Enhancing Detection Model for Multiple Hypothesis Tracking

Jiahui Chen[1], Hao Sheng[1,2], Yang Zhang[1], Zhang Xiong[1,2]
[1]School of Computer Science and Engineering, Beihang University, Beijing, P.R. China
[2]Research Institute in Shenzhen, Beihang University, Shenzhen, P.R. China
{chenjh, shenghao, zhangyang1991, xiongz}@buaa.edu.cn

Abstract

Tracking-by-detection has become a popular tracking paradigm in recent years. Due to the fact that detections within this framework are regarded as points in the tracking process, it brings data association ambiguities, especially in crowded scenarios. To cope with this issue, we extended the multiple hypothesis tracking approach by incorporating a novel enhancing detection model that included detection-scene analysis and detection-detection analysis; the former models the scene by using dense confidential detections and handles false trajectories, while the latter estimates the correlations between individual detections and improves the ability to deal with close object hypotheses in crowded scenarios. Our approach was tested on the MOT16 benchmark and achieved competitive results with current state-of-the-art trackers.

1. Introduction

Multiple object tracking(MOT) automatically estimates the motion status of targets in video sequences, and it is widely applied in many fields, *e.g*. video surveillance[14], human-computer interaction[3], and robot vision[10]. Although a lot of progress was made in MOT, problems such as occlusion, crowded scenario, and illumination variation still persist. Till now, MOT is a challenging task in computer vision.

With the development of object detecting technology, tracking-by-detection becomes one of the most popular tracking frameworks. In this paradigm, the object hypotheses of each frame are detected by object detectors as part of the pre-process, and then the trackers apply data association algorithms to link these object hypotheses into trajectories. Since multiple hypothesis tracking can easily exploit high-order constraints, it is considered to be one of the most attractive tracking approaches.

Recent data association based trackers design energy functions in such a way that they model the tracking status, and find the optimal solution to get the final tracks. These methods help avoid local decisions and improve tracking performance using global optimization. However, since the input detections are often regarded as (center or foot) points in these methods, it brings data association ambiguities, especially in crowded scenarios.

(a) With detection-scene analysis (b) No detection-scene analysis

(c) With detection-detection analysis (d) No detection-detection analysis

Figure 1. Results of multiple hypothesis tracking with/without enhancing detection model. (a)(b):detection-scene analysis, which penalizes detections or tracks which do not fits the scene model, helps to solve false tracks with high confidential false postives. (c)(d):detection-detection analysis, which models the correlation between close targets, helps to increase the recall, especially when detections are overlapped.

In order to address this problem, we propose a novel enhancing detection model, which includes detection-scene analysis and detection-detection analysis. This approach models the scene and correlations between individual detections using analysis on our proposed dense confidential detection set, and following this, we incorporate the model into multiple hypothesis tracking. Tracking results using enhancing model are illustrated in Fig.1.

The main contributions of this paper include:

• A detection-scene analysis which models the scene by using our proposed dense confidential detection set and allows the tracker to handle false trajectories.

• A detection-detection analysis that estimates the correlations between individual detections. It improves the ability to deal with close object hypotheses in crowded scenarios.

• Multiple hypothesis tracking that incorporates our enhancing detection model and allows tracker to handle complex scenes with both fixed and moving cameras.

• A demonstration of our method on the MOT16 benchmark which achieves competitive results with current state-of-the-art trackers.

2. Related Work

Tracking-by-detection is a popular framework in multiple object tracking[28][17], in which object hypotheses are used as input, while trackers associate them into tracks over the whole sequence. Many data-association methods, such as the Hungarian algorithm[30], were introduced to solve tracking problem.

Milan *et al*. [23] proposed a conditional random field based framework, where detection-level and track-level exclusions were modeled to improve tracking performance in crowded scenarios. An extended method [24] obtained a set of tracklets first and then designed a discrete-continuous energy to reconstruct final trajectories. However, the performance of these two approaches relied on the quality of the tracklet proposals, and tracking errors in initial tracklets propagated to the final results.

Zhang *et al*. [31] and Butt *et al*. [2] proposed a min-cost flow based data association method for multiple object tracking, and the optimal solution was solved though either linear programming or Lagrangian methods. Chari *et al*. [3] proposed a pairwise cost to enforce or penalize tracklets, which effectively handled overlapping problems and tracking enhancements. [9] and [29] also followed the network flow framework. McLaughlin *et al*. [21] extended this min-cost network flow method so that the tracking problem was solved in a two-step scheme: 1) An initial result was estimated without motion information, and 2) It was then combined with a motion model to generate a more accurate solution. Accumulative errors could occur during this two-step scheme. More specifically, even though network flow based methods had the benefit of computational efficiency and optimality, the one drawback was that only unary and pairwise terms were taken into consideration.

Cox *et al*. [5] proposed classic multiple hypothesis tracking, in an effort to delay association decisions until they were resolved. However, the number of hypotheses grew exponentially. [13] incorporated appearance model to solve this issue. Then [16] proposed an online appearance model for multiple hypothesis tracking. In the final tracking results, false detections led to false positives.

Other methods used high-quality appearance model or features for tracking. Ma *et al*. [20] exploited a novel deep learning based features trained on object recognition datasets to improve tracking performance. Danelljan *et al*. [8] introduced a real-time tracking framework based on adaptive color channels. Though it handled different types of scenarios, it failed at scaling, so [7] proposed solution to this problem.

