Joint Prediction and Association for Deep Feature Multiple Object Tracking

Wenyuan Qin, Hong Du*, Xiaozheng Zhang, Zhiyang Ma, Xuebin Ren and Tian Luo
China Northern Vehicle Research Institute, Beijing, China
Email: 826673487@qq.com

Abstract. Deep learning (CNN) can significantly improve the accuracy of image recognition with its powerful features, but the low-level network layer also contains important feature information. In order to achieve more stable and efficient tracking in multi-target tracking, this type of deep features will also be used to make the features more expressive by integrating the data from the front and back layers. The deformable convolution is also introduced to overcome the deformation problem caused by the camera motion. And with the increase of time, we predict the position of the target by the motion model, so as to remove the position where the target is impossible to reach in physical space, and further optimize the association before multiple targets. In this paper, we use an end-to-end correlation method to reduce the complexity of the algorithm. We tested it on the open source dataset MOT17 dataset and obtained remarkable results.

Keywords. Neural network; end to end; motion model; multiple object tracking.

1. Introduction
In the current field of multi-target tracking, researchers have naturally divided the multi-target tracking task into two parts. One part is the detection module and the other part is the tracking module [1]. Tracking of multiple targets requires the detection module to pass the detected targets to the tracking module. The correlation is done by the data of each frame passed in by the detector. Thus, the association of the same target from frame to frame is achieved [2]. Therefore the main task of multi-target tracking is in the problem of association of targets and a good association method determines the performance of the tracker.

Currently, the widely used multi-target tracking methods build a representation model of the target detected by the detector [3]. The model will then solve for similarity based on the provided features to correlate between frame and frame targets, and the model generally includes appearance features [4], motion features [5] and composite features [6]. Appearance features focus on encoding each target’s image region as a feature thus achieving the effect of association. Currently it is very common to manually build the features of the target, however, with the development of deep learning applying deep features to build the appearance model can lead to more efficient tracking results.

Motion models accomplish the determination of the future position of the target by creating a dynamic model of the target [7]. These models process the velocity, position or acceleration of the target obtained from the video. Thus, the result of predicting the position is achieved, but such an approach is susceptible to sensors and the algorithm may not have a good predictive performance for the complex motion of the target. Therefore using composite features for correlation is to balance the
motion model of the target with the appearance model of the target so that the algorithm can be well applied to practical problems.

Based on the above idea, a composite feature model is developed in this paper, which is also influenced by the idea of Deep Affinity Network (DAN) [8]. The appearance features and motion model of the target are integrated into the DAN thus completing the end-to-end training of the association, reducing the complexity and improving the accuracy of the algorithm. Also considering the problem of target deformation in the application, we believe that the introduction of variable convolution can be a good solution to the problem of target deformation due to camera motion. At the same time, using deep-level features for the target features will lose some features of the target. We believe that the spatial features of the target are contained in the low-level layer, and as the network deepens, the spatial and other features of the target will be weakened. Therefore, we correlate the low-level target features with the deeper level target features to further improve the correlation of targets. We tested the algorithm on the MOT17 dataset and obtained good results.

2. Related Works
Former multiple object tracking has attracted the interest of many researchers. Many scholars are currently investigating detection-based data association multi-target tracking methods [9]. Such methods are mainly based on the tracking-by-detection idea, which converts the multi-target tracking problem into a data association strategy to form target trajectories by reliably associating detection responses. The tracking algorithm acquires the target foreground in real time and then solves the final cost matrix based on the affinity matrix to determine the optimal association pair between the target history trajectory and the detection response in the current frame. In early studies, in order to obtain the target trajectory accurately, many scholars combined the target detection response and filter techniques to implement the idea of tracking-by-detection in a more direct way. Wang [10] et al. used the inter-frame difference method to detect moving targets, adaptively marking each target’s label, assigning each target an independent Kalman filter. Ge [11] et al. and Zhao [12] combined optical flow histogram features and LBP (Local Binary Patterns) features respectively on the basis of Kalman filter, which can accomplish multiple objects tracking tasks accurately and in real time. Naiel et al. [13] established a new algorithm and Naiel et al. [14] developed a correction model for detectors and trackers within a particle filtering framework, treating each detection region as a sample with different importance, and using frame-by-frame data correlation between detection responses and trackers to achieve online multiple object tracking. Eiselein [15] et al. proposed a Gaussian Mixture Probability Hypothesis Density (GMP) based on the Gaussian Probability Hypothesis Density (GMPHD) based multiple object tracking method, which successfully achieved accurate tracking of multiple targets.

