Multi-Person Tracking By Multi-Scale Detection in Basketball Scenarios

Adria Arbués-Sanguesa
Universitat Pompeu Fabra

Gloria Haro
Universitat Pompeu Fabra

Coloma Ballester
Universitat Pompeu Fabra

Multi-Person tracking by multi-scale detection in Basketball scenarios

Adrià Arbués-Sangüesa[1,2], Gloria Haro[1], Coloma Ballester[1]

[1] *Universitat Pompeu Fabra (Barcelona, Spain) and* [2] *FC Barcelona*

Abstract

Tracking data is a powerful tool for basketball teams in order to extract advanced semantic information and statistics that might lead to a performance boost. However, multi-person tracking is a challenging task to solve in single-camera video sequences, given the frequent occlusions and cluttering that occur in a restricted scenario. In this paper, a novel multi-scale detection method is presented, which is later used to extract geometric and content features, resulting in a multi-person video tracking system. Having built a dataset from scratch together with its ground truth (more than 10k bounding boxes), standard metrics are evaluated, obtaining notable results both in terms of detection (F1-score) and tracking (MOTA). The presented system could be used as a source of data gathering in order to extract useful statistics and semantic analyses *a posteriori*.

Keywords: Multi-Person Detection, Basketball, Tracking, Pose Models, Single-Camera.

1 Introduction

Right in the *Big Data Era*, sports teams are gathering a lot of advanced statistics about players to make the appropriate decisions that may lead to a boost of performance; for instance, hiring a new player or designing optimal tactics for a specific game. In particular, in the basketball field, a strong demand for tracking data has emerged, and video-based companies such as Second Spectrum or STATS extract this information with a set of overhead cameras placed at the ceiling of the stadiums. However, at the moment, there is not an established implemented way to infer tracking statistics from simple single-camera video sequences, due to the many challenges that emerge, such as multiple occlusions, similarity in appearance or fast and erratic motions [Thomas et al., 2017]. Besides, the accurate detection and tracking of persons in a video is a pivotal issue for the automatic analysis and understanding of the actions happening in the scene captured by the camera.In this article, a novel multi-tracker method thought for single-camera basketball sequences is presented, where all targets are properly tracked with 0.67 MOTA confidence; sample results are shown in Figure 1.

2 State of the Art

There has been a considerable effort towards multi-object tracking in video. Frequently, the tracking problem is approached by a previous or simultaneous detection step. Person detection and tracking is used in [Ramanathan et al., 2016] to detect events in multi-person videos. The authors propose to use a CNN-based multibox detector [Szegedy et al., 2013] and a KLT tracker [Veenman et al., 2001]. A human pose estimation method that extends the Mask R-CNN [He et al., 2017] to the video case is presented in [Girdhar et al., 2018], which includes a tracking cost based on three terms: Intersection over Union (IoU) of the bounding boxes, a pose similarity metric, and a similarity metric that uses CNN features. An optimization method for the joint segmentation and tracking of scenes with multiple targets is proposed in [Milan et al., 2015]. The assignment of part of object (or superpixel) detections to trajectory hypothesis is formulated as a multi-label
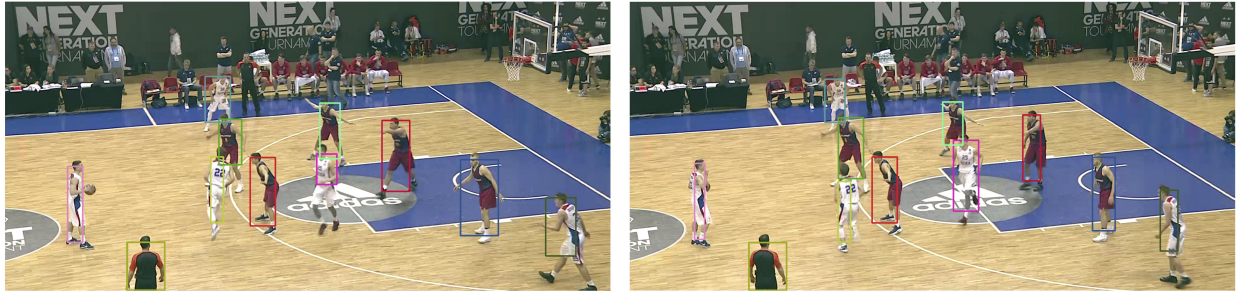
Figure 1: Obtained results in adjacent frames, where all players (and referee) in court are properly detected (bounding box) and tracked (color identifier).