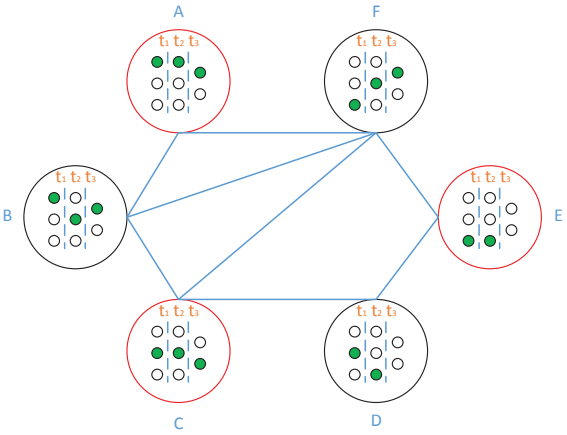


Figure 2. Maximum Weighted Independent Set. Hypotheses with common detection(s) are conflicting. Node A, C and E are selected as a solution in this example.

To summarize, although detections are fundamental to data association approaches, trackers often regard detections as points in the tracking process. This information loss led to final tracking errors. In this paper, we focus on this issue and propose an enhancing detection model, that understands the scene and mutual detection correlation through detection analysis. Moreover, the model is incorporated into multiple hypothesis tracking so that it improves the tracking performance in crowded scenarios with both fixed and moving cameras.

3. Multiple Hypothesis Tracking

We give a tradition formulation of multiple hypothesis tracking for multi-target tracking problem in this section. The details of extended framework are given in Sec.5.

Our goal is to estimate the trajectories of each target within a given sequence. As we adopt the tracking-by-detection framework, the input of the approach is a detection set provided by the object detector, such as DPM detector[12]. Let $\mathcal{D} = \{D_i^j\}$ be the detection set, and $D_i^j = (x_i^j, y_i^j, w_i^j, h_i^j, a_i^j, c_i^j)$ indicates j-th detection in frame i where (x, y) is the position, (w, h) is the scale, a is the appearance and c is the detection confidence. Based on these inputs, the problem is solved by a frame-by-frame method and at each frame, hypothesis updating, hypothesis formation, and hypothesis pruning are processed.

Given the hypothesis set $\mathcal{H}_i = \{H_i^1, H_i^2, ..., H_i^{N_i}\}$, the goal of new hypothesis updating is to estimate hypothe-

sis set $\mathcal{H}_{i+1} = \{H_{i+1}^1, H_{i+1}^2, ..., H_{i+1}^{N_{i+1}}\}$ based on tracking evidences, such as motion, appearance, where N_i is the number of the elements in hypothesis set $\mathcal{H}_i$. Each $H_i^j = \{D_1^{I_{i,1}^j}, D_2^{I_{i,2}^j}, ..., D_i^{I_{i,i}^j}\}$ presents a track(sequence of detections), where D_i^k can be either element in detection set $\mathcal{D}_i$ or dummy detection. Note that the actual starting time of the track is the index of the first detection which is not dummy detection, although all hypotheses start from the first frame. Moreover, each hypothesis H is assigned with a score s, which is defined as follows:

$$s(H) = w_m s_m + w_a s_a \tag{1}$$

,where s_m and s_a are the motion and appearance confidences, and w_m and w_a are their weights.

Hypothesis formation is applied after hypothesis updating, and it is the key process of multiple hypothesis tracking. Given hypothesis set $\mathcal{H}$, the goal of hypothesis formation is to find the most likely track set where tracks fits given constraints. In [16], the problem is formulated as k-dimensional assignment problem:

$$\max_z \ \Sigma_{i_1=1}^{N_1}\Sigma_{i_2=1}^{N_2}...\Sigma_{i_k=1}^{N_k} s_{i_1 i_2 ... i_k} z_{i_1 i_2 ... i_k} \tag{2}$$

$$s.t. \ \Sigma_{i_1=1}^{N_1}...\Sigma_{i_{u-1}=1}^{N_{u-1}}\Sigma_{i_{u+1}=1}^{N_{u+1}}...\Sigma_{i_k=1}^{N_k} z_{i_1 i_2 ... i_k} = 1 \tag{3}$$

$$for \ i_u \in \{1, 2, ..., N_u\}, u \in \{1, 2, ..., k\}$$

,where it has a constraint that each observation can belong to one track at most. It is then solved as a Maximum Weighted Independent Set(MWIS) problem as follows:

$$\max_x \Sigma_l w_l x_l \tag{4}$$

$$s.t. \ x_i + x_j \leq 1, \forall (i,j) \in E \tag{5}$$

$$x_l \in \{0, 1\} \tag{6}$$

,where E is the exclusion set. Each element $(H_i, H_j) \in E$ shows that H_i and H_j cannot be selected simultaneously. An example is shown in Fig.2.

Hypothesis pruning is the last process of each iteration and its purpose is to reduce the number of tracking hypotheses, because the number grows exponentially. An N-scan pruning approach is usually applied for this task[5].

To summarize, multiple hypothesis tracking converts tracking problem into hypothesis generation and selection problems, and provides a flexible framework to model complex mutual exclusion models. Due to this property, we incorporate our enhancing detection model into multiple hypothesis tracking to improve the tracking performance. The enhancing detection model is introduced in later sections.

