FCLT – A Fully-Correlational Long-Term Tracker

Alan Lukežič¹, Luka Čehovin Zajc¹, Tomáš Vojíř², Jiří Matas² and Matej Kristan¹
¹Faculty of Computer and Information Science, University of Ljubljana, Slovenia
²Faculty of Electrical Engineering, Czech Technical University in Prague, Czech Republic

{alan.lukezic, luka.cehovin, matej.kristan}@fri.uni-lj.si
{vojirtom, matas}@cmp.felk.cvut.cz

Abstract

We propose FCLT – a fully-correlational long-term tracker. The two main components of FCLT are a short-term tracker which localizes the target in each frame and a detector which re-detects the target when it is lost. Both the short-term tracker and the detector are based on correlation filters. The detector exploits properties of the recent constrained filter learning and is able to re-detect the target in the whole image efficiently. A failure detection mechanism based on correlation response quality is proposed. The FCLT is tested on recent short-term and long-term benchmarks. It achieves state-of-the-art results on the short-term benchmarks and it outperforms the current best-performing tracker on the long-term benchmark by over 18%.

1. Introduction

Recently, the computer vision community has witnessed significant activity and impressive advances in the area of model-free short-term trackers [49, 30]. Short-term trackers localize a target in a video sequence given a single training example in the first frame. Modern short-term trackers [12, 37, 25, 1, 45, 18, 23] localize the target moderately well even in the presence of significant appearance and motion changes and they are robust to short-term occlusions. Nevertheless, any adaptation at an inaccurate target position leads to gradual corruption of the visual model, drift and irreversible failure. Another major source of failures of short-term trackers are significant target occlusion and target disappearance from the field of view.

These problems are addressed by long-term trackers which combine a short-term tracker with a detector that is capable of reinitializing the tracker. A long-term tracker has to consider several design choices: (i) design of the two core components, (ii) their interaction algorithm, (iii) adaptation strategy, and (iv) estimation of tracking and detection uncertainty.

The design complexity has lead to ad hoc choices and heterogeneous, difficult to reproduce solutions. Initially, memoryless displacement estimators like flock-of-trackers [24] and the flow calculated at keypoints [42] were considered. Later methods applied keypoint detectors [42, 22, 43, 39], but this approach requires large and sufficiently well textured targets. Cascades of classifiers [24, 38] and more recently deep feature object retrieval systems [15] have been proposed to deal with diverse targets. The drawback is in the significant increase of computational complexity and the subsequent reduction in the scope of viable applications. Recent long-term trackers either adapt the detector [38, 22], which makes them prone to failure due to learning from incorrect training examples, or train the detector only on the first frame [42, 15], thus losing the opportunity to use the latest learned target appearance.
The main contribution of this paper is a novel fully correlational long-term (FCLT) tracker. The “fully correlational” refers to the fact that both the short-term tracker and the detector of FCLT are discriminative correlation filters (DCF) operating on the same representation. For some time, DCFs have been the state-of-the-art in short-term tracking, topping a number of recent benchmarks [49, 48, 28, 32, 29, 28, 27, 30]. However, with the standard learning algorithm [21], a correlation filter cannot be used for detection because of two reasons: (i) the dominance of the background in the search regions which necessarily has the same size as the target model and (ii) the effects of the periodic extension on the borders. Only recently theoretical breakthroughs [8, 25, 37, 26] have been made that allow constraining the non-zero filter response to the area covered by the target, effectively decoupling the sizes of the target and the search regions.

The FCLT is the first long-term tracker that exploits the novel DCF learning method, adopting the ADMM optimization from CSRDCF [37], the best performing real-time tracker in a recent short-term benchmark [27]. FCLT is the first tracker to use the optimization technique to build a fast detector. The FCLT thus uses a CSRDCF core [37] to maintain two correlation filters trained on different time scales that act as a short-term tracker and a detector (Figure 1). Since both the detector and the short-term tracker produce the correlation response on the same representation, the localization uncertainty can be estimated by inspecting the correlation response. As another contribution, a stabilizing mechanism is introduced to enable the detector to recover from model contamination.

The interaction between the short-term component and the detector allows long-term tracking even through long-lasting occlusions. Both components enjoy efficient implementation through FFT, which makes our tracker close to real-time. To the best of our knowledge this is the first long-term tracker fully formulated within the framework of discriminative correlation filters.

Extensive experiments show that FCLT tracker outperforms all long-term term-trackers on a long-term benchmark and achieves excellent performance even on short-term benchmarks, while running at 15 fps.

The remainder of the paper is structured as follows. Section 2 overviews most closely related work, the FCLT framework and tracker are presented in Section 3, experimental analysis is reported in Section 4 and Section 5 concludes the paper.

2. Related work

We briefly overview the most closely related work in short-term DCFs and long-term trackers.

**Short-term DCFs.** Since their inception in MOSSE tracker [2], several advancements have been made in discriminative correlation filters that made them the most widely used methodology in short-term tracking [30]. Major boosts in performance followed introduction of kernels by [21], multi-channel formulations [11, 16] and application to scale estimation [7, 35]. Hand-crafted features have been recently replaced with deep features trained for classification [9, 12, 6] as well as features trained for localization [45, 18]. Another line of advancements are constrained filter learning approaches [8, 25, 37] that allow learning a filter with effective size smaller than the training patch.

