SQE: a Self Quality Evaluation Metric for Parameters Optimization in Multi-Object Tracking

Yanru Huang[1] Feiyu Zhu[2] Zheni Zeng[1*] Xi Qiu[2] Yuan Shen[1†] Jianan Wu[2]
[1]Tsinghua University [2]Megvii Inc.
{huangyr18, zengzn16}@mails.tsinghua.edu.cn shenyuan_ee@tsinghua.edu.cn
{zhufeiyu, qiuxi, wjn}@megvii.com

Abstract

We present a novel self quality evaluation metric SQE *for parameters optimization in the challenging yet critical multi-object tracking task. Current evaluation metrics all require annotated ground truth, thus will fail in the test environment and realistic circumstances prohibiting further optimization after training. By contrast, our metric reflects the internal characteristics of trajectory hypotheses and measures tracking performance without ground truth. We demonstrate that trajectories with different qualities exhibit different single or multiple peaks over feature distance distribution, inspiring us to design a simple yet effective method to assess the quality of trajectories using a two-class Gaussian mixture model. Experiments mainly on MOT16 Challenge data sets verify the effectiveness of our method in both correlating with existing metrics and enabling parameters self-optimization to achieve better performance. We believe that our conclusions and method are inspiring for future multi-object tracking in practice.*

1. Introduction

Multi-object tracking (MOT) aims to track all objects of interest categories in a video sequence [18, 26]. It is crucial in applications like video surveillance and autonomous driving, where multiple pedestrians and vehicles need to be tracked simultaneously [6, 25, 3]. In recent years, tracking-by-detection [9, 18, 23, 6, 1, 5] has become the predominant paradigm of MOT. This approach first detects objects in each frame, then extracts discriminative features to quantify the similarities between targets, and finally perform data association to assign detections into their most likely trajectories. During this process, several influential parameters need to be set manually, such as the threshold determining whether to establish associations. To find the optimal parameters, an evaluation procedure is needed to measure the tracking performance. However, existing evaluation metrics like event-based measures CLEAR MOT [3] or identity-based measures IDF_1 [16] all require ground truth annotations, limiting the optimization to training data. Since the optimized parameters could be sub-optimal in test scenes, a self evaluation metric that enables parameters optimization without ground truth is urgently needed.

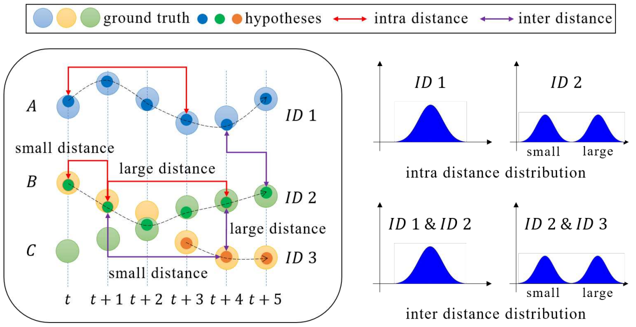


Figure 1. Examples of different intra and inter distance distributions. Ideally, the distance distribution within the same trajectory or between two different trajectories present a single peak with little noise. If false identification takes place, such as the hypothetical identity 2 follows target B at first and then switches to another target C, the distance distribution would present multiple distinguishable peaks. See Section 3.2 for a detailed explanation.

To evaluate the accuracy and stability of a tracker without ground truth, we design a self quality evaluation metric SQE that considers the quantity, length, and feature distance information of the trajectory hypotheses comprehensively. Our method can assess the quality of trajectories owing to the distinctive distance distribution forms as shown in Figure 1. The intra distance denotes the feature distance between each two detection boxes in the same trajectory, and all pairs constitute the intra distance distribution. Similarly, the inter distance denotes the feature distance between each two detection boxes from different trajectories. Intuitively, when a trajectory contains different targets, the distance dis-

[*] This work was done during Zheni Zeng was an intern at Megvii Inc.
[†] Corresponding author.

tribution scatters, and we demonstrate that it has the general characteristic of multiple peaks.

SQE enables automatic parameters adaptation to accommodate different scenes. Designing a tracking algorithm that performs well under various video scenes is hard, yet tuning parameters in existing tracking algorithms can achieve equally outstanding performance in an easier manner. To the best of our knowledge, there is no previous work in this area to date. We believe that our approach is exceedingly instructive and provides new ideas for future research.

In summary, our contributions are as follows: (1) We show that feature distance distributions can reflect trajectory hypotheses quality; (2) We propose a self quality evaluation metric SQE based on two-class Gaussian mixture model, which can primarily fulfill the self-evaluation desire; (3) We test the effectiveness of our method on various data sets and note its drawbacks. A future prospect of using distributions to estimate erroneous frames is discussed in the end.