4. Enhancing Detection Model

Previous sections provide a brief overview of the tracking-by-detection framework. One drawback of current MOT approaches is that detections are regarded as points, and scales are used for non-maximum suppression or detection similarity calculation without position. This paper focuses on digging out tracking evidences from the input detection set, including scene understanding and mutual detection correlation. It is noted that in this paper we use coordinate of foot(middle-bottom) point to describe the position of a detection.

To address this issue, we propose a novel enhancing detection model. Our enhancing detection model consists of two key components: detection-scene analysis and detection-detection analysis. Detection-scene analysis understands the scene information based on input detections, and detection-detection analysis describes the mutual relationship between two individual detections. Details of detection-scene model and detection-detection model are introduced in Sec.4.1 and Sec.4.2 respectively.

4.1. Detection-scene Analysis

The goal for detection-scene analysis is to understand the correlation between detection and scene using input detections. Our initial motivation is to model the correlation between detection and scene so we propose a detection-scene mapping $\mathcal{M} : (x, y) \rightarrow h$ which describes the height variation with the position in the scene. To get a more reliable model, we use the foot points instead of center points to represent the positions of detections so that the detection coordinates are in close positions to the ground plane where objects stand. Compared to using the center ones, this method removes the impact of object height. It is noted that one assumption of our detection-scene analysis is that height and viewpoint of a fixed or moving camera does not change significantly over the time so targets at a certain position in the image space have similar scale based on perspective distortion, even when the camera is moving.

Given detection set $\mathcal{D} = \{D_i^j\}$, where D_i^j indicates j-th detection in frame i. Each detection $D \in \mathcal{D}$ is defined as $D = (x, y, w, h, a, c)$, where (x, y) is the position, (w, h) is the scale, a is the appearance and c is the confidence given by detector.

Data Preprocess. As detection set is generated by an object detector, it contains many noisy detections, so in order to handle the detection noise problem, a data preprocess is needed and we propose a concept of dense confidential detection set.

Let $\mathcal{D}$ be the original detection set, then we first generate the new confidential detection set $\mathcal{D}_{con}$. The confidential detection set $\mathcal{D}_{con}$ is defined as follows:

$$\mathcal{D}_{con} = \{D \in \mathcal{D} | c >= th_{con}\} \tag{7}$$

,where c is the confidence score of detection D and th_{con} is the threshold.

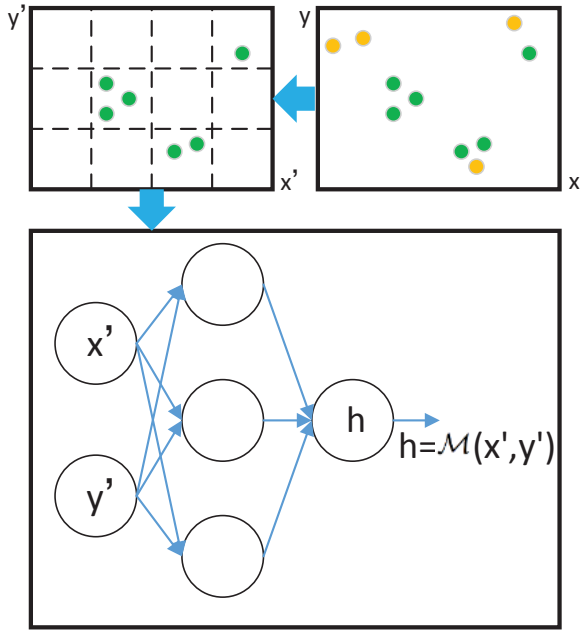


Figure 3. Workflow of Detection-scene Modeling. Detections with high confidence are in green and detections with low confidence are in yellow. x' and y' are the coordinates in dense confidential detection set.

As many detections with low confidence scores are removed from the original detection set, sparse data problems for our scene estimation may occur.

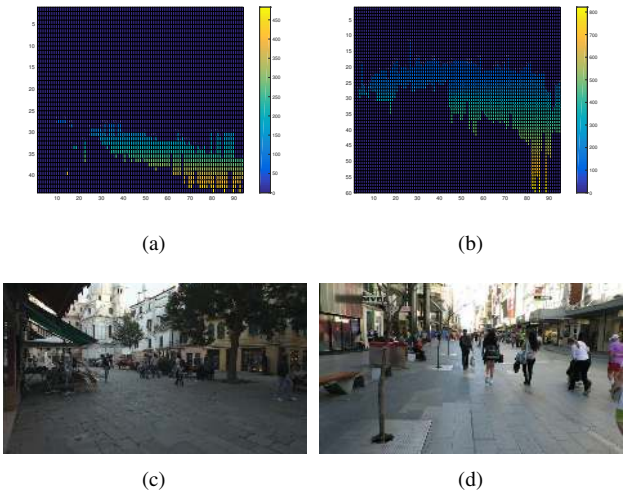
To handle this issue and get a smooth position-height mapping, we divide the whole frame into $M * N$ patches, and each patch has the same size $W_p * H_p$. Therefore, new detection $D' = (x', y', w, h, a, c)$ in dense confidential detection set $\mathcal{D}'_{con}$ is defined based on detection $D = (x, y, w, h, a, c)$ in detection set $\mathcal{D}_{con}$ as follows:

$$\begin{cases} x' = \lfloor \frac{x}{w} \rfloor + 1 \\ y' = \lfloor \frac{y}{h} \rfloor + 1 \end{cases} \tag{8}$$

,while the scale, appearance and confidence remain the same.