conditional random field. [Henschel et al., 2018] propose a method that uses two different detectors, namely, a full-body detector and a head detector. All detections are incorporated in a global optimization formulation which translates in a binary quadratic problem that is solved by relaxation through the Frank-Wolfe algorithm. [Doering et al., 2018] propose a method based on a Siamese network that evaluates the person pose in two frames at a time and a temporal CNN that predicts the so-called Temporal Flow Fields, a graph optimization problem to obtain the tracking. Pose tracking refers in the literature to the task of estimating anatomical human keypoints and assigning unique labels for each keypoint across the frames of a video [Iqbal et al., 2017, Insafutdinov et al., 2017]. Our method for detection and tracking estimates the human pose in each frame and uses the similarity of the detected pose keypoints to reinforce the tracking of a given person. Our focus is on team sports, and in particular, basketball games. For a thorough revision of computer vision techniques applied to the analysis of sports (including players detection and tracking) we refer the reader to the relatively recent survey [Thomas et al., 2017]. In a recent work and for the basketball case, [Senocak et al., 2018] propose a deep learning based method for player identification that incorporates features on body parts which are in turn computed with Convolutional Pose Machines [Wei et al., 2016].

3 Proposed method

Our proposal to track players in a basketball game is based on a tracking-by-detection approach as seen in Figure 2: First, the different individuals on the court are detected in every frame and then, matches are established along time. Given a video of a basketball game, the prior that players are inside the court along the sequence is used to restrict the area where players should be searched, thus avoiding the detection of spectators or bench players. Once this region is detected, players on court are found and, having stabilized the video sequence, geometric and content features are extracted for every single detected instance. Having matched the different detections from adjacent frames, the tracking of players along the video is obtained.
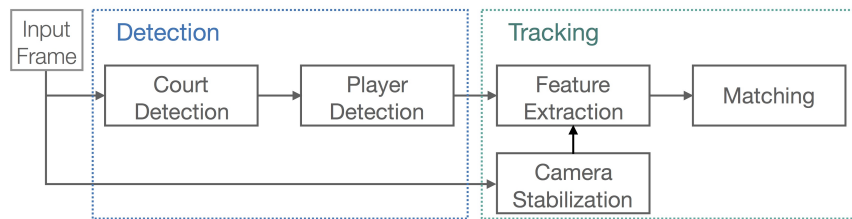

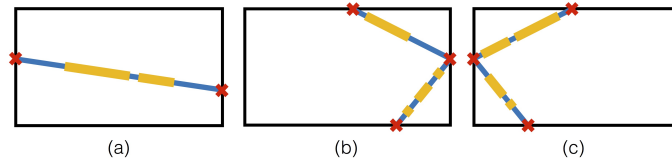Figure 2: Overall pipeline of the presented method.

Figure 3: Line contributions. (a) potential sidelines to be detected. (b)-(c) right-left baselines, respectively.

3.1 Court Detection

The court is a rectangular area whose projection to the camera results in a trapezoid. Thus, the problem is reduced to the identification of visible court boundaries in the image: the sidelines and baselines (from 1 to 4 depending on the camera the point of view). Frequently, some of these court boundaries are only partially visible due to occlusions produced by, *e.g.*, players, referees, the public and the synthetic scoreboard.

