Online Multiple Object Tracking with Reid Feature Extraction Network and Similarity Matrix Network

Qingge Ji\textsuperscript{a}, Haoqiang Yu\textsuperscript{b}

School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510006, China

Guangdong Province Key Laboratory of Big Data Analysis and Processing, Guangzhou 510006, China

\textsuperscript{a}E-mail: issjqg@mail.sysu.edu.cn, \textsuperscript{b}E-mail: qhyu_144s@126.com

ABSTRACT In multiple object tracking (MOT), data association is a crucial part. By constructing similarity loss matrix for trajectories and detections, they can be matched correspondingly by using Hungarian matching algorithm. However, the similarity loss is often obtained by calculating the Euclidean distance or other handcraft distance metrics of features extracted between objects, which may not be robust enough, resulting in matching inaccuracy. In this paper, we propose a novel MOT method with applying deep learning to feature extraction and data association. We firstly design the appearance feature extraction network (AFN) to learn effective features by training it on a large-scale person re-identification dataset (reid). Then, we propose the similarity matrix estimation network (SMN) to obtain reliable similarity by training it on the public MOT Challenge dataset, MOT17. Additionally, the similarity matrix output of SMN includes the dummy objects, which are used to deal with the association problems of object missing and object appearance between frames. In the end, our proposed MOT method is evaluated on MOT15, MOT17 and ablation study is carried out.

CCS Concepts: Computing methodologies~Artificial intelligence~Computer vision~Computer vision problems~Tracking.

1. INTRODUCTION

Because of its high academic potential and commercial value, multiple object tracking plays an indispensable role in the field of computer vision. The purpose of MOT is to establish reliable and accurate tracks for multiple interested objects in video or image sequences. It can not only be used for other tasks such as object location prediction, recognition and relocation but also can be used in video monitoring, virtual reality, pose estimation and other aspects.

Common online multiple object tracking algorithms are based on the tracking-by-detection framework, which usually includes five parts: object detection, feature extraction, similarity measurement, data association and track management. With the rapid development of the field of deep learning, many object detection and feature extraction techniques\cite{1} are more and more mature. Their accuracy and real-time performance have made great progress, which provides great help for tracking-by-detection MOT algorithms. However, in the parts of similarity measurement and data association, most of the work still only use the traditional methods, resulting efficiency and accuracy bottlenecks. They rarely combine MOT evaluation metrics as loss functions, leading to poor performance of tracking method. Moreover, in the data association part, at present, most of the work match tracks and detections by only utilizing Hungarian algorithm according to the similarity loss matrix. However,
when new objects appear or old objects disappear frequently in the video, Hungarian algorithm has some defects, making the algorithm prone to id switch (IDS).

In order to solve some challenging problems mentioned above, in this paper, we design an online MOT method based on the tracking-by-detection framework, taking the advantages of deep learning. We exploit an appearance feature extracting network (AFN) constructed by deep convolution neural network (CNN), along with person re-identification, to extract the discriminative features of detections. Another proposed CNN, namely similarity matrix estimation network, can estimate similarity matrix according to all the detection features in two different frames, which is used as the input of data association matching. It is worth mentioning that the similarity matrix includes dummy targets, which is specially used to deal with the situation that old objects disappear or new objects appear between frames.

The rest of this paper is arranged as follows: in Section 2, we review the related multiple object tracking work. In Section 3, we introduce our tracking algorithm pipeline as well as proposed AFN and SMN. Experiments results, implementation details and ablation study are brought out in Section 4. In Section 5, we summarize our algorithms.

2. RELATED WORK

In general, existing multi-object tracking algorithms can be categorized into online and offline methods. The online methods [2][3] only use the frames up to the current frame to estimate object trajectories. On the contrary, the offline methods [4][5] can use both the past and future frames to indicate object trajectories. Although offline methods can achieve more accurate results relatively easily, this paper pays more attention to the real-time performance and robustness of the algorithm, thus our method is online.