The processes presented below are based on the dense confidential detection set $\mathcal{D}'_{con}$, as shown in Fig.3.

Mapping formation. Since the scale of detection in the image plant varies with the position based on perspective distortion and our model is designed to describe this correlation between detection and scene, so as to fit the position-height surface in the image plane using dense confidential detection set $\mathcal{D}'_{con}$. However, due to the complex scenarios and cameras status(fixed and moving), it is not easy to choose a certain (polynomial) function to model the scale variation so we resort to using a neural network with k hidden layer with universality property to fitting our detection-scene model[6]. Note that we choose the sigmoid function as the active function. Position (x', y') is used as input and height h is used as output. To optimize the minimum of a multivariate function, we apply the iterative technique known as the Levenberg-Marquardt algorithm.



(a) (b) (c) (d)

Figure 4. Result of Detection-scene Model. (a) is the detection-scene model of (c), which is a sequence of fixed camera. (b) is the detection-scene model of (d), which is a sequence of moving camera. The coordinates of (a) and (b) are based on the foot point of detections and the value is the estimated height in pixel.

As a result, we get our detection-scene mapping $\mathcal{M}(x, y)$, where x and y is calculated according to Eq.8. Estimated results of both fixed and moving cameras are shown in Fig.4

4.2. Detection-detection Analysis

Object detectors, such as the DPM detector, often produce more than one response for single target. Current trackers typically solve this problem by adopting non-maximum suppression as part of the pre-process, which keeps the detection with high confidence when overlapping happens. Although NMS works well when tracking sparse scenarios, it always leads to the low recall or fragment problem when tracking objects in crowded sequences. This is due to the fact that NMS is a method based on local decisions instead of global optimization, and in crowded scenes, multiple targets appear close to each other. Detection-detection analysis is proposed to encourage overlapping detections when detections are likely to be different targets or to discourage them when at least one of them is supposed to be a false detection. Detection-detection analysis is based on proposed detection-scene analysis.

Given the detection-scene mapping $\mathcal{M}(x, y)$ introduced in Sec.4.1, the objective of detection-scene mapping is to represent the correlation between individual detections, especially for overlapping detections. For this purpose, we propose that probability $P(D_1 \in \mathcal{T}, D_2 \in \mathcal{T} | D_1, D_2)$,

where $D \in \mathcal{T}$ means detection D is a real detection and belongs to track $\mathcal{T}$. Based on the Bayesian inference, we have

$$\begin{aligned} &P(D_1 \in \mathcal{T}, D_2 \in \mathcal{T} | D_1, D_2) \\ =&\frac{P(D_1, D_2 | D_1 \in \mathcal{T}, D_2 \in \mathcal{T}) P(D_1 \in \mathcal{T}, D_2 \in \mathcal{T})}{P(D_1, D_2)} \end{aligned} \quad (9)$$

, where $P(D_1, D_2 | D_1 \in \mathcal{T}, D_2 \in \mathcal{T})$ is the probability that detection D_1 and D_2 are detected simultaneously assuming these two detections are real targets, $P(D_1 \in \mathcal{T}, D_2 \in \mathcal{T})$ are the probability that both D_1 and D_2 are true detections and $P(D_1, D_2)$ is the probability that D_1 and D_2 are detected simultaneously.

Assuming two objects occupy the exact positions of D_1 and D_2, the probability that two objects are simultaneously detected by the detector is based on the overlapping area. Thus, $P(D_1, D_2 | D_1 \in \mathcal{T}, D_2 \in \mathcal{T})$ is defined as follows:

$$P(D_1, D_2 | D_1 \in \mathcal{T}, D_2 \in \mathcal{T}) = 1 - \frac{D_1 \cap D_2}{D1 \cup D2} \quad (10)$$

, where $D_1 \cap D_2$ is the intersection area of $D1$ and $D2$ and $D1 \cup D2$ is the union area of $D1$ and $D2$. Eq.10 implies that when the less amount of area is overlapped, the more likely the chance that they would be simultaneously detected.

As detections are indpendently generated by the object detector, so we have

$$P(D_1 \in \mathcal{T}, D_2 \in \mathcal{T}) = P(D_1 \in \mathcal{T}) P(D_2 \in \mathcal{T}) \quad (11)$$

We define the probability $P(D \in \mathcal{T})$ that detection $D = (x, y, w, h, a, c)$ is a true detection. Based on our detection-scene model $\mathcal{M}(x, y)$.

$$P(D \in \mathcal{T}) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(h - \mathcal{M}(x, y))^2}{2\sigma^2}) \quad (12)$$

, where (x, y) is the coordinate of detection D in image space and h is the height.

5. MHT Using Enhancing Detection Model

In the previous section, we describe our proposed enhancing detection model. In this section, we show how enhancing detection model, including detection-scene analysis and detection-detection analysis, could be incorporated in multiple hypothesis tracking. We improve hypothesis updating and hypothesis pruning by using detection-scene analysis, which is introduced in Sec.5.1 and Sec.5.3. Moreover, detection-detection analysis is applied in hypothesis formation, which is discussed in Sec.5.2.

