Towards Active Learning for Action Spotting in Association Football Videos

Silvio Giancola[*,1] Anthony Cioppa[*,1,2] Julia Georgieva[3] Johsan Billingham[4]
Andreas Serner[5] Kerry Peek[6] Bernard Ghanem[1] Marc Van Droogenbroeck[2]
[1]IVUL, KAUST [2]TELIM, University of Liège [3]Curtin School of Allied Health, Curtin University
[4]Football Technology & Innovation, FIFA [5]FIFA Medical, FIFA [6]Faculty of Medicine and Health, The University of Sydney

Abstract

Association football is a complex and dynamic sport, with numerous actions occurring simultaneously in each game. Analyzing football videos is challenging and requires identifying subtle and diverse spatio-temporal patterns. Despite recent advances in computer vision, current algorithms still face significant challenges when learning from limited annotated data, lowering their performance in detecting these patterns. In this paper, we propose an active learning framework that selects the most informative video samples to be annotated next, thus drastically reducing the annotation effort and accelerating the training of action spotting models to reach the highest accuracy at a faster pace. Our approach leverages the notion of uncertainty sampling to select the most challenging video clips to train on next, hastening the learning process of the algorithm. We demonstrate that our proposed active learning framework effectively reduces the required training data for accurate action spotting in football videos. We achieve similar performances for action spotting with NetVLAD++ on SoccerNet-v2, using only one-third of the dataset, indicating significant capabilities for reducing annotation time and improving data efficiency. We further validate our approach on two new datasets that focus on temporally localizing actions of headers and passes, proving its effectiveness across different action semantics in football. We believe our active learning framework for action spotting would support further applications of action spotting algorithms and accelerate annotation campaigns in the sports domain.

1. Introduction

Video analysis is a rapidly evolving field with numerous applications in various domains, such as surveillance [50], sports [53], and autonomous driving [39]. One of the essential tasks in video analysis is action spotting, which aims at identifying and precisely localizing specific actions anchored with a single timestamp in video sequences. This task has gained considerable attention in recent years due to its importance in various applications, such as video search [15, 48], video summarization [7, 13], and activity recognition [3, 25].

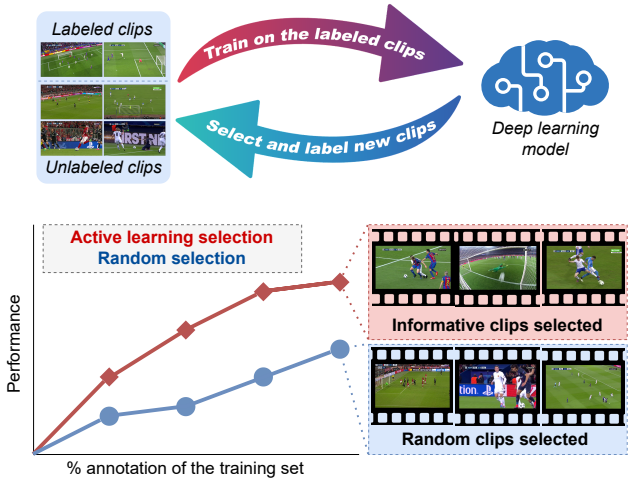

Figure 1. **Active learning for action spotting.** Given a video clip dataset, our active learning framework iteratively **(i)** trains a deep learning model on the labeled clips, and **(ii)** selects the next video clips to be labeled by an oracle. By actively selecting the most informative video samples to annotate next, we accelerate the tagging of unlabeled datasets for training action spotting models.

Traditionally, action spotting is addressed using supervised learning techniques, where a labeled dataset is used to train a classifier that can recognize actions and temporally localize them in the videos. However, annotating large-scale video datasets is time-consuming and expensive, limiting the scalability and applicability of supervised learning approaches. Active learning comes as a promising approach that can mitigate the need for completely annotating large datasets by selecting the most informative samples needed for labeling.

In this paper, we propose an active learning framework

(*) Denotes equal contributions
Contacts: silvio.giancola@kaust.edu.sa, anthony.cioppa@uliege.be.

for action spotting in association football videos, which aims to reduce the annotation effort and improve the overall performance of the system with respect to the number of annotations. Our framework, illustrated in Figure 1, integrates active learning with state-of-the-art action spotting methods, where the labeled set is iteratively expanded with informative samples selected from an unlabeled dataset. We evaluate our approach on several benchmark datasets and compare it with a naive random selection approach that does not leverage active learning. We also analyze the impact of different query strategies on the performance of the system.

Contributions. Our contributions may be summarized as follows: **(i)** We propose the first active learning framework for the task of action spotting that iteratively selects relevant clips to be annotated next. **(ii)** We compare a couple of active learning selection strategies based on uncertainty sampling on several benchmark datasets with state-of-the-art action spotting methods. **(iii)** We provide a comprehensive analysis showing the capability of our framework to significantly reduce the quantity of annotation needed to reach desired performances.

2. Related work

2.1. Sports video understanding

The challenging and nuanced nature of sports video analysis has made it an increasingly popular research topic in recent years [36, 53]. The availability of large-scale datasets has played a crucial role in enabling progress in these tasks. Examples of such large-scale datasets include those developed by Pappalardo *et al*. [38], Yu *et al*. [64], SoccerDB [29], SoccerTrack [42], and DeepSportRadar [57]. The SoccerNet dataset, introduced by Giancola *et al*. [17], has become the most comprehensive resource for labeled data related to video understanding in football. It includes benchmarks for ten different tasks, such as action spotting [11], camera calibration [6], and player tracking [8].

Lately, deep learning-based methods have become the go-to approach for many sports video analysis tasks, thanks to their remarkable performance and ability to extract high-level features from raw data. For instance, automatic methods based on deep learning have shown impressive results in tasks such as player tracking [35] and re-identification with occlusion [49], 3D shuttle trajectory reconstruction in badminton [34], medical risk assessment in rugby [37], tactics analysis [52], pass feasibility [1], and talent scouting [10]. Additionally, deep learning-based methods have also allowed researchers to leverage large-scale datasets effectively. Despite their successes, deep learning-based methods still face several challenges, such as dealing with noisy and incomplete data, accounting for complex game scenarios, and generalizing to new domains. As such, there is still ample room for improvement and research in the field of sports video understanding. Furthermore, annotating large-scale datasets for sports video analysis tasks is a time-consuming and expensive process, requiring significant human resources and expertise. As a solution, Vandeghen *et al*. [58] proposed a semi-supervised method for player detection, leveraging a large unlabeled dataset. Vats *et al*. [61] propose a weakly supervised approach for player identification using transformers. To address this data issue, we propose an active learning approach for action spotting in football that aims at reducing the amount of annotated data needed while maintaining high task performance.

2.2. Action spotting

Action spotting is an important task in football video understanding, as it involves localizing specific events in an untrimmed football broadcast video, including, for instance, penalties, goals, or corners. Unlike temporal activity localization [3], action spotting describes events using a single timestamp, following the definition of actions defined in the football rules [27]. Recent studies have explored the use of large-scale datasets such as SoccerNet [17], which has been expanded from 3 to 17 classes to encompass all possible actions that occur during a game [11]. This dataset has generated significant interest in the research community, as evidenced by the open challenges [18], showing that action spotting is currently experiencing a high level of activity and attention in the research and industrial communities.

