIoU threshold used is very small. In that case, we can compute the recall, precision, Average Precision (AP) for each given class, and a mean Average Precision (mAP) across all classes. For general comparison, we also define an Average-mAP over a set of predefined δ tolerances, in our case below the minute.

4. Data Collection

We build our dataset in three steps: **(i)** we collect videos from online sources; **(ii)** we synchronize the game and video times by detecting and reading the game clock; and **(iii)** we parse match reports available in the web to generate temporal aligned annotations.

4.1. Collecting Videos

We compile a set of 500 games from the main European Championships during the last 3 seasons as detailed in Table 2. Each game is composed of 2 untrimmed videos, one for each half period. The videos come from online providers, in a variety of encodings (MPEG, H264), containers (MKV, MP4, and TS), frame rates (25 to 50 fps), and resolutions (SD to FullHD). The dataset consumes almost 4TB, for a total duration of 764 hours. The games are randomly split into 300, 100, and 100 games for training, validation, and testing ensuring similar distributions of the events between the classes and the datasets.

Table 2. Summary of the video collection for our dataset.

| | Seasons | | | |
League	14/15	15/16	16/17	**Total**
EN - EPL	6	49	40	**95**
ES - LaLiga	18	36	63	**117**
FR - Ligue 1	1	3	34	**38**
DE - BundesLiga	8	18	27	**53**
IT - Serie A	11	9	76	**96**
EU - Champions	37	45	19	**101**
Total	**81**	**160**	**259**	**500**

4.2. Game Synchronization with OCR

The video of the games are untrimmed and contains spurious broadcast content before and after the playing time. Finding a mapping between the game time and the video time is necessary to align the annotations from the web sources to the videos. Soccer games have a continuous game flow, *i.e.* the clock never stops before the end of a half, hence there is a simple linear relationship (with slope 1) between the video and the game time. Wang *et al.* [82] propose a method using the center circle of the field and the sound of the referee whistle to identify the start of the game. We argue that focusing the effort on a single instant is prone to error. In contrast, we focus on detecting the game clock region within multiple video frames and identify the game time through Optical Character Recognition (OCR) at different instants.

The clock is displayed in most of the frames throughout the video, though its shape and position vary between leagues. We leverage a statistical study of the pixel intensity deviation within a set of N random frames to identify candidates for the clock region. We run the Tesseract OCR

Engine [70] on the candidate clocks and look for a coherent time format for each of the N frames. To cope with eventual misreadings in the clock, we use a RANSAC [27] approach to estimate the linear relation between the game clock and the video time, enforcing a unitary gradient to our linear model. Our method also checks for the temporal integrity of the video, reporting temporal inconsistencies. To verify the quality of this game-to-video temporal alignment, we manually annotate the start of the game for all 500 games and report an accuracy of 90% for automatically estimating the start of both halves within a tolerance of two seconds, using a set of $N = 200$ frames.

4.3. Collecting Event Annotations

For our dataset, we obtain event annotations for *free* by parsing match reports provided by league websites[1]. They summarize the main actions of the game and provide the minute at which the actions occur. We categorize these events into our three main categories: *"goals"*, *"cards"* and *"substitutions"*. We parse and mine the annotations for all games of the Big Five European leagues (EPL, La Liga, Ligue 1, Bundesliga and Serie A) as well as the Champions League from 2010 to 2017, for a total of 171,778 annotations corresponding to 13,489 games. For sake of storage, we focus on our subset of videos for the 500 games and use only 6,637 events. To resolve these free annotations to one second level, we manually annotate each event within one second resolution by first retrieving its minute annotation and refining it within that minute window. To do so, we define the temporal anchors for our events from their definitions within the rules of soccer. We define a *"goal"* event as the instant the ball crosses the goal line to end up in the net. We define the *"card"* event as the instant the referee shows a player a yellow or a red card because of a foul or a misbehaviour. Finally, we define the *"substitution"* event as the instant a new player enters in the field. We ensure those definition were coherent when annotating the dataset. Apart for the substitutions that occur during half time break, almost all of our instances follow their definitions.

4.4. Dataset Scalability

We believe that scaling our dataset is cheap and easy, since web annotations are freely available with one minute resolution. Algorithm can either use the weakly annotated events within one minute resolution or generate a complete one second resolution annotation which is estimated to take less than 10 minutes per game. We also argue that broadcast providers can easily scale up such datasets by simply providing more videos and richer annotations.

