Analysis and Evaluation of Deep Learning Based Approaches for Visual UAV-Tracking

Alexander Fratzer
University of Applied Sciences Darmstadt
Department of Computer Science, Department of Mathematics and Natural Sciences
alexander.fratzer@stud.h-da.de

h_da HOCHSCHULE DARMSTADT UNIVERSITY OF APPLIED SCIENCES

DLR

1. Introduction

- Motivation
  - Misuse of Unmanned Arial Vehicles (UAVs) poses a growing thread for security sensitive infrastructures
  - Tracking UAVs in visual surveillance data is a challenging task for counter measurements
  - Visual Object tracking (VOT) includes Single Object Tracking (SOT) and Multiple Object Tracking (MOT)
  - Deep learning is a promising approach detection-task in visual data
  - Most state-of-the-art MOT-algorithms leverage deep learning
  - Lack of comparable research work in UAV-tracking
  - Desirable to identify deep learning based Multiple Object Tracking (MOT) algorithm with UAV-tracking capability
- Research questions
  - What existing deep learning based algorithms perform best for MOT with UAVs as targets?
  - Which aspects does these models leverage and how promising are they for UAV-tracking?

UAV-Tracking Benchmark

Environment	Frames	Frames without UAV	Frames with UAV	Max UAV per Frame	Different UAVs
Wood	27520	4190	23330	1	2
Harbor	21224	5256	15968	1	1
Complete	48744	9446	39298	1	2

Table 1: Properties of the different environments covered in the dataset.

- Dataset
  - Consists of 2 main video-sequences acquired in woodland- and harbor-environment
  - At most one UAV flies in multiple trajectories across the scene
  - Different conditions for lighting, background structures, movement and scale.
- Benchmark
  - Requirements
    * Comparable results to find the best algorithm
    * Overview about the generalization abilities of the models
    * Indication if to little training-data is used
    * Multiple experiments to draw conclusions about suitability of core technologies
  - Experiments
    * Selection of three models based on defined criteria
    * Dataset split into 20 woodland- and 15 harbor-sequences
    * Rating of sequences based on lighting conditions, movement and background complexity
    * Design of multiple training-test subsets for the experiments based on ratings
    * Five experiments to cover all requirements
      · T_{wood}: test on woodland-environment and train on all other
      · T_{harbor}: test on harbor-environment and train on all other
      · $T_{4,7,14,25,27,28,35}$: test on 3 woodland-sequences and 4 harbor-sequences and train on rest
      · $T_{4,7,14}$: test on 3 woodland-sequences and train on rest
      · $T_{25,27,28,35}$: test on 4 harbor-sequences and train on rest

Figure 2: Excerpt from UAV-dataset in harbor-environment.

Multiple Object Tracking Algorithms

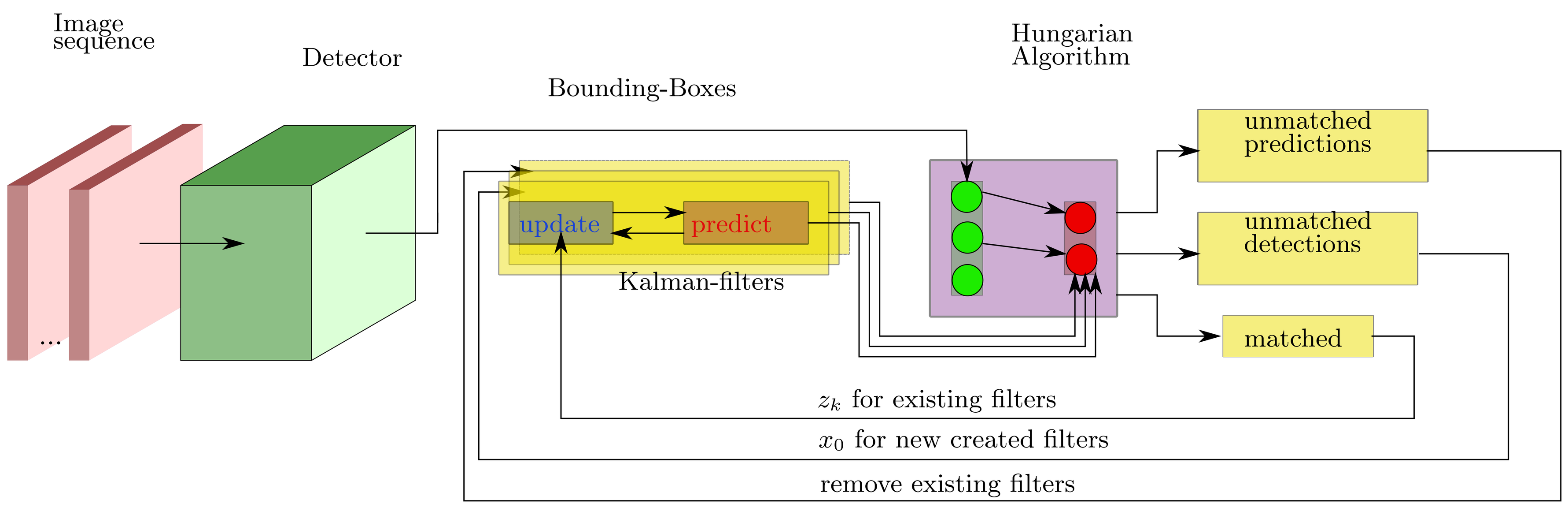

Figure 1: Simple Online Real-Time Tracking (SORT) algorithm.

- MOT-algorithms solve four tasks [1]:
  1. Object-detection: locating all objects in frame k
  2. Feature-extraction: visual- or motion-features or composition [2]
  3. Affinity: creation of distance matrix C_k for extracted features
  4. Association: assignment of detections o_i to tracks t_j based on C_k
- Properties of MOT-algorithms
  - Spatial input dimension: 2-D vs. 3-D
  - Number of targets: SOT vs. MOT
  - Number of representations: single camera vs. multiple cameras
  - Model existence: initial state defined by model vs. model-free
  - Temporal causality: online vs. offline tracking
  - Duration: re-assigned vs. new track
- Performance-metrics for MOT-algorithms
  - ID-metrics [3]

$$P_{id} = \frac{TP_{id}}{TP_{id} + FP_{id}} \quad (1)$$

$$R_{id} = \frac{TP_{id}}{TP_{id} + FN_{id}} \quad (2)$$

$$F_{1_{id}} = \frac{2TP_{id}}{2TP_{id} + FP_{id} + FN_{id}} \quad (3)$$

  - CLEAR-metrics [4]

$$MOTA = 1 - \frac{\sum_k (m_k + FP_k + mme_k)}{\sum_k G_k} \quad (4)$$

$$MOTP = \frac{\sum_{j,k} d_k^j}{\sum_k G_k} \quad (5)$$

  - Classical detection-metrics like precision P_{det} and recall R_{det}
- $F_{1_{id}}$ as primary metric for tracking-performance

SORT [5]

- Object-detection with deep learning
- Kalman-filters predict location of previous tracked objects
- Each Kalman-filter models a track
- Kalman-filter assume linear transformation of object locations
- Intercept of Union (IoU) as distance measure between new detections and predictions.
- Hungarian Algorithm assigns tracks to new detections

FairMOT [6]

- Same principle as SORT except that deep learning component additionally extracts visual features
- Assumes also that the same object looks similar in all frames
- Assignment first based on distances between visual features then on motion predictions from Kalman-filters
- Deep learning component performs center-point object-detection [7] with Deep Layer Aggregation (DLA) architecture

CenterTrack [8]

- Single Convolutional Neuronal Network (CNN) which takes current frame I_K, previous frame I_{k-1} and previous tracks T_{k-1} as input
- CNN performs also center-point object-detection and offset-prediction between objects from I_{k-1} and current frame I_k
- Greedy algorithm assigns tracks based on lowest offsets between adjacent frames
- No assumptions regarding motion and appearances of UAVs

Results

Figure 3: Results of T_{wood}.

- Low scores from all algorithms
- Too little training-data
- SORT
  - Best detection-performance based on P_{det} and R_{det}
  - Best tracking-performance based on $F_{1_{id}}$
  - Generalization-ability starts to manifest
- FairMOT
  - Higher P_{det} than SORT but lower R_{det}
  - Similar tracking-performance as SORT
  - Generalization-ability starts to manifest
- CenterTrack
  - Poor detection-scores
  - Poor tracking-performance
  - No generalization-ability

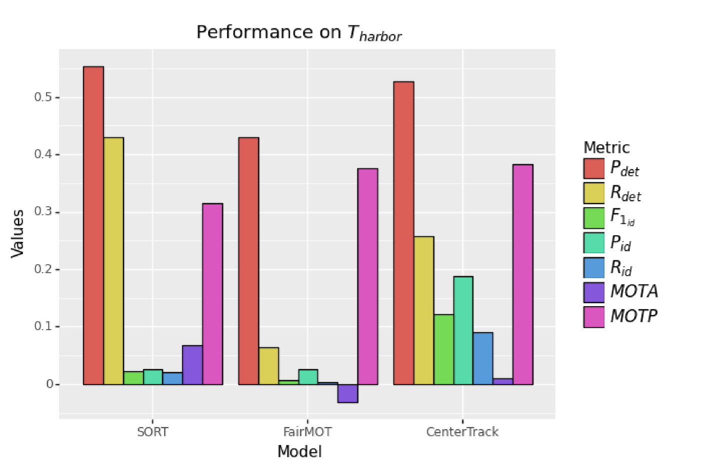

Figure 4: Results of T_{harbor}.

- Low scores from all algorithms
- Too little training-data
- SORT
  - Best detection-performance
  - Only marginally better tracking--performance as in T_{wood}
  - Generalization-ability starts to manifest
- FairMOT
  - Poorest detection-performance
  - Poorest tracking-performance
  - No generalization-ability
- CenterTrack
  - Medium detection-performance
  - Highest tracking-performance
  - Best improvement regarding T_{wood}
  - Generalization-ability starts to manifest

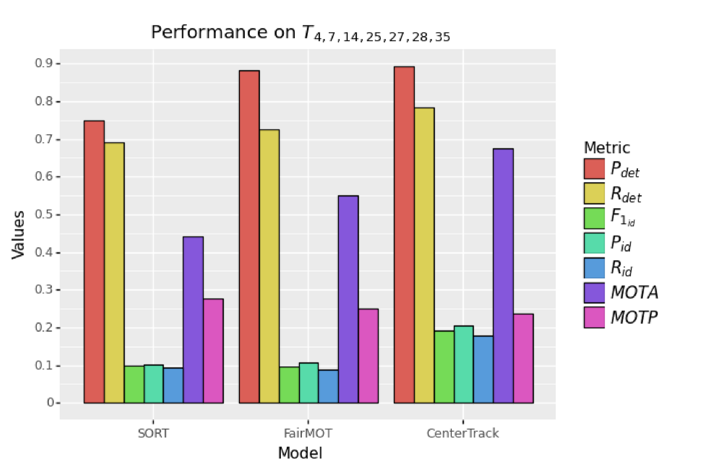

Figure 5: Results of $T_{4,7,14,25,27,28,35}$.

- Much better scores as in T_{harbor} and T_{wood}
- SORT
  - Lowest detection-performance but much higher than in previous experiments
  - Low tracking-performance in contrast to best algorithm
- FairMOT
  - Good detection-performance
  - Similar low tracking-performance as SORT
- CenterTrack
  - Highest detection-performance
  - Highest tracking-performance
  - High improvement
  - Best model in all categories

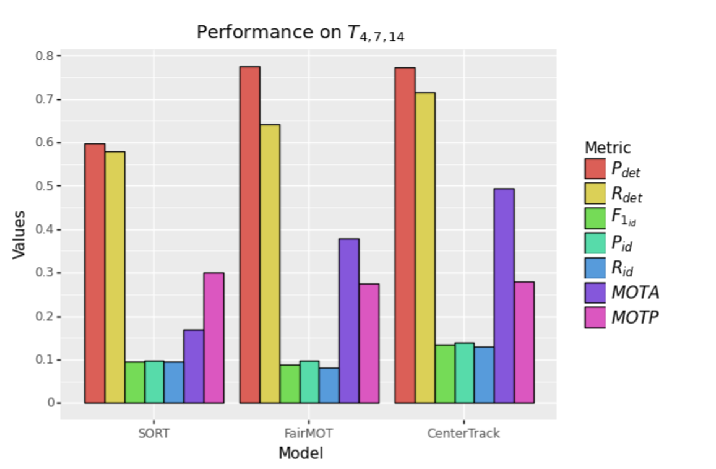

Figure 6: Results of $T_{4,7,14}$.

- More training data available as in $T_{4,7,14,25,27,28,35}$
- Test-data only from woodland-environment
- Results viewed in contrast to $T_{4,7,14,25,27,28,35}$
- SORT
  - Lower detection-performance
  - Similar tracking-performance
- FairMOT
  - Lower detection-performance
  - Similar tracking-performance
- CenterTrack
  - Lower detection-performance
  - Lower tracking-performance

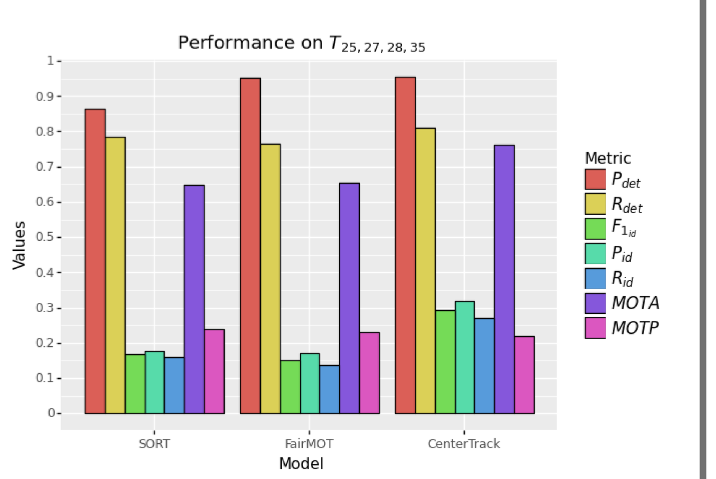

Figure 7: Results of $T_{25,27,28,35}$.

- More training data available as in $T_{4,7,14,25,27,28,35}$
- Test-data only from woodland-environment
- Enough training data available
- Results viewed in contrast to $T_{4,7,14,25,27,28,35}$
- SORT
  - Higher detection-performance
  - Similar tracking-performance
- FairMOT
  - Higher detection-performance
  - Similar tracking-performance
- CenterTrack
  - Higher detection-performance
  - Higher tracking-performance

Discussion

- Main findings
  - CenterTrack performs best on the available data and is considered the baseline technology
  - All three algorithms are good and reliable in the detection-task, if trained on enough data like in $T_{4,7,14,25,27,28,35}$
  - The ability of SORT and FairMOT to assign stable tracks through multiple frames is poor
  - Motion prediction of SORT and FairMOT does not work satisfyingly when faced with fast camera- and UAV-movements
  - The assignment based on appearance-features does not bring any improvement regarding the tracking-performance
  - Good performing models for the task of pedestrian-tracking use assumptions like slow and linear moving objects which are not transferable
- Limitations
  - Model selection excluded Matlab implementations
  - Orientation on MOT-Challenge [9] unpromising
  - Camera movement not considered for model selection
  - Dataset does not reflect MOT because only a single UAV present in each frame

References

1. G. Ciaparrone et al., Neurocomputing **381**, 61–88, ISSN: 18728286, (https://doi.org/10.1016/j.neucom.2019.11.023) (2020).
2. S. Sun et al., IEEE Transactions on Pattern Analysis and Machine Intelligence **13**, 1–1, ISSN: 0162-8828, arXiv: 1810.11780 (2019).
3. E. Ristani et al., presented at the Computer Vision – ECCV 2016 Workshops, ed. by G. Hua et al., vol. 9914 LNCS, pp. 17–35, ISBN: 978-3-319-48881-3, arXiv: **arXiv:1609.01775v2**.
4. K. Bernardin et al., Eurasip Journal on Image and Video Processing **2008**, ISSN: 16875176 (2008).
5. A. Bewley et al., presented at the Proceedings - International Conference on Image Processing, ICIP, vol. 2016-Augus, pp. 3464–3468, ISBN: 9781467399616, arXiv: **1602.00763**.
6. Y. Zhang et al., arXiv: 2004.01888 (2020).
7. X. Zhou et al., arXiv: 1904.07850 (2019).
8. X. Zhou et al., arXiv: 2004.01177 (2020).
9. A. Milan et al., 1–12, arXiv: 1603.00831 (2016).