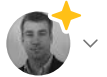


Open in app ↗



Renu Khandelwal

Follow

May 11, 2022 · 9 min read · ✨ · 🎧 Listen



Save



# Evaluation Metrics for Multiple Object Tracking

## Learn different Metrics for evaluating the performance of a Multiple Object Tracking Algorithm

Multiple Object Tracking(MOT) is the task of detecting various objects of interest in a video, tracking these detected objects in subsequent frames by assigning them a unique ID, and maintaining these unique IDs as the objects move around in a video in successive frames.

Multiple Object Tracking(MOT) finds its application in **video surveillance, robotics, or self-driving vehicles.**

*To understand the different metrics used to evaluate MOT algorithms, you first need to understand how MOT works.*

MOT takes a single continuous video as an input and splits it into discrete frames at a specific frame rate(fps). The output of MOT is

- **Detection:** what objects are present in each frame
- **Localization:** where objects are in each frame
- **Association:** whether objects in different frames belong to the same or different objects

*You are performing sports analysis using Multi-object tracking. Would you like to detect every object in the frame precisely or detect the players and their trajectories?*

*What is more critical for a self-driving car? Multi-object tracking algorithm is to detect every pedestrian to avoid collision or to associate the objects detected*



*over time?*

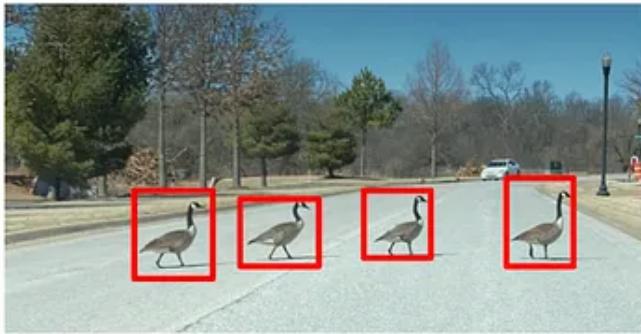
*For video surveillance, is it essential to ensure all objects are detected, and their trajectories tracked accurately?*

---

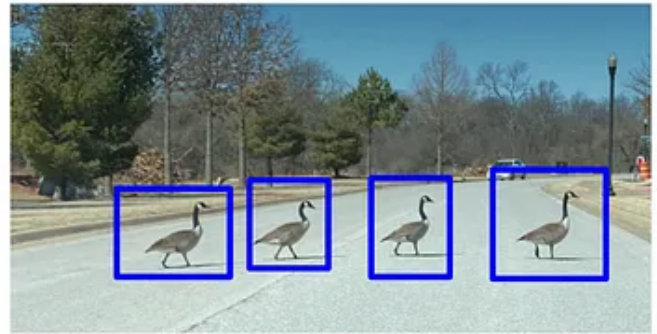
*Read on to find out which MOT evaluation metrics best applies to the above scenarios.*

---

To evaluate the performance of MOT algorithms, measure how well a tracker performs by comparing its predictions to the ground-truth set of tracking results.



**Ground Truth**



**Model Prediction**

Image by author

### **Characteristics of MOT Evaluation Metrics**

MOT evaluation metrics need to exhibit two significant properties

1. **MOT evaluation metrics need to address five error types in MOT.** These five error types are False negatives(FN), False positives(FP), Fragmentation, Mergers, and Deviation.