**Long-term trackers.** The long-term trackers combine a short-term tracker with a detector. The seminal work of Kalal et al. [24] proposes a memory-less flock of flows as a short-term tracker and a template-based detector run in parallel. They propose a P-N learning approach in which the short-term tracker provides training examples for the detector and pruning events are used to reduce contamination of the detector model. The detector is implemented as a cascade to reduce the computational complexity.

Another major paradigm was pioneered by Pernici et al. [43]. Their approach casts localization as local keypoint descriptors matching with a weak geometrical model. They propose an approach to reduce contamination of the keypoints model that occurs at adaptation during occlusion. Nebehay et al. [42] have shown that a keypoint tracker can be utilized even without updating and using pairs of correspondences in a GHT framework to track deformable models. Maresca and Petrosino [39] have extend the GHT approach by integrating various descriptors and introducing a conservative updating mechanism. The keypoint methods require a large and well textured target, which constrains their application scenarios.

Recent long-term trackers have shifted back to the tracker-detector paradigm of Kalal et al. [24], mainly due to advent of DCF trackers [21] which provide a robust and fast short-term tracking component. Ma et al. [38] proposed a combination of KCF tracker [21] and a random ferns classifier as a detector that is used to correct the tracker. Similarly, Hong et al. [22] have proposed a method that combines KCF tracker with a SIFT-based detector that is also used to detect occlusions.

The most extreme example of using a fast tracker and a slow detector is the recent work of Fan and Ling [15]. Their tracker combines a DSST [10] tracker with a CNN detector [44] that verifies and potentially corrects proposals of the short-term tracker. Their tracker achieved excellent results on the challenging long-term benchmark [40], but requires GPU, has a very large memory footprint and requires parallel implementation with backtracking to achieve a reasonable runtime.
3. Fully correlational long-term tracker

In the following we describe our long-term tracking approach based on constrained discriminative correlation filters. The constrained DCF is overviewed in Section 3.1, Section 3.2 overviews the short-term component, Section 3.3 describes the detector, Section 3.4 describes detection of tracking uncertainty and the long-term tracker is described in Section 3.5.

3.1. Constrained discriminative filter formulation

Our tracker is formulated within the framework of discriminative correlation filters. Given a search region of size \( W \times H \) a set of \( N_d \) feature channels \( \mathbf{f} = \{ \mathbf{f}_d \}_{d=1:N_d} \), where \( \mathbf{f}_d \in \mathbb{R}^{W \times H} \), are extracted. A set of \( N_d \) correlation filters \( \mathbf{h} = \{ \mathbf{h}_d \}_{d=1:N_d} \), where \( \mathbf{h}_d \in \mathbb{R}^{W \times H} \), are correlated with the extracted features and the object position is estimated as the location of the maximum of the averaged correlation responses

\[
\mathbf{r} = \sum_{d=1}^{N_d} (\mathbf{f}_d \ast \mathbf{h}_d)w_d, \tag{1}
\]

where \( \ast \) represents circular correlation, which is efficiently implemented by a Fast Fourier Transform and \( \{ \mathbf{w}_d \}_{d=1:N_d} \) are channel weights. The target scale can be efficiently estimated by another correlation filter trained over the scale-space [10].

We use the recently proposed discriminative correlation filter with channel and spatial reliability (CSRDCF) [37] as the basic filter learning method. This tracker constrains the filter learning by a binary mask, resulting in increased robustness and achieves excellent results on a recent benchmark [27]. The tracker also estimates per-channel reliability and uses it in responses averaging for increased robustness. We provide only a brief overview of the learning framework here and refer the reader to the original paper [37] for details.

**Constrained learning.** Since CSRDCF [37] treats feature channels independently, we will assume a single feature channel (i.e., \( N_d = 1 \)) in the following. A channel feature \( \mathbf{f} \) is extracted from a learning region and a fast segmentation [31] is applied to produce a binary mask \( \mathbf{m} \in \{0, 1\}^{W \times H} \) that approximately separates the target from the background. Next a filter of the same size as the training region is learned, with support constrained by the mask \( \mathbf{m} \). CSRDCF [37] learns the discriminative filter \( \mathbf{h} \) by minimizing the following augmented Lagrangian

\[
\mathcal{L}(\hat{\mathbf{h}}, \mathbf{h}, \mathbf{l}|\mathbf{m}) = \|\text{diag}(\hat{\mathbf{f}})\mathbf{h} - \mathbf{g}\|^2 + \frac{\lambda}{2}||\mathbf{h}_m||^2 + \frac{\lambda}{2}||\mathbf{h}_m||^2 \tag{2}
\]

\[
\|I^H(\hat{\mathbf{h}} - \mathbf{h}_m) + \mathbf{l}^H(\hat{\mathbf{h}} - \mathbf{h}_m)\| + \mu \|\mathbf{h}_c - \mathbf{h}_m\|^2,
\]

where \( \mathbf{g} \) is a desired output, \( \mathbf{l} \) is a complex Lagrange multiplier, \( (\cdot) \) denotes Fourier transformed variable, \( \mu > 0 \), and we use the definition \( \mathbf{h}_m = (\mathbf{m} \circ \mathbf{h}) \) for compact notation.