The method starts by detecting all the line segments in the image using a fast and robust parameter-less method [Grompone Von Gioi et al., 2010]. Right after, dominant lines, *i.e.* lines the with longest visible parts, are estimated using a voting procedure. Those lines will correspond, in general, to the sidelines/baselines or, in cases of strong occlusions, to court lines parallel to the sidelines/baselines. The strength of the vote of each line is proportional to the sum of detected segments' length on the line, as seen in Fig. 3, where the detected segments are shown in yellow. Given that in broadcasting sequences only one baseline (or none) can be seen at a time and that even in cases that both sidelines are in the field of view of the camera one of them may appear occluded by the public (*e.g.* Fig. 4), the purpose is to find a *horizontal* dominant line (either a sideline or its orientation) and a *vertical* dominant one (a baseline or its orientation). Horizontal lines are considered to be the ones which intersect the image at the left and right boundaries (Fig. 3(a)), while vertical ones intersect in one of the following pairs of image sides: top-left, bottom-left, top-right or bottom-right (examples in Fig. 3(b)-(c)). In order to find the location of court boundaries, the court is pre-segmented and the set of lines (with the orientation of the horizontal or vertical dominant lines) that better delimits the pre-segmented court is selected. Two different solutions are proposed to pre-segment the court in two different professional basketball scenarios: (a) European, and (b) NBA games.

In European games (Fig. 4-right), court surroundings usually share the same color, and fans sit far from team benches. For this reason, a basic color filter (in the HSV colorspace) can be created; for each possible line candidate, the contribution of pixels that satisfy filter conditions is checked at both right-left (vertical) or above-below (horizontal) sides of the tested candidate. The horizontal and vertical candidates with the highest response in terms of difference will be then considered as court limits.

For NBA games (Fig. 4-left), the scenario is much more challenging, because there is almost no space between sidelines and fans. In order to find the horizontal boundaries, instead of checking for color components, Conditional Random Fields [Zheng et al., 2015] is applied at a coarse resolution to find the total area of people pixels; this estimation is later complemented with Histogram of Oriented Gradients. Once having this rough estimation, an iterative algorithm is applied to delimit court boundaries: at the very beginning, two line candidates with the dominant orientation are placed at the top and bottom of the image; then, for each iteration, these lines are moved towards the middle until convergence. In each iteration, the product of the following percentages is computed: (a) people-pixels above the top line, (b) people-pixels below the bottom line, and (c) non-people-pixels below the top line and above the bottom one. If there is a drop in the first or second percentage, the position of the corresponding line is fixed; convergence is reached when both lines stop moving. Potentially, in the horizontal court limits, the product of these three terms will correspond to a maximum, meaning that there is a large contribution of people pixels above and below the top and bottom line respectively (corresponding to fans), and a small contribution in between (corresponding to the court with a maximum of 10 players plus 3 officials). Having masked the original image given the detected sidelines, the best vertical candidate is found in the same way but scanning only from left to right / right to left.

Figure 4: Court detection results in different scenarios: (left) NBA, and (right) European games
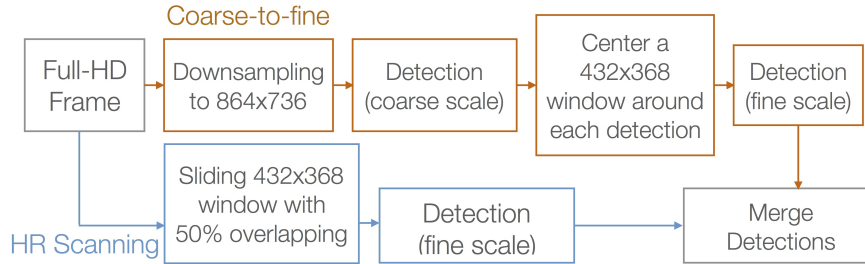

Figure 5: Proposed multi-scale detection strategy.

3.2 Player Detection

Our detection method is based on a robust strategy that relies on pose models techniques [Ramakrishna et al., 2014, Wei et al., 2016, Cao et al., 2017]. Stemming from a TensorFlow implementation of OpenPose[1] thought for low-powered embedded devices, performance on high resolution videos is improved by a multi-scale strategy that refines the detection at the original training scale. More precisely, the method of [Cao et al., 2017] is a bottom-up approach that identifies up to 17 anatomical keypoints (corresponding to parts of the face, body, etc) and joins them in limbs connecting two keypoints and, finally, in the visible person skeleton. It consists on a CNN that incorporates features given by a VGG-19 network and obtains part confidence maps and a nonparametric representation which is referred to as Part Affinity Fields encoding location and orientation of the limbs. The authors use the part confidence maps together with the part affinity fields to extract the person skeleton.

