Camera Calibration and Player Localization in SoccerNet-v2 and Investigation of their Representations for Action Spotting

Anthony Cioppa*
University of Liège

Adrien Deliège*
University of Liège

Floriane Magera*
EVS Broadcast Equipment

Silvio Giancola*
KAUST

Olivier Barnich
EVS Broadcast Equipment

Bernard Ghanem
KAUST

Marc Van Droogenbroeck
University of Liège

Abstract

Soccer broadcast video understanding has been drawing a lot of attention in recent years within data scientists and industrial companies. This is mainly due to the lucrative potential unlocked by effective deep learning techniques developed in the field of computer vision. In this work, we focus on the topic of camera calibration and on its current limitations for the scientific community. More precisely, we tackle the absence of a large-scale calibration dataset and of a public calibration network trained on such a dataset. Specifically, we distill a powerful commercial calibration tool in a recent neural network architecture on the large-scale SoccerNet dataset, composed of untrimmed broadcast videos of 500 soccer games. We further release our distilled network, and leverage it to provide 3 ways of representing the calibration results along with player localization. Finally, we exploit those representations within the current best architecture for the action spotting task of SoccerNet-v2, and achieve new state-of-the-art performances.

1. Introduction

Soccer is often regarded as one of the most unifying activities worldwide, with thousands of professionals entertaining millions of amateurs. Such a large audience makes soccer a very lucrative business, generating billions of dollars of revenue each year from broadcast events [30]. The audiovisual data recorded during the games hides valuable insights about the players positions, the tactics, the strengths and weaknesses of each team. Hence, it is important for clubs and coaches to stay at the top of the data analytics wave, and for the fans, the data can be leveraged to provide customized services, such as personalized replays or enhanced player and game statistics. However, many general challenges of computer vision in sports have to be faced [34, 42]. Besides, the amount of data to process is so large that automated tools need to be developed. This explains the recent rise in deep learning algorithms to perform various tasks such as action spotting [7, 13, 17], player counting [9] and tracking [22, 33], ball tracking [25], tactics analysis [40], pass feasibility [2], talent scouting [12], game phase analysis [10], or highlights generation [1, 37].

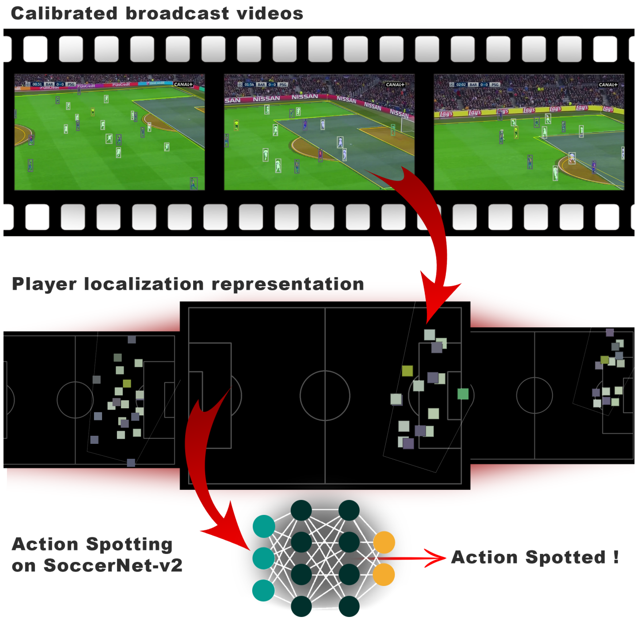

Calibrated broadcast videos
Player localization representation
Action Spotting on SoccerNet-v2
Action Spotted !

Figure 1. **Overview.** We compute and release the camera calibration parameters along player localization in real-world coordinates for the 500 soccer games of the SoccerNet dataset, we generate various types of calibration-based data representations, and we leverage them for the task of action spotting in SoccerNet-v2.

In this work, we investigate the topic of camera cali-

(*) Denotes equal contributions to this project. Contacts: anthony.cioppa@uliege.be, adrien.deliege@uliege.be, f.magera@evs.com, silvio.giancola@kaust.edu.sa. More at https://soccer-net.org/.

bration for researchers in computer vision focused on soccer. Camera calibration serves as a bridge between the images recorded and the physical world. It allows to project any point located on the field of the recorded frame to its real-world coordinates on a plane of the actual soccer field. It can thus provide knowledge about the part of the field recorded by the camera or the localization of the players on that field. One of the main commercial uses of camera calibration is the insertion of graphical elements in augmented reality. Inserting graphical elements may be used to ensure that the rules of the game are respected, such as automatic offside or goal line technologies [15]. However, most common applications aim to improve the viewer experience by fancier storytelling and with game analytics [48].

Given the value of camera calibration tools, it is not surprising that the best methods belong to private companies. This prevents scientific research on that topic to flourish at large scale. For that reason, we leverage a powerful commercial tools [15] to train a neural network on the large-scale SoccerNet dataset [17], and we release the latter to the community, along with calibration estimates for the 500 complete games available. Furthermore, we propose 3 different ways of representing the player localization in real-world coordinates obtained from the camera calibration: a top view image of the game, a feature representation, and a player graph. From an application perspective, we investigate the use of calibration-related information for the task of action spotting in SoccerNet-v2 [13]. Those contributions are illustrated in Figure 1 and further outlined below.

Contributions. We summarize our contributions as follows. **(i) Calibration for SoccerNet.** We provide calibration estimates and player localization for the 500 soccer games of the SoccerNet dataset, and we release the first calibration algorithm trained on such a large-scale soccer dataset. **(ii) Data representations.** We provide top view image-based, compressed feature-based, and player graph-based representations of the calibration data and player localization. **(iii) SOTA on action spotting in SoccerNet-v2.** As use case, we investigate the use of these representations in a state-of-the-art network for the action spotting task of SoccerNet-v2 and we improve its performances.

