Transductive Weakly-Supervised Player Detection using Soccer Broadcast Videos [Supplementary]

Chris Andrew Gadde[1] C.V. Jawahar[2]
International Institute of Information Technology, Hyderabad, India
[1]chris.andrew@research.iiit.ac.in, [2]jawahar@iiit.ac.in

1. Identifying reliable predictions

To construct the similarity graph mentioned in our approach, we use four clustering models, two K-Means and two Gaussian Mixture Models [5], with clusters of 10 and 20 each. We found that the generated similarity graph did not vary much with the clustering models used as long as sufficient number of models and large enough cluster numbers were used. For the *cluster deletion* phase, we set $t_p = 4$ and $t_c = 0.8$. As mentioned before, setting $t_p = p$ allows the cluster deletion algorithm to run in linear time. This was useful since we had close to 500k bounding boxes proposed by our initial inductive model and an $O(n^2)$ algorithm would not be recommended in such cases. The drawback to using $t_p = p$ is that a large number of outlier clusters are created. We remove those clusters by thresholding based on the cluster size. Since we have a large number of proposed bounding boxes from our inductive set, we did not see any lack of training data for the transductive model after pruning. We label representative samples with domain labels and retain only *reliable* predictions for the transductive stage.

2. Training transductive model

Training of the model is done by resizing the input images to 416x416 with padding to maintain the aspect ratio of the bounding boxes. Images are fed in batches of 32, as we found that larger batches lead to poor generalisation and quickly overfit the model. We train the model using the multi-part MSE and cross-entropy loss mentioned in [4], represented in Eq. 1.

$$L = L_c(y) + \lambda_n L_n(C) + L_o(C) + \lambda_l L_l(S) \qquad (1)$$

Where L_c represent the classification loss for the class label y, L_n represents the object confidence loss when no object is present in the bounding box scaled by a factor of λ_n, L_o represent the object confidence loss when an object is present in the bounding box and L_l represents the localisation loss on the bounding box co-ordinates S scaled by a factor of λ_l. Since most boxes do not contain any objects we set $\lambda_n = 0.5$ to equalise this imbalance. We also found $\lambda_l = 8$ to be a good value to ensure better localisation to help predict more accurate bounding boxes. The model is trained for 200 epochs, with an initial learning rate of 0.0001, decayed by a factor of 0.1 at 100 and 150 epoch. Models are trained per video and then tested in the testing phase. Although our method requires models to be trained for every video, the performance gains are significant compared to models trained on several videos such as SoccerDB [3]. This improvement is achieved only with the help of domain-labels and no new bounding box information is given to the model, other than the initial predictions made by the inductive copy.

3. Clustering baselines

For comparison of our clustering approach we train models of different clustering algorithms and compare their performance in identifying reliable predictions in the cluster pruning stage. We use the visual features obtained from the re-identification model f_{ij} for this purpose. Clustering models using K-Means and Gaussian Mixture Models [5] are trained with 20, 40 and 80 clusters, and performance of the best model for each of these methods is compared to our clustering approach. Since our approach determines the number of clusters based on the similarity graph, we need evaluate different number of clusters for K-means and GMM and best performance for each is reported in the main paper. As mentioned earlier, since the similarity graph did not vary as long as p was high enough, t_p values were always set to p to favour efficiency. We found that setting $t_p < p$ yielded larger clusters, but we found in those cases that multiple samples were candidates for the representative sample for the cluster, some even having different domain labels. Experiments were also done to use simpler features from image classification networks such as ResNet-18 [1]. We found that using these features made it difficult to distinguish between players and audience members or staff wearing the team jersey. The visual features generated in these cases were too similar to uniquely identify true-positives from false-positives. These features also give lower values for TPR(True Positive Retention Ratio) and FPR(False Positive Removal Ratio) for the task of identifying reliable samples as compared to features from the re-identification model.