2. Related Work

2.1. MOT algorithms

In the tracking-by-detection paradigm, trackers first detect objects in each frame and then associate detections over time to form trajectories for targeted individuals [9, 26, 12]. Online methods [14, 24, 4, 23] only use previous and current frames and are thus suitable for real-time applications. One straightforward implementation is simple online and real-time tracking (SORT) [4], which predicts the new locations of bounding boxes using Kalman filter, followed by a data association procedure using intersection-over-union (IOU) to calculate the cost matrix. Although SORT achieves favourable speed and accuracy simultaneously, it suffers from heavy identity switches due to short-term motion information. Deep SORT [23], on the other hand, introduces object re-identification (REID) as appearance information to handle long-term occlusions, leading to a more robust and effective algorithm. Due to the rapid development of deep neural networks (DNNs), REID features with powerful discriminative capability have been popularized in MOT algorithms [6, 22, 26, 25]. In addition, the frame-by-frame association problem is often seen as bipartite graph matching solved by Hungarian algorithm [8].

By contrast, offline methods [27, 9, 2] have access to the whole sequence and can perform global optimization on data association. These batch methods generally formulate MOT as a network flow problem [27, 15]. K-shortest paths (KSP) [2], successive shortest-path (SSP) [9], and dynamic programming (DP) [15] can be used to find the optimal solution. Offline methods enable correction of early errors in online methods and often show better performance, but are not applicable to time-critical applications.

In this paper, we focus on a simple, efficient, and easy-to-implement tracking framework. We use REID features to calculate the cost between current object detections and existing tracklets, minimize the total cost by Hungarian algorithm, and employ operations like interpolation and merging to correct previous results. Among all the parameters that need to be set, the REID threshold and merging threshold are the two most dominant parameters, which allows establishing associations and merging tracklets respectively.

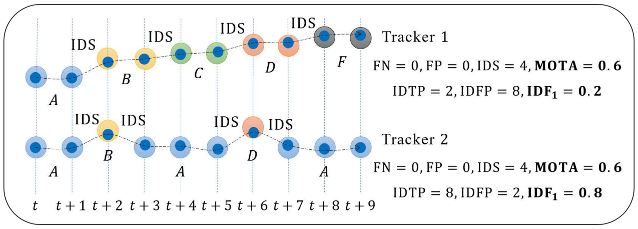

IDS
IDS
IDS
IDS
Tracker 1
A
B
C
D
F
FN = 0, FP = 0, IDS = 4, **MOTA = 0.6**
IDTP = 2, IDFP = 8, **IDF$_1$ = 0.2**
IDS
IDS
IDS
IDS
Tracker 2
A
B
A
D
A
FN = 0, FP = 0, IDS = 4, **MOTA = 0.6**
IDTP = 8, IDFP = 2, **IDF$_1$ = 0.8**
t
$t+1$
$t+2$
$t+3$
$t+4$
$t+5$
$t+6$
$t+7$
$t+8$
$t+9$

Figure 2. Comparison of MOTA and IDF_1. Both Tracker 1 and Tacker2 track target A which lasts 10 frames. Tracker 1, which correctly tracks 2 frames and the remaining frames are assigned to other wrong targets, has 4 identity switches, while Tracker 2 tracks 8 frames with 2 errors has the same number of identity switches. IDF_1 measures how long the identification is correct and give different scores, which can better reflect trackers' performance.

2.2. Evaluation metrics

Quantitative evaluation of tracking performance is challenging due to the complexity of multi-target tracking task. A large number of metrics have been proposed [10, 17, 20, 13], including two main common metrics serving different purposes. One of them is CLEAR MOT metrics [3, 11], which contains multiple object tracking accuracy (MOTA) and multiple object tracking precision (MOTP):

$$\mathrm{MOTA} = 1 - \frac{\sum_t (\mathrm{FN}_t + \mathrm{FP}_t + \mathrm{IDS}_t)}{\sum_t \mathrm{GT}_t}, \tag{1}$$

$$\mathrm{MOTP} = \frac{\sum_{i,t} d_t^i}{\sum_t c_t}, \tag{2}$$

where c_t denotes the number of matched targets in frame t, and d_t^i denotes the matching distance of target i. Comparing to MOTP, which is mainly influenced by localization accuracy of detections, MOTA sums various sources of errors, including false negatives, false positives, and identity switches, providing a better overall performance measure.

The other is ID metrics [16], which contains identification precision (IDP), identification recall (IDR) and corresponding F_1 score IDF_1:

$$\mathrm{IDF}_1 = \frac{2\mathrm{IDTP}}{2\mathrm{IDTP} + \mathrm{IDFP} + \mathrm{IDFN}}, \tag{3}$$

where IDTP, IDFP, and IDFN are calculated by the truth-to-result match, i.e., bipartite graph matching between true trajectories and hypothetical trajectories. Afterwards, each hypothesis is assigned to a unique target. All the frames of hypotheses with small overlap are seen as false positives and that of ground truth are seen as false negatives.

Comparing to MOTA, IDF_1 better measures the consistency of ID matching. A simple example to illustrate its effectiveness is presented in Figure 2. In this paper, we focus on the performance of identification, and thus use IDF_1 as the reference of our self-evaluation metric.

3. Self Quality Evaluation

We design a novel self quality evaluation metric SQE to measure the tracking performance without ground truth annotations that can enable parameters optimization to gain better tracking performance in reality. This metric should be positively correlated with IDF_1 which generally measures the tracking performance the best. The guiding design criteria is provided below in which we highlight some distinctive features an ideal tracker should possess. Both the theoretical and practical facets show that high-quality trajectories present single peaks in the feature distance distribution, while low-quality trajectories present multiple peaks.

