Is First Person Vision Challenging for Object Tracking?

Matteo Dunnhofer[•] Antonino Furnari[⋆] Giovanni Maria Farinella[⋆] Christian Micheloni[•]

[•]Machine Learning and Perception Lab, University of Udine, Udine, Italy
[⋆]Image Processing Laboratory, University of Catania, Catania, Italy

Abstract

Understanding human-object interactions is fundamental in First Person Vision (FPV). Tracking algorithms which follow the objects manipulated by the camera wearer can provide useful cues to effectively model such interactions. Visual tracking solutions available in the computer vision literature have significantly improved their performance in the last years for a large variety of target objects and tracking scenarios. However, despite a few previous attempts to exploit trackers in FPV applications, a methodical analysis of the performance of state-of-the-art trackers in this domain is still missing. In this paper, we fill the gap by presenting the first systematic study of object tracking in FPV. Our study extensively analyses the performance of recent visual trackers and baseline FPV trackers with respect to different aspects and considering a new performance measure. This is achieved through TREK-150, a novel benchmark dataset composed of 150 densely annotated video sequences. Our results show that object tracking in FPV is challenging, which suggests that more research efforts should be devoted to this problem so that tracking could benefit FPV tasks.

1. Introduction

Understanding the interactions between a camera wearer and the surrounding objects is a fundamental problem in First Person Vision (FPV) [19, 87, 57, 33, 20, 73, 12, 4, 5, 13]. To model such interactions, the continuous knowledge of where an object of interest is located inside the video frame is advantageous. The benefits of tracking in FPV have been explored by a few previous works to predict future active objects [32], analyze social interactions [2], improve the performance of hand detection for rehabilitation purposes [83], locate hands and capture their movements for action recognition [44] and human-object interaction forecasting [57]. On a more abstract level, the features computed after the frame by frame localization of objects have been increasingly used for egocentric action recognition [87, 89, 62] and anticipation [33, 76, 78].

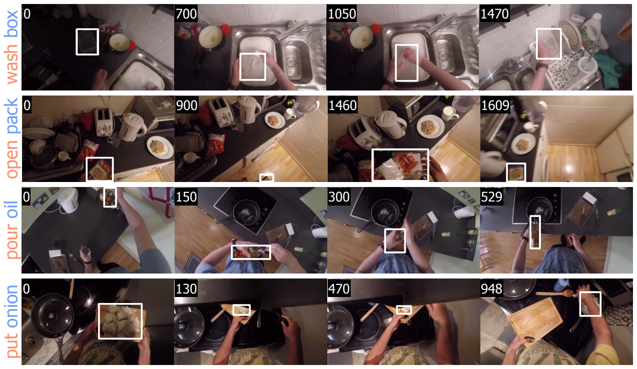


Figure 1: Qualitative examples of some sequences contained in the proposed TREK-150 benchmark dataset. The white rectangle represents the ground-truth bounding box of the target object. Each number in the top left corner identifies the frame index. For each sequence, the action performed by the camera wearer is also reported (verb in orange, noun in blue). As can be noted, objects undergo significant appearance and state changes due to the manipulation by the camera wearer, which makes the proposed setting challenging for current trackers.

Despite the aforementioned attempts to leverage tracking in egocentric vision pipelines, most approaches rely on object detection models that evaluate video frames independently. This paradigm has the drawback of ignoring all the temporal information coming from the object appearance and motion contained in consecutive video frames and generally requires a higher computational cost due to the repeated detection process on every frame. In contrast, visual object tracking aims to exploit past information about the target to infer its position and shape in the next frames of a video [63]. This process is subject to different challenges including occlusions, appearance changes, illumination variation, fast motion, and motion blur. Additionally, many practical applications pose real-time constraints to the computation, which specifically hold in FPV when the localization of objects is needed by higher-level real-time algorithms. While the use cases of object tracking in egocentric vision are manifold as previously discussed, it is clear that tracking is still not a dominant technology in the FPV field. We experimentally show that this is mainly due to the limited performance of current trackers in egocentric videos due to the involved FPV challenges such as camera motion, persistent occlusion, significant scale and state changes, as well as motion blur (see Figure 1). Due to these challenges, previous works have proposed customized approaches to track specific targets like people [3], people faces [1], or hands [44, 83, 65, 37, 81] from the FPV perspective. A solution specifically designed to track arbitrary objects in egocentric videos is still missing. Instead, the computer vision community has made significant progress in the visual tracking of generic objects. This has been possible thanks to development of new and effective tracking principles [11, 40, 7, 21, 8, 16, 98, 18], and to the careful design of benchmark datasets [91, 66, 35, 52, 31, 41] and challenges [49, 48, 50, 47]. Nowadays, the state-of-the-art tracking solutions achieve excellent results on a large variety of tracking domains [91, 66, 35, 41, 50, 47]. However, all these research endeavours have taken into account mainly the classic third person scenario in which objects are observed from an external point of view and are not manipulated by the camera wearer. Additionally, the performance of existing trackers has never been evaluated in the FPV domain, which raises the question of whether current solutions can be used "off-the-shelf" or more domain-specific investigations should be carried out.

To answer the aforementioned questions, in this paper we aim to extensively analyze the problem of visual object tracking in the FPV domain. Given the lack of suitable benchmarks, we follow the standard practice of the visual tracking community that suggests to build an accurate dataset for evaluation [91, 56, 66, 52, 35, 50, 59]. Therefore, we propose a novel visual tracking benchmark, TREK-150 (TRacking-Epic-Kitchens-150), which is obtained from the large and challenging FPV dataset EPIC-KITCHENS-55 (EK-55) [19]. TREK-150 provides 150 video sequences densely annotated with the bounding boxes of a target object the camera wearer interacts with. Additionally, sequences have been labelled with attributes that identify the visual changes the object is undergoing, the class of the target object and the action the person is performing. Using the dataset, we present an in-depth study of the accuracy and speed performance of both non-FPV and FPV visual trackers. A new performance measure is also introduced to evaluate trackers with respect to FPV scenarios.

In sum, the contributions of this paper are: (i) the first systematic analysis of visual object tracking in FPV; (ii) the description and release of the new TREK-150 dataset, which offers new challenges and complementary features with respect to existing visual tracking benchmarks; (iii) two FPV baseline trackers combining a state-of-the-art generic object tracker and FPV object detectors; (iv) a new and improved measure to assess the tracker's ability to maintain temporal reference to targets.

Our results show that FPV offers challenging tracking scenarios for the most recent and accurate trackers [18, 22, 80, 21, 9] and even for FPV trackers. Considering the potential impact of tracking on FPV, we suggest that more research efforts should be devoted to the considered task, for which we believe the proposed TREK-150 benchmark will be a key research tool. Annotations, trackers' results, and code are available at https://machinelearning.uniud.it/datasets/trek150/.

2. Related Work

Visual Tracking in FPV. There have been some attempts to tackle visual tracking in FPV. Alletto et al. [3] improved the TLD tracker [43] with a 3D odometry based module to track people. For a similar task, Nigam et al. [70] proposed a combination of the Struck [38] and MEEM [95] trackers with a person re-identification module. Face tracking was tackled by Aghaei et al. [1] through a multi-object tracking approach termed extended-bag-of-tracklets. Hand tracking was studied in several works [44, 83, 65, 37, 81]. Sun et al. [81] developed a particle filter framework for hand pose tracking. Müller et al. [65] proposed a solution based on an RGB camera and a depth sensor. Kapidis et al. [44] and Visée et al. [83] proposed to combine the YOLO [74] detector trained for hand detection with trackers. The former used the multi-object tracker DeepSORT [88], whereas the latter employed the KCF [40] single object tracker. Han et al. [37] exploited a detection-by-tracking approach on video frames acquired with 4 fisheye cameras. All the presented solutions focused on tracking specific targets (i.e., people, faces, or hands), and thus they are likely to fail in generalizing to arbitrary target objects. Moreover, they have been validated on custom designed datasets, which limits their reproducibility and the ability to compare them to other works. In contrast, we focus on the evaluation of algorithms for the generic object tracking task. We design our evaluation to be reproducible and extendable by releasing TREK-150, a dataset of 150 videos of different objects manipulated by the camera wearer, which we believe will be useful to study object tracking in FPV. To the best of our knowledge, ours is the first attempt to evaluate systematically generic object tracking in the FPV context.