The model being used in the above-mentioned OpenPose implementation was trained with a MobileNet network, with a default image resolution of 432×368 pixels, and it has been experimentally checked that best results are obtained at that resolution. With the purpose of maximizing the accuracy of human pose detection in our full-HD frames of 1920×1080 pixels, the following multi-scale strategy (Figure 5) is proposed:

1A: The frame is downscaled to twice the resolution of the pretrained model images, and pose estimation is applied. Most of the players are detected but their pose is not accurate enough due to the loss of resolution.

1B: Detections at the coarse scale are refined by centering a 432×368 window around the center of each former detection in the original HR image. The human pose is better estimated and new detections may emerge, specially those corresponding to players partially occluded by other players, with only small visible areas.

2: To include those players who have not been detected at the first low-resolution stage (typically corresponding to players with motion blur who even fade away more after downscaling), the detection is carried out with a sliding window of size 432×368 over the full-HD image, with 50% overlapping, both horizontally and vertically. Detections not found at the current stage are added to the former ones.

Notice that the presented person detection method does not include priors such as a maximum number of players in the basketball court or information about the team uniforms and, thus, the set of detected persons might include some referees (an example is displayed in Fig. 1). The detection output is a bounding box placed in the player skeleton position (Fig. 6 shows some examples), and a heatmap of 18 layers (one for each part,

---

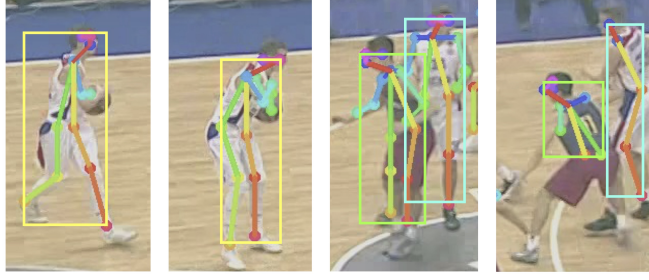[1] https://github.com/ildoonet/tf-pose-estimation

Figure 6: Detected parts with the corresponding bounding box.

plus their combination) for each detection, indicating the confidence of each part being at each particular pixel.

3.3 Multi-Person Tracking

Before tracking the individual players we propose to remove the camera motion by using a camera stabilization method [Sánchez, 2017], which outputs a set of homographies $\{H_t\}_t$, being H_t the matrix that stabilizes the frame at time t. For this reason, it can be assumed that, in general, the player motion within frames is small.

In order to track the bounding boxes corresponding to the players (or the detected part of them, as seen in Figures 1 and 6), a similarity cost is defined, which is made of three terms: two geometric terms and a content-based one. Let us denote by I_{t_1} and I_{t_2} two frames of the input video. Stabilized versions are considered by H_{t_1} and H_{t_2}, respectively. Given two bounding boxes B_{t_1} and B_{t_2}, detected in I_{t_1} and I_{t_2}, respectively, the proposed similarity cost is defined as:

$$C(B_{t_1}, B_{t_2}) = \alpha C_d(B_{t_1}, B_{t_2}) + \beta C_i(B_{t_1}, B_{t_2}) + \gamma C_c(B_{t_1}, B_{t_2}), \tag{1}$$

