Beyond standard benchmarks: Parameterizing performance evaluation in visual object tracking

Luka Čehovin Zajc, Alan Lukežič, Aleš Leonardis, Matej Kristan
University of Ljubljana
Faculty of computer and information science
{luka.cehovin, alan.lukezic, ales.leonardis, matej.kristan}@fri.uni-lj.si

Abstract

Object-to-camera motion produces a variety of apparent motion patterns that significantly affect performance of short-term visual trackers. Despite being crucial for designing robust trackers, their influence is poorly explored in standard benchmarks due to weakly defined, biased and overlapping attribute annotations. In this paper we propose to go beyond pre-recorded benchmarks with post-hoc annotations by presenting an approach that utilizes omnidirectional videos to generate realistic, consistently annotated, short-term tracking scenarios with exactly parameterized motion patterns. We have created an evaluation system, constructed a fully annotated dataset of omnidirectional videos and generators for typical motion patterns. We provide an in-depth analysis of major tracking paradigms which is complementary to the standard benchmarks and confirms the expressiveness of our evaluation approach.

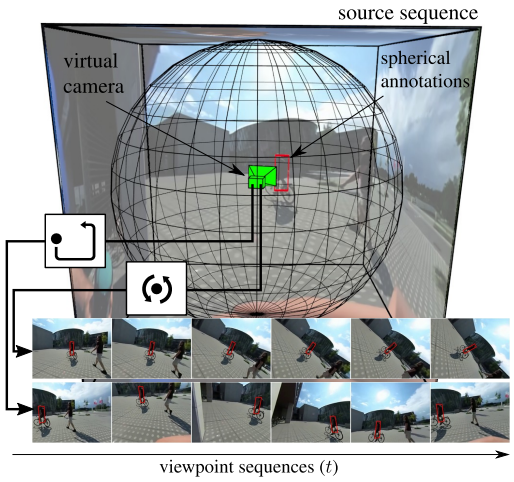

Figure 1: By re-parameterizing camera trajectory, a single 360° video produces various 2D viewpoint sequences with unique apparent motion patterns.

1. Introduction

Single-target visual object tracking has made significant progress in the last decade. To a large extent this can be attributed to the adoption of benchmarks, through which common evaluation protocols, datasets and baseline algorithms have been established. Starting with PETS initiative [28], several benchmarks on general single-target short-term tracking have been developed since, most notably OTB50 [25], VOT2013 [15], ALOV300+ [21], VOT2014 [16, 14], VOT2015 [13] OTB100 [26], TC128 [17] and VOT2016 [12].

The recent benchmarks [12, 18] report that, apart from the obvious situations like full occlusions, the trackers' performance is largely affected by the *apparent motion*, i.e., object motion with respect to the camera. The complexity of apparent motion patterns varies in realistic applications. An automated video-conferencing system largely observes translational motions, a drone circling over a target induces a large off-center rotational pattern, while water movement induces periodic scale changes in underwater robotic vessels. In some trackers, the translational motions are addressed by motion models. But compositions of scale changes, rotations and off-center translations are often assumed to be addressed by the visual models and localization techniques. This aspect is left largely unexplored in standard benchmarks, which are dominated by a handful of motion patterns and cannot fully expose the limitations of existing trackers. Advances in short-term tracking therefore call for accurate parametrization of apparent motion patterns in test sequences.

Dataset variation, systematic organization and low redundancy are crucial for practical evaluation, as argued by the recent work on test-data validation in computer vision [29]. Established benchmarks in visual tracking approach this requirement by increasing the number of sequences [21], applying advanced dataset construction methodologies [12] and by annotating entire sequences or

even individual frames with visual attributes [26, 14]. While such *bottom-up* approach is suitable for determining overall ranking of algorithms it is insufficient to study the performance of modern trackers along different motion patterns. Benchmarks contain annotated frames (or entire sequences) with only few attributes that correspond to motion patterns, which are only binary, non-parameterized and subject to human annotator bias. Additionally, accurate attribute-wise analysis is difficult due to the attribute cross-talk, meaning that multiple attributes occur at the same interval in a sequence (e.g., object rotation and rapid translation), which prohibits establishing a clear causal relation between a single motion pattern and tracker design performance. In principle, computer graphics generated sequences [9, 18] offer full camera control, however the level of realism in object motion and appearance in such sequences still presents a limitation for performance evaluation of general tracking methods.

Our work addresses the limitations of the traditional benchmarks by proposing a framework for *top-down* construction of test sequences through parametrization of apparent motion patterns. A virtual camera model that utilizes omnidirectional videos is introduced to generate photo-realistic, consistently annotated short-term tracking scenarios (Figure 1). The exact specification of parameterized motion patterns guarantees a clear causal relation between the generated apparent motion and the tracking performance change. This enables fine-grained performance analysis and can be used complementary to the existing benchmarks to offer an in-depth analysis of tracking approaches.

Contributions. Our contributions are three-fold. (1) We propose a new performance evaluation paradigm based on generation of realistic sequences with high degree of motion pattern parametrization from annotated omnidirectional videos. (2) We have constructed a new apparent motion benchmark for short-term single-target trackers. A new dataset with per-frame target annotation in omnidirectional videos adding up to 17537 frames and generators of twelve motion patterns are introduced. (3) We have evaluated 17 state-of-the-art trackers from recent benchmarks categorized in major tracking paradigms [26, 12] and provide insights not available in standard benchmarks. The new benchmark, the results and the corresponding software will be made publicly available and are expected to significantly affect future developments in single-target tracking.

Structure. The paper is organized as follows, in Section 2 we review related work, in Section 3 we describe our sequence parameterization framework, in Section 4 we present the proposed benchmark, in Section 5 we present experimental results, and in Section 6 we discuss our findings and make concluding remarks.

2. Related work

Modern short-term tracking benchmarks [26, 15, 14, 13, 12, 21, 18] acknowledge the importance of motion related attributes and support evaluation with respect to these. However, the evaluation capability significantly depends on the attribute presence and distribution, which is often related to the sequence acquisition. In application-oriented benchmarks like [18] the attribute distribution is necessarily skewed by the application domain. Some general benchmarks [26, 21] thus include a large number of sequences from various domains. But since the sequences are post-hoc annotated, the dataset diversity is hard to achieve. A recent benchmark [12] addressed this by considering the attributes already at the sequence collection stage and applied an elaborate methodology for automatic dataset construction.