Visual Tracking for Generic Settings. In recent years, there has been an increased interest in developing accurate and robust single object tracking (SOT) algorithms for generic targets and domains. Preliminary trackers were

Table 1: Statistics of the proposed TREK-150 benchmark compared with other benchmarks designed for SOT evaluation.

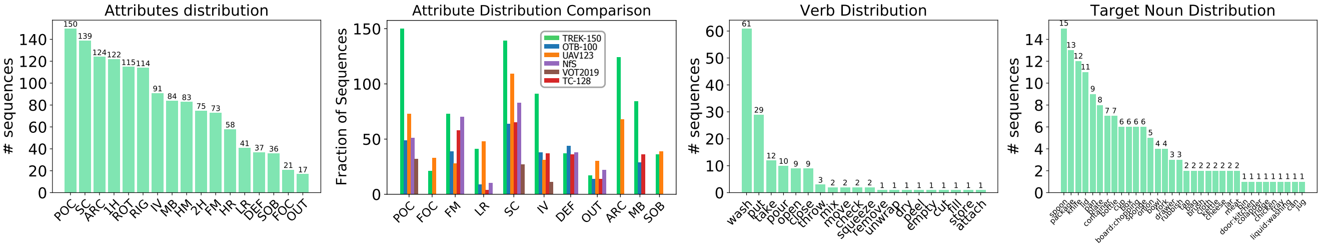
Benchmark	OTB-50	OTB-100	TC-128	UAV123	NUS-PRO	NfS	VOT2019	CDTB	TREK-150
	[90]	[91]	[56]	[66]	[52]	[35]	[50]	[59]	
# videos	51	100	128	123	365	100	60	80	150
# frames	29K	59K	55K	113K	135K	383K	20K	102K	97K
Min frames across videos	71	71	71	109	146	169	41	406	161
Mean frames across videos	578	590	429	915	371	3830	332	1274	649
Median frames across videos	392	393	365	882	300	2448	258	1179	484
Max frames across videos	3872	3872	3872	3085	5040	20665	1500	2501	4640
Frame rate	30 FPS	30 FPS	30 FPS	30 FPS	30 FPS	240 FPS	30 FPS	30 FPS	60 FPS
# target object classes	10	16	27	9	8	17	30	23	34
# sequence attributes	11	11	11	12	12	9	6	13	17
FPV	✗	✗	✗	✗	✗	✗	✗	✗	✓
# action verbs	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	20

based on mean shift algorithms [17], key-point [64], part-based methods [14, 69], or SVM learning [38]. Later, solutions based on correlation filters gained popularity thanks to their processing speed [11, 40, 23, 6, 46]. More recently, algorithms based on deep learning have been proposed to extract efficient image and object features. This kind of representation has been used in deep regression networks [39, 30], online tracking-by-detection methods [68, 80], approaches based on reinforcement learning [94, 15, 27, 36], deep discriminative correlation filters [21, 22, 8, 24, 60, 9], and trackers based on siamese networks [7, 53, 86, 16, 98]. All these methods have been designed for tracking arbitrary target objects in unconstrained domains. However, no solution has been studied and validated on a number of diverse FPV sequences as we propose in this paper.

Visual Tracking Benchmarks. Disparate bounding box level benchmarks are available today to evaluate the performance of SOT algorithms. The Object Tracking Benchmarks (OTB) OTB-50 [90] and OTB-100 [91] are two of the most popular benchmarks in the visual tracking community. They provide 51 and 100 sequences respectively including generic targets like vehicles, people, faces, toys, characters, etc. The Temple-Color 128 (TC-128) dataset [56] comprises 128 videos and was designed for the evaluation of color-enhanced trackers. The UAV123 dataset [66] was constructed to benchmark the tracking of 9 classes of target in 123 videos captured by unmanned aerial vehicle (UAV) cameras. The NUS-PRO dataset [52] contains 365 sequences and aims to benchmark human and rigid object tracking with targets belonging to one of 8 categories. The Need for Speed (NfS) dataset [35] provides 100 sequences with a frame rate of 240 FPS. The aim of the authors was to benchmark the effects of frame rate variations on the tracking performance. The VOT2019 benchmark [50] was the last iteration of the annual Visual Object Tracking challenge that required bounding-boxes as target object representation. This dataset contains 60 highly challenging videos, with generic target objects belonging to 30 different categories. The Color and Depth Tracking Benchmark (CDTB) dataset [59] offers 80 RGB sequences paired with a depth channel. This benchmark aims to explore the use of depth information to improve tracking performance. Following the increased development of deep learning based trackers, large-scale generic-domain SOT datasets have been recently released [67, 41, 31]. These include more than a thousand videos normally split into training and test subsets. The evaluation protocol associated with these sets requires the evaluation of the trackers after they have been trained on the provided training set. Despite the fact that all the presented benchmarks offer various tracking scenarios, limited work has focused on FPV, with some studies tackling the problem of tracking pedestrians or cars from a moving camera [77]. Some datasets of egocentric videos such as ADL [72] and EK-55 [19] contain bounding-box object annotations. But due to the sparse nature of such annotations (typically 1/2 FPS), these datasets cannot be used for the accurate evaluation of trackers in FPV context. To the best of our knowledge, our proposed TREK-150 dataset is the first benchmark for tracking objects which are relevant to (or manipulated by) a camera wearer in egocentric videos. We believe that TREK-150 is tantalizing for the tracking community because it offers complementary tracking situations (which we characterize with a total of 17 attributes) and new target object categories (for a total of 34) that are not present in other tracking benchmarks. Since in this paper we aim to benchmark generic approaches to visual tracking (that would not necessarily consider the deep learning approach), we follow the practice of previous works [91, 56, 66, 52, 35, 50, 59] and set up a well described dataset for evaluation of generic SOT algorithms. We believe that TREK-150 can be a useful research tool for both the FPV and visual tracking research communities.

3. The TREK-150 Benchmark Dataset

The proposed TREK-150 dataset is composed of 150 video sequences. In each, a single target object is labeled with a bounding box which encloses the visible parts of the object. The bounding boxes are given for each frame in which the object is visible (as a whole or in part). To be



(a) (b) (c) (d)

Figure 2: (a) Distribution of the sequences within TREK-150 with respect to the attributes. (b) Comparison of the distributions of common attributes in different benchmarks. Distributions of (c) action verb labels, and (d) target object categories (nouns).

Table 2: Selected sequence attributes. The first block of rows describes attributes commonly used by the visual tracking community. The last four rows describe additional attributes introduced in this paper to characterize FPV tracking sequences.