---

*False Negative or Misses* when ground truth exists but prediction was missed

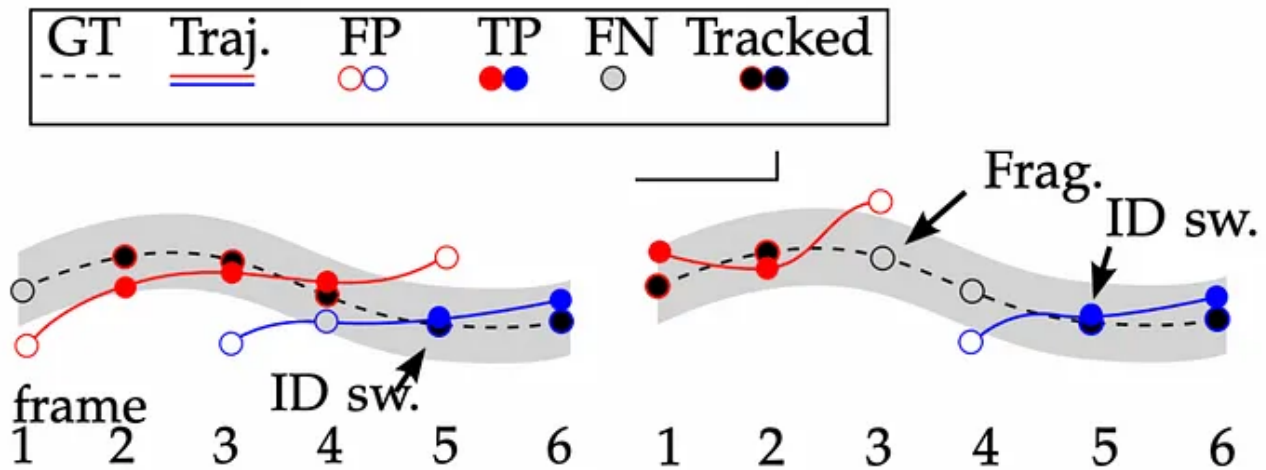
*False Positive* where tracker prediction exists for no ground truth tracker

*Merge or ID switch* when two or more objects tracks are swapped as they pass close to each other

*Deviation* when an object track is reinitialized with a different track ID

*Fragmentation* occurs when a track suddenly stops getting tracked but ground truth track still exists.

---



An ID switch occurs when the mapping switches from the previously assigned red track to the blue one

A track fragmentation is counted in frame 3 because the target is tracked in frames 1-2, then interrupts, and then reacquires its 'tracked' status at a later point with a different ID

Source: MOT16: [A Benchmark for Multi-Object Tracking](#)

2. MOT evaluation metrics should have monotonicity, and error types should be differentiable. The metrics should have information about the tracker's performance for each of the five basic error types.

**Commonly used MOT metrics are**

- Track-mAP
- Multi-Object Tracking Accuracy- MOTA
- Multi-Object Tracking Precision-MOTP
- IDF1
- Higher-Order Tracking Accuracy-HOTA

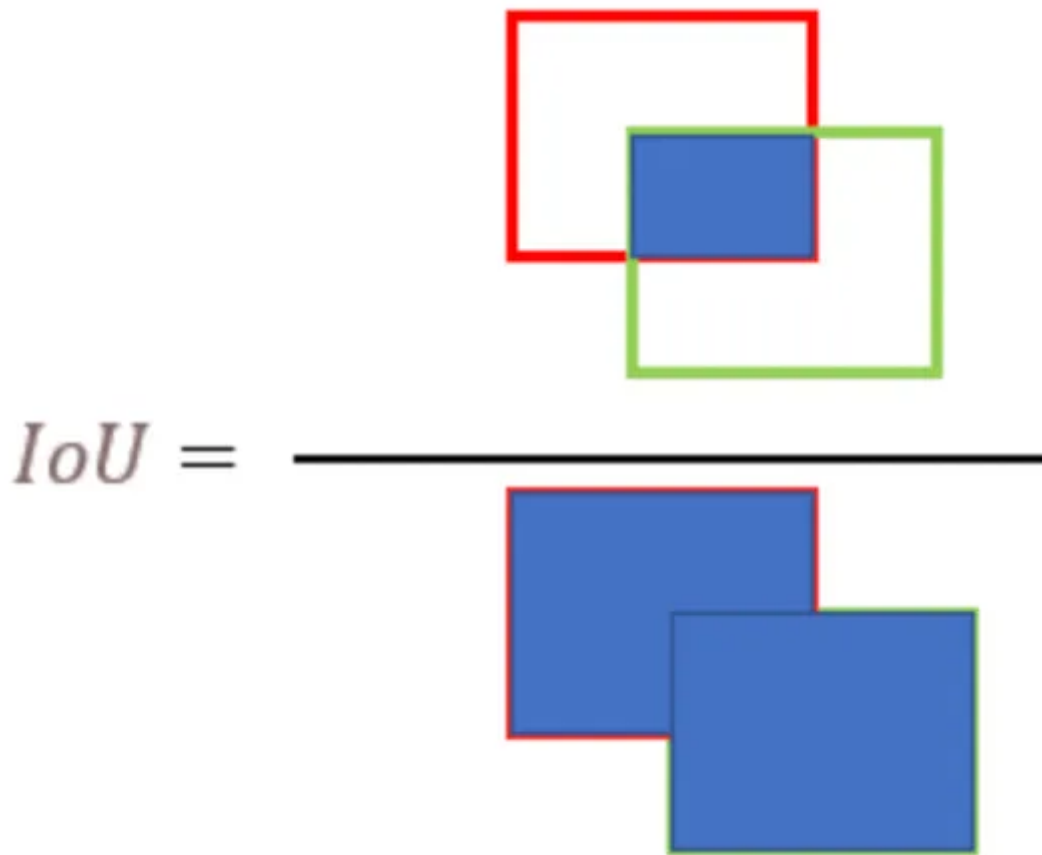
You need to understand the different evaluation metrics as the choice of evaluation metric is extremely important. **Understanding each evaluation metric helps determine how different errors contribute to the final score.** The knowledge of different errors that contribute to the evaluation metrics strongly influences how to improve the MOT scores and the direction for future research.