The strength of per-attribute evaluation depends on the annotation approach. In most benchmarks [26, 21, 18] all frames are annotated by an attribute even if it occupies only a part of the sequence. Kristan et al. [14] argued that this biases per-attribute performance towards average performance and proposed a per-frame annotation to reduce the bias. However, a single frame might still contain several attributes, resulting in the attribute cross-talk bias.

The use of computer graphics in training and evaluation has recently been popularized in computer vision. Mueller et al. [18] propose online virtual worlds creation for drone tracking evaluation, but using only a single type of the object, without motion parametrization, produces a low level of realism. Vig et al. [9] address the virtual worlds realism levels, ambient parametrization learning and performance evaluation, however, only for vehicle detection.

Our work is positioned between the standard benchmarking approaches and the synthetic sequence generation. By using *real imagery* we retain photo-realism of standard benchmarks. Our approach simultaneously enables parameterization of apparent motion thus opening new possibilities for in-depth evaluation that is complementary to existing benchmarks.

3. Sequence parametrization

Two key concepts are introduced by our apparent-motion evaluation methodology: a *source sequence* and a *viewpoint sequence*. A source sequence is an omnidirectional video that simultaneously captures 360 degree field of view. The video is stored as a projection onto a spectator-centered sphere, i.e., $\mathcal{S} = \{S_t\}_{t=1:N}$, where S_t is a projection at frame t. Such representation allows to generate arbitrary views of a 3D scene from the point of observation.

A viewpoint sequence is a sequence of images obtained from a spherical representation by projection into a pinhole camera, i.e., $\mathcal{I} = \{I_t\}_{t=1:N}$. The camera model has ad-

justable rotation and focal length parameters, thereby defining the state of the camera at time t as

$$C_t = [\alpha_t, \beta_t, \gamma_t, f_t], \tag{1}$$

where the first three parameters are the Euler angles and f_t denotes the focal length. Each frame in a viewpoint sequence is therefore the result of the corresponding image in the source sequence and the camera parameters, i.e. $I_t = p_{\text{cam}}(S_t; C_t)$.

The ground truth object state in each frame is specified in a viewpoint-agnostic spherical coordinate system, i.e., $\mathcal{A} = \{A_t\}_{t=1:N}$. Following the VOT Challenge protocol [14] the state is defined as a rectangle using four-points $A_t = \{\theta_t^i, \rho_t^i\}_{i=1:4}$. Given a pinhole camera viewpoint parameters C_t, the ground truth A_t is projected into the image plane by projective geometry, i.e., $G_t = p_{\text{gt}}(A_t; C_t)$.

The camera parameters C_t are continually adjusted during the creation of the viewpoint sequence to keep the projected object within the field of view, thus satisfying the short-term tracking constraint. The camera viewpoint is adjusted via a *camera controller* $p_{\text{con}}(\cdot, \cdot)$ that applies a prescribed motion pattern E and maps the object ground truth state into camera parameters while satisfying the short-term tracking constraint, i.e.,

$$p_{\text{con}}(A_t, E, t) \mapsto C_t. \tag{2}$$

Depending on the pattern type specification, the controller continually adjusts camera-to-object position and generates various apparent object motions.

3.1. Evaluation framework

The evaluation framework implements the *VOT supervised evaluation mode* [6] and the VOT [14, 12] performance evaluation protocol, which allows full use of long sequences. In this evaluation mode, a tracker is initialized and re-set upon drifting off the target. Stochastic trackers are run multiple times and the results are averaged.

The following functionality is required by the supervised experiment mode: (1) reproducible sequence generation and (2) bi-directional tracker-evaluator communication. The viewpoint sequence and the 2D ground truth are therefore generated on the fly during the evaluation and are reproducible for each time-step. The communication between the evaluator and the tracker is implemented through the state-of-the-art TraX [22] communication protocol. Our evaluation framework is summarized in Figure 2.

4. Apparent-motion patterns benchmark

Our motion parametrization framework is demonstrated on a novel single-target visual object tracking benchmark for isolated apparent-motion patterns (AMP). The benchmark contains fifteen very long omnidirectional sequences

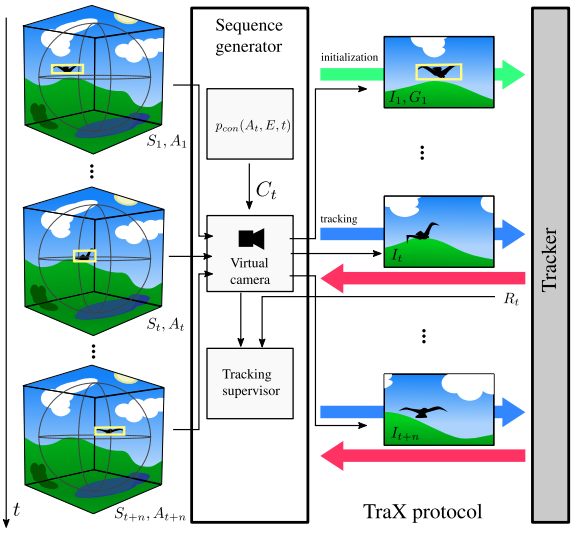

Figure 2: The evaluation framework. The sequence generator constructs a viewpoint sequence with corresponding ground truths according to the motion pattern. Tracker reports predicted region in each frame to the evaluator for automatic failure detection.

(adding up to 17537 frames) and specifies twelve motion types.

4.1. Dataset acquisition

The new dataset contains fifteen omnidirectional videos with an average video length of 1169 frames, amounting to 17537 frames. The videos were mostly selected from a large collection of 360 degree videos available on YouTube. To maximize the target diversity, we recorded additional sequences using Ricoh Theta 360 degree camera. Videos were converted to a cube-map projection and encoded with MP4 H.264 codec. Each frame of the video was manually annotated by a rectangular region encoded in spherical coordinates using an annotation tool specifically designed for this use case. Some of the viewpoint frame examples of individual source sequences are shown in Figure 3.