3.1. Design criteria

To have a better understanding of the proposed metric, we first explain that an ideal MOT tracker should meet the following criteria. It should be able to: (1) track all targets continuously from appearing to leaving the tracking area; (2) track each target consistently, that is, each target should be assigned one and only one track ID over time; (3) locate the position of each target as accurately as possible.

As mentioned in Section 2.2, (3) quantifies the detection performance in the tracking-by-detection paradigm, thus it is not our main focus. For self evaluation metrics design, (1) inspires that the number and length of trajectories are supposed to be appropriate. (2) leads to the assumption that for an outstanding tracker, REID features are as similar as possible if coming from the same trajectory, otherwise are as different as possible. This can be characterized by the intra and inter distance of trajectories. We define the distance between two features$f$ and g as their Euclidean distance:

$$distance = \|f - g\|_2. \tag{4}$$

Based on the above considerations, our self evaluation metric should take the quantity, length, and feature distance information into account comprehensively. Since establishing relationship between the identification quality and the absolute values of distance is hard, distance distribution analysis is considered to be a more reasonable solution.

3.2. Distance distribution analysis

We demonstrate in theory that the intra distance of the same target and the inter distance of different targets obey chi distribution.

For object representation, it is common that low-quality inputs will lead to uncertain estimations, causing the computed REID features to fluctuate around the ideal value. We follow the assumptions in [19], modeling the distribution of features as multivariate Gaussian distribution:

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}, \boldsymbol{\sigma}^2\mathbf{I}), \tag{5}$$

where $\mathbf{z}$ is a N-dimension feature vector, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ represent the ideal value and uncertainty along each dimension respectively. Each dimension obeys an independent Gaussian distribution.

We measure the Euclidean distance between a pair of features $(\mathbf{z}_i, \mathbf{z}_j)$:

$$d(\mathbf{z}_i, \mathbf{z}_j) = \sqrt{\sum_{k=1}^{N}(z_{ik} - z_{jk})^2} = \sqrt{\sum_{k=1}^{N} d_k^2}. \tag{6}$$

According to the nature of independent Gaussian random variables, we have $p(d_k) = \mathcal{N}(d_k; \mu_{ik} - \mu_{jk}, \sigma_{ik}^2 + \sigma_{jk}^2)$. If $(\mathbf{z}_i, \mathbf{z}_j)$ comes from the same target, then $\mu_{ik} - \mu_{jk} = 0$, $\sigma_{ik}^2 + \sigma_{jk}^2 = 2\sigma_k^2$. Thus, the feature distance after standardization obeys chi distribution with a degree of freedom equals to N:

$$\sqrt{\sum_{k=1}^{N}(\frac{d_k}{\sqrt{2\sigma_k^2}})^2} \sim \chi_N, \tag{7}$$

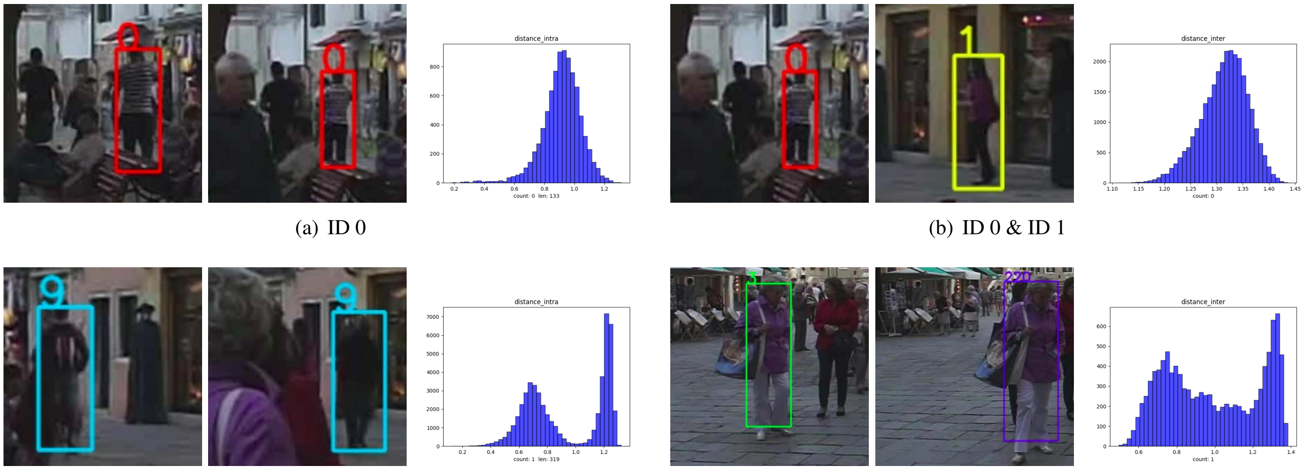
and if $(\mathbf{z}_i, \mathbf{z}_j)$ comes from different targets:

$$\sqrt{\sum_{k=1}^{N}(\frac{d_k - (\mu_{ik} - \mu_{jk})}{\sqrt{\sigma_{ik}^2 + \sigma_{jk}^2}})^2} \sim \chi_N, \tag{8}$$

Therefore, the intra and inter distance distributions of ideal trajectory hypotheses present single peaks. Next we consider a low-quality trajectory containing an identity switch between target A and B. For the ease of analysis, we assume that each target and feature dimension has the same variance. Therefore, the distance of features $(\mathbf{z}_{Ai}, \mathbf{z}_{Bj})$ obeys non-central chi distribution with a positive noncentrality parameter $\lambda = \sum_{k=1}^{N}(\mu_{A_ik} - \mu_{B_jk})^2$. Meanwhile, the distance within each target obeys central chi distribution proved as above. The final distance distribution is indeed the sum of central and non-central chi distributions, thus showing a bimodal form. It can be inferred that the low-quality trajectories with wrong identification would present multiple peaks in the intra and inter distance distributions.