## Track-mAP

Track-mAP (mean average precision) matches predictions and ground truth at a trajectory level. Track-mAP requires a trajectory similarity score,  $Str$ , between trajectories and a threshold  $\alpha_{tr}$ , such that trajectories are matched only if the trajectory similarity score is greater than the threshold.

### $Str \geq \alpha_{tr}$

$Str$  is the sum of the spatial intersection of the boxes across the whole trajectories, divided by the sum of the spatial union of the boxes across the entire trajectories.



$$IoU = \frac{\text{Intersection of boxes}}{\text{Union of boxes}}$$

A prTraj is matched with a gtTraj if it has the highest confidence score of all prTrajs with  $Str \geq \alpha_{tr}$ .

**TPTr:** A True Positive trajectory is when a prTraj is matched with gtTra

**FPTr:** A False Positive trajectory is the remaining prTrajs that are not matched with grTraj.

PrTrajs are ordered by decreasing confidence score. Let the index of this ordering (starting at one) be  $n$ . For each value of  $n$ , the precision ( $Pr_n$ ) and recall ( $Re_n$ ) can

be calculated

$$\text{Pr}_n = \frac{|\text{TPTTr}|_n}{n}$$

$$\text{Re}_n = \frac{|\text{TPTTr}|_n}{|\text{gtTraj}|}$$

The precision values are interpolated (InterpPrn) to decrease monotonically.

$$\text{InterpPr}_n = \max_{m \geq n} (\text{Pr}_m)$$

The Track mAP score is then the integral under the interpolated precision-recall curve created by plotting InterpPrn against Ren for all values of n.

*Track mAP performs both matching and association at a trajectory level and is biased toward measuring association. It operates based on the confidence-ranked potential tracking results. Track-mAP is non-monotonic in detection.*

### Challenges with Track mAP

- The interpretation of tracking outputs using Track mAP is not trivial nor easily visualizable, and track mAP has many overlapping outputs and some of them with low confidence scores. As a result, each trajectory's final score is hidden behind the implicit confidence ranking making interpretation and visualization of outputs challenging.
- The threshold of 0.5 for a trajectory to count as a positive match is high; as a result, the metric ignores a lot of improvement in detection, association, and localization. Even with the best tracking, more than half of its best guess predictions will be counted as errors in Track-mAP, so any improvement in terms of detection and association is not visible in metric scores.

- **Track-mAP** measures trajectories match that mix association, detection, and localization in a way that is error type is non-differentiable and non-separable.

## **Multi-Object Tracking Accuracy: MOTA**

MOTA remains the most representative measure that coincides to the highest degree with human visual assessment.

In MOTA, matching is done at a detection level. A bijective (one-to-one) mapping is constructed between prDets(predicted detection) and gtDets(ground truth detection) in each frame if they are adequately spatially similar to compute True Positive(TP), False Positive, and False Negative.