The solution is obtained via ADMM [3] iterations between two closed-form solutions, i.e.,

\[
\hat{\mathbf{h}}_{c+1} = (\mathbf{f} \circ \mathbf{g} + (\mu \mathbf{h}_m - \hat{\mathbf{l}})) \odot^{-1} \left( \mathbf{f} \circ \mathbf{f} + \mu \hat{\mathbf{l}} \right), \tag{3}
\]

\[
\mathbf{h}_{i+1} = \mathbf{m} \circ \mathcal{F}^{-1}(\hat{\mathbf{l}} + \mu \hat{\mathbf{h}}_{c+1})/ \left( \frac{\lambda}{2\mathbf{D}} + \mu \right), \tag{4}
\]

where \( \mathcal{F}^{-1}(\cdot) \) denotes inverse Fourier transform. In case of multiple channels, the approach independently learns a single filter per channel. Since the support of the learned filter is constrained to be smaller than the learning region, the maximum response on the training region reflects the reliability of the learned filter [37]. These values are used as per-channel weights \( \mathbf{w}_d \) in (1) for improved target localization. After estimating the new filter, CSRDCF [37] updates the segmentation model as well.

3.2. The short-term component

The CSRDCF [37] tracker is used as a short-term component in our long-term tracker. The short-term component is run within a search region centered on the estimated target position from the previous frame. The new target position hypothesis \( x_{t+1} \) is estimated as the location of the maximum response of its DCF within a search region centered at the estimate in the previous frame.

Figure 2. The short-term component estimates the target location at the maximum response of its DCF within a search region centered at the estimate in the previous frame.

The solution is obtained via ADMM [3] iterations between two closed-form solutions, i.e.,

\[
\hat{\mathbf{h}}_{c+1} = (\mathbf{f} \circ \mathbf{g} + (\mu \mathbf{h}_m - \hat{\mathbf{l}})) \odot^{-1} \left( \mathbf{f} \circ \mathbf{f} + \mu \hat{\mathbf{l}} \right), \tag{3}
\]

\[
\mathbf{h}_{i+1} = \mathbf{m} \circ \mathcal{F}^{-1}(\hat{\mathbf{l}} + \mu \hat{\mathbf{h}}_{c+1})/ \left( \frac{\lambda}{2\mathbf{D}} + \mu \right), \tag{4}
\]

where \( \mathcal{F}^{-1}(\cdot) \) denotes inverse Fourier transform. In case of multiple channels, the approach independently learns a single filter per channel. Since the support of the learned filter is constrained to be smaller than the learning region, the maximum response on the training region reflects the reliability of the learned filter [37]. These values are used as per-channel weights \( \mathbf{w}_d \) in (1) for improved target localization. After estimating the new filter, CSRDCF [37] updates the segmentation model as well.

3.3. The detector

The constrained learning [37] (Section 3.1) estimates a filter implicitly padded with zeros to match the learning region size. In contrast to naive learning of filter with a standard approach like [21] and multiplying with a mask post-
A motion model is added to increase robustness. We use a basic random walk, which models the likelihood of target position \( \pi(x_t) = N(x_t; x_c, \Sigma_t) \) at time-step \( t \) by a Gaussian with a diagonal covariance matrix \( \Sigma_t = \text{diag}(\sigma_{xt}^2, \sigma_{yt}^2) \) centered at the last confidently estimated position \( x_c \). The variances in the motion model gradually increase with the number of frames \( \Delta_t \) since the last confident estimation, i.e., \( [\sigma_{xt}, \sigma_{yt}] = [x_w, x_h]\alpha_s^\Delta_t \), where \( \alpha_s \) is scale increase parameter, \( x_w \) and \( x_h \) are the target width and height, respectively.

During detection stage, the filter \( h_t^{DE} \) is constructed according to (6) and correlated with features extracted from entire image. The detected position \( x_t^{DE} \) is estimated as the location maximum of the correlation response multiplied with the motion prior \( \pi(x_t) \) as shown in Figure 3.

### 3.4. Detection of tracking uncertainty

Tracking uncertainty detection is crucial for minimizing short-term visual model contamination as well as for activating target re-detection after events like occlusion. Our tracker is fully formulated within discriminative correlation filter framework, therefore tracking quality can be evaluated by inspecting the correlation response used for target localization.

Confident localization produces in a well expressed local maximum in the correlation response \( r_t \), which can be measured by a peak-to-sidelobe ratio PSR\((r_t)\) [2] as well as by the peak absolute value, \( \max(r_t) \). The localization quality is thus defined as the product of the two, i.e.,

\[
q_t = \text{PSR}(r_t) \cdot \max(r_t).
\]

Detrimental events like occlusion occur on a relatively short time-scale and are reflected in a significant reduction of the current localization quality. Let \( \overline{q}_t \) be the average localization quality computed over the recent \( N_q \) confidently tracked frames. Tracking is considered uncertain if the ratio between \( \overline{q}_t \) and \( q_t \) exceeds a predefined threshold \( \tau_q \), i.e.,

\[
\frac{\overline{q}_t}{q_t} > \tau_q.
\]

Figure 4 shows an example of confident tracking before and after occlusion. The ratio between average and current localization quality significantly increases during occlusion, indicating a highly uncertain tracking.