---

[1]We choose `www.flashscore.info` to get our annotations since they provide a wide number of summaries and have a consistent format across their match reports.

5. Experiments

We focus the attention of our experiments on two tasks: event classification for chunks of one minute duration, and event spotting within an entire video. For these tasks, we report and compare the performance metrics for various baseline methods when trained on weakly annotated data (*i.e.* one minute resolution) and the improvement that is achieved by training on one second resolution annotations.

5.1. Video Representation

Before running any experiments, we extract C3D [77], I3D [13], and ResNet [32] features from our videos to be used by our baselines. The videos are trimmed at the game start, resized and cropped at a 224×224 resolution, and unified at 25fps. Such representation guarantees storage efficiency, fast frame access, and compatible resolution for feature extraction. **C3D** [77] is a 3D CNN feature extractor that stacks 16 consecutive frames and outputs at the *fc7* layer a feature of dimension 4,096. It is pretrained on Sport-1M [42]. **I3D** [13] is based on Inception V1 [75], uses 64 consecutive frames, and is pretrained on Kinetics [43]. In this work, we only extract the RGB features at the *PreLogits* layer of length 1,024 so to maintain a reasonable computational runtime. They have been shown to produce only meager improvements when flow features are used [13]. **ResNet** [32] is a very deep network that outputs a 2,048 feature representation per frame at the *fc1000* layer. In particular, we use ResNet-152 pretrained on ImageNet [22]. Since ResNet-152 applies to single images, it does not intrinsically embed contextual information along the time axis. We use TensorFlow [1] implementations to extract features every 0.5 second (s). In order to simplify and unify the dimension of those features, we reduce their dimensionality by applying Principal Component Analysis (PCA) on the 5.5M features we extract per model. We reduce C3D, I3D, and ResNet-152 features to a dimension of 512 and respectively maintain 94.3%, 98.2%, and 93.9% of their variance. For the benchmark purpose, we provide the original and cropped versions of the videos, as well as the original and the reduced versions of all the features extracted every 0.5s.

5.2. Video Chunk Classification

Similar to the setup in the AVA dataset [31], localization can be cast as a classification problem for densely annotated chunks of video, especially since we gather webly annotations. We split our videos into chunks of duration 1 minute, annotated with all events occurring within this minute, gathering respectively 1246, 1558, 960 and 23750 chunks for cards, substitutions, goals and backgrounds for the training dataset, 115 of which having multiple labels. We aggregate the 120 features within a minute as input for different versions of shallow pooling neural networks. By using a

sigmoid activation function at the last layer of these networks, we allow for multi-labelling across the candidates. We use an Adam optimizer that minimizes a multi binary cross-entropy loss for all the classes. We used a step decay for the learning rate and an early stopping technique based on the validation set performances. Following best practices in the field, the evaluation metric in this case is mAP (classification) across the three classes on the designated testing set. In what follows, we report strong baseline results using different video features, different pooling techniques, and compare solutions to cope with the imbalanced dataset.

Learning How to Pool: We investigate the usage of different feature representations and various pooling methods. We propose shallow neural networks that handles the input matrix of dimension 120×512. We test a **mean** and a **max pooling** operation along the aggregation axis that output 512-long features. We use a custom **CNN** with a kernel of dimension 512×20 that traverses the temporal dimension to gather temporal context. Finally, we use implementations of **SoftDBOW**, **NetFV**, **NetVLAD** and **NetRVLAD** provided by Miech *et al.* [50], who leverage a further context-gating layer. After each of these pooling layer, we stack a fully connected layer with a dropout layer (keep probability 60%) that predicts the labels for the minutes of video and prevent overfitting.

Table 3 summarizes a performance comparison between the various pooling methods when applied to the testing set. First of all, we notice similar results across features by using mean and max pooling, that only rely on a single representation of the set of 120 features and not its distribution. Using the custom CNN layer, which is an attempt to gather temporal context, ResNet-152 performs better than C3D which performs better than I3D. We believe that the I3D and C3D already gather temporal information for 64 and 16 frames.

We can notice that the gap between the features increases by using the pooling methods proposed by Miech *et al.* [50], which is a way to embed context along the temporal dimension. We believe that I3D and C3D features already rely on a temporal characterization within the stack of frames. On the other hand, the ResNet-152 provides a representation that focuses only on the spatial aspect within a frame. We believe that the temporal pooling methods provides more redundant information for I3D and C3D, than for ResNet-152. For this reason, we argue that ResNet-152 features provide better results when coupled with any temporal pooling methods provided by Miech *et al.* [50].