In MOTA, the association is measured with Identity Switch (IDSW), which occurs when a tracker wrongfully swaps object identities or when a track is lost and is reinitialized with a different identity.

MOTA measures three types of tracking errors: False Positive(FP), False Negative(FN), and ID Switch(IDSW)

$$\text{MOTA} = 1 - \frac{|\text{FN}| + |\text{FP}| + |\text{IDSW}|}{|\text{gtDet}|}$$

MOTA doesn't include a measure of localization error, and detection performance significantly outweighs association performance.

## **Multi-Object Tracking Precision: MOTP**

MOTP measures the localization accuracy, and MOTP averages the overlap between all correctly matched predictions and their ground truth.

$$\text{MOTP} = \frac{1}{|\text{TP}|} \sum_{\text{TP}} S$$

It averages the similarity score,  $S$ , over the set of True Positive (TP). It matches prDets with gtDets that have a similarity score greater than the threshold ( $S \geq \alpha$ ) and as well as not cause an ID Switch (IDSW) to maximize the MOTP score. **The threshold  $\alpha$  has a strong influence on the behavior of the MOTP.**

**MOTP mostly quantifies the localization accuracy of the detector, and therefore, it provides little information about the actual performance of the tracker**

MOTP and MOTA evaluation metrics address intuitive characteristics of tracking systems, like precision in localizing objects, their accuracy in recognizing objects, configuring the threshold value, and consistently tracking objects over time.

### **The Identification Metrics: IDF1**

**IDF1 emphasizes Association accuracy rather than detection** and is used as a secondary metric on the MOTChallenge benchmark due to its focus on measuring association accuracy over detection accuracy.

IDF1 calculates a bijective (one-to-one) mapping between the sets of gtTrajs and prTrajs to determine which trajectories are present, unlike MOTA which matches at an object detection level to associate them throughout time.

**IDF1 uses IDTPs (Identity True Positives) where prID is matched with grID when  $S \geq \alpha$  of trajectories. IDF1 is the ratio of correctly identified detections over the average number of ground-truth and computed detections. The Hungarian algorithm selects which trajectories to match to minimize the sum of the number of IDFP and IDFN.**

**IDF1 combines IDP (ID Precision) and IDR (ID Recall).**



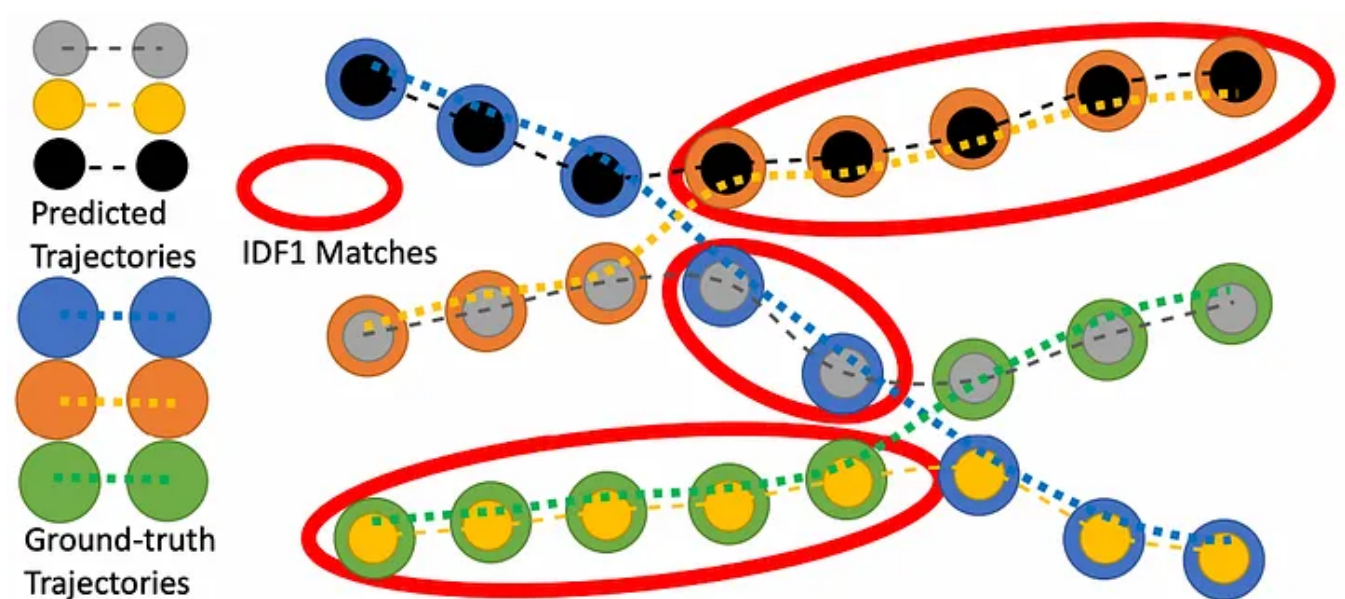
$$\text{ID-Recall} = \frac{|\text{IDTP}|}{|\text{IDTP}| + |\text{IDFN}|}$$

