Self-Supervised Small Soccer Player Detection and Tracking

Supplementary Material

ACM Reference Format:
. 2020. Self-Supervised Small Soccer Player Detection and Tracking: Supplementary Material. In *3rd International Workshop on Multimedia Content Analysis in Sports (MMSports'20), October 16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3422844.3423054

1 TRAINING DATA

Our training images are unlabeled TV broadcast soccer videos taken from a subset of the SoccerNet dataset, where we extract only wide views. Figure 1 and the left column of Figure 2 show some examples of these training images.

Figure 1: Examples of SoccerNet training images.

2 DETAILS OF THE METHOD

In this section, we detail the algorithms used to remove false positives and add false negatives detections to the teacher $\mathcal{T}^{init}$ annotations. These methods have to be as accurate as possible and can be, to a certain extent, computationally expensive as they are not used for the final inference of the student $\mathcal{S}$.

2.1 Field detection

As a first step for detecting the field we find a mask that contains the potential pixels of the field. Then, it is successively refined. As observed in Figure 1, coaches are often present in the green bottom part of the field, and those should not be included in the mask. The field mask is computed thanks to following successive processing steps :

- Application of a green filter to create a first binary image. We use as upper and lower color threshold the values (15,50,50) and (70,255,255).
- Selection of the contour with the biggest connected component.
- Approximation of the contour by a polygon.
- Detection of the bottom line of the field and removal of the mask pixels below it.

To detect the bottom line we use the [1] pix2pix model to get a binary image approximating the lines of the field. We then run a Hough line transform on this binary image and use thresholds on the length, angle and position of the detected lines. Some examples of the final estimated field mask are shown in Figure 2.

Figure 2: Three results obtained by the proposed field detection approach. Left: original image. Right: Output of our field detector.

2.2 Add missed players to the training data

We detail here the blob detection strategy used to add players to the training data. Given an input image u, we work on the pixels belonging to the previously extracted field mask. Note that the teacher $\mathcal{T}^{init}$ has already detected some of the players $\mathcal{T}^{init}(u)$.

We first compute blobs through green filtering and contour selection. We select the contours for which the enclosing bounding box verifies : i) it has a similar area as the bounding boxes of the players already detected by the teacher $\mathcal{T}^{init}$ and ii) it does not exist any detection in $\mathcal{T}^{init}(u)$ with intersection-over-union $IOU < 0.2$.

We then run a human pose estimation algorithm [2] on the corresponding enlarged extracted regions. If the confidence of the pose detection is higher than 0.8, it is considered as a player and it is added to the annotations. We have observed in our experiments that the pretrained version of the human pose estimator [2] performs better on small humans than the multi-person FPN-FasterRCNN model, $\mathcal{T}^{init}$.

We display in Figure 3 some examples of the training images with corrected annotations $\mathcal{Y}^c$.

MMSports'20, October 16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-8149-9/20/10…$15.00
https://doi.org/10.1145/3422844.3423054

Figure 3: Examples of training images with the annotations of $\mathcal{Y}^c$. In white the annotations of the pretrained teacher $\mathcal{T}^{init}$ and in green the added players detected with the strategy detailed in Section 2.2 to include missing players.

Table 1: AP on *SPD* of fine-tuned $\mathcal{T}^f$ and $\mathcal{S}$ before and after fine-tuning on *SPD* images.

Model	$\mathcal{T}^f$	$\mathcal{S}$
Before fine-tuning	92.0	96.0
After fine-tuning	99.3	98.9

3 IMPLEMENTATION DETAILS

We indicate here additional parameter values involved in the proposed framework :
i) We train the teacher and student detectors during 20 epochs using the SGD optimizer with an initial learning rate of 10^{-3} decayed with a factor 0.1 at epochs 8, 14 and 17.
ii) The fine-tuning with triplet loss is done on 5 epochs with initial learning rate of 10^{-4}.
iii) For tracking inference we find the optimal parameters to be : $N_{reID} = 10$; $\alpha = 0.03$; $D_{visual\_max} = 4$ and $D_{spatial\_max} = (1/16) * image\_width$.