Focusing on the pooling, VLAD-based methods are at the top of the ranking, followed by the deep versions of the FV and BoW methods. Such improvement is attributed to the efficient clustering for the 120 features learned in NetVLAD [3] providing state-of-the-art results for action classification [30]. Note that NetRVLAD performs similarly if not better than NetVLAD by relying only on the

average and not the residuals for each clustering, reducing the computational load [50]. For the rest of the experiment we are relying exclusively on ResNet-152 features.

Table 3. Classification metric (mAP) for different combinations of frame representations and pooling methods.

Pooling	Frame features		
	I3D	**C3D**	**ResNet**
Mean Pool.	40.8	40.7	40.2
Max Pool.	50.1	52.4	52.4
CNN	44.1	47.8	53.5
SoftDBOW	46.3	56.9	58.9
NetFV	44.7	59.6	64.4
NetRVLAD	41.3	59.9	**65.9**
NetVLAD	43.4	60.3	65.2

For the various pooling methods, the number of clusters can be fine-tuned. In Table 3, we use $k = 64$ clusters, which can be interpreted as the vocabulary of atomic elements that are learned to describe the events. Intuitively, one can expect that a richer and larger vocabulary can enable better overall performance [30]. We show in Table 4 that this intuition is true within a certain range of values k, beyond which the improvement is negligible and overfitting occurs. The performance of all pooling methods seem to plateau when more than 256 clusters are used for the quantization. The best results are registered when NetVLAD is used with 512 clusters. Nevertheless, the computational complexity increases linearly with the number of clusters, hence computational times grow drastically.

Table 4. Classification metric (mAP) for different number of cluster for the pooling methods proposed by Miech *et al.* [50].

k	Pooling Methods			
	SoftBOW	**NetFV**	**NetRVLAD**	**NetVLAD**
16	54.9	63.0	64.4	65.2
32	57.7	64.0	63.8	65.1
64	58.8	64.1	65.3	65.2
128	60.6	64.4	67.0	65.6
256	61.3	63.8	67.7	67.0
512	62.0	62.1	67.4	**67.8**

Coping with Imbalanced Data: The performance of classifiers are significantly affected when training sets are imbalanced. Due to the sparsity of our events, we have numerous background instances. Here, we present three main techniques to cope with this imbalance. One method focuses on **weighting (Weig)** the binary cross-entropy with the ratio of negative samples to enforce the learning of the positive examples. Another method applies a **random downsampling (Rand)** on the highest frequency classes, or by **hard negative mining (HNM)**, *i.e.* by sampling the examples that are misclassified the most in the previous epoch. The third method uses **Data Augmentation (Augm)** to bal-

ance the classes. In that case, we use the fine annotation of the event and slide the minute window with a stride of 1s within ±20s of the event spot to sample more video segments for the sparsest event classes. We argue that a chunk of 1 minute within ±10s around the anchor of the event still contains this event, and the pooling method should be able to identify it. Although, note that our data augmentation requires the data to be finely annotated.

Table 5 shows the classification mAP for the testing dataset, training with the previous pooling methods on ResNet features, and using the aforementioned strategies to cope dataset imbalance. We see that weighting slightly improves the metric. Both downsampling methods actually lead to the worst results, because of the reduced amount of data the model has been trained on at each epoch. Using the second resolution annotations to augment the data helps to achieve slightly better classification results.

Table 5. Classification metric (mAP) using different solutions to cope with an imbalanced dataset on our pooling methods, using ResNet-152 features.

Pooling	Imbalance				
	Orig	**Weig**	**Rand**	**HNM**	**Augm**
Mean Pool.	41.5	42.4	40.9	42.4	43.9
Max Pool.	39.3	52.4	48.3	52.4	51.5
CNN	52.9	52.1	50.5	49.8	53.5
SoftDBOW	58.4	59.7	48.7	48.8	50.8
NetFV	64.2	63.8	58.0	60.5	65.2
NetRVLAD	65.2	64.9	61.4	61.0	**66.7**
NetVLAD	65.0	64.9	59.4	59.5	65.1

5.3. Spotting

In this section, we discuss the task of event spotting in soccer videos. We use the models trained in the classifier task and apply them in a sliding window fashion on each testing video, with a stride of 1s, thus, leading to a second resolution score along for each event class. We investigate the spotting results of three strong baselines **(i)** a watershed method to compute segment proposals and use the **center** time within the segment to define our candidate; **(ii)** the time index of the **maximum** value of the watershed segment as our candidate; and **(iii)** the local maxima along all the video and apply *non-maximum-suppression* (**NMS**) within a minute window. The evaluation metric is the mAP with tolerance δ as defined for spotting in Section 3, as well as, the Average-mAP expressed as an area under the mAP curve with tolerance ranging from 5 to 60 seconds.

