

Simulation of Synthetically Degraded Tracking Data to Benchmark MOT Metrics

Michael Hartmann¹, Katharina Löffler^{1,2} and Ralf Mikut¹

¹ Institute for Automation and Applied Informatics,

² Institute of Biological and Chemical Systems - Biological Information Processing,

Karlsruhe Institute of Technology

Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen

Abstract

Multiple object tracking (MOT) is an essential task in computer vision, with many practical applications in surveillance, robotics, autonomous driving, and biology. To compare different MOT algorithms efficiently and select the best MOT algorithm for an application, we rely on tracking metrics that reduce the performance of a tracking algorithm to a single score.

However, there is a lack in testing the tracking metrics themselves, which can result in unnoticed biases or flaws in tracking metrics that can influence the decision of selecting the best tracking algorithm. To check tracking metrics for possible limitations or biases towards penalizing specific tracking errors, a standardized evaluation of tracking metrics is needed.

We propose benchmarking tracking metrics using synthetic, erroneous tracking results that simulate real-world tracking errors. First, we select common real-world tracking errors from the literature and describe how to emulate them. Then, we validate our approach by reproducing previously found tracking metric limitations through simulating specific tracking errors. In addition, our benchmark reveals a before unreported limitation in the tracking metric AOGM. Moreover, we make an implementation of our benchmark publicly available.

1 Introduction

Multiple object tracking (MOT) is an essential task in computer vision, with many practical applications in surveillance, robotics, autonomous driving, and biology. As MOT provides the basis for more complex tasks such as scene understanding, human-machine interaction, or analyzing cell behavior, a high tracking quality is needed. To find the most suitable tracking algorithm for an application, we rely on tracking metrics to compare different tracking approaches efficiently. Tracking metrics summarize the performance of a tracker into a single score by comparing the ground truth (GT), perfect tracking of all objects, to the tracker output and penalize any deviation from the GT, which is called a tracking error.

Tracking metrics are considered as an objective measure, thereby ignoring potential biases and limitations in the tracking metrics themselves. This unawareness towards tracking metric limitations can ultimately even lead research into the wrong direction, as the improvement of tracking approaches is quantified by the tracking measure.

Unfortunately, spotting errors in tracking metrics usually happens by chance, and handcrafted examples are generated to demonstrate them [1]. To date, limitations of several tracking metrics, such as AOGM, MOTA and IDF1 [2, 3, 4], have been reported [1, 5, 6, 7]. However, these tracking metrics are still used in benchmarks [8, 9] as developing new tracking metrics requires time. For instance, in 2013, Leichter and Krupka published several problems of the popular MOTA metric [6]. Eight years later, the HOTA metrics was proposed, which aims to replace MOTA by claiming to be more balanced [5].

A systematic approach to evaluate tracking metrics is needed to ensure that tracking metrics align with the established objectives that tracking algorithms should meet, and to facilitate the development of tracking metrics themselves. For example, one concept which is important for the targeted improvement of tracking metrics is error differentiability [6] – the separate quantification of the tracking performance concerning different types of tracking errors. Until now, most metrics only provide a single composite score, which does not indicate how the metric penalizes different types of errors. By providing an approach

that allows to systematically analyze the behavior of the metric, different types of errors, biases and limitations of metrics towards specific error types can be detected.

Moreover, such an approach can help to boost acceptance of a new tracking measure in the community as the consistency of the proposed metric concerning desired properties can be demonstrated. For instance, a desirable property of tracking metrics is monotonicity – a metric score should improve if, for example, a tracking error is removed from the dataset [6].

We propose benchmarking tracking metrics by replacing the tracking algorithm with synthetic tracking results emulating real-world tracking errors. We select a set of frequently occurring real-world tracking errors and provide instructions on how to simulate them. To validate our approach, we reproduce already reported tracking metric limitations by simulating specific tracking errors. Our benchmark reveals a before unreported limitation of the commonly used tracking metric AOGM. In addition, we make an implementation of our benchmark publicly available at: <https://github.com/mrhartmann/benchmark-mot-metrics>. This work presents the method and main results of the Bachelor’s thesis by Hartmann [10].