5.1. Hypothesis Updating with Detection-scene Model

Hypothesis updating generates new hypothesis set $\mathcal{H}_i$ at frame i based on previous hypothesis set $\mathcal{H}_{i-1}$. New hypotheses are divided into two categories as traditional multiple hypothesis tracking: one extends the original hypotheses by adding new observations which fit tracking models, such as motion model and appearance model, while another one regards each new observation as a new object.

After new hypothesis set $\mathcal{H}_i$ is obtained, each hypothesis is needed to assigned with a score to provide the evidence for hypothesis formation. We introduce a new term $w_{ds} s_{ds}$ to Eq.1, the score of hypothesis $H \in \mathcal{H}$ is defined as follows:

$$s(H) = w_m s_m + w_a s_a + w_{ds} s_{ds} \quad (13)$$

, where w_m, w_a and w_{ds} are respectively the weights of motion, appearance and detection-scene factor, and s_m, s_a, s_{ds} are the scores of motion, apperance and detection-scene factor, respectively.

We apply the method mentioned in [16] for scores of motion and appearance. Assuming D is the newest observation of hypothesis H, motion score is defined as follows:

$$s_m = ln \frac{P(D \in H)}{P(D \notin H)} \quad (14)$$

, where $P(D \in H)$ is the probability that detection D is a true detection in H, and $P(D \notin H)$ is the probability that detection D is a false positive in H. Based on previous observations in H, the expected position $(\hat{x}, \hat{y})$ can be estimated by Kalman Filter, so $P(D \in H)$ is defined as follows:

$$P(D \in H) = \mathcal{N}((x, y) | (\hat{x}, \hat{y}), \Sigma xy) \quad (15)$$

, where $\mathcal{N}$ is the normal distribution with $\mu = (\hat{x}, \hat{y})$ and $\Sigma = \Sigma xy$.

The probability $P(D \notin H)$ is defined as a constant:

$$P(D \notin H) = \frac{1}{V} \quad (16)$$

, where V is the pixel number in a frame.

Similar to score of motion, the appearance score is defined as follows:

$$s_a = ln \frac{P(D \in H)}{p(D \notin H)} \quad (17)$$

, where D is current observation of hypothesis H. Based on previous observations, the appearance similarity $S(D, H)$ is calculated using Multi-output Regularized Least Squares(MORLS)[16]. Then $P(D \in H)$ is defined as follows:

$$P(D \in H) = S(D, H) \quad (18)$$

And $P(D \notin H)$ is defined as constant:

$$P(D \notin H) = c_1 \tag{19}$$

Score of detection-scene factor describes how the hypothesis H fits the detection-scene model. Based on Eq.12, s_{ds} is defined as follows:

$$\begin{aligned} s_{ds} &= ln(P(D \in H)) \\ &= ln(\frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(h - \mathcal{M}(x,y))^2}{2\sigma^2})) \\ &= c_2 - (h - \mathcal{M}(x,y))^2 \end{aligned} \tag{20}$$

, where (x, y) is calculated according to Eq.8, and $\mathcal{M}$ is the detection-scene mapping.

5.2. Hypothesis Formation with Detection-detection Model

Hypothesis formation converts hypothesis selection problem into maximum weight independent set problem, and the key component is to model its mutual exclusion set.

Given hypothesis set $\mathcal{H}$, let $s(H)$ be the score of hypothesis H where $H \in \mathcal{H}$. In a traditional multiple hypothesis tracking, the first constraint is that a detection can only belong to at most one track. Then the unique-detection based mutual exclusion set is defined as follows:

$$E_{ud} = \{(H_1, H_2) | H_1 \cap H_2 \neq \emptyset\} \tag{21}$$

, where H_1 and H_2 are two hypotheses. If H_1 and H_2 have common observation(s), then these two detections cannot be simultaneously selected.

Another mutual exclusion set is based on our detection-detection analysis. We models the probability $P(D_1 \in \mathcal{H}, D_2 \in \mathcal{H} | D_1, D_2)$ in Sec.4.2. The probability represents the likelihood that two detections D_1 and D_2 are real detections. Therefore, when the probability is high, it means D_1 and D_2 are likely to be two targets and both detections can be selected. Otherwise, at least one of them is supposed to be a false detection, and we penalize the pair. Our detection-detection analysis based mutual exclusion set is defined as follows:

$$E_{dd} = \{(H_1, H_2) | P(D_1 \in \mathcal{H}, D_2 \in \mathcal{H} | D_1, D_2) < th_{dd}\} \tag{22}$$

, where th_{dd} is a threshold.

Then, the maximum weight independent set problem is formulated as follows:

$$\max_x \ \Sigma_l s(H_l) x_l \tag{23}$$

$$s.t. \ x_i + x_j \leq 1, \forall (i,j) \in E_{ud} \tag{24}$$

$$x_i + x_j \leq 1, \forall (i,j) \in E_{dd} \tag{25}$$

$$x_l \in \{0, 1\} \tag{26}$$

, where x_l is the indicator of H_l. When $x_l = 1$, it indicates H_l is selected. Moreover, this optimization is solved by [25].

5.3. Hypothesis Pruning with Detection-detection Model

In this subsection, we discuss how our proposed model improve the efficiency of hypothesis pruning.

