

Who Did It? Identifying Foul Subjects and Objects in Broadcast Soccer Videos

Chunbo Song^a and Christopher Rasmussen^b

Department of Computer and Information Science, University of Delaware, Newark, DE, U.S.A.

Keywords: Tracking, Detection, Video Analysis.

Abstract: We present a deep learning approach to sports video understanding as part of the development of an automated refereeing system for broadcast soccer games. The task of identifying which players are involved in a foul at a given moment is one of spatiotemporal action recognition in a cluttered visual environment. We describe how to employ multi-object tracking to generate a base set of candidate image sequences which are post-processed to mitigate common mistracking scenarios and then classified according to several two-person interaction types. For this work we created a large soccer foul dataset with a significant video component for training relevant networks. Our system can differentiate foul participants from bystanders with high accuracy and localize them over a wide range of game situations. We also report reasonable accuracy for distinguishing the player who committed the foul, or subject, from the object of the infraction, despite very low-resolution images.

1 INTRODUCTION


Computer vision is becoming ubiquitous for sports video analysis, with applications that include broadcast enhancement; real-time, in-depth player and team performance measurement; and automatic summarization of key events. Across analysis tasks there are several common visual skills such as ball tracking (Tong et al., 2004; Maksai et al., 2016); player segmentation (Canales, 2013; Lu et al., 2017; Huda et al., 2018), recognition (Gerke et al., 2015), and pose estimation (Kazemi and Sullivan, 2012); and recognition of formations, plays, and situations (Assfalg et al., 2003; Tsunoda et al., 2017; Wagenaar et al., 2017; Giancola et al., 2018).


Video-based assistance with officiating, in particular, is proliferating. The metric accuracy of high-speed, multi-camera ball tracking systems (e.g., (Ltd., 2020)) is relied upon in many sports including tennis and volleyball for line calls, baseball for balls and strikes, and soccer for so-called “goal line technology.” In soccer, the Video Assistant Referee (VAR) (FIFA.com, 2019) is commonly used for close and controversial decisions surrounding goals, major fouls, and player expulsions. However, despite the appearance of high technology, it is really nothing more

than an off-field human who flags situations that deserve further review by the head referee via video replays in slow motion from multiple angles.

As deep learning enables more sophisticated understanding of sports video imagery, one can imagine a future *automated refereeing* system, running live or on stored video, that blows a virtual whistle when it detects infractions. Using the sport of soccer as an example, such a system would classify *what* kind of violation occurred—e.g., handball, offside position, tripping or pushing, dangerous high kick, or another misdeed outlined in the FIFA rule book (Fédération Internationale de Football Association (FIFA), 2015)—and *who* was involved in the foul. *Foul events* occur at a location in time and space, and they involve at least one player *participant*. The player who committed the foul is the *foul subject* and the action performed is the *foul type*. Some fouls can be committed by a single player in isolation (such as touching the game ball with one’s hand), but here we focus on events that involve an opposing player, whom we refer to as the *foul object*.

This paper describes work toward a video-based automatic refereeing system. Here we assume that an oracle tells us that a two-player foul has occurred at a certain moment, leaving these two questions: *Who was involved in the foul, and who specifically committed it?* For a full, *live* system, temporal event detec-