4 QUANTITATIVE RESULTS

In Section 4.5.1 of our paper, Table 5 shows that the Resnet18 backbone outperforms Resnet50 on the *SPD* dataset. We made the hypothesis that it is due to the fact that there exists a small difference between the appearance of the players in *SPD* and SoccerNet training images.

In Table 1 we propose to fine-tune $\mathcal{T}^f$ (Resnet50 backbone) and $\mathcal{S}$ (Resnet18 backbone) on 80% of the *SPD* dataset. We then evaluate on the remaining 20%. With this fine-tuning it is now the Resnet50 backbone that outperforms its Resnet18 counterpart. This test confirms that the Resnet50 backbone slightly overfits on the typical TV broadcast player appearances of the training data. Besides providing a faster inference, using a smaller backbone allows the student $\mathcal{S}$ to better generalize.

5 VISUAL RESULTS

5.1 Player detection

We show in Figure 5 additional player detection results on the *SPD* (row 1), *ISSIA* (row 2) and *TV_soccer* (row 3) datasets, as well as on the *panorama* images of [3] (row 4). In order to test the detection on very small players, we down-scale and pad the input images. We show that our network is able to detect almost all the players in very challenging images, even when a player contains only a few pixels.

To show the strength of our network we present in Figure 4 results with the pretrained FPN-FasterRCNN, $\mathcal{T}^{init}$, on the same challenging images. It fails at finding small players and detects a lot of non-player humans.

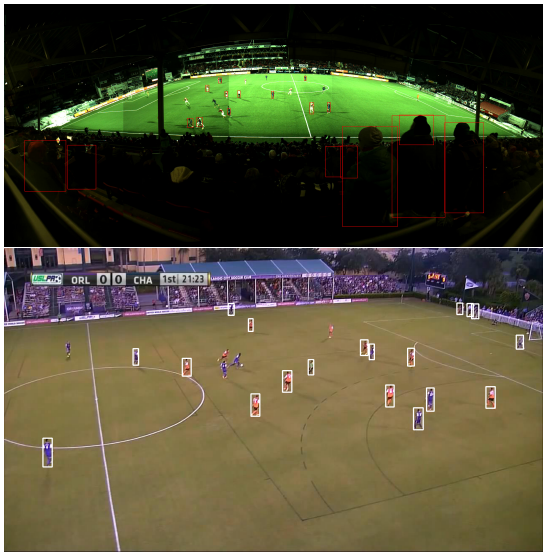


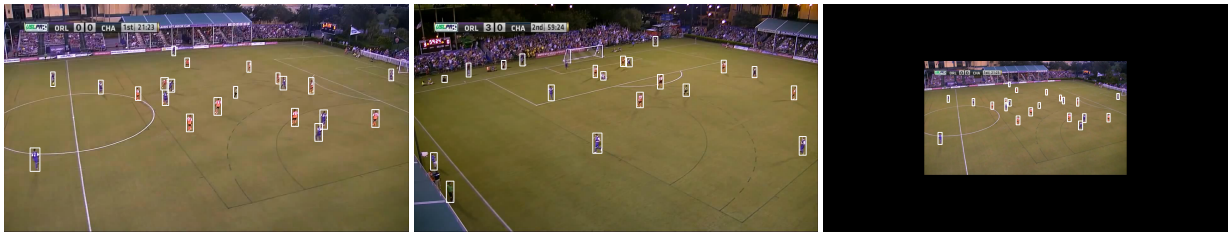
Figure 4: Player detections results with the human annotations of FPN-FasterRCNN pretrained on COCO ($\mathcal{T}^{init}$).

Failure cases. Figure 5 includes some failure cases. For example, in the second image of *SPD* (row 1), non-players are detected on the side of the field. In this particular case, the distinction between players and non-player is challenging because the latter appear on the field. The network can also fail when the background around a player is not green field, like in the third image of *ISSIA* (row 2).