4.2. Motion patterns specification

We consider six motion pattern classes that reflect typical dynamic relations between an object (target) and a camera:
Stabilized setup, denoted as E_b, keeps the object position at image center and adjusts the camera distance to keep the object diagonal constant at 70 pixels. A variant with a diagonal constant at 35 pixels is considered as well to test tracking objects from far away, E_b^s.
Centered rotation setup, denoted as E_r, fixes the object center and the scale as E_b and then rotates the camera around the optical axis. Two variants, with low and high rotation speeds, $\text{E}_\text{r}^\text{s}$ and $\text{E}_\text{r}^\text{f}$, respectively, are considered.

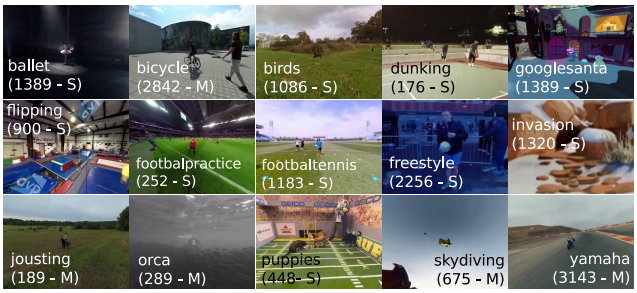

Figure 3: A preview of 360 degree sequences in the dataset from the view that centers the target. The number in brackets is the number of frames, the letter denotes if the camera was Stationary or Moving).

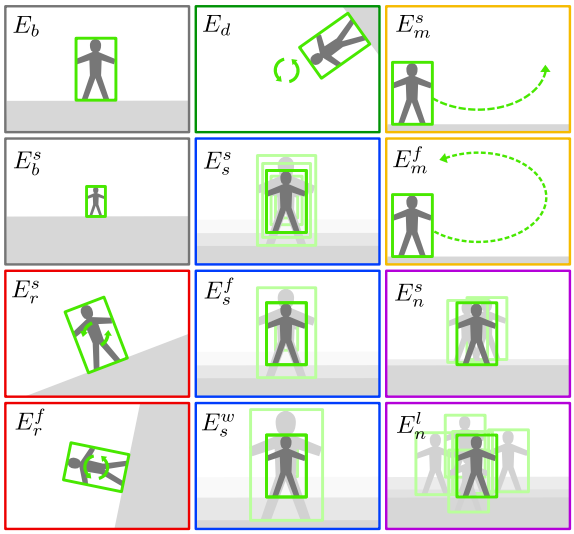

Figure 4: The twelve apparent-motion patterns in the AMP benchmark.

Displaced rotation setup, denoted as $\mathrm{E_d}$, displaces the object center and then rotates the camera around its optical axis.

Scale change setup, denoted as $\mathrm{E_s}$, fixes the center and then periodically changes the scale by a cosine function with amplitude oscillation around the nominal scale of $\mathrm{E_b}$. Two variants, i.e., with a low, $\mathrm{E_s^s}$, and a high frequency, $\mathrm{E_s^f}$, but equally moderate amplitude are considered. Another variant with a moderate frequency but large amplitude, $\mathrm{E_s^w}$, is considered as well.

Planar motion setup, denoted as $\mathrm{E_m}$, displaces the camera from the object center and performs circular motion in the image plane. A variant with low – $\mathrm{E_m^s}$ and high – $\mathrm{E_m^f}$ frequency are considered.

Translation noise setup, denoted as $\mathrm{E_n}$, fixes the center and the scale as in $\mathrm{E_b}$ then randomly displaces the center by drawing a displacement vector from a normal distribution. Two variants, one with small, $\mathrm{E_n^s}$, and one with large, $\mathrm{E_n^l}$, noise are considered.

The variations of six motion classes result in 12 different motion patterns, which are illustrated in Figure 4. While these patterns may seem synthetic, they actually occur in many active-camera robotics scenarios, e.g., a drone circling over an observed target (rotation) or an autonomous boat being swayed by the sea (scale change). Note that each omnidirectional video in our dataset creates a sequence with specific motion parameters. Thus the effect of each motion pattern is evaluated on all frames without being influenced by the presence of other patterns, establishing clear casual relationships between the patterns and the tracker's performances.

4.3. Benchmark comparison

A comparison of our proposed AMP with most popular standard benchmarks is summarized in Table 1. The values under MAC indicate the percentage of frames in the dataset with at least a single motion attribute. The motion attributes are most frequent in the AMP (100% coverage) and UAV123 [18] (96% coverage). To reflect the dataset size in motion evaluation, we compute the number of effective frames per attribute (FPA). This measure counts the number of frames that contain a particular motion attribute, where each frame contributes with weight inversely proportional to the number of motion attributes it contains. The FPA is highest for an application-specific UAV123 [18] (27107). Among the general tracking benchmarks, this value is highest for the proposed AMP (17537), which exceeds the second largest (OTB100 [26]) by over 30%.

The FPA alone does not fully reflect the evaluation strength since it does not account for the attribute cross-talk. A lower bound on the cross-talk is reflected by the INTER measure that shows a percentage of motion-annotated frames with at least two motion attributes. The measure shows that well over half of the frames in UAV123 [18] (78%) and OTB100 [26] (63%) suffer from the attribute cross-talk. The cross-talk is lowest for the proposed AMP (0%), ALOV [21] (0%) and VOT2016 [12] (32%).

Most existing benchmarks are annotated by four motion pattern types. The proposed AMP benchmark contains approximately three times more motion pattern types than existing benchmarks. The existing benchmarks lack motion pattern quantification (e.g., the extent of *speed* in attribute fast motion), which results in inconsistent definitions across benchmarks. In contrast, the motion patterns are objectively defined through their parametrization in the proposed AMP benchmark.