The first method for action spotting was proposed by Giancola *et al*. [17], which is based on temporal pooling. Later, they refined their method by aggregating the temporal context [19]. Rongved *et al*. [41] proposed an approach based on applying a 3D ResNet directly to the video frames in a 5-second sliding window fashion. Vanderplaetse *et al*. [59] and Xarles *et al*. [18] combined visual and audio features in a multimodal approach. Cioppa *et al*. [7] introduced a context-aware loss function to model the temporal context surrounding the actions. Vats *et al*. [60] used a multi-tower CNN that accounts for the uncertainty of the action locations, and Tomei *et al*. [55] fine-tuned a feature extractor and used a masking strategy to focus on the frames after the actions.

The current state-of-the-art on SoccerNet-v2 is held by Soares *et al*. [46, 47], who won the SoccerNet 2022 challenge by proposing an anchor-based approach. They define an anchor as a pair formed by a time instant and an action class, with time instants sampled densely. For each anchor, both a detection confidence and a fine-grained temporal displacement are inferred, with the displacement indicating exactly when an action is predicted to happen. Their approach results in a substantial improvement in temporal precision. Hong *et al*. [24], the runner-ups of the 2022 challenge, proposed the first precise temporal spotting (PTS) method where both the feature extractor and the spotting

head are trained in an end-to-end fashion. They rely on a light-weight RegNet architecture, including a GSM [51] module and a GRU [5] module on top that classifies each frame into an action class or background. Other methods also focused on spatio-temporal encoders [9], graph-based architecture [4], or transformer architectures [65].

Despite their impressive performance, all state-of-the-art methods rely on supervised learning, which requires a large-scale annotated dataset. However, in sports video analysis, the actions to spot may change over time, which would require re-annotating the dataset. In this work, we study how to efficiently re-annotate such datasets with active learning techniques, which aim at selecting the most informative samples for annotation, thereby minimizing the annotation effort while maintaining high task performance.

2.3. Active learning

Active Learning has been successfully applied in a wide range of applications, including image understanding [16,26], video understanding [22], natural language processing [54], speech recognition [21], and chemistry [12]. The main objective of active learning is to select the most informative unlabeled samples for annotation and use the minimal amount of label data to achieve specific performance. The main strategies for active learning include *uncertainty sampling* [30, 33, 56], *diversity maximization* [43, 62], *query-by-committee* [14, 20, 28, 45], and *expected error* [23, 31, 32, 63]. We refer to [40, 44] for a comprehensive and more generic literature review on active learning.

Uncertainty sampling. Those methods sample the unlabeled data that confuses most of the action spotting models trained thus far. Tong *et al*. [56] proposed the use of a support vector machine algorithm for conducting effective relevance feedback for image retrieval. The active learning method introduced by Joshi *et al*. [30] selects unlabeled data that the model finds hardest to classify. The selection is based on the entropy of the output of the classifier or a "Best versus Second Best" (BvSB) paradigm.

Diversity maximization. Those active learning approaches select samples that best represent the whole space of the available unlabeled set. Yang *et al*. [62] proposed a method that maximizes the diversity of the samples. They investigated this approach on diverse visual recognition tasks, including action recognition, object classification, scene recognition, and event detection. Similarly, Sener *et al*. [43] modeled the selection process as a core-set problem. They sample representative subsets of images by minimizing the L2 distance with the remaining samples in the dataset.

Query-by-committee. In this paradigm, the next batch to annotate is chosen according to the principle of maximal disagreement between a committee of student algorithms trained on the same labeled dataset. Seung *et al*. [45] introduced this approach, further analyzed in a Bayesian framework by Freund *et al*. [14]. Houlsby *et al*. [26] further investigate the relationship of Query-by-Committee with information gain theory.

Expected error. Those methods attempt to learn a metric that correlates with the error of classifying specific samples. Learning Active Learning (LAL) [32] learns to regress the error reduction for a candidate sample. The active learning scores are learned in a supervised fashion on the error loss in the training dataset. Similarly, Yoo *et al*. [63] also learn a "loss prediction module" agnostic to any task. By doing so, they actively select samples with higher predicted loss, expecting those samples to provide significant novel information to minimize for on the next train step.

Active learning for temporal video analysis. While active learning has been extensively analyzed on generic setups, only a few works apply those approaches to temporal video analysis. Brandla *et al*. [2] proposed an active learning method for temporal activity localization (TAL) algorithms, based on *uncertainty sampling* [33]. Heilbron *et al*. [22] further investigated active learning in TAL with an empirical study of different active learning paradigms, with LAL [32] performing best.

Following the previous literature, we formalize the first active learning workflow for action spotting. We analyze a couple of *uncertainty sampling* methods and set the ground for more active learning approaches for action spotting.

3. Active learning for action spotting

We propose the first active learning framework for the task of action spotting. Our framework aims at training an accurate action spotting model using a minimal amount of labeled data. Following the literature on *uncertainty sampling* active learning, we identify three key steps to achieve this objective: **(1)** Train an action spotting model on a labeled dataset that grows at each active learning step. **(2)** Select the most informative data from an unlabeled pool using an active learning algorithm. **(3)** Label the selected clips by an oracle and include the new data and annotations in the labeled set. An overview of our complete framework is depicted in Figure 2.

Formally, given a video v, the task of action spotting is to identify all action spots $\mathbf{S} = \{s_1, ..., s_M\}$ inside that video. A spot s_m comprises the action class (*e.g*. penalty, goal, etc.) and a temporal anchor. At each active learning step τ, an action spotting model, whose inference function is denoted f_τ, is trained using a set of labeled samples $\mathcal{L}_\tau$. The model details and training procedure are described in Section 3.1. Then, an active learning algorithm g selects an optimal set of unlabeled samples $\mathbf{C}^*$ from a pool $\mathcal{U}_\tau$. Several active learning algorithms are presented in Section 3.2. Subsequently, an oracle, described in Section 3.3, provides the ground-truth annotations of the actions spots (*i.e*. action

