# WAEF: Weighted Aggregation with Enhancement Filter for Visual Object Tracking

Litu Rout[1], Deepak Mishra[1], and Rama Krishna Sai Subrahmanyam Gorthi[2]

[1] Department of Avionics, Indian Institute of Space Science and Technology
Thiruvananthapuram, Kerala, India - 695 547
liturout1997@gmail.com, deepak.mishra@iist.ac.in

[2] Department of Electrical Engineering, Indian Institute of Technology
Tirupati, Andhra Pradesh, India - 517 506
rkg@iittp.ac.in

**Abstract.** In the recent years, convolutional neural networks (CNN) have been extensively employed in various complex computer vision tasks including visual object tracking. In this paper, we study the efficacy of temporal regression with Tikhonov regularization in generic object tracking. Among other major aspects, we propose a different approach to regress in the temporal domain, based on weighted aggregation of distinctive visual features and feature prioritization with entropy estimation in a recursive fashion. We provide a statistics based ensembler approach for integrating the conventionally driven spatial regression results (such as from ECO), and the proposed temporal regression results to accomplish better tracking. Further, we exploit the obligatory dependency of deep architectures on provided visual information, and present an image enhancement filter that helps to boost the performance on popular benchmarks. Our extensive experimentation shows that the proposed weighted aggregation with enhancement filter (WAEF) tracker outperforms the baseline (ECO) in almost all the challenging categories on OTB50 dataset with a cumulative gain of 14.8%. As per the VOT2016 evaluation, the proposed framework offers substantial improvement of 19.04% in occlusion, 27.66% in illumination change, 33.33% in empty, 10% in size change, and 5.28% in average expected overlap.

**Keywords:** Enhancement Filter · Temporal Regression · Weighted Aggregation · Feature Prioritization · Tikhonov Regularization · Ensembler

## 1 Introduction

Visual object tracking is one of the widely investigated problems by the computer vision community. The goal of this task is to estimate various attributes of an object with the sole supervision of a bounding box given in the first frame of a sequence. A possible approach to address this issue is to learn unique representation of the target object and employ discriminative power of deep similarity networks [28], or correlation filters [14,5] for efficient estimation of target

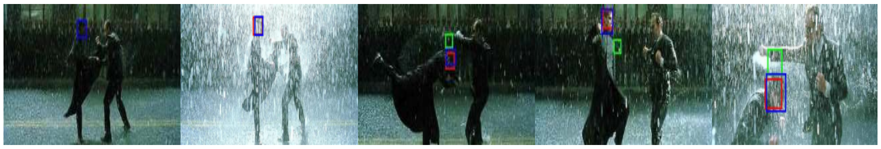

**Fig. 1.** The groundtruth of Matrix sequence of VOT2016 is shown in blue. The ECO (green) tracker fails to track the object because of drastic appearance changes. However, our ECO_EF (red) can handle the abrupt transition in appearance, mainly due to enhanced visual information provided before feature extraction, and tracks successfully.

attributes, often include target position and size. Though the tracking community has achieved significant progress in the recent years, especially after the widespread success of deep CNN in various vision challenges, the complexity of the problem still persists. The difficulty in tracking generic objects in an unconstrained environment still remains at a high level due to several rationale such as occlusion, deformation etc. Getting better at resolving these issues usually has a very good impact on various cross platforms that involves video surveillance, traffic monitoring, human computer interaction etc.

Despite the effort devoted by a large part of the community, there are still several challenges yet to be conquered. To overcome such challenges, most of the previously proposed trackers focus on some of the key components in tracking, including robust feature extraction for learning better representation [20,1,25,34], accurate scale estimation [5], rotation adaptiveness [17,27], motion models [16] etc. There are several other state-of-the-art trackers such as SRDCF [7], and CCOT [8] that implement additional constraint on the residual sum of errors to enforce higher degree of smoothness on the physical movement of the object. In the pursuit of accurate tracking, some of the proposed frameworks [34,9] are predominantly attributed by sophisticated features and complex models. Further, the emergence of deep CNN has replaced the low-level hand-crafted features which are not robust enough to discriminate significant appearance changes. The success of deep learning based trackers such as MDNet [23] and TCNN [22] on popular tracking benchmarks such as OTB [33] and VOT [15] is a clear indication of the distinctive feature extraction ability of deep CNN. In spite of the popularity, these feature extractors still lack high quality visual inputs that can further boost the performance. Therefore, one of the major aspects of this paper is to study the effect of enhancing visual inputs prior to feature extraction. In some sequences like Matrix (ref. Fig. 1), the hand-crafted and CNN features, as used in ECO, also fail to track the target, whereas image enhancement leads to sophisticated feature extraction that helps in tracking under such conditions.

