Proceeding of the IEEE
International Conference on Robotics and Biomimetics
Dali, China, December 2019

Multiple Object Tracking Based on the Deep Neural Networks and Correlation Filter*

Qingyu Zhao, Lu Wang and Zefan Zhou
School of Computer Science and Engineering
Northeastern University, China
wanglu@mail.neu.edu.cn

***Abstract*— Online multi-object tracking is a fundamental problem in video analysis based applications. A major challenge in the popular tracking-by-detection framework is how to make data association between detections and trajectories. To tackle this, we propose an appearance model and a motion model based on the deep neural networks. In particular, we adopt a deeply learned appearance representation, which is trained on large-scale person re-identification datasets using a convolutional neural network, to improve the identification ability of our tracker. Towards the motion model, we design a long-short term memory based network for velocity estimation and use two fully-connected layers to measure the motion affinity between detections and trajectories. To alleviate the problem introduced by missing detections, a kernelized correlation filter is leveraged to localize the missing objects. Extensive experiments show that the integration of these three components is effective, and our tracker is competitive compared with existing methods.**

***Index Terms*— Multiple object tracking, data association,**

I. Introduction

Multiple object tracking (MOT) is one of the most important tasks in computer vision, which has broad applications such as robot navigation [1], spots analysis [2], autonomous vehicles [3], among others. It aims to estimate locations of multiple objects in the video and maintain their identities consistently to produce their individual trajectories. Recently, most approaches follow the popular "tracking-by-detection" strategy [2], [4], [5], [6], [7], where objects are first localized in each frame and then associated across frames to form the target trajectories. The core of this strategy lies in the data association process which is usually treated as three separate parts: feature extraction for candidate representation, affinity metric for estimating the linking probability between candidates or tracklets, and the association algorithm to find the optimal association. These parts need to cooperate with each other to adapt different target status, e.g., occlusion, abrupt motion and missing detections.

Recently, deep neural networks have been employed by MOT works [5], [8], [9], [10]. In general, feature representation is learned by convolutional neural networks (CNNs), and it is used to compute the pairwise affinityto determine whether two detections belong to the same target. This paradigm is widely used in the task of person re-identification (Re-ID) and could be exploited in MOT approaches to improve the feature's discriminability and the data association accuracy. For single object tracking (SOT), correlation filters have demonstrated excellent performance in terms of rapidly tracking objects by utilizing the Fast Fourier Transform (FFT) [11], [12]. In our work, we use the kernelized correlation filters (KCF) [13] to recover missing detections introduced by the imperfectness of the object detectors.

In this paper, we propose a data association method to cope with two common challenges in MOT, i.e., id switch and partial occlusion. Specifically, we employ a CNN to extract appearance cue tailored towards person Re-ID, a Long Short-Term Memory (LSTM) network as motion model to extract motion cue and the KCF to predict the most probable target position when the target is missed by the detector. The appearance model, the motion model and the KCF cooperate with each other when facing different circumstances. When the current target is not occluded or partially occluded, the appearance affinity score and the motion affinity score calculated by the appearance and motion models respectively will be multiplied to generate the affinity score used for determining into which trajectory the detection should join. When the trajectory is short and the target velocity cannot be reliably estimated, only the appearance affinity is exploited for data association. When the target is not detected in the current frame, the KCF is leveraged to estimate a most probable location of the target and based on it, the appearance and the motion affinity scores mentioned above can be generated. When the target is lost or occluded in multiple consecutive frames, we simply terminate this trajectory.

Overall, we make the following contributions in this paper:

- We employ a CNN architecture widely used in the task of person re-identification (Re-ID) to distinguish different targets. With this Re-ID feature, our method can effectively reduce the number of identity switches and handle partial occlusion during tracking, hence improving the multi-target tracking accuracy.
- We propose an LSTM network to estimate the velocity

*This work is partially supported by National Training Program of Innovation and Entrepreneurship for Undergraduates, China, #201810145169, NSF Grant of Liaoning, China, #20170540312, and the Fundamental Research Funds for Central Universities, China #N182410001, #N181604006 and #N161602001. Lu Wang is the corresponding author.