The concept of creating synthetically degraded tracking data for evaluation is established: For instance, Schott synthetically degraded tracks to investigate the robustness of extracted features to describe tracks [11], whereas Löffler et al. synthetically degraded segmentation data to investigate the robustness of tracking algorithms when provided with erroneous segmentation data [12]. However, to the best of our knowledge, we are the first proposing to synthetically degrade tracking data to evaluate tracking metrics.

The remainder of this paper is organized as follows: First, we describe how erroneous tracking data can be used to investigate tracking metrics and introduce how common tracking errors can be emulated. Then, we generate a benchmark data set of erroneous tracking results to evaluate popular MOT metrics. Finally, we discuss our findings and the limitations of our approach.

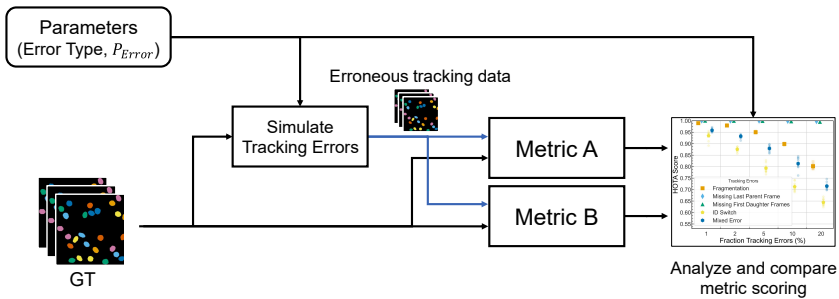


Figure 1: We degrade perfect GT data with simulated tracking errors to create erroneous tracking results, which are evaluated together with the GT by the metrics. The metric can be assessed by comparing the input parameters – the selected tracking errors and their percentage P_{Error} in the data set – with the resulting metric scoring.

2 Methods

We propose to degrade perfect GT data with simulated tracking errors to create erroneous tracking data. The synthetically degraded tracking data and perfect GT are forwarded to the tracking metric for evaluation. The flow of the data is shown in Figure 1. With the simulation of tracking errors, metric development becomes independent of the tracking algorithms that are commonly used to produce the tracking results needed for evaluation. To investigate whether a metric fulfills the property of monotonicity, the fraction of tracking errors can be chosen arbitrarily. We emulate the real-world tracking errors: ID switches, fragmentation, and mitosis errors which can be emulated separately or together to create degraded data covering several types of tracking errors. As GT we assume that each track in the GT is given by its segmentation masks, where segmentation masks belonging to the same track have the same ID. In addition, to model mitosis errors, a lineage file which indicates predecessor-successor links is needed.

In the following, we introduce the selected types of tracking errors by describing where they occur in real-world tracking scenarios and how we simulate them.

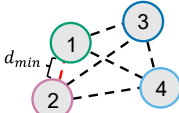

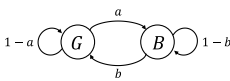
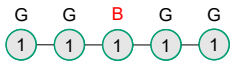

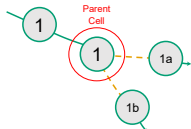
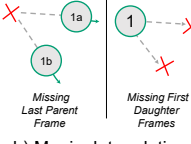
1. Choose Error Percentage $P_{Error} = n \%$			
2. Choose Error Type	ID Switch	Fragmentation	Mitosis
Implementation	 <p>a) Compute distances and sample one of closest pairs</p>  <p>b) Switch mask IDs and repeat in successive frames</p>	 <p>a) Define Markov model parameters</p>  <p>b) Generate and assign sequence of tracking states to select which instance masks to remove</p>  <p>c) Apply fragmentation and save new tracklets</p>	 <p>a) Detect all mother cells and select one for each iteration</p>  <p>b) Manipulate relation between mother and daughter cells</p>

Figure 2: Simulation of tracking errors. An error percentage and tracking error type is selected by the user to degrade the perfect GT data. The implementation is visualized for each error type. The circles are instance masks which are connected by links to form tracks. The processing steps are reiterated until the desired error percentage is reached in the data set.

