Learning Deep Appearance Feature for Multi-target Tracking

Hexi Li, Na Jiang, Chenxin Sun, Zhong Zhou, Wei Wu
State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China
lihexi@buaa.edu.cn

***Abstract*-Multi-target tracking is a worthy studying issue in computer vision. For surveillance video, frequent occlusion and dense crowds complicate the issue. To resolve these difficulties, this paper proposes an effective algorithm of multi-target tracking in videos. Firstly, the faster Rcnn is proposed with the residual network to extract the objects of pedestrians in surveillance videos. The proposedment can effectively eliminate invalid target detection frames, separate peer targets and resist partial occlusions. Then, this paper put forward an accurate and efficient appearance-feature matching network model that is inspired by pedestrian re-identification theory. The deep learning feature-extraction module is composed of the stem Cnn and the Resnet blocks, therefore it can load res-50 caffemodel as pretraining model to increase the accuracy of the feature-extraction. Meanwhile, the proposed network can decrease the time of train and test comparing with Resnet. Finally, the obtained multiple target tracking trajectories are further optimized by the strategy of occlusion distinction, deduplication and merging. The experiment results of the 2D MOT 2015 benchmark、KITTI dataset indicate that this proposed algorithm outperforms alternative multiple objects trackers in terms of multiple indicators.**

Keywords: target detection, appearance match, multi-target tracking, trajectory optimization

I. INTRODUCTION

This article regarded multi-object tracking as a track classification problem mainly based on the characteristics of target is performance and space-time constraints. In other word, the important factor of this problem is extracting and classifying the features of targets and the temporal constraints.

The information of pedestrian is often the key objective understanding of surveillance video content, however, the non-rigid properties of its negatively influence the process of multi-target tracking. In recent five years, the widely application of deep learning in image classification and recognition has seen major advancements since its high efficiency, which provides the contributions made in the areas of target tracking. For instance, the technology of target detection has proposed from RCNN to Faster RCNN which can offer the possibility of the frame detection inside video. However, there are still several challenges in tracking case such as missed, detecting error, Occlusion and Id switch. To analyze these issues is an effect method to research the relationship between the solution for the existing problem of multi-objective and the recognition of the extracting appearance technology and the characteristics of the accurate time and space constraints.

With the inspiration from the similar problem, re-identification, this article also split the research process to two apartments: the extraction of feature and the measure of portions similarity. The difference with re-identification is without the interference from dramatic changes of pose and illumination in the same camera, therefore the method used in this article has less difficulty than re-identification, but it is higher level in requirement of time efficiency due to the need to match the average among frames. Thus, the appearance feature matching model designed in this article can reduces the depth of model to proposed the time efficiency and continuous use the residual model based on the ImageNet pre-trained model to reach the high matching accuracy. Depending on the appearance feature matching method, the result of this study can get closer to the data of the real-time multi-target preliminary tracking trajectory. And then we also use some information about spatial and temporal to constraints which and make contribution to the result. In this way, we has better performance in the platform of 2D MOT 2015 benchmark. In conclusion, the main contribution of this study are the following three points:

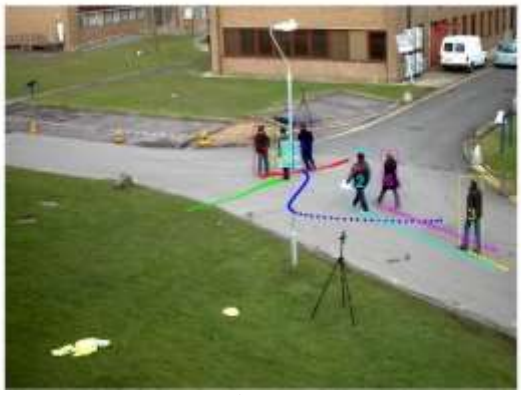

Figure 1. Example result of our method on PETS09-S2L1

1. To realize the exaction of pedestrian target in surveillance videos by improving the Faster RCNN and residual network structure which can effectively eliminate invalid block target detection, separate target peer and resistant partial occlusion.