Though deep learning based models have gained a lot of attention on account of their accuracy and robustness, the inherent scarcity of data, and required time for training these networks online, leave such models a step behind the correlation filter (CF) trackers. For this reason, a proper synthesis of CNN as feature extractor, and CF as detector has been doing exceedingly well in most of the challenging sequences. However, most of these fusion based trackers[8,4], being supervised regressors, learns to maximize the spatial correlation between target and candidate image patches. Due to spatial regularization, as in SRDCF [7], such trackers are capable of searching in a large spatial region that produces a significant gain in performance. But, these classifiers give minimal consideration to regress in temporal domain. Therefore, we exploit the temporal regression (TR) ability of a simple, yet effective model considering weighted aggregation of preceding features. The proposed technical and theoretical contributions can be summarized as following:

- A simple and effective enhancement filter (EF) (Sect. 3.2) is proposed to alleviate the adverse conditions in visual inputs prior to feature extraction. By this approach, the proposed tracker is able to perform against the state-of-the-art on VOT2016 dataset with an improvement of 5.2% in Average Expected Overlap (AEO) over the baseline approach.
- Although a lot of methods have been developed based on spatial regression, TR still remains a relatively less explored method in tracking. Therefore, in this paper, a detailed analysis on impacts of employing TR in single object tracking is undertaken.
- For efficient learning of TR parameters, a weighted aggregation (Sect. 3.3) based approach is proposed to suppress the dominance of un-correlated frames while regressing in temporal domain. Also, the training features are further organised based on average information content (Sect. 3.3). To our knowledge, this is in contrast to the conventional linear regressions in which equal [14], or more preference [30] is given to the historic frames. In order to generalize better, and control over-fitting in temporal domain, we have embedded the whole TR framework in Tikhonov regularization (Sect. 3.3).

Though we have demonstrated the importance of contributions through integrating with ECO, the proposed framework is generic, and can be integrated with other trackers to tackle some of the aforementioned tracking challenges with certain improvement in accuracy. This paper is structured as following. At first we discuss the previous methods which intend to address similar issues as ours (Sect. 2), followed by the proposed methodology (Sect. 3). After describing fundamental concepts of the proposed contributions, we detail our experiments and draw essential inferences (Sect. 4) to assess the overall performance.

## 2 Related Works

Correlation Filter (CF) based trackers have gained a lot of attention due to their low computational cost, high accuracy, and robustness. The regression of circularly shifted input features with a Gaussian kernel makes it plausible for implementation in Fourier domain, which in fact is the predominant cause of low computational cost. The object representation models, as adapted by many such trackers, have emerged gradually with colour attributes [25], HOG [3], SIFT [34], sparse based[20], CNN [6], and hierarchical CNN [18]. These methods have assisted in diminishing the adverse effects of ill-posed visual inputs. In this paper, our proposed enhancement filter, in a loose sense, contributes towards alleviating this issue further by pre-processing the inputs prior to feature extraction.

Among spatio-temporal models, the Spatio-Temporal context model based Tracker (STT) [32] proposes a temporal appearance model that captures historical appearances to prevent the tracker from drifting into the background. Also, STT proposes a spatial appearance model that creates a supporting field which gives much more information than the appearance of the target, and thus, ensures robust tracking. The Recurrently Target-attending Tracker (RTT) [2] exploits the essential components of the target in the long-range contextual cues with the help of a Recurrent Neural Network (RNN). The close form solution used in RTT is computationally less intensive, and more importantly, it helps in mitigating occlusion cases upto a great extent. The deep architecture proposed in [29] consists of three networks: a Feature Net, a Temporal Net, and a Spatial Net which assist in learning better representation model, establishing temporal correspondence, and refining the tracking state, respectively. The Context Tracker [10] explores the context on-the-fly by a sequential randomized forest, an online template based appearance model, and local features. The distracters and supporters, as proposed in Context Tracker, are very much useful in verifying genuine targets in case of resumption. The TRIC-track [31] algorithm uses incrementally learned cascaded regression to directly predict the displacement between local image patches and part locations. The Local Evidence Aggregation [19], as per the discussion in TRIC-track, determines the confidence level which is used to update the model. The Recurrent YOLO (ROLO) [24] tracker studies the regression ability of RNN in temporal domain.

In a nutshell, most of the trackers try to incorporate temporal information either by enforcing filters of previous frames to be somehow similar or by combining the model through a convex combination, which often leads to low performance and high time complexity. In other words, the model possess dual responsibility of detecting the object and maintaining temporal correspondence. However, the proposed method suggests that regularization over the augmented version of two complementary spaces, one encompassing temporal feature space and another enforcing the spatial smoothness through temporal regression over the position variations, can lead to substantial gain in various challenging categories including illumination variation, size change and occlusion. In such case, one model is specifically trained to smoothly localize the object in spatial domain and the other model, to maintain the temporal correspondence in feature space. Thereafter, the mean ensemble of these two models leverage the spatio-temporal information to localize the target object. Though the idea of temporal regularization has been used before in correlation filters, relatively less attention has been paid in decomposing the model so as to enforce higher degree of smoothness on the motion model. Therefore, we propose to reduce the under performance of correlation filter trackers by decomposing the model into two separate models. The detailed description is given in the following Sect. 3.