978-1-7281-6321-5/19//$31.00 ©2019 IEEE

of the current object in the next frame, which can produce a more accurate prediction in complex and crowded scene than linear motion models.
- We integrate the single object tracking algorithm KCF into multiple object tracking to mitigate the missing detection problem.

II. Related Work

Tracking-by-detection has been the most popular strategy for multi-object tracking. The aim of these approaches is to solve the MOT problem following two main steps: object detection and data association. The quality of data association will directly affect the effectiveness of MOT approaches. Most of the existing methods can be roughly categorized into two groups.

The first group treats MOT as an offline global optimization problem that uses observations from both past and future to estimate the status of the targets [14], [15], [16]. These methods often adopt network flow [17], Hungarian algorithm [18] and energy minimization [19] to solve data association. In [14], the appearance and motion information are globally associated and a minimum group graph is used to associate all detection results. Tang et al. [20] introduce a method of joining and clustering target hypothesis over time to overcome a subgraph multi-cut problem. Rezatofighi et al. [21] propose a joint probability method for data association.

The second group works online by estimating the trajectory of the targets only using the detection results of the current and past frames, independent of the detection of the future frames [22], [23]. Most conventional online methods use Kalman filter [9], Particle filter [24] and Markov Decision Method (MDP) [2]. Xiang et al. [2] apply MDP to model MOT as a decision-making process. A coding dependence across multiple cues in a time window is proposed in [9], which learns the expression of multiple cues for calculating similarity scores. Bae et al [25] use both global and local association to produce trajectories.

However, the tracking accuracy of such methods is highly susceptible to occlusion and noisy detections. Therefore, a robust affinity model is needed for reliable data association.

III. Proposed Method

In this section, we present the appearance model, the motion model and the KCF tracker applied in our method, and then the overall data association strategy is described.

A. Appearance Model

To extract the Re-ID feature from a given bounding box, we use the deep neural network proposed in [26], which is trained on several large-scale pedestrian re-identification datasets. The network consists of the convolutional backbone from GoogLeNet [27] followed by K branches of part-aligned fully-connected (FC) layers. When performing multi-target tracking online, we use the cosine distance between detections and tracks in the appearance feature space:

$$d(i,j) = \frac{f_i f_j}{\|f_i\| \, \|f_j\|}, \tag{1}$$

where f_i and f_j denote the appearance representation of the i-th detection and j-th track. The appearance feature of a trajectory is extracted from its last associated detection. As the sum of the cosine distance and the cosine affinity score is 1, the cosine appearance affinity score is computed as $A_{app}(i,j) = 1 - d(i,j)$. The cosine affinity metric pays more attention to the difference in direction between two vectors, rather than the difference in distance or length.

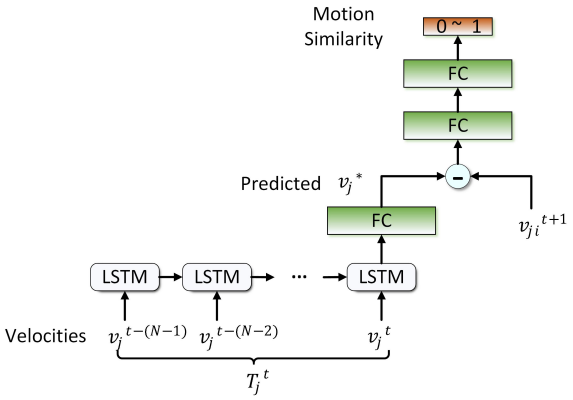

Fig. 1
Motion model. The inputs are 2D coordinates.

B. Motion Model

The motion model includes two parts: the velocity estimation model and the motion similarity model. The schematic diagram of the motion model is shown in Fig. 1.

Due to the powerful capability of deep neural network, we opt to use an LSTM instead of the conventional velocity model, e.g., the Kalman filter, to predict the velocity of target in the next frame. Essentially, as tracking is a sequence problem, it is natural to model the motion of objects by using the LSTM, which is good at handling the sequence problem in complex and crowded scenes, and hence can improve the tracking performance.