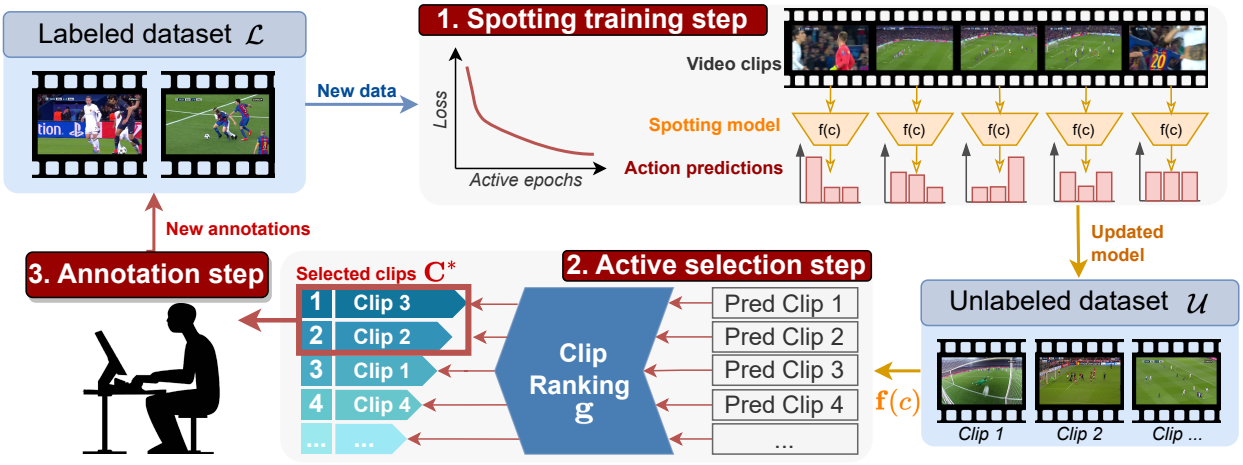

Figure 2. **Active learning pipeline for action spotting.** We start from a small labeled dataset $\mathcal{L}$ on which we train an action spotting model whose inference function is denoted f. With the trained model, we select from an unlabeled dataset $\mathcal{U}$ which sample to annotate next. For that, we first collect the prediction of the model $f(\mathcal{U})$ for each clip and pass the predictions through our selection function g that ranks the clips to select $\mathbf{C}^*$. All selected clips are then passed to the oracle (human annotator) to provide both the class and localization of all actions within that clip. These new annotated data are then added to the labeled dataset and used for the next training iteration. The process is repeated iteratively until the desired performance is reached or the unlabeled dataset is empty.

classes k and temporal anchors t) for the selected samples $\mathbf{C}^*$. The labeled set $\mathcal{L}_\tau$ is then augmented with these newly labeled clip instances $\mathbf{C}^*$ following $\mathcal{L}_{\tau+1} = \mathcal{L}_\tau \cup \mathbf{C}^*$. This process is repeated until the model reaches a desired performance or the set $\mathcal{U}_\tau$ is exhausted. Since the most expensive step consists in labeling the samples, the objective of our framework is to minimize the number of times the oracle is queried by proposing an efficient active selection algorithm.

3.1. Model training step

Datasets. The datasets for action spotting typically consist of a list of L untrimmed videos $\mathbf{V} = \{v_1, ..., v_L\}$, each video being annotated with a set $\mathbf{S}$ of action spots $s = \{k, t\}$ of class k among K classes, anchored with a single timestamp t. Since training action spotting models on long untrimmed video is not yet possible due to hardware constraints (*e.g.* GPU memory or computation time), they are typically trained on clips extracted from the video. In this work, we consider that each video v_l can be viewed as a set of N fixed-length non-overlapping clips $\mathbf{C_l} = \{c_l^1, ..., c_l^N\}$. Each clip c_l^n can be annotated with a list of temporally-anchored action spots $\mathbf{S_l^n} = \{s_{l,1}^n, ..., s_{l,M}^n\}$.

Video encoder. Typical action spotting models are composed of a video encoder $\mathbf{H}$ followed by an action spotting head $\mathbf{A}$. Given a trimmed video clips c_l^n composed of J frames, a video encoder extracts a compact features representation $\mathbf{H}(\mathbf{c_l^n}) = \{h_{l,1}^n, ..., h_{l,J}^n\}$ for each frame. This frame feature encoder is usually pre-trained on an external dataset and then either frozen or fine-tuned on the action spotting dataset. Due to the diversity of features dimension, it is common to homogenize their dimensionality with PCA to produce an even more compact and standardized frame representation. Also, these feature encoders can either be applied on the entire video clip, leveraging the temporal information, or independently on each frame of the clip, commonly requiring less computational power and memory. A typical choice for action spotting baselines is to extract frame features with a learnable CNN-based encoder such as the frame-based ResNet encoder, or the video-based I3D/C3D encoders [17].

Action spotting head. Given a set of compact frame features representation $\mathbf{H}(\mathbf{c_l^n}) = \{h_{l,1}^n, ..., h_{l,J}^n\}$, an action spotting head $\mathbf{A}$ temporally combines the descriptors and outputs a list of predicted action spots $\hat{\mathbf{S}}_\mathbf{l}^\mathbf{n} = \mathbf{A}(\mathbf{H}(\mathbf{c_l^n})) = \{\hat{s}_{l,1}^n, ..., \hat{s}_{l,M'}^n\}$ for the current clip c_l^n. This list of predictions can be obtained in two ways. A first category of action spotting models [7] directly regresses the predicted location and class. In this work, we focus on a second category, that first outputs $K+1$ class scores (including the background) per frame or per clip. The exact localizations t of the actions are then extracted using a non-maxima-suppression algorithm on the predicted class scores over time. The complete mathematical function of action spotting methods f_τ can therefore be expressed as $\mathbf{A} \circ \mathbf{H}$.

Training. We define the labeled dataset $\mathcal{L}_\tau$ at the active learning step n of size $|\mathcal{L}_\tau|$ as $\mathcal{L}_\tau = \{(c_1^{train}, \mathbf{S}_1), ..., (c_{|\mathcal{L}_\tau|}^{train}, \mathbf{S}_{|\mathcal{L}_\tau|})\}$. At each active learning step, the action spotting baseline is trained on $\mathcal{L}_\tau$. In our framework, we consider several training paradigms. The first one consists in training the action spotting module from

scratch at each active learning step. This training may be done until convergence or on a particular number of epochs. The advantage is that the deep learning model usually trains better as it. However, the drawback is that it may require a lot of time to train each epoch. A faster training paradigm consists in fine-tuning the model obtained at the previous active learning step, either until convergence or for a fixed number of epochs. This reduces the training time but does not ensure convergence. For instance, if the network diverges during the first steps due to the low amount of training data, it may be unable to recover later on. We study these training paradigms in the experimental section.

Inference. At test time, the model produces the predictions over a full video while it has been trained only on clips. One common way to solve this mismatch is to split the video into overlapping or non-overlapping clips. Each clip is processed independently and the results are aggregated along the video. The spotting performances are evaluated using the mean average precision (mAP) from successfully spotting an action within a given temporal tolerance δ. The main associated metric is the Avg-mAP, where the mAP are averaged for various values of δ-tolerance between ground truth and predicted action spots. We use the typical metrics [18] *tight* Avg-mAP (with δ ranging from 1 to 5 seconds) and *loose* Avg-mAP (with δ ranging from 5 to 60 seconds).

3.2. Active selection step

The next step consists in selecting clips from the unlabeled dataset, defined as a set of unlabeled video clips $\mathcal{U}_\tau = \{c^u_1, ..., c^u_{|\mathcal{U}_\tau|}\}$ of size $|\mathcal{U}_\tau|$. The objective of the active selection step is to create a function g that selects a new set of clips $\mathbf{C}^*$ from $\mathcal{U}_\tau$. The main challenge is to ensure that the function chooses samples that are likely to have the greatest impact on improving the action spotting model. As described in Section 2, there exists many active learning workflows. In this work, we focus on the particular case of *uncertainty sampling*. In particular, we analyze the predictions of the action spotting model trained at the previous active learning step. The predictions are a set of $(K+1)$ probability values for each class, either per frame or per clip. In the case of black-box models, the class confidence scores are the sole uncertainty information returned by a prediction. Following the literature on active learning for image classification, we construct two selectors leveraging the Uncertainty Measure (UM) and the Entropy Measure (EM). In the following, we show how to implement these methods for action spotting predictions.

