A Benchmark and Simulator for UAV Tracking

Matthias Mueller, Neil Smith, and Bernard Ghanem

King Abdullah University of Science and Technology (KAUST), Saudi Arabia
{matthias.mueller.2, neil.smith, bernard.ghanem}@kaust.edu.sa

Abstract. In this paper, we propose a new aerial video dataset and benchmark for low altitude UAV target tracking, as well as, a photo-realistic UAV simulator that can be coupled with tracking methods. Our benchmark provides the first evaluation of many state-of-the-art and popular trackers on 123 new and fully annotated HD video sequences captured from a low-altitude aerial perspective. Among the compared trackers, we determine which ones are the most suitable for UAV tracking both in terms of tracking accuracy and run-time. The simulator can be used to evaluate tracking algorithms in real-time scenarios before they are deployed on a UAV "in the field", as well as, generate synthetic but photo-realistic tracking datasets with automatic ground truth annotations to easily extend existing real-world datasets. Both the benchmark and simulator are made publicly available to the vision community on our website to further research in the area of object tracking from UAVs.[1]

Keywords: UAV tracking, UAV Simulator, Aerial Object Tracking

1 Introduction

Visual tracking remains a challenging problem despite several decades of progress on this important topic. A broadly adopted evaluation paradigm for visual tracking algorithms is to test them on established video benchmarks such as OTB50 [43], OTB100 [42], VOT2014, VOT2015, TC128 (Temple Color) [26], and ALOV300++ [39]. Since the performance of a tracker is measured against these benchmarks, it is critical that a holistic set of real-world scenarios and a distribution of tracking nuisances (e.g. fast motion, illumination changes, scale changes, occlusion, etc.) are properly represented in the annotated dataset. The benchmark also plays a critical role in identifying future research directions in the field and how to design more robust algorithms. What is currently lacking in these well established benchmarks is a comprehensive set of annotated aerial datasets that pose many challenges introduced by unmanned airborne flight.

Empowering unmanned aerial vehicles (UAVs) with automated computer vision capabilities (e.g. tracking, object/activity recognition, etc.) is becoming a very important research direction in the field and is rapidly accelerating with the increasing availability of low-cost, commercially available UAVs. In fact, aerial tracking has enabled many new applications in computer vision (beyond

[1] https://ivul.kaust.edu.sa/Pages/pub-benchmark-simulator-uav.aspx

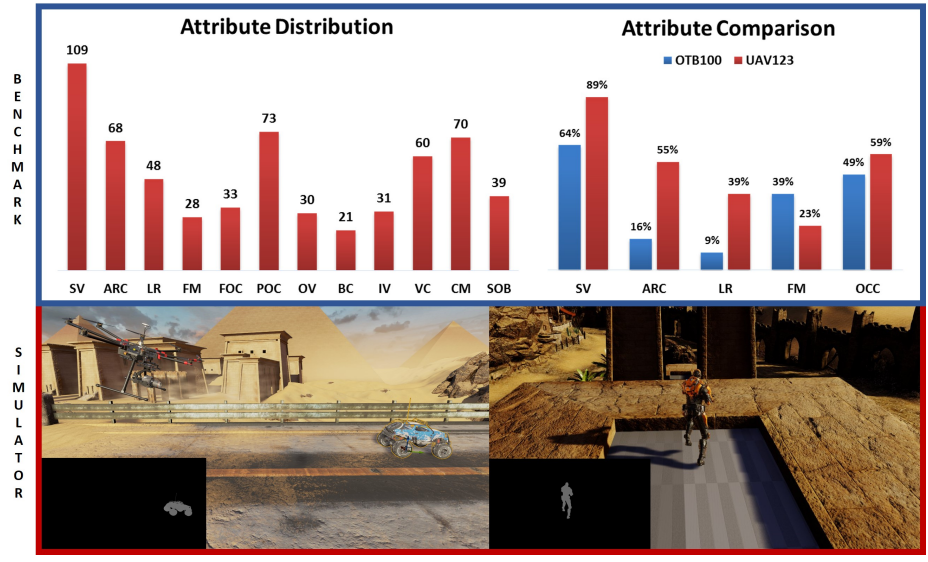

Fig. 1. *Top:* Attribute distribution across UAV123 dataset and a comparison of key attributes with OTB100. *Bottom:* Synthetic dataset generation and online tracker evaluation using the proposed simulator. For a legend of abbreviations, refer to Table 2.

those related to surveillance) including search and rescue, wild-life monitoring, crowd monitoring/management, navigation/localization, obstacle/object avoidance, and videography of extreme sports. Aerial tracking can be applied to a diverse set of objects (e.g. humans, animals, cars, boats, etc.), many of which cannot be physically or persistently tracked from the ground. In particular, real-world aerial tracking scenarios pose new challenges to the tracking problem (see Fig. 1), exposing areas for further research. This paper provides an evaluation of trackers on more than 100 new fully annotated HD videos captured from a professional grade UAV. This benchmark both complements current benchmarks establishing the aerial component of tracking and provides a more comprehensive sampling of tracking nuisances that are ubiquitous in low-altitude UAV videos. To the best of our knowledge, this is the first benchmark to address and analyze the performance of state-of-the-art trackers on a comprehensive set of annotated aerial sequences that exhibit specific tracking nuisances. We anticipate that this dataset and its tracker evaluation will provide a baseline that can be used long into the future as UAV technology advances and target trackers improve.

Visual tracking on UAVs is a very promising application, since the camera can follow the target based on visual feedback and *actively* change its orientation and position to optimize for tracking performance. This marks the defining difference compared to static tracking systems, which passively analyze a dynamic scene. Since current benchmarks are pre-recorded scenes, they cannot provide a quantifiable measure on how slower trackers would affect the performance of the UAV in *shadowing* the target. In this paper, we propose the use of a photo-

realistic simulator to render real-world environments and a variety of life-like moving targets typically found in unmanned aerial recordings. The simulator uses the Unreal Engine 4 to directly feed image frames to trackers and retrieve tracking results to update UAV flight. Any tracker (e.g. written in Matlab or C++) can be tested on the simulator across a diverse set of photo-realistic simulated scenarios. Using this simulator enables the use of new quantitative methods for evaluating tracker performance in the aforementioned aerial feedback loop.