For the purpose of learning the velocity information of objects and making reasonable motion prediction, our LSTM accepts as inputs the velocity of trajectory j of length N, denoted as $T_j^t = \left[v_j^{t-(N-1)}, v_j^{t-(N-2)}, \cdots, v_j^t\right]$, and output the velocity v_j^*, an estimation of velocity of trajectory j for time $t+1$. The velocity of a trajectory in each frame can be obtained by subtracting the center coordinate of the target region in the previous frame from that in the current frame.

After getting the predicted velocity v_j^*, we calculate the coordinate difference between any bounding box detection

Fig. 2

USING KCF TO DEAL WITH MISSING DETECTIONS. LEFT COLUMN: DETECTION RESULTS OF TWO CONSECUTIVE FRAMES; RIGHT COLUMN: THE TRACKING RESULTS. THE LADY IN THE MIDDLE OF THE IMAGE IS DETECTED IN THE FIRST FRAME BUT MISSED IN THE SECOND FRAME. BY USING THE KCF, THE LADY IS SUCCESSFULLY TRACKED IN THE SECOND FRAME (SEE THE RED BOX WITH $id = 5$).

i in the current frame and the j-th trajectory as the actual velocity and denote it as v_{ij}^{t+1}. Then given the estimated velocity v_j^* and the actual velocity v_{ij}^{t+1}, we use the mean square error (MSE) loss to measure their inconsistency.

Based on the value of the loss, we apply the motion similarity model to calculate the motion affinity score $A_{motion}(i, j)$, i.e., the possibility that the i-th bounding box detection belongs to j-th trajectory. The motion similarity model consists of two FC layers and the rang of the output is [0,1].

C. Correlation Filter

Due to the instability of the object detector, a target might appear as a missing detection in the current frame F_t. For a single trajectory, missing detection will cause trajectory break and generate a fragmentation, which also increases the number identity switches. To solve this problem, we adopt the KCF tracker [13], which is widely used in single target tracking to generate bounding box, so as to find back the missing detection and thus extending the trajectory.

In essence, KCF is a method that uses Fast Fourier transform (FFT) to accelerate linear regression and uses cyclic shift to approximate dense sampling to improve the discrimination power of the tracker. Although KCF may not have higher precision than those single target tracking methods based on deep learning, it can convert complex convolution operations in the time domain into simple point multiplication operations in the frequency domain, thus greatly improving the tracking speed and facilitating the implementation of a real-time tracker.

Assuming bounding box i_t is a missing detection of the j-th trajectory. For such a case, not only the motion affinity score but also the appearance affinity scores between the j-th trajectory and all detected bounding boxes in the current frame are low, from which we can determine that a missing detection exists. After that, we get the last entity in the j-th trajectory i_{t-1} and use KCF to generate a tracking box, which can be regarded as a replacement of i_t and will join into the j-th trajectory. Fig. 2 shows an example of using KCF to find back missing detections.

D. Data Association

To achieve good performance for MOT, it is vital to carry out a robust data association procedure. During this stage, appearance and motion affinities are used to determine whether the i-th detection belongs to the j-th trajectory. For each frame, we choose a detection that matches the current trajectory best and record the corresponding score. If this score is higher than a certain threshold, we will use this detection to update the trajectory and finish its data association.

If the length of the j-th trajectory is shorter than N, it does not meet the condition of using the motion model. Therefore, we simply extract the Re-ID feature to compute the appearance affinity. For every detection, we calculate its appearance affinity with the j-th trajectory $A_{app}(i, j)$. If there is some detection not associated yet and its appearance affinity with some trajectory is greater than a threshold σ_{app},

the detection will be associated with that trajectory.

If the length of the j-th trajectory is longer than N, the appearance and motion models will be combined so that both the appearance and motion information is considered in the association process. The combined affinity score $A(i,j)$ is computed as the product of the two affinity scores:

$$A(i,j) = A_{app}(i,j) \cdot A_{motion}(i,j). \quad (2)$$

If $A(i,j)$ is the largest among all unmatched detections and it is also greater than the threshold σ_{app_motion}, the i-th detection will be joined into the j-th trajectory.

If the j-th trajectory cannot be associated with any detection in the current frame while it is associated with a detection in the last frame, KCF based tracking is performed to extend the trajectory. Otherwise, we put the j-th trajectory into the inactive trajectory set.