### 3.5. Tracking with FCLT

The short-term component (Section 3.2), detector (Section 3.3) and uncertainty estimator (Section 3.4) are integrated into a fully correlational long-term tracker as follows.

**Initialization.** The FCLT tracker is initialized in the first frame and the learned initialization model \( h_0^{ST} \) is stored. In the remaining frames, two visual models are maintained at different time-scales for target localization: the short-term visual model \( h_t^{ST} \) and the detector visual model \( h_t^{DE} \).

**Localization.** A tracking iteration at frame \( t \) starts with the target position \( x_{t-1} \) from previous frame as well as tracking quality score \( q_{t-1} \) and the mean \( \overline{q}_{t-1} \) over the recent \( N_q \) confidently tracked frames. A region is extracted
Figure 4. Qualitative example of the localization uncertainty measure (8), which reflects the strength of the peak in correlation response. The measure significantly increases during occlusion and reduces immediately after.

at location $x_{t-1}$ in the current image and the correlation response is computed using the short-term component model $h_{t-1}^{ST}$ (Section 3.2). Position $x_t^{ST}$ and localization quality $q_t^{ST}$ (7) are estimated from the correlation response $r_t^{ST}$. If tracking was confident at $t-1$, i.e., the uncertainty (8) $q_t^{DE}/q_t$ was smaller than $\tau_q$, only the short-term component is run, otherwise the detector (Section 3.3) is activated as well to address potential target disappearance. A detector filter $h_t^{DE}$ is constructed according to (6) and correlated with the features extracted from the entire image. The detection hypothesis $x_t^{DE}$ is obtained as the location of the maximum of the correlation multiplied by the motion model $\pi(x_t)$, while the localization quality $q_t^{DE}$ (7) is computed only on the correlation response.

**Update.** In case the detector has not been activated, the short-term position is taken as the final target position estimate. Alternatively, both position hypotheses, i.e., the position estimated by the short-term component as well as the position estimated by the detector, are considered. The final target position is estimated as the one with higher quality score, i.e.,

$$\{(x_t, q_t) = \begin{cases} (x_t^{ST}, q_t^{ST}) ; & q_t^{ST} \geq q_t^{DE} \\ (x_t^{DE}, q_t^{DE}) ; & \text{otherwise} \end{cases} \} \quad (9)$$

If the estimated position is reliable (8), a constrained filter $h_t^{ST}$ is updated according to CSRDCF [37] and (5). Otherwise the short-term visual model is not updated, i.e., $\eta = 0$ in (5).

4. Experiments

This section provides a comprehensive experimental evaluation of the FCLT tracker. Implementation details are discussed in Section 4.1. FCLT is a long-term tracker, we nevertheless start by an evaluation on challenging short-term benchmarks OTB100 [49], UAV123 [40] and VOT2016 [28] since the transition between long and short term tracking is not abrupt and a long-term tracker without competitive short-term performance is of limited use. The results are presented in Section 4.2. An extensive evaluation on the long-term benchmark UAV20L [40] including per-sequence and attribute-based analysis is reported in Section 4.3, while the detector importance is experimentally evaluated in Section 4.4.

4.1. Implementation details

We use the same standard HOG [5] and colornames [46, 11] features in the short-term component and in the detector. All the parameters of the CSRDCF filter learning are the same as in [37], including filter learning rate $\eta = 0.02$ and regularization $\lambda = 0.01$. The parameter for filter mixing in detector construction was set to $\alpha_D = 0.05$ and the motion model scale increase parameter was set to $\alpha_s = 1.05$. The uncertainty threshold was set to $\tau_q = 2.9$ and the parameter “recent frames” was $N_q = 100$. The parameters did not require fine tuning and were kept constant throughout all experiments. Our Matlab implementation runs on average at 15 frames per second on OTB100 [49] dataset on an Intel Core i7 3.4GHz standard desktop.

4.2. Performance on short-term benchmarks

For completeness, we first evaluate the performance of FCLT on the popular short-term benchmarks: OTB100 [49], UAV123 [40] and VOT2016 [28]. A standard no-reset evaluation (OPE [48]) is applied to focus on long-term behavior: a tracker is initialized in the first frame and left to track until the end of the sequence.

Tracking quality is measured by precision and success plots. The success plot shows all threshold values, the proportion of frames with the overlap between the predicted and ground truth bounding boxes as greater than a threshold. The results are summarized by areas under these plots which are shown in the legend. The precision plots in Figures 5, 6, 7 and 8 show a similar statistics computed from the center error. The results in the legends are summarized by percentage of frames tracked with an center error less than 20 pixels.

The benchmarks results already contain some long-term trackers. We added the most recent PTA V [15] – the currently best-performing published long-term tracker. Since FCLT is derived from the recent CSRDCF [37] we include this tracker as well. We remark that PTA is not causal, i.e. it uses future frames to predict the position of the tracked object which limits its applicability.
4.2.1 OTB100 [49] benchmark

The OTB100 [49] contains results of 29 trackers evaluated on 100 sequences with average sequence length of 589 frames. To reduce clutter in the graphs, we show here only the results for top-performing recent baselines, i.e., Struck [19], TLD [24], CXT [13], ASLA [50], SCM [52], LSK [36], CSK [20], OAB [17], VTS [34], VTD [33], CMT [42] and results for recent top-performing state-of-the-art trackers SRDCF [8], MUSTER [22], LCT [38] and CSRDCF [37].