Contributions. The contributions of our work are threefold. **(1)** We compile a fully annotated high-resolution dataset of 123 aerial video sequences comprising more than 110K frames. It is as large or larger than most recent, generic object tracking datasets. **(2)** We provide an extensive evaluation of many state-of-the-art trackers using multiple metrics [43]. By labeling the videos in the benchmark with various attributes, we can also evaluate each tracker in regards to specific aerial tracking nuisances (e.g. scale/aspect ratio change, camera motion, etc.). **(3)** We provide a novel approach to perform tracker evaluation by developing a high-fidelity real-time visual tracking simulator. We present first results on the performance of state-of-the-art trackers running within its environment. The combination of the simulator with an extensive aerial benchmark provides a more comprehensive evaluation toolbox for modern state-of-the-art trackers and opens new avenues for experimentation and analysis.

Related Work

UAV Datasets. A review of related work indicates that there is still a limited availability of annotated datasets specific to UAVs in which trackers can be rigorously evaluated for precision and robustness in airborne scenarios. Existing annotated video datasets include very few aerial sequences [43]. Surveillance datasets such as PETS or CAVIAR focus on static surveillance and are outdated. VIVID [6] is the only publicly available dedicated aerial dataset, but it is outdated and has many limitations due to its small size (9 sequences), very similar and low-resolution sequences (only vehicles as targets), sparse annotation (only every 10th frame), and focus on higher altitude, less dynamic fixed-wing UAVs. There are several recent benchmarks that were created to address specific deficiencies of older benchmarks and introduce new evaluation approaches [24,25,40], but they do not introduce videos with many tracking nuisances addressed in this paper and common to aerial scenarios.

Generic Object Tracking. In our proposed benchmark, we evaluate classical trackers such as OAB [11] and IVT [38] as baselines and the best-performing recent trackers according to [43]: Struck [13], CSK [17], ASLA [19], and TLD [21]. In the selection process, we reject very slow trackers despite their performance [3, 4, 45–48]. In addition, we include several of the latest trackers such as MEEM [44], MUSTER [18], DSST [8] (winner VOT2014) and SRDCF [7] (winner VOT-TIR2015 and OpenCV challenge). Since current benchmarks provide no more than 1 or 2 real-world scenarios of video capture from a mobile aerial

platform, it is unclear which of these new trackers would perform well in aerial scenarios where certain tracking challenges are amplified, including abrupt camera motion, significant changes in scale and aspect ratio, fast moving objects, as well as, partial and full occlusion.

UAV Tailored Tracking. Despite the lack of benchmarks that adequately address aerial tracking, the development of tracking algorithms for UAVs has become very popular in recent years. The majority of object tracking methods employed on UAVs rely on feature point detection/tracking [30,37] or color-centric object tracking [22]. Only a few works in the literature [33] exploit more accurate trackers that commonly appear in generic tracking benchmarks such as MIL [1,9], TLD [33], and STRUCK [27,28]. There are also more specialized trackers tailored to address specific problems and unique camera systems such as in wide aerial video [34,36], thermal and IR video [10,35], and RGB-D video [29].

UAV Simulation. In recent years, several UAV simulators have been created to test hardware in the loop (HIL). However, the focus is on simulating the physics of the UAV in order to train pilots or improve/tune features of a flight controller (e.g. JMAVSim [41]). The visual rendering in these simulators is often primitive and relies on off-the-shelf simulators (e.g. Realflight, Flightgear, or XPlane). They do not support advanced shading and post-processing techniques, are limited in terms of available assets and textures, and do not support MOCAP or key-frame type animation to simulate natural movement of actors or vehicles. Although simulation is popularly used in machine learning [2] and animation and motion planning [12,20], the use of synthetically generated video or simulation for tracker evaluation is a new field to explore. In computer vision, synthetic video is primarily used for training recognition systems (e.g. pedestrians [14], 3D scenes [31], and 2D/3D objects [15,32]), where a high demand for annotated data exists. The Unreal Engine 4 (UE4) has recently become fully open-source and it seems very promising for simulated visual tracking due in part to its high-quality rendering engine and realistic physics library.

2 Benchmark - Offline Evaluation

2.1 Dataset

Statistics. Video captured from low-altitude UAVs is inherently different from video in popular tracking datasets like OTB50 [43], OTB100 [42], VOT2014, VOT2015, TC128 [26], and ALOV300++ [39]. Therefore, we propose a new dataset (called UAV123) with sequences from an aerial viewpoint, a subset of which is meant for long-term aerial tracking (UAV20L). In Fig. 2, we emphasize the differences between OTB100, TC128, and UAV123. The results highlight the effect of camera viewpoint change arising from UAV motion. The variation in bounding box size and aspect ratio with respect to the initial frame is significantly larger in UAV123. Furthermore, being mounted on the UAV, the camera is able to move with the target resulting in longer tracking sequences on average.