Comparison of Spotting Baselines: We investigate the results for event spotting for our best weakly-trained classifier, to leverage the use of webly-parsed annotations, *i.e.* we train on the imbalanced minute resolution annotated data and do not perform data augmentation. Specifically, we use a NetVLAD model with $k = 512$ cluters based on ResNet
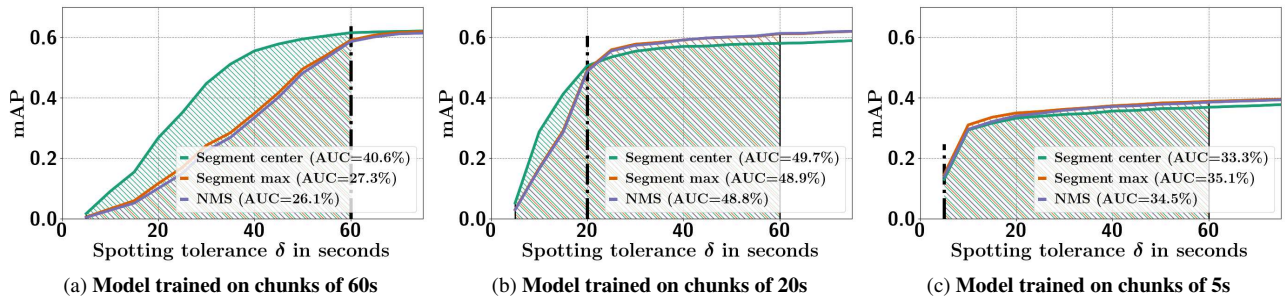
| | (a) **Model trained on chunks of 60s** | (b) **Model trained on chunks of 20s** | (c) **Model trained on chunks of 5s** |

Figure 3. Spotting metric (mAP) in function of the tolerance δ for model trained on chunks of size (a) 60s, (b) 20s and (c) 5s. The Average-mAP is estimated through the area under the curve between 5s and 60s for each baseline.

features and the watershed threshold is set to 50%.

Figure 3a plots the mAP of each spotting baseline as a function of the tolerance δ to the spot. As expected, the mAP decreases with the spot tolerance δ. Above a tolerance δ of 60s, both three baselines plateau at 62.3%. Below 60s, the baseline (ii) and (iii) perform similarly and decrease linearly with the tolerance. On the other hand, baseline (i) decreases more gradually, hence provides a better Average-mAP of 40.6%. Even though the model has been trained using chunks of 1 minute, the method is still able to achieve good spotting results for tolerances below 60s. We argue that our model predicts positively any window that contains an event, creating a plateau.

Training on Smaller Windows: Here, we train our classifiers from Section 5.2 using a smaller chunk size, ranging from 60 seconds to 5 seconds. We expect these models to perform in a similar fashion, with a drop in performance (mAP) occurring for tolerances below the chunk size. Note that we use finely annotated data to train such classifiers.

Figure 3 depicts the spotting mAP in function of the tolerance δ for the models trained on 60, 20 and 5 seconds. They all have similar shape, a metric that plateaus for spotting tolerance δ above the chunk video length they have being trained on, and a decreasing metric below such threshold. By using baseline (i) on chunks of 20s we obtain the best Average-mAP of 50% (see Figure 3b). Also, a drop in performance occurs with models trained with chunks of 5s (see Figure 3c). We believe such gap in performance is related to the amount of context we allow around the event.

With these experiments, we setup a baseline for the spotting task but the best performance is far from satisfactory. Nevertheless, we see our newly compiled and scalable dataset to be a rich environment for further algorithm development and standardized evaluations; especially when it comes to novel spotting techniques.

6. Future Work

Activity detection is commonly solved by proposing candidates that are further classified. We believe that detection can be solved by spotting a candidate and focusing attention around the spot to localize the activity boundaries.