2.1 ID Switches

Fast movements of objects can result in two objects switching positions between frames, causing ID switch tracking errors. Additional sources of ID switches are unpredictable changes of direction, erroneous detection of objects, or wrongly classified mitosis events [13, pp. 2124-2125]. A missing detection creates a scenario, where the tracking algorithm associates another mask to the ID, if it is close to the position of the original mask in the previous frame, causing a switch in tracks that propagates through the following time steps. This error exists even in data with perfect detection, as it is also reliant on the distance measure and linking method of the tracking algorithm. This error is handled separately from fragmentation in metrics [2, 3, 4] since preserving the ID is vital for many MOT fields [2, 3, 4, 5].

Simulation

We simulate ID switches by swapping the IDs of close tracks. Therefore, the Euclidean distance between all pairs of segmentation masks is calculated in each frame. The pair sampling approach is adapted from the merging operation for segmentation masks from Löffler et al. [12]. Based on the Euclidean distances between segmentation masks, we identify for each track its closest neighbor track. Then, for each ID switch, we sample a pair of tracks from the remaining 100 closest pairs of tracks, where each pair of tracks is assigned a sampling weight inversely proportional to their minimum distance. Hence, tracks which are closer are likelier to be sampled for an ID switch. This process, as visualized in Figure 2, is repeated until the desired fraction of ID switches in the dataset is reached. In addition, for cell data sets the daughter tracks need to be assigned to the switched ID to keep the predecessor-successor information intact.

2.2 Fragmentation

Fragmentation results from missing detections, which are often referred to as False Negatives (FN). Missing detections can originate from illumination variation, fluorescent marker wear off, shadows and occlusion, and more [9, p. 22]. As this error occurs frequently, many common performance metrics include a FN tracking error [2, 3, 4, 5]. Moreover, fragmentation can lead to additional tracking errors, for example ID switches and mitosis division ambiguities [11, p. 32].

Simulation

To fragment tracks, a two-state Markov model is used. Like the Markov model introduced by Schott [11], one state represents good tracking quality (G), whereas the other state represents bad tracking quality (B). We use this Markov model to generate fragmented tracks, by generating a sequence of states of the same length as a track, and assigning each instance mask at position k in the track the state at position k of the state sequence. Instance masks that are assigned

to the good tracking quality state G will remain, whereas instance masks that are assigned the bad tracking quality state B will be removed. This process is visualized in Figure 2.

To achieve the desired fragmentation – the fraction of deleted instance masks and the length of the resulting gaps – the transition probabilities a and b between the states can be adjusted. Instead of specifying a and b directly, we propose how transition probabilities can be calculated from two intuitive user input parameters: the desired error percentage (P_{Error}) and the fragmentation gap length (l_{Gap}).

The probability of returning to the good tracking state G in n steps, is $b(1-b)^{n-1}$, as the model stays $n-1$ steps in the bad tracking state B which has probability of $1-b$, before transitioning to G with probability b . Hence, the expected time until the model returns to the good tracking state $E[T_G]$, can be rewritten as a simple fraction by using the geometric sum

$$E[T_G] = \sum_{n=1}^{\infty} nb(1-b)^{n-1} = \frac{1}{b}, \quad \text{for } |1-b| < 1. \quad (1)$$

While $E[T_G]$ is the expected time until the model returns to the G state, it can also be interpreted as the average time spent in the B state. As instance masks which are assigned the B state will be removed, which creates gaps in the track, the average time spent in the B state can also be referred to as the gap length l_{Gap} . Hence, the transition probability b can be computed by specifying the user input parameter l_{Gap}

$$b = \frac{1}{E[T_G]} = \frac{1}{l_{\text{Gap}}}. \quad (2)$$

From b the missing parameter a can be computed using the steady state theorem [14, p. 176]

$$\pi_{\text{eq}} = \left(\frac{b}{a+b}, \frac{a}{a+b} \right). \quad (3)$$

The probabilities of the steady state can be set using the desired error percentage P_{Error} , which is specified by the user. As instance masks are only removed in the bad tracking state, and kept in the good tracking state we set π_{eq}

$$\pi_{\text{eq}} = (1 - P_{\text{Error}}, P_{\text{Error}}), \quad (4)$$

to reach the desired fraction of fragmentation errors. The transition probability for a can be calculated by combining Equation 3 and Equation 4

