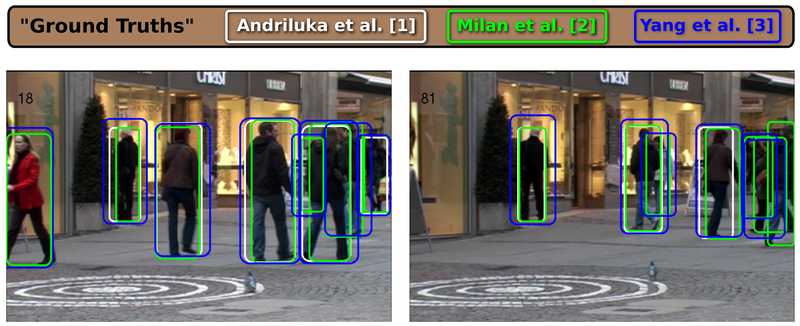
Challenges of Ground Truth Evaluation of Multi-Target Tracking

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

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

TECHNISCHE UNIVERSITÄT DARMSTADT

Status Quo and Overview



Same tracking result [2] – different ground truth

Ground truth	Recall	Precision	GT	MT	ML	ID	FM	MOTA	MOTP
white [1]	90.1	97.1	18	11	4	3	3	87.1	83.3
green [2]	69.3	99.5	10	4	0	7	6	68.3	76.6
blue [3]	72.1	99.1	10	4	0	7	6	70.8	71.9

Quantitative evaluation of multi-target tracking is challenging because:

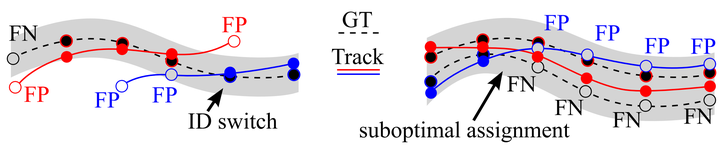
- Multi-target tracking ground truth is not well defined.
- Multiple annotations available for some datasets.
- Multiple (ambiguous) evaluation protocols exist.
- There is no common training/testing dataset.

Metrics

Many sensible metrics possible.

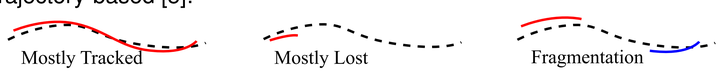
- CLEAR MOT [4]:

MOTA $= 1 - (\#\text{ errors})/(\#\text{ gr. truth obj.})$, MOTP $=$ Avg. alignment precision



Ambiguities: Distance measure, assignment strategy, error weighting, ...

- Trajectory-based [5]:



Further metrics: configuration distance and purity, global mismatch error, ...

Ground Truth

Annotation tools

AnnoTool MultiView VATIC

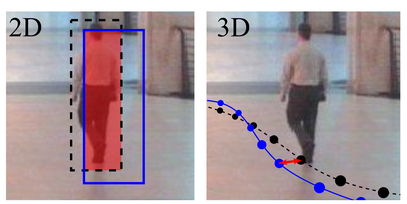
Different level-of-detail **Ambiguity**

How does one ground truth perform with respect to another one?

"Solution"	Ground truth	Recall	Precision	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

→ **One ground truth w.r.t. to another one performs just as well as (or worse than) a state-of-the-art tracker.**

Evaluation Software



The distance between ground truth annotation and tracker output can be computed *e.g.* in 2D as intersection over union of bounding boxes, or in 3D as Euclidean distance on the ground plane.

Same result, same ground truth, different evaluation scripts:

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

The values in parentheses are not part of the script output.

→ **Metrics' definition alone is not enough.**
→ **The same ground truth and evaluation script must be used.**

Parameter Tuning

Tracker	Training	Recall	Precision	ID	FM	MOTA	MOTP
[2]	per sequence	68.6	93.8	49	30	62.8	64.7
	all sequences*	59.1	95.5	29	22	54.9	66.7
	cross validation	60.3	90.9	31	24	49.2	65.2
[8]	per sequence	57.1	95.4	160	124	49.2	66.0
	all sequences*	57.6	92.6	149	123	48.5	65.6
	cross validation	57.1	92.5	144	119	47.7	65.6
[6]	per sequence	64.7	92.4	61	46	58.0	64.5
	all sequences*	60.7	90.7	52	41	52.1	65.4
	cross validation	60.7	90.7	52	41	52.1	65.4

*Most common training procedure.

→ **Tracking performance is very dependent on training data.**
→ **To avoid overfitting, dedicated test data is essential.**

Toward a Benchmark

PETS S2L1 is 'solved'. It is time for a new challenging multi-target tracking benchmark, similar to Middlebury, PASCAL or KITTI.

- **Data:** Variability in camera angle and motion, person count, resolution.
- **Testing / Training:** A clear separation of data (see table above).
- **Detections:** A common set of detections may provide more objective measures of tracking performance.
- **Annotation:** Providing several, independent ground truth annotations may reduce the effect of ambiguities.
- **Evaluation:** One common evaluation metric and script is crucial.
- **"Cheating":** A centralized evaluation server with limited submissions.

References

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

[2] A. Milan, S. Roth, and K. Schindler. Continuous energy minimization for multi-target tracking. *PAMI*. To appear.

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

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

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

[6] A. Milan, S. Roth, and K. Schindler. Detection- and trajectory-level exclusion in multiple object tracking.

[7] 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), 2012.

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