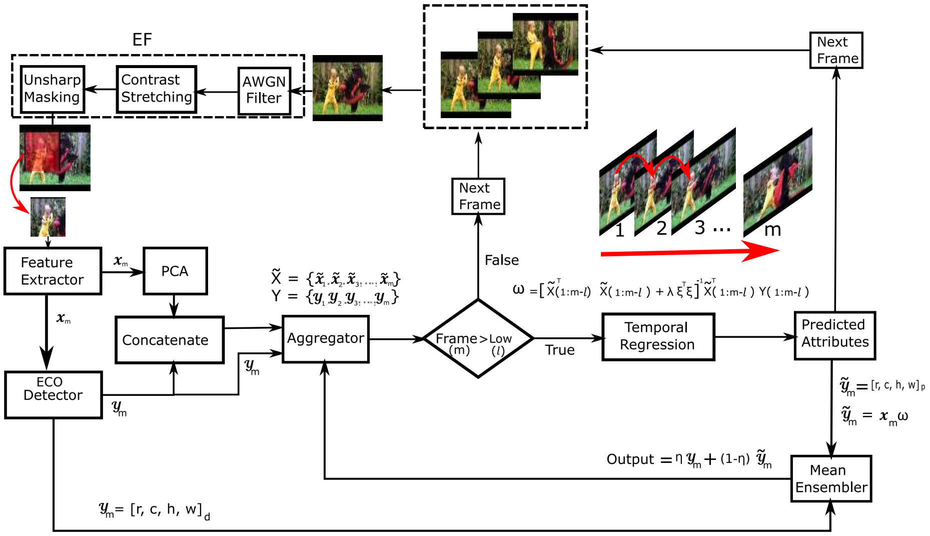
EF
Unsharp Masking
Contrast Stretching
AWGN Filter
Next Frame
Next Frame
1
2
3 ...
m
Feature Extractor
$x_m$
PCA
$x_m$
Concatenate
$\tilde{X} = \{\tilde{x}_1, \tilde{x}_2, \tilde{x}_3, \ldots, \tilde{x}_m\}$
$Y = \{y_1, y_2, y_3, \ldots, y_m\}$
Aggregator
False
$\omega = [\tilde{X}^T_{(1:m-l)} \tilde{X}_{(1:m-l)} + \lambda \xi^T \xi]^{-1} \tilde{X}^T_{(1:m-l)} Y_{(1:m-l)}$
Frame (m) > Low (l)
True
Temporal Regression
Predicted Attributes
ECO Detector
$y_m$
$y_m$
$\tilde{y}_m = [r, c, h, w]_p$
$\tilde{y}_m = x_m \omega$
Output $= \eta\, y_m + (1-\eta)\, \tilde{y}_m$
Mean Ensembler
$y_m = [r, c, h, w]_d$

**Fig. 2.** Temporal regression with weighted aggregation and enhancement filter as proposed in this paper (Sect. 3). Each frame is passed through enhancement filter (EF) before feature extraction. The detector (ECO) uses the extracted features and predicts the target attributes based on spatial correlation. The extracted features are projected into a low dimensional space where these are concatenated with target attributes. The concatenated features are then aggregated based on temporal correspondence and used in learning the parameters ($\omega$) of temporal regression. The TR model predicts the target attributes based on temporal information. Finally, the location of the target object is determined based on weighted mean ensemble of spatial and temporal predictions.

## 3 Proposed Methodology

The overall architecture of our method is shown in Fig. 2. As discussed in the contributions and the preceding sections we enhance the visual inputs before feature extraction through an EF (Sect. 3.2), and thereafter, the essential processing required for TR (Sect. 3.3) is depicted. For the sake of experimentation, we integrate the proposed methodology in ECO tracker, and showcase the efficacy by comparing with various state-of-the-art trackers on various benchmarks. We specifically provide a systematic approach based on well known regularization framework for incorporating temporal information in DCF trackers. The framework provides a proportionate weight-age across the previous frames based on their similarity with current frame and also considers feature prioritization based on the average information content in temporal domain.

At the beginning, we apply EF to each frame. After enhancement of visual information, the search region from each frame is fed to the feature extractor. The search region is decided based on the previous position and scale as implemented in [4]. The high dimensional CNN features, as extracted in [4], are projected onto a low dimensional space, aiming at reduction of time complexity. To achieve this, we have applied principal component analysis (PCA) with 90% captured variance. The compressed features are then concatenated with ECO detector outputs, and thereafter, these concatenated features with weighted aggregation (Sect. 3.3) are accumulated in the aggregator.