4.4. Performance measures

The tracking performance is measured by the VOT [13] measures: tracker accuracy (A), robustness (R) as well as the expected average overlap (EAO). The accuracy mea-

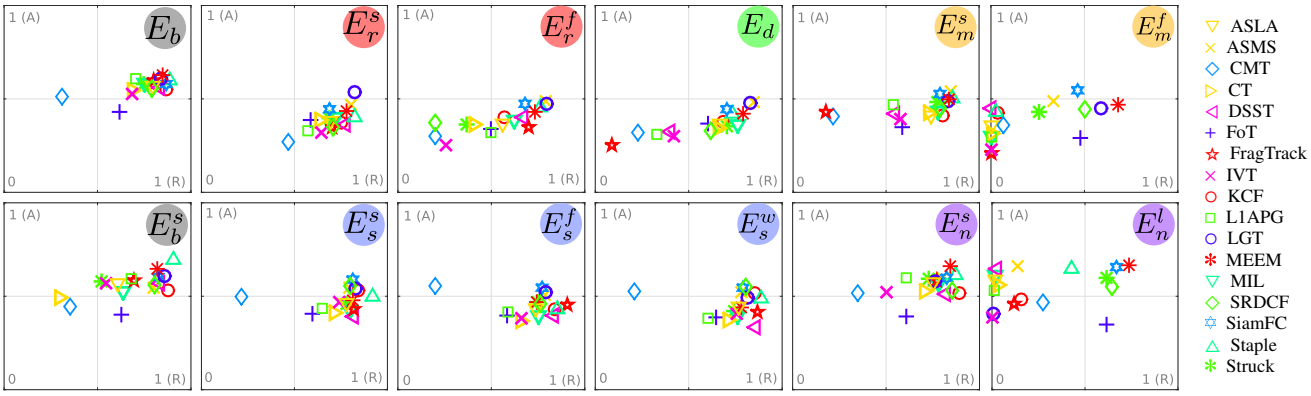

Figure 5: A-R plots for each experiment in the benchmark. The vertical axis denotes accuracy and the horizontal axis denotes robustness. The *sensitivity* visualization parameter was set to 100 frames in all plots.

Table 1: Comparison of AMP with popular recent tracking benchmarks: ALOV300+ [21], OTB100 [26], UAV123 [18] and VOT2016 [12]. Best, second best and third best values are shown in red, blue and green, respectively.

Dataset	[21]	[26]	[18]	[12]	AMP
MAC (%)	19	88	*96*	61	**100**
FPA	4275	12929	**27107**	4366	*17537*
INTER (%)	**0**	63	78	*32*	**0**
Motion classes	3	3	3	3	**6**
Motion patterns	*4*	*4*	*4*	3	**12**
Parameterized	no	no	no	no	**yes**
Per-frame	no	no	no	**yes**	**yes**

sures the overlap between the output of the tracker and the ground truth bounding box during periods of successful tracking, while the robustness measures the number of times a tracker failed and required re-initialization [6]. The expected average overlap score is an estimator of the average overlap on a typical short-term sequence a tracker would obtain without reset [13]. All scores are calculated on per-sequence basis and averaged with weights proportional to the sequence length.

The VOT methodology allows us to use longer sequences since the tracker is restarted after the object is lost which lowers the variance of the results due to after-failure drift [14]. Longer sequences also reduce the chance of saturation (a chance that many trackers track an object for the entire sequence with only minor differences in accuracy).

5. Evaluation and results

To demonstrate the verbosity of the AMP benchmarks we have evaluated 17 trackers. Each tracker was evaluated on a total of 210444 frames, which makes this the largest fine-grained motion-related tracker evaluation to date.

5.1. Trackers tested

A set of 17 trackers was constructed by considering baseline and top-performing representatives on recent benchmarks [26, 12] from the following 6 broad classes of trackers. *(1) Baselines* include standard discriminative and generative trackers MIL [2], CT [31], IVT [20], and FragTrac [1], a state-of-the-art mean-shift tracker ASMS [24], and Struck SVM tracker [10]. *(2) Correlation filters* include the standard KCF [11] and three top-performing correlation filters on VOT2016 [12] – DSST [7], Staple [4] and SRDCF [8]. *(3) Sparse trackers* include top-performing sparse trackers L1APG [3] and ASLA [27]. *(4) Part-based trackers* include the recent state-of-the-art CMT [19], LGT [32], FoT [23]. In addition the set comprises a state-of-the-art *(5) Hybrid tracker* MEEM [30] and *(6) ConvNet tracker* SiamCF [5].

5.2. Experimental results

The results are summarized by A-R plots (Figure 5), general performance graphs (Figure 6, Figure 7) and in Table 2. In the following we discuss the performance with respect to various motion pattern classes and instances.

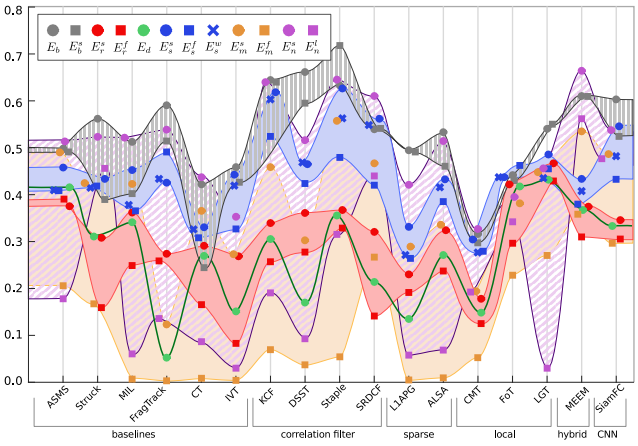

Figure 6: EAO values for all motion patterns over tested trackers.