^a  <https://orcid.org/0000-0001-6775-5553>

^b  <https://orcid.org/0000-0003-2831-8531>

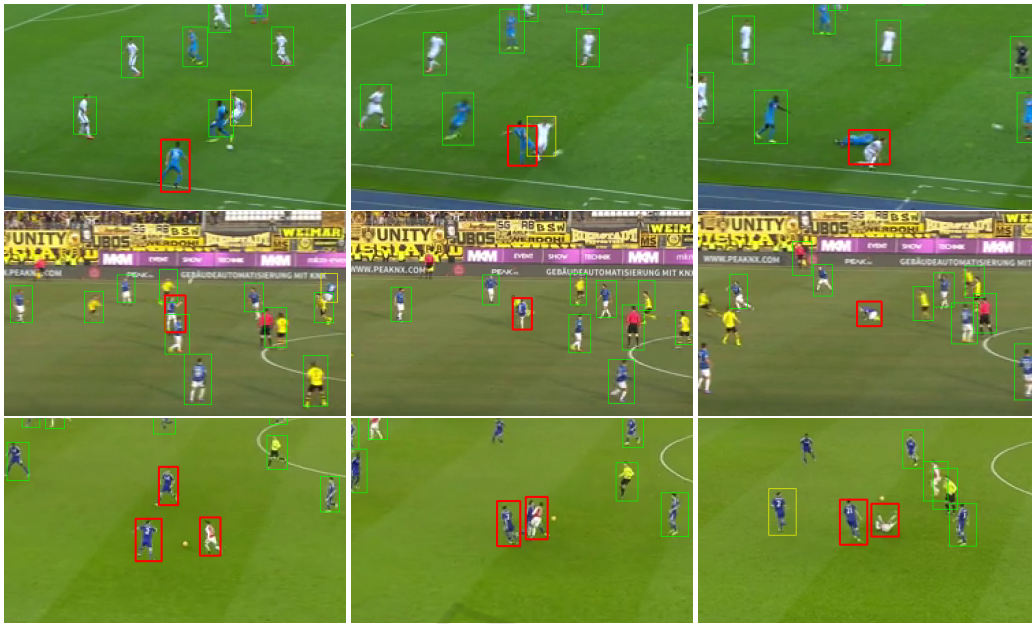


Figure 1: An image sequence for each tracked person, and their activity is classified as foul-related or not. Samples of foul participant detections are shown here with maximum likelihood candidates in red, over threshold in yellow, and non-participants in green (each row spans 2 seconds and the images are cropped to highlight the detections)

tion and foul severity classification would of course be crucial, and we will discuss in a moment how the work here overlaps with (and is therefore usable for) that task. But we argue that the foul oracle assumption is reasonable because non-video shortcuts can simulate it—from audio detection of whistle sounds that signal fouls (Raventos et al., 2014); or, for recorded games, from mining text or audio commentary for key words (as we describe in Sec. 3).

Static image analysis has a certain utility for this problem based on player poses and formations, but we assert that player movement patterns can be exploited to identify and differentiate foul participants. Here we describe an approach to recognizing telltale motions associated with soccer fouls such as slide tackles, pushing and gesturing, and falling to the ground via a three-stage pipeline. First, players are detected and tracked by a state-of-the-art multiple object tracking (MOT) method which we trained to perform well on broadcast soccer videos. Second, raw tracks are cleaned and augmented to account for common tracking errors that could result in crucial players not being covered by a complete track. Finally, processed tracks are fed to two video activity recognition networks to classify whether each person is (a) doing “normal” soccer things vs. exhibiting signs of being involved in a foul, and (b) if they do seem to be involved in a foul, to attempt to discriminate between the person committing the foul and the object of the foul. Fig. 2 shows this three-stage pipeline. The results demon-

strate that our method can achieve promising performance.

2 RELATED WORK

Person detection is one of the main topics in the area of the object detection. It typically applies similar network architectures as standard object detection models like Faster R-CNN (Ren et al., 2015) and Mask R-CNN (He et al., 2017) with some specific modifications for improving localization (Hasan et al., 2020; Zhang et al., 2017).

Thanks to the advantages of deep neural networks, great improvements have been made in action recognition, action detection, human-object interaction (HOI), and multi-object tracking (Peng et al., 2020). Action recognition could either apply 2D convolutions on per-frame input followed by another 1D module for aggregating the features (Karpathy et al., 2014; Simonyan and Zisserman, 2014) or apply stacked 3D convolutions to model temporal and spatial features (Tran et al., 2015; Carreira and Zisserman, 2017). (Feichtenhofer et al., 2019) uses two different pathways to operate on different frame rates for capturing both spatial semantics and temporal motions. Recently there has been more focus on *interactions* (Gkioxari et al., 2018; Ma et al., 2018; Zhou et al., 2020) with the goal of identifying $\{human, verb, object\}$ triplets in static images and videos.