Let $X$ be a collection of feature vectors in $m$ frames $\{x_1, x_2, ..., x_m\} \in \mathbb{R}^{1\times n}$, where $n$ represents the number of features extracted from the highly correlated patch in each frame. Let $Y$ be a collection of regression targets of the corresponding $m$ frames $\{y_1, y_2, ..., y_m\} \in \mathbb{R}^{1\times p}$, where $p$ represents the dimension of attributes in the order of target centroid $(row, column)$ and size $(height, width)$ ,i.e., $(r, c, h, w)$. The matrix $Y$ contains the output $y_m$ of the detector and $X$ contains the corresponding input features to the detector. For robust prediction of $\widetilde{y}_m = x_m\omega$, we learn the regressor parameters $\omega \in \mathbb{R}^{n\times p}$ by accumulating the previous estimates of target attributes $Y(1 : m-1)$, and the associated features with controlled suppression of uncorrelated frames $\widetilde{X}(1 : m-1)$. Then we propose to augment the spatial ECO detector output $y_m$, with temporal regression output $\widetilde{y_m}$, by considering weighted mean ensemble $(\eta y_m + (1-\eta)\widetilde{y_m})$ consistently. The ensemble attributes are then fed back to the aggregator, which are used to update the accumulated attributes in $Y$ and $X$. The main reason for inclduing target attributes as input features is to enhance the degree of smoothness on the trajectory of the target object. However, updating target attributes in both $Y$ and $X$ may unfairly emphasize falsely tracked targets due to marginal inclusion of detector outputs. Therefore, we either update the concatenated detector outputs in $X$ by $x_m(end-p-1 : end) \leftarrow (\eta y_m + (1-\eta)\widetilde{y_m})$ or regression targets in $Y$ by $y_m \leftarrow (\eta y_m + (1-\eta)\widetilde{y_m})$. This is indeed the case as our experiments show that updating $X$ turns out to be more effective than the other counter parts. First, we discuss briefly the fundamental working principles of ECO (Sect. 3.1), and thereafter, the detailed contributions as shown in Fig. 2.

### 3.1 Baseline Approach: ECO

The ECO [4] tracker, which we have adopted as our baseline, has performed well on various benchmarks [33,15,21]. The introduction of factorized convolution operators in ECO, has reduced the parameters in the DCF model drastically. Apart from efficient convolution operators, the ECO tracker proposes a method for feasible memory consumption by reducing the number of training samples, while maintaining diversity. Moreover, the efficient model update strategy, as proposed in ECO, reduces the unfavourable sudden appearance changes as a result of illumination variation, out-of-view, and deformation. As per the comprehensive experimentation, the ECO tracker with deep features outperforms all the previous trackers that rely on DCF formulation. Motivated by these findings, we have integrated the proposed framework into baseline ECO with deep settings in light of further improvement, and demonstrated that the newly developed approach offers significant gain in numerous challenging sequences.

### 3.2 Enhancement Filter (EF)

In real world scenarios, it is intractable to obtain high quality visual information due to stochastic nature of the environment. To combat several random fluctuations, while preserving the fine/sharp details of the information content in images, we employ edge adaptive Gaussian smoothing. The AWGN Filter block in Fig. 2 represents edge preserved Gaussian smoothing of additive white Gaussian noise (AWGN) with three channel or 3D multi variate Gaussian kernel of standard deviation close to 0 each (here, 0.1), in order not to smooth the edges. A detailed description on AWGN filters can be found in [26]. To span the whole intensity from 0 to 255, while rectifying the contrast imbalance in each channel, we have employed linear contrast stretching after AWGN removal.

Low frequency interference arises when the visual information is gathered under variable illumination. This holds in almost all indoor scenes because of the inverse square law of light propagation. Arguably, the outdoor scenes do not suffer from this effect, because the sun is so far away, that all the tiny regions in an image appear to be at equal distance from it. However, other illuminating sources may produce low frequency interference in an unconstrained environment. Also, we may sometimes be interested in minute details of a scene, or scenes that manifest in high frequencies such as object boundaries. Therefore, it is often desirable to suppress the unwanted low frequencies to leverage high variations in a scene. While this issue has been studied extensively in image processing tasks [26], even in state-of-the-art trackers, as per our knowledge, the necessary attention for the same is not paid explicitly. So we intend to introduce the popular algorithm, local unsharp masking on visual object tracking paradigm, which is shown in Eqn. (1). A detail description of these methods along with essential comparisons can be found in [26].

$$g(x, y) = A[f(x, y) - m(x, y)] + m(x, y) \tag{1}$$

where $A = \frac{kM}{\sigma(x,y)}$, $k$ is a scalar, $M$ is the average intensity of the whole image, $\sigma(x, y)$ represents variance of the window. $g(x, y)$, $f(x, y)$, and $m(x, y)$ represent resulting image, input image, and low pass version of $f(x, y)$, respectively.

### 3.3 Temporal Regression by Tikhonov Regularization in Tracking

Here, we elaborate our Temporal Regression (TR) framework with detailed analysis of each key components such as Weighted Aggregation, Feature Prioritization, Tikhonov Regularization, and Mean Ensembler.