Appearance is an important way to calculate similarity in MOT. Generally, appearance models can be divided into visual representation and statistical measurement. Visual representation classes are generally based on single or multiple features to describe an object. Statistical measurement classes are to measure the similarity between different objects. In this paper, we mainly use the more discriminative statistical measurement model.

Probabilistic prediction methods usually regard the object state as an uncertain distribution. And the purpose of tracking algorithm is to estimate the probability distribution with various probability methods based on existing object trajectories. This kind of algorithm usually only needs the past or present object state, thus it is especially suitable for online tracking. The commonly used prediction methods are Kalman filter, particle filter. Compared with probability prediction, deterministic optimization aims to find the optimal posterior solution in MOT. Common methods include greedy bipartite graph matching, Hungarian algorithm[6], linear programming, minimum loss maximum flow network, maximum weight independent set. Among them, Hungarian matching algorithm is widely used because of its convenient, simple, fast and efficient characteristics, but at the same time, it also has obvious disadvantages: when the minimum matching threshold is not set, it will associate all the objects between two frames. In this case, the Hungarian matching algorithm is unsuited to MOT: old objects disappear or new objects appear between frames. When this situation occurs, it is likely to produce mismatch, causing IDS, FP, FN rise significantly. Therefore, if the Hungarian matching algorithm is to be applied perfectly, then the similarity matrix for matching needs to contain all the objects of two frames, as well as additional redundant items, that is, adding a row and a column in the matrix. In our paper, the similarity matrix obtained by SMN contains dummy targets, thus Hungarian algorithm can be well applied.

In recent years, with the development of deep learning, many methods that use neural networks to assist in tracking or extracting features also emerge. For example, Sheng et al. [3] proposed the historical appearance matching method and joint-input siamese network which was trained by 2-step process, as well as providing useful technique to remove noisy detections effectively according to scene condition. Yoon et al. [2] presented an online method that encodes long-term temporal dependencies across multiple cues, using a structure of Recurrent Neural Networks that jointly reasons
on multiple cues over a temporal window. In this paper, we not only apply deep learning to feature extraction, but also innovatively integrate it into data association.

3. PROPOSED METHOD
In this Section, we introduce the overall process of the algorithm, including the AFN based on person re-identification dataset and the SMN based on mot dataset.

3.1 MOT Pipeline
Based on the tracking-by-detection framework, our MOT algorithm pipeline is shown in Figure 1. First of all, the input of the tracking frame is composed of an image frame and a group of position information that can represent all the detections in the current frame, including the coordinates of the top left vertex as well as the width and height of the bounding box. At frame \( t \), the \( i \)-th detection can be expressed as \( D_i^t \), and the whole set can be expressed as \( \{D_i^t, i \in [0,N_t]\} \), in which represents the maximum number of detection at frame \( t \). Secondly, in the stage of feature extraction, the tracking framework extracts the image blocks of each detection, and input them into AFN one by one to get the appearance features based on person re-identification dataset. This process will be described in Section 3.2 in detail. In the phase of estimating similarity matrix, after the tracking framework pairs all the features of detections in frame \( T \) with all the features of track in previous \( \theta \) frames, it then passes each pair into SMN. As the output of SMN, the similarity of each detection-track pair is obtained. It is worth noting that the obtained matrix includes dummy targets, which means the matrix considers the situation that detections and tracks don’t match. It will be described in Section 3.3 in detail. In the stage of data association, according to the similarity matrix, the tracking framework uses Hungarian matching algorithm to match all the detections in frame \( t \) with the current existing track one by one. Since there is no minimum threshold for matching in our paper, detections and tracks can be matched successfully as a result of dummy targets mentioned above in the matrix. The detections matched to the dummy target initializes a new track and updates the features cache with the current feature, while the detections matched to the track only needs to update the feature cache. At the end, the tracking framework outputs the tracks of current frame \( t \) and goes to the next frame.