3.3. Practical verification

We practically verify the above conclusions by visualizing the intra and inter distance distributions of several different tracking cases in Figure 3. The results exhibit that the high-quality trajectories, such as the one labeled with



(a) ID 0

(b) ID 0 & ID 1

(c) ID 9

(d) ID 3 & ID 220

Figure 3. Distance distributions of several different tracking cases and corresponding visualization results.

ID 0 consistently tracks a person moving forward while being separated from the one with ID 1, present single peaks. In contrast, the low-quality trajectories, such as the one containing an identity switch with ID 9 and the overlapped ones with ID 3 and ID 220, present multiple peaks.

To quantify the validity of our Gaussian assumption in Section 3.2, we use the descriptor provided in [21] to perform a normality test on the ground truth of MOT16 train set and find that 74% of the trajectories can be approximate as Gaussian distribution at a significance level of 0.1. Under low-density scenarios like MOT16-05, the percentage raises to 88 %. Considering that counterexamples may occur in practice, such as two similar-dressed people, we have also tested the performance of the descriptor on classifying unique person IDs in MOT16's detection boxes. When the precision is set to 0.95, the recall and mAP can reach 0.94 and 0.98 respectively. Therefore, we consider the counterexamples only make up a small portion.

However, due to non-ideal factors, the final distances do not fully obey the theoretical chi distribution. We take ID 0 for example. Although a similar overall shape is shown, the hypothesis test has an extremely low p-value of 0, indicating a statistically significant difference. This may have two reasons: (1) Bias is introduced when using sample statistics to replace the true mean and variance for standardization; (2) Features extracted by the REID model are not independent in each dimension. The second reason is very common, since deep neural networks tend to cause strong correlations between multiple dimensions.

It is encouraging that the trajectories of different qualities still retain the distinctive single or multiple peaks. The more frames with wrong identification, the more obvious the two peaks, and the larger interval between them. In practice we found that fitting a two-class Gaussian distribution and setting a threshold for the mean difference can qualitatively detect those low-quality trajectories which significantly affect tracking performance. According to the visualization results, we also found that the false alarm trajectory is usually short in length, large in variance, and may interfere with the inter distances to produce multiple peaks. These trajectories for which no real target exists are also categorized as low-quality trajectories.

3.4. Metric

Based on the above criteria and distance distribution analyses, we propose a novel self quality evaluation metric SQE, which can be expressed as:

$$\text{SQE} = \frac{n \times L}{n + k_1 \times L + k_2 \times (fp + dif + sim)} \tag{9}$$

The specific explanation is detailed below. The evaluation process is summarized in Algorithm 1 and mainly divided into four steps:

(1) For a trajectory with short length and large standard deviation, we mark it as false alarm and accumulate fp.

(2) For the rest trajectories we utilize a two-class Gaussian mixture model to fit the intra distances, and judge whether it is a low-quality trajectory according to the mean difference. If it exceeds a certain threshold, we assert that this trajectory contains more than one target and accumulate a difference error, denoted by dif.

(3) Similarly, the inter distances of each two non-false alarm trajectories are also fitted. They are considered to match the same target with a large mean difference, and the similarity error is denoted by sim.

(4) Other internal characteristics like the number n and mean length L of trajectories are also embedded.

When the REID threshold is set too strict, there are so many detection boxes being excluded that n and L are both

Algorithm 1: self quality evaluation

```
Input: set of trajectory hypotheses $\mathcal{T} = \{T_k\}$
Output: self evaluation result
$n$ = number of trajectories in $\mathcal{T}$;
$L$ = mean length of trajectories in $\mathcal{T}$;
initialization;
foreach $T_k \in \mathcal{T}$ do
    calculate the intra distance by Equation 4;
    if $L_{Tk} < \delta_L$ and $Std > \delta_D$ then
        mark the track as false alarm;
        $fp = fp + 1$;
    else
        fit a two-class Gaussian mixture model;
        if $\Delta_{mean} > \delta_m$ then
            $dif = dif + 1$;
        end
    end
end
foreach $T_l, T_m \in \mathcal{T}$ do
    calculate the inter distance by Equation 4;
    if not false alarm tracks then
        fit a two-class Gaussian mixture model;
        if $\Delta_{mean} > \delta_m$ then
            $sim = sim + 1$;
        end
    end
end
calculate SQE by Equation 9;
return SQE;
```

small; when $n \times L$ remains almost constant, the two variables have opposite trend, and extreme situations including excessively fragmented or concatenated trajectories will lead to imbalance between them. To downgrade these poor tracking results, we employ the form of harmonic mean, and set k_1 to accommodate moving speed and density of tracking objects. For pedestrian tracking task on street videos, the magnitude of n and L is approximately equivalent, and thus k_1 could be set to 1 concisely.