2. To propose an accurate and efficient appearance feature matching network model based on the pedestrian re-recognition theory.

3. To propose a data correlation matching algorithms which focus on the modeling on different states of the target, the re-association of the preliminary results and the completion of blocked targets to obtain an accurate and complete tracking trajectory.

II. REALTED WORK

Multi-target tracking algorithm can be divided into two categories: the multi-target tracking without detection[1][2] method and the tracking multi-target with detection method. The first one requires to mark the bounding box of the first frame to do the initialization of system.

In the multi-target tracking based on the detection algorithm, the result of detection plays an important role in the accuracy of tracking, this article classifies the method without application of deep learning as the traditional methods. For example, DPM[3][4] is a robustness component-based detection method which firstly calculates a gradient direction histogram to obtained proposed features of HOG. Secondly, using the sliding window to construct the scale pyramid after obtaining the gradient model trained by SVM to get detecting result. With the development of the deep neural networks, a series of RCNN[5][6][7] methods gain positive results in the detection of target exceed DPM rapidly. For instance, aster RCNN[5] is an end-to-end frame composed by two-part. The first step of this method is to extracts features and the areas of interest. Secondly, the targets in these areas will be classified. There are several advantages of this frame. Compared with general object detection, it does not show a same high performance in the pedestrian detection because of the small scale of objectives and background interference.

Based on the detection[24] and feature extraction algorithm[26], the multiple target tracking problem can be transformed into the problem about data association between the frames.

It focuses on applying statistical theory and fuzzing mathematics on the association between the targets. In recent years, the emergence of new appearance information and suitable approximate model has recovered this data association applies the JPDA[3] frame revisited the JPDA formulation in visual MOT with the goal to address the combinatorial complexity issue with an efficient approximation of the JPDA by exploiting recent developments in solving integer programs[4] also applies an appearance model for each target to prune the MHT[9][12] graph to achieve state-of-the-art performance[6] uses Kuhn-Munkres algorithm[25] to optimize the track association and obtain an online tracker . Geiger used Hungarian algorithm to perform data association, he divided the tracking process into two steps, in the first, and these data among frames should be correlated to produce the track segments. Secondly, the completed tracking trajectory could be obtained after connecting these segments together. Inspired by this reference, the study of this article applied the separating segment approach to associate data and correct the deviation of the test results, but the 3D information would not be used. This article will provide detailed explanation in the following chapter.

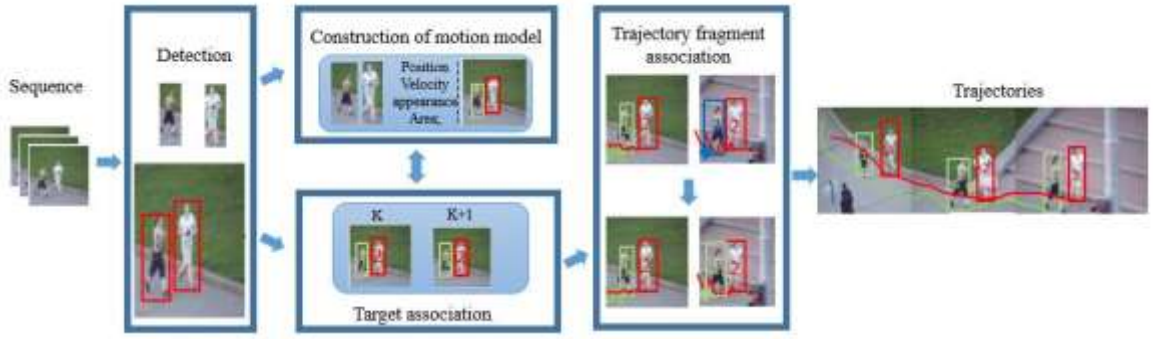

Figure 2. Outline of our approach

III. METHODOLOGY

This article adopts the proposed detection network. In the first, the detection targets in the surveillance video will be detected among all frames and the appearance feature of target will be extracted. The next process is that the data will be associated, combined and the redundant will be removed after building the model of target movement, which could obtain the multi-target trajectory. The multi-target tracking algorithm proposed in this article consists of the following three aspects: the detection algorithm, feature extraction, object association. The figure2 shows the process:

A. Target detection

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

In multi-target tracking algorithms, the algorithm based on the detection is to be the mainstream. The tracking algorithm applied on this article firstly use the detection algorithm to detect each frame in the video stream. However, most detection algorithms based on deep learning are not only for pedestrians. Inspired by He Kaiming[23], the study in this article select RPN in Faster RCNN as the detecting basis and do several modifications. For instance, the process for pedestrians is to modify the length-width ratio of reference box from RPN network multiple aspect ratios to the 2 aspect ratio which is more adapted for human. In addition, after several tests and result comparing, the authors found the utilization of pre-training model has a positive effect in the classification of small sample dataset. Therefore, the resnet-

50 with pre-training model is selected as a feature for the feature extraction process. Last but not least, after considering the generalized flaw of the deep learning algorithm model, this study applies the dataset of ETH and Daimier pedestrian detection to finetune.

TABLE 1. DETECTION PERFORMANCE ON MOT16 TRAIN SET

Tracker	FP	FN	AP
ACF[24]	86341	16013	0.3
DPM[5]	62353	28839	0.6
Faster RCNN[8]	46738	4476	0.592
Ours	42695	4231	0.62

For verify the effectiveness of the proposed detection algorithm, the pedestrian is selected as the example and these algorithm mentioned before will be compared on the MOT16 bench mark platform. The result can be showed on the table 1: the detecting result is obtained from the training set of MOT16, it has 7 video sequences which includes 5316 frames and 11407 bounding boxes. From the table, the proposed detecting algorithm has better performance than other algorithm in terms of the number of missing and false and AP. This result could demonstrate that this method can proposed the accuracy of detecting. Specially, when associating the target in the following process, the occurrence of missing may lead to no related associate target. The error association and the increase of the number of targets may be resulted in the occurrence of the false data. Therefore reduce of the number of these errors could positively influence the following association and the tracking result.

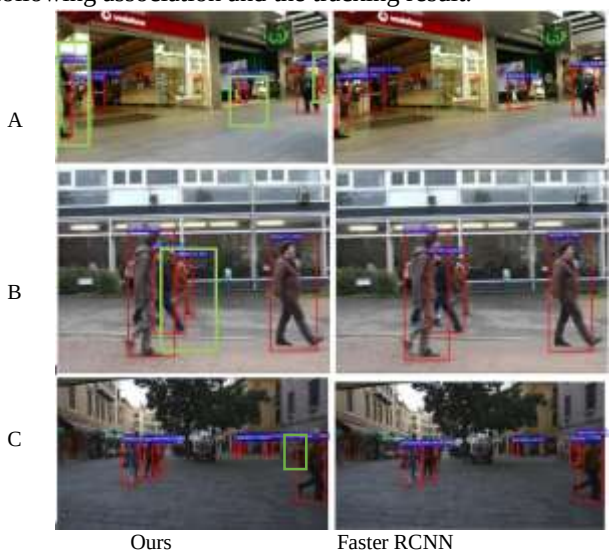

Figure 3. Compare our approach with Faster RCNN

As shown in figure 3, compared with Faster-RCNN, the proposed algorithm can have higher performance when meeting problems, such as interleaving targets, blocked, missed and false detection. The red bounding boxes in the figure 3 indicate the detection result and the green and bold bounding boxes show the result of the proposed algorithm is more accurate than the Faster-RCNN algorithm.

It is clear that the woman and children (surrounded by the green bounding boxes) were combined as the one target in the Faster-RCNN. However, the proposed algorithm can separate two target samples. For the left-most man, the scale of the bounding boxes of the proposed algorithm is more accurate than the Faster-RCNN algorithm because later has the missed detection. B, C Similarly.

B. Featurematching appearance

In multi-target tracking process, this article applies the target detection method mentioned before to obtain all pending tracking targets, and introduce the re-id method to associate the adjacent frames. After analyzing feature extraction algorithms generated in recent years, the re-id method could be divided into two parts: appearance feature extraction and similarity measure. The CNN based algorithm has higher performance than the traditional algorithms, especially in the appearance feature extraction part. Inspired by the CVPR paper[14][26][27], this article proposes proposed appearance matching model, the detailed network construction is shown in the figure 4.