![Figure 1. Proposed MOT Pipeline](image)

In addition, before input, non maximum suppression (NMS) and minimum confidence filtering are carried out to ensure the reliability of the detections.

3.2 Appearance Feature Network
Without using other cues such as motion information and interactive information, a discriminative appearance feature is of crucial importance for similarity measurement. Especially in the multiple object tracking task, with the passage of time, objects are prone to pose variation or appearance change in the video frames, which requires the feature extractor to extract robust and reliable features. In order to achieve this goal, we design a CNN to extract the appearance features of detections, and train it on a large-scale person re-identification dataset[7]. The dataset contains more than 32,000 images of pedestrians, including 1501 different identities. Each person is composed of 2 to 6 cameras, which is suitable for deep learning.
Figure 2. Region Segmentation. The size of Global is 128x64. The size of a sub-block of Local is 32x64.

As shown in Figure 2, the structure of AFN is mainly divided into two parts: Global part and local part. The global part corresponds to the complete pedestrian image, whose size is 128x64. It aims to extract the overall attributes of the image, such as texture, shape, color and other depth information. The local part corresponds to four different longitudinal areas of pedestrian, whose size is 32x64. Four areas are divided equally, each of which can be expressed as head, upper body, lower body and foot respectively. The reason of the idea is that in actual tracking, detections usually collides or occludes. Merely using the global part to extract features is easy to mix with those occluded parts, which will cause certain interference to the discrimination of features, while joining the local part can enhance the interpret-ability according to the human body structure.

The specific network structure of AFN is shown in Table 1. Due to the few parameters, high speed and accuracy of wide residual network[8], we design AFN based on the residual block in the network. It can be seen that the global part of the network consists of seven blocks, and the local part consists of four blocks. Both parts get a 128 dimensional features respectively through their convolution blocks, and then pass them into FC layer for batch and L2 normalization operation, and finally connect them into a 256 dimensional feature output. The training process of AFN and the evaluation results on the person re-identification dataset will be described in Section 4.3.

Table 1. proposed appearance feature network structure.

| Layer | Output size | Patch Size, Channel, Stride |
|-------|-------------|-----------------------------|
|       |             | Global                      | Local                      |
| Conv1 | 128x64      | [3x3, 32]                   | [3x3, 32]                  |
| Conv2 | G: 64x32, L: 32x64 | 3x3 MP, S-2, [3x3, 32] | SL-4, 2x2 MP, S-1 |
| Conv3 | G: 64x32, L: 16x32 | [3x3, 32], [3x3, 32]| S-2, [3x3, 32] |
| Conv4 | G: 32x16, L: 8x16 | [3x3, 64], S-2, [3x3, 64]| [3x3, 32] |
| Conv5 | G: 32x16, L: 8x16 | [3x3, 64]| [3x3, 64] |
| Conv6 | G: 16x8, L: 8x16 | [3x3, 128], S-2, [3x3, 128]| |
| Conv7 | G: 16x8, L: 8x16 | [3x3, 128]| [3x3, 128] |
| Fc    | 1x1         | 16x8 MP, S-1                | 8x16 MP, S-1               |
|       |             | CA-4                        | CA-2                       |
|       |             | 16x8 MP, S-1                | 8x16 MP, S-1               |
|       |             | CA-2                        | 1x1, 256                   |
3.3 Similarity Matrix Network

Similarity matrix has a pivotal role in MOT process. Reasonable construction enables the discriminative features to be effectively used. In the past, most tracking algorithms used traditional distance measurements such as Euclidean distance, Mahalanobis distance, Manhattan distance and even Intersection over Union (IoU), to calculate similarity between objects. Although these methods are simple and fast, they are often more suitable for measuring motion feature information between objects, while they may not be effective for appearance information. In our paper, we innovatively utilize deep learning to construct similarity loss matrix with deep appearance information.