And with the advent of deep learning algorithms, pre-trained classification tasks have received a lot of attention [16]. Therefore, the use of deep learning for multiple object tracking and thus target association started. Li et al. proposed the use of deep learning techniques for single-target tracking using twin neural networks for real-time tracking. Zhang proposed a long-term tracking approach based on two network structures. The first network uses regression to generate a series of candidate targets and
calculates their similarity, and the second network evaluates them for tracking. Wang was the first to use two-layer self-coding to extract visual features for input into a support vector machine and model the data association part to solve the minimum spanning tree problem in 2014 [17]. After this, deep neural networks started to be widely used for multi-target tracking. On the one hand, many scholars obtained good tracking results by simply adding convolutional features. Wojke et al. extended the convolutional features on the classical SORT [18] algorithm, which not only achieved real-time operation but also improved the tracking performance [19]. Kim added the classical Multiple Hypothesis Tracking (MHT) Kim added deep performance features trained on ImageNet to the classical MHT (Multiple Hypothesis Tracking) and achieved better performance [20]. Chen et al. continued this research and added the Enhanced Detection Model to further improve the results [21]. On the other hand, many scholars started to study network structures more suitable for tracking. Chu et al. combined single-target tracking and multiple object tracking, used ROI pooling to build a network branch for each potential target to track, and introduced an attention mechanism to solve the occlusion and target interaction problems [22]. Zhu et al. proposed a dual-matching attention network, which used the matching of input picture pairs in the spatial attention module to patterns to obtain dual attention maps, and adaptively assign different attention mechanisms to different samples for tracking in the temporal attention module [23]. Leal-TAIX’e proposed twin convolutional neural networks (Siamese CNN) with input picture pairs and combined them with gradients of extracted content features to match predictions, and obtained the final trajectory after linear programming [24]. Zhang et al. proposed pixel-level motion and performance modeling to extract primary features in the low-level network, while using ResNet18 as the base module in the high-level structure using a non-weight sharing strategy to continue Re-ID output tracking results [25]. xu considered that none of the current trackers are implemented at the metric level of MOTA and MOTP for algorithm optimization, and proposed the differentiable pair target tracking framework deepMOT algorithm, which optimizes data association directly by combining a bi-directional RNN with a single-target tracker [26]. S.Y. Tang at the Max Planck Institute in Germany modeled the global data association problem as a Minimum Cost Multicut approach and applied it to multi-target tracking and human pose estimation [27]. Emami et al. [28] viewed multi-target tracking as an allocation task and used a number of multiple object tracking methods to construct this equation.

3. Method
Our tracking algorithm is detection-based tracking. We utilize the features from the middle layer of the network as part of our final features, which are enriched by the up-sampling integration of the individual features, thus ensuring the trustworthiness of the features in the matching module. We feed the obtained feature modules into the matching module. During training the error will be passed backwards from the matching module to the feature extraction module for joint learning, thus improving the performance of the feature extraction module. Our model also uses Kalman filter as our motion module from a practical point of view. We use Kalman filter prediction to remove positions that are impossible to reach for the physical space.

In the inference stage as in figure 1, after the features are extracted to the corresponding features, the bounding box and features are used as input to further extract the features of the target. The matching module calculates the similarity by the features of the target in the current frame and the target in the previous frame. And the model prediction module, by Kalman filtering removes the targets that are impossible to exist in the physical space to calculate a loss. Then the optimal match is found according to the Hungarian algorithm to obtain the optimal association pair between frames. The details will be explained next.
3.1. Feature Extractor
For the input image, we first go through feature extraction on it. The features of the image are extracted as much as possible. First we consider that the lower network layer contains more spatial and texture information while the higher network layer ignores this part. Therefore, we integrate the feature maps extracted from these two parts to make the features more expressive. We also consider the deformation of the target due to the transformation of the angle during the target movement or camera movement. Therefore, based on the above idea, we use deformable convolutional [29] to extract the features to make the network more expressive. Due to the influence of the Deep Affinity Network (DAN) idea, we feed the obtained features into the deformable DAN for deep association. The details of the network are as follows: First, we deform the input image to 900, and then pass through a convolutional neural network with a convolutional kernel size of 3 and a stride 1. Our framework is somewhat similar to Deep Layer Aggregation (DLA) [30], but we have changed the structure of some of its parts. There are five channels of 512, 256, 128, 64, and 32, respectively. A total of ten feature maps are obtained, and the obtained ten feature maps are fed into the DAN network for deep association.