IV. Experimental Results

In order to evaluate the performance of the proposed method, experiments are carried out on the MOT17 [28] dataset, which is a widely used MOT benchmark. First, we divide the training set of MOT17 into two subsets, one for training and the other for validation, to analyze the contribution of each component in our approach. Next, the performance of our method on the test set of MOT17 is compared with other existing methods. Visualization results are given at last.

A. Dataset

The MOT17 benchmark dataset consists of 14 video sequences taken in an unrestricted environment, among which 7 are used for training and the other 7 are for testing. The MOT17 dataset provides three sets of detections, from DPM [29], Faster-RCNN [30] and SDP [31] respectively, for a more comprehensive assessment of the tracking algorithm.

B. Evaluation Metrics

We consider the metrics used by the MOT Challenge benchmarks [32] for evaluation, including Multiple Object Tracking Accuracy (MOTA) [33], Multiple Object Tracking Precision (MOTP) [34], the ratio of correct detections over the average number of ground-truth and computed detections (IDF1 score) [35], the ratio of Mostly Tracked targets (MT), the ratio of Mostly Lost targets (ML), the number of False Positives (FP), the number of False Negatives (FN), the number of ID Switches (IDSw), the number of fragments (Frag).

C. Implementation Details

The proposed method is implemented using Tensorflow and Pytorch [33], with a single GTX1080 GPU. For feature extraction, we exploit the same Re-ID features as the MOTDT [36] to get the feature vector from a given bounding box and the corresponding image.

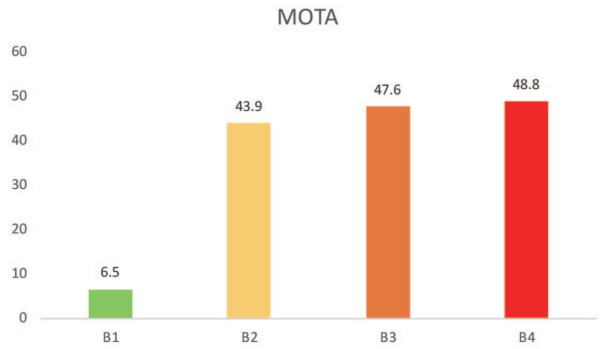

Fig. 3
Contributions of each component.

We form a validation set containing 3 video sequences from the training set, i.e., MOT17-09, MOT17-10 and MOT17-11, to analyze the contribution of each component in our framework. We submit the tracking results of our approach on the MOT17 test set to the MOT Challenge website and get the quantitative evaluation results.

D. Ablation Studies

To demonstrate the contribution of each module in our algorithm, we set up four baseline approaches by disabling each module at one time. Each baseline approach is described as follows:

B1: We disable the proposed appearance model including the MOTDT Re-ID feature extractor and just rely on the motion affinity model to measure the similarity to link the trajectory and the detections.

B2: We disable the KCF module, i.e., we do not handle the missing detections.

B3: We disable the motion model and use the appearance affinity model to finish data association.

B4: Our proposed approach.

Fig. 3 shows the MOTA of each baseline approach on the MOT17 validation set. As can be seen, all the proposed modules make contributions to the performance, with the appearance affinity being the most important one, while KCF also promotes the MOT accuracy significantly.

E. Performance on the MOT17 Benchmark Dataset

We evaluate our approach on the test set of MOT17 benchmark against other online methods, as shown in Table I. Our approach performs favorably against existing methods in terms of the MOTA score. Our method also excels in IDSw, due to the reliability of the Re-ID feature based appearance model. In addition, the FN produced by our approach is the fewest, demonstrating that the KCF tracker can successfully recover many missing detections.

Fig. 4

VISUALIZATION RESULTS OF THE PROPOSED METHOD. LEFT AND MIDDLE: OCCLUSION HANDLING (SEE PEOPLE WITH $id = 66$ IN THE FIRST COLUMN AND $id = 118$ IN THE SECOND COLUMN). RIGHT: ROBUST LONG TERM TRACKING (THE MAN WITH $id = 29$ IS RELIABLY TRACKED EVEN WHEN HE UNDERGOES GREAT SCALE CHANGE).

TABLE I

TRACKING PERFORMANCE ON THE MOT17 DATASET.

