Tracking Handball Players with the DeepSORT Algorithm

Kristina Host[a], Marina Ivašić-Kos[b] and Miran Pobar[c]
Department of Informatics, University of Rijeka, Rijeka, Croatia

Keywords: Object Detection, YOLO, Sport, Handball, Tracking, DeepSORT.

Abstract: In team sports scenes, such as in handball, it is common to have many players on the field performing different actions according to the rules of the game. During practice, each player has their own ball, and sequentially repeats a particular technique in order to adopt it and use it. In this paper, the focus is to detect and track all players on the handball court, so that the performance of a particular athlete, and the adoption of a particular technique can be analyzed. This is a very demanding task of multiple object tracking because players move fast, often change direction, and are often occluded or out of the camera field view. We propose a DeepSort algorithm for player tracking after the players have been detected with YOLOv3 object detector. The effectiveness of the proposed methods is evaluated on a custom set of handball scenes using standard multiple object tracking metrics. Also, common detection problems that have been observed are discussed.

1 INTRODUCTION

Handball, along with football and basketball, is one of the most represented team sports in Europe. For that reason, it would be useful to develop a system that performs a complete analysis of players' movements and actions. The performance of players can be greatly enhanced by tracking their movements. Therefore, object detection and tracking in a sports environment are rapidly gaining importance.

Just by looking at an image or a video, humans can instinctively detect and track an object in it and differentiate between them. Computers, on the other hand, need human guidance in order to learn to do so themselves. For that reason, numerous algorithms for detection and tracking objects have been created.

For the analysis of player activities, a player must be located concerning its environment and other players, as well as tracked in time. For detecting the players, an object detection algorithm such as YOLO (Redmon & Farhadi, 2018) can be used on each frame. After having been detected, a player can be tracked and labeled with a corresponding ID.

Many approaches for object tracking have been proposed, but lately, due to advances in object detection, tracking-by-detection has become more prevalent (Ciaparrone et al., 2019). One of such algorithms, The Simple Online and Realtime Tracking (SORT) was proposed by Bewley, Ge, Ott, Ramos, & Upcroft (2016), which performed favorably in comparison with other tracking algorithms such as TDAM(Yang, & Jia, 2016), MDP(Xiang, Alahi & Savarese, 2015), SMOT (Dicle, Camps & Sznaier, 2013). An extension of that algorithm, SORT with deep association metric (Deep SORT) was proposed in (Wojke, Bewley & Paulus, 2017) and tuned for pedestrian detection. In Burić, Ivašić-Kos & Pobar (2019), a preliminary qualitative evaluation of these algorithms on the sports domain was considered. The DeepSORT achieved the most stable tracking results, so it is tested here using different metrics on the task of tracking handball players.

The paper is organized as follows: in the next section, the tracking of players is elaborated, and the Deep SORT algorithm is described. In Section 3, the experimental setup and the prepared dataset are presented. The results of the experiment and the discussion of different problems that were noticed are given in Section 4, followed by the conclusion and future research directions in Section 5.

[a] https://orcid.org/0000-0002-1829-8023
[b] https://orcid.org/0000-0002-1940-5089
[c] https://orcid.org/0000-0001-5604-2128

Host, K., Ivašić-Kos, M. and Pobar, M.
Tracking Handball Players with the DeepSORT Algorithm.
DOI: 10.5220/0009177605930599
In *Proceedings of the 9th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2020)*, pages 593-599
ISBN: 978-989-758-397-1; ISSN: 2184-4313

2 TRACKING PLAYERS

Handball is played by two teams of seven players, using hands to pass the ball to each other in order to score a goal. This causes multiple players to be involved in the scene at the same time, which makes it difficult to detect and track them the whole time.

In the handball footages, depending on whether it is a match or a practice, approximately 14 to 25 players need to be tracked. For that reason, an algorithm for multiple object tracking (MOT) must be used (Burić et al. 2019).

In the Deep SORT algorithm, which is used here, the detections obtained with an object detector, are used to steer the tracking process. The goal of the tracker itself is then to associate the obtained bounding boxes in different frames together so that the same unique ID is assigned to those boxes that contain the same target. To this end, the tracker may use the information it can obtain from the detected bounding boxes, e.g. the locations of box centroid, their dimensions, the relative position from the boxes in previous frames, or some visual features extracted from the image.