3.2. Matching Module
The features we obtained from the feature extraction module were used for subsequent feature embedding, with a total of ten convolutional network layers with a channel number output of 32. This allowed us to use the obtained feature representations for tracking. We obtained different embedding from different network layers. we obtained the bounding box \( B_t = \{b_1^t, b_2^t, ..., b_N^t\} \) for each frame from the dataset. \( N_t \) is the number of targets at moment.

For each detection target provided by the dataset, we extract the embedding features of the target from the center of the bounding box. If the center of the ith target is at (x,y) and the size of the input is \( W \times H \), the size of the obtained feature map is \( W_m \times H_m \times C_m \), we scale the coordinates of the center point \( \left( \frac{y}{H} H_m, \frac{x}{W} W_m \right) \) represents the \( m \)th feature map. After that, we integrate the m features embedding \( f_i = f_{i1}^1 \cdot f_{i2}^2 \cdot ... \cdot f_{iM}^M \). Since we want to finally integrate into one feature vector, we add the last layer of the convolutional neural network whose purpose is to change our number of channels from multidimensional to one-dimensional so that we can use the final one-channel feature map as our final feature map. channel feature map as the feature vector of our final target.

Since our deep matching framework is based on the idea of Deep Affinity Network (DAN), the obtained target embedding features are used to perform the similarity calculation. We set \( N_{\text{max}} = 80 \), and we construct a matrix \( E_{t-\eta} \in \mathbb{R}^{N_{\text{max}} \times N_{\text{max}}} \). In an image, we cannot get a fixed 80 targets, so we performed a filtering, if there is not enough we will fill the remaining positions with 0, and if there are more we exclude the redundant targets. For the obtained similarity matrix, we find the corresponding trajectory labeled with the same position set to 1, for different positions set to 0 so as to correlate, and calculate a similarity matrix \( A \in \mathbb{R}^{N_{\text{max}} \times N_{\text{max}}} \). Of course, since there is a target of matching error in matching, we extend the obtained A by deforming it to \( A' \in \mathbb{R}^{(N_{\text{max}}+1) \times (N_{\text{max}}+1)} \).

We define a set \( T \) of trajectories then for the detection and its bounding box \( b^t \) at time \( t \), we obtain a distance value between the trajectory and the detection according to the following equation.
\[ d(b^t, T) = \frac{1}{|T|} \sum_{b^{t-n} \in T} A' \]

\[ 3.3. \text{Motion Model} \]
We use the appearance model to model a similarity that is obtained. However, for targets with very similar appearance, it is not appropriate to use only appearance features for association. Therefore we established a geometric constraint. We used Kalman Filter \cite{31} for predicting the location of the target, and we obtained a distance value for the obtained bounding box location and the actual location by finding the IOU:

\[ d(b^t, b^{t-n}) = IOU(b^t, b^{t-n}) \]

\[ 3.4. \text{Training} \]
For training we construct a data module where we take a series of frames as input, and we randomly select the image data at temporal distance. We believe that in the practical use of multiple object tracking, we cannot guarantee that the target can appear in every frame. Therefore, we randomly select data within five frames for training, which can solve the problem of tracking failure caused by occlusion or target disappearance. Also for the association matrix we obtained using the deep association method, we construct the matching loss as follows:

\[ L_1 = \sum_{i}^{N_{\text{max}}} \sum_{j}^{N_{\text{max}} + 1} reg_p \log (A'_{i,j}) \]

\[ L_2 = \sum_{i}^{N_{\text{max}} + 1} \sum_{j}^{N_{\text{max}}} reg_n \log (A'_{i,j}) \]

\[ L_3 = \sum_{i}^{N_{\text{max}}} \sum_{j}^{N_{\text{max}}} reg_t \log (A'_{i,j}) \]

\[ L_4 = \frac{L_1 + L_2 + L_3 + L_4}{4} \]

\[ regp \text{ represents the real existing target of the current frame contains misclassification}, regn \text{ represents the real existing target of the next frame contains misclassification}, \text{and regt represents the target that does not contain misclassification.} \]

\[ 4. \text{Experiment} \]

\[ 4.1. \text{Experimental Implementation Details} \]
This method resizes the training image to 900*900, and also performs random cropping, flipping, image brightness, and saturation enhancement on the data. In the deep learning training we use SGD optimizer, we set the momentum to 0.9 and set the decay factor of learning rate $5e^{-4}$, we start with a learning rate of 0.01. As the training time increases the learning rate will be reduced to one tenth of the previous rate at the 50th, 80th and 100th times. We tested our algorithm on the open-source dataset MOT17.