Scale adaptation: Slow scale changes (E_s^s, Figure 6) are

addressed best by correlation filters that apply scale adaptation (i.e., KCF, DSST, Staple, SRDCF). Their performance is not significantly affected as long as the change is gradual enough, even for large amplitudes (E_s^w). However, fast changes (E_s^f) significantly reduce performance, implying that the number of scales explored should be increased in these trackers. The ConvNet tracker SiamCN does not suffer from this discrepancy, which is likely due to a large set of scales it explores. The difference in performance drop for fast (E_s^f) and large (E_s^w) scale change is low for scale-adaptive mean shift ASMS and part-based trackers (i.e., CMT, FoT and LGT). In contrast to correlation filters, these trackers do not greedily explore the scale space but apply blob size estimation (ASMS) or apply key-point-like matching approaches (CMT, FoT, LGT). The average performance at moderate scale change is better for correlation filters than part-based trackers. Struck and MEEM are least affected by the scale change among the trackers that do not adapt their scale. From the AR plots in Figure 5 it is apparent that the performance drops are due to a drop in accuracy, but not in failures.

Rotation: Rotation (E_r) significantly affects the performance of all classes of the trackers. Figure 7 and the AR plots in Figure 5 show that the drop comes from a reduced accuracy as well as increased number of failures across most trackers. The drop is least apparent with ASMS, FoT and LGT which is likely due to their object visual models. The visual model in ASMS is rotation invariant since it is based on color histograms, while FoT and LGT explicitly address rotation by geometric matching. Rotation most significantly affects performance of correlation filters and ConvNets (Figure 6). These trackers apply templates for tracking and since rotation results in significant discrepancies between the template and object, the trackers fail. In particular, from the AR plots in Figures 5 we see that slow rotation (E_r^s) only results in decreased accuracy, but fast rotation (E_r^f) results in increased failures as well (e.g., SRDCF). On the other hand, the performance of correlation filters, ConvNet tracker (SiamFC) and hybrid tracker (MEEM) surpasses the part-based models when no rotation is observed (E_b in Figures 5 and Figure 6).

Motion: From the AR plots in Figure 5 we see that slow planar motion (E_m^s) only slightly reduces performance in general, but this reduction is significant for most trackers in case of fast motion (E_m^f). LGT is the only tracker resilient to fast motion. A likely reason is the use of nearly-constant-velocity motion model in the LGT. However, the performance significantly drops for this tracker when extensive random motion is observed (E_n^f in Figure 6). Trackers like SiamFC and MEEM are least affected by all patterns of fast motions. The reason is likely in their very large search region for target localization. The AR plots in Figure 5 indicate that SiamCF fails much more often at fast motions (E_m^f) than MEEM implying that MEEM is more robust at local search.

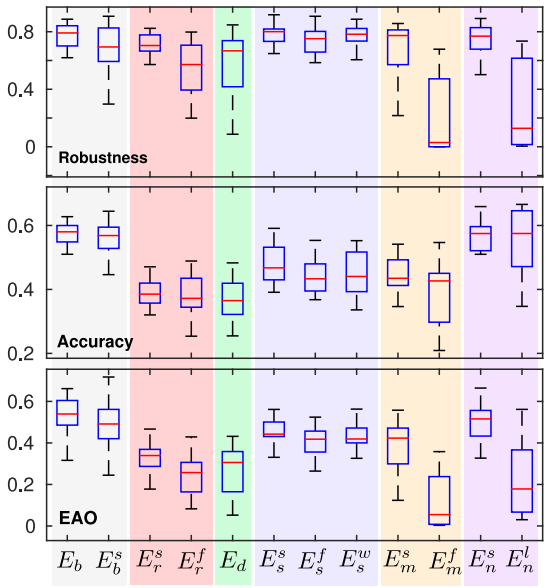

Figure 7: Motion patterns difficulty levels according to robustness, accuracy, and EAO. Motion patterns are grouped by motion classes: stabilized (E_b), centered rotation (E_r), displaced rotation (E_d), scale change (E_s), planar motion (E_m) and noise (E_n).

Object size: All trackers perform very well in the baseline setup (E_b) in which the object is kept centered and of constant size (Figure 7 and Figure 6). In fact, top performance is achieved by the correlation filter trackers. The reason is that the visual model assumptions that these trackers make exactly fit this scenario. When considering smaller objects (E_b^s) the following trackers appear unaffected: ASMS, KCF, SRDCF, L1APG, CMT, FoT, LGT and MEEM. This implies that the level of detail of target representation in these trackers is unaffected by the reduced object size. Note that these trackers come from different classes. The AR plots in Figure 5 show that performance drop in tracking small objects is most significant for baselines like CT, IVT, MIL and Fragtrac as well as a sparse tracker ASLA and the Struck tracker. The performance drop comes from increased failures, which means that their representation is not discriminative enough on this scale which leads to frequent drifts.

General observations: All trackers exhibit a large performance variance across the apparent-motion patterns (Figure 6). The variance appears lowest over most motion patterns for the part-based trackers, although their average performance is moderate. Table 2 shows the average performance over the motion patterns without the baseline motion pattern (E_b). Among the trackers whose average EAO is within 70% of best EAO are three out for four correla-

tion filters, a hybrid tracker (MEEM), a ConvNet tracker (SiamFC), two out of three part-based trackers and two baselines (ASMS and Struck). The top three trackers in average performance are Staple (0.470 EOA), MEEM (0.468 EAO) and SiamCF (0.448 EAO). These trackers are also performing well on the recent benchmarks, however, our analysis shows that the weak spot of these trackers are target rotations, as well as fast movements and shaky videos.

Motion class difficulty: Considering the average EAO in Figure 7, the most difficult classes are rotation (both patterns—central and displaced) as well as planar motion and translation noise, but the distribution of difficulty within individual classes as well as the degradation modes vary. The AR plots in Figure 5 show that performance drops in rotation are due to inaccurate bounding box estimation, leading to reduced accuracy but not to complete failure. This figure also shows that trackers generally well address planar motion, but tend to fail at fast nonlinear motions due to large inter-frame displacement.

5.3. Relation to existing benchmarks

A relation of AMP to the existing tracking benchmarks was established by comparing the ranks of common trackers on a well known OTB100 [26] and the recent UAV123 [18] benchmark.