$$\text{ID-Precision} = \frac{|\text{IDTP}|}{|\text{IDTP}| + |\text{IDFP}|}$$

$$\text{IDF1} = \frac{|\text{IDTP}|}{|\text{IDTP}| + 0.5 |\text{IDFN}| + 0.5 |\text{IDFP}|}$$

IDFN(Identity False Negative) is the remaining gtID that is not matched with prID.

IDFP(Identity False Positive) is the remaining prID trajectories that are not matched with any gtID.



A tracking example displaying the single best trajectory matching performed by IDF1(Source: [HOTA: A Higher-Order Metric for Evaluating Multi-Object Tracking](#))

A high IDF1 score estimates the total number of unique objects in a scene than giving information about good detection or association. It also does not evaluate the localization accuracy of trackers.

## High Order Tracking Accuracy: HOTA



*HOTA is a single unified metrics that explicitly evaluates all of these aspects of tracking i.e., accurate detection, association and localization.*

All evaluation metrics, MOTA, IDF1, and HOTA, perform a bijective matching between gtDets and prDets using the Jaccard Index or IOU score, which measures the spatial similarity between gtDet and prDet, any extra or missed predictions are penalized.

$$\text{Jaccard Index} = \frac{|\text{TP}|}{|\text{TP}| + |\text{FN}| + |\text{FP}|}$$

A bijective mapping is calculated between all pairs of gtDet and prDet using the Hungarian algorithm to find the match that optimizes the sum of the matching score.

### Detection

**Detection Accuracy(DetA)** is the percentage of aligning detections. A detection is TP when the Loc-IoU > 0.5 between the ground truth detection(grDet) and predicted detection(prDet).

$$\text{DetA} = \text{Det-IoU} = \frac{|\text{TP}|}{|\text{TP}| + |\text{FN}| + |\text{FP}|}$$

The Hungarian algorithm is applied between the predicted detection and ground truth to determine a one-to-one matching, which helps when one prediction detection overlaps with more than one ground truth and vice-versa.

### Association

**Association Accuracy(AssA)** is the average alignment between matched trajectories, averaged over all detections.

$$\begin{aligned}
 \text{AssA} &= \frac{1}{|\text{TP}|} \sum_{c \in \text{TP}} \text{Ass-IoU}(c) \\
 &= \frac{1}{|\text{TP}|} \sum_{c \in \text{TP}} \frac{|\text{TPA}(c)|}{|\text{TPA}(c)| + |\text{FNA}(c)| + |\text{FPA}(c)|}
 \end{aligned}$$