\[ 4.2. \text{Datasets and Metrics} \]
MOT17: This dataset is a publicly available dataset of multi-target tracking dataset. The dataset contains both indoor and outdoor pedestrian tracking sequences. The frame rates of the videos are in the range of 14-30 FPS. The dataset contains seven training sets and seven test sets. The metrics contain MOTA-Multi-target tracking accuracy, MOTP-Multi-target tracking accuracy, IDF1-Feature F1 score MT-Most tracked ML-Most lost FP-False positive FN-FALSE negative IDS-ID mismatch. More details can be obtained from \cite{2}. Each sequence of the MOT17 dataset represents a different scene, and for each scene we divide it into two parts, one as a training set and one as a test set. The dataset provides the detection targets for each frame using SDP, Faster-RCNN and DPM.
4.3. Performance Analysis

Our algorithm is trained on the server. Above we show the results of our algorithm for association on the dataset in figure 2. In figure 3, the same color is assigned to the same target to indicate the association. We selected the top three frames of scene 1 and scene 6 for tracking, each track corresponds to a color, and the number on the head represents the label of the track and the last frame number that appeared. You can see that our algorithm can associate the targets well. We have selected some algorithms that are currently in the leaderboard for comparison, as shown in table 1, the upper arrows represent better with high parameters and the lower arrows represent better with low parameters, overall our algorithm has a high performance in terms of tracking accuracy.

![Figure 2. Association representation of algorithm.](image)

![Figure 3. Three consecutive frame algorithm demonstration results.](image)

| Tracker            | MOTA↑ | HOTA↑ | MT↑ | ML↓ | IDF1↑ | IDS↓ |
|--------------------|-------|-------|-----|-----|-------|------|
| HISP_DAL17 [32]    | 45.4  | 34.0  | 349 | 922 | 39.9  | 8727 |
| FPSN [33]          | 44.9  | 38.1  | 388 | 844 | 48.4  | 7136 |
| HISP_T17 [34]      | 44.6  | 33.3  | 355 | 913 | 38.8  | 10617|
| GMPHD_DAL [35]     | 44.4  | 31.4  | 350 | 927 | 36.2  | 11137|
| Ours               | 44.3  | 34.3  | 370 | 908 | 42.1  | 8332 |

5. Conclusion

In this paper, for feature extraction network only uses deep level features will lose the spatial information of the image. Also, as the camera moves, the change of observation angle will cause the
object to deform, so we introduce variable convolution to further optimize. Using only appearance features for association will greatly increase the probability of mismatch for similar targets. From this point of view, we introduce a motion model that uses Kalman filtering to predict the location of the target and match it, removing locations that are impossible for the target to reach in physical space. The advantage of our algorithm is that it can integrate low-level and high-level features, and introduce variable convolution to reduce tracking inaccuracies and mismatches caused by target deformation. It also introduces an end-to-end associative learning method, which greatly reduces the complexity of the algorithm. Our algorithm is tested on the MOT17 dataset and has a high performance.

References
[1] Impiombato D, Giarrusso S, Mineo T, Catalano O, Gargano C, La Rosa G, Russo F, Sottile G, Billotta S, Bonanno G, Garozzo S, Grillo A, Marano D and Romeo G 2015 You only look once: Unified, real-time object detection *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **794** 185-192.
[2] Luo W, Xing J, Milan A, Zhang X, Liu W, Zhao X and Kim T K 2017 Multiple object tracking: A literature review arXiv:1409.7618v4 pp 1-18.
[3] Felzenszwalb P, McAllester D and Ramanan D 2008 A discriminatively trained, multiscale, deformable part model *Proc. CVPR*.
[4] Nam H and Han B 2015 Learning multi-domain convolutional neural networks for visual tracking, *CVPR* pp 4293-4302.
[5] Bae S H and Yoon K J 2014 Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning *Proc. CVPR* pp 1218-1225.
[6] Z Fu, P Feng, S M Naqvi and J A Chambers 2017 Particle PHD filter based multi-target tracking using discriminative group structured dictionary learning *ICASSP* pp 4376-4380.
[7] Shafique K, Mun W L and Haering N 2008 A rank constrained continuous formulation of multi-frame multi-target tracking problem *Proc. CVPR*.
[8] Sun S J, Akhtar N, Song H S, Mian A S and Shah M 2021 Deep affinity network for multiple object tracking *IEEE Transactions on Pattern Analysis and Machine Intelligence* **43** (1) 104-119.
[9] Zhang L and Nevatia R 2008 Global data association for multi-object tracking using network flows *IEEE Conference on Computer Vision and Pattern Recognition* pp 1-8.
[10] Wang Z, Jiang X, Bo X, et al. 2015 An online multi-object tracking approach by adaptive labelling and Kalman filter *International Conference on Research in Adaptive and Convergent Systems* pp 146-151.
[11] Ge Z, Chang F and Liu H 2018 Multi-target tracking based on Kalman filtering and optical flow histogram *Chinese Automation Congress* pp 2540-2545.
[12] Z Zhao, S Yu and X Wu, 2009 A multi-target tracking algorithm using texture for realtime surveillance *IEEE International Conference on Robotics and Biomimetics* pp 2150-2155.
[13] M A Naiel, M O Ahmad, M N S Swamy, et al. 2017 Online multi-object tracking via robust collaborative model and sample selection *Computer Vision and Image Understanding* **154** 94-107.
[14] Eiselein V, Arp D, Patzold M, et al. 2012 Real-time multi-human tracking using a probability hypothesis density filter and multiple detectors *International Conference on Advanced Video and Signal-Based Surveillance* pp 325-330.
[15] Bae S H and Yoon K J 2014 Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning *IEEE Conference on Computer Vision and Pattern Recognition* pp 1218-1225.
[16] Bertinetto L, Valmadre J, Henriques J F, Vedaldi A and Torr P H 2016 Fully convolutional Siamese networks for object tracking *LNCS* **9914** 850-865.
[17] Wang B, Wang L, Shuai B, et al. 2016 Joint learning of convolutional neural networks and temporally constrained metrics for tracklet association. *IEEE Conference on Computer Vision & Pattern Recognition Workshops*.