Mode	Method	MOTA ↑	IDF1 ↑	MT(%) ↑	ML(%) ↓	FP ↓	FN ↓	IDSw ↓	Frag ↓
online	GM_ PHD [37]	36.4	33.9	4.10	57.30	23723	330767	4607	11317
	GMPHD_KCF [38]	39.6	36.6	8.80	43.30	50903	284228	5811	7414
	GMPHD_N1Tr [39]	42.1	33.9	11.90	42.70	**18214**	297646	10698	10864
	EAMTT [40]	42.6	**41.8**	12.70	42.70	30711	288474	4488	**5720**
	Ours	**44.5**	40.1	**14.10**	**41.20**	32193	**276734**	**4326**	6660

F. Visualization of the Tracker

Fig. 4 shows the visualization results of our method. From these examples, we see that our algorithm can have good tracking performance when the scene is crowded and the targets undergo occlusion and scale variation. The results shown on the left two columns demonstrate that our algorithm is able to correctly track partially occluded pedestrians. The robustness of our algorithm in long-term tracking is shown by the people on the right column.

V. CONCLUSION

In this paper, we present a MOT approach to handle the tracking problem in a unified online MOT framework. For data association, we propose an affinity model, including the motion and appearance models, which takes full advantage

of recent deep neural networks, i.e. LSTM and CNN. To deal with missing detections, the KCF tracker is exploited to extend the trajectory. Experimental results on the MOT17 benchmark dataset show the effectiveness of our approach.

REFERENCES

[1] W. Choi, "Near-online multi-target tracking with aggregated local flow descriptor," IEEE International Conference on Computer Vision, 2015, pp. 3029-3037.

[2] Y. Xiang, A. Alahi, and S. Savarese, "Learning to track: Online multi-object tracking by decision making," International Conference on Computer Vision, 2015, pp. 4705–4713.

[3] B. Lee, E. Erdenee, S. Jin and P. Rhee, "Multi-class multi-object tracking using changing point detection," ECCV Workshops, 2016.

[4] Q. Chu, et al, "Online multi-object tracking using cnn-based single object tracker with spatial-temporal attention mechanism," IEEE International Conference on Computer Vision, 2017, pp. 4846–4855.

[5] L. Leal-Taixe, C. Canton-Ferrer, and K. Schindler. "Learning by tracking: Siamese cnn for robust target association," IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016, pp. 33–40.

[6] A. Milan, S. H. Rezatofighi, A. R. Dick, I. D. Reid, and K. Schindler. "Online multi-target tracking using recurrent neural networks," AAAI, 2017.

[7] R. Sanchez-Matilla, F. Poiesi, and A. Cavallaro. "Online multi-target tracking with strong and weak detections," ECCV 2016 Workshops, 2016, pages 84–99.

[8] C. Kim, F. Li, A Ciptadi, and J. M. Rehg. "Multiple hypothesis tracking revisited," IEEE International Conference on Computer Vision, 2015, pp. 4696–4704.

[9] A. Sadeghian, A. Alahi, and S. Savarese. "Tracking the untrackable: Learning to track multiple cues with long-term dependencies," IEEE International Conference On Computer Vision, 2017, pp. 300-311.

[10] S. Tang, M. Andriluka, B. Andres and B. Schiele." Multiple people tracking by lifted multicut and person re-identification," IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3539–3548.

[11] S. Liu, T. Zhang, X. Cao, and C. Cao. "Structural Correlation Filter for Robust Visual Tracking," IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4312-4320.

[12] M. Danelljan, G. Hager, F. Shahbaz Khan, and M. Felsberg. "Convolutional features for correlation filter based visual tracking," IEEE International Conference on Computer Vision Workshops, 2015, pp. 58–66.

[13] J. F. Henriques, R. Caseiro, P. Martins, J. Batista. "High-speed tracking with kernelized correlation filters," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, pp. 583-596.

[14] A. R. Zamir, A. Dehghan, and M. Shah. "Gmcp-tracker: global multi-object tracking using generalized minimum clique graphs," European Conference on Computer Vision, 2012.

[15] H. U. Kim and C. S. Kim. "Cdt: Cooperative detection and tracking for tracing multiple objects in video sequences," In ECCV, 2016.

