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].

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 (E-K-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.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td># videos</td>
<td>51</td>
<td>100</td>
<td>128</td>
<td>123</td>
<td>365</td>
<td>100</td>
<td>60</td>
<td>80</td>
<td>150</td>
</tr>
<tr>
<td># frames</td>
<td>29K</td>
<td>59K</td>
<td>55K</td>
<td>113K</td>
<td>135K</td>
<td>383K</td>
<td>20K</td>
<td>102K</td>
<td>97K</td>
</tr>
<tr>
<td>Min frames across videos</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>109</td>
<td>146</td>
<td>169</td>
<td>41</td>
<td>406</td>
<td>161</td>
</tr>
<tr>
<td>Mean frames across videos</td>
<td>578</td>
<td>590</td>
<td>429</td>
<td>915</td>
<td>371</td>
<td>3830</td>
<td>332</td>
<td>1274</td>
<td>649</td>
</tr>
<tr>
<td>Median frames across videos</td>
<td>392</td>
<td>393</td>
<td>365</td>
<td>882</td>
<td>300</td>
<td>2448</td>
<td>258</td>
<td>1179</td>
<td>484</td>
</tr>
<tr>
<td>Max frames across videos</td>
<td>3872</td>
<td>3872</td>
<td>3872</td>
<td>3085</td>
<td>5040</td>
<td>20665</td>
<td>1500</td>
<td>2501</td>
<td>4640</td>
</tr>
<tr>
<td>Frame rate</td>
<td>30 FPS</td>
<td>30 FPS</td>
<td>30 FPS</td>
<td>30 FPS</td>
<td>30 FPS</td>
<td>240 FPS</td>
<td>30 FPS</td>
<td>30 FPS</td>
<td>60 FPS</td>
</tr>
<tr>
<td># target object classes</td>
<td>10</td>
<td>16</td>
<td>27</td>
<td>9</td>
<td>8</td>
<td>17</td>
<td>30</td>
<td>23</td>
<td>34</td>
</tr>
<tr>
<td># sequence attributes</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>12</td>
<td>12</td>
<td>9</td>
<td>6</td>
<td>13</td>
<td>17</td>
</tr>
<tr>
<td>FPV</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td># action verbs</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>20</td>
</tr>
</tbody>
</table>

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
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.

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], MCCOTH [84]) and deep features (ECO [21], ATOM [32], 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 [58], 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 SPLIT [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. 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. 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 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...
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).

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
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%. 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. 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. 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-
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].

<table>
<thead>
<tr>
<th>Hands-in-contact</th>
<th>LTMU-H</th>
<th>LTMU-F</th>
<th>ATOM</th>
<th>LTMU Ocean</th>
<th>SiamBAN</th>
<th>SiamRPN++</th>
<th>MetaCrest</th>
<th>D3S</th>
<th>DiMP</th>
<th>KYS</th>
<th>VITAL</th>
<th>GlobalTrack</th>
<th>MOSSE</th>
<th>ECO</th>
<th>Staple</th>
<th>KCF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.354</td>
<td>0.368</td>
<td>0.367</td>
<td>0.364</td>
<td>0.354</td>
<td>0.340</td>
<td>0.311</td>
<td>0.293</td>
<td>0.292</td>
<td>0.292</td>
<td>0.279</td>
<td>0.253</td>
<td>0.251</td>
<td>0.231</td>
<td>0.197</td>
<td>0.177</td>
</tr>
</tbody>
</table>

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. 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.

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.

---

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


[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.


[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


[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


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