In future works, we encourage the usage of RNNs to embed a further temporal aspect that will understand the evolution of the game. We will also include more classes for soccer events to enrich its contents and enable learning potential causal relationships between events. We believe for instance that the event *"card"* is mostly the result of an event *"foul"*. Also, embedding semantic relationship information from the players, the ball and the field can improve soccer video understanding. Our video also contains an audio track that should be used; visual and audio sentiment analysis could localize the salient moments of the game.

The match reports from our online provider also includes match commentaries. We collected and will release a total of 506,137 commentaries for the six aforementioned leagues with a one second resolution. We believe such data can be used for captioning events in soccer videos.

7. Conclusion

In this paper, we focus on soccer understanding in TV broadcast videos. We build this work as an attempt to provide a benchmark for soccer analysis, by providing a large-scale annotated dataset of soccer game broadcasts. We discussed the concept of *event* within the soccer context, proposed a definition of *"goal"*, *"card"* and *"substitution"* and parse a large amount of annotation from the web. We defined the task of *spotting* and provide a baseline for it. For the minute classification task, we have shown performance of 67.8% (mAP) using ResNet-152 features and NetVLAD pooling along a 512-long vocabulary and using a coarse annotation. Regarding the spotting task, we have establish an Average-mAP of 49.7% with fine annotation and 40.6% by using only weakly annotated data. We believe that focusing effort on spotting, new algorithms can improve the state-of-the-art in detection tasks.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org. 6

[2] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016. 2, 3

[3] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *CVPR*, pages 5297–5307, 2016. 2, 6

[4] R. Arandjelovic and A. Zisserman. All about VLAD. In *CVPR*, pages 1578–1585, 2013. 2

[5] G. Awad, J. Fiscus, M. Michel, D. Joy, W. Kraaij, A. F. Smeaton, G. Quénot, M. Eskevich, R. Aly, and R. Ordelman. Trecvid 2016: Evaluating video search, video event detection, localization, and hyperlinking. In *Proceedings of TRECVID*, volume 2016, 2016. 3, 4

[6] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt. Action classification in soccer videos with long short-term memory recurrent neural networks. *Artificial Neural Networks–ICANN 2010*, pages 154–159, 2010. 2

[7] V. Bettadapura, C. Pantofaru, and I. Essa. Leveraging contextual cues for generating basketball highlights. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 908–917. ACM, 2016. 1, 2

[8] S. Buch, V. Escorcia, B. Ghanem, L. Fei-Fei, and J. C. Niebles. End-to-end, single-stream temporal action detection in untrimmed videos. In *BMVC*, 2017. 2

[9] S. Buch, V. Escorcia, C. Shen, B. Ghanem, and J. Carlos Niebles. Sst: Single-stream temporal action proposals. In *CVPR*, July 2017. 2

[10] F. Caba Heilbron, W. Barrios, V. Escorcia, and B. Ghanem. Scc: Semantic context cascade for efficient action detection. In *CVPR*, pages 1454–1463, 2017. 1, 2

[11] F. Caba Heilbron, J. Carlos Niebles, and B. Ghanem. Fast temporal activity proposals for efficient detection of human actions in untrimmed videos. In *CVPR*, pages 1914–1923, 2016. 2

[12] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015. 3, 4

[13] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. *arXiv preprint arXiv:1705.07750*, 2017. 6

[14] W. Chen, C. Xiong, R. Xu, and J. J. Corso. Actionness ranking with lattice conditional ordinal random fields. In *CVPR*, pages 748–755, 2014. 2, 3

[15] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, pages 1–2. Prague, 2004. 2

[16] X. Dai, B. Singh, G. Zhang, L. S. Davis, and Y. Qiu Chen. Temporal context network for activity localization in videos. In *ICCV*, Oct 2017. 2, 3

[17] 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

[18] 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

[19] 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

[20] 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

[21] Deloitte. Top-20 european football clubs breakdown of revenues 2015/16 season (in million euros). In *Statista - The Statistics Portal*, 2017. Retrieved October 30, 2017, from https://www.statista.com/statistics/271636/revenue-distribution-of-top-20-european-soccer-clubs/. 1

[22] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE, 2009. 6

[23] T. DOrazio and M. Leo. A review of vision-based systems for soccer video analysis. *Pattern recognition*, 43(8):2911–2926, 2010. 2