Attribute	Meaning
SC	Scale Change: the ratio of the bounding-box area of the first and the current frame is outside the range [0.5, 2]
ARC	Aspect Ratio Change: the ratio of the bounding-box aspect ratio of the first and the current frame is outside the range [0.5, 2]
IV	Illumination Variation: the area of the target bounding-box is subject to light variation
SOB	Similar Objects: there are objects in the video of the same object category or with similar appearance to the target
RIG	Rigid Object: the target is a rigid object
DEF	Deformable Object: the target is a deformable object
ROT	Rotation: the target rotates in the video
POC	Partial Occlusion: the target is partially occluded in the video
FOC	Full Occlusion: the target is fully occluded in the video
OUT	Out Of View: the target completely leaves the video frame
MB	Motion Blur: the target region is blurred due to target or camera motion
FM	Fast Motion: the target bounding-box has a motion change larger than its size
LR	Low Resolution: the area of the target bounding-box is less than 1000 pixels in at least one frame
HR	High Resolution: the area of the target bounding-box is larger than 250000 pixels in at least one frame
HM	Head Motion: the person moves their head significantly thus causing camera motion
1H	1 Hand Interaction: the person interacts with the target object with one hand for consecutive video frames
2H	2 Hands Interaction: the person interacts with the target object with both hands for consecutive video frames

compliant with other tracking challenges, every sequence is additionally labeled with one or more of 17 attributes describing the visual variability of the target in the sequence, plus two additional action verb and noun attributes indicating the action performed by the camera wearer and the class of the target. Qualitative examples of the video sequences are shown in Figure 1, whereas Table 1 reports key statistics of our dataset in comparison with existing benchmarks.[1]

Data Collection. The videos have been sampled from EK-55 [19], which is a public, large-scale, and diverse dataset of egocentric videos focused on human-object interactions in kitchens. EK-55 provides videos annotated with the actions performed by the camera wearer in the form of temporal bounds and verb-noun labels. The dataset also contains sparse bounding-box references of manipulated objects annotated at 2 frames per second in a temporal window around each action. To obtain a suitable pool of video sequences interesting for object tracking, we cross-referenced the original verb-noun temporal annotations of EK-55 to the sparse bounding box labels. This allowed to select sequences in which the camera wearer manipulates an object. Each sequence is composed of the video frames contained within the temporal bounds of the action, extracted at the original 60 FPS frame rate and at the original full HD frame size [19]. According to the authors of [19], this frame rate is necessary in FPV to contrast the fast motion and motion blur happening due to the proximity of the main scene and the camera point of view. From the initial pool, we selected 150 video sequences which were characterized by attributes such as scale changes, partial/full occlusion and fast motion, which are commonly considered in standard tracking benchmarks [91, 66, 67, 31, 50]. The top part of Table 2 reports the 13 attributes considered for the selection.

Data Labeling. After selection, the 150 sequences were associated to only 3000 bounding boxes, due to the sparse nature of the object annotations in EK-55. Since it has been shown that visual tracking benchmarks require dense and accurate annotations [50, 66, 31, 82], we re-annotated the bounding boxes of the target objects on the 150 sequences. Batches of sequences were delivered to annotators who were explicitly instructed to perform the labeling. Such initial annotations were then carefully checked and refined by a visual tracking expert. This process produced 97296 frames labeled with bounding boxes related to the position and visual presence of objects the camera wearer is interacting with. Following the initial annotations, we employed axis-aligned bounding boxes. This kind of representation is widely used in many FPV pipelines [32, 34, 33, 19, 45, 83, 79], and thus it allows us to give immediate results on the impact of trackers in such contexts. Moreover, the recent progress of trackers on various benchmarks that use this state representation [91, 66, 35, 59, 67, 31, 41] demonstrates that it provides sufficient information about the target for consistent and reliable performance evaluation.

Along with the bounding boxes, the sequences have been

[1] Please see Appendix A of the supplementary material for additional motivations and details.

labeled considering 17 attributes which define the motion and visual appearance changes the target object is subject. These include the aforementioned 13 standard tracking attributes, plus 4 additional ones (High Resolution, Head Motion, 1-Hand Interaction, 2-Hands Interaction) which have been introduced to characterize FPV sequences and are summarized in the bottom part of Table 2. Figure 2(a) reports the distributions of the sequences with respect to the 17 attributes. Figure 2(b) compares the distributions of the most common SOT attributes in TREK-150 and in other well-known benchmarks. Our dataset provides a larger number of sequences affected by partial occlusions (POC), changes in scale (SC) and/or aspect ratio (ARC), and motion blur (MB). We claim that these peculiarities, which are complementary to those of existing datasets, are due to the particular first person viewpoint, camera motion, and the human-object interactions contained in the videos. Based on EK-55's verb-noun labels, sequences were also associated to 20 verb labels (e.g., "wash" - see Figure 1) and 34 noun labels indicating the category of the target object (e.g., "box"). Figures 2(c-d) show the distributions of the videos relative to verbs and target nouns. As can be noted, TREK-150 reflects the EK-55's long-tail distribution of labels.

4. Trackers

We considered 33 trackers in our benchmark evaluation. 31 of these trackers have been selected to represent different popular approaches to SOT, for instance with respect to the matching strategy, type of image representations, learning strategy, etc. Specifically, in the analysis we have included short-term trackers [50] based on both correlation-filters with hand-crafted features (MOSSE [11], DSST [23], KCF [40], Staple [6], BACF [46], DCFNet [85], STRCF [54], MCCTH [84]) and deep features (ECO [21], ATOM [22], DiMP [8], PrDiMP [24], KYS [9]). We also considered deep siamese networks (SiamFC [7], GOTURN [39], DSLT [58], SiamRPN++ [53], SiamDW [97], UpdateNet [96], SiamFC++ [92], SiamBAN [16], Ocean [98]), tracking-by-detection methods (MDNet [68], VITAL [80]), as well as trackers based on target segmentation representations (SiamMask [86], D3S [60]), meta-learning (MetaCrest [71]), and fusion strategies (TRASFUST [29]). The long-term [50] trackers SPLT [93], GlobalTrack [42], and LTMU [18] have been also taken into account in the study. These trackers are designed to address longer target occlusion and out of view periods by exploiting object re-detection modules. All of the selected trackers are state-of-the-art approaches published between the years 2010-2020.

In addition to the aforementioned generic object trackers, we developed 2 baseline FPV trackers that combine the LTMU tracker [18] with (i) the EK-55 trained Faster-R-CNN [19] and (ii) the Faster-R-CNN-based hand-object detector [79]. We refer to them as LTMU-F and LTMU-H respectively. These baseline trackers exploit the respective detectors as object re-detection modules according to the LTMU scheme [18]. In short, the re-detection happens when a verification module notices that the tracker is not following the correct target. In such a case, the module triggers the execution of the respective FPV detector which proposes candidate locations of the target object. Each of the candidates is evaluated by the verification module, and the location with highest confidence is used to re-initialize the tracker.[2] The two modules implement conceptually different strategies for FPV-based object localization. The first aims to find objects in the scene, while the second looks for the interaction between the camera wearer and objects.

5. Evaluation

Evaluation Protocols. We employed three standard protocols to perform our analysis.[3] The first is the one-pass evaluation (OPE) protocol detailed in [91], which implements the most realistic way to execute trackers. It consists in initializing a tracker with the ground-truth bounding box of the target in the first frame and let the tracker run on every subsequent frame until the end of the sequence.

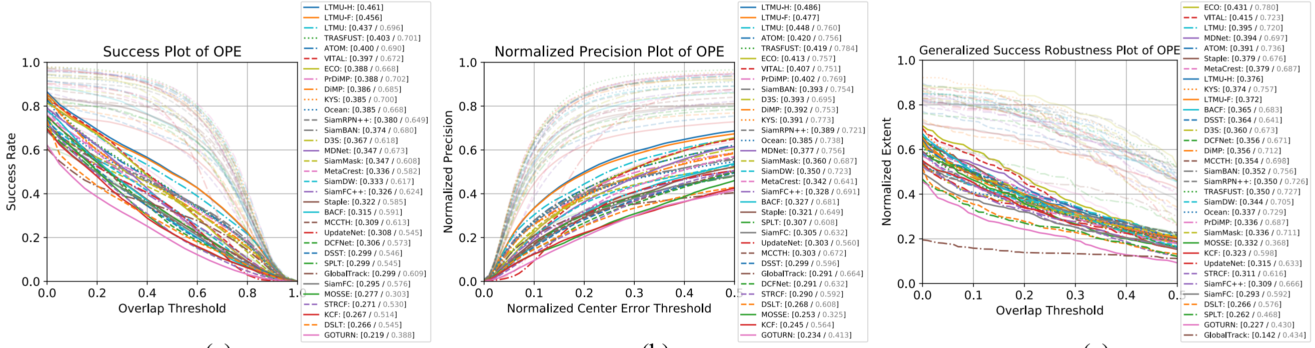
To obtain a more robust evaluation [51], especially for the analysis over sequence attributes and action verbs, we employ the recent protocol of [47] which defines different points of initialization along a sequence. A tracker is initialized with the ground-truth in each point and let run either forward or backward in time (depending on the longest sub-sequence yielded by the initialization point) until the end of the sub-sequence. This protocol allows a tracker to better cover all the situations happening in the sequences, ultimately leading to more robust evaluation scores. We refer to this setup as multi-start evaluation (MSE).