[18] Bewley A, Ge Z, Ott L, et al. 2016 Simple online and realtime tracking. *arXiv:1602.00763*.

[19] Wojke N, Bewley A and Paulus D 2017 Simple online and realtime tracking with a deep association metric. *IEEE International Conference on Image Processing ICIP2017* pp 3645-3649.

[20] Kim C, Li F, Ciptadi A, et al. 2015 Multiple hypothesis tracking revisited. *Proceedings of the IEEE International Conference on Computer Vision ICCV2015* pp 4696-4704.

[21] Chen J, Sheng H, Zhang Y, et al. 2017 Enhancing detection model for multiple hypothesis tracking. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops CVPRW 2017* pp 18-27.

[22] Chu Q, Ouyang W, Li H, et al. 2017 Online multi-object tracking using CNN-based single object tracker with spatial-temporal attention mechanism. *Proceedings of the IEEE International Conference on Computer Vision ICCV 2017* pp 4836-4845.

[23] Zhu J, Yang H, Liu N, et al. 2018 Online multi-object tracking with dual matching attention networks. *Proceedings of the European Conference on Computer Vision ECCV 2018* pp 366-382.

[24] Leal-Taixé L, Canton-Ferrer C and Schindler K 2016 Learning by tracking: Siamese CNN for robust target association. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops CVPRW 2016* pp 33-40.

[25] Zhang J, Zhou S, Wang J, et al. 2019 Frame-wise motion and appearance for realtime multiple object tracking. *arXiv preprint arXiv: 1905.02292*.

[26] Xu Y, Ban Y, Alameda-Pineda X, et al. 2019 DeepMOT: A differentiable framework for training multiple object trackers. *arXiv preprint arXiv: 1906.06618*.

[27] Tang S, Andres B, Andriluka M, et al. 2015 Subgraph decomposition for multi-target-tracking. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[28] Emami P, Pardalos P M, Eleftheriadou L and S Ranka 2018 Machine learning methods for solving assignment problems in multi-target tracking. *arXiv:1802.06897*.

[29] Dai J, Qi H, Xiong Y, Li Y, Zhang G, Hu H and Wei Y 2017 Deformable convolutional networks. *IEEE International Conference on Computer Vision (ICCV)* pp 764-773.

[30] Yu F, Wang D, Shelhamer E and Darrell T 2018 Deep layer aggregation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* pp 2403-2412.

[31] Wojke N, Bewley A and Paulus D 2017 Simple online and realtime tracking with a deep association metric. *IEEE International Conference on Image Processing (ICIP)*.

[32] Baisa N 2019 Robust online multi-target visual tracking using a HISP filter with discriminative deep appearance learning. *arXiv:1908.03945*.

[33] Lee S and Kim E 2019 Multiple object tracking via feature pyramid Siamese networks. *IEEE ACCESS* 7 8181-8194.

[34] Baisa N 2018 Online multi-target visual tracking using a HISP filter. *Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5 (VISAPP)*.

[35] Baisa N 2019 Online multi-object visual tracking using a GM-PHD filter with deep appearance learning. *22th International Conference on Information Fusion (FUSION)*.