[24] A. Ekin, A. M. Tekalp, and R. Mehrotra. Automatic soccer video analysis and summarization. *IEEE Transactions on Image processing*, 12(7):796–807, 2003. 2

[25] V. Escorcia, F. C. Heilbron, J. C. Niebles, and B. Ghanem. DAPs: Deep action proposals for action understanding. In *ECCV*, pages 768–784. Springer, 2016. 2

[26] P. Felsen, P. Agrawal, and J. Malik. What will happen next? forecasting player moves in sports videos. In *ICCV*, Oct 2017. 1, 2

[27] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 5

[28] J. Gao, Z. Yang, K. Chen, C. Sun, and R. Nevatia. Turn tap: Temporal unit regression network for temporal action proposals. In *ICCV*, Oct 2017. 2

[29] J. Gemert, M. Jain, E. Gati, and C. G. Snoek. Apt: Action localization proposals from dense trajectories. In *BMVC*. BMVA Press, 2015. 2

[30] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, and B. Russell. ActionVLAD: Learning spatio-temporal aggregation for action classification. *arXiv preprint arXiv:1704.02895*, 2017. 2, 6, 7

[31] C. Gu, C. Sun, S. Vijayanarasimhan, C. Pantofaru, D. A. Ross, G. Toderici, Y. Li, S. Ricco, R. Sukthankar, C. Schmid, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. *arXiv preprint arXiv:1705.08421*, 2017. 3, 4, 6

[32] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6

[33] N. Homayounfar, S. Fidler, and R. Urtasun. Sports field localization via deep structured models. In *CVPR*, July 2017. 2

[34] C. Hori, T. Hori, T.-Y. Lee, Z. Zhang, B. Harsham, J. R. Hershey, T. K. Marks, and K. Sumi. Attention-based multimodal fusion for video description. In *ICCV*, Oct 2017. 2

[35] 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. 2

[36] T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *Advances in neural information processing systems*, pages 487–493, 1999. 2

[37] H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *CVPR*, pages 3304–3311. IEEE, 2010. 2

[38] H. Jiang, Y. Lu, and J. Xue. Automatic soccer video event detection based on a deep neural network combined cnn and rnn. In *Tools with Artificial Intelligence (ICTAI), 2016 IEEE 28th International Conference on*, pages 490–494. IEEE, 2016. 2

[39] Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. http://crcv.ucf.edu/THUMOS14/, 2014. 3, 4

[40] V. Kalogeiton, P. Weinzaepfel, V. Ferrari, and C. Schmid. Action tubelet detector for spatio-temporal action localization. In *ICCV*, Oct 2017. 2

[41] R. Kapela, K. McGuinness, A. Swietlicka, and N. E. OConnor. Real-time event detection in field sport videos. In *Computer vision in Sports*, pages 293–316. Springer, 2014. 2

[42] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014. 1, 2, 3, 6

[43] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijaya-narasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 3, 6

[44] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. Carlos Niebles. Dense-captioning events in videos. In *ICCV*, Oct 2017. 2

[45] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *ICCV*, 2011. 3

[46] G. Lev, G. Sadeh, B. Klein, and L. Wolf. Rnn fisher vectors for action recognition and image annotation. In *ECCV*, pages 833–850. Springer, 2016. 2

[47] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *CVPR*, pages 2929–2936. IEEE, 2009. 3

[48] A. Mazaheri, D. Zhang, and M. Shah. Video fill in the blank using lr/rl lstms with spatial-temporal attentions. In *ICCV*, Oct 2017. 2

[49] P. Mettes, J. C. van Gemert, S. Cappallo, T. Mensink, and C. G. Snoek. Bag-of-fragments: Selecting and encoding video fragments for event detection and recounting. In *Proceedings of the 5th ACM on ICMR*, pages 427–434. ACM, 2015. 2

[50] A. Miech, I. Laptev, and J. Sivic. Learnable pooling with context gating for video classification. *arXiv preprint arXiv:1706.06905*, 2017. 2, 6, 7

[51] T. V. Nguyen, Z. Song, and S. Yan. Stap: Spatial-temporal attention-aware pooling for action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 25(1):77–86, 2015. 2

[52] J. C. Niebles, C.-W. Chen, and L. Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *ECCV*, pages 392–405. Springer, 2010. 3

[53] W. Pei, T. Baltrusaitis, D. M. Tax, and L.-P. Morency. Temporal attention-gated model for robust sequence classification. In *CVPR*, July 2017. 2