As shown in Figure 3, the input of SMN is a matrix composed of all the appearance features of the detections in two frames, \( t_1 \) and \( t_2 \), where \( t_1, t_2 \in [0, t_{max}] \). \( t_{max} \) represents the maximum video frame, and the interval between \( t_1 \) and \( t_2 \) is no more than \( \alpha \), which is set to 40 in this paper. That is to say, in the interval of 40 frames, two image frames are randomly selected by SMN to extract features as the input. Each element in the matrix is a new feature \( \{F_{ij}^{t_1}, i \in [0,N_{t_1}], j \in [0,N_{t_2}]\} \) formed by connecting each detection feature in the two frames, \( F_{i}^{t_1} \) and \( F_{j}^{t_2} \). The dimension of new features is 512, so the dimension of the whole matrix is \( 512 \times N_{max} \times N_{max} \), in which \( N_{max} \) represents the maximum number of objects in the current frame, which is set as 80 in this paper. Then, the tracking framework input the matrix into a CNN with eight convolution layers. The network structure is shown in Table 2. Through forward propagation, the matrix is mapped to a matrix with dimension \( 1 \times N_{max} \times N_{max} \). At this time, the matrix only integrates the detections features information in these two frames, however, it does not include the situation of new objects appearing and old objects disappearing. Therefore, we need to add a row and a column to the original matrix to represent the above situation, and then row-wise softmax and column-wise softmax are applied to the extended matrix, to get the matrix \( S_{col} \) and \( S_{row} \). The result is substituted into the following formula (1) ~ (4), and the loss \( L \) is the loss of the whole SMN:

\[
L = L_4 + L_2 + L_3
\]  

![Figure 3. Proposed SMN Pipeline](image)

| Layer Index | Output Channel | Patch Size |
|-------------|----------------|------------|
| 1           | 512            | 1x1        |
| 2           | 256            | 1x1        |
| 3           | 128            | 1x1        |
| 4           | 64             | 1x1        |
| 5           | 32             | 1x1        |
| 6           | 16             | 1x1        |
| 7           | 8              | 1x1        |
| 8           | 1              | 1x1        |
\[ L_1 = -\sum \frac{G_{row} \otimes \log(S_{row})}{G_{row}} \]  
\[ L_2 = -\sum \frac{G_{col} \otimes \log(S_{col})}{G_{col}} \]  
\[ L_3 = \left\| \frac{\sum (S_{row} - S_{col})}{\sum S} \right\| \]

where \( \otimes \) represents matrix dot multiplication, \( G_{it} \) represents the correlation ground truth matrix between all the objects in frame \( t_1 \) and frame \( t_2 \). For example, if the i-th object in the frame \( t_1 \) is associated with the j-th object in the frame \( t_2 \), the value of j-th element in the i-th row of the matrix is ‘1’. \( G_{row} \) and \( G_{it} \) represent the new truth matrix after row expansion and column expansion respectively, that is, if the i-th object in frame \( t_1 \) is not associated with the object in frame \( t_2 \), the value of \((N_{max} + 1) - \text{th}\) element in i-th row of the matrix \( G_{col} \) is ‘1’. Similarly to \( G_{row} \). \( S_{row} \) and \( S_{col} \), respectively, indicate that the extended row is removed and the extended column is removed. The training process of SMN will be described in Section 4.3.

4. EXPERIMENTS

In this section, we evaluate our proposed algorithm on the public MOT challenge dataset: MOT15[9][10] and MOT17[11]. Meanwhile, the implementation details and ablation study are also carried out.

4.1 Dataset

The datasets we use include MOT15 and MOT17. MOT15 includes eleven different scene of videos. Each is divided into two clips, one for training and the other for testing. Similar to MOT15, MOT17 is among the latest online challenges in tracking, which contains seven different scene of videos, with 3 different detectors of objects, namely SDP[12], Faster-RCNN[1] and DPM[13] respectively. These different videos include indoor and outdoor, dense and sparse scenes, which are very comprehensive datasets for the tracking algorithm evaluation.