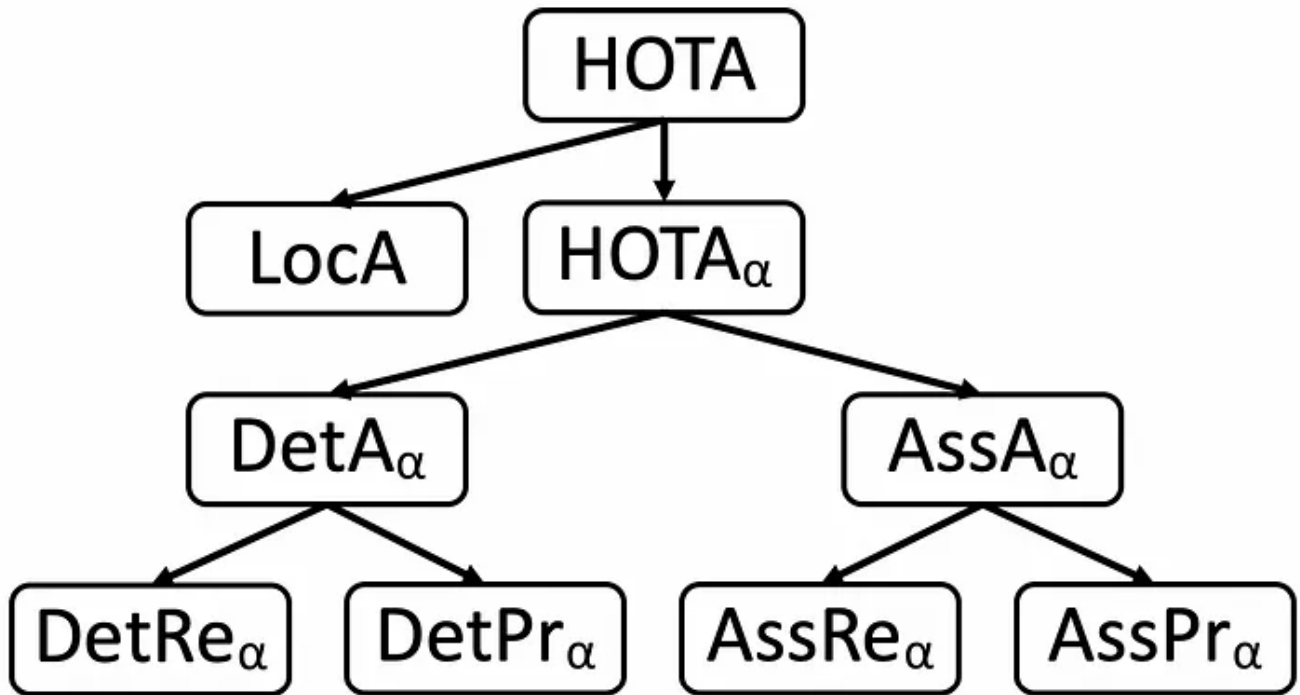
### Localization

Localization Accuracy is the average Loc-IoU over all pairs of matching predicted and ground-truth detections in the whole dataset.