Comparison with OTB100: The OTB100 contains relatively old trackers, therefore the intersection is in the following six trackers: ASLA, CT, FoT, IVT, L1-APG, MIL, and Struck. Figure 8 shows the ranking differences between these trackers for the different ranking modes. The ranking by average performance differs mainly for L1-APG and FoT trackers. The possible reasons for this are different implementations, algorithm parameters, as well as different evaluation methodology[1]. Three motion patterns in OTB100 are compatible with AMP: scale variation, fast motion and in-plane rotation. The performance over three scale changing motion patterns on AMP was averaged to obtain a scale change ranking. While the FoT achieves top performance on AMP it is positioned relatively low on OTB100, which is likely due to different implementations (ours is from the authors) and interaction of other attributes on OTB100. Both rankings place Struck at the top, while ranks of other trackers vary. The fast motion ranking on AMP was obtained by averaging fast motions, i.e., E_m^f and E_n^w. Both benchmarks rank Struck as top performing and IVT as worst performing. The in-plane rotation attribute was compared with combined ranking of center and displaced rotation (E_r and E_d). The situation is similar to scale change, where FoT, which explicitly addresses rotation, is ranked much lower according to OTB100.

[1]The OTB100 methodology does not restart a tracker on failure which can lead to large differences between trackers that support re-detection and those that do not.

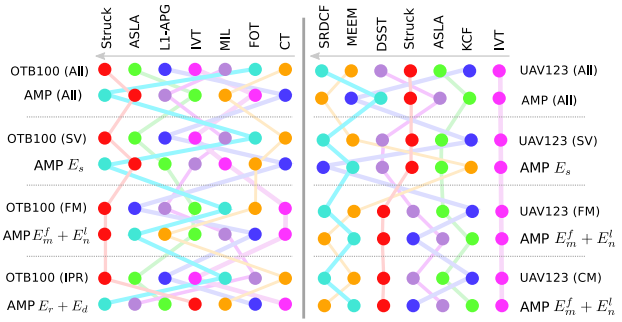

Figure 8: Tracker ranking comparison of AMP with OTB100 (left) and UAV123 (right). The trackers are sorted from left (best) to right. Attribute abbreviations: SV – Scale Variation, FM – Fast Motion, IPT – In-Plane Rotation, CM – Camera Motion.

Comparison with UAV123: The AMP and UAV123 intersect in the following seven trackers: ASLA, DSST, IVT, KCF, MEEM, SRDCF, and Struck. The comparison of average performance as well as with respect to three types of motion patterns: scale variation, camera motion and fast motion is shown in Figure 8. The ranks are mostly consistent with the best two trackers mostly being SRDCF and MEEM. A discrepancy is observed for the MEEM tracker at scale change attribute. MEEM does not adapt the scale, which results in a low rank at AMP. However, it is ranked high on UAV123, which is likely due to attribute cross-talk. The discrepancy in KCF is due to implementation – our KCF adapts the scale. Notice that the UAV123 ranks on camera motion are equal to scale variation ranks. We therefore compare both with ranks obtained by averaging fast in-plane motion (E_m^f) and large translation noise (E_n^w) performance on AMP. The ranks match very well, which means that AMP offers a significant level of granularity in analysis.

6. Discussion and conclusions

We have proposed a novel approach for single-target tracker evaluation on parameterized motion-related attributes. At the core of our approach is the use of 360 degree videos to generate annotated realistically-looking tracking scenarios. We have presented a novel benchmark AMP, composed of an annotated dataset of fifteen such videos and the results of 17 state-of-the-art trackers. We have experimentally verified the realism of the generated sequences by reproducing partial ranks available in standard benchmarks.

The results of our experiments provide a detailed overview of strengths and limitations of modern short-term visual trackers. The scale change appears to be well addressed by many tracking approaches. Even trackers that do not adapt scale do not fail often. Nevertheless, in practice scale change is often accompanied by appearance change or

Table 2: Overview of the EAO scores and their relative differences according to the baseline score. The top value in each cell represents the absolute EAO score while the bottom one represents the EAO difference in relation to the baseline experiment. Green text denotes relative increase, orange text relative decrease, and red and bold red text decrease greater than 25% and 50% of the baseline score. The baseline experiment is not used for computing the average tracker score.