$$a = \frac{P_{\text{Error}}b}{1 - P_{\text{Error}}}. \quad (5)$$

The fragmentation is applied iteratively until the desired fraction of fragmentation errors is reached. To simulate different gap lengths, the fragmentation gap length parameter l_{Gap} can be adjusted. If no l_{Gap} is provided, the steady state of a Markov model π_{eq} is used, as given a long enough time, it provides an approximation of how long the Markov model will stay in each state. We set the probability a – switching to the bad tracking quality state B – equal to P_{Error} , whereas b – switching to the good tracking quality state G – is set to $1 - P_{\text{Error}}$, resulting in

$$b = \pi_{\text{eq},1} = 1 - P_{\text{Error}}, \quad a = \pi_{\text{eq},2} = P_{\text{Error}}. \quad (6)$$

2.3 Mitosis Tracking Errors

In cell data, an additional type of tracking error exists, which can occur due to missing detections or False Positives. During mitosis, cells face large changes in their scale and shape. As the temporal resolution of cell data sets is usually very low, the changes between two successive frames can be substantial and therefore can lead to erroneous detections or links [13, p. 1]. Moreover, the simultaneous division of nearby cells can lead to a high density of cells, further complicating the correct association between predecessor and successor tracks.

Simulation

We base the simulation of the mitosis tracking errors on the showcases first described by Chen et al.: Single Daughter Frame Missing, Last Mother Frame Missing and Both Daughter Frames Missing [1]. In addition, we added the No Mitosis Detection and Single Daughter Link Detected cases, which can be modeled by manipulating the lineage file only. To degrade mother-daughter links, first all mother cells are extracted using the lineage information from the GT. The mother-daughter links are then altered by removing the link from the lineage information or also adding fragmentation errors to the mother and daughter tracks. This process is visualized in Figure 2.

3 Experiment

To demonstrate our approach, we generate synthetically degraded tracking data using the just introduced tracking errors and evaluate four MOT metrics on them.

3.1 Data

To create synthetically degraded data sets, we select microscopy data showing cells as these data comprise all challenges encountered in general MOT and in addition contain splitting objects (cell divisions). We convert the synthetically degraded tracking data into two different file formats for metric evaluation, to evaluate specialized cell tracking metrics and general MOT metrics.

The synthetically degraded tracking data is generated by modifying the ground truth masks of the Fluo-N2DH-SIM+ 02 data set, from the Cell Tracking Challenge (CTC) [9]. We generate fractions of $n = 1, 2, 5, 10, 20\%$ of errors for each tracking error type and create for each combination of tracking error type and error fraction $N = 10$ runs. Every synthetically created tracking dataset is evaluated on the tracking metrics and the tracking score is averaged over the ten runs for each combination of tracking error type and error fraction.

3.2 Metrics

For evaluation, we select the metrics MOTA [3], IDF1 [4], HOTA [5], and TRA [2]. All metrics range between 0 and 1, where a higher score refers to a better tracking result. MOTA and IDF1 have been the most popular general MOT metrics and are widely used in benchmarks such as the MOTChallenge [8]. The recently proposed HOTA metric has been claimed to resolve issues of IDF1 and MOTA [5]. TRA is a normalized version of the AOGM [2] metric, which is used in the Cell Tracking Challenge benchmark [9]. Some flaws of TRA were already reported by Chen et al. [1].

3.3 Results

In the following, we analyze the impact of different tracking errors on the selected MOT metrics.

Effect of Different Errors on the four Metrics

First, the effect of different tracking errors – fragmentation, ID switches, mitosis errors, and a mixed error – on the metric scores of TRA, HOTA, MOTA, and IDF1 is analyzed, which is shown in Figure 3. For the mixed errors, the error percentage is split equally between the three tracking error types: $\frac{1}{3}$ of fragmentation, $\frac{1}{3}$ of ID switches, and $\frac{1}{3}$ of mitosis missing daughter frames errors.

For all tracking errors, a higher error percentage leads to a reduced metric score. For all metrics, mitosis errors are penalized the least. For IDF1 and HOTA, in Figure 3, the ID switch error has the biggest impact on the final metric scores. In contrast to MOTA and TRA, the same ID switch data sets are scored considerably lower by IDF1 and HOTA, with a score as low as 0.6 for IDF1. This is accompanied by notable differences in the metric scores between each error type for IDF1 and HOTA.