**Weighted Aggregation (WA) in Temporal Regression:** Here, we illustrate the weighted aggregation strategy, which brings substantial gain on a diverse set of tough sequences. Let $\alpha \in \mathbb{R}^{m\times 1}$ represent the coefficients for modulating the $m$ frames in temporal domain. The elements of $\alpha$ are computed based on the projection of $x_m$ onto $X$ which consists of $m$ vectors in $\mathbb{R}^{1\times n}$. An important point to remember here is, even if $m$ frames are modulated based on this correlation metric, the frame $x_m$ remains unaltered due to maximal correlation, and also, it is excluded from training set. The underlying hypothesis is to learn from the weighted aggregation of preceding features based on similarity measure with the test frame $x_m$, and predict the current attributes $\widetilde{y_m}$. Thereby, we inhibit the dominance of dissimilar frames in voting for target attributes in the current frame. In other words, features from only those frames are amplified which have a contextual correspondence with the test frame in the temporal domain. We squash the elements of $\alpha$ using sigmoid activation in order to map the correlation values to a fixed smooth range between 0 and 1 for all frames, reason for which is understandable. Thus, the coefficients $\alpha$ can be computed using Eqn. (2).

$$\alpha = sigmoid(\frac{Xx_m^T}{n}), \tag{2}$$

where $X \in \mathbb{R}^{m\times n}$, $x_m \in \mathbb{R}^{1\times n}$, and $\alpha \in \mathbb{R}^{m\times 1}$.

The features from preceding $m$ frames are modulated by $\alpha$ to enhance the contribution of highly correlated frames, while suppressing the contribution of uncorrelated ones. Thereby, efficient aggregation of past information is utilized in learning the parameters of regressor, which leads to robust prediction of target attributes in the subsequent frames. The modulated training samples are computed by Eqn. (3).

$$\widetilde{X} = X. * \alpha \tag{3}$$

where $.*$ represents row wise multiplication with corresponding scalar value of $\alpha$, i.e., $\widetilde{X}(i,:) = X(i,:) * \alpha(i), i = 1,2,\ldots,m$ and $*$ represents element wise multiplication.

In a nutshell, the temporal regression model uses the information over several frames to determine which frames it should pay more, or less attention to. The proposed modulating factor determines the attention values while learning the representation. Thus, the WA block enforces selective learning of representation based on temporal correspondence. Fig. 3 shows the aggregation coefficients of Ironman sequence from OTB50. After obtaining $\widetilde{X} = \{\widetilde{x_1}, \widetilde{x_2}, \ldots, \widetilde{x_m}\}$, the training features are further regulated based on entropy of the associated random variables (Sect. 3.3).

**Feature Prioritization through Entropy Estimation (FPEE):** In this section, we briefly discuss an efficient feature engineering approach as part of WA, taking into account the uncertainty preserved in each feature in the temporal domain. The hypothesis is to estimate the entropy of each feature in $\widetilde{X}$ across all $m$ frames, and use this information content to enhance the contribution of that particular set of features towards estimation of target attributes. This can be achieved by modulating each column of $\widetilde{X}$, which is in contrast to row wise modulation, as done by $\alpha$. Let $f_i \in \mathbb{R}^{1\times m}, i = 1,2,\ldots,n$ represent a random variable with observations drawn from the $i^{\text{th}}$ feature of all $m$ frames. For the ease of experimentation, the observations of these random variables are used to estimate the distribution based on normalized histogram counts. For better

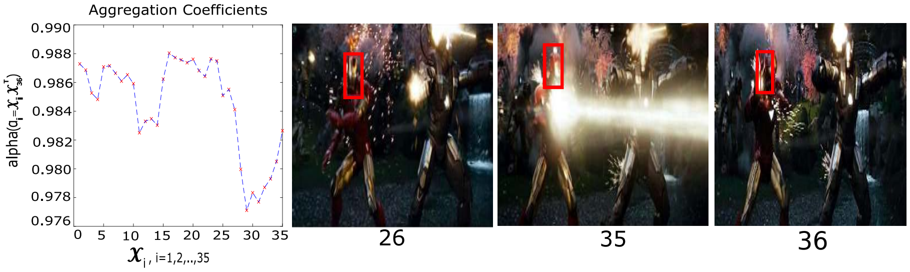

**Fig. 3.** Coefficients of aggregation $\alpha$, which are used to modulate the preceding features of the corresponding frames based on similarity rational. Here, $x_{36}$ has been projected onto $X(1:35)$, where $n = 3140, m = 36$, i.e., $x_i \in \mathbb{R}^{1\times 3140}, i = 1, 2, \dots, m$, $X \in \mathbb{R}^{36\times 3140}$, $Y \in \mathbb{R}^{36\times 4}$, and $\omega \in \mathbb{R}^{3140\times 4}$. Note that the current frame has higher correlation with the distant frames than the immediate previous ones.