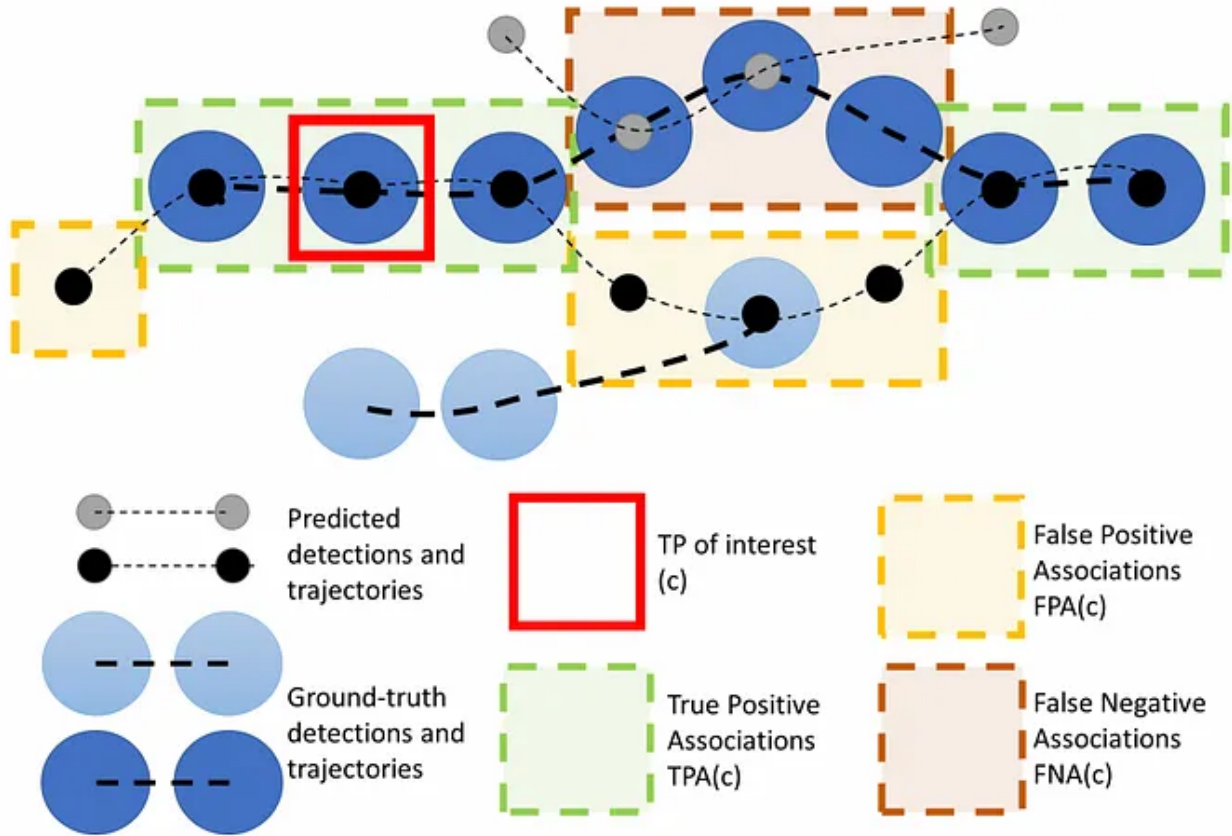
$$\text{LocA} = \frac{1}{|\text{TP}|} \sum_{c \in \text{TP}} \text{Loc-IoU}(c)$$

HOTA decomposes into a family of sub-metrics, which enables the evaluation of different aspects of tracking separately to provide insights into the different types of errors that trackers are making. This knowledge about different error types of tracking enables trackers to be fine-tuned based on use case requirements.

HOTA track errors into three categories:



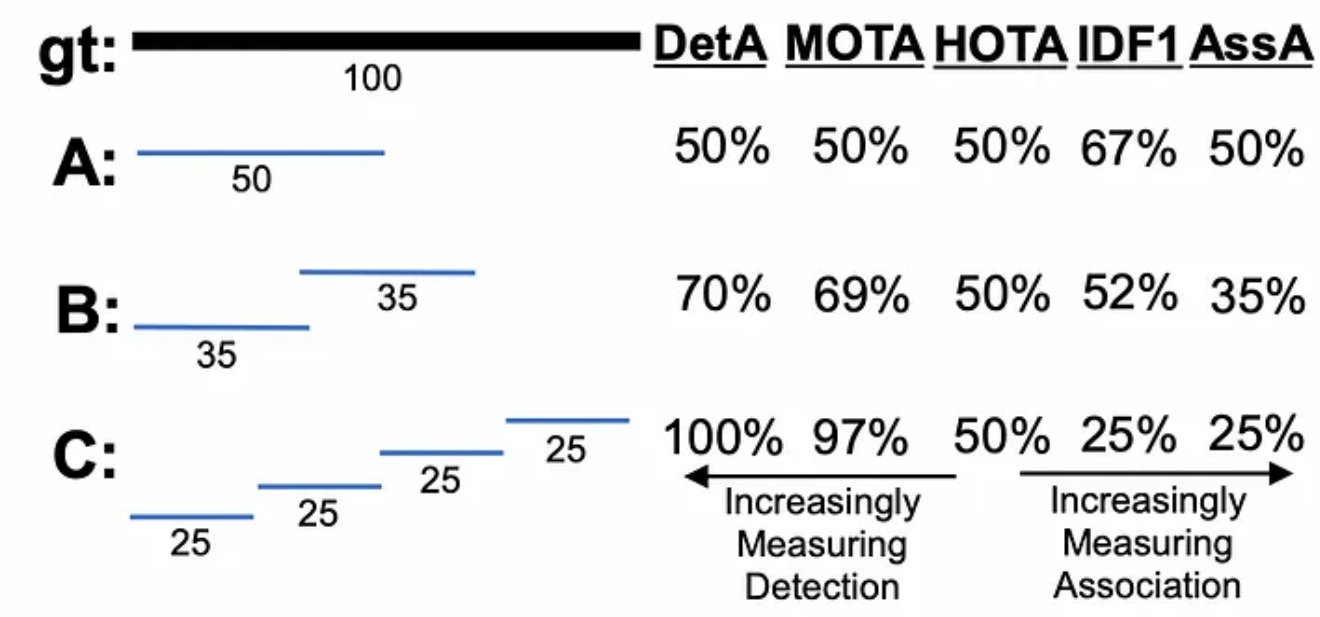
1. **Detection error** occurs when a tracker predicts detections that don't exist in the ground truth or fails to predict detections in the ground truth. Detection errors are further categorized into detection recall (measured by FNs) and detection precision (measured by FPs)
2. **Association error** occurs when trackers assign the same prID to two detections with different gtIDs or assign different prIDs to two detections that should have the same gtID. Association errors are further categorized into errors of association recall (measured by FNAs) and association precision (measured by FPAs)
3. **Localization error** occurs when prDets are not perfectly spatially aligned with gtDets.