2.1 Detection Algorithm - YOLOv3

In order to track an object, it must first be detected. When it comes to object detection, there are many algorithms such as Mask R CNN (He, Gkioxari, Dollár & Girshick, 2017), Faster R CNN (Girshick, 2015), SSD (Liu et al. 2016), YOLO (Redmon, Divvala, Girshick, & Farhadi, 2016), etc. Following the results of (Burić, Pobar & Ivašić-Kos 2019), in the course of which various algorithms were tested, YOLOv3 (Redmon & Farhadi, 2018) was chosen because it gave the best results for detecting persons.

Yolo is an algorithm based on a single Convolutional neural network (CNN) and can detect objects in real-time. It predicts bounding boxes and confidence values for grid cells into which an image or frame is divided. In the cases when an object is spread across more than one grid cell, the holder of its prediction will be the center cell.

In this particular research, it was important to track the handball players. In consequence of that, for the detection of those players (Burić, Pobar & Ivašić-Kos 2018), the bounding boxes for the objects corresponding to the class "person" were taken into consideration, but only if the confidence for that class was higher than 70%.

2.2 Tracking Algorithm - DeepSORT

DeepSORT (Wojke et al., 2017) is a tracking-by-detection algorithm that considers both the bounding box parameters of the detection results, and the information about appearance of the tracked objects to associate the detections in a new frame with previously tracked objects. It is an online tracking algorithm. Therefore it only considers information about the current and previous frames to make predictions about the current frame without the need to process the whole video at once.

At the beginning of the footage, i.e., in the first frame, a unique track ID is assigned to each bounding box that represents a player and has a confidence value higher than a set threshold. The Hungarian algorithm is used to assign the detections in a new frame to existing tracks so that the assignment cost function reaches the global minimum.

The cost function involves the spatial (Mahalanobis) distance $d^{(1)}$ of the detected bounding box from the position predicted according to previously known position of that object, and a visual distance $d^{(2)}$ that considers the appearance of the detected object and the history of appearance of the tracked object. The cost function of assigning a detected object j to a track i is given by the expression:

$$c_{i,j} = \lambda d^{(1)}(i,j) + (1-\lambda) d^{(2)}(i,j) \quad (1)$$

where λ is a parameter that can be set to determine the influence of the spatial distance $d^{(1)}$ and the visual distance $d^{(2)}$.

The spatial distance $d^{(1)}$ is given by the expression:

$$d^{(1)}(i,j) = \left(d_j - y_i\right)^T S_i^{-1} \left(d_j - y_i\right) \quad (2)$$

where y_i and S_i represent the mean and the covariance matrix bounding box observations for the i-th track, and d_j represents the j-th detected bounding box.

The visual distance $d^{(2)}$ relies on appearance feature descriptors and is given by the expression:

$$d^{(2)}(i,j) = \min\left\{1 - r_j^T r_k^{(i)} | r_k^{(i)} \in \mathcal{R}_i\right\}, \quad (3)$$

where r_j is the appearance descriptor extracted from the part of the image within the j-th detected bounding box, and $\mathcal{R}_i$ is the set of last 100 appearance descriptors r_k^i associated with the track i.

The $d^{(2)}$ measure uses the cosine distance between the j-th detection and i-th track to select the

track where visually the most similar detection was previously found to the current detection.

The appearance descriptors are extracted using a wide residual neural network comprising two convolutional layers followed by six residual blocks that output a 128-element vector. The network was pre-trained on a person re-identification dataset of more than a million images of 1261 pedestrians (Wojke et al., 2017). The feature vectors are normalized to fit within a unit hypersphere so that the cosine distance can be used.

New track IDs are generated whenever there are more detections in a frame than already tracked players, i.e., when a new player is detected in a frame. A new track is also created when a detection cannot be assigned to any track, because the detection is too far from any track, or not visually similar to any previous detection. This is controlled by thresholds that set the maximum $d^{(1)}$ and $d^{(2)}$ when an assignment is still possible. A track may be abandoned if no assignment has been made to it for n consecutive frames. In that case, a new track ID will be assigned if the same object re-appears later in the video.