The cell tracking metric TRA, in Figure 3a, is effected strongly by the fragmentation, whereas ID switches have a low impact on the score. The TRA

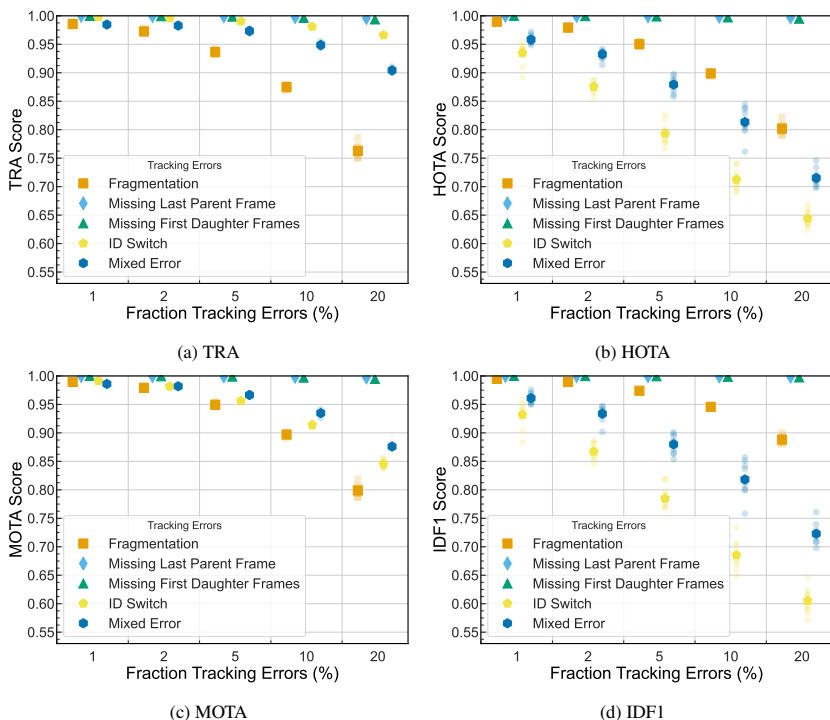


Figure 3: Evaluation of erroneous tracking data sets on the metrics. The link to the predecessor is set for all cases. For each run, a fixed percentage of ground truth tracks is modified to simulate tracking errors. The $N = 10$ single runs are shown in translucent markers, whereas the average of all runs is shown in solid color.

metric scores ID switches and mixed errors similarly and penalizes these errors only slightly, whereas the fragmentation scoring declines rapidly for increasing fractions of tracking errors.

For MOTA, Figure 3c shows a similar decline in score for ID switches, fragmentation and mixed errors.

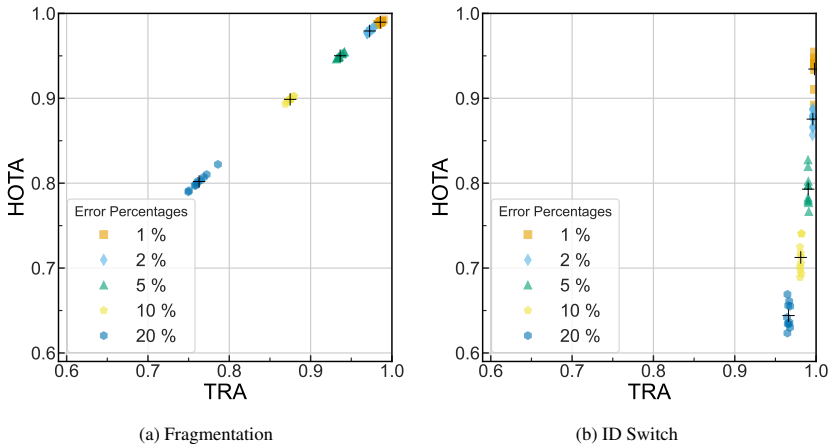


Figure 4: Comparison of the effect of fragmentation and ID switches on the TRA and HOTA metric. For each run, a fixed percentage of ground truth tracks is modified to generate tracking errors. The $N = 10$ single runs are shown as circles, whereas the average of all runs is shown as a black cross.