Since many FPV tasks such as object interaction [20] and early action recognition [34], or action anticipation [19], require real-time computation, we evaluated the ability of trackers to provide their object localization in such a setting. This was achieved by following the details given in [48, 55]. In short, this protocol, which we refer to as RTE, runs an algorithm considering its running time. The protocol skips all the frames, considered to occur regularly according to the frame rate, which appeared during the interval between the algorithm's execution start and end times.

Performance Measures. To quantitatively assess the performance of the trackers on the proposed dataset, we used different measures that compare all tracker's predicted bounding boxes with respect to the temporally aligned ground-truth bounding boxes. To evaluate the localization accuracy of the trackers, we employ the success plot

[2] More details are given in Appendix B of the supplementary material.

[3] See Appendix C of the supplementary material for further details.



(a) (b) (c)

Figure 3: Performance of the selected trackers on the proposed TREK-150 benchmark under the OPE protocol. The curves in solid colors report the performance of the 33 benchmarked trackers on TREK-150, whereas the curves overlaid in semi-transparent colors outline the performance obtained by the same trackers on the standard OTB-100 [91] dataset. In brackets, next to the trackers' names, we report the SS, NPS and GR values achieved on TREK-150 (in black) and on OTB-100 [91] (in gray). As can be noted, all the trackers exhibit a significant performance drop when tested on our challenging FPV benchmark. LTMU-H and LTMU-F achieve marginally better performance, while we expect significant boosts to be achievable with a careful design of FPV trackers.

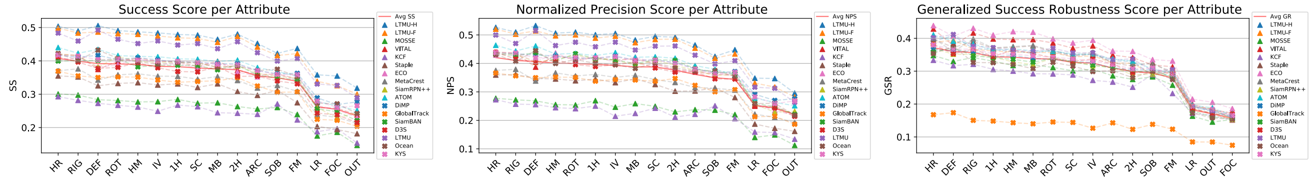


Figure 4: SS, NPS, and GSR of 17 of the benchmarked trackers on the sequence attributes of proposed TREK-150 benchmark under the MSE protocol. The red plain line highlights the average performance. (The results for POC are not reported because this attribute is present in every sequence).

[91], which shows the percentage of predicted bounding boxes whose intersection-over-union with the ground-truth is larger than a threshold varied from 0 to 1 (Figure 3 (a)). We also use the normalized precision plot [67], that reports, for a variety of distance thresholds, the percentage of bounding boxes whose center points are within a given normalized distance (in pixels) from the ground-truth (Figure 3 (b)). As summary measures, we report the success score (SS) [91] and normalized precision scores (NPS) [67], which are computed as the Area Under the Curve (AUC) of the success plot and normalized precision plot respectively.

Along with these standard metrics, we employ a novel plot which we refer to as generalized success robustness plot (Figure 3 (c)). We take inspiration from the robustness metric proposed in [47] which measures the normalized extent of a tracking sequence before a failure. But differently from [47], which uses a fixed overlap threshold to detect a collapse, we propose to use different thresholds ranging in [0, 0.5]. This allows to assess the length of tracking sequences for different application scenarios. We consider 0.5 as the maximum threshold as higher overlaps are usually associated to positive predictions in many computer vision tasks. Similarly as [91, 67], we use the AUC of the generalized robustness plot to obtain an aggregate score which we refer to as generalized success robustness (GSR). This new measure evaluates trackers' capability of maintaining long temporal reference to targets. We think this aspect is especially important in FPV as longer references to the target can lead to a better modeling of the camera viewer's actions and interactions with objects.

Finally, we evaluate the trackers' processing speed in frames per second (FPS) to quantify their efficiency.

6. Results

How Do the Trackers Perform in the FPV Scenario? Figure 3 reports the performance of the selected trackers on TREK-150 using the OPE protocol. For reference, we also report the performance of the trackers on the popular OTB-100 [91] benchmark (semi-transparent curves - gray numbers in brackets). It can be clearly noted that the overall performance of the trackers is decreased across all measures when considering the challenging FPV scenario of TREK-150. For example, the SS, NPS, and GSR scores of LTMU on TREK-150 are 43.7% , 44.8%, and 43.1%, which

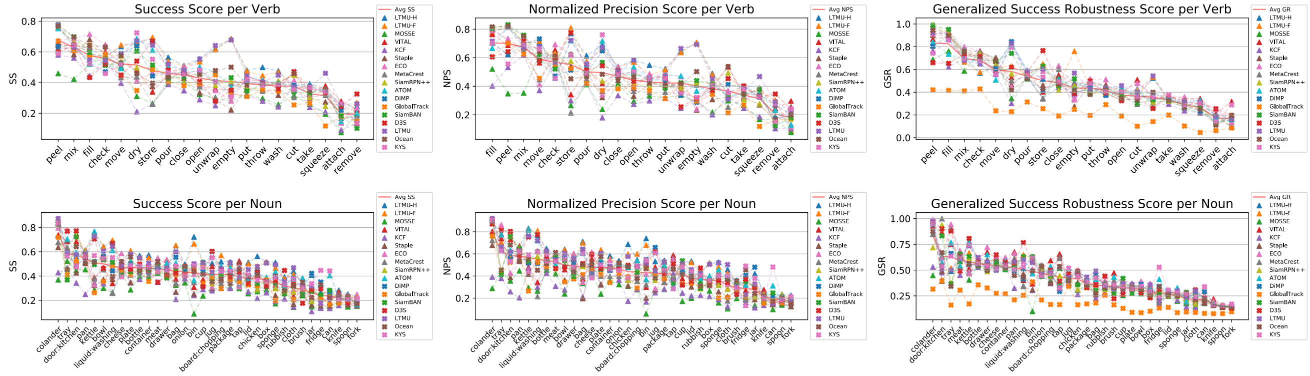


Figure 5: SS, NPS, and GSR performance of 17 among the 33 selected trackers with respect to the action verbs (first row of plots) and target nouns (second row of plots) in TREK-150. The red plain line highlights the average performance.

Table 3: Performance achieved by 17 of the benchmarked trackers on TREK-150 using the RTE protocol.