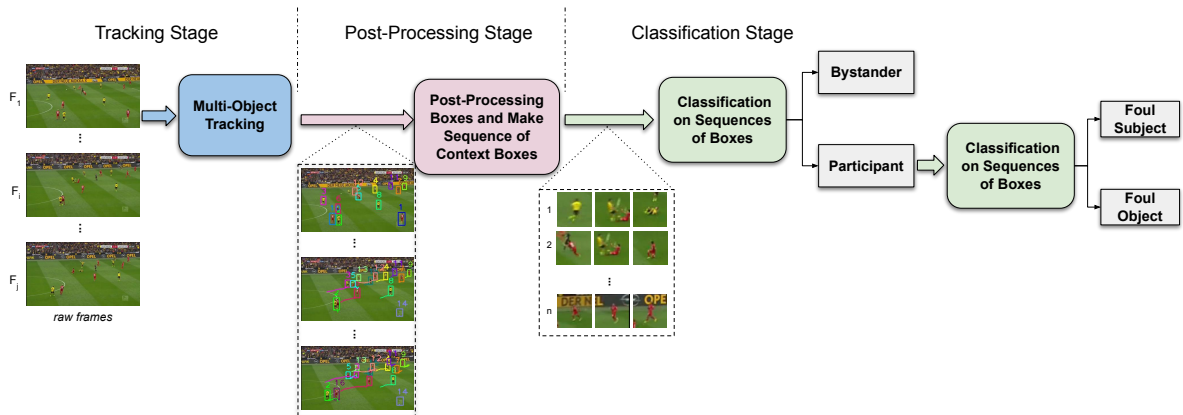


Figure 2: The three-stage pipeline of our work. At the first stage, we input the sequence of raw frames to Multi-Object Tracker to get players’ tracks. Then, each sequence of patches are extracted with post-processing instead of resizing to get context ROIs. At the last stage, the sequences go into the 3D classifier for identifying bystanders, foul subjects and objects.

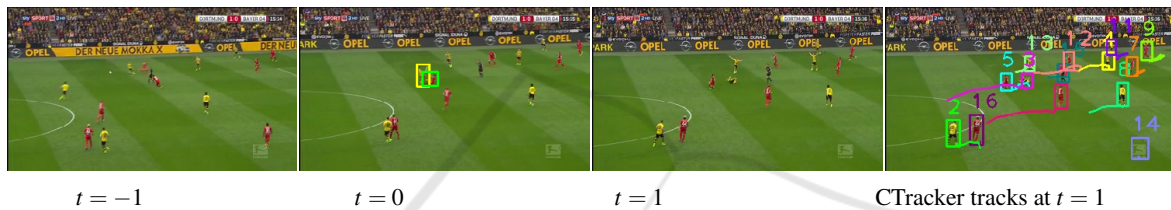


Figure 3: A sample two-person foul with 2-second temporal context around the *foul moment* at $t = 0$. The *foul subject* is denoted with a green bounding box (track 5 in the last column) and the *foul object* is marked with a yellow box (track 3).

(Liu et al., 2020) demonstrates high-quality spatial-temporal activity detection in a surveillance video scenarios, and more and more state-of-the-art methods have been utilized in the area of sports. (Vats et al., 2020) introduces a multi-tower temporal 1D convolutional network to detect events in ice hockey game and soccer game videos. (Hu et al., 2020) constructed their model based on deep reinforcement learning that shows only part of people’s activities have impacts on the entire group and tests their model on volleyball videos. (Sanford et al., 2020) use self-attention models to learn and extract relevant information from a group of soccer players for activity detection from both trajectory and video data. (Giancola et al., 2018) try to “spot” three soccer event categories: *goal*, *card*, and *substitution*.

3 DATASETS

Our foul dataset is built upon SoccerNet (Giancola et al., 2018), which comprises 500 complete soccer games from six European professional leagues, covering three seasons from 2014 to 2017, encoded mostly at 25 fps with a total duration of 764 hours. The footage is from broadcasts, so it includes camera pans and zooms, cuts between cameras, graphics

overlays, and replays. Both high-definition and lower-resolution (224p) versions are available; here we use the low-resolution version for all learning, evaluation, and paper figures except where noted.