2. Related work

Calibration. In the context of sports events, camera calibration often benefits from the presence of a field whose layout is specified by the rules of the game. The camera may be parameterized using the full perspective projection model, but also using a homography model. Indeed, the field being most often planar, it is a convenient calibration rig to estimate the homography between the field plane and the image. Hereafter, "camera calibration" means the estimation of the intrinsic and extrinsic camera parameters.

For soccer, existing methods are assessed on the World Cup 2014 dataset [21], which introduces a metric based on the Intersection over Union (IoU) between the reference field model and its predicted deprojection from an image. This work leverages the segmentation of horizontal and vertical lines to derive a set of plausible field poses from the vanishing points, and selects the best field after a branch-and-bound optimization. However, it requires at least two of both vertical and horizontal lines to estimate the vanishing points. Some areas of the field contain few line markings, restricting the practical use of the method to goal areas.

Another common approach is to rely on a dictionary of camera views. The dictionary associates an image projection of a synthetic reference field model to a homography used to produce said projection. Each input image is first transformed to resemble a projection of the synthetic field, typically by a semantic segmentation of the field lines [5, 39] or of the areas defined by the field lines [38]. That segmentation is then associated with its closest synthetic view in the dictionary, giving a rough estimate of the camera parameters, which is eventually refined to produce the final prediction. One drawback of this kind of approach is that the processing time scales poorly with the size of the dictionary. Some applications require a large dictionary, which may become a bottleneck if real-time processing is required.

Some other calibration methods rely on tracking algorithms. Lu *et al.* [32] use an extended Kalman filter to track the pan-tilt-zoom (PTZ) camera parameters. Citraro *et al.* [11] use a particle filter to track the camera orientation and position. Due to the nature of tracking, these methods are restricted to deal with uncut, single-sequence video streams, making them inappropriate for a dataset of broadcast videos with many discontinuities, as in SoccerNet.

Kendall *et al.* [26] introduced the concept of training a neural network to directly predict the camera parameters from an image. This approach was further investigated successfully by Jiang *et al.* [24] where the predicted homography is further refined by iterative differentiation through a second neural network that predicts the error. Due to the amount of computation needed in this latter step, this method is quite slow (0.1 fps). Sha *et al.* [38] also use a neural network to refine the camera parameters found within the dictionary for the input image. They use a spatial transform network, trained to predict the homographic correction necessary to align two segmented images. In our work, we opt for the latter method because it does not involve tracking, reports a processing rate of up to 250 fps, and achieves good performances on the World Cup dataset.

Action Spotting. The task of action spotting in soccer considered in this work was introduced by Giancola *et al.* [17] along with the large-scale SoccerNet dataset. The objective is to identify at which moment various salient game actions occur, such as goals, corners, free-kicks, and more. Retrieving such information is valuable for downstream tasks such as camera selection in live game production, post-game soccer analytics, or automatic highlights generation. While detecting players on broadcast images can now be achieved with existing deep learning algorithms [8, 19], combining spatio-temporal information about their localization to infer the occurrence of game actions remains challenging as it requires a high level of cognition. Besides, in broadcast videos, several cameras are used and important actions are replayed, breaking the continuity of the stream.

In SoccerNet [17], Giancola *et al*. focus on three types of actions: goals, cards, and substitutions, which are temporally annotated with single anchors to retrieve. Several baselines are proposed, all of which rely either on ResNet [20], I3D [4], or C3D [44] frame features computed at 2 frames per second followed by temporal pooling methods (NetVLAD and MaxPool), with the ResNet features yielding the best results. Several works followed, building on the same set of pre-computed ResNet features. Cioppa *et al*. [7] develop a particular loss function that takes into account the context surrounding the actions in the temporal domain. They use it to perform a temporal segmentation of the videos before using a spotting module, achieving state-of-the-art results. Similarly, Vats *et al*. [47] handle the temporal information around the actions with a multi-tower CNN that takes into account the noise due to the single anchor annotation scheme. Tomei *et al*. [43] randomly mask a portion of the frames before the actions to force their network to focus on the following frames, as those may contain the most discriminative features to spot actions. By further fine-tuning the last block of the ResNet backbone, they achieve a strong state-of-the-art results on SoccerNet-v1. Rongved *et al*. [35] directly learn a whole 3D ResNet applied to the video frames on 5-seconds clips. This turns out to be an ambitious approach with moderate results, given the huge volume of data to process from scratch. Vanderplaetse *et al*. [46] propose a multimodal approach by including audio features, first extracted with a pre-trained VGGish network, then averaged over 0.5 seconds windows and synced with the 2 fps original ResNet features. They are processed in parallel streams before undergoing a late fusion, yielding the best results in several action classes.

Besides those works, the literature is rich in papers using either small custom datasets, such as [16, 23], or focusing on event recognition from pre-cut clips and selected frames rather than spotting actions in untrimmed videos, such as [27, 28, 29], or even targeting a single class, such as goals [45]. In this work, we tackle the large-scale action spotting task of SoccerNet-v2, the extension of SoccerNet proposed by Deliège *et al*. [13]. It covers 17 classes of actions, annotated for the 500 untrimmed SoccerNet games, and constitutes the most appropriate public benchmark for research on action spotting in soccer.

