Tracking Data Visual Representations for Sports Broadcasting Enrichment

Murilo Couceiro[1,3][a], Inês Rito Lima[3][b], Alexandre Ulisses[3][c], Tiago Mendes-Neves[1,2][d] and João Mendes-Moreira[1,2][e]

[1]*Faculdade de Engenharia, Universidade do Porto, Porto, Portugal*

[2]*LIAAD - INESC TEC, Porto, Portugal*

[3]*MOG Technologies, Department of Innovation, Maia, Portugal*

Keywords: Metadata Enhancement, Player Tracking, Computer Vision, Sports Broadcasting.

Abstract: The broadcast of audio-video sports content is a field with increasingly larger audiences demanding higher quality content and involvement. This growth creates the necessity to develop more content to engage the users and keep this trend. Otherwise, it may stall or even diminish. Therefore, enhancing the user experience, engagement, and involvement during live sports event broadcasts is of utmost importance. This paper proposes a solution to extract event's information from video, resorting to Computer Vision techniques and Deep Learning algorithms. More specifically, the project encompassed the definition and implementation of field registration, object detection and tracking tasks. Focusing on football sports events, a novel dataset combining several video sources was created and used for analysis and metadata extraction. In particular, the proposed solution can detect and track players with acceptable precision using state-of-the-art methods, like YOLOv5 and DeepSORT. Furthermore, resorting to unsupervised learning techniques, the system provides team segmentation based on the colour of the players' kits. A series of visual representations regarding the players' movements on the field enables broadcast enrichment and increased user experience. The presented solution is framed in the H2020 DataCloud project and will be deployed in a cloud environment simplifying its access and utilisation.

1 INTRODUCTION

As the most popular sport in our society, with millions of fans, football has very high market value and has a growing amount of visual data generated, leading to a growth in the research of this topic. Football clubs, coaches and fans demand customised football match analysis services, creating the need within the broadcasting industry for automated video annotation and metadata extraction tools.

The proposed work is framed in the DataCloud project[1], particularly on the MOGSports business case. MOGSports focuses on the media and entertainment business and aims to enhance the user experience, engagement, and involvement during live sports event broadcasts.

This paper explores the development of a system that assesses multi-camera video content and provides visual representations to enhance the user experience on live sports broadcasting. Additionally, the presented system extracts tracking data as a first step towards event detection. The proposed data acquisition setup includes two cameras to extract information from the entire field at all times. The system extract metadata by a detection and tracking algorithm for each camera. This uni-camera information is then fused, creating a final multi-camera global result. The solution must be deployed in a Cloud environment along with the project objectives. Moreover, the project relates to complementing a broadcast, so real-time or near-real-time extraction is necessary. Thus, the size and velocity of the methods used are critical aspects.

Football matches video contains prominent and diverse information that can be extracted to enrich the

[a] https://orcid.org/0000-0002-8451-9680
[b] https://orcid.org/0000-0002-9681-4740
[c] https://orcid.org/0000-0003-4802-1508
[d] https://orcid.org/0000-0002-4802-7558
[e] https://orcid.org/0000-0002-2471-2833

[1]Horizon2020 grant agreement nº 101016835

Couceiro, M., Rito Lima, I., Ulisses, A., Mendes-Neves, T. and Mendes-Moreira, J.
Tracking Data Visual Representations for Sports Broadcasting Enrichment.
DOI: 10.5220/0011551900003321
In *Proceedings of the 10th International Conference on Sport Sciences Research and Technology Support (icSPORTS 2022)*, pages 125-131
ISBN: 978-989-758-610-1; ISSN: 2184-3201

content of a broadcast. This paper presents a solution for extracting player position and team information, in near real-time, throughout the match, providing the equivalent positional representations as output. The proposed methodology is crucial to enable further complex analysis such as goals, passes, player velocity and other statistics extraction. Furthermore, the applications of the retrieved positional data are manifold since it is the basis for most sport-related Machine Learning work, making it an excellent contribution to the Sports Data Mining Community.

Following the context provided in the current section, a literature overview on sports analytics, computer vision and machine learning is given in Section 2. Section 3 presents the pipeline of the project at hand and detailed information about its modules. Section 4 presents the results obtained with the proposed solution, and Section 5 describes its positive and negatives sides as well as recommendations and ideas for future work. Finally, Section 6 summarises the work done for this paper.