Metric	Ocean	SiamBAN	SiamRPN++	DiMP	KYS	ATOM	LTMU	D3S	ECO	GlobalTrack	Staple	MOSSE	LTMU-H	MetaCrest	LTMU-F	VITAL	KCF
FPS	21	24	23	16	12	15	8	16	15	8	13	26	4	8	4	4	6
SS	0.365	0.360	0.362	0.336	0.327	0.319	0.284	0.276	0.252	0.253	0.249	0.227	0.213	0.207	0.205	0.204	0.186
NPS	0.358	0.366	0.356	0.331	0.317	0.312	0.257	0.263	0.231	0.227	0.236	0.190	0.174	0.175	0.161	0.165	0.157
GSR	0.294	0.313	0.293	0.224	0.237	0.179	0.169	0.182	0.173	0.139	0.169	0.141	0.161	0.165	0.162	0.158	0.177

are much lower than the respective 69.6%, 76%, and 78%, achieved on OTB-100. With the MSE protocol, LTMU achieves the respective scores of 46.9%, 48.3%, 38.6%.[4] These results show that the particular characteristics of FPV present in TREK-150 introduce challenging scenarios for visual trackers. Some qualitative examples of the trackers' performance are shown in Figure 11 of the Appendix.

Generally speaking, trackers based on deep learning (e.g. LTMU, TRASFUST, ATOM, KYS, Ocean) perform better in SS and NPS than those based on hand-crafted features (e.g. BACF, MCCTH, DSST, KCF). Among the first class of trackers, the ones leveraging online adaptation mechanisms (e.g. LTMU, ATOM, VITAL, ECO, KYS, DiMP) are more accurate than the ones based on single-shot instances (e.g. Ocean, D3S, SiamRPN++). The generalized success robustness plot in Figure 3(c) and the GSR results of Figure 10 of the supplementary report a different rankings of the trackers, showing that more spatially accurate trackers are not always able to maintain longer reference to targets.

Under both the OPE and MSE protocols, the proposed FPV trackers LTMU-H and LTMU-F are largely better in SS and NPS, while they lose some performance in GSR. Such outcome shows that adapting a state-of-the-art method to FPV allows to marginally improve results, while we expect significant performance improvements to be achievable by a tracker accurately designed to tackle the FPV challenges introduced by this benchmark.

In Which Conditions Do the Trackers Work Better? Figure 4 reports the SS, NPS, and GSR scores, computed with the MSE protocol, of 17 trackers with respect to the attributes introduced in Table 2.[5] We do not report results for the POC attribute as it is present in every sequence, as shown in Figure 2 (a). It stands out clearly that full occlusion (FOC), out of view (OUT) and the small size of targets (LR) are the most difficult situations for trackers. The fast motion of targets (FM) and the presence of similar objects (SOB) are also critical factors that cause drops in performance. Trackers show to be less vulnerable to rotations (ROT) and to the illumination variation (IV). Generally, tracking rigid objects (RIG) results easier than tracking deformable ones (DEF). With respect to the new 4 sequence attributes related to FPV, it results that tracking objects held with two hands (2H) is more difficult than tracking objects held with a single hand (1H). This is probably due to the additional occlusions generated in the 2H scenario. Trackers are instead quite robust to head motion (HM) and seem to cope better with objects appearing in larger size (HR).

How Do the Trackers Perform With Respect to the Actions? The first row of plots in Figure 5 reports the MSE protocol results of SS, NPS, and GSR with respect to the associated verb action labels.[5] Actions that mainly cause a spatial displacement of the target (e.g. "move", "store", "check") generally have less impact on the performance. Actions that change the state, shape, or aspect ratio of an object (e.g. "remove", "squeeze", "cut", "attach") generate harder tracking scenarios. Also the sequences characterized by the "wash" verb lead trackers to poor performance. In-

[4] See Appendix D for the overall MSE results of all trackers.

[5] The analysis was restricted to 17 trackers for a better visualization of the plots/tables. The 17 trackers were selected to represent various methodologies. The results for all trackers are available in Appendix D.

Table 4: Accuracy results on TREK-150 of a video-based hand-object detection solution which considers each of the considered trackers as localization method for the object involved in the interaction. As a baseline, we employ the object detection capabilities of the hand-object interaction solution Hands-in-contact [79].

Hands-in-contact [79]	LTMU-H	LTMU-F	ATOM	LTMU	Ocean	SiamBAN	SiamRPN++	MetaCrest	D3S	DiMP	KYS	VITAL	GlobalTrack	MOSSE	ECO	Staple	KCF
0.354	0.368	0.367	0.361	0.354	0.340	0.340	0.311	0.293	0.292	0.292	0.279	0.253	0.251	0.231	0.230	0.197	0.177

deed, the wash action can cause many occlusions and make the object harder to track.

The second row of the same figure presents the performance scores of the trackers with respect to the associated noun labels. Rigid, regular-sized objects such as "pan", "kettle", "bowl", "plate", and "bottle" are among the ones associated with high average scores. On the other hand, some rigid objects such as "knife", "spoon", "fork" and "can" are harder to track, probably due to their particularly thin shape and the light reflectance they are easily subject to. Deformable objects such as "sponge", "onion", "cloth" and "rubbish" are in general also difficult to track.

How Fast Are the Trackers? Table 3 reports the FPS performance of the trackers and the SS, NPS, and GSR scores achieved under the RTE protocol.[5] None of the trackers achieve the frame rate speed of 60 FPS. We argue that this is due the full HD resolution of frames which requires demanding image crop and resize operations with targets of considerable size. Thanks to their non-reliance of online adaptation mechanisms, trackers based on siamese networks (e.g. Ocean, SiamBAN, SiamRPN++) emerge as the fastest trackers and exhibit a less significant performance drop of the proposed scores. Trackers using online learning approaches (e.g. ATOM, DiMP, ECO, KYS) generally achieve a below real-time speed, consequently causing a major accuracy loss when deployed to real-time scenarios. In general, we observe that the GSR score is the measure on which all trackers present the major drop in the real-time setting, suggesting that particular effort should be spent to better model actions and interactions in such scenarios.

Do Trackers Already Offer Any Advantage in FPV? Despite we are demonstrating that FPV is challenging for current trackers, we assess whether these already offer an advantage in the FPV domain to obtain information about the objects' locations and movements in the scene [87, 32, 33, 78, 79]. To this aim, we performed two experiments.[6] First, we evaluated the performance of a Faster R-CNN [75] instance trained on EK-55 [19] when used as a naive tracking baseline. Such a solution achieves an SS, NPS, and GSR of 0.323, 0.369, 0.044, by running at 1 FPS. Comparing these results with the ones presented in Figure 3, we clearly notice that trackers, if properly initialized by a detection module, can deliver faster, more accurate and much more temporally long object localization than detectors.

As a second experiment, we evaluated the accuracy of a video-based hand-object interaction detection solution [79] whose object localisation is given by a tracker rather than a detector. The tracker is initialized with the object detector's predicted bounding-box at the first detection of the hand-object interaction, and let run until its end. By this setting, we created a ranking of the trackers which is presented in Table 4. The results demonstrate that stronger trackers can improve the accuracy and efficiency of current detection-based methodologies [79]. Interestingly, the trackers' ranking differs from what shown in Figure 3, suggesting that trackers can manifest other capabilities when deployed into application scenarios.

Given these preliminary results, we hence expect that trackers will likely gain more importance in FPV as new methodologies explicitly considering the first person point of view are investigated.

7. Conclusions

In this paper, we proposed the first systematic evaluation of visual object tracking in FPV. The analysis has been conducted with standard and novel measures on the newly introduced TREK-150 benchmark, which contains 150 video sequences extracted from the EK-55 [19] FPV dataset. TREK-150 has been densely annotated with 97K bounding-boxes, 17 sequence attributes, 20 action verb attributes and 34 target object attributes. The performance of 31 state-of-the-art visual trackers and two baseline FPV trackers was analysed extensively on the proposed dataset. The results show a generalized drop in accuracy with respect to the performance achieved on existing tracking benchmarks. Furthermore, our analysis provided insights about which scenarios and actions cause the performance to change. Finally, we have shown that object tracking gives an advantage in terms of object localization accuracy and efficiency over object detection. These results suggest that FPV is a challenging scenario for current trackers and that tracking will likely get more importance in this domain as new FPV-specific solutions will be investigated. Annotations, results, and code, are available at https://machinelearning.uniud.it/datasets/trek150/.