[16] N. McLaughlin, J. M. Del Rincon, and P. Miller. "Enhancing linear programming with motion modeling for multi-target tracking," IEEE Winter Conference on Applications of Computer Vision, 2015, pp. 71-77.

[17] A. Butt, and R. T Collins. "Multi-target tracking by lagrangian relaxation to mincost network flow," IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 1846-1853.

[18] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft. "Simple online and realtime tracking," IEEE International Conference on Image Processing, 2016, pp. 3464-3468.

[19] A. Milan, S. Roth and K. Schindler. "Continuous energy minimization for multi-target tracking[J]," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, pp. 58-72.

[20] S. Tang, B. Andres, M. Andriluka, and B. Schiele. "Multi-person tracking by multicuts and deep matching," IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[21] S. H. Rezatofighi, A. Milan, A. Zhang, Q. Shi, A. Dick, and L. Reid. "Joint probabilistic data association revisited," IEEE International Conference on Computer Vision, 2015, pp. 3047-3055.

[22] W. Choi, C. Pantofaru, and S. Savarese. "A general framework for tracking multiple people from a moving camera," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, pp. 1577-1591.

[23] S. Pellegrini, A. Ess, K. Schindler, and L. V. Gool. "You'll never walk alone: modeling social behavior for multi-target tracking," IEEE International Conference on Computer Vision, 2009, pp. 261-268.

[24] K. Okuma et al. "A boosted particle filter: multitarget detection and tracking," European Conference on Computer Vision, 2004.

[25] S.H. Bae, and K.J. Yoon, "Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning," IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1218-1225.

[26] L. Zhao, X. Li, J. Wang, and Y. Zhuang, "Deeply-learned part-aligned representations for person reidentification," International Conference on Computer Vision, 2017, pp. 3239-3248.

[27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke,and A. Rabinovich, "Going deeper with convolutions," IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1-9.

[28] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. "MOT16: A benchmark for multi-object tracking," CoRR abs/1603.00831, 2016, in press.

[29] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan. "Object detection with discriminatively trained part-based models," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, pp. 1627-1645.

[30] S. Ren, K. He, R.B. Girshick, and J. Sun. "Faster r-cnn: Towards real-time object detection with region proposal networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, pp. 1137-1149.

[31] F. Yang, W. Choi, and Y. Lin. "Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers," IEEE Conference on Computer Vision and Pattern Recognition , 2016, pp. 2129-2137.

[32] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler. "MOTchallenge 2015: Towards a benchmark for multi-target tracking," CoRR abs/1504.01942, 2015, in press.

[33] M. Abadi et al., "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," CoRR abs/1603.04467, 2016, in press.

[34] L. Chen et al. "Real-Time Multiple People Tracking with Deeply Learned Candidate Selection and Person Re-Identification," IEEE International Conference on Multimedia and Expo, 2018, pp. 1-6.

[35] N. Wojke, B. Alex, and Dietrich Paulus. "Simple online and realtime tracking with a deep association metric." IEEE International Conference on Image Processing, 2017, pp. 3645-3649.

[36] K. Bernardin, and R. Stiefelhagen. "Evaluating multiple object tracking performance: the CLEAR MOT metrics," EURASIP J.Image and Video Processing, 2008.

[37] V. Eiselein, D. Arp, M. Pätzold, and T. Sikora. "Real-time Multi-Human Tracking using a Probability Hypothesis Density Filter and multiple detectors," IEEE International Conference on Advanced Video and Signal-Based Surveillance, 2012, pp. 325-330.

[38] T. Kutschbach, E. Bochinski, V. Eiselein, and T. Sikora. "Sequential Sensor Fusion Combining Probability Hypothesis Density and Kernelized Correlation Filters for Multi-Object Tracking in Video Data," IEEE International Conference on Advanced Video and Signal Based Surveillance, 2017, pp. 1-5.

[39] N. Baisa, and A. Wallace. "Development of a N-type GM-PHD filter for multiple target, multiple type visual tracking," J. Visual Communication and Image Representation, 2019, pp. 257-271.

[40] R. Sanchez-Matilla, F. Poiesi, and A. Cavallaro. "Online Multi-target Tracking with Strong and Weak Detections," ECCV 2016 Workshops, 2016.