The appearance information is used in particular to assist in re-identification and prevent new IDs generation for objects that have not been tracked for some time, either because they were under occlusion, have temporarily left the scene, or were not detected because of detector error.

3 EXPERIMENT

The goal of the experiment is to test the performance of the Deep SORT tracker on the handball player tracking task.

In order to evaluate the tracking results, a dataset from the handball domain, containing ground truth tracking annotations had to be prepared first, which is described below in more detail.

The tracking results were then evaluated using multiple common object tracking performance measures. No single performance measure exists that can uniquely describe the complex behavior of trackers, so different measures have been designed for specific applications in mind. Here, the number of identity switches (ID) (Milan et al., 2016), identification precision (IDP), identification recall (IDR) and the identification F1 (IDF1) measures (Ristani et al., 2016) are used.

An identity switch happens if a ground truth target is matched to a track j and the last known assignment was to a track $k \neq j$.

In contrast, the set of IDP, IDR, and IDF1 measures focus on how long a target is correctly identified, regardless of the number of mismatches. Identification precision (IDP) is the fraction of computed detections that are correctly identified in all frames, while identification recall (IDR) is the fraction of ground truth detections that are correctly identified. The IDF1 measure is the ratio of correctly identified detections over the average number of ground-truth and computed detections.

Both the number of identity switches and the IDP/R/F1, measures can be useful for gaining insight about the performance of the tracker in the handball domain, however, being able to re-identify a player after the tracker lost him for a while may be more important than the number of mismatches. Other commonly used measures, the multiple objects tracking precision (MOTP), and the multiple objects tracking accuracies (MOTA) (Bernardin & Stiefelhagen, 2008) are also reported here, but since they heavily award detection performance, they are less relevant here since the same detector was used to generate the ground truth detections and the detections for the tracker.

3.1 Dataset

The dataset contains a subset of high-quality video recordings of handball practice and matches recorded indoors during a handball school (Ivasic-Kos & Pobar, 2018). The recordings were made using a stationary Nikon D7500 DSLR camera, with a Nikon 18-200mm VR lens, in full HD resolution (1920x1080) at 60 frames per second. The camera was positioned on the border of the playing field, on a tripod at 1.5m height. From the spectators' point of view, the height of the camera, which was 10m away from the border, was at 3.5m.

The participants were young handball players and their coaches.

To obtain the ground truth annotation data, the videos were first automatically processed using the YOLOv3 object detector and the DeepSORT tracker to bootstrap the annotation process, and then manually corrected. Since the object detection already performed satisfactory (Pobar & Ivašić-Kos, 2019), the focus of the work was to evaluate the tracking performance, i.e. the ability of the tracker to assign the correct IDs to detections, and not the object detection performance. For this reason, object detection errors such as a missed detection of a player in a frame were not corrected in the ground truth files, but only incorrect assignments of detections to tracks.

To facilitate manual correction and annotation, a custom tool was made using Python and OpenCV.

The tool allows easy swapping of IDs of two tracks, which is a common tracker error when a player moves in front of the other. For example, to correct the error of swapping players, the first player is selected with the left mouse click, and the second one with the right and their IDs would be swapped from that frame forward. An incorrect ID, which often happens in case of occlusion, or when previously tracked players re-enter the frame, can also be edited by right-clicking the bounding box which ID needs changing. Since a single incorrect ID in a frame usually corresponds to a wrongly created new track, it would repeat in the next frames, so correcting a tracking ID in a frame also changes the ID in the following frames. Given that the IDs in a frame must be unique, the program does not allow accidental duplication of an already existing ID. Furthermore, if the swapped IDs need to be changed, but the bounding boxes for both players were not detected at some time later, the user enters the desired value, and the corrections of the IDs occur only until both of the bounding boxes appear again.

The total duration of the annotated dataset is 6min and 18s. The Fig. 1 shows the distribution of the number of detected players on each frame in the tested videos. Most frames in the video contained 10-11 players.