442 SoccerNet games have text transcripts of audio commentary on game events which are timestamped by half and game clock with one-second precision. A sample foul is shown in Fig. 3 (and in more detail in Fig. 4) which corresponds to the following comment: *1 - 15:33: This yellow card was deserved. The tackle [...] was quite harsh and [the referee] didn’t hesitate to show him a yellow card.* We roughly located fouls by searching all transcripts for relevant words and phrases such as: “foul”, “violate”, “trip”, “bad challenge”, “rough challenge”, “handball”, “blows [...] whistle”, and “offside.” Video frames in the temporal neighborhood of each candidate’s timestamp were then manually examined to determine a precise *foul moment*. Clues from the commentary about which player committed the foul were used to resolve any visual ambiguities about the placement of the foul subject and object bounding boxes (green and yellow, respectively, in Fig. 3).

In all, 6492 foul events were labeled, of which 4862 were two-player fouls, as well as 1507 offside offenses and 123 handball offenses. Almost all of these events occurred in “far” camera views such as



Tight tracker ROI sequence for subject (top), object (bottom) in Fig. 3



Medium context ROIs on same subject and object

Figure 4: Sample *tight* and *context* ROI sequences derived from tracker output as input to the action recognition network.

shown in Fig. 3, but some were in close-ups or “near” views.

MOT Subset. 85-100 frame bounding box sequences (*tracks*) for all people ($n = 309$) present in 17 randomly-selected person detection frames (16 far, 1 near) were annotated over 4-second temporal windows $[-1, +3]$ s surrounding the foul moment. Tracks were manually trimmed at any shot boundaries (e.g., near/far transitions).

Action Recognition Subset. Complete 50-frame tracks for the foul subject and object were annotated over 2-second temporal windows $[-1, +1]$ s surrounding 833 randomly-selected two-person foul moments (all far views with no shot boundaries). Furthermore, 50-frame tracks for people ($n = 5006$) not involved in the foul, whom we call *bystanders* (e.g. other players, coaches, and referees) were obtained from CTracker (Peng et al., 2020) tracks that spanned the entire clip and did not overlap the ground truth subject or object bounding boxes.

4 METHODS

For identifying the player actions “committing a foul” and “being fouled,” we adopt the SlowFast network (Feichtenhofer et al., 2019) for video recognition. To adapt this network for our spatiotemporal task, we stabilize the video around each candidate player by assembling image sequences from tracker bounding boxes derived from an MOT tracker’s output. Here we use Chained-Tracker (CTracker) (Peng et al., 2020), which combines object detection, feature extraction,

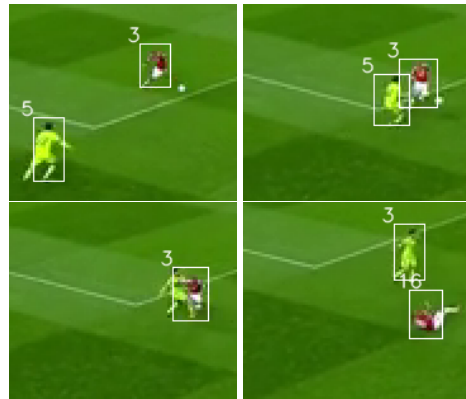


Figure 5: Example of CTracker mistacking: Track 5 disappears when the two players come together, and when they separate, track 3 follows the wrong player. Our post-processing corrects this: One *candidate* track is created via a *join* of the truncated 5 and the “wrong” ending of 3, and another track is made via a *branch* from the middle of 3 to the new track 16. The complete, erroneous track 3 also remains as a candidate.

and data association in a single end-to-end model that chains paired bounding box regression results estimated from overlapping nodes, of which each node covers two adjacent frames. CTracker achieves fast tracking speed (30+ Hz) and a Multiple Object Tracking Accuracy (MOTA) on MOT17 online of 66.6, which is highly competitive with other state-of-the-art algorithms.

As an example, the foul subject and object in Fig. 3 (green and yellow bounding boxes, respectively, at $t = 0$) are followed in tracks 5 and 3, respectively, produced by CTracker. Synopses of the sequences resulting from this *tight* tracking box, cropped and scaled to SlowFast’s 224×224 input, are shown in the top two rows of Fig. 4.