The FCLT ranks among the top on this benchmark (Figure 5) outperforming all baselines as well as state-of-the-art SRDCF, CSRDCF and MUSTER. Using only handcrafted features, the FCLT achieves comparable performance to the non-causal PTAV [15] which uses deep features for re-detection and applies backtracking.

4.2.2 VOT2016 [28] benchmark

The VOT2016 [28] is the most challenging recent short-term tracking benchmark which contains results of 70 trackers evaluated on 60 sequences with the average sequence length of 358 frames. The dataset was created using a methodology that selected sequences which are difficult to track, thus the target appearance varies much more than in other benchmarks. In the interest of visibility, we show only top-performing trackers on no-reset evaluation, i.e., SSAT [41, 28], TCNN [41, 28], CCOT [12], MDNetN [41, 28], GGTv2 [14], MLDF [47], DNT [4], Deep-SRDCF [9], SiamRN [1] and FCF [28]. In addition we add CSRDCF [37] and the long-term trackers TLD [24], LCT [38], MUSTER [22], CMT [42] and PTAV [15].

The FCLT is ranked fifth on this benchmark according to the tracking success measure, outperforming 65 trackers, including DeepSRDCF [8] with deep features, CSRDCF and PTAV. Note that four tracker that achieve better performance than FCLT (SSAT, TCNN, CCOT and MDNetN) are CNN-based trackers and are computationally very expensive. They are optimized for accurate tracking on short sequences, without an ability for re-detection. The FCLT outperforms all long-term trackers on this benchmark (TLD, CMT, LCT, MUSTER and PTAV).

4.2.3 UAV123 [40] benchmark

The UAV123 [40] contains results of 14 trackers evaluated on 123 sequences with average sequence length of 915 frames. To reduce clutter in the graphs, we show here only the results for top-performing recent baselines, i.e., ASLA [50], Struck [19], SAMF [35], MEEM [51], LCT [38], TLD [24], CMT [42] and results for recent top-performing state-of-the-art trackers SRDCF [8], MUSTER [22], PTAV [15] and CSRDCF [37].

Results are shown in Figure 7. The FCLT outperforms by a margin all recent short-term trackers, i.e., SRDCF and CSRDCF as well as the long-term trackers (PTAV, MUSTER, LCT and TLD) in both measures.

4.3. Evaluation on a long-term benchmark

The long-term performance of the FCLT is analyzed on the recent long-term benchmark UAV20L [40] that contains results of 14 trackers on 20 long term sequences with average sequence length 2934 frames. To reduce clutter in the plots we include top-performing trackers SRDCF [8], OAB [17], SAMF [35], MEEM [51], Struck [19], DSST [7] and all long-term trackers in the benchmark (MUSTER [22], TLD [24]). We include the most recent state-of-the-art long-term trackers CMT [42], LCT [38], and PTAV [15] in the analysis. Additionally, we add recent state-of-the-art short-term DCF trackers.
Results in Figure 8 show that FCLT by far outperforms all top baseline trackers on benchmark as well as all the recent long-term state-of-the-art. In particular FCLT outperforms the recent long-term correlation filter LCT [38] by 93% in precision and success measures. The FCLT also outperforms the currently best-performing published long-term tracker PTAV [15] by over 17% and 18% in precision and success measures, respectively. This is an excellent result especially considering that FCLT does not apply deep features and backtracking like PTAV [15] and that it runs in near-realtime on a single thread CPU.

Table 2 shows tracking performance with respect to the twelve attributes annotated in the UAV20L benchmark. The FCLT is top performing tracker across all attributes, except fast motion, where PTAV exploits its backtracking mechanism. On the other hand, the FCLT achieves 18% and 12% better performance in tracking success at full occlusion and out-of-view comparing to PTAV. These attributes are the most specific for long-term tracking since target re-detection is required after they happen.

Figure 9 shows qualitative tracking examples for the proposed FCLT and four state-of-the-art trackers: PTAV [15], CSRDCF [37], MUSTER [22] and TLD [24]. In Group2 and Person19 a long-lasting full occlusion happens during tracking. Only FCLT is able to redetect the target in both situations. In sequence Person10, the target disappears from the image and only FCLT, TLD and MUSTER are able to re-detect it, but TLD and MUSTER are tracking it with much lower accuracy. In Boat2 sequence trackers suffer from background clutter, while FCLT and CSRDCF are able to track the target to the end. Sequence Human3 contains many partial occlusions. Only the FCLT and PTAV are able to successfully track.

4.4. Impact of the detector in FCLT

To study the importance of the detector in FCLT, we have compared per-sequence performance with the CSRDCF [37] which is used as the short-term tracker in the FCLT. For each sequence in UAV20L [40] we have calculated the average overlap between the predicted and ground truth bounding box and fraction of the tracked frames with overlap greater than zero as global performance measures. The number of tracking recoveries is quantified by counting the number of times the overlap increased from zero to a positive value. The results show that the detector and recovery system in FCLT is successful in many sequences. As a result, FCLT manages to track significantly longer than CSRDCF. In some cases, like the person14 sequence, the recovery dramatically improves FCLT tracking performance with a 1800% increase in the tracked sequence length compared to CSRDCF. On average, the FCLT outperforms the CSRDCF in overlap by over 38%.