where $\alpha, \beta \in [0,1]$, $\gamma = 1 - (\alpha + \beta)$, and C_d, C_i, C_c in Equation (1) are defined as follows:
(a) $C_d(B_{t_1}, B_{t_2})$ is the normalized distance between the transformation by H_{t_1} and H_{t_2}, respectively, of the centroids of the bounding boxes, $\mathbf{x}_{B_{t_1}}$ and $\mathbf{x}_{B_{t_2}}$. That is, $C_d(B_{t_1}, B_{t_2}) = \frac{1}{\sqrt{w^2+h^2}} \|H_{t_1}(\mathbf{x}_{B_{t_1}}) - H_{t_2}(\mathbf{x}_{B_2})\|$, where w and h are the width and the height of the frame domain.
(b) $C_i(B_{t_1}, B_{t_2}) = \text{IoU}(H_{t_1}(B_{t_1}), H_{t_2}(B_{t_2}))$ is the Intersection over Union (IoU) value of the transformation by H_{t_1} and H_{t_2} of the two bounding boxes, respectively.
(c) The content of B_{t_1} and B_{t_2} is compared by considering only the pairs of anatomical keypoints (joints between limbs) present or detected in both B_{t_1} and B_{t_2}, denoted here as $\mathbf{p}_1^k$ and $\mathbf{p}_2^k$, respectively. The joint $\mathbf{p}_i^k$ is obtained by finding the maximum value in the associated k-th heatmap of each bounding box B_{t_i}. For all parts detected both in B_{t_1} and B_{t_2}, the color and texture content in a neighborhood around these two keypoints are compared. Let $\mathscr{E}$ be a squared neighborhood of 24×24 pixels centered at $\mathbf{0} \in \mathbb{R}^2$. Then,

$$C_c(B_{t_1}, B_{t_2}) = \frac{1}{255|S||\mathscr{E}|} \sum_{k \in S} \sum_{\mathbf{y} \in \mathscr{E}} \|I_{t_1}(\mathbf{p}_1^k + \mathbf{y}) - I_{t_2}(\mathbf{p}_2^k + \mathbf{y})\| \tag{2}$$

where S in Equation (2) denotes the set of mentioned pairs of corresponding keypoints detected in both frames. In the presented experiments, $\alpha = 0.65$, $\beta = 0.05$, $\gamma = 0.3$.
To establish the matching assignments between bounding boxes along time, a memory criterion is used to set back on track those targets that have not been detected in particular single frames. For each frame t, the values of $C(B_t, B_{t-1})$ and $C(B_t, B_{t-2})$ are considered (for both frames' bounding boxes) and, by using a variant of the Hungarian algorithm, the bounding boxes of the current frame are assigned both to the ones detected in the $t-1$ and $t-2$. Those boxes that have only been detected in either $t-1$ or $t-2$ will be directly associated to the respective candidates, while the ones that have been properly found in both previous frames will be associated to the assignment that minimizes the overall matching cost. The output of this computation is an existing unique identifier for each matched bounding box, and a new identifier for boxes that could not be assigned.

	Precision	Recall	F1-Score
Coarse-to-Fine	**0.9959**	0.8095	0.8923
High-Res. Scanning	0.9947	0.8109	0.8926
YOLO	0.8401	**0.9426**	0.8876
Proposed Method	0.9900	0.8563	**0.9178**

Table 1: Detection performance: Average precision, recall and F1-Score of each strategy, over all sequences.

		MOTA	MOTP
no mem.	Coarse-to-fine	0.5837	0.6109
	High-Res. Scanning	0.5711	0.5729
	Proposed Combination	0.6237	0.6086
mem.	Coarse-to-fine	0.6259	0.6110
	High-Res. Scanning	0.6134	0.5871
	Joint Track. + Segm.	**0.7142**	0.3375
	Proposed Combination	0.6704	**0.6138**

Table 2: Tracking and memory performance.

4 Results

This section focuses on the quantitative and qualitative evaluation of the proposed method. An ablation study assessing the contribution of each of the ingredients is included.

Dataset: A dataset of 22 Basketball sequences has been built from several European basketball games corresponding to the *Final Four* of the *ANGT 2017*. The sequences have been directly extracted from the broadcasting video by manually setting the beginning and end times, hence avoiding having different camera shots; the original videos have full-HD resolution (1920 × 1080) and 25 fps, but for the purposes of this experiment, only 4 frames are extracted per second, thus reducing computational expenses. The content of these sequences involve many different basketball offensive plays, such as isolation, *pick and roll*, or even sideline strategies; besides, all 22 sequences include different jersey colors and different color-skinned players. The main limitation of this dataset is that it does not contain offensive transitions where players run from one side of the court to the other one, as camera stabilization is not able to handle this situation. The mean duration of these sequences is 11.07 seconds, resulting in a total of 1019 frames. A ground truth has been manually generated with the by dragging bounding boxes over each player and all 3 referees (taking the minimum visible X and Y coordinates of each individual) in every single frame (when visible), which results in a total of 11339 boxes.