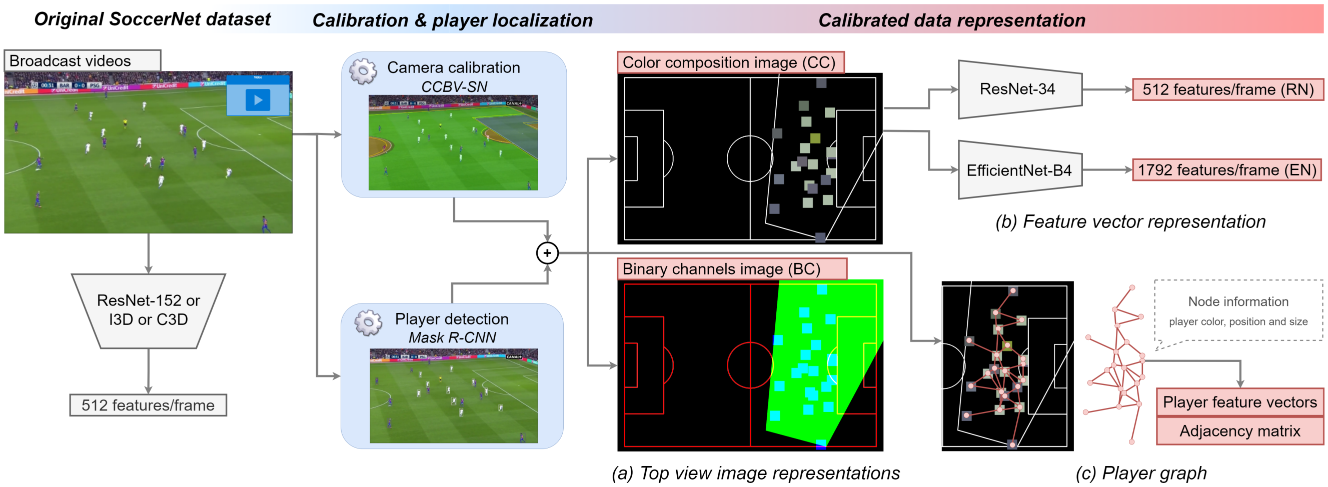
3. Calibration and player localization

Contribution. In SoccerNet [17], the frames of the raw videos are subsampled at 2 fps, then transformed into feature vectors, by passing through a ResNet-152 [20], I3D [4], or C3D [44] network pre-trained on ImageNet [14], all of which are released with the dataset. Hence, those vectors only encode generic information about the frames. As first contribution, shown in Figure 2, we enrich the SoccerNet dataset with actionable camera calibration estimates, along with players and referee localization. Such information provides a soccer-specific insight and is explicitly linked with the game in real-world coordinates. Besides releasing the largest set of calibration estimates to date, we are also the first to deliver a calibration algorithm trained on a large scale dataset such as SoccerNet. For synchronization purposes, we compute the calibration, player and referee localization for the 2-fps-subsampled set of frames considered in SoccerNet. In the following, we make no difference anymore between players and referees, all of which are called "players", and we call "per-frame information" any information computed for each of those subsampled frames.

Calibration algorithm. We base our calibration on the Camera Calibration for Broadcast Videos (CCBV) of Sha *et al*. [38], but we write our own implementation, given the absence of usable public code. They use as calibration parameterization the homography between the field plane and the image, which is valid under the assumption of a planar field [18]. First, we describe their original algorithm, then we give the details of our changes.

The algorithm relies on a dictionary, *i.e*. a set of pairs of artificial field zone segmentations, called "templates", and homographies. The dictionary is built in a pre-processing step, according to the camera parameters distribution over the training dataset. Since this distribution is unknown, it is estimated with a clustering algorithm based on Gaussian Mixture Models, that also determines the number of modes necessary to fit the distribution. The mean of each mode corresponds to a homography of the dictionary, that defines a camera perspective from which its corresponding template is generated as an artificial semantic image of the field.

CCBV itself consists of three steps, each performed by a specific neural network. First, a zone segmentation of the field is computed with a U-Net architecture [36], where a zone is a field area enclosed by field lines. Second, a rough estimate of the homography between the field plane and the image is obtained. A siamese network [3, 6] encodes the zone segmentation and the templates of the dictionary in feature vectors. The homography associated with the template encoding with the shortest L^2 distance to the zone segmentation encoding is the rough estimate of the sought homography. Third, this template homography is refined, in two steps. A Spatial Transform Network first regresses the



(b) Feature vector representation

(a) Top view image representations

(c) Player graph

Figure 2. **Calibration and player localization representations.** The original SoccerNet dataset (left) provides raw videos of 500 complete soccer games as well as generic per-frame feature vectors. We distill a commercial calibration tool into a recent network architecture on SoccerNet, which we release along all the calibrations. We combine Mask R-CNN player detections with the calibration to provide 3 representations of the calibrated data, thus enriching the dataset with specific player-based information: (a) top view representations, (b) feature vectors representations, (c) a player graph representation. The red boxes, also released, further serve as inputs in neural networks to investigate the usefulness of calibration for the task of action spotting in SoccerNet-v2, leading to a new state-of-the-art performance.

homography between the zone segmentation and the template. Then, the final homography prediction is obtained by multiplying the regressed homography with the template homography, giving the estimated calibration parameters.

Our training process. Given the absence of a large-scale corpus of ground-truth camera calibrations in the literature, we opt for a student-teacher distillation approach. We consider a commercial tool [15] as teacher to generate a dataset of 12,000 pseudo-ground-truth calibrations on the SoccerNet dataset, which we use to train our student calibration algorithm. Our training dataset is 60x larger than the World Cup 2014 dataset [21] used in [38] and contains a larger variety of camera viewpoints, making our student calibration network a valuable candidate for universal camera calibration in the context of soccer. In fact, during the creation of the dictionary, more than 1000 modes are found by the clustering algorithm. Besides, during the training phase of the Spatial Transform Network, we notice vanishing gradient issues. To overcome this problem, we first pre-train it with a MSE loss and use leaky ReLU activations instead of ReLUs. After convergence, we compute the calibration estimates of the SoccerNet video frames with our trained calibration network. A binary score about the relevance of the calibration, set to 1 for frames with a plausible estimated calibration, is also computed by our student. This allows to discard cameras views that are not recorded by the main camera, such as close-up views, or public views. We release those estimates along our trained calibration network, which can be used with a wide variety of soccer videos. We denote CCBV-SN our student trained on SoccerNet.