5. Conclusion

We proposed a fully-correlational long-term tracker (FCLT). The FCLT is the first long-term tracker that exploits the novel DCF learning method from CSRDCF [37], the best performing real-time tracker in a recent short-term benchmark [27]. The method is used in FCLT to maintain two correlation filters trained on different time scales that act as a short-term and a detector component. The short-term component localizes the target within a limited search range in each frame. On the other hand, the detector exploits properties of the recent constrained filter learning [37] and is able to re-detect the target in the whole image efficiently. A failure detection mechanism based on correlation response quality is proposed and used for tracking uncertainty detection. The interaction between the short-term
component and the detector allows long-term tracking even through long-lasting occlusions.

Experimental evaluation on short-term benchmarks [49, 28, 40] showed state-of-the-art performance. On long-term benchmark [40] the FCLT outperform the best method by over 18%. The FCLT also consistently outperforms the short-term state-of-the-art CSRDCF [37], while running at the same frame-rate. Our Matlab implementation runs at 15 fps and it will be made publicly available.

References

[1] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr. Fully-convolutional siamese networks for object tracking. arXiv preprint arXiv:1606.09549, 2016. 1, 6
[2] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui. Visual object tracking using adaptive correlation filters. In Comp. Vis. Patt. Recognition, pages 2544–2550. IEEE, 2010. 2, 4
[3] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends in Machine Learning, 3(1):1–122, 2011. 3
[4] Z. Chi, H. Li, H. Lu, and M. H. Yang. Dual deep network for visual tracking. IEEE Trans. Image Proc., 26(4):2005–2015, 2017. 6
[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In Comp. Vis. Patt. Recognition, volume 1, pages 886–893, June 2005. 5
[6] M. Danelljan, M. Felsberg. Eco: Efficient convolution operators for tracking. In Comp. Vis. Patt. Recognition, pages 6638–6646, 2017. 2
[7] M. Danelljan, M. Häger, F. S. Khan, and M. Felsberg. Accurate scale estimation for robust visual tracking. In Proc. British Machine Vision Conference, pages 1–11, 2014. 2, 6
[8] M. Danelljan, M. Häger, F. Shahbaz Khan, and M. Felsberg. Learning spatially regularized correlation filters for visual tracking. In Int. Conf. Computer Vision, pages 4310–4318, 2015. 2, 6
[9] M. Danelljan, M. Häger, F. S. Khan, and M. Felsberg. Convolutional features for correlation filter based visual tracking. In IEEE International Conference on Computer Vision Workshop (ICCVW), pages 621–629, Dec 2015. 2, 6
[10] M. Danelljan, M. Häger, F. S. Khan, and M. Felsberg. Discriminative scale space tracking. IEEE Trans. Pattern Anal. Mach. Intell., 39(8):1561–1575, 2017. 2, 3
[11] M. Danelljan, F. S. Khan, M. Felsberg, and J. van de Weijer. Adaptive color attributes for real-time visual tracking. In 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014, pages 1090–1097, 2014. 2, 5
[12] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg. Beyond correlation filters: learning continuous convolution operators for visual tracking. In Proc. European Conf. Computer Vision, pages 472–488. Springer, 2016. 1, 2, 6, 7
[13] T. B. Dinh, N. Vo, and G. Medioni. Context tracker: Exploring supporters and distracters in unconstrained environments. In Comp. Vis. Patt. Recognition, pages 1177–1184, 2011. 6
[14] D. Du, H. Qi, L. Wen, Q. Tian, Q. Huang, and S. Lyu. Geometric hypergraph learning for visual tracking. IEEE Transactions on Cybernetics, 47(12):4182–4195, 2017. 6
[15] H. Fan and H. Ling. Parallel tracking and verifying: A framework for real-time and high accuracy visual tracking. In Int. Conf. Computer Vision, pages 5486–5494, 2017. 1, 2, 5, 6, 7
[16] H. K. Galoogahi, T. Sim, and S. Lucey. Multi-channel correlation filters. In Int. Conf. Computer Vision, pages 3072–3079, 2013. 2
[17] H. Grabner, M. Grabner, and H. Bischof. Real-time tracking via on-line boosting. In Proc. British Machine Vision Conference, volume 1, pages 47–56, 2006. 6
[18] E. Gundogdu and A. A. Alatan. Good features to correlate for visual tracking. arXiv preprint arXiv:1704.06326, 2017. 1, 2
[19] S. Hare, A. Saffari, and P. H. S. Torr. Struck: Structured output tracking with kernels. In Int. Conf. Computer Vision, pages 263–270, Washington, DC, USA, 2011. IEEE Computer Society. 6
[20] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista. Exploiting the circulant structure of tracking-by-detection with kernels. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, editors, Proc. European Conf. Computer Vision, pages 702–715, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. 6
[21] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista. High-speed tracking with kernelized correlation filters. IEEE Trans. Pattern Anal. Mach. Intell., 37(3):583–596, 2015. 2, 3
[22] Z. Hong, Z. Chen, C. Wang, X. Mei, D. Prokhorov, and D. Tao. Multi-store tracker (muster): A cognitive psychology inspired approach to object tracking. In Proc. Vis. Patt. Recognition, pages 749–758, June 2015. 1, 2, 6, 7
[23] C. Huang, S. Lucey, and D. Ramanan. Learning policies for adaptive tracking with deep feature cascades. In Int. Conf. Computer Vision, number 105–114, 2017. 1
[24] Z. Kalal, K. Mikolajczyk, and J. Matas. Tracking-learning-detection. IEEE Trans. Pattern Anal. Mach. Intell., 34(7):1409–1422, July 2012. 1, 2, 6, 7
[25] H. Kiani Galoogahi, A. Fagg, and S. Lucey. Learning background-aware correlation filters for visual tracking. In Int. Conf. Computer Vision, number 1135–1143, 2017. 1, 2
[26] H. Kiani Galoogahi, T. Sim, and S. Lucey. Correlation filters with limited boundaries. In Comp. Vis. Patt. Recognition, pages 4630–4638, 2015. 2
[27] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, L. Cehovin Zajc, T. Vojir, G. Hager, A. Lukezic, A. Eldesokey, and G. Fernandez. The visual object tracking vot2016 challenge results. In The IEEE International Conference on Computer Vision (ICCV), 2017. 2, 3, 7
[28] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, L. Cehovin, T. Vojir, G. Häger, A. Lukežič, and G. et al. Fernandez. The visual object tracking vot2017 challenge results. In Proc. European Conf. Computer Vision, 2016. 2, 5, 6, 8
Table 2. Tracking performance (AUC measure) for fourteen tracking attributes and seventeen trackers on the UAV20L [40].