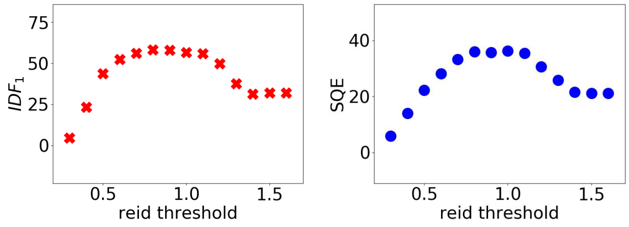
Based on this rough constraint form, a correction item is added to the denominator. We have demonstrated that the accumulated dif, sim and fp can reflect the number of low-quality trajectories. Therefore, their sum is expected to be small, and meanwhile the value of SQE is large. The correction item actually plays a key role within the range of moderate values of n and L. k_2 is used to adjust the ratio between n, L and sum of errors.

Parameters in SQE are not difficult to set. δ_L is comparable to the video's frame rate. With a high-precision ReID model, randomly selecting false alarms and ID switch examples from reference videos is adequate to observe Std and Δ_m, so as to set δ_D and δ_m accordingly. Additionally, when the tracker and task (vehicle/pedestrian) are given, k_1 and k_2 could be set empirically.

4. Experiments

Implementation details. We assess our self evaluation method mainly on the MOT16 Challenge data sets [11], which contains 14 video sequences (7 for training, 7 for testing) taken by both static and moving cameras from different angles in different scenes. We focus our study on pedestrian tracking and make use of the person ReID model provided by [21]. All the experiments are completed with the same parameter setting: $\delta_L = 15$, $\delta_D = 0.2$, $\delta_m = 0.3$, $k_1 = 1$ and k_2 takes 2 and 10 for the REID threshold and the merging threshold, respectively. The REID threshold varies from 0.3 to 1.6, beyond which IDF_1 remains invariant. Similarly, the merging threshold varies from 0.5 to 1.5. The parameter optimization process is based on grid search. The rest of this section prove the accuracy, universality, and effectiveness of our self quality evaluation metric SQE.

Comparison with supervised metrics. To demonstrate the effectiveness of our self evaluation metric in evaluating tracking performance, we compare its score with existing commonly used supervised metrics on MOT16-02 training video, and visualize IDF_1 and SQE in Figure 4. We found that as the the REID threshold ascends, both SQE and IDF_1 increase at first and decrease afterwards, and reach the highest value at 0.8 with relatively high IDP, IDR, and MOTA. These two items present a very similar trend, which indicates that our designed metric can primarily fulfill the desired positive correlations with IDF_1 which generally measures the performance of identification the best.



(a) (b)

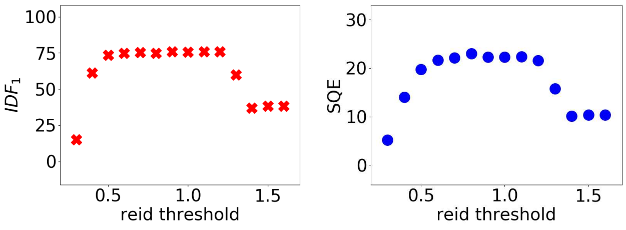
Figure 4. Visualization of IDF_1 and SQE on MOT16-02 (complex scene) when changing the REID threshold.

MOT16-02 video records a complex scene with a large number of people walking around a large square. We further analyse the result on MOT16-09 video, a simpler street scene with low density and the least number of tracks from a low angle, in Figure 5. The favourable similarity illustrates that our self evaluation method can be generalized to different viewpoints and scenarios. The detailed results on other videos are provided in the supplementary material.

video	method	parameter	IDF_1^*	IDP*	IDR*	MOTA*	IDS*	Δ (IDF_1)
02	baseline(gt)	0.80	58.3	79.3	46.0	51.9	69	0.0
	SQE(ours)	0.80	58.3	79.3	46.0	51.9	69	
04	baseline(gt)	1.05	82.0	93.5	73.0	77.3	21	1.1
	SQE(ours)	0.80	80.9	93.0	71.5	76.2	32	
05	baseline(gt)	0.90	71.2	79.2	64.6	62.0	23	0.1
	SQE(ours)	1.00	71.1	78.3	65.2	61.5	32	
09	baseline(gt)	1.20	76.0	88.8	66.4	73.5	8	1.3
	SQE(ours)	0.80	74.7	88.6	64.6	72.2	7	
10	baseline(gt)	0.95	72.4	76.6	68.7	71.5	79	1.9
	SQE(ours)	0.90	70.5	74.5	66.8	71.2	81	
11	baseline(gt)	0.85	80.1	89.7	72.4	75.0	29	4.6
	SQE(ours)	1.00	75.5	83.8	68.7	73.5	34	
13	baseline(gt)	0.75	58.2	74.6	47.7	47.0	73	2.3
	SQE(ours)	1.05	55.9	68.9	47.0	45.6	90	

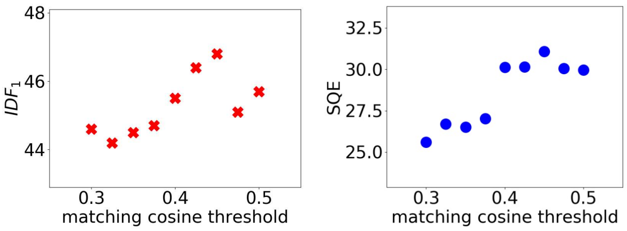
* denotes that the score for SQE parameters is only calculated after the parameters are determined by SQE, but not used to tune the parameters.