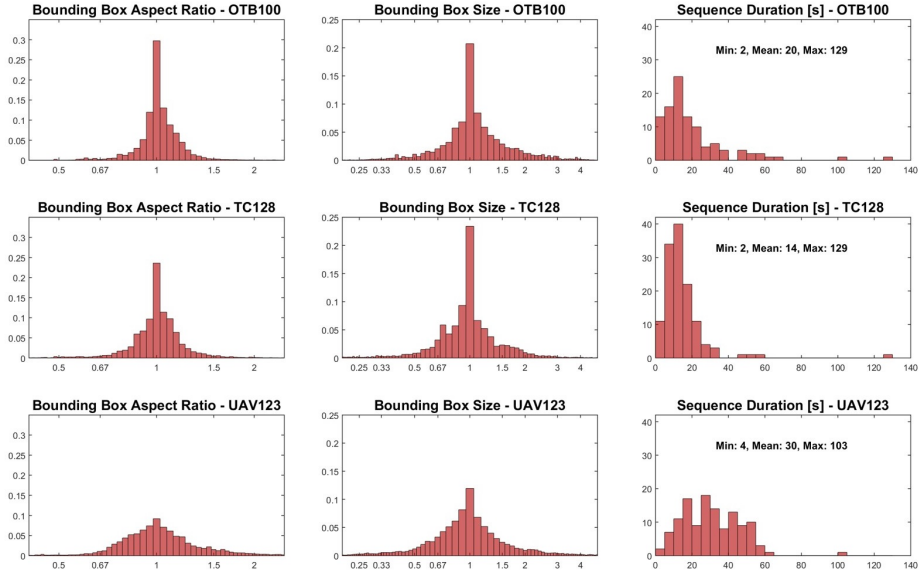

Fig. 2. *Column 1 and 2:* Proportional change of the target's aspect ratio and bounding box size (area in pixels) with respect to the first frame and across three datasets: OTB100, TC128, and UAV123 (ours). Results are compiled over all sequences in each dataset as a histogram with log scale on the x-axis. *Column 3:* Histogram of sequence duration (in seconds) across the three datasets.

Our new UAV123 dataset contains a total of 123 video sequences and more than 110K frames making it the second largest object tracking dataset after ALOV300++. The statistics of our dataset are compared to existing datasets in Table 1. Note that OTB50 is a subset of both OTB100 and TC128, so the total number of unique frames contained in all three datasets combined is only around 90K. The datasets VOT2014 and VOT2015 are both subsets of existing datasets too. Hence, while there is a number of datasets available to the tracking community, the number of distinct sequences is smaller than expected and sequences specific to tracking from a UAV vantage point are very sparse.

Table 1. Comparison of tracking datasets in the literature. Ranking: R(1),G(2),B(3)

Dataset	UAV123	UAV20L	VIVID	OTB50	OTB100	TC128	VOT14	VOT15	ALOV300
Sequences	123	20	9	51	100	129	25	60	314
Min frames	109	1717	1301	71	71	71	171	48	19
Mean frames	915	2934	1808	578	590	429	416	365	483
Max frames	3085	5527	2571	3872	3872	3872	1217	1507	5975
Total frames	112578	58670	16274	29491	59040	55346	10389	21871	151657

Fig. 3. First frame of selected sequences from UAV123 dataset. The red bounding box indicates the ground truth annotation.

Acquisition. The UAV123 dataset can be divided into 3 subsets. **(i)** Set1 contains 103 sequences captured using an off-the-shelf professional-grade UAV (DJI S1000) following different objects at altitudes varying between 5-25 meters. Video sequences were recorded at frame rates between 30 and 96 FPS and resolutions between 720p and 4K using a Panasonic GH4 with Olympus M.Zuiko 12mm f2.0 lens mounted on a fully stabilized and controllable gimbal system (DJI Zenmuse Z15). All sequences are provided at 720p and 30 FPS and annotated with upright bounding boxes at 30 FPS. The annotation was done manually at 10 FPS and then linearly interpolated to 30 FPS. **(ii)** Set2 contains 12 sequences captured from a boardcam (with no image stabilization) mounted to a small low-cost UAV following other UAVs. These sequences are of lower quality and resolution and contain a reasonable amount of noise due to limited video transmission bandwidth. The sequences are annotated in the same manner as in Set1. **(iii)** Set3 contains 8 synthetic sequences captured by our proposed UAV simulator. Targets move along predetermined trajectories in different worlds rendered with the Unreal4 Game Engine from the perspective of a flying UAV. Annotation is automatic at 30fps and a full object mask/segmentation is also available.

Attributes. As illustrated in Fig. 3, UAV123 contains a wide variety of scenes (e.g. urban landscape, roads, buildings, fields, beaches and a harbor/marina), targets (e.g. cars, trucks, boats, persons, groups, and aerial vehicles), and activities (e.g.

Table 2. Attributes used to characterize each sequence from a tracking perspective.

Attr	Description
ARC	Aspect Ratio Change: the fraction of ground truth aspect ratio in the first frame and at least one subsequent frame is outside the range [0.5, 2].
BC	Background Clutter: the background near the target has similar appearance as the target.
CM	Camera Motion: abrupt motion of the camera.
FM	Fast Motion: motion of the ground truth bounding box is larger than 20 pixels between two consecutive frames.
FOC	Full Occlusion: the target is fully occluded.
IV	Illumination Variation: the illumination of the target changes significantly.
LR	Low Resolution: at least one ground truth bounding box has less than 400 pixels.
OV	Out-of-View: some portion of the target leaves the view.
POC	Partial Occlusion: the target is partially occluded.
SOB	Similar Object: there are objects of similar shape or same type near the target.
SV	Scale Variation: the ratio of initial and at least one subsequent bounding box is outside the range [0.5, 2].
VC	Viewpoint Change: viewpoint affects target appearance significantly.

walking, cycling, wakeboarding, driving, swimming, and flying). Naturally, these sequences contain common visual tracking challenges including long-term full and partial occlusion, scale variation, illumination variation, viewpoint change, background clutter, camera motion, etc. Table 2 shows an overview of all tracking attributes present in UAV123. Fig. 1 shows the distribution of these attributes over the whole dataset and a comparison to the very popular OTB100 dataset for a selection of key attributes.

Long-term tracking. Object tracking in an aerial surveillance setting usually requires long-term tracking, since the camera can follow the target in contrast to the static surveillance scenario. During the dataset design, some fully annotated long sequences captured in one continuous shot were split into subsequences to ensure that the difficulty of the dataset remains reasonable. For long-term tracking, we merge these subsequences and then pick the 20 longest sequences among them. Table 1 shows the statistics of the resulting dataset (UAV20L).

2.2 Evaluated Algorithms

We consider tracking algorithms for comparison on our benchmark according to their performance in OTB50 [43] and give preference to popular and reasonably fast trackers. Code for these trackers is either available online or from the authors. All selected trackers incorporate some form of model update and are discriminative, except for IVT and ASLA which use generative models. For fair evaluation, we run all trackers with standard parameters on the same server-grade workstation (Intel Xenon X5675 3.07GHz, 48GB RAM).