Acknowledgements. Research at the University of Udine has been supported by the ACHIEVE-ITN H2020 project. Research at the University of Catania has been supported by MIUR AIM - Attrazione e Mobilita Internazionale Linea 1 - AIM1893589 - CUP: E64118002540007.

[6] Details are given in the Appendix C of the supplementary material.

References

[1] Maedeh Aghaei, Mariella Dimiccoli, and Petia Radeva. Multi-face tracking by extended bag-of-tracklets in egocentric photo-streams. *Computer Vision and Image Understanding*, 149:146–156, aug 2016. 2

[2] Maedeh Aghaei, Mariella Dimiccoli, and Petia Radeva. With whom do i interact? Detecting social interactions in egocentric photo-streams. In *Proceedings - International Conference on Pattern Recognition*, volume 0, pages 2959–2964. Institute of Electrical and Electronics Engineers Inc., jan 2016. 1

[3] Stefano Alletto, Giuseppe Serra, and Rita Cucchiara. Egocentric object tracking: an odometry-based solution. In *International Conference on Image Analysis and Processing*, pages 687–696. Springer, 2015. 2

[4] Gedas Bertasius, Hyun Soo Park, Stella X. Yu, and Jianbo Shi. First-person action-object detection with egonet. In *Proceedings of Robotics: Science and Systems*, July 2017. 1

[5] Gedas Bertasius, Hyun Soo Park, Stella X Yu, and Jianbo Shi. Unsupervised learning of important objects from first-person videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1956–1964, 2017. 1

[6] Luca Bertinetto, Jack Valmadre, Stuart Golodetz, Ondrej Miksik, and Philip H.S. Torr. Staple: Complementary learners for real-time tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2016-Decem, pages 1401–1409, 2016. 3, 5, 14, 17

[7] Luca Bertinetto, Jack Valmadre, João F. Henriques, Andrea Vedaldi, and Philip H.S. Torr. Fully-convolutional siamese networks for object tracking. *European Conference on Computer Vision*, 9914 LNCS:850–865, 2016. 2, 3, 5, 14, 17, 20

[8] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning Discriminative Model Prediction for Tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 2, 3, 5, 14, 17

[9] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Know Your Surroundings: Exploiting Scene Information for Object Tracking. In *European Conference on Computer Vision*, mar 2020. 2, 3, 5, 14, 17

[10] Goutam Bhat, Felix Järemo Lawin, Martin Danelljan, Andreas Robinson, Michael Felsberg, Luc Van Gool, and Radu Timofte. Learning what to learn for video object segmentation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 777–794, Cham, 2020. Springer International Publishing. 14

[11] David S Bolme, J Ross Beveridge, Bruce A Draper, and Yui Man Lui. Visual object tracking using adaptive correlation filters. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2544–2550. IEEE, 2010. 2, 3, 5, 14, 17

[12] Minjie Cai, Kris M Kitani, and Yoichi Sato. Understanding hand-object manipulation with grasp types and object attributes. In *Robotics: Science and Systems*, volume 3. Ann Arbor, Michigan;, 2016. 1

[13] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. *arXiv preprint arXiv:2012.09856*, 2020. 1

[14] Luka Čehovin, Matej Kristan, and Aleš Leonardis. Robust visual tracking using an adaptive coupled-layer visual model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(4):941–953, 2013. 3

[15] Boyu Chen, Dong Wang, Peixia Li, Shuang Wang, and Huchuan Lu. Real-time 'Actor-Critic' Tracking. In *European Conference on Computer Vision*, pages 318–334, 2018. 3

[16] Zedu Chen, Bineng Zhong, Guorong Li, Shengping Zhang, and Rongrong Ji. Siamese Box Adaptive Network for Visual Tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6667–6676, 2020. 2, 3, 5, 14, 17, 20

[17] Dorin Comaniciu, Visvanathan Ramesh, and Peter Meer. Real-time tracking of non-rigid objects using mean shift. *IEEE Conference on Computer Vision and Pattern Recognition*, 2:142–149, 2000. 3

[18] Kenan Dai, Yunhua Zhang, Dong Wang, Jianhua Li, Huchuan Lu, and Xiaoyun Yang. High-Performance Long-Term Tracking With Meta-Updater. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6297–6306. Institute of Electrical and Electronics Engineers (IEEE), aug 2020. 2, 5, 14, 17, 18

[19] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018. 1, 2, 3, 4, 5, 8, 14, 15, 17, 18, 19

[20] Dima Damen, Teesid Leelasawassuk, and Walterio Mayol-Cuevas. You-do, i-learn: Egocentric unsupervised discovery of objects and their modes of interaction towards video-based guidance. *Computer Vision and Image Understanding*, 149:98–112, 2016. 1, 5

[21] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. ECO: Efficient Convolution Operators for Tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, nov 2017. 2, 3, 5, 14, 17

[22] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. ATOM: Accurate Tracking by Overlap Maximization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2, 3, 5, 14, 17

[23] Martin Danelljan, Gustav Hager, Fahad Shahbaz Khan, and Michael Felsberg. Discriminative Scale Space Tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(8):1561–1575, 2017. 3, 5, 14, 17

[24] Martin Danelljan, Luc Van Gool, and Radu Timofte. Probabilistic Regression for Visual Tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7181–7190, 2020. 3, 5, 14, 17

[25] Patrick Dendorfer, Aljosa Osep, Anton Milan, Konrad Schindler, Daniel Cremers, Ian Reid, Stefan Roth, and Laura Leal-Taixé. Motchallenge: A benchmark for single-camera multiple target tracking. *International Journal of Computer Vision*, 129(4):845–881, 2021. 15

[26] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 20

[27] Matteo Dunnhofer, Niki Martinel, Gian Luca Foresti, and Christian Micheloni. Visual Tracking by means of Deep Reinforcement Learning and an Expert Demonstrator. In *Proceedings of The IEEE/CVF International Conference on Computer Vision Workshops*, 2019. 3

[28] Matteo Dunnhofer, Niki Martinel, and Christian Micheloni. An exploration of target-conditioned segmentation methods for visual object trackers. In Adrien Bartoli and Andrea Fusiello, editors, *Computer Vision – ECCV 2020 Workshops*, pages 618–636, Cham, 2020. Springer International Publishing. 14

[29] Matteo Dunnhofer, Niki Martinel, and Christian Micheloni. Tracking-by-Trackers with a Distilled and Reinforced Model. In *Asian Conference on Computer Vision*, 2020. 5, 17

[30] Matteo Dunnhofer, Niki Martinel, and Christian Micheloni. Weakly-supervised domain adaptation of deep regression trackers via reinforced knowledge distillation. *IEEE Robotics and Automation Letters*, 6(3):5016–5023, 2021. 3, 22

[31] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. LaSOT: A High-quality Benchmark for Large-scale Single Object Tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, sep 2019. 2, 3, 4, 14, 15, 20

[32] Antonino Furnari, Sebastiano Battiato, Kristen Grauman, and Giovanni Maria Farinella. Next-active-object prediction from egocentric videos. *Journal of Visual Communication and Image Representation*, 49:401–411, nov 2017. 1, 4, 8, 14, 18

[33] Antonino Furnari and Giovanni Farinella. Rolling-Unrolling LSTMs for Action Anticipation from First-Person Video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, may 2020. 1, 4, 8, 14, 18

[34] Antonino Furnari and Giovanni Maria Farinella. What Would You Expect? Anticipating Egocentric Actions With Rolling-Unrolling LSTMs and Modality Attention. In *IEEE/CVF International Conference on Computer Vision*, 2019. 4, 5, 14, 18