In the hypothesis pruning process, besides standard N-scan pruning approach introduced in Sec.3 is applied, we also prune hypotheses according to our proposed detection-scene analysis. For each detection $D = (x, y, w, h, a, c)$, relative height of detection h_{rd} is defined as follows:

$$h_r(D) = \frac{h}{\mathcal{M}(x,y)} \tag{27}$$

, where $\mathcal{M}$ is our detection-scene mapping. Then the relative height of hypothsis $h_r(H)$ is defined as follows:

$$h_r(H) = \frac{\Sigma_{i=1}^{N} h_r(D_i)}{N} \tag{28}$$

, where D_i is the $i-$th observation in H, and N is the number of observations in H. Then, pruning hypothesis set $\mathcal{D}$ is given:

$$\mathcal{D} = \{H | H \in \mathcal{H}, h_r(H) < th_{low} \ or \ r_r(H) > th_{high}\} \tag{29}$$

Then hypotheses in $\mathcal{D}$ are also removed in each iteration.

6. Experiments

Dataset. We tested our approach on the MOT16 Benchmark[22] and achieved very competitive results. There are seven training sequences and seven test sequences in the benchmark. We first demonstrate the evaluation results on training sequences to verify the effectiveness of our framework in Sec.6.2. Moreover, in Sec.6.3, we compare our method with other state-of-art tracking methods. For a fair comparison, we use the public detection set given by MOT16 as our algorithm input. So all tracking approaches are based on the same input.

Implementation Details. We trained a convolution neural network to extract the appearance feature and in our implementation, the network contains three convolutional layers and six inception modules. The input size is 14456, and the output layer is a fully-connected layer which outputs the 256-dimensional feature. In our tracking method, detections are resized to 144 56 for feature extraction. For training the convolution neural network, we use data from CUHK03[19], Market-1501[32], CUHK01[18], VIPeR[26] and i-LIDS[26]. We use the softmax classifier during training.

Parameters. th_{con} in Eq.7 is 0. w_m, w_a and w_{ds} in Eq.13 are repectively 0.7, 0,1 and 0.2. c_1 in Eq.19 is 0.3. th_{low} and th_{high} in Eq.29 are repectively 0.6 and 1.8.

Evaluation metrics. We use CLEAR MOT to measure the tracking results:

Table 1. Results on training sequences

Method	Rcll↑	Prcn↑	GT	MT↑	ML↓	FP↓	FN↓	IDs↓	FM↓	MOTA↑	MOTP↑
baseline	38.4	**93.4**	517	62	284	**3014**	68044	**228**	**312**	35.4	**78.7**
NMS with high threshold	**45.9**	74.7	517	87	**226**	17195	**59675**	340	463	30.1	77.7
+detection-detection analysis	45.7	85.4	517	84	232	8653	59957	331	502	37.6	77.7
+detection-scene scoring	45.6	87.3	517	**88**	230	7347	60009	334	483	38.7	77.7
+detection-scene pruning	45.1	89.0	517	**88**	237	6187	60580	310	455	**39.2**	77.8

Table 2. Results from 2D MOT 2016 Challenge(accessed on 03/28/2017)

Method	Rcll↑	Prcn↑	GT	MT↑	ML↓	FP↓	FN↓	IDs↓	FM↓	MOTA↑	MOTP↑
NOMT[4]	-	-	759	**139**	314	9753	**87565**	**359**	**504**	**46.4**	**76.6**
JMC[28]	-	-	759	118	**301**	6373	90914	657	1114	46.3	75.7
olCF[15]	-	-	759	86	368	6651	96515	381	1404	43.2	74.3
MHT_DAM[16]	-	-	759	103	356	**5668**	97919	499	659	42.9	**76.6**
LINF[11]	-	-	759	88	389	7896	99224	430	963	41.0	74.8
EAMTT[27]	-	-	759	60	373	8114	102452	965	1657	38.8	75.1
OVBT[1]	-	-	759	57	359	11517	99463	1321	2140	38.4	75.4
LTTSC-CRF[17]	-	-	759	73	419	11969	101343	481	1012	37.6	75.9
ours(EDMT)	51.8	89.5	759	129	303	11122	87890	639	946	45.3	75.9

Multiple object tracking precision (MOTP ↑), Multiple object tracking accuracy (MOTA ↑), Recall(Rcll ↑), Precision(Prcn ↑), the number of mostly tracked trajectories (MT ↑), the number of mostly lost trajectories(ML ↓), the number of false positives (FP ↓), the number of false negatives (FN ↓), the number of identity switching (IDs ↓) and the number of trajectory fragments (FM ↓).

The symbol ↑ is a positive indicator that means that the higher the value, the better, while ↓ means the lower the value, the better.

6.1. Computational Time

We implemented our approach in Matlab without code optimization or parallelization and tested it on a PC with 3.0GHz CPU and 16 GB memory. It took 4345.2 seconds for all seven training sequences and 3207.3 seconds for all seven testing sequences. Note that the time of object detection and appearance feature extraction are not included.

6.2. Framework Verification

We first verify our method on MOT16 training sequences. One baseline method, three intermediate results and final result are shown in Tab.1.

The baseline method is the implementation of traditional multiple hypothesis tracking[16]. To verify the ability to handle crowded scenes and avoid the errors caused by non-maximum suppression, we first increase the threshold of non-maximum suppression so as to remove fewer detections. As a result, the recall increases from 38.4% to 45.9%, while false positive grows from 3014 to 17195. MOTA decreases by 5.3%, because the high threshold of non-maximum suppression brings too many false trajectories.