| Tracker | Scale Var. | Aspect Change | Low Res. | Fast Motion | Full Occ. | Partial Occ. | Out-of-View | Back. Clutter | Illum. Var. | Viewp. Change | Camera Motion | Similar Object |
|---------|------------|---------------|----------|-------------|-----------|--------------|-------------|---------------|-------------|---------------|---------------|----------------|
| FCLT    | 0.492      | 0.464         | 0.420    | 0.328       | 0.423     | 0.488        | 0.435       | 0.537         | 0.518       | 0.456         | 0.493         | 0.545          |
| PTA V   | 0.416      | 0.410         | 0.390    | 0.349       | 0.357     | 0.415        | 0.389       | 0.435         | 0.435       | 0.430         | 0.418         | 0.420          |
| CCOT    | 0.378      | 0.322         | 0.277    | 0.277       | 0.183     | 0.368        | 0.352       | 0.188         | 0.382       | 0.330         | 0.380         | 0.463          |
| CSRDCF  | 0.346      | 0.293         | 0.232    | 0.194       | 0.210     | 0.339        | 0.326       | 0.227         | 0.359       | 0.308         | 0.348         | 0.403          |
| SRDCF   | 0.332      | 0.270         | 0.228    | 0.197       | 0.170     | 0.320        | 0.329       | 0.156         | 0.295       | 0.303         | 0.327         | 0.397          |
| MUSTER  | 0.314      | 0.275         | 0.278    | 0.206       | 0.200     | 0.305        | 0.309       | 0.230         | 0.242       | 0.318         | 0.307         | 0.342          |
| TLD     | 0.193      | 0.196         | 0.159    | 0.235       | 0.154     | 0.201        | 0.212       | 0.111         | 0.167       | 0.188         | 0.202         | 0.225          |
| LCT     | 0.244      | 0.201         | 0.183    | 0.112       | 0.151     | 0.244        | 0.249       | 0.156         | 0.232       | 0.225         | 0.245         | 0.283          |
| CMT     | 0.208      | 0.169         | 0.139    | 0.199       | 0.134     | 0.173        | 0.184       | 0.104         | 0.146       | 0.212         | 0.187         | 0.203          |
| OAB     | 0.301      | 0.262         | 0.261    | 0.198       | 0.258     | 0.298        | 0.308       | 0.332         | 0.303       | 0.303         | 0.307         | 0.310          |
| SAMF    | 0.298      | 0.251         | 0.212    | 0.140       | 0.174     | 0.288        | 0.262       | 0.201         | 0.323       | 0.276         | 0.294         | 0.333          |
| MEEM    | 0.277      | 0.231         | 0.195    | 0.166       | 0.163     | 0.274        | 0.253       | 0.212         | 0.334       | 0.243         | 0.283         | 0.302          |
| Struck  | 0.270      | 0.228         | 0.191    | 0.148       | 0.198     | 0.287        | 0.282       | 0.229         | 0.301       | 0.232         | 0.280         | 0.276          |
| DSST    | 0.251      | 0.203         | 0.159    | 0.123       | 0.159     | 0.249        | 0.241       | 0.211         | 0.288       | 0.206         | 0.257         | 0.274          |
| ASLA    | 0.274      | 0.231         | 0.192    | 0.129       | 0.160     | 0.254        | 0.266       | 0.162         | 0.236       | 0.250         | 0.250         | 0.325          |
| IVT     | 0.229      | 0.193         | 0.173    | 0.099       | 0.140     | 0.219        | 0.222       | 0.138         | 0.191       | 0.191         | 0.209         | 0.267          |
| CSK     | 0.173      | 0.134         | 0.117    | 0.102       | 0.082     | 0.182        | 0.209       | 0.074         | 0.153       | 0.157         | 0.188         | 0.215          |

Figure 9. Qualitative results of the FCLT and four state-of-the-art trackers on five sequences from [40] and [49].