4.2 Evaluation Metrics

We use the evaluation metrics[9][14][15] of MOT benchmark dataset as our measurement metrics. The higher the score, the better the metric with (↑) and the worse the metric with (↓). These metrics include:

MOTA(↑): Multiple Object Tracking Accuracy, combines three error sources: false positives, missed objects and identity switches.
MT(↑): Mostly tracked objects. The ratio of ground-truth trajectories that are covered by a track hypothesis for at least 80% of their respective life span.
ML(↓): Mostly lost objects. The ratio of ground-truth trajectories that are covered by a track hypothesis for at most 20% of their respective life span.
FP(↓): The total number of false positives.
FN(↓): The total number of false negatives.
IDS(↓): The total number of identity switches.
Frag(↓): The total number of times a trajectory is fragmented.
4.3 Implementation Detail
The neural networks in this paper are all trained on NVIDIA GeForce GTX 1070 GPU. Besides, for the appearance feature extraction network, we set its batch size as 32, learning rate as 1e-3, optimizer as adaptive moment estimation (Adam) and loss function as cosine softmax [16]. We trained it about 50,000 times to get the convergence model. To test the performance of the re-id model, we evaluated it on the Market1501 test set and achieved a result of mAP=0.61 and rank-1=0.79, proving that our model has certain discrimination for pedestrian.

For the similarity matrix network, we set its batch size as 8, learning rate as 1e-4 and optimizer as stochastic gradient descent (SGD). We trained it about 30,000 times to get the convergence model.

4.4 Ablation Study
In order to prove the validity of each part of our algorithm, we disable the functions of each part of the algorithm and evaluate on the MOT17 training set. As shown in Table 3, AFSM is the result of our proposed algorithm in our paper.

Table 3. Tracking performance of our method disable different component on the MOT17 training dataset.

| Method          | MOTA(↑) | MT(↑) | ML(↓) | FP(↓) | FN(↓) | IDS(↓) | Frag(↓) |
|-----------------|---------|-------|-------|-------|-------|--------|---------|
| B1              | 44.2    | 285   | 573   | 16475 | 165021| 6416   | 7855    |
| B2              | 47.5    | 327   | 600   | 9674  | 165231| 1969   | 3750    |
| B3              | 47.7    | 329   | 590   | 9435  | 164996| 1907   | 3807    |
| AFSM(Ours)      | 50.4    | 392   | 565   | 13286 | 151901| 1965   | 2802    |

B1: We disable the AFN proposed in Section 3.2 and replace the network with a single convolution layer.
B2: We disable the SMN proposed in Section 3.3 and replace the network with a single convolution layer.
B3: We disable the dummy targets included in and replace the matrix with as the output similarity matrix.

As can be seen from the Table 3, each part of the algorithm has a certain impact on the whole pipeline. Throughout the whole Table, AFN is the most influential one. Since once the effective and discriminative features has lost, the role of subsequent parts will be greatly declined. Additionally, the influence of SMN and output matrix changes are not much difference, both of which have obvious reduction to the whole algorithm, proving the contribution of depth similarity information and dummy targets to the algorithm.

4.5 Results
We take mot15 public detection set as input, and the generated tracking results are shown in Table 4. Also, we take mot17 public detection set as input, the generated tracking results are shown in Table 5. In order to make a fair and effective comparison, we gather the online methods published in recent years as the baseline methods. These online methods also made full use of appearance feature, but didn’t attach great importance to similarity.