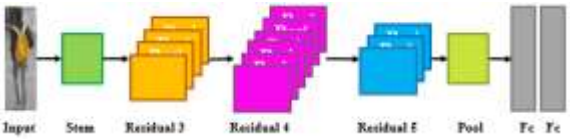

Figure 4. Framework of appearance features network

For the input data with different sizes, the raw image will be expended to the fixed length 128*64 in the data preprocessing stage and move into the feature extraction model. The detailed output blob and the change will be shown in the table 2.

TABLE 2. NETWORK STRUCTURE OF APPEARANCE MODEL

Name	Patch Size/Stride	Unites number	Output size
input	-	-	3*128*64
Stem	3*3/1(C),2*2/2(P)	CNN*3+Pool*1	512*64*32
Residual block(3)	1*1/2	4	1024*16*8
Residual block(4)	1*1/2	6	2048*8*4
Residual block(5)	1*1/1	3	2048*8*4
Pool2	8*4/1	-	2048*1*1
Fc1	-	-	4096
Fc2	-	-	M

From the table2, it is clear that this network is consisted of three convolutional layers, a reservoir layer, three residual blocks, a pooling layer and two FC layers. These three residual blocks include 4, 6, 3 residual unites. Due to the deep learning network frame is more dependent on the parameters of initialization, this article applies the resnet-50[28] as the pre-training model. Considering the fact that the pending associating-tracking targets of the multi-target tracking problems belong to the same camera, and the type of the movement of pedestrian is less than its in re-id problem, the complexity of appearance matching model could be reduced with the utilization of the last three residual blocks on the resnet-50 network structure. This method could ensure the

efficiency of target association and the decrease of the number of the network layers could positively reduce the time for training.

For the high generalization of appearance model network in the training process and the robustness generated in the un-training tracking dataset, this article applies DGD[26] multiple datasets hybrid approach which could mix the four relatively large dataset in the re-id field, respectively shows in figure 5.

The time for extracting every picture features is 4.95ms (cosine distance calculation time can be ignored). For example, the 10 candidate frames and 10 targets in the current frame could be extracted during 60ms. This advantage could result that the proposed multi-target tracking algorithm can be treated as a real- time method without the trajectory optimization. The experiment result is shown as the figure 5, this appearance model not only has high performance in the re-id dataset, but also in the multi-target tracking process. The red bonding box is the tracking result, the digit represents the id number of character and green box indicates the attitude changes in the larger figure.

In figure 5 A shows a man in a red dress from the front to the side, B shows the unfolded action of the man who is with black clothes changes to vertical side-to-side motion, C shows all body of the woman appeared in the scene and the part of body away from the scene could be observed and matched in the appearance model.

C. Target ssociation and optimization

In this chapter, we mainly introduce two aspects: target association method and trajectory optimization strategy.

1) Target association

Conversion probabilities of the detected value between two consecutive frames could be gained by calculating the detected value, position, size, speed, and the similarity of surface.

$$\mathrm{P}\left(x_{t_{k+1}}^{j}\middle|x_{t_k}^{i}\right)=\gamma_1 A_p\left(x_{t_{k+1}}^{j}\middle|x_{t_k}^{i}\right)+\gamma_2 A_s\left(x_{t_{k+1}}^{j}\middle|x_{t_k}^{i}\right)+\gamma_3 A_v\left(x_{t_{k+1}}^{j}\middle|x_{t_k}^{i}\right)+\gamma_1 A_\theta\left(x_{t_{k+1}}^{j}\middle|x_{t_k}^{i}\right) \quad (1)$$

where in A_p, A_s, A_v, A_θ are the position, size, speed, the appearance phase between two detected values, the formula as below:

$$A_p\left(x_{t_{k+1}}^{j}\middle|x_{t_k}^{i}\right)=\exp\left[-\frac{\left(p_{t_{k+1}}^{j}-p_{t_k}^{i}\right)}{\sigma_p^2}\right] \quad (2)$$

