2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops

Challenges of Ground Truth Evaluation of Multi-Target Tracking

Anton Milan[1] Konrad Schindler[2] Stefan Roth[1]
[1]Department of Computer Science, TU Darmstadt [2]Photogrammetry and Remote Sensing Group, ETH Zürich

Abstract

Evaluating multi-target tracking based on ground truth data is a surprisingly challenging task. Erroneous or ambiguous ground truth annotations, numerous evaluation protocols, and the lack of standardized benchmarks make a direct quantitative comparison of different tracking approaches rather difficult. The goal of this paper is to raise awareness of common pitfalls related to objective ground truth evaluation. We investigate the influence of different annotations, evaluation software, and training procedures using several publicly available resources, and point out the limitations of current definitions of evaluation metrics. Finally, we argue that the development an extensive standardized benchmark for multi-target tracking is an essential step toward more objective comparison of tracking approaches.

1. Introduction

Measuring the performance of novel methods is not only important for monitoring the progress compared to previous approaches in absolute terms, but also for assessing which contribution has the largest influence on the targeted application. However, quantitatively evaluating computer vision algorithms is not a straightforward task. The reasons for this are varied. On one hand it is not always obvious what the 'correct' solution should look like. For some applications this question may be easier to answer than for others. In low-level tasks, such as image restoration or deblurring, the goal is usually to precisely reconstruct the original, artifact free image. But even in this seemingly clear case the ground truth might be either unavailable, or contain some level of noise itself [31]. For higher-level problems, *e.g.* image classification or object detection, it may seem easy to manually determine whether a certain object is present in the image, or not. The answer becomes ambiguous, however, if the object is only partially visible, either due to occlusion or cropping [28]. For tasks such as segmentation, the situation becomes even more challenging. When multiple people are asked to annotate the outline of the same object in the same image, one will get many different contours [21].

This paper considers the problem of quantitatively evaluating multiple target tracking. Here, the ground truth is not always well-defined either. Although most human annotators would agree on the presence or absence of a person in a certain image region, pinpointing the precise location poses a more difficult task. As a matter of fact, we will see in the following how large the spatial displacement between independent ground truth annotations can be (*cf*. Fig. 1). The second challenge of evaluation is measuring the similarity between the obtained solution and the ground truth. To that end, several protocols and metrics have been proposed and have in fact become widely accepted; we will review these in Sec. 2. Nonetheless, their definition remains somewhat ambiguous and involves meta-parameters, such as the overlap threshold. Another important issue specifically concerns tracking-by-detection methods. These methods heavily rely on the output of an object detector. As a consequence, a better detector will most likely yield better tracking results. Therefore, it is essential that the same input, *i.e*. the same set of detections, is used if one is interested in only comparing the merits of different tracking algorithms themselves.

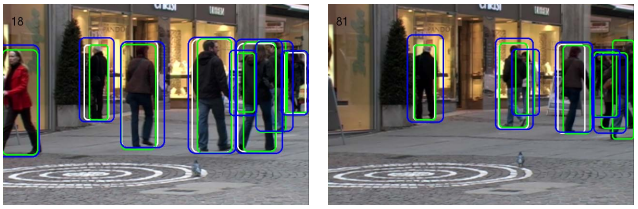


Figure 1. Three different publicly available ground ruth annotations on the *TUD-Stadtmitte* sequence. The original annotations [1] (white) do not contain any occluded pedestrians, while the other two sets [4] (green) and [30] (blue) show large deviations regarding the bounding box size.

This paper makes the following contributions: *(i)* We summarize important multi-target tracking evaluation metrics and discuss their respective advantages and disadvantages; *(ii)* we present the challenges of obtaining the ground truth and investigate its influence on the reported performance; *(iii)* we systematically compare different implementations of evaluation software; *(iv)* we discuss the impact of parameter tuning on a limited ground thruth dataset; and *(v)* we raise the question of the importance of more standardized ground truth benchmarks toward enabling a fair comparison of future multi-target tracking research.

978-0-7695-4990-3/13 $26.00 © 2013 IEEE
DOI 10.1109/CVPRW.2013.111

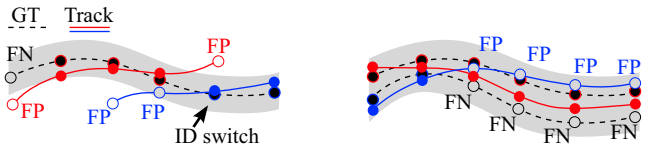


Figure 2. Illustration of the *CLEAR MOT* components. Events that are classified as correct are denoted with solid circles. Errors are indicated with empty circles. The influence of track-to-ground-truth assignments is illustrated on the right: A 'wrong' decision at the beginning of a trajectory leads to persistent errors over the whole sequence.

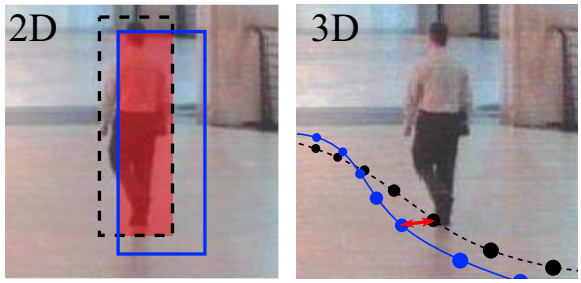


Figure 3. Measuring correspondence as bounding box overlap (2D) or as distance on the ground plane (3D).

2. Metrics for Quantitative Evaluation

To quantitatively measure the performance of one multi-target tracking method and optionally to compare it with others, a clearly defined protocol is required. Unfortunately, objectively assessing the quality of a multi-target tracking solution is not an easy task. Furthermore, the 'perfect' solution, or *ground truth*, is needed to serve as reference. We will discuss these issues and related challenges in Sec. 3. Before doing so, we first review various protocols that are currently used for evaluating multi-target tracking.

2.1. CLEAR MOT

To evaluate the correctness of any tracker at least three entities need to be defined:

- the tracker output (or hypothesis) $\mathcal{H}$, which is the result of the tracking algorithm;
- the correct result, or ground truth $\mathcal{GT}$; and
- a distance function $\overline{d}$ that measures the similarity between the true target and the prediction.

Note that these requirements are kept very general without any assumptions on the concrete representation or on the exact definition of the distance function. Intuitively, one wishes to incorporate and rate every possible error that a solution may contain. One of the protocols that follow this goal is the *CLEAR MOT* evaluation [9]. It emerged from the *CLEAR* Workshop[1] in 2006 and has since been widely accepted as a standard evaluation tool by the tracking community. The two proposed quantities, *MOTA* and *MOTP* on the one hand measure the number of errors that occur during tracking, and on the other hand assess the tracker's precision, *i.e.* its ability to localize the target in the image. Let us now take a closer look at the different components that give rise to these quantities.

MOT Accuracy. As in object detection, the two most common errors in multi-target tracking are false positives *(FP)* and false negatives *(FN)*. The former correspond to spurious tracking results that do not match any ground truth trajectory, while the latter ones are annotated targets that are not identified by the tracker. To determine whether a target is being tracked, a correspondence between true targets and hypotheses must be established. This is usually done in a greedy manner in consideration of temporal matching, rather than independently in each frame. More precisely, if and only if a target is not tracked, it is assigned the closest unmatched hypothesis. Otherwise, the correspondence from the previous frame is maintained. To decide, whether a track is a potential candidate for a match, a distance between all hypotheses and all targets must be computed. If the distance between a track-object pair is small enough, they can potentially be matched. Note that this procedure to compute the correspondences is application and representation specific. If both the output and the annotations are described by bounding boxes, then usually the *PASCAL* criterion

$$\overline{d}(\mathcal{H},\mathcal{GT}) = \frac{\text{bbox}(\mathcal{H}) \cap \text{bbox}(\mathcal{GT})}{\text{bbox}(\mathcal{H}) \cup \text{bbox}(\mathcal{GT})}, \tag{1}$$

i.e. the intersection over union (Jaccard index) or the relative overlap of the true and the predicted bounding boxes, determines the similarity between the two. Here, 0 means no overlap and 1 means that both bounding boxes are identical. The most common threshold for considering a pair as correct is 0.5. For 3D tracking, it may be more reasonable to compute the correspondence directly in world coordinates (*cf*. Fig. 3). In this case, the Euclidean distance between the centroids of two objects gives a suitable estimate. For people tracking, the foot position, *i.e.* the center of the bottom edge of the bounding box, defines the target's centroid and a threshold of 1 meter is typically used.

Recall that the goal of multi-target tracking is not only to find all objects and suppress all false alarms, but also to correctly follow each object over time. In other words, the reconstructed trajectory should adhere to one specific object from the moment of entry until it exits the scene. Whenever there is a mismatch between a hypothesis and the corresponding ground truth trajectory, an identity switch *(ID)* occurs, which is counted as an error. A simple example illustrating these three error types is depicted on the left-hand side of Fig. 2. Although temporally-aware target-to-tracker matching suppresses unnecessary identity switches, it may lead to undesirable artifacts, as illustrated in Fig. 2 *(right)*.

[1] http://clear-evaluation.org

To formally define *MOTA*, let $FP(t)$, $FN(t)$ and $ID(t)$ denote the number of false positives, missed targets and identity switches at time t, respectively. Further, let $N_{\text{GT}}(t)$ denote the number of annotated targets at time t. Then the *MOTA* score is computed as

$$MOTA = 1 - \frac{\sum_t \Big(FP(t) + FN(t) + ID(t) \Big)}{\sum_t N_{\text{GT}}(t)}. \qquad (2)$$

Note that if a solution contains no errors, *i.e*. the numerator sums up to 0, then the accuracy equals 100%. This value decreases as the number of failures increases. The *MOTA* score can also result in negative values and is in fact unbounded (from below). Allowing for a negative accuracy may seem unnatural, but this can only occur when the number of errors is larger than the number of targets in the scene, which only rarely happens in practice. Combining the quality of a tracking result into a single number has both positive and negative consequences. On the one hand, it enables a simple comparison. On the other hand, the strengths and weaknesses of a particular method may become concealed. It is therefore preferable to present all available numbers.

MOT Precision. The *MOTA* described above measures the discrete number of errors made by the tracker. On the contrary, the *MOTP* avoids such hard decisions and instead estimates, how well a tracker localizes the targets. Again, in its general form it is defined as

$$MOTP = \sum_{t,i} \bar{d}\Big(\mathcal{GT}_i^t, \mathcal{H}_{g(i)}^t \Big) \Big/ \sum_t m_t, \qquad (3)$$

where $\mathcal{GT}_i^t$ and $\mathcal{H}_{g(i)}^t$ are the target and its associated hypothesis, respectively, and m_t is the number of matches at time t. Intuitively, it provides the average distance over all matched pairs. In 2D, this number directly represents the average overlap of matched bounding boxes, while for the evaluation in 3D it is usually normalized to the hit/miss threshold such that it provides a percentage value between 0 and 100%. We point out that *MOTP* is a rather rough estimate of the performance, because it heavily relies on the quality of the annotations, which are often inaccurate or even ambiguous as we will see below.

2.2. Further metrics

Next to the widely used *CLEAR* metrics, other performance measures have been introduced in the literature.

Trajectory-based measures [29] assess the performance on entire trajectories rather than on a frame-by-frame basis. Their definition has later been refined [18] to capture some ambiguous cases. A target is often tracked correctly only for a certain period and not for its entire presence in the scene. To quantify this property, a trajectory can be classified as mostly tracked *(MT)*, partially tracked *(PT)* and mostly lost *(ML)*, see Fig. 4. A target is considered mostly lost when it is found during less than 20% of its presence. Similarly, a target is mostly tracked when at least 80% of its ground truth trajectory is found. Consequently, all other trajectories are partially tracked. Note that identity switches do not play any part in the computation of these figures. Finally, track fragmentations *(FM)* count how many times a ground truth trajectory changes its status from 'tracked' to 'not tracked', *i.e*. each time it is lost by the current hypothesis.

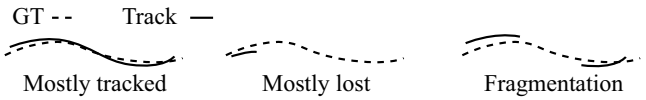


Figure 4. Trajectory-level measures [18].

Configuration distance and purity proposed by [24] provide a more detailed inspection of each tracker, each trajectory and the configuration state. In particular, they allow multiple tracker-to-target assignments, but count these as multiple trackers or multiple objects errors. The configuration distance measures the difference between the number of predicted and true targets, and indicates the bias toward more false alarms or toward missed targets. Further measures, such as tracker or object purity, are somewhat related to the mostly tracked definition above, but provide a more detailed evaluation on the produced hypotheses and not only on the ground truth trajectories. Since these metrics are rarely used, we do not employ them here.

Global mismatch error *(gmme)* [7] is an extension of the traditional count of ID switches. Instead of only counting the number of instantaneous swaps, *gmme* counts all frames *after* the swap as erroneous.

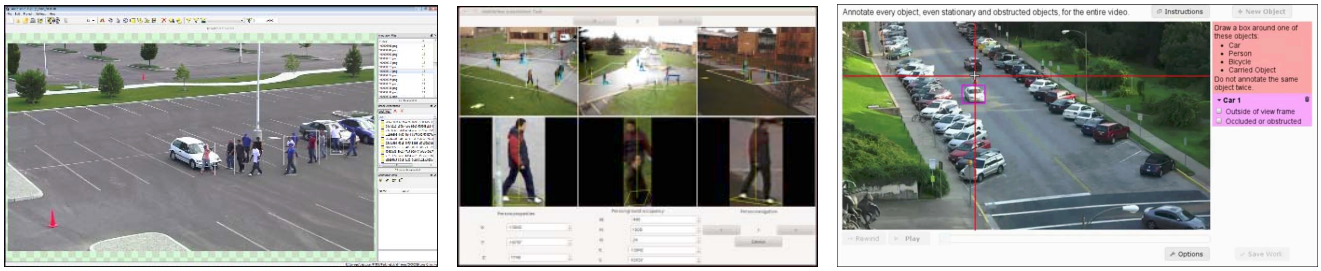
Related applications, such as people counting or queue length estimation also require quantitative evaluation, but a standardized protocol has not been defined yet.

To summarize so far, there is no single objective measure for the quantitative ground-truth evaluation of multi-target tracking algorithms that incorporates all important aspects. Many proposed protocols follow a similar intuition, but are somewhat ambiguous in their exact definitions. As a result, the computed numbers usually give a fair assessment of the overall performance, but may vary depending on the concrete implementation of the evaluation software. We will discuss this aspects further in Sec. 3.4.

3. Ground Truth for Multi-Target Tracking

3.1. Obtaining ground truth

Annotating images is a tedious task. The most naive, but very common way is to draw rectangles around the objects of interest to define their bounding boxes frame by frame. There are, however, several software packages that assist the user in order to facilitate the annotation process.



(a) AnnoTool (b) MultiView (c) VATIC

Figure 6. User interfaces of three different annotation tools.

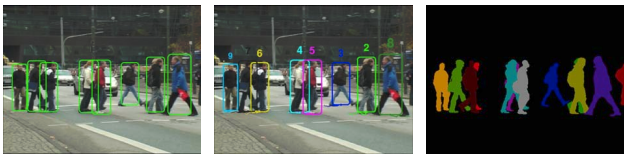
(a) (b) (c) (d) (e)

Figure 5. Ambiguous cases. The mannequin in the display window *(a)*, a man inside a phone booth *(b)*, a clearly visible reflection *(c)*, a child *(d)* and a partially occluded person *(e)* are all missing in the ground truth and can potentially cause erroneous false positives.

Annotation tools. Annotations for *TUD*, *EPFL* and many of the *PETS2009* sequences[2] were created using AnnoTool2. It allows one to linearly interpolate the location and the size of the bounding box between key frames, which leads to a significant speed up. The tracks were smoothed afterwards to reflect natural people motion. While keyframe interpolation facilitates the annotation process, one must bear in mind that it also leads to an approximation, since the 'true' target motion hardly ever precisely follows a linear (or a higher order polynomial) pattern. VATIC[3] is a more recent annotation tool [27]. It offers an integrated interface to Amazon's Mechanical Turk such that one can leverage the power of crowdsourcing for the annotation task. Finally, [26] provide an annotation software specifically designed for a multi-camera setup[4]. Interestingly, there a target is defined by its actual height in world units and the rectangular area that it occupies on the ground plane instead of the usual bounding box representation. Screenshots of all three annotation tools are shown in Fig. 6.

3.2. Annotation quality

As we already briefly discussed above, different annotations of the same video sequence may vary quite severely, both in terms of quality and in terms of the actual information that is provided (see Fig. 7). Many of the widely used tracking datasets, including the *ETHMS* [12] and the *TUD* [1, 2] sequences, were originally annotated for the purpose of evaluating person detection. The annotations provided by the authors of these datasets only included bounding boxes of people without their corresponding IDs. Moreover, partially occluded pedestrians (approximately 50% and more) are ignored by the annotators, since they are not expected to be found by the detector. An important ability of a multi-target tracker, however, is to keep track of individuals over time, even through complete occlusions. Therefore, performance results reported on these sequences either ignored the number of identity switches [10] or resorted to manual counting [19], which is both tiresome and inaccurate.



(a) (b) (c)

Figure 7. Different level-of-detail. Next to unordered bounding boxes *(a)*, annotations for multi-target tracking should also provide the corresponding ID of each box *(b)*. In some cases even a pixel-level segmentation mask is available *(c)*.

Annotations can also be provided on different levels-of-detail, both spatially and in terms of temporal resolution. For example, [17] provide pixel-level segmentation masks for each person in the *TUD-Crossing* sequence. Due to the required effort in obtaining such detailed information, it is only available every 10^{th} frame. The authors of the *ParkingLot* sequence [23] annotate every 3^{rd} frame in two versions; one includes only fully visible targets, the other also includes occluded ones. The ground truth for the *EPFL* datasets [14] is discretized both spatially and temporally. These annotations include the cell occupancy of a ground plane grid every 25^{th} frame, *i.e.* only once every second.

Ambiguities are inevitable in annotations of real-world sequences. Some of the more common ones are illustrated in Fig. 5. While objects that look like targets, such as the

[2] http://www.gris.tu-darmstadt.de/~aandriye/data
[3] http://mit.edu/vondrick/vatic
[4] http://web.eee.sztaki.hu/~ucu/mvatool

Table 1. Evaluating the same tracking result obtained with the public implementation of [3] w.r.t. different ground truth annotations.

Gr. truth	Rcll	Prcn	GT	MT	ID	FM	MOTA	MOTP
white [1]	90.1	97.1	18	11	3	3	87.1	83.3
green [4]	69.3	99.5	10	4	7	6	68.3	76.6
blue [30]	72.1	99.1	10	4	7	6	70.8	71.9

mannequin, or reflections should not be annotated, small, occluded or blurred targets ought not be ignored. One difficulty arises at image borders where targets become partially cropped. Especially in crowded scenarios where targets frequently enter and exit the field of view, such errors tend to accumulate, preventing any tracking method to achieve 100% accuracy. To mitigate this effect, we propose to use several annotation sets and average the performance. To analyze how much different annotations affect the measured performance we conduct two experiments: *(i)* We evaluate the identical tracker output on three different sets of ground truth data. *(ii)* We evaluate the accuracy of one ground truth annotation w.r.t. the others for all three combinations.

The *TUD-Stadtmitte* sequence [1] has become fairly popular and is frequently used for evaluating detection as well as tracking quality. Somewhat surprisingly, several 'ground truths' are publicly available for this short sequence, which differ significantly from one another [1, 4, 30]. The reasons for this may be that the original annotations do not contain target IDs and that occluded pedestrians are not annotated. For the following experiment we obtained the IDs by greedy nearest neighbor linking, but did not connect trajectories across occlusion gaps. The other two sets were annotated independently by two different groups [4, 30]. Bounding boxes from all three ground truth sets are overlaid and shown in Fig. 1. A coarse qualitative assessment reveals that the boxes in the dataset from [30] (blue) are much larger than those in the other two. Quantitative results are listed in Tab. 1. The numbers are computed in 2D with an overlap threshold of 0.5. As expected, the recall is much higher on a ground truth with fewer annotated bounding boxes (white). But there is still a noticeable gap in tracking accuracy *MOTA*, and an even larger one in tracking precision *MOTP* between the two other annotation sets that were created specifically for multi-target tracking evaluation. This observation clearly demonstrates that the computed figures may vary greatly depending on what ground truth annotation is used.

In our second experiment we use one of the three sets of annotations as the "solution" and evaluate it with respect to the other two. Obviously, one cannot expect that the bounding boxes are always perfectly aligned to each other across various sets. However, it is reasonable to assume that at least different annotations would agree on the presence or absence of targets in the image. The figures shown in Table 2 are rather disillusioning. For instance, the top two rows show how the *white* ground truth scores when evaluated on the *green* and on the *blue* one. Obviously, the recall stays low since occluded people are not present in this annotation. But even when comparing the more complete annotations to each other (rows 4 and 6), the overall accuracy *(MOTA)* remains below 70%. The reason here is that the difference in bounding box sizes leads to an overlap that is less than 50% in many cases, hence the annotations are counted as false positives. This is particularly problematic, since the output of the tracker given in Table 1 actually produces better quantitative results than a different ground truth. This once again shows that bounding box annotations are in fact quite ambiguous.

To conclude, both the quality and the level-of-detail can vary significantly across annotations, even for the same video sequence. A misalignment of bounding boxes in different annotation sets may not only lead to a lower tracking precision, but can severely impair the overall performance numbers due to wrongly counted errors. It is therefore always important to state which ground truth data was used for measuring performance of a certain tracker output.

3.3. Metrics ambiguity

Having analyzed the impact of different ground truth annotations on the resulting performance, we now take a closer look at the protocols themselves. In Sec. 2, we formally defined several methods for measuring the performance of a tracking system and discussed some of the problems related to the quantitative evaluation. Here, we will follow up on this issue and point out concrete deficits of the existing definitions. Throughout this paper, we employed two sets of evaluations metrics, *CLEAR MOT* [9] and the trajectory-based measures of [18]. As we will see in Sec. 3.4, computing the same error measure is not clearly defined since various evaluation scripts do not produce identical numbers. Besides possible implementation discrepancies, the metrics' definitions themselves carry ambiguities.

Distance. To establish correspondences between the true objects and the produced results, a distance measure is required to assess how similar or how close the hypothesis is to the ground truth object. One possible choice is the *PASCAL VOC* criterion, which measures the overlap between two bounding boxes (*cf*. Eq. (1)). When tracking is performed directly in the world coordinate system, the standard Euclidean distance between the objects' centers can be employed. In both cases, a threshold is required that determines whether a target-hypothesis pair constitutes a potential match or not. In other words, the evaluation procedure itself is dependent on at least one parameter that should always be stated. For the overlap criterion, a threshold of 0.5 has been widely accepted. For measuring distances in world coordinates, [25] propose 500mm. However, the main application there is to track multiple people in meetings in a

Table 2. A quantitative comparison of various ground truth annotations with respect to one another.

"Solution"	Ground truth	Rcll	Prcn	GT	MT	ML	ID	FM	MOTA	MOTP
white	green	75.1	100.0	10	6	0	8	288	74.4	81.1
	blue	77.2	98.5	10	6	0	10	252	75.2	68.9
green	white	100.0	75.1	18	18	0	0	0	66.8	81.1
	blue	85.1	81.5	10	9	1	0	165	65.8	66.7
blue	white	98.5	77.2	18	18	0	2	13	69.2	68.9
	green	81.5	85.1	10	8	1	0	214	67.2	66.7

rather small area. We found that such a threshold is too conservative for outdoor scenes for two reasons: First, in surveillance settings cameras are usually far away from the scene showing a much larger area of interest, such that targets only occupy a small image region. Second, the camera calibration may be unreliable, *e.g.* due to a low view point. In both cases targets that are only slightly misplaced on the image induce a large 3D error. Consequently, a threshold that is too small will lead to an undesirable behavior when correct results are counted as false alarms, while the true target remains untracked. We therefore use a 1 meter hit/miss threshold throughout all experiments.

Assignment. One further ambiguity of tracking metrics lies related to how the output hypotheses are assigned to the ground truth objects, which is not specified explicitly. A greedy assignment strategy is arguably the simplest choice, but does not lead to the best matching. A typical case of non-optimal assignment is illustrated in Fig. 2 (right). One way to avoid this is to perform a two-pass matching with the Hungarian algorithm, as is done, *e.g.*, by [30].

Error weighting. Recalling the definition of *MOTA* from Eq. (2), all three types of errors (FP, FN and ID) are weighted equally as suggested by [9, 25]. Naturally, each error type can be weighted individually according to its importance for the respective application. For offline motion analysis it may be important to reconstruct correct, identity preserving trajectories, while finding absolutely all present targets is less crucial. A higher weight for identity switches may therefore be more desirable. On the contrary, a driver assistance system should detect every single pedestrian and at the same time maintain a low number of false positives to avoid unnecessary warnings. On the other hand it is less relevant to keep the identity of each person over time. In such case, the aim is to achieve the highest possible precision and recall while less attention is paid to the number of ID switches. This may also be the motivation of [11], who impose a logarithmic weight on the number of mismatch errors when computing the *MOTA* score.

3.4. Evaluation software

We will now investigate whether the particular implementation of the evaluation protocol has an impact on the computed measures. To that end we evaluate the same tracking result from Tab. 1 on one particular ground truth, but with different evaluation scripts. All tested scripts provide the raw number of false alarms and missed targets, such that precision and recall can easily be computed. An evaluation script by Masi and Lisanti[5] computes the *CLEAR MOT* metrics, but not the trajectory-based ones [5]. Yang's software [30], which operates on bounding boxes in 2D, additionally computes the number of mostly tracked and mostly lost trajectories, but does not provide the average overlap. Unfortunately, these are difficult to extract, since only the binary executables are available. We also employ Bernardin's implementation provided for the original *CLEAR* challenge [9]. Finally, our own Matlab script[6] computes all sets of metrics and can operate on bounding boxes as well as on the ground plane. All available numbers are listed in Table 3. The values in parentheses are not part of the script output but are rather computed based on the provided number of false positives, false negatives and identity switches. Note the extremely high number of detected mismatches in Masi & Lisanti's implementation. This number is probably not very reliable because the authors state in their documentation that "ID switches should be carefully counted by visual inspection". Other than that, the figures in Table 3 do not deviate substantially. Nonetheless, for a meaningful comparison it is crucial to use exactly the same evaluation software.

3.5. Training and testing

Many tasks in computer vision are approached by designing models that need to be trained or tuned, *i.e.* fitted to the annotated training data, to make predictions about unseen data. To enable a fair comparison between various methods, some areas offer well-established benchmarks with pre-defined training and test sets. To name a few, there is the *PASCAL* challenge for object detection or segmentation [13], the Middlebury benchmark for multi-view stereo [22], or KITTI for stereo or optical flow [15]. Although several multi-target tracking datasets are frequently used in the literature [3, 20, 23], there is no established consensus of how to separate the data into training and testing sets. The

[5] http://www.micc.unifi.it/masi/code/clear-mot
[6] http://goo.gl/8ZTrM

Table 3. Evaluating the same tracking result [3] with respect to the same ground truth [4], but with different evaluation scripts. The top part considers an evaluation in 2D, while the bottom results are computed in 3D on the ground plane.

Space	Evaluation software	Rcll	Prcn	FP	FN	GT	MT	ML	ID	FM	MOTA	MOTP
2D	Andriyenko et al. [4]	69.3	99.5	4	355	10	4	0	7	6	68.3	76.6
	Bagdanov et al. [5]	67.9	99.7	4	355	10	-	-	16	-	67.6	77.0
	Yang & Nevatia [30]	67.6	98.0	16	373	10	2	1	2	3	(66.0)	-
3D	Andriyenko et al. [4]	59.4	85.3	118	469	10	2	0	9	9	48.4	59.8
	Bernardin & Stiefelhagen [9]	(59.4)	(85.3)	118	469	10	-	-	10	-	48.4	(59.8)

Table 4. Influence of training procedure.

Tracker	Training	Rcll	Prcn	ID	FM	MOTA	MOTP
[3]	per seq.	68.6	93.8	49	30	62.8	64.7
	global	59.1	95.5	29	22	54.9	66.7
	cross val.	60.3	90.9	31	24	49.2	65.2
[20]	per seq.	57.1	95.4	160	124	49.2	66.0
	global	57.6	92.6	149	123	48.5	65.6
	cross val.	57.1	92.5	144	119	47.7	65.6
[4]	per seq.	64.7	92.4	61	46	58.0	64.5
	global	60.7	90.7	52	41	52.1	65.4
	cross val.	60.7	90.7	52	41	52.1	65.4

common strategy to report the performance of a tracking method is to tune the parameters to a fixed set of sequences, thereby treating them as training and test data at the same time. Obviously, this is not ideal since the model is overfitted to the chosen data and will usually perform considerably worse on unseen data. To nonetheless reduce the effect of overfitting, it is considered good practice to choose several datasets that exhibit strong variations in person count, view point and resolution, while keeping the parameters fixed.

To examine the influence of training, we perform an experiment on six datasets: five *PETS* sequences and *TUD-Stadtmitte*. We tune the parameters for three tracking methods [3, 4, 20] in three different ways: A per-sequence search, a global tuning over all sequences simultaneously, and leave-one-out cross validation. Parameter tuning is performed by a random search [8] w.r.t. *MOTA* starting from the default set in all cases. The results are summarized in Tab. 4. Our intention here is not to compare the performance of different tracking approaches to each other, but rather to point out that the particular choice of training data and training procedure may have a large impact on the computed performance. Note that [3] is more flexible than [4, 20] and can be tuned much more accurately to each specific sequence. However, using cross validation, the mean accuracy (*MOTA*) drops by over 10 percentage points. This case study shows that the two other methods generalize better to unseen data, and suggests that cross-validation may need to be considered seriously for evaluating the robustness of multi-target tracking algorithms.

3.6. Toward a benchmark

We believe that a standardized multiple target tracking benchmark consisting of a variety of diverse video sequences is needed to facilitate comparison between state-of-the-art methods. The only currently existing method (that we are aware of) to objectively measure the performance of a tracking algorithm is to send the results on the *S2L1* sequence (represented by bounding boxes) to the *PETS* organizers [11]. The computed *CLEAR MOT* metrics, evaluated with respect to unpublished ground truth, are then sent back to the authors. Provided that current methods achieve near perfect results on that particular sequence, it is time to move on to more challenging datasets. Clearly, such benchmarks entail the risk of shifting the research goals from developing innovative techniques to pushing the numbers higher on that particular data. However, previous benchmarks, such as Middlebury [6] or *PASCAL* [13] for example, show that the raw ranking is not the only criterion for how a specific method is valued in the community. In fact, despite caveats of benchmarks both projects considerably boosted research in their respective area of vision. In the following, we point out some of the main issues to be considered in future when compiling a benchmark for multi-target tracking.

Data selection. Most of the current approaches can be easily tuned to a specific sequence to achieve good performance. To assess the general applicability of a particular method, it is crucial to test it on a wide range of various scenarios using a single set of parameters. A benchmark should therefore contain several scenes that show substantial variability in camera angle and motion, person count and resolution. Furthermore, a clear separation of data for training and validation on the one hand, and for testing on the other hand should be defined to enable consistent parameter tuning or learning.

Detections. Another issue further complicates elementary comparison. Most current multi-target trackers perform tracking-by-detection, *i.e.* the actual input data are not the raw images, but a set of independently precomputed detections. Clearly, the performance of both the data association and the reconstruction of trajectories will greatly depend on the quality of the detector. One way to evaluate various trackers fairly might be to provide a standard detection set

for each method. However, this is not straightforward to implement in practice, since different methods require different types of input. Some rely on plain bounding boxes [20], others also consider the confidence value of each detection [3, 4] – which is non-trivial to calibrate in general – while other approaches work on contours of pedestrians [16]. A solution may be to offer several test categories so that each method can be compared in a meaningful way.

Evaluation. Obviously, one single ground truth and evaluation script should be made available for a fair comparison. It is conceivable to withhold the annotations of the test set to avoid overfitting. This would, however, require a centralized evaluation tool. It may be beneficial to restrict the number of evaluations of the same method, similar to the practice followed by the Middlebury benchmark.

4. Conclusion and Outlook

In this paper we presented several common pitfalls related to evaluating multiple target tracking. We systematically investigated the influence of different ground truth annotations, evaluation scripts, and training procedures on publicly available data. We found that all of these aspects may have a significant impact on the resulting numbers, making a fair comparison rather challenging. Hence, we argue that a unified benchmark is important toward a more meaningful quantitative evaluation. At least it is essential to state which ground truth and which evaluation script the reported numbers are based on, or, even better to make the actual results publicly available alongside every publication.

References

[1] M. Andriluka, S. Roth, and B. Schiele. Monocular 3D pose estimation and tracking by detection. *CVPR 2010*.

[2] M. Andriluka, S. Roth, and B. Schiele. People-tracking-by-detection and people-detection-by-tracking. *CVPR 2008*.

[3] A. Andriyenko and K. Schindler. Multi-target tracking by continuous energy minimization. *CVPR 2011*.

[4] A. Andriyenko, K. Schindler, and S. Roth. Discrete-continuous optimization for multi-target tracking. *CVPR 2012*.

[5] A. Bagdanov, A. Del Bimbo, F. Dini, G. Lisanti, and I. Masi. Compact and efficient posterity logging of face imagery for video surveillance. *IEEE Multimedia*, 19(4):48–59, 2012.

[6] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski. A database and evaluation methodology for optical flow. *Int. J. Comput. Vision*, 92(1):1–31, 2011.

[7] H. Ben Shitrit, J. Berclaz, F. Fleuret, and P. Fua. Tracking multiple people under global appearance constraints. *ICCV 2011*.

[8] J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *JMLR*, 13:281–305, 2012.

[9] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: The CLEAR MOT metrics. *Image and Video Processing*, 2008(1):1–10, 2008.

[10] W. Choi and S. Savarese. Multiple target tracking in world coordinate with single, minimally calibrated camera. *ECCV 2010*.

[11] A. Ellis and J. Ferryman. PETS2010 and PETS2009 evaluation of results using individual ground truthed single views. *AVSS*, 2010.

[12] A. Ess, B. Leibe, K. Schindler, and L. Van Gool. A mobile vision system for robust multi-person tracking. *CVPR 2008*.

[13] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. *The PASCAL VOC 2012 Results*. 2012.

[14] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua. Multicamera people tracking with a probabilistic occupancy map. *IEEE T. Pattern Anal. Mach. Intell.*, 30(2):1806–1819, 2008.

[15] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? The KITTI Vision Benchmark Suite. *CVPR 2012*.

[16] J. Henriques, R. Caseiro, and J. Batista. Globally optimal solution to multi-object tracking with merged measurements. *ICCV 2011*.

[17] E. Horbert, K. Rematas, and B. Leibe. Level-set person segmentation and tracking with multi-region appearance models and top-down shape information. *ICCV 2011*.

[18] Y. Li, C. Huang, and R. Nevatia. Learning to associate: Hybridboosted multi-target tracker for crowded scene. *CVPR 2009*.

[19] D. Mitzel, E. Horbert, A. Ess, and B. Leibe. Multi-person tracking with sparse detection and continuous segmentation. *ECCV 2010*.

[20] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes. Globally-optimal greedy algorithms for tracking a variable number of objects. *CVPR 2011*.

[21] B. Russell, A. Torralba, K. Murphy, and W. Freeman. Labelme: A database and web-based tool for image annotation. *Int. J. Comput. Vision*, 77(1-3):157–173, 2008.

[22] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. *CVPR 2006*.

[23] G. Shu, A. Dehghan, O. Oreifej, E. Hand, and M. Shah. Part-based multiple-person tracking with partial occlusion handling. *CVPR 2012*.

[24] K. Smith, D. Gatica-Perez, J.-M. Odobez, and S. Ba. Evaluating multi-object tracking. *Workshop on Empirical Evaluation Methods in Computer Vision (EEMCV)*, 2005.

[25] R. Stiefelhagen, K. Bernardin, R. Bowers, J. S. Garofolo, D. Mostefa, and P. Soundararajan. The CLEAR 2006 evaluation. *CLEAR*, 2006.

[26] A. Utasi and C. Benedek. A multi-view annotation tool for people detection evaluation. *Workshop on Visual Interfaces for Ground Truth Collection in Computer Vision Applications*, Capri, Italy, 2012.

[27] C. Vondrick, D. Patterson, and D. Ramanan. Efficiently scaling up crowdsourced video annotation. *Int. J. Comput. Vision*, 101(1):184–204, 2013.

[28] S. Walk, N. Majer, K. Schindler, and B. Schiele. New features and insights for pedestrian detection. *CVPR 2010*.

[29] B. Wu and R. Nevatia. Tracking of multiple, partially occluded humans based on static body part detection. *CVPR 2006*.

[30] B. Yang and R. Nevatia. An online learned CRF model for multi-target tracking. *CVPR 2012*.

[31] D. Zoran and Y. Weiss. Scale invariance and noise in natural images. *ICCV 2009*.