Table 4. Tracking performance on the MOT15 dataset. Best results in each category appear in bold.

| Method          | MOTA(↑) | MT(↑) | ML(↓) | FP(↓) | FN(↓) | IDS(↓) | Frag(↓) |
|-----------------|---------|-------|-------|-------|-------|--------|---------|
| TSDA_OAL[17]    | 18.6%   | 9.4%  | 42.3% | 16350 | 32853 | 806    | 1544    |
| TC_SIAMESE[18]  | 20.2%   | 2.6%  | 67.5% | 6127  | 42596 | 294    | 825     |
| MTSTracker[19]  | 20.6%   | 9.0%  | 36.9% | 15161 | 32212 | 1387   | 2357    |
| EAMTTPub[20]    | 22.3%   | 5.4%  | 52.7% | 7924  | 38982 | 833    | 1485    |
| DEEPPDA_MOT[21] | 22.5%   | 6.4%  | 62.0% | 7346  | 39092 | 1159   | 1538    |
| AFSM(Ours)      | 23.2%   | 6.9%  | 39.2% | 8101  | 39421 | 1025   | 1289    |
Table 5: Tracking performance on the MOT17 dataset. Best results in each category appear in bold.

| Method        | MOTA(↑) | MT(↑) | ML(↓) | FP(↓)  | FN(↓)  | IDS(↓) | Frag(↓) |
|---------------|---------|-------|-------|--------|--------|--------|---------|
| GMPHD[22]     | 36.4%   | 4.1%  | 57.3% | 23723  | 330767 | 4607   | 11317   |
| GMPHD_KCF[23] | 39.6%   | 8.8%  | 43.3% | 50903  | 284228 | 5811   | 7414    |
| EAMTT[20]     | 42.6%   | 12.7% | 42.7% | 30711  | 288474 | 4488   | 5720    |
| FPSN[24]      | 44.9%   | 16.5% | 35.8% | 33757  | 269952 | 7136   | 14491   |
| HISP_DAL17[25]| 45.4%   | 14.8% | 39.2% | 21820  | 277473 | 8727   | 7147    |
| AFSM(Ours)    | 46.9%   | 17.1% | 33.9% | 25887  | 272299 | 4503   | 6175    |

On the most important comprehensive metric MOTA, the proposed algorithm surpasses most online algorithms in MOT15 but achieved an average score in the other metrics. In contrast, our method achieve a comparable MOTA score in MOT17 and high MT score and low ML score in the same time. It attributes to the detections provided in MOT17 dataset are much precise, showing that our method relies on detection more heavily. The more accurate the detection, the better the performance of the feature extracted by AFN.

5. CONCLUSION
In this paper, we exploited an online multiple object tracking algorithm, which combines deep learning with feature extraction and data association, containing two different convolutional neural networks: AFN and SMN. AFN was trained on person re-identification dataset, in order to extract the appearance features of detections. SMN was trained on MOT Challenge dataset, in order to learn the similarity matrix between detections and tracks. In addition, SMN also establishes dummy targets terms for similarity matrix, which are used to deal with the problem of association between vanishing detections or new-born detections and tracks in consecutive frames. Last but not least, we analyzed each part of the modules independently, and evaluate our tracking algorithm on the public datasets MOT15 and MOT17. The evaluation results verify the feasibility and superiority of our proposed algorithm.

ACKNOWLEDGMENTS
This research was supported by the Natural Science Foundation of Guangdong Province, China (No. 2016A030313288). We thank the anonymous reviewers for their suggestions and comments.