$$A_s\left(x_{t_{k+1}}^{j}\middle|x_{t_k}^{i}\right)=\exp\left[-\frac{\left(s_{t_{k+1}}^{j}-s_{t_k}^{i}\right)}{\sigma_s^2}\right] \quad (3)$$

$$A_v\left(x_{t_{k+1}}^{j}\middle|x_{t_k}^{i}\right)=\exp\left[-\frac{\left(v_{t_{k+1}}^{j}-v_{t_k}^{i}\right)}{\sigma_s^2}\right] \quad (4)$$

$$A_\theta\left(x_{t_{k+1}}^{j}\middle|x_{t_k}^{i}\right)=E\left(a_{t_{k+1}}^{j},a_{t_k}^{i}\right) \quad (5)$$

where $E\left(a_{t_{k+1}}^{j},a_{t_k}^{i}\right)$ denotes the detected value of the appearance color histogram Euclidean distance. $a_{t_{k+1}}^{j}$ represented by k id number is the position of the figure in the time t_k, $a_{t_{k+1}}^{j}$ denotes figures (unassigned character id number) in the position of the time t_k , $A_p\left(x_{t_{k+1}}^{j}\middle|x_{t_k}^{i}\right)$ indicates the position of the similarity between two detection values.

A detected value which can be associated to the target track has two conditions: the correct detected value and the match of detected conversion conditions which includes the matching surface, the related size and uniform motion direction between this detected value and the value for the end of the trajectory. This article defines the detected value is related to one trajectory. In addition, the optimal solution for the association between the value and the trajectory could be gained from the Hungarian algorithm.

2) Optimization strategy

The parts of frames may not be matched and recognized as a new target after the data association among frames, because there are several problems on the bounding boxes, such as the complex background, the occlusion and the change of attitudes, which could lead to the fragmented target trajectory.

The probability of the correct detected value is in a high level. According to the fragmented target trajectories, the calculate method is that the 5 bounding boxes with highest confidence coefficient from the big trajectories (includes 20 bounding boxes) and the representative point will be chose to calcite the feature value of this trajectory. Considering the relatively non-constant area of the bounding boxes, the relatively consistency of the appearance features, the position and speed (a sudden change of direction) of the target in the video sequence, the average feature vector of this sequence will be calculated with the confidence coefficient chosen as

Figure 5. Associate results of tracking target with difference

TABLE 3. TRACKING RESULTS ON THE MOT15 CHALLENGE

Tracker	Det	MOTA	MOTP	MT	ML	FP	FN	IDs	Frag
TRID[15]	Pri	**55.7**	**76.5**	40.6%	25.8	6273	**20611**	**351**	**667**
NOMTwSDP[16]	Pri	**55.5**	**76.6**	**39.0%**	25.8%	5594	21322	**427**	**701**
AP_RCNN	Pri	53.0	75.5	29.1%	**20.2%**	**5159**	22984	708	1476
EAMTT[17]	Pri	53.0	75.3	35.9%	**19.6%**	7538	**20590**	776	1269
CDA_DDAL[18]	Pri	51.3	74.2	**36.3%**	22.2%	7110	22271	544	1335
KCFKF	Pri	51.0	76.4	33.0%	14.1%	6426	22546	1131	2649
MDP_SubCNN[20]	Pri	47.5	74.2	30.0%	18.6%	8631	22969	628	1370
TSML_CDE[19]	Pri	49.1	74.3	30.4%	26.4%	**5204**	25460	637	1034
APRCNN_Pub	Pub	38.5	72.6	8.7%	37.4%	**4005**	33203	586	1263
MDPNN[21]	Pub	37.6	71.7	15.8%	26.8%	7933	29397	1026	2024
JointMC[22]	Pub	35.6	71.9	23.2%	39.3%	10580	28508	457	969
Ours	Pri	**56.2**	**77.3**	**38.0%**	**19.8%**	7239	**19217**	**403**	**628**

the calculating coefficient and the appearance feature vectors of only 5 bounding boxes.