Quantitative results: Quantitative assessment of both the proposed detection and tracking methods is provided in Tables 1 and 2. Table 1 focuses on evaluating the proposed detection strategy; in this context, bounding boxes assignments are computed in each frame from scratch according to the maximum (and non-zero) IoU between ground truth boxes and current annotations, thus leading to TPs, FPs, FNs and TNs. Besides, a comparison with the state-of-the-art YOLO network [Redmon et al., 2016] is given; for a fair comparison, only the *person* detections within the court boundaries are kept. The tracking results are discussed in Table 2, showing widely used error metrics: MOTA, which takes into account false positives, misses and missmatches; and MOTP, which measures detection distances (details in [Bernardin and Stiefelhagen, 2008]). To provide an ablation study of each of the contributions, results for the three stages described in Section 3.2 are included: the coarse-to-fine strategy of stages 1 and 1A, the sliding window method on the original (full-HD) video frames described in stage 2, and their combination. Moreover, tracking results have been compared with [Milan et al., 2015] as a State-of-the-Art method; in all tests, our detections have been used, thus starting off with the same conditions. It has to be mentioned that results obtained with the proposed tracking method were obtained only with CPU usage, whilst the *Joint Tracking + Segmentation* method had to be run on a High Performance Cluster.

On the one hand, Table 1 shows that the first two individual strategies have high precision values, but those suffer a drop in recall that can be compensated by merging detections. This improvement indicates that non-detected players (false negatives) are not the same ones when using the sliding window method than when implementing the coarse-to-fine approach; as mentioned, the second method deals better with motion blur. Comparing with YOLO detections, it can be seen that lower recall is obtained, but it comes at a cost of precision; while our method does not detect some players, YOLO detects some non-existing players. This trade-off is compensated with the F1-score, which shows that the proposed method outperforms the SoA network.

On the other hand, tracking metric results in Table 2 lead to several conclusions. First, it is proved that introducing memory (with a tolerance of only 2 frames) in the matching procedure increases the MOTA metric by

5%. Second, the combination of approaches results in less missed players/frame (less false negatives), thus providing better MOTA results. It can also be seen that the performance in terms of MOTP, which considers the IoU between matched instances, is better when using the full-HD scan technique. As seen in Figure 6, the bounding boxes surrounding the human skeleton given by the pose model do not include the top-bottom parts of the human body, such as the forehead or feet; besides, the area of the bounding boxes depends on how many human parts have been found. As mentioned, the coarse-to-fine approach is better than the full-HD scanning at detecting partially occluded players, which are the most likely ones not to have all parts detected, thus resulting in notable area changes. Finally, better MOTA results are obtained with the *Joint Tracking + Segmentation* method, but these come at a cost of MOTP for a main reason: their method performs segmentation inside every bounding box jointly with tracking, but our boxes contain really challenging situations, where players have random and stretched poses (plus occlusions), thus resulting in a poor box refinement and a MOTP drop. For this reason, the SoA method is an optimal tracker in constrained situations, where targets' bounding boxes contain almost no background and segmentation barely changes their size (*i.e* scenes where the camera is further away than in basketball broadcasting sequences). Note that all performed tests have been done at 4 fps due to GT limitations; it is logic to hypothesize that with higher frame rates, MOTA results would improve due to closer proximity of targets, but it would come at a cost of computational expenses (and vice versa with lower rates).

5 Conclusions

In this article, a novel basketball player tracker based on multiple detections at different scales has been detailed. The proposed method is thought for single-camera systems and low-powered embedded devices, and uses – once the court area is segmented – a customized version of existing pose models, combining a coarse-to-fine approach and a sliding window technique at full resolution. Contextual and geometric features are then extracted for every single detection, and matched across consecutive frames. Having gathered a dataset from scratch, and having labelled thousands of ground truth bounding boxes, detection and tracking results can be independently obtained. Results are encouraging in both cases, proving that this technique could be used as the basis of a data analysis system for professional basketball teams. As future work, fast basketball transitions should be included in the dataset, and GPU-based pose models' performance should be tested with this multi-scale method together with other quantitative assessment tests; with more computational resources, deep learning features (*i.e* output of a Convolutional Layer) could be used instead of contextual ones.