2 RELATED WORK

Investigation into sports-related areas has been getting much attention. Many studies have been conducted in Computer Vision and Deep Learning to extract football matches information through video. This paper covers the areas of Field Registration, Object Detection and Object Tracking. All to extract information from a video of a football match.

The task of sports field registration aims at determining a homography matrix H for the transformation of an image into two-dimensional sports field coordinates (Theiner et al., 2022). The homography is defined by a 3x3 matrix with eight degrees of freedom that performs mapping between two images of a planar surface from different perspectives. There are two traditional homography estimation approaches, direct and feature-based methods, which can also be used together to increase robustness (Nie et al., 2021; Nguyen et al., 2018). The direct methods use pixel-to-pixel matching by shifting or warping the images relative to each other and comparing the pixel intensity values using an error metric. The feature-based approach first extracts keypoints in each image using local feature detectors to establish correspondences between the two sets. Also, inspired by the data-driven Convolution Neural Networks (CNN) in CV, approaches for homography estimation have emerged using supervised methods. For this extent, (Nie et al., 2021) regresses a set of dense features through a deep network, to estimate the homography. Another approach is to use unsupervised methods (Nguyen et al., 2018). Generative Adversarial Networks (GAN) and synthetic data have also been proposed (Chen and Little, 2019)

Player detection is a case of object detection but with additional challenges. When performing player detection, several associated challenges must be faced to obtain successful results: the occlusions between players, the abrupt movements of the camera, the changes in lighting, the lack of resolution in very distant players, the blurring of the players that are moving, or the players remaining long static periods (Cuevas et al., 2020). When (Sah and Direkoglu, 2021) compares object detection models such as Faster R-CNN, YOLO and SSD using different backbones and other parameters for a field hockey dataset, it concludes that these complex neural networks fail to detect occluded players and players with low resolution (far from the camera).

Player Tracking is a Multi-Object-Tracking (MOT) problem. The MOT methods can be grouped regarding the initialisation method into two sets: Detection-Based Tracking (DBT) and Detection-Free Tracking (DFT). Briefly, DBT, commonly named tracking-by-detection, happens when objects are first detected and then linked into trajectories. Thanks to improvements in the performance of object detectors and the ability to deal with challenges such as cluttered scenes or dynamics of tracked objects, tracking-by-detection has become a leading paradigm for MOT (Ivasic-Kos et al., 2021). This approach is followed by methods like SORT, DeepSORT, TRACKTOR++, FairMOT, TransMOT, and StrongSORT. These methods make the tracking step an association step for the detections received by the detector.

The game analysis applications have a variety of topics: action spotting or event detection, player and ball tracking, tactics analysis, statistics collection, pass feasibility, talent scouting, game phase analysis, or highlights generation (Cioppa et al., 2021). Furthermore, there is work on the enhancement and experience area, like (Rematas et al., 2018) that uses augmented reality to present a match in a 3D model on a flat surface.

3 PROPOSED SOLUTION

This section describes the MOGSports modular pipeline shown in Figure 1, focusing on the AI Event Detection and Metadata Fusion modules.

The pipeline was designed to enable the use of N cameras, each undertaking an AI Event Detection module. This module includes three steps: Field De-

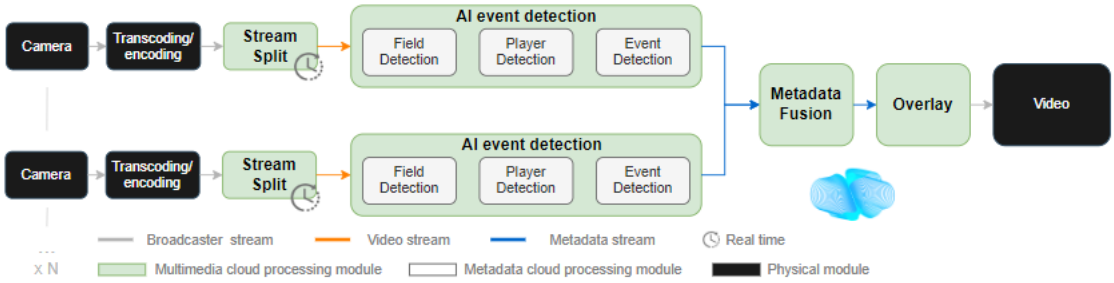

Figure 1: MOGSports pipeline.

tection, Player Tracking and Event Detection. This paper presents the first two modules.