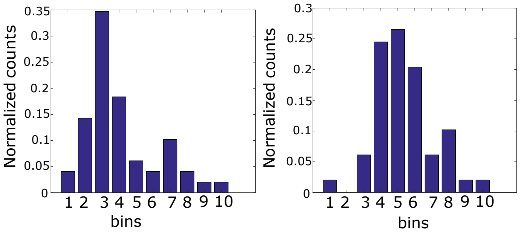

**Fig. 4.** The histogram of features are computed with fixed number of bins(here, 10). The normalized count is used as probability density $\mathbb{P}_{f_i}$. The distributions of $f_1$(left) and $f_{104}$(right) are used to quantify the average information content.

understanding, we have visualized the histogram of two random variables, $f_1$ and $f_{104}$ in Fig. 4.

The basic intuition is, learning that an unlikely event has occurred is more informative than a likely event has occurred. Therefore, we define self-information of event $f = \mathrm{f}$ by $I(\mathrm{f}) = -\log \mathbb{P}_f(\mathrm{f})$, with base $e$, as characterized in information theory. The self-information deals with a single outcome which leads to several drawbacks, such as an event with unity density has zero self-information, despite it is not guaranteed to occur. Therefore, we have opted Shannon entropy,

$$H(f) = \mathbb{E}_{f\sim\mathbb{P}_f}\left[I(\mathrm{f})\right] = -\mathbb{E}_{f\sim\mathbb{P}_f}\left[\log \mathbb{P}_f(\mathrm{f})\right],$$

which is used to deal with such issues [12], to quantify the amount of uncertainty conserved in the entire distribution. We use this uncertainty measure to enhance, or suppress the training features in $\widetilde{X} = f_1, f_2, \dots, f_n$ by Eqn. (4).

$$\widetilde{f_i} = f_i * H(f_i), i = 1, 2, \dots, n \tag{4}$$

Consequently, the parameters $(\omega)$ of temporal regression are computed with the updated training features $\widetilde{X} = \left\{\widetilde{f_1}, \widetilde{f_2}, \dots, \widetilde{f_n}\right\}$.

**Tikhonov Regularization in Temporal Regression:** Here, we describe the context in which we employ standard Tikhonov regularization. To ensure smooth variation of temporal weights ($\omega$), we have penalized the coefficients with larger norms. In our formulation, $\lambda\xi$ represents the standard Tikhonov operator. For equal preference, we have set $\xi$ to be an identity matrix $I \in \mathbb{R}^{m \times n}$, and $\lambda$ to be 1000. Thus, after incorporating temporal correspondence by WA and FPEE, the standard ridge regression has been updated to Eqn. (5).

$$J = \left\| \widetilde{X}\omega - Y \right\|_2^2 + \lambda \left\| \xi\omega \right\|_2^2 \tag{5}$$

The closed-form solution of $J$ can be obtained as following.

$$\boxed{\nabla_\omega \left\{ \left\| \widetilde{X}\omega - Y \right\|_2^2 + \lambda \left\| \xi\omega \right\|_2^2 \right\} = 0 \implies \omega = \left[ \widetilde{X}^T \widetilde{X} + \lambda \xi^T \xi \right]^{-1} \widetilde{X}^T Y,}$$

where $\omega \in \mathbb{R}^{n \times p}$, and the predicted attributes are computed by $\widetilde{y_m} = x_m \omega$.

**Mean Ensembler for Spatio-Temporal Aggregation:** This section depicts the theoretical background on the efficacy of mean ensemble. The proposed dynamic model comprises two models having minimal interdependence in their way of implementation. The detector works in the spatial domain with efficient training and robust model update strategy. On the contrary, the regression model operates in the temporal domain maximizing the correspondence with visual features from the current frame, and capturing the physically meaningful movement variables, such as position and angular displacement. Hence, the composition of these two models with bootstrap aggregation would be beneficial in lessening the overall error [12]. Assume there are $k$ models with error $\delta_i \sim \mathcal{N}(\mu = 0, \sigma^2 = v),\ i = 1, 2, \ldots, k$. Let the covariance $\mathbb{E}[\delta_i \delta_j] = c$. The error made by the mean ensembler output would be $\frac{1}{k}\sum_{i=1}^{k} \delta_i$. The expected squared error predicted by the ensembler would be

$$\mathbb{E}\left[ \left( \frac{1}{k} \sum_{i=1}^{k} \delta_i \right)^2 \right] = \mathbb{E}\left[ \frac{1}{k^2} \sum_{i=1}^{k} \left( \delta_i^2 + \sum_{j=1, j \neq i}^{k} \delta_i \delta_j \right) \right] = \frac{v}{k} + \frac{k-1}{k} c.$$

If the models are perfectly correlated, i.e., $\mathbb{E}\left[\delta_i \delta_j\right] = c = v$, then there will not be any improvement in expected squared error $v$. However, the uncorrelated models, i.e., $\mathbb{E}\left[\delta_i \delta_j\right] = 0$ would shrink the expected squared error by $k$ times. Thus, the proposed dynamic model would perform significantly better than the individual models due to ensemble of two partially uncorrelated models. In addition, the speed will not degrade much due to closed-form solution of the temporal weights.