Firstly, detection-detection analysis is introduced to hypothesis formation, and the result is shown as '+detection-detection analysis'. Compared to the previous result, false positive drops to almost half, while false negative slightly increases by 0.5%. Moreover, the recall keeps a high-level while MOTA dramatically increases from 30.1% to 37.6%, higher than the baseline method. The results prove that our detection-detection analysis can handle noisy detections in crowded scenes more effectively.

After we incorporate detection-scene analysis based scoring into the approach, the result is shown as '+detection-scene scoring' and we find the false positive keep falling while false negative remains stable. Moreover, more trajectories become mostly tracked, and fewer ones are mostly lost. Then, the MOTA reaches 38.7%, and our results prove that our detection-scene model can help to penalize false detections and improve the tracking performance.

Finally, detection-scene analysis based hypothesis pruning is added into tracking. Now the whole enhancing detection model is incorporated into the tracker, and the result is shown as '+detection-scene pruning'. Almost all of the evaluation indicators get the best of all these five results. The reason is that our novel pruning method can remove the trajectories which are not reasonable in the scenes.

Compared to the original baseline method, our final MOTA grew from 35.4% to 39.2%. It proves our effectiveness of our model.

Table 3. Results on testing sequences

Sequence	Method	Rcll↑	Prcn↑	GT	MT↑	ML↓	FP↓	FN↓	IDs↓	FM↓	MOTA↑	MOTP↑
MOT16-01	MHT_DAM[16]	-	-	23	**6**	11	164	4294	**15**	**30**	30.1	**72.8**
	ours(EDMT)	37.8	94.2	23	**6**	**10**	**150**	**3975**	**15**	41	**35.3**	72.3
MOT16-03	MHT_DAM[16]	-	-	148	28	40	**3591**	49521	**230**	**304**	49.0	**76.5**
	ours(EDMT)	58.4	89.5	148	**36**	**28**	7182	**43487**	319	479	**51.2**	75.9
MOT16-06	MHT_DAM[16]	-	-	221	35	118	**247**	5840	62	**84**	46.7	**75.1**
	ours(EDMT)	54.0	93.5	221	**46**	**104**	430	**5308**	**53**	91	**49.8**	74.6
MOT16-07	MHT_DAM[16]	-	-	54	6	23	**408**	9896	**39**	**60**	36.6	**75.8**
	ours(EDMT)	50.3	93.0	54	**8**	**15**	614	**8114**	72	103	**46.1**	74.9
MOT16-08	MHT_DAM[16]	-	-	63	**8**	28	**331**	10903	**76**	**77**	**32.4**	**80.9**
	ours(EDMT)	37.9	88.2	63	7	**21**	846	**10396**	85	87	32.3	80.0
MOT16-12	MHT_DAM[16]	-	-	86	14	43	**266**	4480	**15**	**27**	42.6	**78.5**
	ours(EDMT)	52.4	85.5	86	**17**	**35**	739	**3948**	31	40	**43.1**	77.7
MOT16-14	MHT_DAM[16]	-	-	164	6	93	**661**	12985	**62**	**77**	**25.8**	**75.6**
	ours(EDMT)	31.5	83.4	164	**9**	**90**	1161	**12662**	64	105	24.9	75.0

6.3. MOT16 Benchmark Comparison

Our approach are tested on the MOT16 Benchmark which contains seven training sequences and seven testing sequences. The parameters are tuned on the training set and the final result of testing sequences is submitted to the benchmark. Tab.2 shows the quantitative evaluations of our approach and the best previous published approaches on MOT16 benchmark. The comparison is also found in the MOT Challenge website, and our tracker is named EDMT(Enhancing Detection Model based Tracker). Our tracker achieved competitive results as opposed to the published state-of-the-art trackers.

[16] is the best published multiple hypotheses tracking method, and we also follow this framework. Compared to [16], our tracking result has significant improvements and outperforms [16] by 2.4% on MOTA.

Tab.3 demonstrates the performance of our approach with the best published multiple hypothesis tracking method[16] on each testing sequence. Compared to MHT_DAM[16], our tracking results have significant improvement for most testing sequences, except MOT16-08 and MOT16-14. Especially, on sequence MOT16-07 our MOTA increase by 9.5%. On MT, ML, and FN, our method is better for all testing sequences. The results prove that our enhancing detection model can considerably benefit multiple hypothesis tracking.

7. Conclusion

This paper proposed a novel enhancing detection model that simultaneously modeled the detection-scene and detection-detection correlation. The detection-scene analysis modeled the scene by using our proposed dense confidential detection set and it allowed the tracker to handle false trajectories. In addition to that, the detection-detection analysis estimated the correlations between individual detections, which improved the ability to deal with close object hypotheses in crowded scenarios. We incorporated our model into the multiple hypothesis tracking, by adding a new term to hypothesis scoring, mutual detection exclusion constraints and detection-scene strategy pruning, all of which helped to handle complex scenes with both fixed and moving cameras. The results on the MOT16 benchmark were provided. We first verified the effectiveness of each process on training sequences, followed by a comparison with the best published trackers is shown, and we achieved competitive performance. We also demonstrated our method with the best published multiple hypothesis tracking on the MOT16 benchmark, and the results proved that our model significantly improved the tracking results, which are available on the MOT16 website.