Uncertainty measure. The Uncertainty Measure (UM) solely considers the confidence scores associated with each clip or frame. Given a confidence score p_k, the active learning score is inversely proportional to its distance to a confused confidence of 0.5. The Entropy Measure is formally defined as follows:

$$\text{UM} = 1 - 2 \times |p_k - 0.5| \,. \tag{1}$$

In the particular case of action spotting, this score is computed per frame and then averaged or max pooled over all frames.

Entropy measure. The Entropy Measure (EM) considers the distribution of the confidence for all the classes. Such estimation requires access to the confidence score for all classes, which might not be accessible in the case of black-box algorithms, that only returns the highest confidence for the selected class. Based on the list of confidence scores $p_1, ..., p_K$, we extract an active learning score inversely proportional to the uniformity of the distribution between the predictions. The Entropy Measure is formally defined as:

$$\text{EM} = -\sum_{i=1}^{K} p_i \log(p_i) \,. \tag{2}$$

Selecting samples. We leverage the function g to select the top-k most informative clips $\mathbf{C}^*$ with the highest active score. In this work, we study several approaches to select the number of clips $|\mathbf{C}^*|$ at each active learning step. A first approach consists in selecting a fixed number of clips at each active learning step. A second approach consists in selecting an increasing number of clips. This has the advantage of selecting only relevant clips at the beginning, when the model still requires highly informative clips.

3.3. Annotation step

Once the clips have been selected by the active learning step, they need to be annotated by an oracle (which is a human annotator in a real scenario), that will provide the ground-truth action spots. The set of clips $\mathbf{C}^*$ is manually annotated and both the clips and their corresponding annotations are added to $\mathcal{L}_\tau$. In a passive learning setup, the oracle would usually randomly select a few clips in $\mathcal{U}_\tau$, potentially annotating redundant information. In this work, we show that our active learning framework allows us to select relevant clips that increase the performance compared to a random selection, therefore saving time and money.

4. Experiments

4.1. Experimental setup

Our active learning framework is agnostic to the datasets, action spotting training parameters, and active learning selection algorithms. In this section, we provide the technical details describing our experiments in various settings.

Datasets. In this study, we leverage three datasets to evaluate our active learning framework for action spotting in football videos: SoccerNet-v2, SoccerNet-ball (public), and

Dataset	Games	Annotations	Classes	Density
SoccerNet-v2	550	110,458	17	2.23/min
SoccerNet-ball	9	11,041	2	13.62/min
FWWC19-header	52	6,527	1	1.39/min

Table 1. **Datasets.** We investigate our active learning framework on three datasets for action spotting on football videos.

FWWC19-header (private). Table 1 provides an overview of the main characteristics of each dataset.

SoccerNet-v2 consists of 550 games annotated with 110,458 action spots from 17 classes of generic actions such as goals, penalties, cards, and free-kicks. These timestamped annotations provide a comprehensive understanding of the actions that occur in football videos.

SoccerNet-ball consists of 9 public games annotated with 11,041 ball-related events such as passes and drives. This dataset provides valuable information on the actions related to the ball, which is a crucial aspect of the game. Moreover, the density of the events in the game requires precise temporal spotting capabilities.

FWWC19-header is a private dataset of 52 games from the FIFA Women World Cup 2019 (FWWC19), annotated for 5 classes of head impacts, including purposeful headers, unintentional headers, header duels, attempted headers, and other head impacts. This dataset provides insights into the events surrounding head impacts, which are a significant medical concern in football and other contact sports.

Action spotting methods. In this study, we investigate two action spotting baselines to support our active learning framework for action spotting in football videos: NetVLAD++ [19] and PTS [24]. Table 2 provides an overview of the main characteristics of each baseline.

Baseline	Encoder	AS Head	Training
NETVLAD++	ResNet152	NetVLAD	Head
PTS	ResNet18	GRU	Encoder+Head

Table 2. **Action spotting baselines for football videos.** We investigate our active learning framework on two baselines for action spotting on football videos, namely NetVLAD++ and PTS.

NetVLAD++ [19] learns to pool temporally contiguous frame features to identify which action class occurs in a clip. Since the feature encoder is frozen and the spotting head is lightweight, it is very fast to train in an active learning fashion. One major feature is that it is trained in a weakly supervised manner that does not take into account the precise localization of the action in the clip, which significantly speeds up the annotation process by the oracle. However, the drawback is that it is less precise in spotting actions.

Precise Temporal Spotting (PTS) [24] learns to combine dense frame features with a GRU, to identify if specific actions occur on specific frames. The compact encoder is trained end-to-end with the GRU, producing state-of-the-art performances on SoccerNet-v2. In this work, we select the ResNet 18 feature encoder that runs on a single GPU, as it is much faster than the RegNet encoder with GRU.

Metrics. For action spotting, we rely on the loose Avg-mAP [17], unless stated otherwise. For active learning, we analyze the learning curve of the action spotting performance as a function of the ratio of data used to train the model, *i.e.* the size of the labeled dataset. Following [22], we estimate the Area Under the Learning Curve (AULC). A good active learner is expected to have higher AULC than a random sampler. Moreover, we propose two more metrics: (i) $\mathcal{M}^{10\%}_{data}$: the Avg-mAP performance when using only 10% of the data, and (ii) $\mathcal{M}^{90\%}_{perf}$: the ratio of data required to reach 90% of the final Avg-mAP performance.

Technical details. Unless specified otherwise, We focus our experiments on the action spotting model NetVLAD++ [19] with the ResNET_PCA512 features and train the model until convergence using the same training parameters as defined in the original implementation. At each active learning iteration step, we select an amount of sample $|\mathbf{C}^*|$ equivalent of 1% of the dataset. At each action spotting training step, we restart the training from scratch.

4.2. Initial results

We first compare our framework with two active learning selection algorithms: the Uncertainty Measure (UM) and Entropy Measure (EM), against a random sampler (RS). Figure 3 shows the action spotting performances (loose Avg-mAP) as a function of the labeled dataset size. Table 3 reports the main metrics AULC, $\mathcal{M}^{\%}_{data}$ and $\mathcal{M}^{\%}_{perf}$. The initial results show that our learning framework significantly accelerates the training. With Entropy Measure (EM), the performance of the model converges at a faster pace, thus requiring less annotated data to reach higher performance. In particular, our setup with EM leads to an AULC metric of 47.96% *vs*. 45.11% with RS. Moreover, the $\mathcal{M}^{90\%}_{perf}$ of 16% for EM *vs*. 45% for RS indicates that we only need a third of the data to reach 90% of the action spotting performance. Finally, the $\mathcal{M}^{5\%}_{data}$ shows that with only 5% of the data, we reach an action spotting metric Avg-mAP of 37.24% *vs*. 32.06% when sampled randomly. Interestingly, the Uncertainty Measure (UM) provided only a limited improvement compared to Random Sampling (RS).