2.3 Evaluation Methodology

Following the evaluation strategy of OTB50 [43], all trackers are compared using two measures: precision and success. Precision is measured as the distance between the centers of a tracker bounding box (bb_tr) and the corresponding ground truth bounding box (bb_gt). The precision plot shows the percentage of tracker bounding boxes within a given threshold distance in pixels of the ground truth. To rank the trackers, we use the conventional threshold of 20 pixels [43]. Success is measured as the intersection over union of pixels in box bb_tr and those in bb_gt. The success plot shows the percentage of tracker bounding boxes whose overlap score is larger than a given threshold. Moreover, we rank trackers using the area under the curve (AUC) measure [43]. Besides one-pass evaluation (OPE), we perform a spatial robustness evaluation (SRE) [43]. For SRE, the initial bounding box is spatially shifted by 4 center shifts, 4 corner shifts and scaled by 80, 90, 110 and 120 percent, as done in [43].

3 Simulator - Online Evaluation

3.1 Setup and Limitations

The UE4 based simulator allows real-time tracker evaluation with the ability to simulate the physics of aerial flight, produce realistic high-fidelity renderings (similar to if not better than professional rendering software, e.g. 3DSMax and Maya), and automatically generate precise ground truth annotation for offline or real-time use cases (see Fig. 1). The UAV is modeled after the DJI S1000+, which was used to capture the majority of the benchmark. An accurate 3D model (same geometry/weight and thrust vectors) is subjected to game physics (UE4) and real-world conditions (e.g. wind and gravity). The ground truth trajectory and orientation of the target and UAV are recorded at every frame. The PID controllers for stabilization and visual servoing (gimbal) mimic the Pixhawk FC. For further details on the implementation, see the simulator documentation.

UE4 allows for a large variety of post-processing rendering steps to create realistic and challenging scene images that simulate real-world UAV data. Although not implemented for this work, motion blur, depth of field, over/under exposure, HDR and many more features can be enabled. UE4 post-processing rendering allows assignment of custom depth maps to any mesh in the engine. The depth maps allows extraction of segmented annotation of the tracked target as seen through the camera viewpoint. We simulate the movement of both a human character and a 4WD vehicle moving along set trajectories within a detailed off-road race track with palm trees, cacti, mountains, historical buildings, lakes, and sand dunes (see Fig. 3). This is one example of many photo-realistic UE4 worlds created by the developer community in which our UAV simulator can be used. The UAV simulator enables the integration of any tracker (MATLAB or C++) into the tracking-navigation loop; at every frame, the output bounding box of the tracker is read and used to correct the position of the UAV.

3.2 Novel Approaches for Evaluation

Our UE4 based simulator provides new possibilities for online performance measurement (see Fig. 4). Advantages include a controlled environment for isolation of specific tracking attributes, a higher degree of repeatability with rapid experiments, and generation of large annotated datasets for testing and learning. Unlike real-world scenarios where the UAV and target location are imprecisely known (e.g. error of 5-10m), it quantitatively compares position, orientation, and velocity of the UAV at each time-step to understand the impact of the tracker on flight dynamics. For evaluation, we develop several new approaches to measure tracker performance: (1) the impact of a dynamic frame rate (trackers are fed frames at the rate of computation), (2) trajectory error between target and UAV motion, (3) accumulative distance between ground truth and tracker, and (4) long-term tracking within a controlled environment where attribute influence can be controlled and clearly measured.

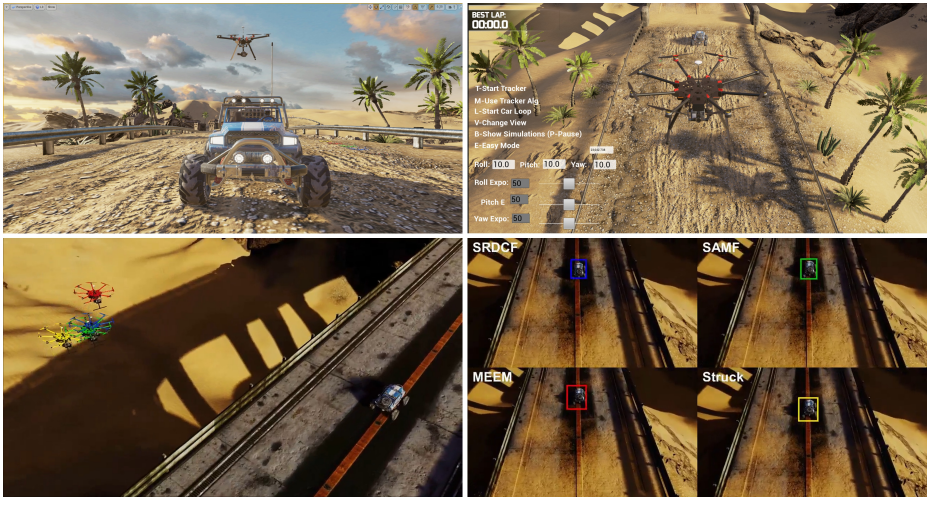

Fig. 4. *Top:* Third person view of simulator environment. *Bottom:* Four UAVs are controlled by different trackers indicated by the different colors.

3.3 Evaluation Methodology

Four trackers are selected for evaluation, namely SRDCF, MEEM, SAMF, and STRUCK. The ground truth bounding box generated from the custom depth map of the target is called GT. We first optimize the UAV visual servoing using the GT tracker (see supplementary material on our visual servoing technique). Despite absolute accuracy of the GT, the flight mechanics of the UAV limit its ability to always keep the target centered, since it must compensate for gravity,

air resistance, and inertia. After evaluating the performance of the UAV with the GT, each tracker is run multiple times within the simulator provided with the same starting initialization bounding box. The target follows a pre-defined path and speed profile. The UAV tracks and follows the target for 3.5 minutes (ca. 6000 frames at 30 FPS). The target speed varies but is limited to 6m/s, the UAV speed is limited to 12m/s (similar to the real UAV). For evaluation, we measure the distance between the trajectory of the target and the UAV.