Table 1. Comparison of the optimal REID threshold determined by ground truth and our self evaluation method on MOT16-02∼13.



(a) (b)

Figure 5. Visualization of IDF_1 and SQE on MOT16-09 (simple scene) when changing the REID threshold.



(a) (b)

Figure 6. Visualization of IDF_1 and SQE on MOT16-02 when changing the matching cosine threshold in Deep SORT algorithm.

We summarize the optimal REID threshold determined by IDF_1 and SQE in Table 1, with corresponding evaluation scores under these parameters. Our self evaluation method can approximately quantify tracking performance, specifically, 85% of the optimal parameter differences do not exceed 0.25, and 85% of the corresponding IDF_1 differences do not exceed 3.

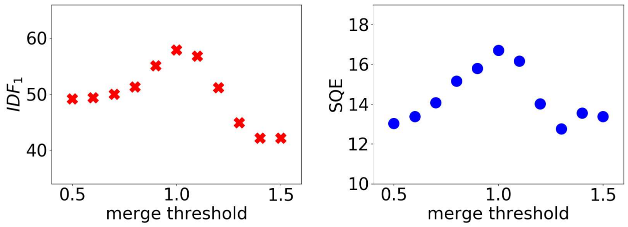
Generalization to other tracking algorithms. To illustrate the robustness and universality of our method, other tracking algorithms are supposed to be tested as a supplementary experiment. We choose Deep SORT, which is one of the highly recognized and open source MOT algorithms in recent years. The REID threshold corresponds to the matching cosine threshold in Deep SORT. This algorithm replaces our interpolation logic with IOU matching, causing the features during occlusion time period to exhibit a small interference peak in the intra distance distribution; therefore, we remove the feature information of these frames when performing self evaluation. As shown in Figure 6, a strong correlation between IDF_1 and SQE is presented, demonstrating the success of our method on other trackers.

Generalization to other parameters. We further test the universality of our method on other parameters. Except for the REID threshold, the merging threshold is another dominant factor affecting final tracking performance. Similarly, we visualize the comparison of IDF_1 and SQE of both complex and simple scenes in Figure 7 and 8. The results still maintain positive correlations. Table 2 shows a high accuracy, with 5 out of 7 videos have an optimal parameter difference below 0.1, and almost all the corresponding IDF_1 differences do not exceed 3.

Practical testing. Our ultimate goal is to find the optimal parameters in realistic scenes where ground truth is unavailable. Additionally in reality the training data is relatively small in scale comparing to the unknown test environment. To test our method in a pragmatic manner, we regard the first 4 training videos as our test set and the last 3 training videos as our training set. Conventionally, the parameters are tuned on the training set and remain constant during testing. In our simulation, we name these pa-

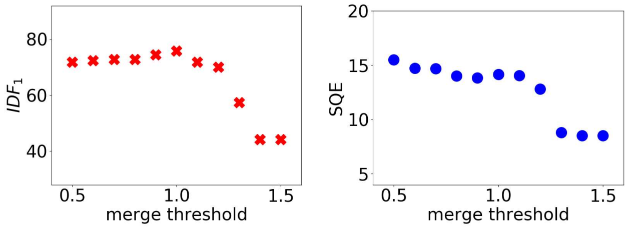
video	method	parameter	IDF$_1^*$	IDP*	IDR*	MOTA*	IDS*	Δ (IDF$_1$)
02	baseline(gt)	1.00	57.9	72.7	46.1	51.7	82	0.0
	SQE(ours)	1.00	57.9	72.7	46.1	51.7	82	
04	baseline(gt)	1.05	82.3	94.2	73.1	76.8	23	0.6
	SQE(ours)	1.00	81.7	93.7	72.4	76.5	27	
05	baseline(gt)	0.85	73.6	82.5	66.5	61.9	34	0.5
	SQE(ours)	0.75	73.1	82.8	65.5	61.4	46	
09	baseline(gt)	1.00	75.9	89.2	66.0	73.2	7	3.1
	SQE(ours)	0.70	72.8	85.8	63.2	72.8	12	
10	baseline(gt)	1.05	71.5	75.1	68.2	70.8	77	1.6
	SQE(ours)	0.95	69.9	74.5	65.9	71.5	82	
11	baseline(gt)	1.10	78.2	87.6	70.5	75.0	30	2.8
	SQE(ours)	0.85	75.4	84.7	68.0	73.9	35	
13	baseline(gt)	1.05	56.7	70.7	47.3	46.2	68	0.6
	SQE(ours)	1.10	56.1	69.2	47.1	45.4	72	

Table 2. Comparison of the optimal merging threshold determined by ground truth and our self evaluation method on MOT16-02∼13.