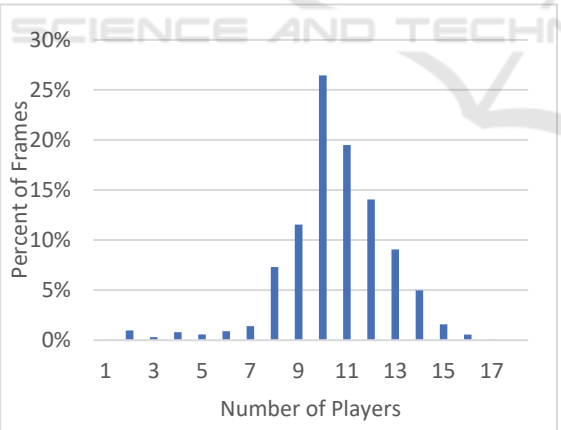

Figure 1: Histogram of the number of players per frame.

4 RESULTS AND DISCUSSION

The quantitative evaluation results are presented in Table 1.

Table 1: Quantitative evaluation results.

Measure	Value
#tracks in the ground truth	279
#tracks	1554
Identity switches	1483
IDF1	24.7%
IDP	24.7%
IDR	24.7%
MOTA	99.3%
MOTP	99%

The results show that for each player that should be tracked, the identity switches caused the creation of, on average, 5-6 additional tracks by the Deep SORT algorithm, so there are 5 times more tracks than in ground-truth data. Due to the relatively large number of players in the video, frequently changed positions, and occlusion, a large number of identity switchers are present (1483). Also, the number of players simultaneously present in the frame obviously affects the tracking performance so that the players can be correctly identified for 24,7% of the time, according to the IDF1 measure. Player identifications that are correctly identified in frames (IDP) are balanced to identification recall (IDR).

Measures MOTA and MOTP show high accuracy and precision results of 99% but are not relevant here since the same detector was used to generate the ground truth detections and the detections for the tracker.

Tracking mistakes can be attributed to several factors. As in all tracking-by-detection algorithms, the accuracy of tracking is greatly influenced by the accuracy of the object detector. If a player is inaccurately detected, the tracking will be inaccurate as well. For example, false positives of the detector, i.e. the bounding boxes that are detected where there are no objects to detect, can confuse the tracker to assign an ID to that box that would otherwise have been assigned to a correct detection.

In other cases, the false positive will produce spurious object IDs with short track durations. In Figure 2, the left shows an example of detecting a player where there is not one, and assigning the ID with the value 31 to it. Furthermore, when the players are at a significant distance, the detector may not recognize them, and in such cases they cannot be tracked. In Figure 2, the middle shows an example of a player not being detected due to the distance from the camera. Another problem regarding the detector is the occlusion, e.g. when a player is covered by another player or by the goal frame.

An example of that is shown in Figure 2, right, where a player was not detected because she was occluded by another player with the ID 5.

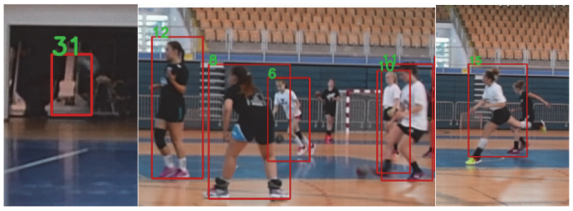

Figure 2: Example of false-positive detection (left), missed detection (middle) and missed detection due to occlusion (right).

Partial occlusion can also make re-identification based on appearance more difficult as it can change the seeming appearance of a player. In these cases, the player will be recognized as a new player and assigned a new ID, which means that the previous data about tracking will not be accessible. Figure 3 shows one such example in three consecutive frames. On the left, there are three players with the IDs 5,7, and 15 in the frame, in the middle the player with the ID 7 is occluded and not detected, and on the right the tracker didn't re-identify the player with the ID 7, but it assigned to it a new ID with the value 22.

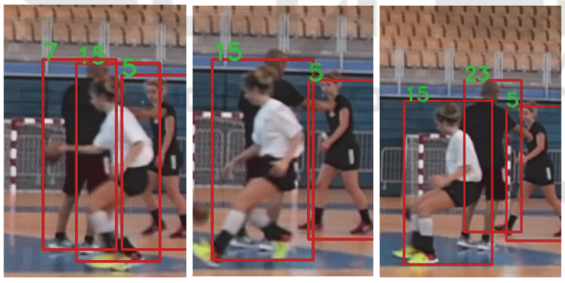

Figure 3: The problem of re-identification after occlusion.