4. Field heat-maps

For generating the heatmaps for the match, we follow the method described in [2], to perform top-view registration of the play field. This is done by estimating a homography matrix for the frames in our video to warp them onto a template image of the field. We use the inverse homography to warp the image onto the playing field template. Figure 1 shows the various stages in the top-view registration method used. We then warp the centers of the bounding boxes obtained from our detection model for every player in the frame onto a blank template. This is done for every frame in the video, adding detections onto histogram, and a heat-map is generated for the playing field over the entire video. The heatmap shows the distribution of the player positions on the playing field across the entire video. With a proper player recognition model, these templates can be generated for individual players, giving much more insights into the game.

References

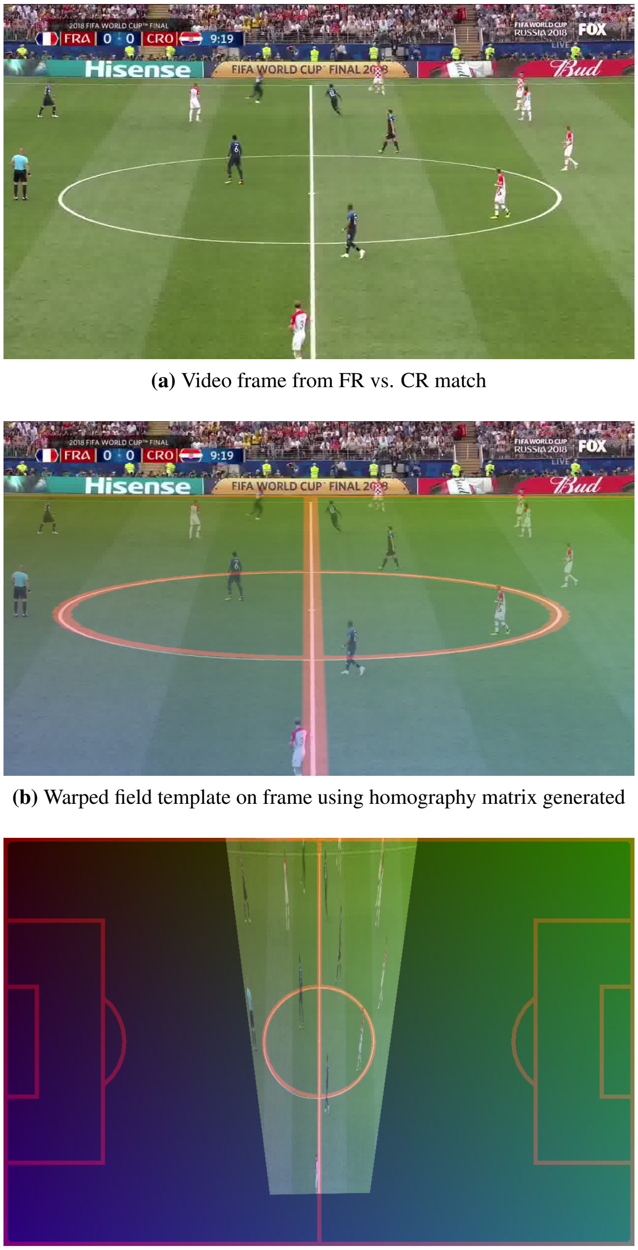
[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[2] 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 *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 201–210, 2020.

[3] Yudong Jiang, Kaixu Cui, Leilei Chen, Canjin Wang, and Changliang Xu. Soccerdb: A large-scale database for comprehensive video understanding. In *Proceedings of the 3rd International Workshop on Multimedia Content Analysis in Sports*, MMSports '20, page 1–8, New York, NY, USA, 2020. Association for Computing Machinery.

[4] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[5] Douglas A Reynolds. Gaussian mixture models. *Encyclopedia of biometrics*, 741:659–663, 2009.


2018 FIFA WORLD CUP FINAL
FRA 0 0 CRO 9:19
FIFA WORLD CUP RUSSIA 2018
FOX
Hisense
FIFA WORLD CUP FINAL 2018
Bud

(a) Video frame from FR vs. CR match

(b) Warped field template on frame using homography matrix generated

(c) Frame warped onto field template using inverse homography

Figure 1: Various stages of mapping player bounding boxes onto a field template