Acknowledgments

References

[Bernardin and Stiefelhagen, 2008] Bernardin, K. and Stiefelhagen, R. (2008). Evaluating multiple object tracking performance: the clear mot metrics. *Journal on Image and Video Processing*, 2008:1.

[Cao et al., 2017] Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1302–1310.

[Doering et al., 2018] Doering, A., Iqbal, U., and Gall, J. (2018). Joint flow: Temporal flow fields for multi person tracking. *arXiv preprint arXiv:1805.04596*.

[Girdhar et al., 2018] Girdhar, R., Gkioxari, G., Torresani, L., Paluri, M., and Tran, D. (2018). Detect-and-track: Efficient pose estimation in videos. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 350–359.

[Grompone Von Gioi et al., 2010] Grompone Von Gioi, R., Jakubowicz, J., Morel, J.-M., and Randall, G. (2010). Lsd: A fast line segment detector with a false detection control. *IEEE transactions on pattern analysis and machine intelligence*, 32(4):722–732.

[He et al., 2017] He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn. In *IEEE International Conf. on Computer Vision*, pages 2980–2988.

[Henschel et al., 2018] Henschel, R., Leal-Taixé, L., Cremers, D., and Rosenhahn, B. (2018). Fusion of head and full-body detectors for multi-object tracking. In *IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, pages 1509–150909.

[Insafutdinov et al., 2017] Insafutdinov, E., Andriluka, M., Pishchulin, L., Tang, S., Levinkov, E., Andres, B., and Schiele, B. (2017). Arttrack: Articulated multi-person tracking in the wild. In *IEEE Conf. on Computer Vision and Pattern Recognition*, volume 4327.

[Iqbal et al., 2017] Iqbal, U., Milan, A., and Gall, J. (2017). Posetrack: Joint multi-person pose estimation and tracking. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2011–2020.

[Milan et al., 2015] Milan, A., Leal-Taixé, L., Schindler, K., and Reid, I. (2015). Joint tracking and segmentation of multiple targets. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 5397–5406.

[Ramakrishna et al., 2014] Ramakrishna, V., Munoz, D., Hebert, M., Andrew Bagnell, J., and Sheikh, Y. (2014). Pose Machines: Articulated Pose Estimation via Inference Machines. In *IEEE European Conf. Computer Vision*, pages 33–47.

[Ramanathan et al., 2016] Ramanathan, V., Huang, J., Abu-El-Haija, S., Gorban, A., Murphy, K., and Fei-Fei, L. (2016). Detecting events and key actors in multi-person videos. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3043–3053.

[Redmon et al., 2016] Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788.

[Sánchez, 2017] Sánchez, J. (2017). Comparison of motion smoothing strategies for video stabilization using parametric models. *Image Processing On Line*, 7:309–346.

[Senocak et al., 2018] Senocak, A., Oh, T.-H., Kim, J., and Kweon, I. S. (2018). Part-based player identification using deep convolutional representation and multi-scale pooling. In *IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, pages 1732–1739.

[Szegedy et al., 2013] Szegedy, C., Toshev, A., and Erhan, D. (2013). Deep neural networks for object detection. In *Advances in neural information processing systems*, pages 2553–2561.

[Thomas et al., 2017] Thomas, G., Gade, R., Moeslund, T. B., Carr, P., and Hilton, A. (2017). Computer vision for sports: current applications and research topics. *Computer Vision and Image Understanding*, 159:3–18.

[Veenman et al., 2001] Veenman, C. J., Reinders, M. J., and Backer, E. (2001). Resolving motion correspondence for densely moving points. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (1):54–72.

[Wei et al., 2016] Wei, S. E., Ramakrishna, V., Kanade, T., and Sheikh, Y. (2016). Convolutional Pose Machines. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 4724–4732.

[Zheng et al., 2015] Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., and Torr, P. (2015). Conditional Random Fields as Recurrent Neural Networks. *arXiv.org*, cs.CV:2015.