Field Detection encapsulates the task of Field Registration. Resorting to a homography matrix, Field Registration allows the transformation from the game image to a top view perspective of the field. With the players' detection representations, this top model allows a better understanding of the game.

The Metadata Fusion module combines the information retrieved from each camera video processing. In this sense, the fusion module must adjust accordingly, accompanying the camera scalability.

3.1 Camera Setup

Whilst using more cameras enables detailed field coverage, it also increases the computation effort when matching the extracted data. Hence, the counterbalance between the number of cameras and the fusion algorithm complexity must be evaluated.

In this sense, the proposed setup comprises two static cameras placed along one side-line, as shown in Figure 2, guaranteeing the coverage of the entire field with significant detail. Furthermore, this setup allows the design of innovative strategies to fight occlusion, one of the problematic aspects of payer detection and tracking task

This setup was tested at Estádio Municipal de Aveiro stadium, and the cameras were around 20 meters high and had around 15°tilt.

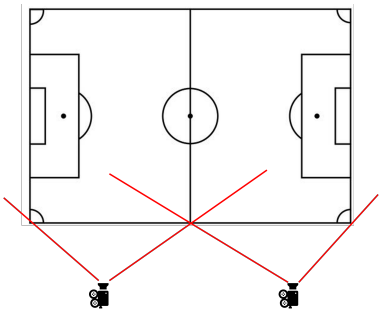

Figure 2: Proposed Camera Setup.

3.2 Field Detection

The Field Detection problem relates to the homography matrix computation between the captured image and a top-view field image.

Since the setup cameras are static, the homography computation only needs to happen once at the match's beginning. The premise of a constant homography makes the registration process exponentially cheaper when compared to a moving camera, where it can change several times per second.

Furthermore, this premise also enables a lightweight manual registration. Thus, the proposed solution integrates a manual registration process, requiring the user to select four or more corresponding points in one video frame and on the top view field image.

3.3 Player Tracking

Player Tracking allows the identification and re-identification of the detected players throughout the frames, as shown in Figure 5. Tracking can create and represent more valuable information that augments the user experience, such as other visual representation options or individual player statistics. The chosen tracking algorithm is based on a tracking-by-detection methodology, so player detection is a fundamental step. Without a robust detection model, the tracking algorithm will perform poorly.

3.3.1 Player Detection

The detection algorithm must be as accurate and fast as possible, enabling the pipeline to run in real-time. Also, it must be precise to strengthen the tracking step. These requirements drove the choice of the one-stage algorithm YOLOv5 (Jocher et al., 2022). This object detector is chosen as it has the best speed and accuracy counterbalance, as well as an easy integration. As very fast inference is needed, only one-stage detectors were considered.

The players' position in the field is set where they stand, at their feet. Foreseeing that the player is in the centre of the retrieved bounding box, its position is defined with the bounding box's bottom centre coordinates. Once mapped by the homography, this point is the player's position in the top-view representation. The top-view field image is predetermined, with known field lines coordinates. Thus, the players projected coordinates can be easily filtered around the field area. Hence, the off-field detections are removed before being forwarded to the tracking algorithm.

3.3.2 DeepSORT

The tracking algorithm integrated with the solution is DeepSORT (Wojke et al., 2017). Despite sporadic accuracy issues, it is one of the most suitable tracking algorithm for the time-quality balance among the state-of-the-art.

This algorithm contains a motion and a deep association metric. The proposed solution uses an OSNet, trained for person re-identification, to extract the features needed for the deep association metric.

3.4 Metadata Fusion

This module is responsible for receiving the cameras' metadata and merging it into a single, complete output. It is also responsible for assigning the tracked players to their teams, creating visual representations and storing the information created for other possible objectives.

To fuse both cameras' information, the detections are filtered after mapping the players' position into the top view according to their field side - the left camera detects the players on the left side and right on the right side. This strategy requires the definition of the midfield line pixels, in the top view representations, as a separator of the cameras' input.