Player localization. For each calibrated frame, we use Mask R-CNN [19] to obtain a bounding box and a segmentation mask per detected person instance. Then, we compute a field mask following [10] to filter out the bounding boxes that do not intersect the field, thus removing *e.g.* staff and detections in the public. We use the homography computed by CCBV-SN to estimate the player localization on the field in real-world coordinates from the middle point of the bottom of their bounding box. Finally, we also store the average RGB color of each segmented player to keep track of a color-based information per person. As for the calibrations, we release this raw player-related information.

4. Calibrated data representation

Contribution. The calibration estimates and player localization are not easy to handle efficiently. Hence, to encourage their use in subsequent soccer-related works, we propose and release various easy-to-use representations of the calibration data extracted from the previous section. We illustrate these representations in Figure 2 and describe them in this section. We also discuss their pros and cons.

4.1. Top view image representations

In this section, we provide image representations of the player localization information. We use the calibration of CCBV-SN to generate a synthetic top view of the game containing generic field lines, the players represented by small squares, and the polygon delimiting the portion of the field seen by the camera. We represent that top view in two ways.

Color composition (CC). We generate a RGB image where we first set field pixels in black and line pixels in white. Then, we superimpose with white pixels the contour of the polygon of the field seen by the camera. Finally, we represent the players by squares filled with their associated RGB color, overriding previous pixels in case of intersection.

Binary channels (BC). We generate an "image" composed of 3 binary channels: one for the generic field lines, one for the filled polygon of the field seen by the camera, and one for the players without their color information.

Pros and cons. A major advantage of image representations is their interpretability, since a human observer can directly understand relevant information about the game from such top views. Besides, they can be easily processed with convolutional neural networks in deep learning pipelines. As a drawback, they have a relatively large memory footprint compared with the low amount of actionable information that they actually contain. The color composition view has the advantage over the binary channels of keeping track of the color of the players, necessary for team discrimination and tactics analysis. On the other hand, the representation of a player in the binary channels is not influenced by a poor segmentation in the raw image or a color shift due to *e.g.* an occlusion. Also, players located on field lines do not prevent those lines to be encoded properly in their binary channel, while they hide the lines in the color composition.

4.2. Feature vector representation

Inspired by Giancola *et al.* [17], we compress our top views as frame feature vectors extracted by pre-trained backbones. This is common practice in deep learning approaches, as universal networks trained on *e.g.* ImageNet have an excellent transfer capability to encode meaningful visual information about any given image. We use top views of 224 × 224 pixels, with field lines of 4 pixels width and players of 8 × 8 pixels. We consider two backbones with similar number of parameters, both trained on ImageNet.

ResNet-34 (RN). This network has 21.8 million parameters and achieves 73.27% top-1 accuracy on ImageNet. We use a frozen ResNet-34 [20] and collect the feature vectors of dimension 512 in its penultimate layer.

EfficientNet-B4 (EN). This more recent network has 19 million parameters and achieves 82.6% top-1 accuracy on ImageNet. We use EfficientNet-B4 [41], which yields feature vectors of dimension 1792 in its penultimate layer.

Pros and cons. We choose these networks for their good trade-off between performance on ImageNet and inference time. Indeed, they allow for a much faster training of neural networks compared with the top views, as computationally expensive spatial convolutions have already been performed. As a drawback, the features collected from these networks are not interpretable anymore, which may reduce the possibilities of developing explainable models.

4.3. Player graph representation

Player graph (PG). Our third approach consists in encoding per-frame players information in a graph. Each player is represented with a node, whose features are defined by their associated RGB color, their position in real-world coordinates, and the area of the detected bounding box in the image frame. Two players are linked to each other with an edge if their real-world distance is below 25 meters, which we consider sufficient to pass contextual information between the nodes in the graph (*i.e.* the players in the field).

Pros and cons. The player graph is a compromise between the compactness of feature representations and the interpretability of top views. Indeed, it explicitly encodes in a compact way the interpretable information that we want to embed in our descriptive features: the players color, their position in the field and their interactions with each other. Contrary to top view images, it does not encode any empty portion of the field, nor considers the field lines that are constant across the videos, which makes the learning focusing more on the interesting player features. The graph convolutional network (see next section) that processes the player graph aggregates features from neighboring players, which helps it understand real-world distances by discarding players further away. Yet, that aggregation does not consider different clusters of neighbors, which could lead to a misunderstanding between teammates and adversaries.

5. Experiments

Contribution. In this section, we first validate with performance metrics the effectiveness of CCBV-SN as calibration algorithm. Then, we leverage our various calibration data representations in the particular use case of the action spotting task in SoccerNet-v2. We build on top of the current best network to achieve a new state-of-the-art performance.

5.1. Validating the camera calibration distillation

In order to validate our calibration-based data representations and their use for an action spotting task, we first validate CCBV-SN as camera calibration algorithm.

Dataset. The World Cup 2014 dataset [21] stands as reference for evaluating soccer camera calibration methods. The test set comprises 186 calibrated images taken from 10 different games in various stadiums, perspectives, lighting conditions, and moments of the day.

Metric. Following [5, 38], for each test image, we compute the entire intersection over union (*IoU entire*) between the top view projections of the field model by the ground-truth camera and by the estimated camera, as well as the IoU restricted to the part of the field actually shown on the image (*IoU part*). For both metrics, we report the mean and the median value across the test images.