Comparison of Metric Scores for Fragmentation and ID Switches

Next, the HOTA and TRA metric are compared concerning their scoring of the tracking errors fragmentation and ID switches, which is shown in Figure 4. Both metrics show a similar decline when fragmentation errors increase. In contrast, ID switches affect HOTA stronger – 1% of ID switch errors result in a score under 0.95 – which is followed by a steep drop in the metric score for higher percentages. The effect of ID switches on the TRA score is considerably lower – 20% of ID switch errors result in a score larger than 0.95.

Difference of Keeping or Ignoring Predecessor Information for TRA Metric

Chen et al. first spotted an issue with the TRA measure concerning mother-daughter links around mitosis events with FN errors [1] based on small showcase examples. Using the different implemented mitosis errors, we reproduce this error scenario to investigate how much this limitation influences the final

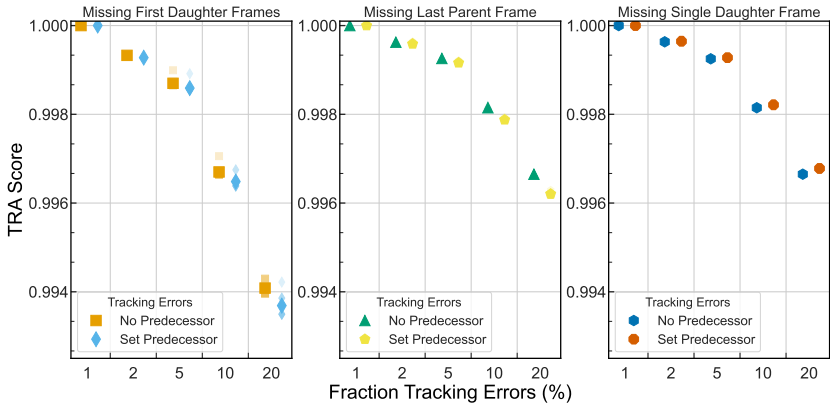


Figure 5: Difference of Linking the Predecessor for the Mitosis evaluation on TRA. For each run, a fixed percentage of ground truth tracks is modified to simulate tracking errors. The predecessor ID is either set or removed for modified tracks. The $N = 10$ single runs are shown in translucent markers. The average of all runs is shown in solid color.

metric score in a larger dataset. Each of the five plots in Figure 5 and Figure 6 includes two separate evaluations of synthetically degraded tracking data, one with and one without keeping the predecessor information for the erroneous tracks. Starting with two identical data sets, while applying tracking errors, the predecessor of the manipulated track is kept in the first and removed in the second.

A comparison for mitosis, fragmentation, and mixed errors is done, to analyze whether the TRA metric wrongly penalizes this correct information of the predecessor.

Overall, mitosis tracking errors, as shown in Figure 5, have a small effect on the TRA metric score. There is also nearly no variation of the scores between runs. The first two plots in Figure 5 score mitosis cases where either the last frame of the predecessor track is missing, or the first frames of both successor tracks are missing. In both cases, keeping the predecessor information results in worse scores by the TRA metric. The last plot in Figure 5 shows the case of a single missing successor frame, which is scored higher for keeping the predecessor information. The influence of the different mitosis errors on the TRA score decreases from left to right in Figure 5.

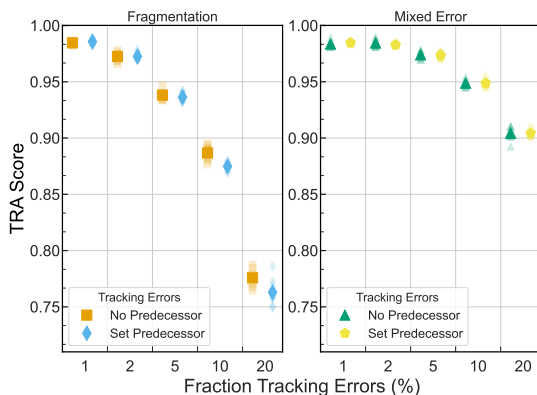


Figure 6: Difference of Linking the Predecessor for fragmentation and mixed tracking errors on the TRA metric. For each run, a fixed percentage of ground truth tracks is modified to simulate tracking errors. The predecessor ID is either set or removed for modified tracks. The $N = 10$ single runs are shown in translucent markers. The average of all runs is shown in solid color.