3.4.1 Team Labelling

The players' positional representations must illustrate their respective teams, which are recognised through their outfit's colour. Thus, the objective is to analyse the bounding boxes' colours and distribute them to the teams. For this, a K-means clustering algorithm allows the aggregation of team players based on their outfit colours. As the field players' colours are the most meaningful to differentiate, the number of clusters is two. Regularly, as the players are in the centre of the bounding boxes, the value used to represent them is the average of the pixels' colour of a 5x5 centred area. The model is built with the colour values of the bounding boxes from both cameras' first frame. After fitting the model, the following frames' bounding boxes' colour values are fed to the model to predict its cluster and, consequently, its team. After predicted, the colour is the same throughout the tracking ID lifetime. The team labelling algorithm operates within the fusion module as the k-means model should be the same for both cameras.

3.4.2 Visual Representations

Three types of visual representations are defined for the proposed solution based on the players' positional information, as shown in Figure 3. The first two provide a spatial-temporal representation of the players' position at a given moment. However, the second also represents the past positions of that player, giving it a notion of direction. The third is a heat map that indicates the field areas with more presence throughout the game.

These visual representations can be used by the media companies as screen overlays to enrich the broadcasting of a football match.

4 RESULTS

The results of the proposed solution are promising, while unveiling technical and scientific challenges that need to be addressed in the future. The proposed solution can perform player detection and according representation, namely the positional plot and heatmap, in real-time. However, due to hardware limitations, the system can not perform online tracking.

To evaluate the overall system behaviour, Figure 4 illustrates the number of tracked players per frame for the right, left and fused outputs, during a significant time span video ($\tilde{5}$ minutes) using the setup described in Figure 2. The setup employed two cameras with different lens exposures, which changes the image's brightness. This will prove significant in the further presented results. The usage of two different cameras enabled testing the system's behaviour facing various image qualities and delimiting a minimum quality threshold. The cameras used were CANON's 250D and 700D, where the former presents better image quality leading to better performance. Figure 4 shows many variations in the detection and tracking models. Overall, the number of tracks is around 22 persons, which is a good indicator. However, when the left camera (CANON 250D) has fewer detections, the right camera (CANON 700D) can not cope with the players on its side, decreasing the total number players tracked. This is a consequence of a video quality discrepancy between cameras, since the right

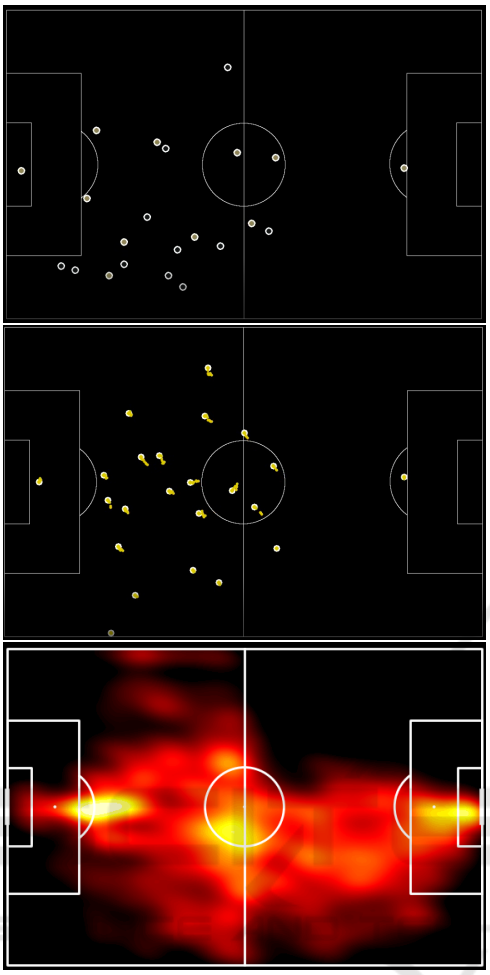

Figure 3: Visual Representations.

camera quality is significantly lower than the left (see Figure 5 and 6). In fact, all results are affected by this limitation, which will be taken into account for future data acquisition experiments (with CANON's 250D image quality as the minimum threshold).

The main consequence of these variations is the number of ID switches. DeepSORT was selected based on its ability to fight this problem. In light of these results, the deep metric used may be less relevant than expected. The auto-encoder used for this metric is explicitly trained for re-identification, so this is presumably not the faulty step. The reason may be the distance between the players and the camera. Despite the camera placement advantages relative to the field coverage, the players become small, leading to less significant features, which may affect the deep metric precision. Another aspect affecting the number of ID switches is the players switching midfields.