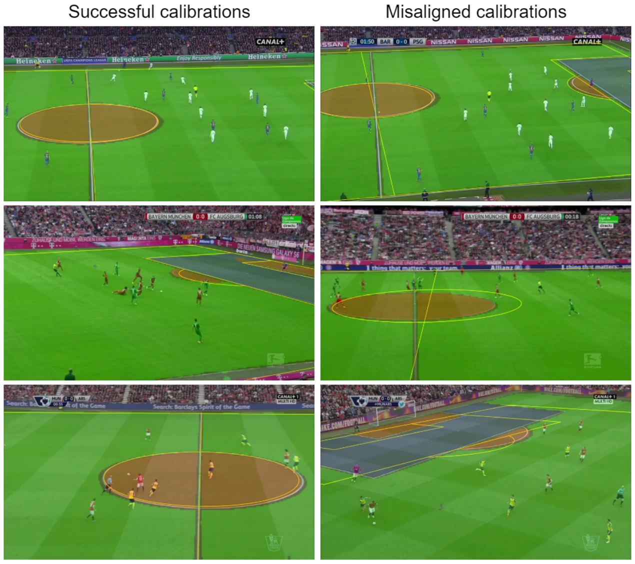


Figure 3. **Examples** of calibrations obtained with CCBV-SN. Globally, the results are satisfying and allow for an effective use of the calibration for downstream tasks such as action spotting.

Results. We report the calibration performances in Table 1. The private teacher achieves the best results on 2 out of 4 metrics, which validates its use as a teacher in our distillation approach. It is topped by Citraro *et al.* [11] on the IoU (entire) metrics, which finetune their method with additional manual annotations on the dataset. In comparison, none of our methods are trained on the that evaluation dataset. Thus we actually measure the generalization capabilities of our teacher and CCBV-SN on a completely new dataset. This evaluation also allows us to quantify the performance drop induced by our distillation procedure. CCBV-SN loses 6 to 12 points in the distillation process, making its performances close to [39], especially on the IoU (part). This metric is actually the most relevant for us, as our use of the calibration is limited to the visible part of the field for the calibration data representations. Therefore, CCBV-SN is legitimately usable in the rest of our experiments, and is presumably even better on SoccerNet, since it is the dataset on which it has been trained. Some calibration results obtained with CCBV-SN are shown in Figure 3.

5.2. Use case: calibration-aware action spotting

In this section, we investigate a possible use case of our calibration representations, by leveraging a state-of-the-art network for the task of action spotting in SoccerNet-v2.

Dataset. The action spotting dataset of SoccerNet-v2 [13] consists in 110,458 action timestamps spread over 17 classes within the 500 complete games of the SoccerNet [17] dataset, with 22,551 actions related to the 100 test games. Each action is annotated with a single timestamp, that must be retrieved as precisely as possible.

Table 1. **Calibration results** on the World Cup 2014 dataset [21]. The teacher tool outperforms the other methods on the IoU (part) metric. Our publicly released student network CCBV-SN, obtained by distilling the teacher in the CCBV architecture on SoccerNet, achieves acceptable transfer learning results.

	IoU (entire)		IoU (part)	
Method	Mean	Med.	Mean	Med.
DSM [21]	83	-	-	-
Sharma *et al.* [39]	-	-	91.4	92.7
Chen *et al.* [5]	89.4	93.8	94.5	96.1
Sha *et al.* [38]- CCBV	88.3	92.1	93.2	96.1
Jiang *et al.* [24]	89.8	92.9	95.1	96.7
Citraro *et al.* [11] $^+$	**93.9**	**95.5**	-	-
Teacher [15] *	91.7	93	**96.7**	**98.7**
Our CCBV-SN *	79.8	81.7	88.5	92.3

* no finetuning on WC14 $^+$ used extra annotations on WC14

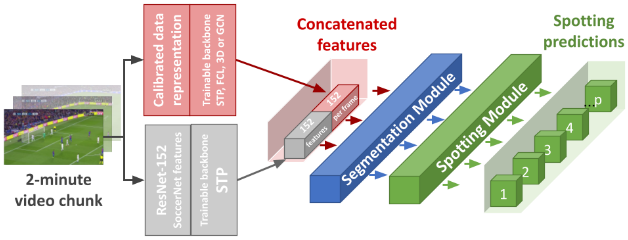


Figure 4. **Our action spotting pipeline** for the patterned actions. We include calibration information within CALF [7], by concatenating frame feature vectors extracted from our various representations. This allows us to mix generic information from the SoccerNet features with player-specific information from the calibration. For each chunk, the network outputs p spotting predictions.

CALF architecture. We focus on integrating the calibration information along the original SoccerNet features in the Context-Aware Loss Function (CALF) architecture of Cioppa *et al.* in [7]. This architecture achieves state-of-the-art performances on the task of action spotting in SoccerNet-v2. As original features, we choose the ResNet features further reduced from 2048 to 512 components by PCA, as they yield the best results both in [7] and in [17], which we also noticed in our preliminary experiments.

CALF is composed of three trainable modules: a frame feature extractor, a temporal segmentation module, and an action spotting module. The first one is a convolutional spatio-temporal pyramid (STP) that aggregates the ResNet features across various time scales, and outputs a feature vector of user-defined dimension d per frame. Our goal is to concatenate such features judiciously along frame feature vectors extracted from our calibration representations, as shown in Figure 4. The remaining two modules and the training protocol are kept as is to assess the improvement brought by only the calibration information.

Processing our representations. Each calibration data representation must be processed appropriately for a seamless integration within the network. We proceed as follows.