[29] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, L. Čehovin, G. Fernandez, T. Vojar, G. Häger, G. Nebehay, and R. et al. Pflugfelder. The visual object tracking vot2015 challenge results. In Int. Conf. Computer Vision, 2015. 2
[30] M. Kristan, J. Matas, A. Leonardis, T. Vojir, R. Pflugfelder, G. Fernandez, G. Nebehay, F. Porikli, and L. Cehovin. A novel performance evaluation methodology for single-target trackers. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2016. 1, 2

[31] M. Kristan, J. Perš, V. Sulič, and S. Kovacíč. A graphical model for rapid obstacle image-map estimation from unmanned surface vehicles. In *Proc. Asian Conf. Computer Vision*, pages 391–406, 2014. 3

[32] M. Kristan, R. Pflugfelder, A. Leonardis, J. Matas, L. Čehovin, G. Nebehay, T. Vojir, and G. et al. Fernandez. The visual object tracking vot2014 challenge results. In *Proc. European Conf. Computer Vision*, pages 191–217, 2014. 2

[33] J. Kwon and K. M. Lee. Visual tracking decomposition. In *Comp. Vis. Patt. Recognition*, pages 1269–1276, 2010. 6

[34] J. Kwon and K. M. Lee. Tracking by sampling and integrating multiple trackers. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(7):1428–1441, July 2014. 6

[35] Y. Li and J. Zhu. A scale adaptive kernel correlation filter tracker with feature integration. In *Proc. European Conf. Computer Vision*, pages 254–265, 2014. 2

[36] B. Liu, J. Huang, L. Yang, and C. Kulikowski. Robust tracking using local sparse appearance model and k-selection. In *Comp. Vis. Patt. Recognition*, pages 1313–1320, June 2011. 6

[37] A. Lukežič, T. Vojíř, L. Čehovin Zajc, J. Matas, and M. Kristan. Discriminative correlation filter with channel and spatial reliability. In *Comp. Vis. Patt. Recognition*, pages 6309–6318, 2017. 1, 2, 3, 4, 5, 6, 7, 8

[38] C. Ma, X. Yang, C. Zhang, and M.-H. Yang. Long-term correlation tracking. In *Comp. Vis. Patt. Recognition*, pages 5388–5396, 2015. 1, 2, 6, 7

[39] M. E. Maresca and A. Petrosino. Matrioska: A multi-level approach to fast tracking by learning. In *Proc. Int. Conf. Image Analysis and Processing*, pages 419–428, 2013. 1, 2

[40] M. Mueller, N. Smith, and B. Ghanem. A benchmark and simulator for uav tracking. In *Proc. European Conf. Computer Vision*, pages 445–461, 2016. 2, 5, 6, 7, 8, 9

[41] H. Nam and B. Han. Learning multi-domain convolutional neural networks for visual tracking. In *Comp. Vis. Patt. Recognition*, pages 4293–4302, June 2016. 6

[42] G. Nebehay and R. Pflugfelder. Clustering of static-adaptive correspondences for deformable object tracking. In *Comp. Vis. Patt. Recognition*, pages 2784–2791, 2015. 1, 2, 6

[43] F. Pernici and A. Del Bimbo. Object tracking by oversampling local features. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(12):2538–2551, 2013. 1, 2

[44] R. Tao, E. Gavves, and A. W. M. Smeulders. Siamese instance search for tracking. In *Comp. Vis. Patt. Recognition*, pages 1420–1429, 2016. 2

[45] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, and P. H. S. Torr. End-to-end representation learning for correlation filter based tracking. In *Comp. Vis. Patt. Recognition*, pages 2805–2813, 2017. 1, 2

[46] J. van de Weijer, C. Schmid, J. Verbeek, and D. Larlus. Learning color names for real-world applications. *IEEE Trans. Image Proc.*, 18(7):1512–1523, July 2009. 5

[47] L. Wang, W. Ouyang, X. Wang, and H. Lu. Visual tracking with fully convolutional networks. In *Int. Conf. Computer Vision*, pages 3119–3127, Dec 2015. 6

[48] Y. Wu, J. Lim, and M.-H. Yang. Online object tracking: A benchmark. In *Comp. Vis. Patt. Recognition*, pages 2411–2418, 2013. 2, 5

[49] Y. Wu, J. Lim, and M. H. Yang. Object tracking benchmark. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(9):1834–1848, Sept 2015. 1, 2, 5, 6, 8, 9

[50] M.-H. Y. Xu Jia, Huchuan Lu. Visual tracking via adaptive structural local sparse appearance model. In *Comp. Vis. Patt. Recognition*, pages 1822–1829, 2012. 6

[51] J. Zhang, S. Ma, and S. Sclaroff. MEEM: robust tracking via multiple experts using entropy minimization. In *Proc. European Conf. Computer Vision*, pages 188—203, 2014. 6

[52] W. Zhong, H. Lu, and M. H. Yang. Robust object tracking via sparse collaborative appearance model. *IEEE Trans. Image Proc.*, 23(5):2356–2368, 2014. 6