4.3. Accelerating the active learning framework

In this section, we share a few findings that speed up our active learning framework. In particular, **(i)** we introduce a faster scheduler for NetVLAD++ that lead to similar performance, **(ii)** we introduce an Adaptive Active Learning scheduler (AdapAL), **(iii)** we investigate a continual training that fine-tunes the model instead of training from scratch at each active learning step.

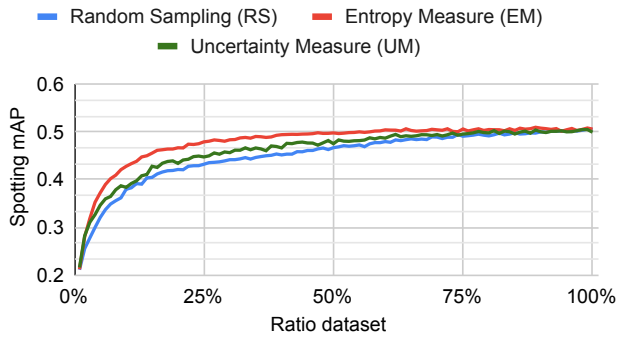

Figure 3. **Active learning *vs*. random sampling.** Our uncertainty sampling using the Entropy Measure (EM) converges to the optimal solution at a faster pace, using fewer data. In practice, active learning only needs 36% of the data needed by a random sampler to reach similar performances ($\mathcal{M}^{90\%}_{perf}$), and a similar amount of data could lead to up to 18% performance improvement ($\mathcal{M}^{4\%}_{data}$).

Method	RS	UM	EM
AULC (↑)	45.11	46.24	**47.96**
$\mathcal{M}^{90\%}_{perf}$ (↓)	45.00	31.00	**16.00**
$\mathcal{M}^{99\%}_{perf}$ (↓)	99.00	-	**64.00**
$\mathcal{M}^{5\%}_{data}$ (↑)	32.06	34.64	**37.24**
$\mathcal{M}^{10\%}_{data}$ (↑)	37.98	38.38	**42.79**

Table 3. **Active learning *vs*. random sampling.** Our proposed active learning framework based on Entropy Measure (EM) outperforms Random Sampling (RS) and Uncertainty Measure (UM).

Faster model training. First, we leverage a faster scheduler for the learning rate. Instead of starting from 10^{-3}, and reducing the learning rate on each validation loss plateau with a ratio of 10, until we reach 10^{-8}, we start with a learning rate of 10^{-2} and reduce it down to 10^{-4}. Also, we reduce the patience to identify the plateau from 10 to 5 epochs. By doing so, we practically cut the training time by two, still producing performance on par with the original training scheduler, as shown in Figure 4.

Adaptive active learning scheduler. Second, we adapt the active learning (AL) steps, gradually increasing the number of samples selected and annotated per AL step. At regime, we can increase the dataset $\mathcal{L}_t$ by more than only 1% of the dataset. In practice, we chose to increment 2% after 15% of the dataset, 5% after 25% of the dataset, and 10% after 40% of the dataset. By doing so, we reduced the AL steps from 100 to 30, saving 70% of the time. Figure 4 illustrates that the Adaptive AL step size does not impact the performance of the training of NetVLAD++ on SoccerNet-v2.

Continual training. Third, we investigate whether resuming the training from the previous active learning step would be beneficial to reduce the training time. The stopping series of triangles in Figure 5 illustrates that a naive implementation of continual training with the original parameters leads to a divergent loss that impedes any further fine-tuning. Instead, we propose to bootstrap the training with 20 epochs and fine-tune the model for 5 epochs at each active learning step, with a LR fixed at 10^{-2}. We can see that

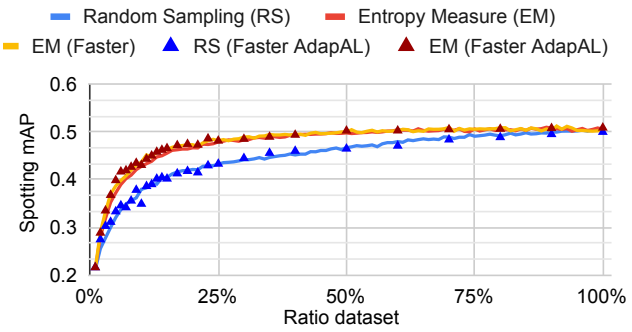

Figure 4. **Faster training and adaptive active learning (AdapAL) paradigms.** We show here that we can significantly decrease the active learning time for our experiments without reducing in any way the performance of the active learning training.

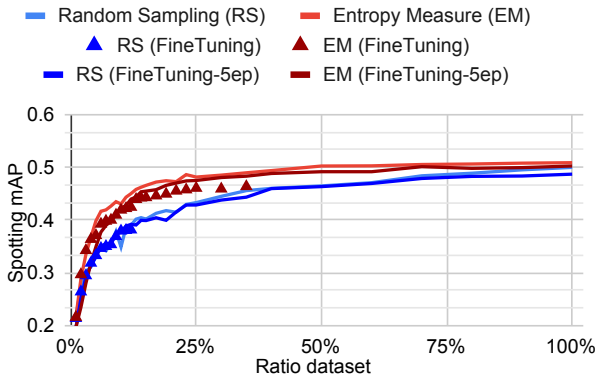

Figure 5. **Effect of fine-tuning a limited number of epochs.** Fine-tuning from a model with a limited number of epochs leads to more stability in the training for the next active learning step.

AL Setup	Train	AULC (↑)	$\mathcal{M}^{10\%}_{data}$ (↑)	$\mathcal{M}^{90\%}_{perf}$ (↓)
$RS_{1\%}$	orig.	45.11	37.98	45.00
$EM_{1\%}$	orig.	47.96	42.79	16.00
$EM_{1\%}$	fast	**48.28**	**43.68**	**14.00**
RS_{AdapAL}	fast	44.64	34.91	40.00
EM_{AdapAL}	fast	**48.01**	**43.06**	**13.00**
RS_{AdapAL}	5 ep.	44.13	37.78	40.00
EM_{AdapAL}	5 ep.	**46.94**	**42.30**	**19.00**

Table 4. **Ablation.** Our proposed active learning framework based on Entropy Measure (EM) outperforms Random Sampling (RS) on all active learning setups.

Data	Metric	RS	EM	UM
SoccerNet-ball	AULC (↑)	40.47	42.18	**42.88**
SoccerNet-ball	$\mathcal{M}_{data}^{10\%}$ (↑)	36.55	**42.41**	41.88
SoccerNet-ball	$\mathcal{M}_{perf}^{90\%}$ (↓)	23.00	**8.00**	9.00
FWWC19-header	AULC (↑)	42.67	44.28	**44.59**
FWWC19-header	$\mathcal{M}_{data}^{10\%}$ (↑)	35.18	42.32	**42.39**
FWWC19-header	$\mathcal{M}_{perf}^{90\%}$ (↓)	35.00	**12.00**	**12.00**

Table 5. **Dataset generalization.** EM and UM outperform RS on the other two datasets. With less class diversity, the gap between EM and UM is smaller.

continuing the training for 5 epoch in every active learning step still preserve the same trend of EM outperforming RS.