## 4 Experiments

Here, we detail our experiments and draw essential inferences to validate our methodology. In all our experiments, we use VOT toolkit and OTB toolkit for

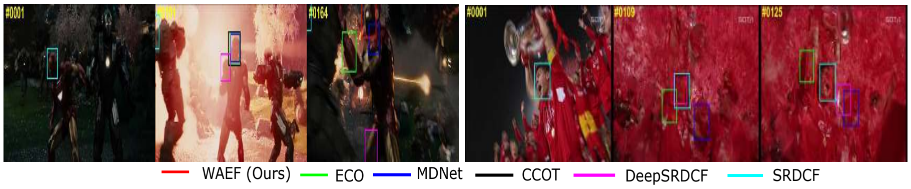

**Fig. 5.** Comparison on two of the toughest sequences from OTB50 dataset: Ironman(left) and Soccer(right). The WAEF tracker localizes the target under severe deformation, occlusion and illumination variations, unlike the compared trackers.

evaluation on VOT2016 and OTB50 benchmark, respectively. We develop our algorithm by progressively integrating the contributions into baseline. We demonstrate the impact of individual components by performing ablation studies on OTB50. We compare our top-performing trackers with state-of-the-art trackers and show compelling results in all the challenging categories of OTB50. Fig. 5 shows the qualitative analysis of the proposed framework.[3]

### 4.1 Implementation Details

To avoid the ambiguity caused by numerical computation of different machines, we evaluate both the baseline and our proposed trackers on the same machine with exactly same experimental setup. We use the exact parameter settings of ECO [4], including feature extraction, factorized convolution and optimization, for generating detector output. All the experiments are conducted on a single machine: Intel(R) Xeon(R) CPU E3-1225 v2 @ 3.20GHz, 4 Core(s), 4 Logical Processor(s), 44GB RAM and NVIDIA GPU (GeForce GTX 1080 Ti). The proposed tracker has been implemented on MATLAB with Matconvnet. We observed that elimination of immediate past frame ($(m-1)^{\text{th}}$) during training of the TR model provides improvement over inclusion of that particular frame. One possible hypothesis is that the output of the tracker may sometimes lead to false positive bounding box which will incrementally allow it to drift away from the actual target. In other words, the trajectory of an object, moving in a straight line, may become curved during regression due to the outlier in $(m-1)^{\text{th}}$ frame. To avoid this, one can eliminate few past frames from TR, but this would restrain the learning of recent appearance changes. Therefore, we propose to remove only the last frame from training TR model, which would capture the actual straight line trajectory, and thus, will assist in few scenarios where drastic change is a major concern. We have eliminated the experiments with removal of more immediate frames based on qualitative analysis, and showcase the efficacy of removing immediate past frame on whole OTB50 dataset. However, this approach may become troublesome when the actual trajectory has abrupt deviation from previous estimates. So, the weighted mean ensemble of spatial detector, which is mostly right (more weightage, $\eta = 0.7$), and TR would be

[3] For more results on OTB and VOT, please refer to supplementary material.

**Table 1.** The success and precision area under the curve (AUC) of the individual components of our proposed framework on OTB50.

| Tracker | WAEF | TREF | ECO | TR2 | TR1 | WAEF1 | WAEF2 |
|---|---|---|---|---|---|---|---|
| Success Rate | 0.651 | 0.648 | 0.643 | 0.627 | 0.619 | 0.615 | 0.610 |
| Precision | 0.880 | 0.877 | 0.874 | 0.849 | 0.839 | 0.825 | 0.814 |

useful to tackle this issue. The weights have been determined by employing a grid search from 0 to 1 with step size 0.1. The TR model requires a minimum of Low($l$)= 2 frames for a meaningful regression. We consider only past 50 frames for training TR model to meet the computational requirement.

### 4.2 Ablation Studies

In Table 1, we analyse the performance of ablative trackers on OTB50 benchmark. TR1 and TR2 denote the temporal regression with training features from max($m-50, 1$) to $m-1$ and $m-2$, respectively. Note that the TR1 and TR2 do not use weighted aggregation while computing $\omega$. It is evident that TR2 is better than TR1 both in accuracy and robustness, which validates our hypothesis of excluding immediate previous frame from training TR model in order to supress the adverse effect of outliers up to some extent. Despite the weak performance of TR, the composition tracker TREF outperforms the baseline in Success rate and Precision. Further, the WA and TREF consolidate into Weighted Aggregation with Enhancement Filter (WAEF) which again achieves substantial gain over baseline. In WAEF1, WAEF2 and WAEF, we update $x_m$& $y_m$, $y_m$, and $x_m$, respectively. It is evident that WAEF performs better than its counterparts, which validates our claim of updating $x_m$ alone in order to enforce smooth transition from previous frame. We report that the WAEF tracker exceeds the baseline with a gain of 1.24% in success rate, and 0.69% in precision.