The formula is shown:

$$\bar{a}_{id_j} = \frac{\sum P(x_{t_k}^j) a_{t_k}^i}{n} \tag{6}$$

$$\mathrm{P}\left(t_{kd_i} \middle| t_{kd_j}\right) = E(\bar{a}_{kd_j}, \bar{a}_{kd_i}) \tag{7}$$

where $\mathrm{P}\left(t_{id_i} \middle| t_{id_j}\right)$ is represented as id number character bounding boxes of confidence in the time t_{kd_i}, $a_{t_k}^i$ represents the appearance characteristics.The average feature vector should be pairwise matched and the result of matching will determine whether the trajectory is from the same target, then the same tracking segments connect each other to form a complete tracking trajectory. For disappear bounding boxes, the Kalman filter prediction position could be applied to complete them.

IV. EVALUATION AND ANALYSIS

We evaluate the performance of our tracking implementation on a diverse set of testing sequences as set by the MOT benchmark database which contains both moving and static camera sequences.

A. 2D MOT 2015 benchmark evaluation

We utilize different evaluation indicators in the evaluation criteria MOTA, MOTP, FAF, MT, ML and so on. And First result in red, blue means second, green means third, shown in table 3. There are several on-line algorithms, such as EAMTT[17], CDA_DDAL[18], MDPNN[21], MDP_SubCNN[21], NOMTwSDP[16], KCFKF. There are two off-line algorithm, such as TRID[15], JointMC[22], NOMTwSDP[16], TSML_CDE[19]. Compared with these algorithm, the MOTA value and MOTP value of the proposed algorithm in this article is in the highest level, which indicates the result of this algorithm has the heightset accuracy in the benchmark. The MOTA and ML is minimal, because of the relative association among the subsequent frames and the completion for the missing frames.

B. KITTI dataset objective evaluation

Shown in table 4, the proposed algorithm has achieved a higher MOTA value for pedestrians on the KITTI dataset, which can prove that this proposed algorithm has higher performance than other algorithm. Moreover, the performance of the number of trajectory segments is in the same level due to the proposed feature matching and the trajectory association method.

Figure 7. Examples of tracking for MOT15

TABLE 4. RESULTS OF KITTI PEDESTRIAN

Tracker	MOTA	MOTP	MT	ML	ID	FG
MCMOT-CPD[14]	78.9%	82.13%	52.31%	11.69%	228	536
NMOT*[13]	78.15%	79.46%	57.23%	13.23%	**31**	207
LP-SSVM[11]	77.63%	77.80%	56.31%	**8.46%**	62	539
CEM[10]	51.94%	77.11%	20.00%	31.54%	125	396
Ours	**80.64%**	**82.74%**	**61.37%**	8.50%	43	**198**

V. CONCLUSION

This article proposed an accurate target tracking algorithm which proposed the current detection method, efficiently eliminated the invalid target detection bounding boxes, separated the target peer and minimized the effect of partial occlusion. In addition, an accurate and efficient appearance feature matching network model has been introduced based on the re-recognition pedestrian theory, which can ensure the effectiveness of the features after the relative short training and feature extracting process. In the last part of this article, the data association method based on classification has been described. It divided the targets into three types: active, inactive and missing targets. For these types, the proposed algorithm can implement related methods to solve the omission and occlusion problems, such as the optimization of trajectory, combination and optimization strategies. This algorithm had high performance in both benchmark dataset and KITTI dataset.

ACKNOWLEDGMENT

This work is supported by the National 863 Program of China under Grant No.2015AA016403 and the Natural Science Foundation of China under Grant No.61572061, 61472020.

REFERENCES

[1] Lucas B D, Kanade T. An iterative image registration technique with an application to stereo visio[C] //IJCAI.1981,81:674-679

[2] Zhang L, Maaten L V D. Structure Preserving Object Tracking[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2013:1838-1845.