[54] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, pages 1–8. IEEE, 2007. 2

[55] F. Perronnin and D. Larlus. Fisher vectors meet neural networks: A hybrid classification architecture. In *CVPR*, pages 3743–3752, 2015. 2

[56] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. *ECCV*, pages 143–156, 2010. 2

[57] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *CVPR*, pages 1–8. IEEE, 2008. 2

[58] V. Ramanathan, J. Huang, S. Abu-El-Haija, A. Gorban, K. Murphy, and L. Fei-Fei. Detecting events and key actors in multi-person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3043–3053, 2016. 2, 4

[59] Reely. Reely. In *Reely*, 2017. Retrieved October 30, 2017, from http://www.reely.ai. 1

[60] R. Ren and J. M. Jose. Football video segmentation based on video production strategy. In *European Conference on Information Retrieval*, pages 433–446. Springer, 2005. 2

[61] M. D. Rodriguez, J. Ahmed, and M. Shah. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, pages 1–8. IEEE, 2008. 3

[62] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele. A database for fine grained activity detection of cooking activities. In *CVPR*, pages 1194–1201. IEEE, 2012. 3

[63] S. Saha, G. Singh, and F. Cuzzolin. Amtnet: Action-micro-tube regression by end-to-end trainable deep architecture. In *ICCV*, Oct 2017. 2

[64] Z. Shou, D. Wang, and S.-F. Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *CVPR*, pages 1049–1058, 2016. 2

[65] G. A. Sigurdsson, O. Russakovsky, and A. Gupta. What actions are needed for understanding human actions in videos? In *ICCV*, Oct 2017. 3

[66] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, 2016. 3, 4

[67] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014. 2

[68] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *null*, page 1470. IEEE, 2003. 2

[69] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *8th ACM international workshop on Multimedia information retrieval*, pages 321–330. ACM, 2006. 3

[70] R. Smith. An overview of the tesseract ocr engine. In *Document Analysis and Recognition, 2007. ICDAR 2007*, volume 2, pages 629–633. IEEE, 2007. 5

[71] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 3

[72] SportVU. Sportvu. In *SportVU*, 2017. Retrieved October 30, 2017, from https://www.stats.com/sportvu-football. 1

[73] S. Strassel, A. Morris, J. G. Fiscus, C. Caruso, H. Lee, P. Over, J. Fiumara, B. Shaw, B. Antonishek, and M. Michel. Creating havic: Heterogeneous audio visual internet collection. In *LREC*, volume 4, page 2, 2012. 3

[74] V. Sydorov, M. Sakurada, and C. H. Lampert. Deep fisher kernels-end to end learning of the fisher kernel gmm parameters. In *CVPR*, pages 1402–1409, 2014. 2

[75] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016. 6

[76] 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. 2

[77] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497, 2015. 6

[78] G. Tsagkatakis, M. Jaber, and P. Tsakalides. Goal!! event detection in sports video. *Electronic Imaging*, 2017(16):15–20, 2017. 2

[79] H. Wang and C. Schmid. Action recognition with improved trajectories. In *ICCV*, pages 3551–3558, 2013. 2

[80] L. Wang, Y. Qiao, and X. Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *CVPR*, pages 4305–4314, 2015. 2

[81] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, pages 20–36. Springer, 2016. 2

[82] Z. Wang, J. Yu, and Y. He. Soccer video event annotation by synchronization of attack–defense clips and match reports with coarse-grained time information. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(5):1104–1117, 2017. 5

[83] WyScout. Wyscout. In *WyScout*, 2017. Retrieved October 30, 2017, from http://www.wyscout.com. 1

[84] G. Ye, Y. Li, H. Xu, D. Liu, and S.-F. Chang. Eventnet: A large scale structured concept library for complex event detection in video. In *23rd ACM international conference on Multimedia*, pages 471–480. ACM, 2015. 3

[85] S. Yeung, O. Russakovsky, N. Jin, M. Andriluka, G. Mori, and L. Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *arXiv preprint arXiv:1507.05738*, 2015. 3, 4

[86] G. Yu and J. Yuan. Fast action proposals for human action detection and search. In *CVPR*, pages 1302–1311, 2015. 2

[87] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, X. Tang, and D. Lin. Temporal action detection with structured segment networks. In *ICCV*, Oct 2017. 2