4 Experiments

4.1 Benchmark Evaluation

Overall Performance. To determine the overall performance of the different trackers on the new challenges in the UAV123 dataset, we use the evaluation paradigm proposed in [43], as outlined in Section 2.3. In the one-pass evaluation (OPE), each tracker processes over 110K frames from all 123 sequences, each with a variety of attributes as shown in Table 2.

The top performing tracker on the UAV123 dataset in terms of precision and success is SRDCF [7]. This is primarily due to its high fidelity scale adaptation that is evident across every success plot. Although MEEM [44] is the top performing tracker in precision on OTB100, it cannot keep up in our dataset, primarily due to the fact that it does not have scale adaptation. SAMF [23], MUSTER [18], DSST [8], Struck [13], and ASLA [19] group into a second tier of close performing trackers, while the remaining trackers IVT [38], TLD [21], MOSSE [5], CSK [17], OAB [11], KCF [16] and DCF [16] achieve consistently lower performance. In general, with the exception of MEEM, the top five performers in terms of success exploit scale adaptation. However, since they are only adapting to scale and not aspect ratio, there is still much room for improvement. In general, the recently developed correlation based trackers perform very well in the OPE and rank in the top five in terms of precision (SRDCF, SAMF, MUSTER, DSST) and success (SRDCF, SAMF, MUSTER). Owing to their manipulation of circulant structure in the Fourier domain, these trackers require low computational cost, making them attractive for onboard UAV tracking.

In comparison with OTB100, all trackers perform much worse in OPE on the more challenging UAV123 dataset and several trackers change rankings (notably MEEM to SRDCF and MUSTER to SAMF). The difference in performance between the top trackers in OTB100 is marginal suggesting that this benchmark is getting closer to saturation. To obtain a global view of overall performance on both datasets, we plot the success results of all trackers per video in Fig. 5 as a color gradient map, where red corresponds to 0 and dark green to 1. The score of the best performing tracker per video is shown in the last row and the average across all videos per tracker is shown in the last column. In OTB100, most videos have at least one tracker that performs well; however, there exist many sequences in UAV123 where none of the trackers are successful. For example, all these trackers perform poorly on low resolution videos of one UAV tracking another, an important aerial tracking scenario.

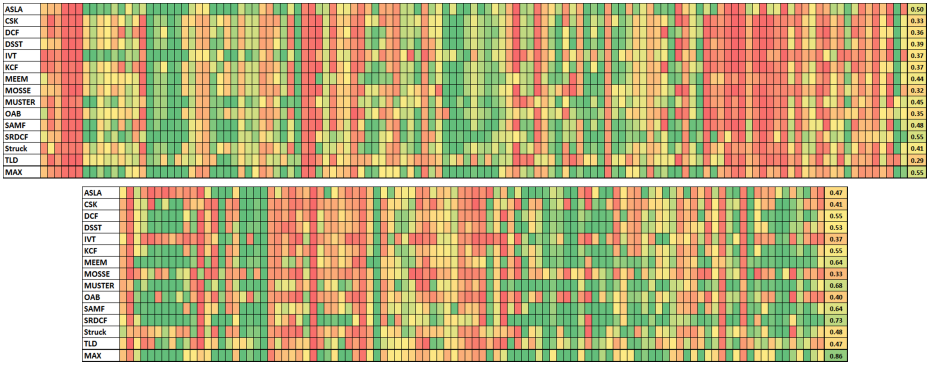

Fig. 5. *Top:* OPE success per video on UAV123. Bottom: OPE success per video for OTB100.

Speed Performance. In Fig. 6, most of the top performing trackers have a frame rate lower than 10 FPS and as low as 0.9 FPS (MUSTER). Note that each tracker predicts a bounding box for each frame regardless of their actual speed. Of course, this is very different when tracking is required in real-time (e.g. when tracker output is needed for persistent UAV navigation). If frames are not processed fast enough, intermediate frames are dropped resulting in larger target displacement between frames, thus, making tracking more difficult. Therefore, if the tracker has a low frame rate, its tracking performance in real-time applications is expected to degrade. In order to investigate the impact of speed on performance, we compare all trackers on the same UAV123 dataset but now temporally downsampled to 10 FPS (refer to Fig. 6). The degradation in performance ranges from 21%-36% for ASLA, DSST, and SAMF, and 11%-15% for SRDCF, STRUCK, and MUSTER. MEEM becomes the top-performing tracker in this case, although its performance degradation (7%) is still noticeable.

Long-Term Tracking. In order to evaluate a tracker's performance in long-term tracking scenarios, we evaluate their performance on UAV20L (see Section 2.1). Tracking results in Fig. 6 show that all trackers perform much worse on UAV20L than on UAV123, indicating that long-term tracking remains a difficult challenge with much room for improvement. In long-term tracking cases, tracker drift is more likely to cause complete loss of the object, especially in occlusion scenarios, where the model update for the target is contaminated by the occluder. The top performer on this dataset is MUSTER, due to its short-term/long-term memory strategy that can correct past tracking mistakes.

Discussion. Throughout the evaluation, trackers perform consistently across attributes; however, we find that trackers struggle more with attributes common to aerial tracking. The most difficult attributes seem to be scale variation and aspect ratio changes but also to a lesser extent low resolution, background clutter, fast motion, and full occlusion. Scale variation is the most dominant attribute

in the aerial tracking dataset, so trackers that incorporate scale adaptation are typically the top performers. There is still much room for improvement especially for attributes common in our dataset, but not very common in current datasets. Moreover, for automated tracking to be integrated on a UAV, tracking speeds must be higher, ultimately reaching real-time speeds of 30 FPS. We also