REFERENCES
[1] S. Ren, K. He, R. Girshick, and J. Sun, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137–1149, 2017
[2] A.Sadeghian, A.Alahi, S.Savarese. Tracking the Untrackable: Learning to Track Multiple Cues with Long-Term Dependencies. in ICCV, 2017
[3] Y.-C. Yoon, A. Boragule, Y. min Song, K.Yoon, and M.Jeon, Online multi-object tracking with historical appearance matching and scene adaptive detection filtering, in AVSS, 2018.
[4] H.Sheng, Y.Zhang, J.Chen, Z.Xiong and J.Zhang. Heterogeneous association graph fusion for target association in multiple object tracking. IEEE Transactions on Circuits and Systems for Video Technology, PP(99), 1-1. 2018
[5] M.Keuper, S. Tang, B. Andres, T.Brox and B. Schiele. Motion segmentation & multiple object tracking by correlation co-clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1-1, 2018
[6] J. Xing, H. Ai, and S. Lao. Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses, in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2009, pp. 1200–1207.
[7] L.Zheng, L.Shen, LTian, S.Wang, and Q.Tian. Scalable Person Re-identification: A Benchmark. 2015 IEEE International Conference on Computer Vision (ICCV). IEEE.
[8] S. Zagoruyko and N. Komodakis, Wide residual networks, in BMVC, 2016, pp. 1–12
[9] E. Ristani, F. Solera, R. Zou, et al. Performance measures and a data set for multi-target, multi-camera tracking[C]//European Conference on Computer Vision. Springer, Cham, 2016: 17-35.
[10] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler, MOTChallenge 2015: Towards a Benchmark for Multi-Target Tracking, pp. 1–15, 2015.
[11] A. Milan, L. Leal-Taixé, I. D. Reid, S. Roth, and K. Schindler, MOT16: A Benchmark for Multi-Object Tracking, CoRR, vol. abs/1603.0, 2016.
[12] 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, in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016 pp. 2129-2137.
[13] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, Object detection with discriminatively trained partbased models,IEEE transactions on pattern analysis and machine intelligence, vol. 32, no. 9, pp. 1627–1645, 2010.
[14] K. Bernardin and R. Stiefelhagen. Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics. Image and Video Processing, 2008(1):1-10, 2008.
[15] Y. Li, C. Huang and R. Nevatia. Learning to associate: HybridBoosted multi-object tracker for crowded scene. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2009.
[16] N. Wojke, A. Bewley. Deep Cosine Metric Learning for Person Re-identification. In IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018: 748-756.
[17] J. Ju, D. Kim, B. Ku, et al. Online multi-person tracking with two-stage data association and online appearance model learning[J]. Iet Computer Vision, 2017, 11(1):87-95.
[18] Chul, Yoon & Song, Young-Min & Yoon, Kwangjin & Jeon, Moongu. (2018). Online Multi-Object Tracking Using Selective Deep Appearance Matching. 206-212. 10.1109/ICCE-ASIA.2018.8552105.
[19] N.T.L. Anh, F.M. Khan, F. Negin, et al. Multi-Object tracking using multi-channel part appearance representation[C]//Advanced Video and Signal Based Surveillance (AVSS), 2017 14th IEEE International Conference on. IEEE, 2017: 1-6.
[20] R. Sanchez-Matilla, F. Poiesi, A. Cavallaro. Online multi-target tracking with strong and weak detections[C]//European Conference on Computer Vision. Springer, Cham, 2016: 84-99.
[21] K.Yoon, D.Kim, Y. C Yoon, & M.Jeon. (2019). Data association for multi-object tracking via deep neural networks. Sensors, 19(3).
[22] Eiselein, Volker & Arp,Daniel & Patzold, M & Sikora, Thomas. (2012). Real-Time Multi-human Tracking Using a Probability Hypothesis Density Filter and Multiple Detectors. 325-330. 10.1109/AVSS.2012.59
[23] Kutschbach, Tino & Bochinski, Erik & Eiselein, Volker & Sikora, Thomas. (2017). Sequential Sensor Fusion Combining Probability Hypothesis Density and Kernelized Correlation Filters for Multi-Object Tracking in Video Data. 10.1109/AVSS.2017.8078517.
[24] Lee, Sangyun & Kim, Euntai. (2018). Multiple Object Tracking via Feature Pyramid Siamese Networks. IEEE Access. PP. 1-1. 10.1109/ACCESS.2018.2889442.
[25] N. Baisa. Robust Online Multi-target Visual Tracking using a HISP Filter with Discriminative Deep Appearance Learning. In CoRR, 2019.