[35] Hamed Kiani Galoogahi, Ashton Fagg, Chen Huang, Deva Ramanan, and Simon Lucey. Need for Speed: A Benchmark for Higher Frame Rate Object Tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2017-Octob, pages 1134–1143. Institute of Electrical and Electronics Engineers Inc., mar 2017. 2, 3, 4, 14, 15

[36] Qing Guo, Ruize Han, Wei Feng, Zhihao Chen, and Liang Wan. Selective Spatial Regularization by Reinforcement Learned Decision Making for Object Tracking. *IEEE Transactions on Image Processing*, 29:2999–3013, 2020. 3

[37] Shangchen Han, Beibei Liu, Randi Cabezas, Christopher D. Twigg, Peizhao Zhang, Jeff Petkau, Tsz Ho Yu, Chun Jung Tai, Muzaffer Akbay, Zheng Wang, Asaf Nitzan, Gang Dong, Yuting Ye, Lingling Tao, Chengde Wan, and Robert Wang. MEgATrack: Monochrome Egocentric Articulated Hand-Tracking for Virtual Reality. *ACM Transactions on Graphics*, 39(4):13, jul 2020. 2

[38] Sam Hare, Stuart Golodetz, Amir Saffari, Vibhav Vineet, Ming Ming Cheng, Stephen L Hicks, and Philip H.S. Torr. Struck: Structured Output Tracking with Kernels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(10):2096–2109, 2016. 2, 3

[39] David Held, Sebastian Thrun, and Silvio Savarese. Learning to Track at 100 FPS with Deep Regression Networks. *European Conference on Computer Vision*, abs/1604.0, 2016. 3, 5, 14, 17

[40] Joao F. Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. High-speed tracking with kernelized correlation filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):583–596, 2015. 2, 3, 5, 14, 17

[41] Lianghua Huang, Xin Zhao, and Kaiqi Huang. GOT-10k: A Large High-Diversity Benchmark for Generic Object Tracking in the Wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, oct 2019. 2, 3, 4, 14, 15, 20

[42] Lianghua Huang, Xin Zhao, and Kaiqi Huang. GlobalTrack: A Simple and Strong Baseline for Long-term Tracking. In *AAAI Conference on Artificial Intelligence*, dec 2020. 5, 14, 17

[43] Zdenek Kalal, Krystian Mikolajczyk, and Jiri Matas. Tracking-learning-detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(7):1409–1422, 2012. 2

[44] Georgios Kapidis, Ronald Poppe, Elsbeth Van Dam, Lucas Noldus, and Remco Veltkamp. Egocentric hand track and object-based human action recognition. In *2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pages 922–929. IEEE, 2019. 1, 2

[45] Georgios Kapidis, Ronald Poppe, Elsbeth Van Dam, Lucas Noldus, and Remco Veltkamp. Egocentric hand track and object-based human action recognition. In *Proceedings - 2019 IEEE SmartWorld, Ubiquitous Intelligence and Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Internet of People and Smart City Innovation, SmartWorld/UIC/ATC/SCALCOM/IOP/SCI 2019*, pages 922–929. Institute of Electrical and Electronics Engineers Inc., may 2019. 4, 14

[46] Hamed Kiani Galoogahi, Ashton Fagg, and Simon Lucey. Learning Background-Aware Correlation Filters for Visual Tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 3, 5, 14, 17

[47] Matej Kristan, Aleš Leonardis, Jiří Matas, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kämäräinen, Martin Danelljan, Luka Čehovin Zajc, Alan Lukežič, Ondrej Drbohlav, Linbo He, Yushan Zhang, Song Yan, Jinyu Yang, Gustavo Fernández, et al. The eighth visual object tracking vot2020 challenge results. In Adrien Bartoli and Andrea

Fusiello, editors, *Computer Vision – ECCV 2020 Workshops*, pages 547–601, Cham, 2020. Springer International Publishing. 2, 5, 6, 18, 19

[48] Matej Kristan, Ales Leonardis, Jiri Matas, Michael Felsberg, Roman Pflugfelder, Luka Cehovin Zajc, Tomas Vojir, Gustav Hager, Alan Lukezic, Abdelrahman Eldesokey, Gustavo Fernandez, et al. The Visual Object Tracking VOT2017 Challenge Results. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1949–1972. IEEE, oct 2017. 2, 5, 19

[49] Matej Kristan, Jiri Matas, Ales Leonardis, Michael Felsberg, Luka Cehovin, Gustavo Fernandez, Tomas Vojir, Gustav Hager, Georg Nebehay, Roman Pflugfelder, et al. The Visual Object Tracking VOT2015 Challenge Results. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 564–586. IEEE, dec 2015. 2

[50] Matej Kristan, Jiří Matas, Aleš Leonardis, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kämäräinen, LukaČehovinLuka˘LukaČehovin Zajc, Ondrej Drbohlav, Alan Lukežič, Amanda Berg, Abdelrahman Eldesokey, Jani Käpylä, Gustavo Fernández, et al. The Seventh Visual Object Tracking VOT2019 Challenge Results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019. 2, 3, 4, 5

[51] M. Kristan, J. Matas, A. Leonardis, T. Vojíř, R. Pflugfelder, G. Fernández, G. Nebehay, F. Porikli, and L. Čehovin. A novel performance evaluation methodology for single-target trackers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(11):2137–2155, 2016. 5, 18

[52] Annan Li, Min Lin, Yi Wu, Ming Hsuan Yang, and Shuicheng Yan. NUS-PRO: A New Visual Tracking Challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):335–349, feb 2016. 2, 3, 15

[53] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. SIAMRPN++: Evolution of siamese visual tracking with very deep networks. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2019-June, pages 4277–4286, 2019. 3, 5, 14, 17, 20

[54] Feng Li, Cheng Tian, Wangmeng Zuo, Lei Zhang, and Ming Hsuan Yang. Learning Spatial-Temporal Regularized Correlation Filters for Visual Tracking. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 4904–4913. IEEE Computer Society, mar 2018. 5, 14, 17

[55] Mengtian Li, Yu-Xiong Wang, and Deva Ramanan. Towards streaming perception. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 473–488, Cham, 2020. Springer International Publishing. 5, 19

[56] Pengpeng Liang, Erik Blasch, and Haibin Ling. Encoding Color Information for Visual Tracking: Algorithms and Benchmark. *IEEE Transactions on Image Processing*, 24(12):5630–5644, dec 2015. 2, 3

[57] Miao Liu, Siyu Tang, Yin Li, and James Rehg. Forecasting human object interaction: Joint prediction of motor attention and actions in first person video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 2, 2020. 1

[58] Xiankai Lu, Chao Ma, Bingbing Ni, Xiaokang Yang, Ian Reid, and Ming Hsuan Yang. Deep regression tracking with shrinkage loss. In *European Conference on Computer Vision*, volume 11218 LNCS, pages 369–386, 2018. 5, 17

[59] Alan Lukezic, Ugur Kart, Jani Kapyla, Ahmed Durmush, Joni Kristian Kamarainen, Jiri Matas, and Matej Kristan. CDTB: A color and depth visual object tracking dataset and benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2019-Octob, pages 10012–10021. Microsoft Research Asia, jul 2019. 2, 3, 4, 14

[60] Alan Lukežič, Jiří Matas, and Matej Kristan. D3S – A Discriminative Single Shot Segmentation Tracker. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, nov 2020. 3, 5, 14, 17

[61] Alan Lukežič, Luka Čehovin Zajc, Tomáš Vojíř, Jiří Matas, and Matej Kristan. Now you see me: evaluating performance in long-term visual tracking, 2018. 16, 17

[62] Minghuang Ma, Haoqi Fan, and Kris M Kitani. Going deeper into first-person activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1894–1903, 2016. 1

[63] Dr Emilio Maggio and Dr Andrea Cavallaro. *Video Tracking: Theory and Practice*. Wiley Publishing, 1st edition, 2011. 1