The small scale of some objects in the distance can also be a problem. Although the detector can detect a player in the distance, a different ID may be assigned by the tracker in subsequent frames due to the similarity between the clothes of the player and the background, which makes it difficult for the tracker to recognize that the player is the same as in the previous frames. Figure 4 shows an example of that situation, where the distant player with ID 11 is being detected and tracked, but when it comes to the background with similar color, the tracker recognizes it as a new player.

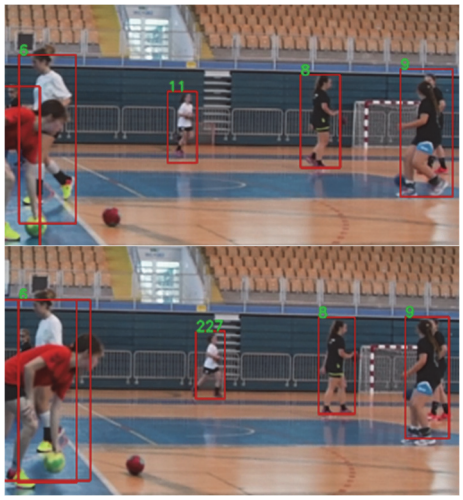

Figure 4: Identity switch due to small scale and similar colors.

Players entering and exiting the frame pose similar problems like occlusion. When a player has left the frame at one moment and returns after some time, the tracker often assigns him a new ID. In some cases, this has been overcome thanks to the information of appearance, but in others the issue is still present. Moreover, when a player exits a frame, and a different player enters, the previous player's ID is often assigned to the new player, when in fact a new ID should be assigned. Another example is when one player is leaving the frame, and the second one is situated on the edge of the frame. In this case when the first player leaves the second one takes its ID.

The IDs can also be swapped between two similarly dressed players, between whom the tracker does not differentiate, or when their movements coincide.

Figure 5 shows two problems of ID swapping due to exiting the frame and occlusions. The player with ID 9 exits the frame (Fig. 5 left), while at the same time the player with ID 3 occupies his position and his ID value (Fig. 5 right). On the same frame, a third player who was not previously detected on the left frame due to occlusion, takes the ID of second player (becames 3), which is probably due to the same color of clothes and similar position on the frame. The coach and the player with ID 10 are well detected on both frames, although large occlusions exist. A player with ID 5 (Fig. 5 left) has left the frame and his ID is correctly excluded from the frame on the right (Fig. 5 right).

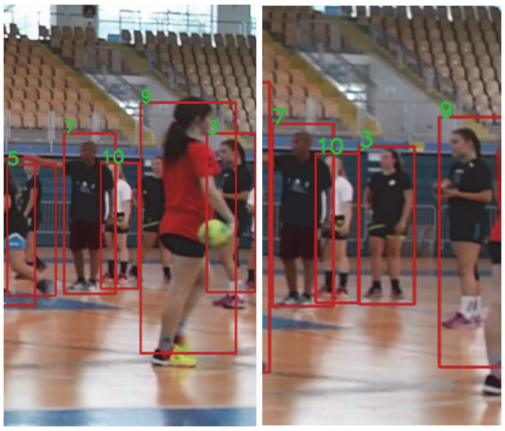

Figure 5: Identity switches due to exiting the frame.

5 CONCLUSIONS

This paper deals with the tracking of young handball players during handball practice. The goal was to detect and track all players on the handball court so that the performance of a particular athlete and the adoption of a particular technique can be analyzed.

This is a very demanding task of multiple object tracking since players move fast, often change direction, and are very often occluded and out of the camera field view. For detection of player the YOLOv3 detector was used and DeepSORT algorithm for player tracking. The results were evaluated on custom dataset that contains handball videos with marked player ID-s. The performances of the algorithm were tested according to common multiple object tracking measures: IDF1, IDP, IDR, MOTA, MOTP. The results of MOTA and MOTP are excellent but not relevant because the same detector was used for ground truth detections and in tracking. Due to the relatively large number of players on the field that are often occluded, and the demanding scenario, players were correctly identified 24.7% of the time, according to the IDF1 measure.