Another cause for variations are occlusions, illustrated in the top sequence of Figure 6. These are hard to oppose with a single view. However, as shown in the bottom sequence of Figure 6, the opposite camera can detect the previously occluded players, which is an advantage of the proposed setup. Therefore, future work may include the usage of both cameras to overcome this phenomenon.

Regarding the current fusion approach, even though the detection works suitably, some loss of information may occur. As Figure 5 shows, if one camera detects the players that the other does not, if on the opposite midfield, these detections are discarded.

Team labelling has satisfactory results. It assigns the expected colours for a high number of tracks, with occasional errors, specially when noise is in the bounding box.

5 DISCUSSION AND FUTURE WORK

As previously mentioned, the presented solution is a work-in-progress in the scope of MOGSports and DataCloud. In this section are discussed some of the positive and negative aspects of the implemented solution and of the next steps. The current solution encompasses modules that perform field registration, player detection and tracking, team labelling and fusion. Manual registration is a straightforward method for computing the homography matrix, which benefits from static cameras. Despite being user-friendly for few cameras, automation is a project goal and requirement for scalability. Accordingly, an automated approach is being implemented and tested. The Two-GAN model (Chen and Little, 2019) is used for field line segmentation. Then, the solution attempts a feature-based approach, comparing the top-view and the segmented image. Along with the tests conducted until now, the Two-GAN line segmentation is expected to improve with more training data, primarily similar to the ones acquired in the proposed setup.

The Player Detection model presents promising results, still requiring further fine-tuning. Players' unusual positions or fast changing movements increase the difficulty of accurate detections. Furthermore, the many occlusions during a match lead to unstable positional information retrieval and consequent unstable visual representations.

Future work is expected to improve the used YOLOv5 model performance to detect the players and add the capacity to detect the ball resorting to the acquired dataset.

As for the Player Tracking module, due to hardware constraints, DeepSORT is not yet thoroughly tested for real-time processing. Nonetheless, an unwanted ID switching issue has been identified. Conse-

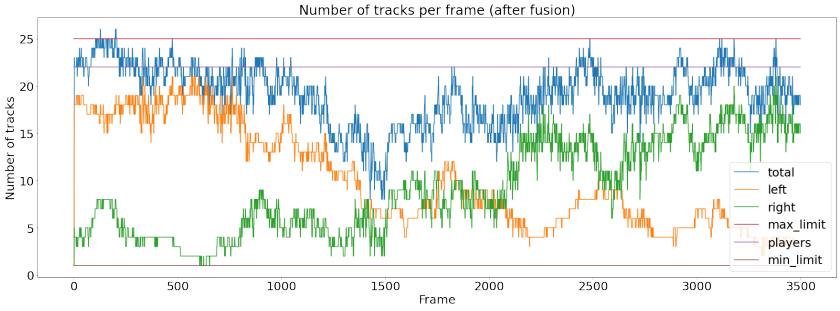

Figure 4: Number of tracks per frame, fused tracks (blue), left camera (orange), right camera (green). Results obtained with reduced sample (around 3 minutes) of a custom dataset with proposed camera setup.

Figure 5: Player tracking example for a frame inside the interval [1400-1600] in each camera (cropped image). Blue: Predicted tracks. Red: Miss detections manually annotated.

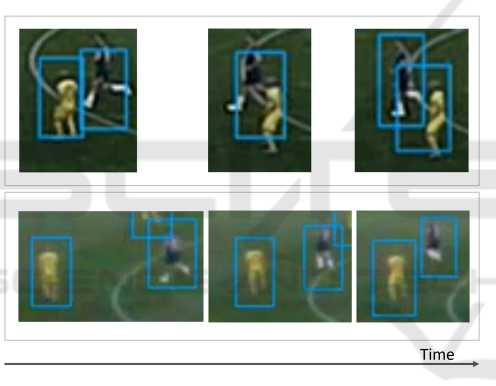

Figure 6: Camera Setup advantages for occlusion handling. Frame sequences for both cameras representing the same moment, Top: Left camera sequence. Bottom: Right camera sequence.

quently, the algorithm can suffer some modifications. As presented in Section 4, the deep metric may have a lower impact than initially thought in the association step. Team information could be a valuable metric for this step, neglecting possible matches with players from the opposite team. This strategy would require a very robust team labelling model.

Alternatively, tracking can occur after data fusion, resorting to algorithms like graph-based tracking, using the transformed top-view detection coordinates. This strategy would demand a robust fusion algorithm as a previous step.