(a) (b)

Figure 7. Visualization of IDF$_1$ and SQE on MOT16-02 (complex scene) when changing the merging threshold.



(a) (b)

Figure 8. Visualization of IDF$_1$ and SQE on MOT16-09 (simple scene) when changing the merging threshold.

rameters as the baseline parameters. Conversely, our SQE metric can guide the self-optimization of parameters without ground truth. Thus, it is employed directly to tune the 4 testing videos individually.

In reality we can first acquire baseline parameters as reference on small-scale training data, then conduct self evaluation to further optimize the parameters in a relatively small range. The procedure of computing the customized parameters is as follows: (1) Find baseline parameters; (2) For each testing video, fix one parameter with reference to baseline, and then tune the other according to SQE alternately; (3) Combine them to be the customized parameters.

Our method is considered to be effectual if the tracker using the customized parameters outperform the tracker using constant training-set-tuned parameters. The result is shown in Table 3, where gt denotes the true optimal parameters on each video. To be rigorous, we use the best parameters found by grid search on the 3 assumed training videos as the baseline. It is apparent that the parameters tuned by SQE achieve considerable improvement comparing to the baseline, and the results are much closer to the the true optimum, showing the effectiveness of our method when implemented in a practical manner.

To further illustrate the performance of self-optimization using SQE, we experiment on MOT16 test set and KITTI train set [7]. The baseline parameters are the best parameters found by grid search on MOT16 training set, which outperform empirical parameters by 5.8% IDF$_1$ already. This setup is based on the updated submitting policy of KITTI [1], and we believe it can simulate pedestrian tracking in reality where test scenes varies greatly compared to annotated videos. As shown in Table 4, the parameter self-optimization enabled by SQE elevates the performance of the tracker on these data sets.

Drawbacks and prospects. The above experiments reflect the effectiveness of our proposed SQE metric, while there are still some drawbacks worth noting. Firstly, due to the randomness during model fitting, dif and sim possess several units of uncertainty, resulting in insufficient sensitivity to small changes in IDF$_1$. Secondly, current metrics lack physical consistency explanation. IDF$_1$ is calculated by

[1] http://www.cvlibs.net/datasets/kitti/eval_tracking.php

video	method	parameter	IDF_1^*	IDP^*	IDR^*	MOTA^*	IDS^*
02	gt	0.85, 0.95	59.2	79.9	47.0	52.8	80
	baseline	0.75, 1.00	56.2	**78.3**	43.8	50.1	**61**
	SQE(ours)	0.90, 1.00	**57.9**	75.5	**47.0**	**52.8**	82
04	gt	0.85, 1.05	82.6	94.6	73.3	76.7	23
	baseline	0.75, 1.05	80.9	92.9	71.6	76.3	25
	SQE(ours)	0.75, 1.10	**82.3**	**94.4**	**72.9**	**76.6**	**21**
05	gt	0.90, 0.85	73.6	82.5	66.5	61.9	34
	baseline	0.75, 1.05	68.4	80.1	59.7	57.6	**24**
	SQE(ours)	1.00, 0.95	**72.2**	**81.1**	**65.1**	**62.6**	25
09	gt	1.20, 0.70	76.0	88.8	66.4	73.5	8
	baseline	0.75, 1.05	71.4	84.7	61.7	**72.2**	**6**
	SQE(ours)	0.85, 0.90	**73.0**	**87.1**	**62.8**	71.2	10
overall	gt	-	76.4	90.3	66.2	69.7	145
	baseline	-	74.0	88.5	63.5	68.4	**116**
	SQE(ours)	-	**75.5**	**88.8**	**65.6**	**69.5**	138

Table 3. Comparison of different REID and merging threshold selections based on ground truth, optimum of the remaining 3 videos and our SQE metric. Adjust two parameters at the same time.

dataset	method	IDF_1^*	IDP^*	IDR^*	IDS^*
MOT16 test	baseline	66.6	75.8	59.4	442
	ours	**68.3**	83.4	57.8	456
KITTI train**	baseline	67.4	67.2	67.7	37
	ours	**68.5**	67.9	69.1	44

** we use the 5 videos with the most pedestrians in KITTI train set.

Table 4. Practical testing on MOT16 test set and KITTI train set.

IDTP, IDFP and IDFN, while our method simply records the number of low-quality trajectories. A more precise idea is to estimate IDTP and IDFP relying on the quantity information. Assume that for a trajectory where an identity switch occurs, target A appears n_1 frames while target B appears n_2 frames. The total length is L and the number of distances in the class with larger values is N. Then A and B satisfy the following conditions:

$$\begin{cases} n_1 + n_2 = L, \\ n_1 \times n_2 = N, \end{cases} \tag{10}$$

which can be easily solved. We can make estimations by:

$$\mathrm{IDTP}, \mathrm{IDFP} = \max(n_1, n_2), \min(n_1, n_2). \tag{11}$$

Such processing for the intra distance distribution can accurately estimate the number of erroneous frames. Furthermore, the inter distance distribution can help refine the estimations. For example, if there is another trajectory that also tracks A, we only keep the longer one as IDTP according to the calculation rule of IDF_1. However, more detailed considerations are needed for global precise estimations. In addition, categorizing low-quality trajectories and estimating erroneous frames may also be conducive to tracker's post-processing so as to improve tracking performance. Finally, the adjustable parameters k_1 and k_2 need to be defined more strictly. We plan to investigate these downsides in the future.