Table 4 summarizes how (i) training faster, (ii) adapting the active learning scheduler, and (iii) continuing the training actually performs in terms of metrics. Despite the minimal difference in the performances, the gap between RS and EM is maintained. Most importantly, those tricks lead to an order of magnitude acceleration in running the experiments.

4.4. Generalization analyses

Datasets generalization. We successively experimented our framework on two other datasets, namely SoccerNet-ball and FWWC19-header. Since SoccerNet-ball has denser actions, the hyper-parameters of NetVLAD++ were refined with a temporal window of 1s and an NMS of 1s. We chose the accelerated active learning settings, with a faster training scheduler, adaptive active learning step, and continual fine-tuning for 5 epochs per active learning step, after a bootstrap of 20 epochs. Table 5 details the results and shows, in particular, that UM and EM significantly accelerate the training efficiency for both datasets compared to Random Sampling. Interestingly, the gap between UM and EM is smaller in these two datasets than it was on SoccerNet-v2. We hypothesize that this behavior originates from the lower number of classes in SoccerNet-Headers and FWWC19-Header, respectively 2 and 1 (see Table 1). In fact, the ranking for the samples from UM and EM are actually similar in the case of a binary classifier (see Equations (1) and (2)).

Architecture generalization. We analyzed the generalization capability of our active learning framework to other action spotting methods, in particular PTS [24]. Unlike NetVLAD++, PTS produces class prediction scores per frame instead of per clip. To estimate an active learning score per clip, we aggregate the active learning score per frame with mean or max pooling. The former will consider an average uncertainty along all frames of the clip, the latter will sample clips containing single uncertain frames to train next. Similarly, we chose the same accelerated active learning settings, as PTS is way slower to train than NetVLAD++. Aligned with the findings from NetVLAD++, we show in Table 4 that UM and EM outperform RS on all three datasets. Moreover, we identify a similar trend showing that with a lower number of classes, UM tends to outperform EM. Finally, max pooling appears to work better, which means that clips with single uncertain frames are generally more informative to train the model.

Dataset	AL	AULC (↑)	$\mathcal{M}_{data}^{10\%}$ (↑)	$\mathcal{M}_{perf}^{90\%}$ (↓)
SoccerNet-v2	RS	27.76	13.59	60.00
SoccerNet-v2	mean-EM	28.53	16.14	**50.00**
SoccerNet-v2	max-EM	**28.83**	**17.62**	**50.00**
SoccerNet-v2	mean-UM	28.26	13.58	60.00
SoccerNet-v2	max-UM	28.80	15.64	**50.00**
SoccerNet-ball	RS	56.37	52.28	19.00
SoccerNet-ball	mean-EM	54.26	46.57	40.00
SoccerNet-ball	max-EM	57.64	51.79	25.00
SoccerNet-ball	mean-UM	54.92	49.53	35.00
SoccerNet-ball	max-UM	**58.48**	**53.80**	**12.00**
FWWC19-header	RS	29.77	21.77	-
FWWC19-header	mean-EM	34.81	18.45	**40.00**
FWWC19-header	max-EM	33.92	23.43	70.00
FWWC19-header	mean-UM	**35.93**	19.14	50.00
FWWC19-header	max-UM	35.26	**24.86**	**40.00**

Table 6. **Architecture generalization.** The EM and UM active selection function also outperform the RS selection when coupled with PTS [24].

5. Conclusion

In conclusion, our proposed active learning framework selects the most informative video samples to be annotated next, thus reducing the annotation effort and accelerating the training of action spotting models. We leveraged uncertainty sampling to select the most challenging video clip to train on next, which speeds up the learning process of the models. We show that our framework effectively reduces the required training data for accurate action spotting in football videos, achieving similar performance with NetVLAD++ on SoccerNet-v2 using only one-third of the dataset. This indicates significant capabilities for reducing annotation effort. Furthermore, we validated our approach on two new datasets that focus on localizing in time the actions of headers and passes. In future works, we will investigate the use of other active learning paradigms for the task of action spotting, such as diversity maximization, query-by-committee, and expected error.

Acknowledgement. This work was partly supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research through the Visual Computing Center (VCC) funding and the SDAIA-KAUST Center of Excellence in Data Science and Artificial Intelligence (SDAIA-KAUST AI). A. Cioppa is funded by the F.R.S.-FNRS. We thank Eloise Arnold, who helped design the reliability protocol for the FWWC19-header dataset.

References

[1] Adrià Arbués Sangüesa, Adrián Martín, Javier Fernández, Coloma Ballester, and Gloria Haro. Using player's body-orientation to model pass feasibility in soccer. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 3875–3884, Seattle, WA, USA, Jun. 2020. Inst. Electr. Electron. Eng. (IEEE). 2

[2] Sunil Bandla and Kristen Grauman. Active learning of an action detector from untrimmed videos. In *IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 1833–1840, Sydney, NSW, Australia, Dec. 2013. Inst. Electr. Electron. Eng. (IEEE). 3

[3] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 961–970, Boston, MA, USA, Jun. 2015. Inst. Electr. Electron. Eng. (IEEE). 1, 2

[4] Alejandro Cartas, Coloma Ballester, and Gloria Haro. A graph-based method for soccer action spotting using unsupervised player classification. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, pages 93–102, Lisbon, Port., Oct. 2022. ACM. 3

[5] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. In *Proc. SSST-8, Eighth Work. Syntax. Semant. Struct. Stat. Transl.*, pages 103–111, Doha, Qatar, 2014. Association for Computational Linguistics. 3

[6] Anthony Cioppa, Adrien Deliège, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. Scaling up SoccerNet with multi-view spatial localization and re-identification. *Sci. Data*, 9(1):1–9, Jun. 2022. 2

[7] 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/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 13123–13133, Seattle, WA, USA, Jun. 2020. Inst. Electr. Electron. Eng. (IEEE). 1, 2, 4

[8] Anthony Cioppa, Silvio Giancola, Adrien Deliege, Le Kang, Xin Zhou, Zhiyu Cheng, Bernard Ghanem, and Marc Van Droogenbroeck. SoccerNet-tracking: Multiple object tracking dataset and benchmark in soccer videos. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*, pages 3490–3501, New Orleans, LA, USA, Jun. 2022. Inst. Electr. Electron. Eng. (IEEE). 2

[9] Abdulrahman Darwish and Tallal El-Shabrway. STE: Spatio-temporal encoder for action spotting in soccer videos. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, pages 87–92, Lisbon, Port., Oct. 2022. ACM. 3

[10] Tom Decroos, Lotte Bransen, Jan Van Haaren, and Jesse Davis. Actions speak louder than goals. In *ACM SIGKDD Int. Conf. Knowl. Discov. & Data Min.*, page 1851–1861. ACM, Jul. 2019. 2

[11] Adrien Deliège, Anthony Cioppa, Silvio Giancola, Meisam J. Seikavandi, Jacob V. Dueholm, Kamal Nasrollahi, Bernard Ghanem, Thomas B. Moeslund, and Marc Van Droogenbroeck. SoccerNet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*, pages 4508–4519, Nashville, TN, USA, Jun. 2021. Best CVSports paper award. 2