Top views. We process our top views with our own 3D-convolutional network **(3D)**. We choose the same structure as the STP module but where the kernels are extended to take into account the extra spatial dimension of the top view compared to the original ResNet features. The output is a d-dimensional vector for each frame that gathers the spatial and temporal information of the top view representation.

Feature vectors. We investigate two ways of further processing the feature vectors obtained from the pre-trained backbones: (1) we use the trainable STP of CALF to extract d-dimensional frame feature vectors **(STP)**, (2) we fully connect our feature vectors through a trainable layer directly to feature vectors of dimension d followed by a ReLU activation **(FCL)**. In the second case, we obtain per-frame feature vectors solely based on the raw frame information, without any temporal aggregation.

Player graph. We design a graph convolutional neural network **(GCN)** to extract per-frame features from the player graph. For that purpose, we follow DeeperGCN [31]. In particular, we build our architecture with 14 GCN blocks with residual skip connections. We leverage two layers of GENeralized Graph Convolution (GENConv) per block, that aggregate the lifted neighboring nodes using a softmax with a learnable temperature parameter. Then, a max operation across the node pools a global feature for the player graph. This feature is later lifted with a single fully connected layer to the desired dimension d.

Class separation. Intuitively, the player localization extracted with the calibration can prove more helpful for spotting some classes (*e.g*. penalty) than others (*e.g*. shot off target). Hence, we leverage our domain knowledge to split the 17 action classes of SoccerNet-v2 into two sets: "patterned" and "fuzzy" actions. We consider an action as "patterned" when its occurrence is systematically linked with typical player placements: penalty, kick-off, throw-in (one player outside the field), direct free-kick (player wall), corner, yellow card, red card, yellow then red card (players grouped around the referee for the card-related actions). On the other hand, a "fuzzy" action may occur in many different player configurations: goal, substitution, offside, shot on target, shot off target, clearance, ball out of play, foul, indirect free-kick. Given our class separation, we train two networks: one on the patterned classes that uses the calibration information and the original ResNet features, one on the fuzzy classes that only uses those ResNet features.

Feature fusion. For the network trained on the patterned classes, we input SoccerNet's ResNet features to the STM, collect d-dimensional feature vectors, and concatenate them with our d-dimensional vectors extracted by one of the above processing steps. This is illustrated in Figure 4. We set d =152, which allows us to simply plug a calibration-related branch next to the original branch of CALF working on SoccerNet's ResNet features. The concatenation yields feature vectors of dimension 304 and is performed just before the temporal segmentation module of the whole network. For the network trained on the fuzzy classes, we use SoccerNet's ResNet features only as in CALF, and set $d = 304$ after the STM to have the same input dimension for the segmentation modules of the two networks.

Training. Following CALF, we process 2-minute video chunks. We extract frame feature vectors as described above, concatenate them when necessary, and input them to a temporal segmentation module, that provides per-class features and per-class actionness scores per frame. This module is trained with a context-aware loss that aggregates the temporal context around the actions. Those features and scores are concatenated and sent to an action spotting module, which provides predicted action vectors for the chunk, containing a localization estimate and a classification per predicted action. An iterative one-to-one matching connects those predictions with ground-truth action vectors, allowing to train the module with an element-wise MSE loss.

Metric. As defined in [17], we measure the action spotting performance with the Average-mAP. First, predicted action spots are said positive when they fall within a margin δ of a ground-truth timestamp from their predicted class. Then, the Average Precision (AP) is computed from Precision-Recall curves, then averaged over the classes (mAP). Finally, the Average-mAP is the AUC of the mAP obtained at margins δ varying from 5 to 60 seconds. Given our class separation, we merge the predictions of our two networks before computing the Average-mAP.

Results. We achieve our best result with the color composition reduced to frame features by ResNet-34 as calibration data representation, further bridged to d-dimensional feature vectors with a fully connected layer. This yields an Average-mAP of 46.8% on the test set, reported in Table 2, the current SoccerNet-v2 action spotting leaderboard. We achieve a novel state-of-the-art performance, outperforming the other methods by a comfortable margin. In particular, we prevail on 15 of the 17 classes, only topped by Vanderplaetse *et al*. [46] for kick-offs and penalties. Besides, kick-offs are the only actions for which our performances degrade compared to the original network, most probably because those actions are regularly unshown in soccer broadcasts [13]. We illustrate some action spotting results in Figure 5. We manage to spot actions that CALF misses, and some false positives of CALF are correctly avoided. On the current open competition of action spotting in SoccerNet-v2, organized on EvalAI, we achieve an Average-mAP of 46.4% on the private challenge dataset. This validates the generalization capabilities of our network.

For completeness, we give additional results with the dif-

Table 2. **Leaderboard for action spotting** (Average-mAP %) on SoccerNet-v2. Patterned actions are indicated with a * . We report the results of our best method, based on [7], which outperform the other techniques on almost all the classes.