Raw tracker output can be noisy, exhibiting sudden shifts and scale changes that present challenges for video recognition, especially when the source ROIs are on the order of $\sim \times 30$ pixels. Moreover, the entire player might not be shown, losing valuable information about leg and hand motion, and certainly any depiction of *interactions* with nearby players is lost. Therefore, we expand the spatial context around each tracked bounding box on the hypothesis that it will aid the video recognition task. We define *context* ROIs as squares with sidelength proportional to the median max dimension of all tracker bounding boxes over an entire clip ($1.5 \times$ scaling for *medium*). Samples are shown in Fig. 4.

Track Post-processing. Tracks may be incomplete. In order to supply the video recognition network with sequences that span the full temporal context T and to mitigate mistacking and track merging and splitting

(see Fig. 5 for an example), we transform CTracker’s output to create a modified set of *candidate* tracks. First, tracks with small “gaps” of up to 5 or 6 frames are **patched** with linear interpolation between adjacent bounding boxes. In a second pass, tracks which end near another viable track are **joined** to them in order to extend them. Also in this pass, **branches** may be created between continuing tracks and new tracks that start nearby, increasing the overall number of tracks. In clips with high player densities, this may result in enlarged sets of candidate tracks with subsections in common.

Inference. Two SlowFast networks are used. SF_PvB classifies each candidate track video as either a foul *participant* (without regard to subject or object) or *bystander*, and SF_SvO classifies each candidate track video as a foul *subject* or a foul *object*. Because of the oracle assumption, we know that there is exactly one subject and one object per clip, transforming detection into a maximum likelihood problem. However, as seen in Fig. 5, there is not necessarily a one-to-one correspondence between tracks and people – we must always allow for the possibility that two players are being tracked by one box.

Participant detections are the bounding boxes at the foul moment from those tracks with the highest likelihood according to SF_PvB . There may be a tie due to floating point precision and the network output saturating; these are broken first by voting in the case that multiple maximum likelihood tracks share the same foul moment bounding box, and second randomly. Subject and object detections are maximum likelihood classifications according to SF_SvO , but they are only considered if already recognized as participants.

5 EXPERIMENTS

5.1 Training Details

CTracker. A CTracker network with a ResNet-101 backbone pre-trained on the MOT dataset (Milan et al., 2016; Peng et al., 2020) was fine-tuned on 10 4-second clips (9 far, 1 near) from our dataset in which all player tracks were manually annotated, with standard data augmentation.

SlowFast. We used the ResNet-50 8×8 variant of the network, pre-trained on the Kinetics dataset, for both of our video action classifiers. 666 48-frame, 2-second clips (with ground truth for 666 subjects and objects and 7996 bystanders) were randomly selected from our foul action recognition subset and SF_PvB

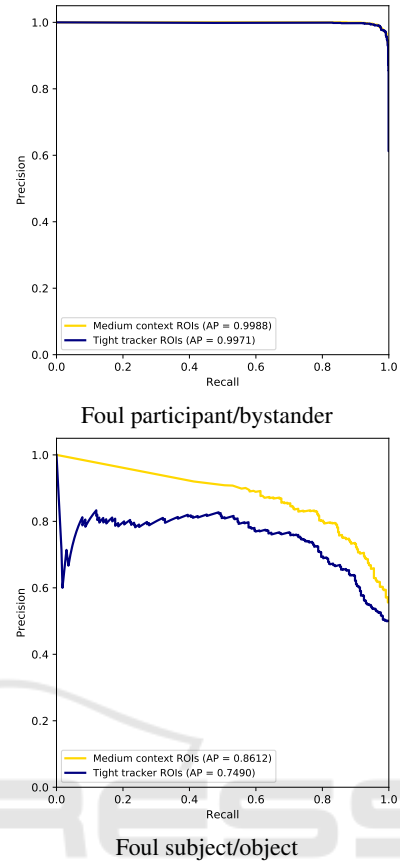


Figure 6: Precision-recall curves for SlowFast action classifiers.

and SF_SvO were fine-tuned for 10 and 40 epochs, respectively.