[12] Vasilios Duros, Jonathan Grizou, Weimin Xuan, Zied Hosni, De-Liang Long, Haralampos N Miras, and Leroy Cronin. Human versus robots in the discovery and crystallization of gigantic polyoxometalates. *Angewandte Chemie*, 129(36):10955–10960, 2017. 3

[13] A. Ekin, A.M. Tekalp, and R. Mehrotra. Automatic soccer video analysis and summarization. *IEEE Trans. Image Process.*, 12(7):796–807, Jul. 2003. 1

[14] Yoav Freund, Sebastian H. Seung, Eli Shamir, and Naftali Tishby. Selective sampling using the query by committee algorithm. *Machine learning*, 28(2-3):133–168, Aug. 1997. 3

[15] Gregory G. Castanon, Castañón. *Exploratory search through large video corpora*. PhD thesis, Boston University, USA, 2016. 1

[16] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep Bayesian active learning with image data. In *Int. Conf. Mach. Learn. (ICML)*, pages 1183–1192. PMLR, 2017. 3

[17] Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. SoccerNet: A scalable dataset for action spotting in soccer videos. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 1792–179210, Salt Lake City, UT, USA, Jun. 2018. Inst. Electr. Electron. Eng. (IEEE). 2, 4, 6

[18] Silvio Giancola, Anthony Cioppa, Adrien Deliège, Floriane Magera, Vladimir Somers, Le Kang, Xin Zhou, Olivier Barnich, Christophe De Vleeschouwer, Alexandre Alahi, Bernard Ghanem, Marc Van Droogenbroeck, Abdulrahman Darwish, Adrien Maglo, Albert Clapés, Andreas Luyts, Andrei Boiarov, Artur Xarles, Astrid Orcesi, Avijit Shah, Baoyu Fan, Bharath Comandur, Chen Chen, Chen Zhang, Chen Zhao, Chengzhi Lin, Cheuk-Yiu Chan, Chun Chuen Hui, Dengjie Li, Fan Yang, Fan Liang, Fang Da, Feng Yan, Fufu Yu, Guanshuo Wang, H. Anthony Chan, He Zhu, Hongwei Kan, Jiaming Chu, Jianming Hu, Jianyang Gu, Jin Chen, João V. B. Soares, Jonas Theiner, Jorge De Corte, José Henrique Brito, Jun Zhang, Junjie Li, Junwei Liang, Leqi Shen, Lin Ma, Lingchi Chen, Miguel Santos Marques, Mike Azatov, Nikita Kasatkin, Ning Wang, Qiong Jia, Quoc Cuong Pham, Ralph Ewerth, Ran Song, Rengang Li, Rikke Gade, Ruben Debien, Runze Zhang, Sangrok Lee, Sergio Escalera, Shan Jiang, Shigeyuki Odashima, Shimin Chen, Shoichi Masui, Shouhong Ding, Sin-wai Chan, Siyu Chen, Tallal El-Shabrawy, Tao He, Thomas B. Moeslund, Wan-Chi Siu, Wei Zhang, Wei Li, Xiangwei Wang, Xiao Tan, Xiaochuan Li, Xiaolin Wei, Xiaoqing Ye, Xing Liu, Xinying Wang, Yandong Guo, Yaqian Zhao, Yi Yu, Yingying Li, Yue He, Yujie Zhong, Zhenhua Guo, and Zhiheng Li. SoccerNet 2022 challenges results. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, pages 75–86, Lisbon, Port., Oct. 2022. ACM. 2, 5

[19] Silvio Giancola and Bernard Ghanem. Temporally-aware feature pooling for action spotting in soccer broadcasts. In

IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR), pages 4490–4499, Nashville, TN, USA, Jun. 2021. 2, 6

[20] Ran Gilad-Bachrach, Amir Navot, and Naftali Tishby. Query by committee made real. In *NeurIPS*, volume 18, 2005. 3

[21] Dilek Hakkani-Tür, Giuseppe Riccardi, and Allen Gorin. Active learning for automatic speech recognition. In *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, volume 4. IEEE, 2002. 3

[22] Fabian Caba Heilbron, Joon-Young Lee, Hailin Jin, and Bernard Ghanem. What do I annotate next? an empirical study of active learning for action localization. In *ECCV*, volume 11215 of *Lect. Notes Comput. Sci.*, pages 212–229. Springer Int. Publ., 2018. 3, 6

[23] Steven C. Hoi, Rong Jin, Jianke Zhu, and Michael R. Lyu. Batch mode active learning and its application to medical image classification. In *Int. Conf. Mach. Learn. (ICML)*, pages 417–424, 2006. 3

[24] James Hong, Haotian Zhang, Michaël Gharbi, Matthew Fisher, and Kayvon Fatahalian. Spotting temporally precise, fine-grained events in video. *CoRR*, abs/2207.10213, 2022. 2, 6, 8

[25] Kristina Host and Marina Ivašić-Kos. An overview of human action recognition in sports based on computer vision. *Heliyon*, 8(6), Jun. 2022. 1

[26] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *CoRR*, abs/1112.5745, 2011. 3

[27] IFAB. Laws of the game. Technical report, The International Football Association Board, Zurich, Switzerland, 2022. 2

[28] Juan Eugenio Iglesias, Ender Konukoglu, Albert Montillo, Zhuowen Tu, and Antonio Criminisi. Combining generative and discriminative models for semantic segmentation of CT scans via active learning. In *Information Processing in Medical Imaging*, volume 6801 of *Lect. Notes Comput. Sci.*, pages 25–36. Springer Berlin Heidelberg, 2011. 3

[29] Yudong Jiang, Kaixu Cui, Leilei Chen, Canjin Wang, and Changliang Xu. SoccerDB: A large-scale database for comprehensive video understanding. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, page 1–8. ACM, Oct. 2020. 2

[30] Ajay J. Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. Multi-class active learning for image classification. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 2372–2379, Miami, FL, USA, Jun. 2009. Inst. Electr. Electron. Eng. (IEEE). 3

[31] Ajay J. Joshi, Fatih Porikli, and Nikolaos P. Papanikolopoulos. Scalable active learning for multiclass image classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(11):2259–2273, Nov. 2012. 3

[32] Ksenia Konyushkova, Raphael Sznitman, and Pascal Fua. Learning active learning from data. In *NeurIPS*, volume 30, 2017. 3

[33] David D. Lewis and Jason Catlett. Heterogeneous uncertainty sampling for supervised learning. *Mach. Learn. Proc.*, pages 148–156, 1994. 3

[34] Paul Liu and Jui-Hsien Wang. MonoTrack: Shuttle trajectory reconstruction from monocular badminton video. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 3512–3521, New Orleans, LA, USA, Jun. 2022. Inst. Electr. Electron. Eng. (IEEE). 2

[35] Adrien Maglo, Astrid Orcesi, and Quoc-Cuong Pham. Efficient tracking of team sport players with few game-specific annotations. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 3460–3470, New Orleans, LA, USA, Jun. 2022. Inst. Electr. Electron. Eng. (IEEE). 2