	E_b	E_b^s	E_r^s	E_r^f	E_d	E_s^s	E_s^f	E_s^w	E_m^s	E_m^f	E_n^s	E_n^w	**Average**
△ Staple	0.633	0.718	0.367	0.329	0.356	0.627	0.479	0.563	0.558	0.055	0.646	0.315	0.456
		0.085	*-0.266*	*-0.304*	*-0.277*	*-0.006*	*-0.154*	*-0.070*	*-0.075*	**-0.578**	**0.013**	**-0.318**	*-0.177*
✻ MEEM	0.610	0.609	0.375	0.310	0.367	0.434	0.380	0.408	0.535	0.358	0.664	0.562	0.455
		-0.001	*-0.235*	*-0.300*	*-0.243*	*-0.176*	*-0.230*	*-0.202*	*-0.075*	*-0.252*	**0.054**	*-0.048*	*-0.155*
✡ SiamFC	0.603	0.525	0.346	0.305	0.334	0.546	0.433	0.482	0.487	0.296	0.538	0.477	0.433
		-0.078	*-0.257*	*-0.298*	*-0.269*	*-0.057*	*-0.170*	*-0.120*	*-0.116*	**-0.307**	*-0.065*	*-0.126*	*-0.169*
○ KCF	0.644	0.640	0.339	0.257	0.306	0.618	0.524	0.603	0.459	0.069	0.640	0.191	0.422
		-0.004	*-0.305*	**-0.387**	**-0.338**	*-0.026*	*-0.120*	*-0.041*	*-0.185*	**-0.575**	*-0.005*	**-0.453**	*-0.222*
◇ SRDCF	0.539	0.542	0.320	0.141	0.214	0.562	0.420	0.548	0.467	0.267	0.610	0.440	0.412
		0.002	*-0.219*	**-0.398**	**-0.326**	**0.022**	*-0.119*	**0.009**	*-0.073*	**-0.272**	**0.071**	*-0.099*	*-0.128*
○ LGT	0.540	0.550	0.467	0.429	0.432	0.486	0.455	0.435	0.448	0.271	0.456	0.030	0.405
		0.010	*-0.074*	*-0.111*	*-0.109*	*-0.055*	*-0.085*	*-0.105*	*-0.092*	*-0.270*	*-0.085*	**-0.510**	*-0.135*
× ASMS	0.498	0.491	0.375	0.390	0.415	0.458	0.409	0.410	0.490	0.206	0.514	0.178	0.394
		-0.007	*-0.123*	*-0.107*	*-0.082*	*-0.040*	*-0.089*	*-0.088*	*-0.008*	**-0.292**	**0.016**	**-0.320**	*-0.104*
+ FoT	0.442	0.430	0.422	0.297	0.417	0.437	0.462	0.437	0.381	0.228	0.395	0.342	0.386
		-0.011	*-0.020*	*-0.145*	*-0.024*	*-0.005*	**0.021**	*-0.004*	*-0.060*	*-0.213*	*-0.046*	*-0.100*	*-0.055*
✻ Struck	0.562	0.390	0.308	0.160	0.311	0.433	0.418	0.414	0.425	0.167	0.524	0.456	0.364
		-0.172	*-0.254*	**-0.403**	*-0.251*	*-0.129*	*-0.144*	*-0.148*	*-0.137*	**-0.395**	*-0.038*	*-0.106*	*-0.198*
◁ DSST	0.662	0.595	0.361	0.278	0.170	0.465	0.423	0.468	0.303	0.038	0.516	0.093	0.337
		-0.066	*-0.301*	**-0.384**	**-0.492**	*-0.197*	*-0.238*	*-0.193*	**-0.359**	**-0.624**	*-0.146*	**-0.568**	*-0.324*
▽ MIL	0.513	0.402	0.362	0.249	0.341	0.452	0.366	0.378	0.423	0.006	0.522	0.061	0.324
		-0.110	*-0.151*	**-0.264**	*-0.172*	*-0.060*	*-0.147*	*-0.135*	*-0.089*	**-0.506**	**0.009**	**-0.452**	*-0.189*
▽ ASLA	0.533	0.461	0.324	0.237	0.271	0.433	0.385	0.416	0.336	0.009	0.514	0.069	0.314
		-0.072	*-0.209*	**-0.296**	*-0.262*	*-0.100*	*-0.148*	*-0.117*	*-0.196*	**-0.524**	*-0.019*	**-0.464**	*-0.219*
☆ FragTrack	0.590	0.514	0.274	0.259	0.052	0.426	0.491	0.434	0.123	0.003	0.539	0.136	0.296
		-0.076	**-0.316**	**-0.331**	**-0.538**	*-0.164*	*-0.099*	*-0.157*	**-0.467**	**-0.587**	*-0.052*	**-0.454**	*-0.295*
▷ CT	0.422	0.245	0.291	0.166	0.269	0.331	0.308	0.326	0.366	0.008	0.437	0.086	0.258
		-0.177	*-0.131*	**-0.256**	*-0.153*	*-0.091*	*-0.114*	*-0.096*	*-0.056*	**-0.413**	**0.015**	**-0.335**	*-0.164*
× IVT	0.458	0.427	0.269	0.083	0.151	0.442	0.327	0.419	0.273	0.004	0.353	0.030	0.252
		-0.032	*-0.190*	**-0.376**	**-0.308**	*-0.016*	*-0.131*	*-0.039*	*-0.186*	**-0.455**	*-0.106*	**-0.428**	*-0.206*
□ L1APG	0.495	0.491	0.230	0.192	0.135	0.332	0.264	0.272	0.289	0.005	0.421	0.058	0.244
		-0.004	**-0.265**	**-0.303**	**-0.360**	*-0.163*	*-0.231*	*-0.223*	*-0.206*	**-0.490**	*-0.074*	**-0.437**	**-0.251**
◇ CMT	0.316	0.297	0.178	0.125	0.148	0.304	0.280	0.277	0.194	0.053	0.327	0.192	0.216
		-0.019	*-0.139*	**-0.191**	**-0.168**	*-0.012*	*-0.036*	*-0.039*	*-0.122*	**-0.264**	**0.011**	*-0.124*	*-0.100*
Average	0.533	0.490	0.330	0.247	0.276	0.458	0.401	0.429	0.386	0.120	0.507	0.219	
		-0.043	*-0.203*	**-0.286**	*-0.257*	*-0.075*	*-0.131*	*-0.104*	*-0.147*	**-0.413**	*-0.026*	**-0.314**	

fast motion, which increase chances of failures. We believe that this is the reason why scale change is perceived as a challenging attribute in related benchmarks. The state-of-the-art trackers perform reasonably well in tracking small targets. Rotation and abrupt motion are two of the most challenging motion classes. Due to their scarcity on existing benchmarks they remain poorly addressed by most modern trackers. Our results have shown that non-random motions are well addressed by motion models, which have also become quite rare in modern trackers. We believe that future research in tracker development should focus on these topics to make further improvements.

We have demonstrated the usefulness of the proposed approach for evaluating trackers in a controlled, yet realistic environment. The approach is complementary to existing benchmarks allowing better insights into tracking behavior on various apparent-motion patterns. Moreover, capturing omnidirectional videos is nowadays possible with commodity equipment. Therefore our dataset adaptation to a specific tracking scenario may in fact be easier than in traditional approaches since it does not require careful planning before the acquisition to cover all possible motion patterns. Our framework allows a straightforward quantified simulation of arbitrary attribute crosstalk across a sequence. Our future work will therefore focus on evaluation of complex motion patterns and their effects on tracking performance. We also plan to explore adaptation of our evaluation methodology to active tracking.

Acknowledgements This work was in part supported by the Slovenian research agency ARRS program P2-0214 and the Slovenian research agency ARRS research project J2-8175. Aleš Leonardis was supported in part by MoD/Dstl and EPSRC MURI project.