5.2 Results

CTracker’s tracking performance was assessed on 6 test video clips (all far views), resulting in an 88.6 MOTA.

The classification performance of SF_PvB and SF_SvO were measured on a test set of 167 clips (with ground truth ROI sequences for 167 subjects and objects and 1008 bystanders). Precision-recall curves for each network trained on *tight* tracker ROIs vs. the looser context ROIs discussed in Sec. 4 are plotted in Fig. 6. For both training regimens, SF_PvB is nearly perfect, with an average precision (AP) of 0.997 for tight ROIs and 0.999 for context ROIs after 10 epochs. The subject vs. object task seems harder, as blame is hard to assign to two tussling players, and while foul objects often wind up sprawled on the ground, so do the foul subjects whether intentionally or not. This assessment is borne out by SF_SvO ’s lower performance after 40 epochs of training, with an AP = 0.749 for

tight ROIs and 0.861 for context ROIs. Training on high-res videos with context ROIs did not improve performance, yielding an AP of 0.848.

The context ROI of a player who is the object of a foul often includes, completely or partially, the subject who is fouling him or her—and vice versa (e.g., the 3rd and 4th columns of Fig. 4). Therefore, a binary subject-object label seems improper and may slow training. So we propose a multi-label task in which each video clip is labeled with two floating-point values between 0 and 1 indicating *subjectness* and *objectness* by computing the median IoU between the ROI and subject and object bounding boxes over every frame of the sequence. In this case the AP rises to 0.980.

The context variant of SF_PvB successfully detected 64.24% of foul participants @ 0.5 IoU threshold at the foul moment over a test set of 167 clips (vs. 52.51% for the tight variant with the same tracks). Fig. 1 shows three examples of such detections. The second row demonstrates the detector’s ability to pick out one anomalous motion in a crowd (in this case the foul object sinking to the ground). Subjects and objects were detected at the same IoU threshold with 30.15% and 45.21% accuracy, respectively (16.39% and 30.06% for tight). The detection accuracy is considerably higher at lower IoU thresholds (e.g. 84.34% @ 0.1 IoU), indicating that this approach locates the rough foul area quite robustly.

6 CONCLUSION AND FUTURE WORK

We report strong performance on a sports spatiotemporal video activity recognition task. There are a number of directions to take before removing the foul oracle assumption and working on the scale of entire games, including extending the system to near-view clips with more training examples, dealing with shot boundaries automatically, and incorporating foul-relevant information outside of subject/object bounding boxes. Filtering subject/object hypotheses by making sure candidate pairs are on opposite teams could boost performance, but require a per-game learning of jersey colors and patterns using, for example, deep image clustering (Li et al., 2021). Using high-res versions of the game videos would enable further analysis such as ball tracking and reading player names/jersey numbers to correlate with roster data and/or commentary. Finally, camera pose estimation and parsing of field line features (Cuevas et al., 2020) would help filter off-field person detections and recognize foul-relevant game situations.