	SN-v2	Ball out	Throw-in *	Foul	Ind. f.-kick	Clearance	Shot on tar.	Shot off tar.	Corner *	Substitution	Kick-off *	Yel. card *	Offside	Dir. f.-kick *	Goal	Penalty *	Yel.→Red *	Red card *
Counts (test set)	22551	6460	3809	2414	2283	1631	1175	1058	999	579	514	431	416	382	337	41	14	8
MaxPool [17]	18.6	38.7	34.7	26.8	17.9	14.9	14.0	13.1	26.5	40.0	30.3	11.8	2.6	13.5	24.2	6.2	0.0	0.9
NetVLAD [17]	31.4	47.4	42.4	32.0	16.7	32.7	21.3	19.7	55.1	51.7	45.7	33.2	14.6	33.6	54.9	32.3	0.0	0.0
AudioVid [46]	39.9	54.3	50.0	55.5	22.7	46.7	26.5	21.4	66.0	54.0	**52.9**	35.2	24.3	46.7	69.7	**52.1**	0.0	0.0
CALF [7]	40.7	63.9	56.4	53.0	41.5	51.6	26.6	27.3	71.8	47.3	37.2	41.7	25.7	43.5	72.2	30.6	0.7	0.7
Ours (CC + RN + FCL)	**46.8**	**68.7**	**59.9**	**56.2**	**45.5**	**55.4**	**32.5**	**33.0**	**78.7**	**60.4**	34.8	**50.4**	**33.6**	**48.6**	**76.2**	50.5	**3.1**	**8.5**

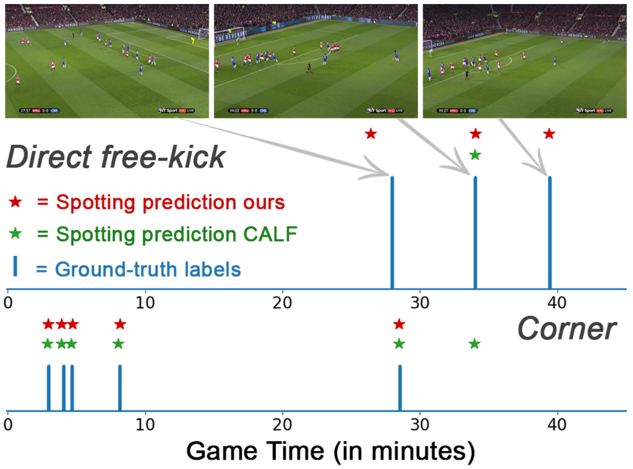


Figure 5. **Examples** of action spotting results on a game between Manchester United and Chelsea in December 2015. In this case, we spot correctly two more direct free-kicks than the original network, and we rightly avoid predicting a corner around 35 minutes.

ferent combinations of calibration data representation and feature extraction in Table 3. We see that the color composition with the 3D network and the player graph representation yield performances that are practically equivalent to our best result, while other variants are less effective. Hence, each calibration data representation is able to reach competitive performances. We do not report any result with extracted feature representations from top view images composed of binary channels as they globally yield much lower performances. Finally, fusing features from the top view and the player graph does not appear useful either as these contain essentially the same type of information.

6. Conclusion

In this paper, we examine the problem of computing, representing, and exploiting the camera calibration information for the large-scale SoccerNet dataset, composed of 500 soccer games. We leverage a powerful commercial tool to generate pseudo ground truths and manage to distill it into a recent deep learning algorithm. As first contribution, we release our distilled network, which is the first public soccer calibration algorithm trained on such a large dataset, along with its calibration estimates for the SoccerNet videos to enrich the dataset. We use our calibration and a player detection algorithm to obtain the player localization in real-world coordinates. To further serve the scientific community, our second contribution is to provide three actionable ways of representing those calibration data: top view images, feature vectors representations, and player graphs. Eventually, we investigate the benefit of using these representations in a deep learning network for the task of action spotting in SoccerNet-v2. Standing for our third contribution, we design an appropriate concatenation of generic video and specific calibration information within the current best network to achieve a novel state-of-the-art performance.

Table 3. **Action spotting results** (Average-mAP %) obtained with our various data representations, feature vectors, and networks.

Data repres.	Features	Network	Av.-mAP
Binary channels	-	3D	44.7
Color compos.	-	3D	46.7
Color compos.	ResNet-34	STP	43.5
Color compos.	ResNet-34	FCL	**46.8**
Color compos.	Effic.Net-B4	STP	42.5
Color compos.	Effic.Net-B4	FCL	45.5
Player graph	-	GCN	46.7

Acknowledgments. This work is supported by the DeepSport project of the Walloon Region and the FRIA, EVS Broadcast Equipment, and KAUST Office of Sponsored Research (OSR) under Award No. OSR-CRG2017-3405.

References

[1] Rockson Agyeman, Rafiq Muhammad, and Gyu Sang Choi. Soccer video summarization using deep learning. In *IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 270–273, 2019. 1

[2] Adrià Arbués Sangüesa, Adriàn Martín, Javier Fernández, Coloma Ballester, and Gloria Haro. Using player's body-orientation to model pass feasibility in soccer. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 3875–3884, 2020. 1

[3] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a "siamese" time delay neural network. In *Adv. Neural Inform. Process. Syst.*, pages 737–744, 1993. 3

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4724–4733, 2017. 3

[5] Jianhui Chen and James J. Little. Sports camera calibration via synthetic data. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 2497–2504, 2019. 2, 5, 6

[6] Davide Chicco. Siamese neural networks: An overview. *Artificial Neural Networks*, 2190:73–94, 2021. 3

[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 Conf. Comput. Vis. Pattern Recog.*, pages 13126–13136, 2020. 1, 3, 6, 8

[8] Anthony Cioppa, Adrien Deliège, Maxime Istasse, Christophe De Vleeschouwer, and Marc Van Droogenbroeck. ARTHuS: Adaptive Real-Time Human Segmentation in Sports Through Online Distillation. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 2505–2514, 2019. 3

[9] Anthony Cioppa, Adrien Deliège, Noor Ul Huda, Rikke Gade, Marc Van Droogenbroeck, and Thomas B. Moeslund. Multimodal and multiview distillation for real-time player detection on a football field. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 3846–3855, 2020. 1

[10] Anthony Cioppa, Adrien Deliège, and Marc Van Droogenbroeck. A Bottom-Up Approach Based on Semantics for the Interpretation of the Main Camera Stream in Soccer Games. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 2018. 1, 4