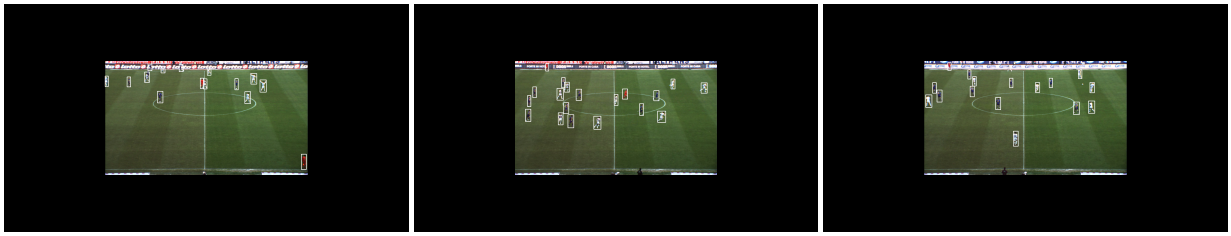
5.2 Player tracking

Figure 6 shows two tracking results on the most challenging *SPD* and panorama sequences. We present four images of each sequence, one every 5 frames, from top to bottom.

We observe that almost all the players are correctly followed. One failure case appears in the first column, in a typical crowded scene. Due to occlusion, the white player 20 is lost between the frames 2 and 3. The re-ID module fails to re-identify him at frame 4 and it is assigned a new id 21. The reason is that the appearance of the player between frame 2 and 4 has changed substantially. Moreover, the bounding box given by the tracker is not really precise for a small player, it can involve only one part of the player, or its close neighbours. In these cases, the visual embedding extracted from this bounding box is not good enough to perform an association.

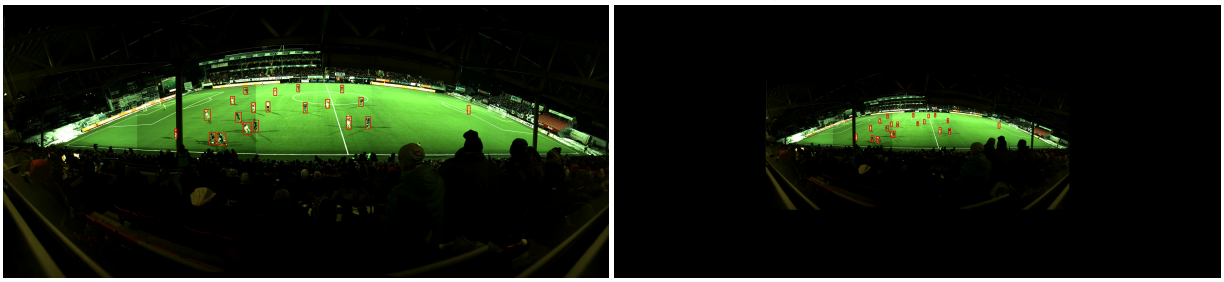


(a) *SPD* **dataset**



(b) *ISSIA* **dataset with down-scaling** $x0.5$

(c) *TV_soccer* **dataset with down-scaling** $x0.5$



(d) images from dataset [3]