[3] S. H. Rezatofighi, A. Milan, Z. Zhang, A. Dick, Q. Shi,and I. Reid, "Joint Probabilistic Data Association Revisited," in International Conference on Computer Vision,2015

[4] Object detection with discriminatively trained partbased models. IEEE Trans. PAMI, 32(9):1627‑1645, 2010.DPM

[5] Rodrigo Benenson, Mohamed Omran, Jan Hosang, Bernt Schiele. Ten Years of Pedestrian Detection, What Have We Learned ? In ECCV, CVRSUAD workshop, 2014.

[6] Girshick R, Donahue J, Darrell T, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[C]// Computer Vision and Pattern Recognition. IEEE, 2014:580-587.

[7] Girshick, Ross. "Fast r-cnn." Proceedings of the IEEE International Conference on Computer Vision. 2015.

[8] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6):1137.

[9] C. Kim, F. Li, A. Ciptadi, and J. M. Rehg, "Multiple Hypothesis Tracking Revisited," in International Conference on Computer Vision, 2015.

[10] A. Milan, S. Roth and K. Schindler: Continuous Energy Minimization for Multitarget Tracking. IEEE TPAMI 2014.

[11] S. Wang and C. Fowlkes: Learning Optimal Parameters for Multi-target Tracking with Contextual Interactions. International Journal of Computer Vision 2016..

[12] D. Reid, "An Algorithm for Tracking Multiple Targets," Automatic Control, vol. 24, pp. 843‑854, 1979

[13] W. Choi: Near-Online Multi-target Tracking with Aggregated Local Flow Descriptor . ICCV 2015..

[14] B. Lee, E. Erdenee, S. Jin, M. Nam, Y. Jung and P. Rhee: Multi-class Multi-object Tracking Using Changing Point Detection. ECCVWORK 2016

[15] S. Manen, M. Gygli, D. Dai, L. Van Gool. PathTrack: Fast Trajectory Annotation with Path Supervision. In ArXiv e-prints, 2017.

[16] W. Choi. Near-Online Multi-target Tracking with Aggregated Local Flow Descriptor. In ICCV, 2015.

[17] R. Sanchez-Matilla, F. Poiesi, A. Cavallaro "Multi-target tracking with strong and weak detections" in BMTT ECCVw 2016

[18] S. Bae and K. Yoon, Confidence-Based Data Association and Discriminative Deep Appearance Learning for Robust Online Multi-Object Tracking, In IEEE TPAMI, 2017.

[19] B. Wang, G. Wang, K. L. Chan, L. Wang. Tracklet Association by Online Target-Specific Metric Learning and Coherent Dynamics Estimation. In arXiv:1511.06654, 2015.

[20] Y. Xiang, A. Alahi, S. Savarese. Learning to Track: Online Multi-Object Tracking by Decision Making. In International Conference on Computer Vision (ICCV), 2015.

[21] A. Sadeghian, A. Alahi, S. Savarese. Tracking The Untrackable: Learning To Track Multiple Cues with Long-Term Dependencies. In arXiv preprint arXiv:1701.01909, 2017.

[22] M. Keuper, S. Tang, Z. Yu, B. Andres, T. Brox, B. Schiele. A Multi-cut Formulation for Joint Segmentation and Tracking of Multiple Objects. In CoRR, 2016.

[23] Zhang L, Lin L, Liang X, et al. Is Faster R-CNN Doing Well for Pedestrian Detection[J]. 2016:443-457.

[24] P. Dollar, R. Appel, S. Belongie, and P. Perona, "Fast Feature Pyramids for Object Detection," Pattern Analysis and Machine Intelligence, vol. 36, 2014

[25] A. Geiger, M. Lauer, C. Wojek, C. Stiller, and R. Urtasun, "3D Traffic Scene Understanding from Movable Platforms," Pattern Analysis and Machine Intelligence,2014.

[26] Xiao T, Li H, Ouyang W, et al. Learning Deep Feature Representations with Domain Guided Dropout for Person Re-identification[J]. 2016.

[27] Cho Y J, Yoon K J. Improving Person Re-identification via Pose-Aware Multi-shot Matching[C]// Computer Vision and Pattern Recognition. IEEE, 2016:1354-1362.

[28] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.