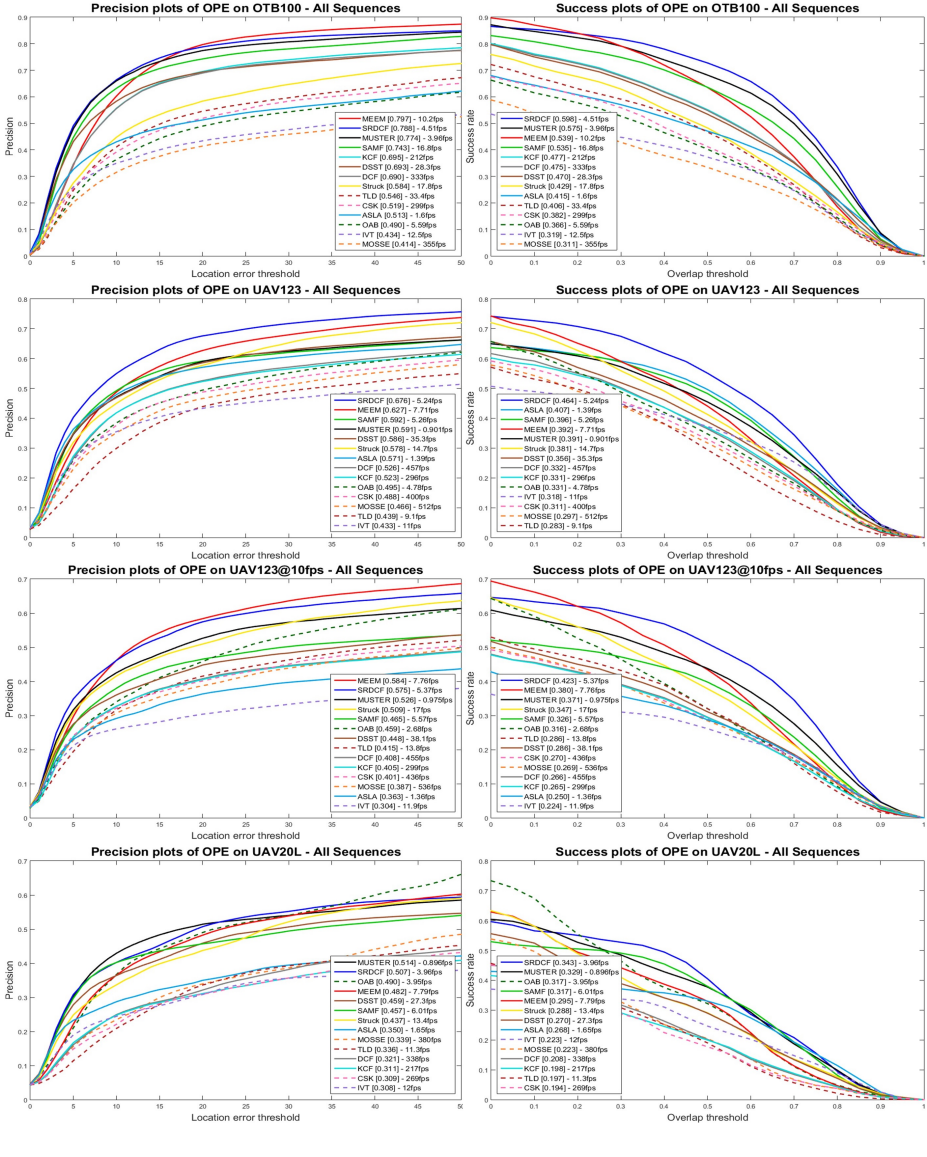

Fig. 6. From top to bottom: Precision and success plots for OPE on OTB100, UAV123, UAV123@10fps and UAV20L.

observe that trackers, which have a robust update method that can help correct past mistakes (MEEM, MUSTER) or suppress background (SRDCF), perform better than those that do not. The spatial robustness evaluation which measures robustness to noise in the initialization is consistent with the OPE plots and trackers rank similarly with overall lower scores. For a detailed evaluation and discussion of all trackers for each prevalent attribute and spatial robustness, please refer to the supplementary material.

4.2 Simulator Evaluation (Quantitative and Qualitative Results)

Overall Performance. Several challenges such as significant change in scale, aspect ratio and viewpoint, illumination variation, and fast motion occur throughout the test course. Despite noticeable drift, all trackers maintain tracking at least throughout half of the course. At this point, the vehicle takes a sharp turn and accelerates down a hill; the conservative default PID setting limits the UAVs' response and most of the trackers fail (see frame 3000 in Fig. 7). However, when the PID controller is set to be more responsive, the tracking results vary significantly. SRDCF already fails at the very beginning of the course, since it is not able to handle the rapid acceleration of the object and overshoots due to the latency introduced by the tracker. The other trackers welcome the more responsive PID setting and follow the target with much more ease than before. This shows that the PID controller and tracker complement each other.

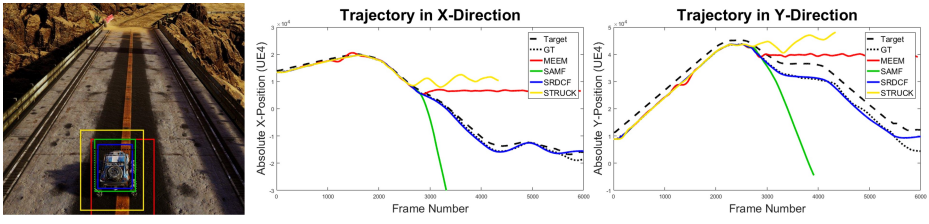

Fig. 7. Trajectory of tracker controlled UAV over the period of the simulation and multiple trackers bounding boxes layered over the tracked synthetic frame.

Speed Performance. The tested trackers vary in computational time with STRUCK and MEEM being the fastest. The bounding boxes of slower trackers (SCRDF and SAMF) have noticeable lag and do not remain consistently centered on the target, especially during rapid acceleration. The UAV altitude, wide vertical FOV, and PID setting can compensate for some latency, allowing the UAV to sync its speed to the vehicle. As altitude increases between the UAV and the target, the precision of the trackers improves. This is an important observation. In real-world scenarios, increasing altitude can be a UAV strategy to enhance tracking performance of slower trackers attempting to follow fast targets.

Long-Term Tracking. At some point, all of the trackers start to drift and usually become locked onto highly salient features of the target. Despite inaccurate bounding boxes, all trackers succeed to follow the target for more than one minute. Only SRDCF completes the course, but it only tracks a portion of the vehicle towards the end.