[64] Mario Edoardo Maresca and Alfredo Petrosino. MATRIOSKA: A multi-level approach to fast tracking by learning. In *International Conference on Image Analysis and Processing*, volume 8157 LNCS, pages 419–428, 2013. 3

[65] Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. Real-Time Hand Tracking Under Occlusion from an Egocentric RGB-D Sensor. In *Proceedings - 2017 IEEE International Conference on Computer Vision Workshops, ICCVW 2017*, volume 2018-Janua, pages 1284–1293, 2017. 2

[66] Matthias Mueller, Neil Smith, and Bernard Ghanem. A Benchmark and Simulator for UAV Tracking. In *European Conference on Computer Vision*, pages 445–461. Springer, Cham, 2016. 2, 3, 4, 14, 15

[67] Matthias Müller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. TrackingNet: A Large-Scale Dataset and Benchmark for Object Tracking in the Wild. In *European Conference on Computer Vision*, volume 11205 LNCS, pages 310–327. Springer Verlag, mar 2018. 3, 4, 6, 14, 15, 18, 20

[68] Hyeonseob Nam and Bohyung Han. Learning Multi-domain Convolutional Neural Networks for Visual Tracking. *IEEE Conference on Computer Vision and Pattern Recognition*, 2016-Decem:4293–4302, 2016. 3, 5, 14, 17

[69] Hyeonseob Nam, Seunghoon Hong, and Bohyung Han. Online graph-based tracking. In *European Conference on Computer Vision*, volume 8693 LNCS, pages 112–126. Springer Verlag, 2014. 3

[70] Jyoti Nigam and Renu M Rameshan. EgoTracker: Pedestrian Tracking with Re-identification in Egocentric Videos. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, volume 2017-July, pages 980–987, 2017. 2

[71] Eunbyung Park and Alexander C. Berg. Meta-tracker: Fast and Robust Online Adaptation for Visual Object Trackers. In *European Conference on Computer Vision*, volume 11207 LNCS, pages 587–604. Springer Verlag, jan 2018. 5, 14, 17

[72] Hamed Pirsiavash and Deva Ramanan. Detecting activities of daily living in first-person camera views. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2847–2854. IEEE, 2012. 3

[73] Francesco Ragusa, Antonino Furnari, Salvatore Livatino, and Giovanni Maria Farinella. The meccano dataset: Understanding human-object interactions from egocentric videos in an industrial-like domain. In *IEEE Winter Conference on Application of Computer Vision (WACV)*, 2020. 1, 18

[74] Joseph Redmon, Santosh Kumar Divvala, Ross B Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016. 2

[75] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015. 8

[76] Ivan Rodin, Antonino Furnari, Dimitrios Mavroedis, and Giovanni Maria Farinella. Predicting the future from first person (egocentric) vision: A survey. *Computer Vision and Image Understanding*, 2021. 1

[77] Ricardo Sanchez-Matilla and Andrea Cavallaro. A predictor of moving objects for first-person vision. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 2189–2193. IEEE, 2019. 3

[78] Fadime Sener, Dipika Singhania, and Angela Yao. Temporal aggregate representations for long-range video understanding. In *European Conference on Computer Vision*, pages 154–171. Springer, 2020. 1, 8

[79] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F. Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 4, 5, 8, 14, 17, 18, 19, 20

[80] Yibing Song, Chao Ma, Xiaohe Wu, Lijun Gong, Linchao Bao, Wangmeng Zuo, Chunhua Shen, Rynson W.H. Lau, and Ming Hsuan Yang. VITAL: VIsual Tracking via Adversarial Learning. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 8990–8999. IEEE Computer Society, apr 2018. 2, 3, 5, 14, 17

[81] Li Sun, Ulrich Klank, and Michael Beetz. EYEWATCHME-3D Hand and object tracking for inside out activity analysis. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 9–16. Institute of Electrical and Electronics Engineers (IEEE), mar 2010. 2

[82] Jack Valmadre, Luca Bertinetto, João F. Henriques, Ran Tao, Andrea Vedaldi, Arnold W.M. Smeulders, Philip H.S. Torr, and Efstratios Gavves. Long-Term Tracking in the Wild: A Benchmark. In *European Conference on Computer Vision*, volume 11207 LNCS, pages 692–707. Springer Verlag, mar 2018. 4

[83] Ryan J. Visee, Jirapat Likitlersuang, and Jose Zariffa. An Effective and Efficient Method for Detecting Hands in Egocentric Videos for Rehabilitation Applications. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(3):748–755, mar 2020. 1, 2, 4, 14

[84] Ning Wang, Wengang Zhou, Qi Tian, Richang Hong, Meng Wang, and Houqiang Li. Multi-cue Correlation Filters for Robust Visual Tracking. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 4844–4853, 2018. 5, 14, 17

[85] Qiang Wang, Jin Gao, Junliang Xing, Mengdan Zhang, and Weiming Hu. DCFNet: Discriminant Correlation Filters Network for Visual Tracking. apr 2017. 5, 17

[86] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip H S Torr. Fast Online Object Tracking and Segmentation: A Unifying Approach. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3, 5, 14, 17, 20

[87] Xiaohan Wang, Yu Wu, Linchao Zhu, and Yi Yang. Symbiotic attention with privileged information for egocentric action recognition. In *AAAI Conference on Artificial Intelligence*, 2020. 1, 8

[88] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *Proceedings - International Conference on Image Processing, ICIP*, volume 2017-Septe, pages 3645–3649. IEEE Computer Society, mar 2018. 2

[89] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 284–293, 2019. 1

[90] Yi Wu, Jongwoo Lim, and Ming Hsuan Yang. Online object tracking: A benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2411–2418. IEEE Computer Society, 2013. 3

[91] Yi Wu, Jongwoo Lim, and Ming Hsuan Yang. Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1834–1848, sep 2015. 2, 3, 4, 5, 6, 14, 15, 18

[92] Yinda Xu, Zeyu Wang, Zuoxin Li, Ye Yuan, and Gang Yu. SiamFC++: Towards Robust and Accurate Visual Tracking with Target Estimation Guidelines. In *AAAI Conference on Artificial Intelligence*, nov 2020. 5, 14, 17

[93] Bin Yan, Haojie Zhao, Dong Wang, Huchuan Lu, and Xiaoyun Yang. 'Skimming-perusal' tracking: A framework for real-time and robust long-term tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2019-Octob, pages 2385–2393, 2019. 5, 14, 17

[94] Sangdoo Yun, Jongwon Choi, Youngjoon Yoo, Kimin Yun, and Jin Young Choi. Action-decision networks for visual tracking with deep reinforcement learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2017-Janua, pages 1349–1358. IEEE, jul 2017. 3

[95] Jianming Zhang, Shugao Ma, and Stan Sclaroff. MEEM: Robust tracking via multiple experts using entropy minimization. In *European Conference on Computer Vision*, volume 8694 LNCS, pages 188–203. Springer Verlag, 2014. 2

[96] Lichao Zhang, Abel Gonzalez-Garcia, Joost Van De Weijer, Martin Danelljan, and Fahad Shahbaz Khan. Learning

the model update for siamese trackers. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2019-Octob, pages 4009–4018. Institute of Electrical and Electronics Engineers Inc., oct 2019. 5, 17

[97] Zhipeng Zhang and Houwen Peng. Deeper and Wider Siamese Networks for Real-Time Visual Tracking. *IEEE Conference on Computer Vision and Pattern Recognition*, jan 2019. 5, 14, 17

[98] Zhipeng Zhang, Houwen Peng, Jianlong Fu, Bing Li, and Weiming Hu. Ocean: Object-aware Anchor-free Tracking. In *European Conference on Computer Vision*, jun 2020. 2, 3, 5, 14, 17

[99] L. Čehovin, A. Leonardis, and M. Kristan. Visual object tracking performance measures revisited. *IEEE Transactions on Image Processing*, 25(3):1261–1274, 2016. 18