A detailed analysis of the results showed that the scale of an object, occlusion, swapping IDs, and the similar color of the players' clothes with the background, many times appear as problems. Those issues are challenging even for people familiar with players and the rule of the game, so in the future, we will consider different methods to focus monitoring only on players who are active, perform a given action, or are carriers of the game.

Also, we will consider defining an appropriate multiplayer tracking metric that would appropriately evaluate those elements of athlete tracking that are relevant to the task of monitoring and analyzing athlete activity, and performing a particular action.

ACKNOWLEDGMENTS

This research was fully supported by the Croatian Science Foundation under the project IP-2016-06-8345 "Automatic recognition of actions and activities in multimedia content from the sports domain" (RAASS) and by the University of Rijeka under the project number uniri-drustv-18-222.

REFERENCES

Bernardin, K., & Stiefelhagen, R. (2008). Evaluating multiple object tracking performance: the CLEAR MOT metrics. Journal on Image and Video Processing, 2008, 1.

Bewley, A., Ge, Z., Ott, L., Ramos, F., & Upcroft, B. (2016, September). Simple online and realtime tracking. In 2016 *IEEE International Conference on Image Processing (ICIP)* (pp. 3464-3468). IEEE.

Burić, M., Ivašić-Kos, M. & Pobar, M. (2019) (in press) Player Tracking in Sports Videos. In *19th IEEE International Conference on Computer and Information Technology (CIT 2019*)

Burić, M., Pobar, M., & Ivašić-Kos, M. (2019, January). Adapting YOLO network for ball and player detection. In *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods* (ICPRAM 2019) (pp. 845-851).

Burić, M., Pobar, M., & Ivašić-Kos, M. (2018, January). Ball detection using YOLO and Mask R-CNN. In *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*.

Ciaparrone, G., Sánchez, F. L., Tabik, S., Troiano, L., Tagliaferri, R., & Herrera, F. (2019). Deep Learning in Video Multi-Object Tracking: A Survey. arXiv preprint arXiv:1907.12740.

Dicle, C., Camps, O. I., & Sznaier, M. (2013). The way they move: Tracking multiple targets with similar appearance. In *Proceedings of the IEEE international conference on computer vision* (pp. 2304-2311).

Girshick, R. (2015). Fast r-CNN. In *Proceedings of the IEEE international conference on computer vision* (pp. 1440-1448).

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). *Mask r-CNN*. In *Proceedings of the IEEE international conference on computer vision* (pp. 2961-2969).

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016, October). SSD: Single shot multibox detector. In *European conference on computer vision* (pp. 21-37). Springer, Cham.

Ivasic-Kos, M., Pobar, M. (2018). "*Building a labeled dataset for recognition of handball actions using mask R-CNN and STIPS*," in 7th IEEE European Workshop

on Visual Information Processing (EUVIP), Tampere, Finland, pp. 1-6

Milan, A., Leal-Taixé, L., Reid, I., Roth, S., & Schindler, K. (2016). MOT16: A benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831.

Pobar, M., & Ivašić-Kos, M. (2019, March). Detection of the leading player in handball scenes using Mask R-CNN and STIPS. In *Eleventh International Conference on Machine Vision* (ICMV 2018) (Vol. 11041, p. 110411V). International Society for Optics and Photonics.

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

Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767.

Ristani, E., Solera, F., Zou, R., Cucchiara, R., & Tomasi, C. (2016, October). Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision* (pp. 17-35). Springer, Cham.

Wojke, N., Bewley, A., & Paulus, D. (2017, September). Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)* (pp. 3645-3649). IEEE.

Xiang, Y., Alahi, A., & Savarese, S. (2015). Learning to track: Online multi-object tracking by decision making. In *Proceedings of the IEEE international conference on computer vision* (pp. 4705-4713).

Yang, M., & Jia, Y. (2016). Temporal dynamic appearance modeling for online multi-person tracking. *Computer Vision and Image Understanding*, 153, 16-28.