5. Conclusion

In this paper, we propose a self quality evaluation metric SQE to enable the parameters optimization in the test environment and realistic scenes where ground truth is unavailable. This new perspective can bypass the difficulty of designing an algorithm that perform well in various scenes. We demonstrate that trajectories with different qualities exhibit different single or multiple peaks in feature distance distribution, inspiring us to use a two-class Gaussian mixture model to estimate identification errors. Experiments mainly on the MOT16 Challenge data sets demonstrate the effectiveness of our method in both correlating with existing metrics and enabling parameters self-optimization to achieve better tracking performance. In the end, the drawbacks and prospects for future work are summed up. We believe that our work is instructive for further MOT research.

6. Acknowledgement

This research was supported by National Key R&D Program of China (No. 2017YFA0700800).

References

[1] Anton Andriyenko, Konrad Schindler, and Stefan Roth. Discrete-continuous optimization for multi-target tracking. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1926–1933. IEEE, 2012.

[2] Jerome Berclaz, Francois Fleuret, Engin Turetken, and Pascal Fua. Multiple object tracking using k-shortest paths optimization. *IEEE transactions on pattern analysis and machine intelligence*, 33(9):1806–1819, 2011.

[3] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *Journal on Image and Video Processing*, 2008:1, 2008.

[4] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3464–3468. IEEE, 2016.

[5] Michael D Breitenstein, Fabian Reichlin, Bastian Leibe, Esther Koller-Meier, and Luc Van Gool. Online multiperson tracking-by-detection from a single, uncalibrated camera. *IEEE transactions on pattern analysis and machine intelligence*, 33(9):1820–1833, 2010.

[6] Weitao Feng, Zhihao Hu, Wei Wu, Junjie Yan, and Wanli Ouyang. Multi-object tracking with multiple cues and switcher-aware classification. *arXiv preprint arXiv:1901.06129*, 2019.

[7] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[8] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[9] Philip Lenz, Andreas Geiger, and Raquel Urtasun. Followme: Efficient online min-cost flow tracking with bounded memory and computation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4364–4372, 2015.

[10] Yuan Li, Chang Huang, and Ram Nevatia. Learning to associate: Hybridboosted multi-target tracker for crowded scene. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2953–2960. IEEE, 2009.

[11] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016.

[12] Anton Milan, S Hamid Rezatofighi, Anthony Dick, Ian Reid, and Konrad Schindler. Online multi-target tracking using recurrent neural networks. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[13] Anh-Tuan Nghiem, Francois Bremond, Monique Thonnat, and Valery Valentin. Etiseo, performance evaluation for video surveillance systems. In *2007 IEEE Conference on Advanced Video and Signal Based Surveillance*, pages 476–481. IEEE, 2007.

[14] Songhwai Oh, Stuart Russell, and Shankar Sastry. Markov chain monte carlo data association for general multiple-target tracking problems. In *2004 43rd IEEE Conference on Decision and Control (CDC)(IEEE Cat. No. 04CH37601)*, volume 1, pages 735–742. IEEE, 2004.

[15] Hamed Pirsiavash, Deva Ramanan, and Charless C Fowlkes. Globally-optimal greedy algorithms for tracking a variable number of objects. In *CVPR 2011*, pages 1201–1208. IEEE, 2011.

[16] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision*, pages 17–35. Springer, 2016.

[17] Dominic Schuhmacher, Ba-Tuong Vo, and Ba-Ngu Vo. A consistent metric for performance evaluation of multi-object filters. *IEEE transactions on signal processing*, 56(8):3447–3457, 2008.

[18] Samuel Schulter, Paul Vernaza, Wongun Choi, and Manmohan Chandraker. Deep network flow for multi-object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6951–6960, 2017.

[19] Yichun Shi and Anil K. Jain. Probabilistic face embeddings. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[20] Kevin Smith, Daniel Gatica-Perez, Jean-Marc Odobez, and Sileye Ba. Evaluating multi-object tracking. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops*, pages 36–36. IEEE, 2005.

[21] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 480–496, 2018.

[22] Siyu Tang, Mykhaylo Andriluka, Bjoern Andres, and Bernt Schiele. Multiple people tracking by lifted multicut and person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3539–3548, 2017.

[23] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3645–3649. IEEE, 2017.

[24] Yu Xiang, Alexandre Alahi, and Silvio 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.

[25] Young-Chul Yoon, Du Yong Kim, Kwangjin Yoon, Youngmin Song, and Moongu Jeon. Online multiple pedestrian tracking using deep temporal appearance matching association. *arXiv preprint arXiv:1907.00831*, 2019.

[26] Jimuyang Zhang, Sanping Zhou, Jinjun Wang, and Dong Huang. Frame-wise motion and appearance for real-time multiple object tracking. *arXiv preprint arXiv:1905.02292*, 2019.

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