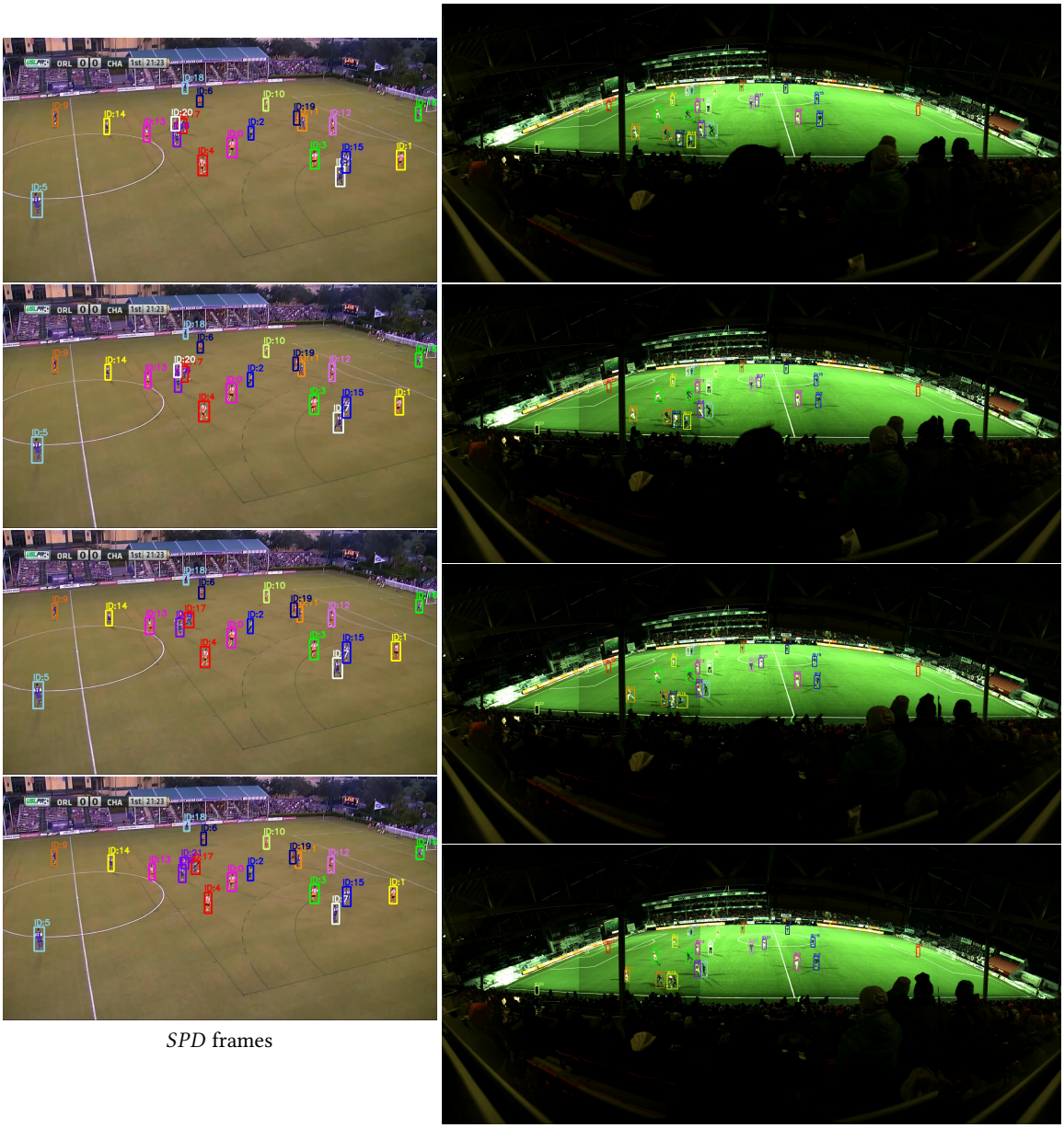
Figure 5: Player detection results on different datasets, for different down-scaling levels.

REFERENCES

[1] Jianhui Chen and James J Little. 2019. Sports camera calibration via synthetic data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 0–0.

[2] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. 2018. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7103–7112.

[3] Svein Arne Pettersen, Dag Johansen, Håvard Johansen, Vegard Berg-Johansen, Vamsidhar Reddy Gaddam, Asgeir Mortensen, Ragnar Langseth, Carsten Griwodz, Håkon Kvale Stensland, and Pål Halvorsen. 2014. Soccer Video and Player Position Dataset. In *Proceedings of the 5th ACM Multimedia Systems Conference* (Singapore, Singapore) *(MMSys '14)*. Association for Computing Machinery, New York, NY, USA, 18–23. https://doi.org/10.1145/2557642.2563677



SPD frames

panorama frames

Figure 6: Player tracking results on a *SPD* sequence and a *panorama* sequence.