[36] Thomas B. Moeslund, Graham Thomas, and Adrian Hilton. *Computer vision in sports*. Springer, 2014. 2

[37] Naoki Nonaka, Ryo Fujihira, Monami Nishio, Hidetaka Murakami, Takuya Tajima, Mutsuo Yamada, Akira Maeda, and Jun Seita. End-to-end high-risk tackle detection system for rugby. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3550–3559, 2022. 2

[38] Luca Pappalardo, Paolo Cintia, Alessio Rossi, Emanuele Massucco, Paolo Ferragina, Dino Pedreschi, and Fosca Giannotti. A public data set of spatio-temporal match events in soccer competitions. *Sci. Data*, 6(1):1–15, Oct. 2019. 2

[39] Monirul Islam Pavel, Siok Yee Tan, and Azizi Abdullah. Vision-based autonomous vehicle systems based on deep learning: A systematic literature review. *Appl. Sci.*, 12(14):1–51, Jul. 2022. 1

[40] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B. Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM Comput. Surv.*, 54(9):1–40, Oct. 2021. 3

[41] Olav Rongved, Markus Stige, Steven Hicks, Vajira Thambawita, Cise Midoglu, Evi Zouganeli, Dag Johansen, Michael Riegler, and Pål Halvorsen. Automated event detection and classification in soccer: The potential of using multiple modalities. *Machine Learning and Knowledge Extraction*, 3(4):1–25, Dec. 2021. 2

[42] Atom Scott, Ikuma Uchida, Masaki Onishi, Yoshinari Kameda, Kazuhiro Fukui, and Keisuke Fujii. SoccerTrack: A dataset and tracking algorithm for soccer with fish-eye and drone videos. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 3568–3578, New Orleans, LA, USA, Jun. 2022. Inst. Electr. Electron. Eng. (IEEE). 2

[43] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *CoRR*, abs/1708.00489, 2017. 3

[44] Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, 2009. 3

[45] H. Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proc. Fifth Annu. Work. Comput. Learn. Theory*, page 287–294. ACM, Jul. 1992. 3

[46] João V. B. Soares and Avijit Shah. Action spotting using dense detection anchors revisited: Submission to the SoccerNet challenge 2022. *CoRR*, abs/2206.07846, 2022. 2

[47] Joao V. B. Soares, Avijit Shah, and Topojoy Biswas. Temporally precise action spotting in soccer videos using dense detection anchors. In *IEEE Int. Conf. Image Process. (ICIP)*, pages 2796–2800, Bordeaux, France, Oct. 2022. Inst. Electr. Electron. Eng. (IEEE). 2

[48] Mattia Soldan, A. Pardo, Juan Le'on Alc'azar, Fabian Caba Heilbron, Chen Zhao, Silvio Giancola, and Bernard

Ghanem. Mad: A scalable dataset for language grounding in videos from movie audio descriptions. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5016–5025, 2021. 1

[49] Vladimir Somers, Christophe De Vleeschouwer, and Alexandre Alahi. Body part-based representation learning for occluded person Re-Identification. In *IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, pages 1613–1623, Waikoloa, HI, USA, Jan. 2023. Inst. Electr. Electron. Eng. (IEEE). 2

[50] G. Sreenu and M. A. Saleem Durai. Intelligent video surveillance: a review through deep learning techniques for crowd analysis. *J. Big Data*, 6(1), Jun. 2019. 1

[51] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Gate-shift networks for video action recognition. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 1099–1108, Seattle, WA, USA, Jun. 2020. Inst. Electr. Electron. Eng. (IEEE). 3

[52] Genki Suzuki, Sho Takahashi, Takahiro Ogawa, and Miki Haseyama. Team tactics estimation in soccer videos based on a deep extreme learning machine and characteristics of the tactics. *IEEE Access*, 7:153238–153248, 2019. 2

[53] Graham Thomas, Rikke Gade, Thomas B. Moeslund, Peter Carr, and Adrian Hilton. Computer vision for sports: current applications and research topics. *Comput. Vis. Image Underst.*, 159:3–18, Jun. 2017. 1, 2

[54] Cynthia A. Thompson, Mary Elaine Califf, and Raymond J. Mooney. Active learning for natural language parsing and information extraction. In *Int. Conf. Mach. Learn. (ICML)*, pages 406–414. Citeseer, 1999. 3

[55] Matteo Tomei, Lorenzo Baraldi, Simone Calderara, Simone Bronzin, and Rita Cucchiara. RMS-net: Regression and masking for soccer event spotting. In *IEEE Int. Conf. Pattern Recognit. (ICPR)*, pages 7699–7706, Milan, Italy, Jan. 2021. Inst. Electr. Electron. Eng. (IEEE). 2

[56] Simon Tong and Edward Chang. Support vector machine active learning for image retrieval. In *Proc. Ninth ACM Int. Conf. Multimedia*, page 107–118. ACM, Oct. 2001. 3

[57] Gabriel Van Zandycke, Vladimir Somers, Maxime Istasse, Carlo Del Don, and Davide Zambrano. DeepSportradar-v1: Computer vision dataset for sports understanding with high quality annotations. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, pages 1–8, Lisbon, Port., Oct. 2022. ACM. 2

[58] Renaud Vandeghen, Anthony Cioppa, and Marc Van Droogenbroeck. Semi-supervised training to improve player and ball detection in soccer. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*, pages 3480–3489, New Orleans, LA, USA, Jun. 2022. Inst. Electr. Electron. Eng. (IEEE). 2

[59] Bastien Vanderplaetse and Stephane Dupont. Improved soccer action spotting using both audio and video streams. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*, pages 3921–3931, Seattle, WA, USA, Jun. 2020. Inst. Electr. Electron. Eng. (IEEE). 2

[60] 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. *CoRR*, abs/2004.06172, 2020. 2

[61] Kanav Vats, William McNally, Pascale Walters, David A. Clausi, and John S. Zelek. Ice hockey player identification via transformers and weakly supervised learning. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 3450–3459, New Orleans, LA, USA, Jun. 2022. Inst. Electr. Electron. Eng. (IEEE). 2

[62] Yi Yang, Zhigang Ma, Feiping Nie, Xiaojun Chang, and Alexander G. Hauptmann. Multi-class active learning by uncertainty sampling with diversity maximization. *IJCV*, 113(2):113–127, Nov. 2014. 3

[63] Donggeun Yoo and In So Kweon. Learning loss for active learning. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 93–102, Long Beach, CA, USA, Jun. 2019. Inst. Electr. Electron. Eng. (IEEE). 3

[64] Junqing Yu, Aiping Lei, Zikai Song, Tingting Wang, Hengyou Cai, and Na Feng. Comprehensive dataset of broadcast soccer videos. In *IEEE Conf. Multimedia Inf. Process. Retr. (MIPR)*, pages 418–423, Miami, FL, USA, Apr. 2018. Inst. Electr. Electron. Eng. (IEEE). 2

[65] He Zhu, Junwei Liang, Chengzhi Lin, Jun Zhang, and Jianming Hu. A transformer-based system for action spotting in soccer videos. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, pages 103–109, Lisbon, Port., Oct. 2022. ACM. 3