References

[1] A. Adam, E. Rivlin, and I. Shimshoni. Robust Fragments-based Tracking using the Integral Histogram. In *Comp. Vis. and Patt. Recognition*, volume 1, pages 798–805. IEEE Computer Society, jun 2006. 5

[2] B. Babenko, M.-H. Yang, and S. Belongie. Robust object tracking with online multiple instance learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(8):1619–1632, Aug. 2011. 5

[3] C. Bao, Y. Wu, H. Ling, and H. Ji. Real time robust l1 tracker using accelerated proximal gradient approach. In *Comp. Vis. Patt. Recognition*, pages 1830–1837, June 2012. 5

[4] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P. H. S. Torr. Staple: Complementary learners for real-time tracking. In *Comp. Vis. Patt. Recognition*, pages 1401–1409, June 2016. 5

[5] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr. Fully-convolutional siamese networks for object tracking. *arXiv preprint arXiv:1606.09549*, 2016. 5

[6] L. Čehovin, A. Leonardis, and M. Kristan. Visual object tracking performance measures revisited. *IEEE Trans. Image Proc.*, 25(3):1261–1274, 2016. 3, 5

[7] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg. Accurate scale estimation for robust visual tracking. In *Proc. British Machine Vision Conference*, pages 1–11, 2014. 5

[8] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg. Learning spatially regularized correlation filters for visual tracking. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 4310–4318, 2015. 5

[9] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Comp. Vis. and Patt. Recognition (CVPR)*, pages 4340–4349, June 2016. 2

[10] S. Hare, A. Saffari, and P. H. S. Torr. Struck: Structured output tracking with kernels. In *Int. Conf. Computer Vision*, pages 263–270, Washington, DC, USA, 2011. IEEE Computer Society. 5

[11] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista. High-speed tracking with kernelized correlation filters. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(3):583–596, 2014. 5

[12] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, L. Čehovin, T. Vojir, G. Häger, A. Lukežič, G. Fernandez, and et. al. The visual object tracking VOT2016 challenge results. In *Proc. European Conf. Computer Vision*, 2016. 1, 2, 3, 4, 5

[13] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, L. Čehovin, G. Fernandez, T. Vojir, G. Häger, G. Nebehay, and R. P. et. al. The visual object tracking vot2015 challenge results. In *Int. Conf. Computer Vision*, 2015. 1, 2, 4, 5

[14] M. Kristan, J. Matas, A. Leonardis, T. Vojir, R. Pflugfelder, G. Fernandez, G. Nebehay, F. Porikli, and L. Cehovin. A novel performance evaluation methodology for single-target trackers. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2016. 1, 2, 3, 5

[15] M. Kristan, R. Pflugfelder, A. Leonardis, J. Matas, F. Porikli, L. Čehovin, G. Nebehay, G. Fernandez, and T. Vojir et al. The visual object tracking vot2013 challenge results. In *Vis. Obj. Track. Challenge VOT2013, In conjunction with ICCV2013*, pages 98–111, Dec 2013. 1, 2

[16] M. Kristan, R. Pflugfelder, A. Leonardis, J. Matas, L. Čehovin, G. Nebehay, T. Vojir, and G. F. et al. The visual object tracking vot2014 challenge results. In *Proc. European Conf. Computer Vision*, pages 191–217, 2014. 1

[17] P. Liang, E. Blasch, and H. Ling. Encoding color information for visual tracking: Algorithms and benchmark. *IEEE Tran. on Im. Proc.*, 24(12):5630–5644, Dec 2015. 1

[18] M. Mueller, N. Smith, and B. Ghanem. A benchmark and simulator for uav tracking. In *Proc. European Conf. Computer Vision*, 2016. 1, 2, 4, 5, 7

[19] G. Nebehay and R. Pflugfelder. Clustering of static-adaptive correspondences for deformable object tracking. In *Comp. Vis. Patt. Recognition*, pages 2784–2791, June 2015. 5

[20] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang. Incremental learning for robust visual tracking. *Int. J. Comput. Vision*, 77(1-3):125–141, May 2008. 5

[21] A. Smeulders, D. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah. Visual tracking: An experimental survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(7):1442–1468, July 2014. 1, 2, 4, 5

[22] L. Čehovin. Trax: The visual tracking exchange protocol and library. *Neurocomputing*, 260:5 – 8, 2017. 3

[23] T. Vojir and J. Matas. The enhanced flock of trackers. In *Regist. and Recog. in Images and Videos*, volume 532 of *Studies in Comput. Intell.* January 2014. 5

[24] T. Vojir, J. Noskova, and J. Matas. Robust scale-adaptive mean-shift for tracking. In J.-K. Kämäräinen and M. Koskela, editors, *Image Analysis: 18th Scandinavian Conference, SCIA 2013, Espoo, Finland, June 17-20, 2013. Proceedings*, pages 652–663, 2013. 5

[25] Y. Wu, J. Lim, and M.-H. Yang. Online object tracking: A benchmark. In *Comp. Vis. Patt. Recognition*, pages 2411–2418, 2013. 1

[26] Y. Wu, J. Lim, and M. H. Yang. Object tracking benchmark. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(9):1834–1848, Sept 2015. 1, 2, 4, 5, 7

[27] M.-H. Y. Xu Jia, Huchuan Lu. Visual tracking via adaptive structural local sparse appearance model. In *Comp. Vis. Patt. Recognition*, pages 1822–1829, 2012. 5

[28] D. P. Young and J. M. Ferryman. Pets metrics: On-line performance evaluation service. In *ICCCN '05 Proceedings of the 14th International Conference on Computer Communications and Networks*, pages 317–324, 2005. 1

[29] O. Zendel, M. Murschitz, M. Humenberger, and W. Herzner. Cv-hazop: Introducing test data validation for computer vision. In *Int. Conf. Computer Vision*, December 2015. 1

[30] J. Zhang, S. Ma, and S. Sclaroff. MEEM: robust tracking via multiple experts using entropy minimization. In *Proc. European Conf. Computer Vision*, pages 188–203, 2014. 5

[31] K. Zhang, L. Zhang, and M.-H. Yang. Real-time compressive tracking. In *Proc. European Conf. Computer Vision*, pages 864–877, 2012. 5

[32] L. Čehovin, M. Kristan, and A. Leonardis. Robust visual tracking using an adaptive coupled-layer visual model. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(4):941–953, Apr. 2013. 5