Figure 6 shows the influence of keeping the predecessor information on the TRA score in case of fragmentation and mixed errors. The fragmentation tracking errors in the left plot of Figure 6, are scored lower for setting the predecessor information, for error percentages of 5 – 20%.

4 Discussion

As mentioned by Luiten et al., IDF1 and MOTA have a bias towards detection and association, respectively [5]. Using the proposed benchmark, we could reproduce this observation as shown in Figure 3. The IDF1 score is strongly effected by ID switches, whereas fragmentation has a very low impact, as shown in Figure 3d. For MOTA, the mentioned bias towards detection is also visible in Figure 3c, as fragmentation is penalized the most.

Concerning the proposed alternative HOTA [5], ID switches and mixed errors are penalized similarly to IDF1, whereas fragmentation is penalized similarly to MOTA. These initial observations suggest that HOTA has a better balance

between detection and association, in contrast to MOTA or IDF1 alone, although the influence of ID switches on the HOTA score is still very strong.

Association errors have a very low impact on the TRA score, whereas fragmentation errors have a large impact, as shown in Figure 3a. The strong bias towards detection in the TRA metric is caused by high penalties for FN errors. The experiments with different types of mitosis errors, shown in Figure 5, support the statement of [1] that for AOGM and hence for TRA, which is derived from AOGM, it is more advantageous to ignore the information about the predecessor than to keep it in certain cases of mitosis errors. Although TRA is developed for cell tracking, erroneous mitosis detection is penalized very little, although the lineage of a cell is of high interest for instance during embryonic development [15, 16]. Additional metrics should be taken in consideration when comparing the quality of different cell tracking algorithms.

Moreover, using the proposed benchmark we could discover a yet unreported limitation of the AOGM and hence the TRA metric. If a track is fragmented, storing this information, for evaluation with the TRA metric, requires creating several tracks and link each track to its previous track. However, keeping this information is scored worse than discarding this information by not linking to the previous track. This observation matches with the already mentioned, flawed scoring of mitosis errors. Taken together, these observations reveal a weakness of the file format used in the TRA metric: the parent ID column indicates fragments belonging to the same track as well as mother-daughter relationships after cell division.

Limitations

Although we emulated real-world tracking errors, the simulated tracking results can in some cases appear artificial – e.g. long gaps (fragmentation) but the predecessor link is kept. Comparing the impact of tracking errors on different metrics should be done cautiously, especially when comparing different metrics which each other, as the method by which the error percentage is achieved differs. For fragmentation, each track has on average $n\%$ of its segmentation masks removed. For mitosis errors, only $n\%$ of the mitosis events

are modified. For ID switches, $n\%$ of tracks are selected for which their ID is switched pairwise. The results were computed with a limited amount of $N = 10$ runs for each combination of tracking error type and fraction of tracking errors. A larger number of runs is required to examine the variation between the runs in more detail. Moreover, we applied our method just to data from the microscopy image domain, so the method should be applied to data from other domains as well. Also, different tracking metrics can require a different storing of track and lineage information. When comparing metrics, the file format might not always include the same information and thus result in an unfair comparison.

5 Conclusion

We propose benchmarking tracking metrics by replacing the tracking algorithm with a method to synthetically degrade tracking data, emulating a set of frequently occurring real-world tracking errors. In a first analysis, we reproduced already reported limitations of popular tracking metrics and discovered a limitation of the AOGM metric.

Directions of future work are the extension of the proposed concept to provide a standardized approach to evaluate tracking metrics, using the approach to investigate tracking metrics that rank tracking algorithms without requiring a ground truth, and the adaptation of the proposed idea to develop benchmarks for metrics used in other tasks than MOT.