HOTA combines are

$$\begin{aligned}
 \text{HOTA}_\alpha &= \sqrt{\text{DetA}_\alpha \cdot \text{AssA}_\alpha} \\
 &= \sqrt{\frac{\sum_{c \in \text{TP}_\alpha} \text{Ass-IoU}_\alpha(c)}{|\text{TP}_\alpha| + |\text{FN}_\alpha| + |\text{FP}_\alpha|}} \\
 \text{HOTA} &= \int_{0 < \alpha \leq 1} \text{HOTA}_\alpha \\
 &\approx \frac{1}{19} \sum_{\substack{\alpha=0.05 \\ \alpha+=0.05}}^{0.95} \text{HOTA}_\alpha
 \end{aligned}$$

HOTA includes the localization accuracy in the tracking results not present in either MOTA or IDF1.



The diagram shows the differences between evaluation metrics.

- MOTA performs both matching and association scoring at a local detection level but accentuates detection accuracy, whereas IDF1 performs at a trajectory level by emphasizing the effect of association.*
- Track-mAP is similar to IDF1 as it performs both matching and association at a trajectory level and is biased toward measuring association.*
- HOTA balances both detection and association by being an explicit combination of a detection score and an association score by performing matches at the detection level while scoring association globally over trajectories.*

	MOTA	IDF1	Track-mAP	HOTA
Representation	Final Tracks	Final Tracks	Potential Tracks with Conf. Score	Final Tracks
Matching Mechanism	Bijjective	Bijjective	Highest Conf.	Bijjective
Matching Domain	Detection	Trajectory	Trajectory	Detection
Association Domain	Prev. One Det	Matched Dets	Matched Dets	All Dets
Scoring Function	$1 - \frac{\sum \text{Err}}{ \text{GTDet} }$	F1 Score	Av. Precision	Doub. Jaccard
Bias Toward	Detection	Association	Association	Balanced

An overview of different evaluation metrics for MOT

Conclusion:

Evaluation metrics like IDF1, MOTA, and MOTP summarize the performance into one single number for comparison, or metrics like HOTA provides you granular information about the errors made by the algorithm, which can be a piece of critical information to improve the performance. Choosing the right metrics will depend on the use case you are working on and will heavily influence the direction of improvement.

References:

[HOTA: A Higher-Order Metric for Evaluating Multi-Object Tracking](#)

**How to evaluate tracking with the HOTA metrics**

HOTA (Higher Order Tracking Accuracy) is a novel metric for evaluating multi-object tracking (MOT) performance. It is...

[autonomousvision.github.io](https://autonomousvision.github.io)

[MOT16: A Benchmark for Multi-Object Tracking](#)

[Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics](#)

[Evaluating Multi-Object Tracking](#)

[An Introduction to Object Tracking](#)



## **ByteTrack: A Simple Yet Effective Multi-Object Tracking Technique**

Technology

Computer Vision

Multiple Object Tracking

Robotics

Artificial Intelligence