[11] Leonardo Citraro, Pablo Márquez-Neila, Stefano Savare, Vivek Jayaram, Charles Dubout, Félix Renaut, Andres Hasfura, Horesh Ben Shitrit, and Pascal Fua. Real-time camera pose estimation for sports fields. *Machine Vision and Applications*, pages 1–13, 2020. 2, 6

[12] Tom Decroos, Lotte Bransen, Jan Van Haaren, and Jesse Davis. Actions speak louder than goals: Valuing player actions in soccer. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, page 1851–1861, 2019. 1

[13] 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. *CoRR*, 2020. 1, 2, 3, 6, 7

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 248–255, 2009. 3

[15] EVS Broadcast Equipment. Xeebra product of EVS Broadcast Equipment. `https://evs.com/en/product/xeebra`. 2, 4, 6

[16] Babak Fakhar, Hamidreza Rashidy Kanan, and Alireza Behrad. Event detection in soccer videos using unsupervised learning of spatio-temporal features based on pooled spatial pyramid model. *Multimedia Tools and Applications*, 78(12):16995–17025, 2019. 3

[17] Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. SoccerNet: A Scalable Dataset for Action Spotting in Soccer Videos. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 1711–1721, 2018. 1, 2, 3, 5, 6, 7, 8

[18] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004. 3

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Int. Conf. Comput. Vis.*, pages 2980–2988, 2017. 3, 4

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016. 3, 5

[21] Namdar Homayounfar, Sanja Fidler, and Raquel Urtasun. Sports field localization via deep structured models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4012–4020, 2017. 2, 4, 5, 6

[22] Samuel Hurault, Coloma Ballester, and Gloria Haro. Self-supervised small soccer player detection and tracking. In *The 3rd International Workshop on Multimedia Content Analysis in Sports (MMSports)*, page 9–18, 2020. 1

[23] Haohao Jiang, Yao Lu, and Jing Xue. Automatic soccer video event detection based on a deep neural network combined cnn and rnn. In *IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 490–494, 2016. 3

[24] Wei Jiang, Juan Camilo Gamboa Higuera, Baptiste Angles, Weiwei Sun, Mehrsan Javan, and Kwang Moo Yi. Optimizing through learned errors for accurate sports field registration. In *IEEE Winter Conference on Applications of Computer Vision*, 2020. 2, 6

[25] Paresh R. Kamble, Avinash G. Keskar, and Kishor M. Bhurchandi. A deep learning ball tracking system in soccer videos. *Opto-Electronics Review*, 27(1):58–69, 2019. 1

[26] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Int. Conf. Comput. Vis.*, 2015. 2

[27] Abdullah Khan, Beatrice Lazzerini, Gaetano Calabrese, and Luciano Seraf. Soccer event detection. In *International Conference on Image Processing and Pattern Recognition (IPPR)*, 2018. 3

[28] Muhammad Zeeshan Khan, Summra Saleem, Muhammad A. Hassan, and Muhammad Usman Ghanni Khan. Learning deep C3D features for soccer video event detection. In *International Conference on Emerging Technologies (ICET)*, pages 1–6, 2018. 3

[29] Victor Khaustov and Maxim Mozgovoy. Recognizing events in spatiotemporal soccer data. *Applied Sciences*, 10(22), 2020. 3

[30] David Lange. Market size of the european professional football market from 2006/07 to 2018/19. https://www.statista.com/statistics/261223/european-soccer-market-total-revenue/. Accessed: March 3, 2021. 1

[31] Guohao Li, Chenxin Xiong, Ali Thabet, and Bernard Ghanem. Deepergcn: All you need to train deeper gcns. *CoRR*, 2020. 7

[32] Jikai Lu, Jianhui Chen, and James J. Little. Pan-tilt-zoom SLAM for sports videos. *Brit. Mach. Vis. Conf.*, 2019. 2

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

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

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

[36] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pages 234–241, 2015. 3

[37] Melissa Sanabria, Sherly, Frédéric Precioso, and Thomas Menguy. A deep architecture for multimodal summarization of soccer games. In *ACM Int. Conf. Multimedia Worksh.*, page 16–24, 2019. 1

[38] Long Sha, Jennifer Hobbs, Panna Felsen, Xinyu Wei, Patrick Lucey, and Sujoy Ganguly. End-to-end camera calibration for broadcast videos. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 2, 3, 4, 5, 6

[39] Rahul Anand Sharma, Bharath Bhat, Vineet Gandhi, and C. V. Jawahar. Automated top view registration of broadcast football videos. In *IEEE Winter Conference on Applications of Computer Vision*, pages 305–313, 2018. 2, 6

[40] 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. 1

[41] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Int. Conf. Machine Learning*, 2019. 5

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

[43] Matteo Tomei, Lorenzo Baraldi, Simone Calderara, Simone Bronzin, and Rita Cucchiara. Rms-net: Regression and masking for soccer event spotting. In *Int. Conf. Pattern Recog.*, 2020. 3

[44] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Int. Conf. Comput. Vis.*, pages 4489–4497, 2015. 3

[45] Grigorios Tsagkatakis, Mustafa Jaber, and Panagiotis Tsakalides. Goal!! event detection in sports video. *Electronic Imaging*, 2017:15–20, 2017. 3

[46] Bastien Vanderplaetse and Stéphane Dupont. Improved soccer action spotting using both audio and video streams. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 3921–3931, 2020. 3, 7, 8

[47] Kanav Vats, Mehrnaz Fani, Pascale Walters, David A Clausi, and John Zelek. Event detection in coarsely annotated sports videos via parallel multi-receptive field 1d convolutions. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 882–883, 2020. 3

[48] Vizrt. Viz libero product of Vizrt. https://www.vizrt.com/en/products/viz-libero. 2