Discussion. Several insights can be obtained from the live tracking results within the simulator. Despite latency, trackers remain locked on the target throughout a large portion of the course. At higher altitudes latency has less impact on performance, since the UAV has more time to respond to target movement. Tracker performance is noticeably impacted by the flight dynamics and control system of the UAV. The failure of several trackers can be overcome by a more agile UAV. SRDCF's robustness and the UAV's ability to compensate for its latency make it the only tracker to complete the entire course. A major achievement however, is that all the tested state-of-the-art trackers autonomously move the UAV across a complex course. Over longer periods, the predicted center and size of the target drift primarily due to poor adaptation to scale and aspect ratio. Appearance change and partial occlusion lead to loss of the target by all trackers. The benchmark helps identify which trackers are most suitable for aerial tracking and the simulator provides insights for the best integration on a UAV. It provides many avenues to rapidly test trackers and clearly delineate their shortcomings and advantages in real-world scenarios.

5 Conclusions and Future Work

In this paper, we provide extensive empirical evidence of the shortcomings of current datasets for aerial tracking and propose a new benchmark with fully annotated sequences from the perspective of a UAV. The new dataset is similar in size to the largest available datasets for generic object tracking and the benchmark evaluates 14 state-of-the-art trackers. Extensive experiments suggest that sequences with certain tracking attributes (namely scale variation, aspect ratio change, and low resolution), which tend to be under-represented in other benchmarks and are quite common in aerial tracking scenarios, pose significant challenges to current state-of-the-art trackers. This builds the stage for further improvements in precision and speed.

Our proposed UAV simulator along with novel evaluation methods enables tracker testing in real-world scenarios with live feedback before deployment. We will make this simulator publicly available to support more progress in the realm of UAV tracking, as well as, other computer vision tasks including aerial Structure-from-Motion (SfM), aerial localization, dynamic scene monitoring, etc. The simulator is not limited to UAVs alone but can be easily extended to simulate autonomous vehicles and evaluate their performance with algorithms designed for navigation and pedestrian detection.

Acknowledgments. Research in this paper was supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research.

References

1. Babenko, B., Yang, M.H., Belongie, S.: Visual Tracking with Online Multiple Instance Learning. IEEE transactions on pattern analysis and machine intelligence 33(8), 1619–1632 (Dec 2010)
2. Battaglia, P.W., Hamrick, J.B., Tenenbaum, J.B.: Simulation as an engine of physical scene understanding. Proceedings of the National Academy of Sciences 110(45), 18327–18332 (2013), http://www.pnas.org/content/110/45/18327.abstract
3. Bibi, A., Ghanem, B.: Multi-template scale-adaptive kernelized correlation filters. In: 2015 IEEE International Conference on Computer Vision Workshop (ICCVW). pp. 613–620 (Dec 2015)
4. Bibi, A., Mueller, M., Ghanem, B.: Target response adaptation for correlation filter tracking. In: European Conference on Computer Vision (ECCV 2016) (October 2016)
5. Bolme, D.S., Beveridge, J.R., Draper, B.A., Lui, Y.M.: Visual object tracking using adaptive correlation filters. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. pp. 2544–2550 (June 2010)
6. Collins, R., Zhou, X., Teh, S.K.: An open source tracking testbed and evaluation web site. In: IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS 2005), January 2005 (January 2005)
7. Danelljan, M., Hager, G., Shahbaz Khan, F., Felsberg, M.: Learning spatially regularized correlation filters for visual tracking. In: The IEEE International Conference on Computer Vision (ICCV) (Dec 2015)
8. Danelljan, M., Hger, G., Shahbaz Khan, F., Felsberg, M.: Accurate scale estimation for robust visual tracking. In: Proceedings of the British Machine Vision Conference. BMVA Press (2014)
9. Fu, C., Carrio, A., Olivares-Mendez, M., Suarez-Fernandez, R., Campoy, P.: Robust real-time vision-based aircraft tracking from unmanned aerial vehicles. In: Robotics and Automation (ICRA), 2014 IEEE International Conference on. pp. 5441–5446 (May 2014)
10. Gaszczak, A., Breckon, T.P., Han, J.: Real-time people and vehicle detection from uav imagery. In: Röning, J., Casasent, D.P., Hall, E.L. (eds.) IST/SPIE Electronic Imaging. vol. 7878, pp. 78780B–78780B–13. International Society for Optics and Photonics (January 2011)
11. Grabner, H., Grabner, M., Bischof, H.: Real-time tracking via on-line boosting. In: Proceedings of the British Machine Vision Conference. pp. 6.1–6.10. BMVA Press (2006), doi:10.5244/C.20.6
12. Hamalainen, P., Eriksson, S., Tanskanen, E., Kyrki, V., Lehtinen, J.: Online motion synthesis using sequential monte carlo. ACM Trans. Graph. 33(4), 51:1–51:12 (Jul 2014), http://doi.acm.org/10.1145/2601097.2601218
13. Hare, S., Saffari, A., Torr, P.H.S.: Struck: Structured output tracking with kernels. In: 2011 International Conference on Computer Vision. pp. 263–270. IEEE (Nov 2011)
14. Hattori, H., Naresh Boddeti, V., Kitani, K.M., Kanade, T.: Learning scene-specific pedestrian detectors without real data. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2015)
15. Hejrati, M., Ramanan, D.: Analysis by synthesis: 3d object recognition by object reconstruction. In: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. pp. 2449–2456 (June 2014)

16. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.: High-speed tracking with kernelized correlation filters. Pattern Analysis and Machine Intelligence, IEEE Transactions on (2015)
17. Henriques, J., Caseiro, R., Martins, P., Batista, J.: Exploiting the circulant structure of tracking-by-detection with kernels. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) Computer Vision ECCV 2012, Lecture Notes in Computer Science, vol. 7575, pp. 702–715. Springer Berlin Heidelberg (2012)
18. Hong, Z., Chen, Z., Wang, C., Mei, X., Prokhorov, D., Tao, D.: Multi-store tracker (muster): A cognitive psychology inspired approach to object tracking. In: Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on. pp. 749–758 (June 2015)
19. Jia, X., Lu, H., Yang, M.H.: Visual tracking via adaptive structural local sparse appearance model. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. pp. 1822–1829 (June 2012)
20. Ju, E., Won, J., Lee, J., Choi, B., Noh, J., Choi, M.G.: Data-driven control of flapping flight. ACM Trans. Graph. 32(5), 151:1–151:12 (Oct 2013), http://doi.acm.org/10.1145/2516971.2516976
21. Kalal, Z., Mikolajczyk, K., Matas, J.: Tracking-Learning-Detection. IEEE transactions on pattern analysis and machine intelligence 34(7), 1409–1422 (Dec 2011)
22. Kendall, A., Salvapantula, N., Stol, K.: On-board object tracking control of a quadcopter with monocular vision. In: Unmanned Aircraft Systems (ICUAS), 2014 International Conference on. pp. 404–411 (May 2014)
23. Kristan, M., Pflugfelder, R., Leonardis, A., Matas, J., Čehovin, L., Nebehay, G., Vojíř, T., Fernandez, G., Lukežič, A., Dimitriev, A., et al.: The visual object tracking vot2014 challenge results. In: Computer Vision-ECCV 2014 Workshops. pp. 191–217. Springer (2014)
24. Li, A., Lin, M., Wu, Y., Yang, M.H., Yan, S.: Nus-pro: A new visual tracking challenge. IEEE Transactions on Pattern Analysis and Machine Intelligence 38(2), 335–349 (Feb 2016)
25. Liang, P., Blasch, E., Ling, H.: Encoding color information for visual tracking: Algorithms and benchmark. IEEE Transactions on Image Processing 24(12), 5630–5644 (2015)
26. Liang, P., Blasch, E., Ling, H.: Encoding color information for visual tracking: Algorithms and benchmark. Image Processing, IEEE ... pp. 1–14 (2015), http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7277070
27. Lim, H., Sinha, S.N.: Monocular localization of a moving person onboard a quadrotor mav. In: Robotics and Automation (ICRA), 2015 IEEE International Conference on. pp. 2182–2189 (May 2015)
28. Mueller, M., Sharma, G., Smith, N., Ghanem, B.: Persistent aerial tracking system for uavs. In: Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference (October 2016)
29. Naseer, T., Sturm, J., Cremers, D.: Followme: Person following and gesture recognition with a quadrocopter. In: Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on. pp. 624–630 (Nov 2013)
30. Nussberger, A., Grabner, H., Van Gool, L.: Aerial object tracking from an airborne platform. In: Unmanned Aircraft Systems (ICUAS), 2014 International Conference on. pp. 1284–1293 (May 2014)
31. Papon, J., Schoeler, M.: Semantic pose using deep networks trained on synthetic RGB-D. CoRR abs/1508.00835 (2015), http://arxiv.org/abs/1508.00835

32. Pepik, B., Stark, M., Gehler, P., Schiele, B.: Teaching 3d geometry to deformable part models. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. pp. 3362–3369 (June 2012)
33. Pestana, J., Sanchez-Lopez, J., Campoy, P., Saripalli, S.: Vision based gps-denied object tracking and following for unmanned aerial vehicles. In: Safety, Security, and Rescue Robotics (SSRR), 2013 IEEE International Symposium on. pp. 1–6 (Oct 2013)
34. Pollard, T., Antone, M.: Detecting and tracking all moving objects in wide-area aerial video. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on. pp. 15–22 (June 2012)
35. Portmann, J., Lynen, S., Chli, M., Siegwart, R.: People detection and tracking from aerial thermal views. In: Robotics and Automation (ICRA), 2014 IEEE International Conference on. pp. 1794–1800 (May 2014)
36. Prokaj, J., Medioni, G.: Persistent tracking for wide area aerial surveillance. In: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. pp. 1186–1193 (June 2014)
37. Qadir, A., Neubert, J., Semke, W., Schultz, R.: chap. On-Board Visual Tracking With Unmanned Aircraft System (UAS). Infotech@Aerospace Conferences, American Institute of Aeronautics and Astronautics (Mar 2011), 0
38. Ross, D., Lim, J., Lin, R.S., Yang, M.H.: Incremental learning for robust visual tracking. International Journal of Computer Vision 77(1-3), 125–141 (2008)
39. Smeulders, A.W.M., Chu, D.M., Cucchiara, R., Calderara, S., Dehghan, A., Shah, M.: Visual tracking: An experimental survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 36(7), 1442–1468 (July 2014)
40. Smeulders, A.W.M., Chu, D.M., Cucchiara, R., Calderara, S., Dehghan, A., Shah, M.: Visual tracking: An experimental survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 36(7), 1442–1468 (July 2014)
41. Trilaksono, B.R., Triadhitama, R., Adiprawita, W., Wibowo, A., Sreenatha, A.: Hardwareintheloop simulation for visual target tracking of octorotor uav. Aircraft Engineering and Aerospace Technology 83(6), 407–419 (2011), http://dx.doi.org/10.1108/00022661111173289
42. Wu, Y., Lim, J., Yang, M.H.: Object tracking benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence 37(9), 1834–1848 (Sept 2015)
43. Wu, Y., Lim, J., Yang, M.H.: Online Object Tracking: A Benchmark. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition. pp. 2411–2418. IEEE (June 2013)
44. Zhang, J., Ma, S., Sclaroff, S.: MEEM: robust tracking via multiple experts using entropy minimization. In: Proc. of the European Conference on Computer Vision (ECCV) (2014)
45. Zhang, T., Bibi, A., Ghanem, B.: In defense of sparse tracking: Circulant sparse tracker. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
46. Zhang, T., Ghanem, B., Liu, S., Xu, C., Ahuja, N.: Robust visual tracking via exclusive context modeling. IEEE Transactions on Cybernetics 46(1), 51–63 (Jan 2016)
47. Zhang, T., Ghanem, B., Xu, C., Ahuja, N.: Object tracking by occlusion detection via structured sparse learning. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 1033–1040 (June 2013)
48. Zhang, T., Liu, S., Xu, C., Yan, S., Ghanem, B., Ahuja, N., Yang, M.H.: Structural sparse tracking. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 150–158 (June 2015)