Towards the evaluation of reproducible robustness in tracking-by-detection

Francesco Solera Simone Calderara Rita Cucchiara
Department of Engineering Enzo Ferrari
University of Modena and Reggio Emilia
name.surname@unimore.it

Abstract

Conventional experiments on MTT are built upon the belief that fixing the detections to different trackers is sufficient to obtain a fair comparison. In this work we argue how the true behavior of a tracker is exposed when evaluated by varying the input detections rather than by fixing them. We propose a systematic and reproducible protocol and a MATLAB toolbox for generating synthetic data starting from ground truth detections, a proper set of metrics to understand and compare trackers peculiarities and respective visualization solutions.

1. Introduction

After a decade of studies on Multiple Target Tracking (MTT) the assessment of reproducible experimental protocols and stable benchmarks for trackers evaluation is still controversial. Most of the MTT approaches, especially when used in surveillance and crowd analysis, adopt the *tracking-by-detection* scheme due to the strong improvements in people detection witnessed over recent years [2]. Nevertheless, detector performances are not stable across different sequences, ranging even among the well known PETS dataset from 0.75 to 0.91 for precision and from 0.52 to 0.88 for recall [15]. Being the detection set the initial tracker input, performances cannot be detached from the detection quality. This consideration is guiding most of the recent research on MTT with many open issues about the metrics and the sets of detections that should be used for evaluation. An example is the recent MOT Challenge website [11], a crowdsourced repository that collects the most famous benchmark sequences and provides a unified set of detections for comparison purpose. This is a good starting point.

However, in this paper the purpose is to expose the limits of current test protocols and propose a novel experimental procedure to improve the understanding of trackers behavior. The first contribution of the paper is a reproducible procedure to generate synthetic detection data starting from the ground truth annotation. This enables a controlled simulation of different degrees of detectors reliability in terms of precision/recall and occlusions. Furthermore we propose a proper set of metrics to evaluate a tracker on this potentially vast amount of detections sets. Accordingly, specific colored grids and plots are also suggested to help visualization and ease comparison between different trackers. Eventually, all these contributions resulted in a MATLAB toolbox made available to the community for future MTT analysis and evaluation.

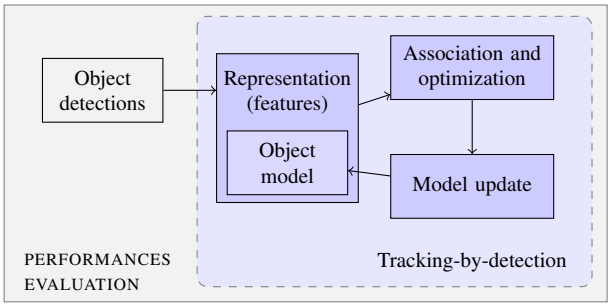
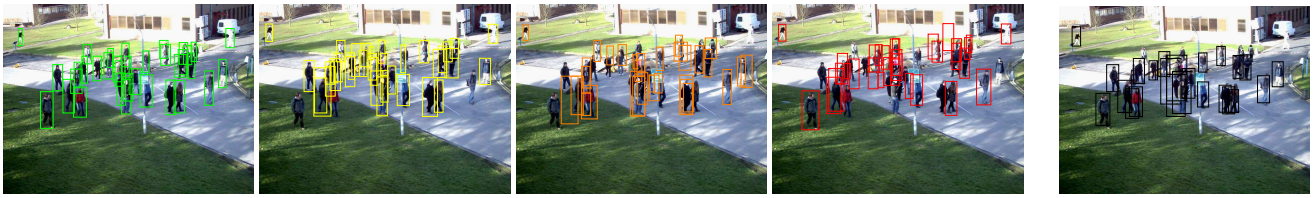


Figure 1: Tracking-by-detection overview scheme. Tracking evaluation cannot be decoupled from detections.

2. Problem statement

The basic scheme of tracking-by-detection, is depicted in Fig. 1. Modern trackers differ from each other by the way they pre-process trajectories (or post-process detections), extract features and perform the association step. Pre-processing usually involves the definition of thresholds on the detector confidence, on the returned bounding boxes size or on the position of the objects inside the image, [5, 9]. To cope with the uncertainties in the detections, complex appearance models [1] or trajectory dynamics [3, 6] are often employed in the representation of the targets. Data association is then tackled globally [9, 3, 6], hierarchically [8, 5] or through online optimization methods [1, 13].



(a) P and R = 1 (b) P and R = 0.8 (c) P and R = 0.6 (d) P and R = 0.4 (e) MOT Challenge

Figure 2: Synthetic detections at different levels of precision and recall. (e) are the detection currently used in the public MOT Challenge benchmark.

Conventional experiments on MTT are built upon the belief that fixing the detections is sufficient to obtain a fair comparison among different trackers. However, each set of detections may exhibit many joint challenges ranging from undetected occluded targets to different levels of false positives and false negatives. These challenges all derive from the scene peculiar issues: clutter, confusion, luminance variation, camera points of view, targets shape variations and more - all of them must be taken into account in the evaluation of both single target tracking [14] and, even more, multiple target tracking. Given an unique set of detections, the evaluation of a tracker doesn't explicitly describe its robustness to any of these challenges as it is not possible to assess how much the results are due to the tracker quality rather than the detections. Thus, a first motivation to make available a parametric evaluation framework is to prevent a tracker to overfit on a specific dataset, denying the possibility to reproduce the declared robustness in even simpler scenarios. Moreover we aim at following the positive performances trend of object detection algorithms: as detectors are increasingly enhancing their ability to precisely localize objects [2], also the quality of available detections for tracking is better than the one provided by decade old benchmarks. This work assumes that a deeper characterization of the tracker peculiarities can be obtained by testing the trackers on a set of controlled disruptions of ground truth detections. By deviating the detection from the ground truth exposing one challenge at a time, it is possible to precisely describe the tracker behavior and its robustness by observing the expected performance degradation.
The proposal of this work does not controvert the standard experiments, generally input with people detector responses: these data are still considered for performance evaluation, being only another possible instance of the disrupted input set used in the proposed test protocol. Consequently, this paper adds another and more complete view of tracking evaluation.

3. Data degradation

In this section, we propose the protocol for generating artificial detections by varying the purity of the ground truth. In general, detections can diverge from ground truth due to:

(a) detection errors, mainly false positives (nonexistent detected objects) and false negatives (miss detection);

(b) occlusions, which can be considered as a continuous sequence of miss detection on the same track;

(c) other localization errors such as size or position of the bounding box.

However, here (c) is jointly modeled within type (a) errors.

3.1. Detection errors

The first proposed experimental setup consists in different variations of the ground truth detections, controlling typical detector errors in terms of false positives (FP) and false negatives (FN). The aim is to understand how a tracker is sensible to different values of detector precision and recall. We create different detections sets $D(P, R)$ at precision $P \in [0.5, 1]$ and recall $R \in [0.5, 1]$; results outside this range are rarely interesting. Each $D(P, R)$ contains d different instances of detections to account for the randomness in the modification of the ground truth.

Protocol Consider GT the full set of ground truth detections for the considered sequence and let TP $=$ GT $-$ FN be the number of true positives. From the definition of precision and recall, the number of false positives FP and false negatives FN that should be added to GT to obtain a pair of (P, R) values follows this relation:

$$FN = GT \cdot (1 - R) \quad (1)$$

$$FP = GT \cdot R \cdot (1 - P)/P \quad (2)$$

The procedure to obtain the proposed set of detections can be summarized as:

- Add FP detections at locations $(\bar{x}, \bar{y}) \sim \mathcal{N}((x, y), \sigma_1)$ close to random GT detections (x, y). By sampling every new location from a gaussian distribution centered on a GT point, the sparsity of artificial detections is reduced and the data mimics the usual behavior of people detector in crowded scenes.

- Alter the FP detections bounding box size by scaling the associate ground truth detections in the interval $[\frac{1}{2}, \frac{3}{2}]$ with uniform probability.
- Remove FN randomly selected detections from GT.
- Resize TP detections by sampling the new bounding boxes from $\mathcal{N}((w,h),\sigma_2)$ with (w,h) the original size and keeping fixed its center coordinates.

Moreover, since false positives are added close to true objects before their possible removal, we also account for localization errors through the parameter σ_1, which we suggest to fix at 4. The final resize renders the difficulty of a detector to precisely identify the targets size. We fix $\sigma_2 = 2$ and the number of instances $d = 5$.
From the visual point of view in Fig. 2, the current protocol produces at the same level of precision and recall, a set of detection fairly indistinguishable from the one obtained by the most popular recent people detector [7].

3.2. Occlusions

A tracker considers continuous miss detections, occlusions, as a different problem from punctual false negative and false positive detections. Handling occlusions becomes more difficult as their duration increases. Dealing with prolonged occlusion is directly related to the ability of a tracker to re-identify targets. In this setup, $N \in [0,1]$ is the percentage of occluded targets and $L \in [0,1]$ the ratio between the occluded points and the track length. The final set S is obtained by generating d instances for each pair (N, L).

Protocol Start from the GT tracks, to obtain a set of detections for (N, L):

- Randomly choose $(\text{GT} \cdot N)$ tracks to affect. Selected tracks must be at least $n \geq \tau$ frames long.
- For each selected trajectory, remove the points between the occlusion starting frame f_{out} and the reappearance frame f_{in}, defined as

$$f_{\text{out}} \sim \mathcal{U}([0, (1-L)\cdot n]) \quad (3)$$
$$f_{\text{in}} = f_{\text{out}} + L \cdot n \quad (4)$$

  where $\mathcal{U}$ is the uniform distribution and n is the track length.

During this experiment we set $\tau = 10$ to consider only significant occlusions and $d = 5$.

4. Evaluation metrics

In last five years, the evaluation metrics for MTT have converged to the CLEAR MOT [4]. Despite its fame, these metrics are of complex use as many values have to be considered altogether to establish tracking quality [12].

MOTA Among the CLEAR MOT metrics, MOT accuracy (MOTA) is perhaps the most complete; it comprises false positives, false negative and identity switches (IDS) errors over all frames t in a unique measure:

$$\text{MOTA} = 1 - \frac{\sum_t (\text{FP}(t) + \text{FN}(t) + \text{ID}(t))}{\sum_t \text{GT}(t)} \quad (5)$$

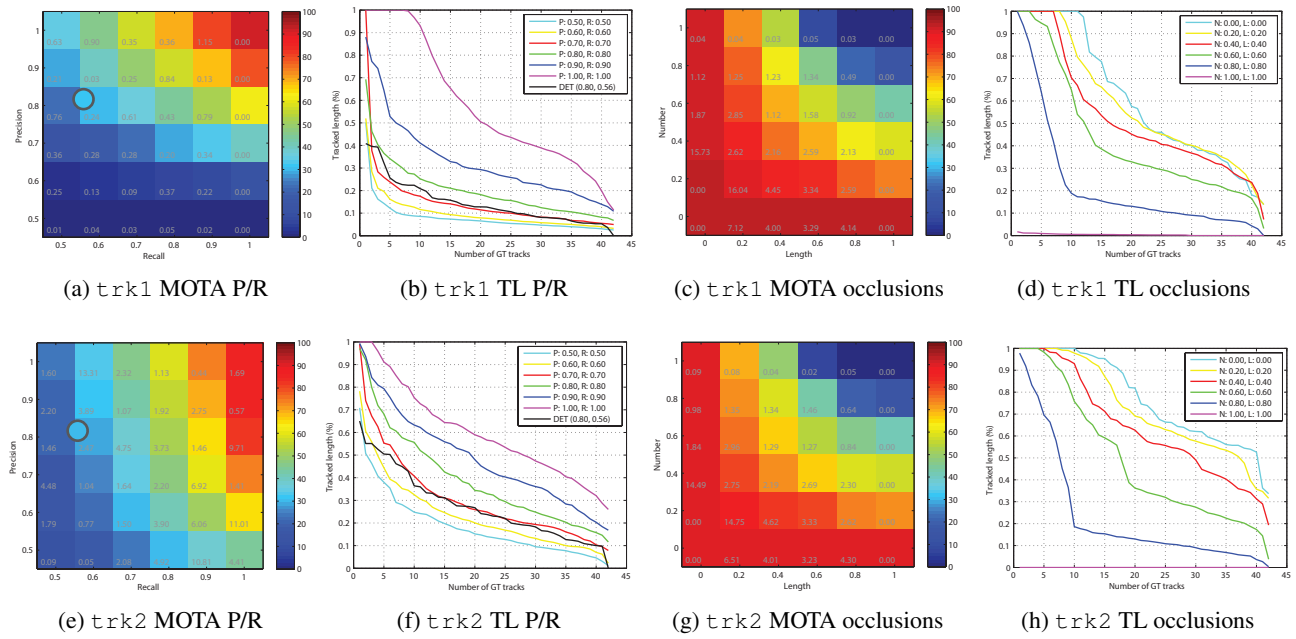
where GT is the number of ground truth trajectories. This measure gives an idea of how much a tracker is able to find targets, reject false alarms proposed by the detector and keep the correct identities throughout the sequence. In the remainder of this paper only MOTA will be considered, but all the proposed evaluation and visualization protocols can be employed for the other CLEAR MOT metrics as well.

Tracks length Other than CLEAR MOT, trajectory-based metrics are jointly used in literature [10]. These metrics include the number of mostly tracked (MT) and mostly lost (ML) trajectories, as well as the number of times a ground truth trajectory was re-associated to different constructed tracks (FRAG). A ground truth trajectory is said to be mostly tracked (lost) if at least the 80% (less than 20%) of its points were associated during tracking. MT and ML are not influenced by the number of FRAG or IDS. As a result, these metrics give more information about the coverage of the trajectories rather than the ability of the tracker to reproduce them. They become informative only when considered strictly together with the number of IDS and FRAG, which makes them difficult to interpret for large number of targets as in MTT case.

To this end we propose to compute, for each ground truth trajectory, the percentage of its overall length correctly tracked. We call this measure track length (TL). When an object path is completely reproduced by the tracker, then its TL value is 1.

5. Performances visualization

The proposed protocol, based on a grid of parameters, produces a set of detections for each grid element and requires a proper solution for a quick visualization and interpretation of the results. For clarity, we restrict the analysis on two different tracking-by-detection algorithms, `trk1` and `trk2`. These methods are very different. The former, from more than a decade ago, is based on the hungarian algorithm for global detections association using bounding box coordinates; the latter is very recent and proposed in a 2014 top computer vision conference [1]. How far did MTT methods advance? The answer is: it depends on the detection quality. For demonstration purposes, we set the experiment on the well known PETS09-S2L2 sequence that is present in all the MTT benchmarks. The sequence is widely



(a) `trk1` MOTA P/R (b) `trk1` TL P/R (c) `trk1` MOTA occlusions (d) `trk1` TL occlusions

(e) `trk2` MOTA P/R (f) `trk2` TL P/R (g) `trk2` MOTA occlusions (h) `trk2` TL occlusions

Figure 3: MOTA matrices and TL curves are shown for `trk1` and `trk2` when varying precision/recall and number/length of occlusions as in Sec 3. The filled circle in (a) and (e) represent the MOTA value obtain by the trackers when using real world detections, position is set according to precision and recall of the detector. In (b) and (f) a black curve represent the tracks length starting from the same real world detections.

adopted but extremely difficult and exposes the typical challenges of most of the MTT sequences used for benchmarking, *e.g.* self occlusion, a high number of targets, adversarial motion flows and a poor visual representation of the targets themselves. Under this premises, conclusions can be generalized easily on other similar sequences as well[1].

5.1. MOTA matrices

Given the sets of detections from D, for each value of precision and recall we average the MOTA scores over the d experiments. Results can be visualized in a matrix format, where the cells color is based on the MOTA value. In every cell the standard deviation is reported, giving an idea of the reproducibility of the results, Fig 3a and 3e. The number of cells depends on the selected precision and recall steps. The same visualization can also be performed for the occlusions data S, Fig. 3c and 3g.

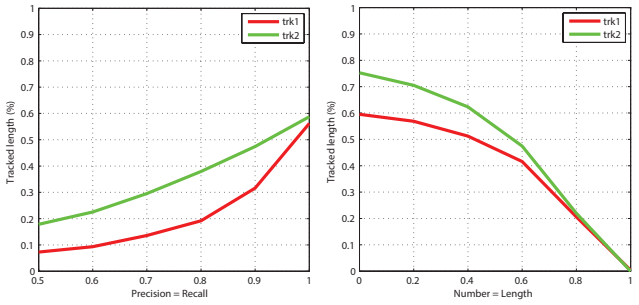
5.2. TL curves

A TL curve is a survival curve where tracks expectation to survive is the longest sequence of frames in which they were correctly and continuously tracked. The representation derives from the Kaplan-Meier estimator employed in human mortality analysis, but was already applied for single target tracking in Smeulders *et al.* [14]. Starting from TL values from Sec. 4, tracks length need to be sorted in descending order. In the TL plot, given a specific point (x, y), it is directly observable how many tracks x were correctly and continuously tracked for at least the y of their total length. The first set of proposed curves studies the robustness of the tracker w.r.t. detector precision and recall variations. A TL curve could be constructed for each pair of precision and recall values, anyway to summarize the results we suggest to plot only the results corresponding to a simultaneous increase in both values of P and R. Sorted results were averaged over the d different set of experiments. Results are shown in Fig. 3b and 3f. In Fig. 3d and 3h results are computed with the detections corrupted by occlusions from experimental setup S.

5.3. TL areas

In Fig. 3 each survival plot is composed of a set of per-tracker curves corresponding to increasing values of the control parameters. This visualization is helpful to understand the tracker robustness at different complexity levels, but it is not suitable to compare different methods. To this end, we introduce the AUC TL plot derived from the area under the survival curves. Through this plot, each tracker is represented as a curve of its performances degradation and its robustness is visually comparable with other trackers as well. Given a survival curve, the greater the AUC is the better trajectories were correctly and continuously

[1] `website omitted` - additional results on other sequences.



(a) AUC TL P/R (b) AUC TL occlusions

Figure 4: Through the area under the TL curves at different level of GT degradation, trackers can be easily compared.

tracked. The AUC values do not distinguish between few well tracked trajectories and many poorly tracked trajectories, but it gives an overall perception of the tracker ability. Fig. 4 shows `trk1` and `trk2` comparison in terms of precision and recall and occlusions handling.

6. It's like a finger pointing to the moon

One of the advantages of the proposed experimental protocol is that it is not incompatible with the standard evaluation scheme that uses a single set of fixed detections. In particular, we show how results obtained through the use of input from a real detector can be thought of as a specific instance inside our framework. To this end, we employ real detections provided by MOT Challenge benchmark, for which complete results in terms of CLEAR MOT are reported in Tab 1. The precision and recall values of this set of detections are 0.80 and 0.56, respectively. In Fig. 3a and 3d a small circle indicates the performance for these trackers obtained with the MOT Challenge detections. The color similarity between the filled circle and the surrounding MOTA matrix cells highlights that we were able to artificially produce a realistic augmented experimental set. Moreover, the same behavior is met when measuring TL on the MOT Challenge detections, as shown in Fig. 3b and 3e through the black curve. The TL results obtained on the these detections matches the expectations emerging from the experiments following the proposed protocol, being bounded by surrounding curves at close P and R values. Given a benchmark detection set, it is not possible to decouple precision and recall errors from occlusions errors. Nevertheless, the compatibility of the obtained results, using our protocol, suggests that in this scenario the detector was much more prone to frequent and discontinued errors rather than occlusions. This brief comparison between results, both in terms of MOTA and TL, underlines how little we know of a tracker when we evaluate it only over a single set of benchmark detections. It's a mere point in an arbitrarily detailed MOTA matrix and it's only a curve in the possibly many of TL plots.

	MOTA	MOTP	MT	ML	IDS	FRAG
`trk1`	**32.6**	**69.1**	**4**	3	981	876
`trk2`	31.8	67.8	1	3	**261**	**505**

Table 1: CLEAR MOT and standard trajectory-based evaluation of `trk1` and `trk2` on the real world detections provided in the MOT Challenge for PETS09-S2L2. MT and ML must be normalized w.r.t. the number of GT tracks (43).

By looking at the CLEAR MOT results for `trk1` and `trk2` on PETS09-S2L2 in Tab. 1, it is not a trivial task to firmly describe the difference in the behaviors of the two methods. Tracker `trk1` presents a slightly higher value of MOTA and a lower value of MOTP, while strong differences can be found in MT and, above all, in the number of IDS and FRAG. Different people could rate these methods differently depending on which measures are valued the most.

On the opposite, by looking at our results in Fig. 3 and 4, there is no doubt about which tracker is the best one. In particular, `trk1` is highly sensible to precision drops (darker blue on bottom rows of Fig. 3a) while `trk2` is more robust to precision variations and modestly robust to lower values of recall as well (leftmost column in Fig. 3e). This makes `trk2` a more reliable tracker when used in conjunction with modern detectors, as appreciable from the better spaced curves in Fig. 3f. Nevertheless, `trk1` seems to behave better when approaching ideal values of P and R (darker red on top right corner and nearby cells in Fig. 3a) and is intended to become a better tracker than `trk2` in a few years, supporting our claim that in tracking-by-detection best trackers really are a matter of detections quality.

The experiments on occlusions provide some new understanding of the considered trackers that could not be otherwise distilled from Tab. 1. As already anticipated in Sec. 6, it is not possible decouple precision/recall errors from occlusions in real detections. Typically, careful visual inspection is needed in order to assess a tracker robustenss to occlusions. Conversely, the TL curves related to occlusions from Fig. 3 and its compact visualization in Fig. 4b, quantitatively reveal how `trk2` is more robust to occlusions than `trk1`. Eventually, when the intensity of the occlusions become too heavy, both trackers fail at achieving good results.

7. Tools and guidelines for future experiments

To help researchers in the subtle and important part of evaluating future tracking-by-detection methods, we publicly provide a MATLAB toolbox[2]. This toolbox is composed of three main components:

[2] `link omitted for db review` - see additional material

- **Data degradation (Sec. 3):** this module is required to generate new detections from ground truth. It should only be employed for training, while generated data should be kept fixed for future comparison.
- **Evaluation (Sec. 4):** this code partially extends the DEVKIT proposed at MOT Challenge[3] with the ability to measure TL and automatically process a whole set of detections at different pairs of control parameter values.
- **Result visualization (Sec. 5):** is needed to reproduce the exact same plots we reported in Fig. 3 and 4.

Moreover, we describe in the remainder of this section how to evaluate a tracker on a whole dataset, not just one sequence, by following our protocol. It is important to avoid averaging results over different sequences or many important aspects of a tracker could be hidden by the strong contribution of easier scenes. To this end, MOTA and TL should be computed only at the end of the whole evaluation:

$$\text{MOTA}_{\text{all}} = 1 - \frac{\sum_v \sum_t (\text{FP}_v(t) + \text{FN}_v(t) + \text{ID}_v(t))}{\sum_v \sum_t \text{GT}_v(t)} \quad (6)$$

where the index v accounts for all the different video sequences. Similarly, TL from many sequences can be appended and sorted altogether, resulting in a more comprehensive view of the tracking results. The code provided with the toolbox is already capable of these evaluations.

8. Conclusion

With this paper we are not criticizing the standard tracking evaluation protocol. Our purpose is to open a deep discussion on what it is important for tracker to be robust against. Moreover we argue that controlled experiments, that are often employed in different scientific fields, could be profitably applied for MTT as well. The proposed protocol has the advantage of be controllable, reproducible and tailored for evaluating different tracking peculiarities. In our opinion, it gives important insights on the robustness of the tracker, independently from the chosen detector. This aspect is a double advantage for research in MTT allowing the community to better present the individual strengths of new solutions, and consumers to choose the proper method according to the requirements of their tracking system.

References

[1] S.-H. Bae and K.-J. Yoon. Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1218–1225, 2014. 1, 3

[2] R. Benenson, M. Omran, J. Hosang, and B. Schiele. Ten years of pedestrian detection, what have we learned? In *European Conference on Computer Vision Workshops*, pages 613–627. 2015. 1, 2

[3] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua. Multiple object tracking using k-shortest paths optimization. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 33(9):1806–1819, Sept. 2011. 1

[4] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: the CLEAR MOT metrics. *EURASIP Journal on Image and Video Processing*, 2008. 3

[5] M. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool. Robust tracking-by-detection using a detector confidence particle filter. In *IEEE International Conference on Computer Vision*, pages 1515–1522, 2009. 1

[6] C. Dicle, O. Camps, and M. Sznaier. The way they move: Tracking multiple targets with similar appearance. In *IEEE International Conference on Computer Vision*, pages 2304–2311, 2013. 1

[7] P. Dollar, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 36(8):1532–1545, 2014. 3

[8] C. Huang, Y. Li, and R. Nevatia. Multiple Target Tracking by Learning-Based Hierarchical Association of Detection Responses. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35(4):898–910, 2013. 1

[9] A. Milan, S. Roth, and K. Schindler. Continuous energy minimization for multitarget tracking. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 36(1):58–72, 2014. 1

[10] A. Milan, K. Schindler, and S. Roth. Challenges of ground truth evaluation of multi-target tracking. In *IEEE Computer Vision and Pattern Recognition Workshops*, pages 735–742, 2013. 3

[11] L. Milan, A. Leal-Taix, K. Schindler, S. Roth, and I. Reid. MOT Challenge. http://www.motchallenge.net, 2014. 1

[12] T. Nawaz, F. Poiesi, and A. Cavallaro. Measures of effective video tracking. *IEEE Trans. on Image Processing*, 23(1):376–388, 2014. 3

[13] H. Possegger, T. Mauthner, P. M. Roth, and H. Bischof. Occlusion geodesics for online multi-object tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1306–1313, 2014. 1

[14] A. Smeulders, D. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah. Visual tracking: An experimental survey. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 36(7):1442–1468, 2014. 2, 4

[15] J. Zhang, L. Presti, and S. Sclaroff. Online multi-person tracking by tracker hierarchy. In *IEEE Advanced Video and Signal-Based Surveillance*, pages 379–385, 2012. 1

[3] http://motchallenge.net/data/devkit.zip