REFERENCES

- Assfalg, J., Bertini, M., Colombo, C., Bimbo, A. D., and Nunziati, W. (2003). Semantic annotation of soccer videos: automatic highlights detection. *Computer Vision and Image Understanding*, 92(2):285–305.
- Canales, F. (2013). *Automated Semantic Annotation of Football Games from TV Broadcast*. PhD thesis, Department of Informatics, TUM Munich.
- Carreira, J. and Zisserman, A. (2017). Quo vadis, action recognition? a new model and the Kinetics dataset. In *IEEE Conference on Computer Vision and Pattern Recognition*.
- Cuevas, C., Quilón, D., and García, N. (2020). Automatic soccer field of play registration. *Pattern Recognition*, 103.
- Fédération Internationale de Football Association (FIFA) (2015). Laws of the game. <https://img.fifa.com/image/upload/datz0pms85gbnqy4j3k.pdf>.
- Feichtenhofer, C., Fan, H., Malik, J., and He, K. (2019). Slowfast networks for video recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 6202–6211.
- FIFA.com (2019). Video assistant referees (VAR). <https://football-technology.fifa.com/en/media-tiles/video-assistant-referee-var>.
- Gerke, S., Muller, K., and Schafer, R. (2015). Soccer jersey number recognition using convolutional neural networks. In *IEEE International Conference on Computer Vision Workshop*.
- Giancola, S., Amine, M., Dghaily, T., and Ghanem, B. (2018). Soccernet: A scalable dataset for action spotting in soccer videos. In *CVPR Workshop on Computer Vision in Sports*.
- Gkioxari, G., Girshick, R., Dollár, P., and He, K. (2018). Detecting and recognizing human-object interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8359–8367.
- Hasan, I., Liao, S., Li, J., Akram, S. U., and Shao, L. (2020). Generalizable pedestrian detection: The elephant in the room.
- He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969.
- Hu, G., Cui, B., He, Y., and Yu, S. (2020). Progressive relation learning for group activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 980–989.
- Huda, N., Jensen, K., Gade, R., and Moeslund, T. (2018). Estimating the number of soccer players using simulation-based occlusion handling. In *CVPR Workshop on Computer Vision in Sports*.
- Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732.
- Kazemi, V. and Sullivan, J. (2012). Using richer models for articulated pose estimation of footballers. In *British Machine Vision Conference*.

- Li, Y., Hu, P., Liu, Z., Peng, D., Zhou, J., and Peng, X. (2021). Contrastive clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Liu, W., Kang, G., Huang, P.-Y., Chang, X., Qian, Y., Liang, J., Gui, L., Wen, J., and Chen, P. (2020). Argus: Efficient activity detection system for extended video analysis. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision Workshops*, pages 126–133.
- Ltd., H.-E. I. (2020). Products: Ball tracking. <https://www.hawkeyeinnovations.com/products/ball-tracking>.
- Lu, K., Chen, J., Little, J. J., and He, H. (2017). Light cascaded convolutional neural networks for accurate player detection. In *British Machine Vision Conference*.
- Ma, C.-Y., Kadav, A., Melvin, I., Kira, Z., AlRegib, G., and Peter Graf, H. (2018). Attend and interact: Higher-order object interactions for video understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6790–6800.
- Maksai, A., Wang, X., and Fua, P. (2016). What players do with the ball: A physically constrained interaction modeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.
- Milan, A., Leal-Taixé, L., Reid, I., Roth, S., and Schindler, K. (2016). Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*.
- Peng, J., Wang, C., Wan, F., Wu, Y., Wang, Y., Tai, Y., Wang, C., Li, J., Huang, F., and Fu, Y. (2020). Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking. In *European Conference on Computer Vision*, pages 145–161. Springer.
- Raventos, A., Quijada, R., Torres, L., Tarres, F., Carsusan, E., and Giribet, D. (2014). The importance of audio descriptors in automatic soccer highlights generation. In *Proceedings of the IEEE International MultiConference on Systems, Signals, and Devices*.
- Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.
- Sanford, R., Gorji, S., Hafemann, L. G., Pourbabae, B., and Javan, M. (2020). Group activity detection from trajectory and video data in soccer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.
- Simonyan, K. and Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576.
- Tong, X., Lu, H., and Liu, Q. (2004). An effective and fast soccer ball detection and tracking method. In *International Conference on Pattern Recognition*.
- Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497.
- Tsunoda, T., Komori, Y., Matsugu, M., and Harada, T. (2017). Football action recognition using hierarchical lstm. In *CVPR Workshop on Computer Vision in Sports*.
- Vats, K., Fani, M., Walters, P., Clausi, D. A., and Zelek, J. (2020). Event detection in coarsely annotated sports videos via parallel multi-receptive field 1d convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 882–883.
- Wagenaar, M., Okafor, E., Frencken, W., and Wiering, M. (2017). Using deep convolutional neural networks to predict goal-scoring opportunities in soccer. In *International Conference on Pattern Recognition Applications and Methods*.
- Zhang, S., Benenson, R., and Schiele, B. (2017). Citypersons: A diverse dataset for pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3221.
- Zhou, T., Wang, W., Qi, S., Ling, H., and Shen, J. (2020). Cascaded human-object interaction recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4263–4272.