Despite being accurate for numerous players, team labelling also has limitations. As players move rapidly, their bounding box may temporarily not fit its body. Hence, background or other artefacts can influence the colour value calculation. In this sense, the method could be improved by assigning each player to the team associated with the mode of a periodic check prediction or of the N initial frames predictions. Furthermore, to counter the cluster issue, as the goalkeepers have a generally fixed position, their colour could be assigned by it.

Fusion is critical, and a scalable solution is necessary for this module. The cameras provide a good overlap area that can be used for occlusion handling and to complete miss detections from one another. Incorporating a matching algorithm into this module could leverage the full potential of the proposed camera setup. One possible strategy would be associating the players from each camera in each frame. This method could consider all players, calculate the euclidean distance between the transformed points, and then associate them optimally. Being a computationally demanding task, other parameters, such as the team label, could be used to reduce the distances calculations and consecutive matching. The quality discrepancy between cameras used for the data acquisition affects the results negatively. Consequently, the debated problems may be softened when considering a dataset with equally high-quality images.

Event Detection is a comprehensive problem that MOGSports intends to solve, which reflects the culmination of the developments from the previous modules and its valuable metadata extracted. It may comprise match highlights (e.g. goals), match events (e.g. passes, offsides), as well as statistic insights.

6 CONCLUSION

In response to audience demand, user experience enhancement is an imperative action for the sports broadcasting industry. This paper presented a step forward to accomplish this enhancement, based on players' positional data extraction and representation. Extracting this data has complex associated problems and the results reflect this complexity. However, there are one or more proposed alternatives for each problem encountered. The proposed solution reflects ongoing work. The current solution can be integrated into sports broadcasting. Future work includes the overlay of specific game-related events.The results achieved so far create good expectations for the upcoming developments and for the potential of the MOGSports business case within the media and entertainment industries.

ACKNOWLEDGEMENTS

This work has been supported by the H2020 DATACLOUD project, funded by European Union's Horizon 2020 research and innovation programme under grant agreement No 101016835, as well as by the National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project UIDB/50014/2020.

REFERENCES

Chen, J. and Little, J. J. (2019). Sports camera calibration via synthetic data. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2497–2504.

Cioppa, A., Deliege, A., Magera, F., Giancola, S., Barnich, O., Ghanem, B., and Van Droogenbroeck, M. (2021). Camera calibration and player localization in soccernet-v2 and investigation of their representations for action spotting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4537–4546.

Cuevas, C., Quilon Gonzalez, D., and García, N. (2020). Techniques and applications for soccer video analysis: A survey. *Multimedia Tools and Applications*, 79.

Ivasic-Kos, M., Host, K., and Pobar, M. (2021). Application of deep learning methods for detection and tracking of players. In *Deep Learning Applications*. IntechOpen.

Jocher, G., Chaurasia, A., Stoken, A., Borovec, J., NanoCode012, Kwon, Y., TaoXie, Fang, J., imyhxy, Michael, K., Lorna, V, A., Montes, D., Nadar, J., Laughing, tkianai, yxNONG, Skalski, P., Wang, Z., Hogan, A., Fati, C., Mammana, L., AlexWang1900, Patel, D., Yiwei, D., You, F., Hajek, J., Diaconu, L., and Minh, M. T. (2022). ultralytics/yolov5: v6.1 - TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference.

Nguyen, T., Chen, S. W., Shivakumar, S. S., Taylor, C. J., and Kumar, V. (2018). Unsupervised deep homography: A fast and robust homography estimation model. *IEEE Robotics and Automation Letters*, 3(3):2346–2353.

Nie, X., Chen, S., and Hamid, R. (2021). A robust and efficient framework for sports-field registration. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1935–1943.

Rematas, K., Kemelmacher-Shlizerman, I., Curless, B., and Seitz, S. (2018). Soccer on your tabletop. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4738–4747.

Sah, M. and Direkoglu, C. (2021). Review and evaluation of player detection methods in field sports: Comparing conventional and deep learning based methods. *Multimedia Tools and Applications*.

Theiner, J., Gritz, W., Müller-Budack, E., Rein, R., Memmert, D., and Ewerth, R. (2022). Extraction of positional player data from broadcast soccer videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 823–833.

Wojke, N., Bewley, A., and Paulus, D. (2017). Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE.