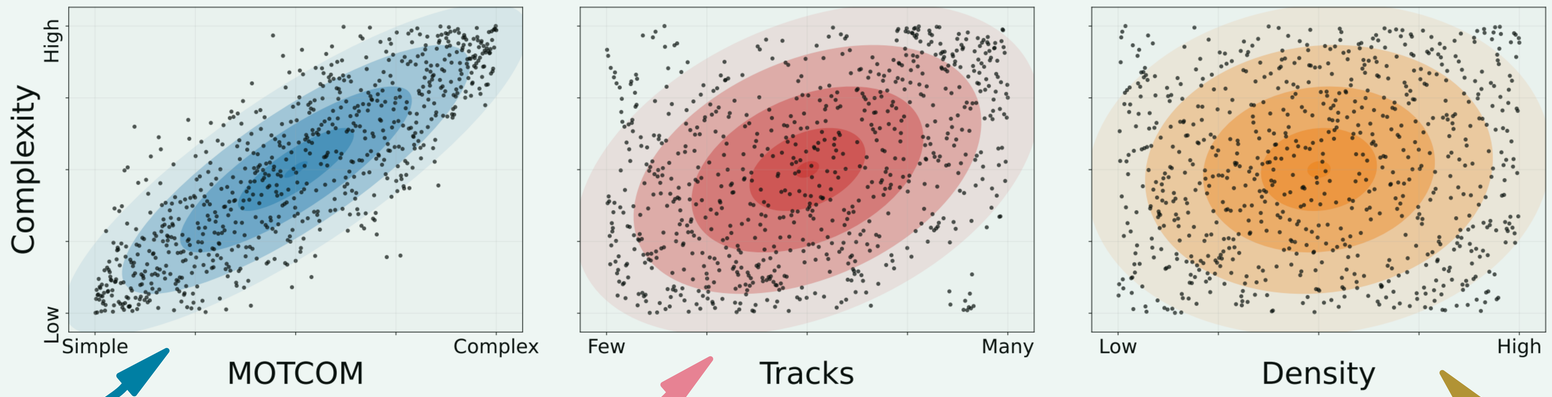
MOTCOM:
The Multi-Object Tracking Dataset Complexity Metric

Malte Pedersen[1], Joakim B. Haurum[1,2], Patrick Dendorfer[3], and Thomas B. Moeslund[1,2]

[1]Visual Analysis and Perception Lab, Aalborg University, Denmark

[2] Pioneer Center for AI, Denmark

[3]Dynamic Vision and Learning Group, Techical University of Munich, Germany



What do we mean by 'complexity'?

We use the term complexity to describe the general difficulty level of MOT sequences.
There exists no ground truth for the complexity of a MOT sequence, so we use a proxy based on the performance of state-of-the-art trackers. There are many performance metrics for evaluating trackers and we use the recent and comprehensive **HOTA**[a] metric.
In this figure we illustrate how **MOTCOM** correlates negatively with the performance of **CenterTrack**[b] when evaluated on the **MOTSynth**[c] sequences.

Conventional Metrics

Number of tracks:
The total number of objects in a sequence.

Density:
The average number of objects per frame.

MOTCOM

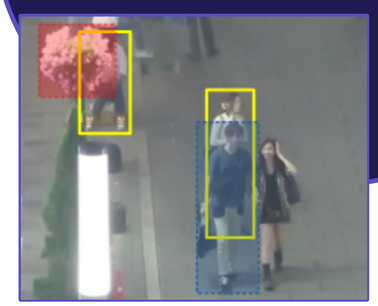
We propose a novel metric for estimating the complexity of MOT sequences. The metric is an average of three sub-metrics based on three factors that complicates MOT sequences:
occlusion, **non-linear motion** and **visual similarity**.

OCOM

The occlusion sub-metric takes *scene-occlusion* and *inter-object-occlusion* into account.

$$\mathrm{OCOM} = \frac{1}{|K|}\sum_{k}^{K}\bar{v}^{k}$$

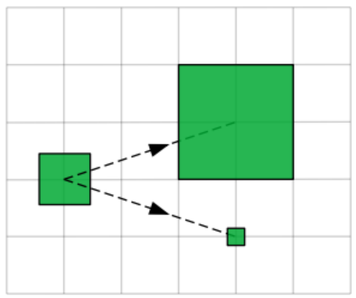
K is the set of objects and $\bar{v}$ is the mean occlusion level.

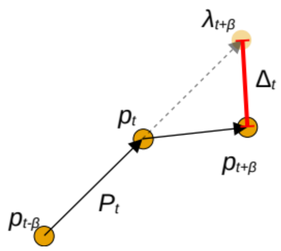


Sample from the MOT17-04 sequence[d].

MCOM

We assume objects move linearly between frames. We use the error Δ_t between the predicted $\lambda_{t+\beta}$ and actual position $p_{t+\beta}$ as the basis for the MCOM score.





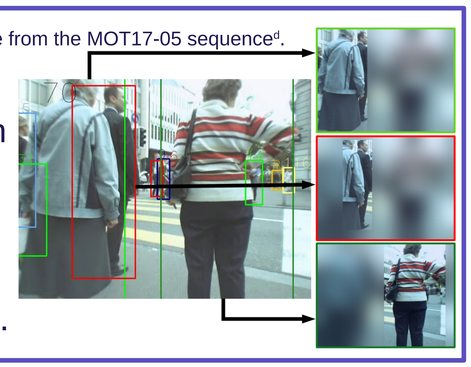
Furthermore, we take the object size and change in size into account (ρ) and uses a Sigmoid function to normalize the output.

$$\mathrm{MCOM} \approx g\left(\frac{1}{\sum |F^k|}\sum_{k}^{K}\sum_{t}^{F^k}\frac{\Delta_t^k}{\rho_t^k}\right)$$

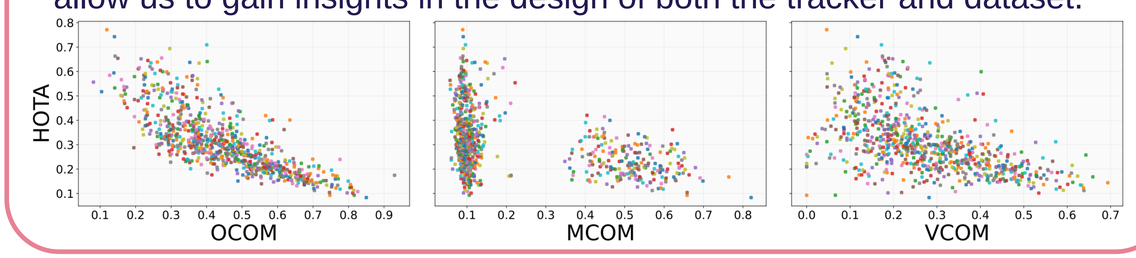
VCOM

We blur the frame, except for the object, and extract features. Then we measure the distance between the target and all the objects in the next frame. More objects in the proximity of the target indicates a harder problem.

Sample from the MOT17-05 sequence[d].



Results from CenterTrack[b] evaluated on MOTSynth[c]. The sub-metrics allow us to gain insights in the design of both the tracker and dataset.



References

[a] Luiten, J., Osep, A., Dendorfer, P., Torr, P., Geiger, A., Leal-Taixé, L. and Leibe, B., 2021. Hota: A higher order metric for evaluating multi-object tracking. International journal of computer vision (IJCV), 129(2), pp.548-578.

[b] Zhou, X., Koltun, V. and Krähenbühl, P., 2020, August. Tracking objects as points. In European Conference on Computer Vision (ECCV) (pp. 474-490). Springer, Cham.

[c] Fabbri, M., Brasó, G., Maugeri, G., Cetintas, O., Gasparini, R., Ošep, A., Calderara, S., Leal-Taixé, L. and Cucchiara, R., 2021. Motsynth: How can synthetic data help pedestrian detection and tracking?. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ECCV) (pp. 10849-10859).

[d] Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K.: Mot16: A benchmark for multi-object tracking. ArXiv (2016) https://doi.org/10.48550/ARXIV.1603.00831