In future work, we plan to integrate more constraints, including detection-detection, detection-tracklet and tracklet-tracklet correlations, into multiple hypothesis tracking to achieve a better tracking performance, thereby further raising the tracking performance in complex scenarios. Moreover, our proposed enhancing detection model can also try to be incorporated into other tracking approaches to handle the detection information loss problem, and it may also benefit the tracking performance.

Acknowledgement: This study is partially supported by the National Natural Science Foundation of China(No.61472019), the National Science & Technology Pillar Program (No.2015BAF14B01), the Programme of Introducing Talents of Discipline to Universities, the Open Fund of the State Key Laboratory of Software Development Environment under grant #SKLSDE-2015ZX-21 and HAWK-EYE Group.

References

[1] Y. Ban, S. Ba, X. Alameda-Pineda, and R. Horaud. Tracking multiple persons based on a variational bayesian model. In *European Conference on Computer Vision*, pages 52–67. Springer, 2016.

[2] A. A. Butt and R. T. Collins. Multi-target tracking by lagrangian relaxation to min-cost network flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1846–1853, 2013.

[3] V. Chari, S. Lacoste-Julien, I. Laptev, and J. Sivic. On pairwise costs for network flow multi-object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5537–5545, 2015.

[4] W. Choi. Near-online multi-target tracking with aggregated local flow descriptor. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3029–3037, 2015.

[5] I. J. Cox and S. L. Hingorani. An efficient implementation of reid's multiple hypothesis tracking algorithm and its evaluation for the purpose of visual tracking. *IEEE Transactions on pattern analysis and machine intelligence*, 18(2):138–150, 1996.

[6] B. C. Csáji. Approximation with artificial neural networks. *Faculty of Sciences, Etvs Lornd University, Hungary*, 24:48, 2001.

[7] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg. Discriminative scale space tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.

[8] M. Danelljan, F. Shahbaz Khan, M. Felsberg, and J. Van de Weijer. Adaptive color attributes for real-time visual tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1090–1097, 2014.

[9] A. Dehghan, Y. Tian, P. H. Torr, and M. Shah. Target identity-aware network flow for online multiple target tracking. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1146–1154. IEEE, 2015.

[10] A. Ess, B. Leibe, K. Schindler, and L. Van Gool. A mobile vision system for robust multi-person tracking. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[11] L. Fagot-Bouquet, R. Audigier, Y. Dhome, and F. Lerasle. Improving multi-frame data association with sparse representations for robust near-online multi-object tracking. In *European Conference on Computer Vision*, pages 774–790. Springer, 2016.

[12] P. F. Felzenszwalb, R. B. Girshick, D. Mcallester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Software Engineering*, 32(9):1627–45, 2010.

[13] M. Han, W. Xu, H. Tao, and Y. Gong. An algorithm for multiple object trajectory tracking. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE, 2004.

[14] X.-Y. Jing, X. Zhu, F. Wu, X. You, Q. Liu, D. Yue, R. Hu, and B. Xu. Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 695–704, 2015.

[15] H. Kieritz, S. Becker, W. Hübner, and M. Arens. Online multi-person tracking using integral channel features. In *Advanced Video and Signal Based Surveillance (AVSS), 2016 13th IEEE International Conference on*, pages 122–130. IEEE, 2016.

[16] C. Kim, F. Li, A. Ciptadi, and J. M. Rehg. Multiple hypothesis tracking revisited. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4696–4704, 2015.

[17] N. Le, A. Heili, and J.-M. Odobez. Long-term time-sensitive costs for crf-based tracking by detection. In *European Conference on Computer Vision*, pages 43–51. Springer, 2016.

[18] W. Li, R. Zhao, and X. Wang. Human reidentification with transferred metric learning. In *Asian Conference on Computer Vision*, pages 31–44. Springer, 2012.

[19] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159, 2014.

[20] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang. Hierarchical convolutional features for visual tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3074–3082, 2015.

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

[22] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016.

[23] A. Milan, K. Schindler, and S. Roth. Detection-and trajectory-level exclusion in multiple object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3682–3689, 2013.

[24] A. Milan, K. Schindler, and S. Roth. Multi-target tracking by discrete-continuous energy minimization. *IEEE transactions on pattern analysis and machine intelligence*, 38(10):2054–2068, 2016.

[25] P. R. Östergård. A new algorithm for the maximum-weight clique problem. *Nordic Journal of Computing*, 8(4):424–436, 2001.

[26] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary. Person re-identification by support vector ranking. In *BMVC*, volume 2, page 6, 2010.

[27] R. Sanchez-Matilla, F. Poiesi, and A. Cavallaro. Online multi-target tracking with strong and weak detections. In *European Conference on Computer Vision*, pages 84–99. Springer, 2016.

[28] S. Tang, B. Andres, M. Andriluka, and B. Schiele. Multi-person tracking by multicut and deep matching. In *European Conference on Computer Vision*, pages 100–111. Springer, 2016.

[29] X. Wang, E. Turetken, F. Fleuret, and P. Fua. Tracking interacting objects using intertwined flows. 2015.

[30] Y. Xiang, A. Alahi, and S. Savarese. Learning to track: Online multi-object tracking by decision making. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4705–4713, 2015.

[31] L. Zhang, Y. Li, and R. Nevatia. Global data association for multi-object tracking using network flows. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[32] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *Computer Vision, IEEE International Conference on*, 2015.