References

- [1] Ye Chen and Yuankai Huo. “Limitation of acyclic oriented graphs matching as cell tracking accuracy measure when evaluating mitosis”. In: *Medical Imaging 2021: Image Perception, Observer Performance, and Technology Assessment*, pp. 30–35, 2021.
- [2] Pavel Matula, Martin Maška, Dmitry V. Sorokin, Petr Matula, Carlos Ortiz-de-Solórzano, and Michal Kozubek. “Cell tracking accuracy

- measurement based on comparison of acyclic oriented graphs”. In: *PLOS ONE* 10(12): 1–19, 2015.
- [3] Keni Bernardin and Rainer Stiefelwagen. “Evaluating multiple object tracking performance: the CLEAR MOT metrics”. In: *EURASIP Journal on Image and Video Processing* 2008(1): 1–10, 2008.
- [4] Ergys Ristani, Francesco Solera, Roger S. Zou, Rita Cucchiara, and Carlo Tomasi. “Performance measures and a data set for multi-target, multi-camera tracking”. arXiv: 1609.01775. Sept. 6, 2016.
- [5] Jonathon Luiten, Aljos A. Os Ep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. “HOTA: a higher order metric for evaluating multi-object tracking”. In: *International Journal of Computer Vision* 129(2): 548–578, 2021.
- [6] Ido Leichter and Eyal Krupka. “Monotonicity and error type differentiability in performance measures for target detection and tracking in video”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(10): 2553–2560, 2013.
- [7] Martin Maška, Vladimír Ulman, David Svoboda, Pavel Matula, Petr Matula, Cristina Eder, Ainhoa Urbiola, Tomás España, Subramanian Venkatesan, Deepak M.W. Balak, et al. “A benchmark for comparison of cell tracking algorithms”. In: *Bioinformatics* 30(11): 1609–1617, 2014.
- [8] Patrick Dendorfer, Aljoša Ošep, Anton Milan, Konrad Schindler, Daniel Cremers, Ian Reid, Stefan Roth, and Laura Leal-Taixé. “MOTChallenge: a benchmark for single-camera multiple target tracking”. In: *International Journal of Computer Vision* 129(4): 845–881, 2021.
- [9] Vladimír Ulman, Martin Maška, Klas E. G. Magnusson, Olaf Ronneberger, Carsten Haubold, Nathalie Harder, Pavel Matula, Petr Matula, David Svoboda, Miroslav Radojevic, et al. “An objective comparison of cell-tracking algorithms”. In: *Nature Methods* 14(12): 1141–1152, 2017.

- [10] Michael Hartmann. “Simulation of Synthetically Degraded Tracking Data to Benchmark Multi Object Tracking Metrics”. Bachelor’s thesis. Karlsruhe: Karlsruhe Institute of Technology (KIT), 2022.
- [11] Benjamin Schott. “Interactive and Quantitative Knowledge-Discovery in Large-Scale 3D Tracking Data”. PhD thesis. Karlsruhe: Karlsruhe Institute of Technology (KIT), 2018.
- [12] Katharina Löffler, Tim Scherr, and Ralf Mikut. “A graph-based cell tracking algorithm with few manually tunable parameters and automated segmentation error correction”. In: *PLOS ONE* 16(9): 1–28, 2021.
- [13] Seungil Huh, Sungeun Eom, Ryoma Bise, Zhaozheng Yin, and Takeo Kanade. “Mitosis detection for stem cell tracking in phase-contrast microscopy images”. In: *2011 IEEE International Symposium on Biomedical Imaging: from Nano to Macro (ISBI 2011)*, pp. 2121–2127, 2011.
- [14] Robert G. Gallager. “Discrete Stochastic Processes”. Vol. 321. The Springer International Series in Engineering and Computer Science, Communications and Information Theory, 1996.
- [15] Khaled Khairy and Philipp J. Keller. “Reconstructing embryonic development”. In: *genesis* 49(7): 488–513, 2011.
- [16] Andrei Y. Kobitski, Jens C. Otte, Masanari Takamiya, Benjamin Schäfer, Jonas Mertes, Johannes Stegmaier, Sepand Rastegar, Francesca Rindone, Volker Hartmann, Rainer Stotzka, et al. “An ensemble-averaged, cell density-based digital model of zebrafish embryo development derived from light-sheet microscopy data with single-cell resolution”. In: *Scientific Reports* 5(1): 8601, 2015.