### 4.3 Comparison with the State of the Arts

**Evaluation on OTB50** In Fig. 6, we compare our top-performing trackers with the state-of-the-art trackers. Among the compared trackers, our WAEF tracker does exceedingly well, outperforming the winner on OTB50. We observe that the proposed framework is robust enough to tackle the typical challenging issues in object tracking. In Table 2, we show the categorical comparison of area under the curve (AUC) and success rate, which are the standard metrics on benchmark results. The WAEF tracker provides substantial cumulative gain of 14.8% over all the crucial categories on OTB50. Moreover, the proposed architecture does not deteriorate the baseline performance in either of the aforementioned categories.

**Evaluation on VOT2016** We also evaluate the WAEF tracker on VOT2016 dataset, and compare the results in Table 3. The WAEF tracker offers remarkable achievement, improving 5.28% AEO, 6.31% accuracy rank, and 7.75% robustness

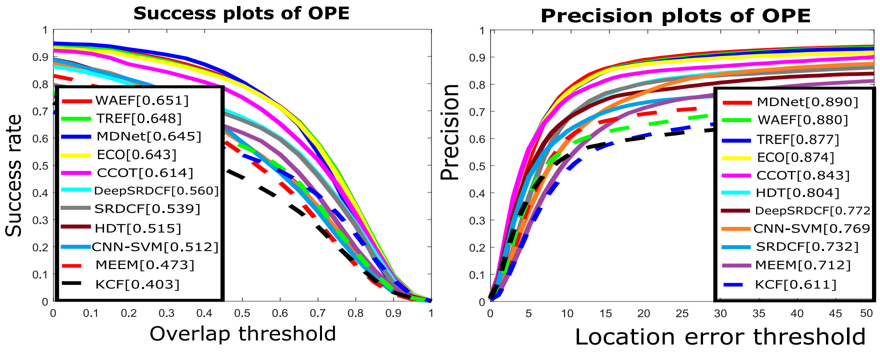

**Fig. 6.** The success and precision plots of our proposed WAEF, TREF, and several state-of-the-art trackers on OTB50 dataset.

**Table 2.** The success and precision plots in various category of our proposed WAEF, TREF, and several state-of-the-art trackers on OTB50 dataset.

| Tracker | **WAEF** | **TREF** | MDNet | ECO | CCOT | DeepSRDCF | SRDCF | HDT | KCF |
|---|---|---|---|---|---|---|---|---|---|
| Out of view | 0.657 | 0.654 | 0.617 | 0.644 | 0.636 | 0.551 | 0.512 | 0.479 | 0.368 |
| Occlusion | 0.654 | 0.652 | 0.631 | 0.643 | 0.632 | 0.555 | 0.532 | 0.504 | 0.405 |
| Illumination Variation | 0.632 | 0.628 | 0.625 | 0.623 | 0.594 | 0.530 | 0.509 | 0.488 | 0.386 |
| Low Resolution | 0.626 | 0.623 | 0.608 | 0.617 | 0.613 | 0.511 | 0.486 | 0.471 | 0.334 |
| Background Clutter | 0.638 | 0.636 | 0.625 | 0.629 | 0.588 | 0.535 | 0.517 | 0.494 | 0.388 |
| Deformation | 0.634 | 0.634 | 0.627 | 0.621 | 0.602 | 0.532 | 0.520 | 0.488 | 0.399 |
| Out-of-plane rotation | 0.646 | 0.642 | 0.627 | 0.636 | 0.605 | 0.549 | 0.516 | 0.503 | 0.399 |
| FastMotion | 0.645 | 0.643 | 0.620 | 0.637 | 0.625 | 0.554 | 0.523 | 0.499 | 0.365 |

rank relative to baseline. In particular, the WAEF tracker provides substantial improvement of 19.04% in occlusion, 27.66% in illumination change, 33.33% in empty, and 10% in size change category of VOT2016, as can be inferred from Fig. 7. Also, to validate the usefulness of EF, we have experimented ECO with EF alone. We observe that the enhancement filter assists in shaping the visual information which eventually leads to a notable gain of 1.48% in AEO. This implicates that the robust feature extractors still lack high quality visual inputs that may boost the overall performance.

**Evaluation on VOT2018** Here, we build the proposed TR around a different framework CFCF [13], namely Correlation Filter with Temporal Regression (CFTR) and show that the performance consistently improves irrespective of the framework. The CFTR tracker achieves 3.44% and 7.27% gain in AEO and Robustness relative to baseline CFCF, respectively. The decomposed network runs almost double the speed of baseline without degrading the overall performance.

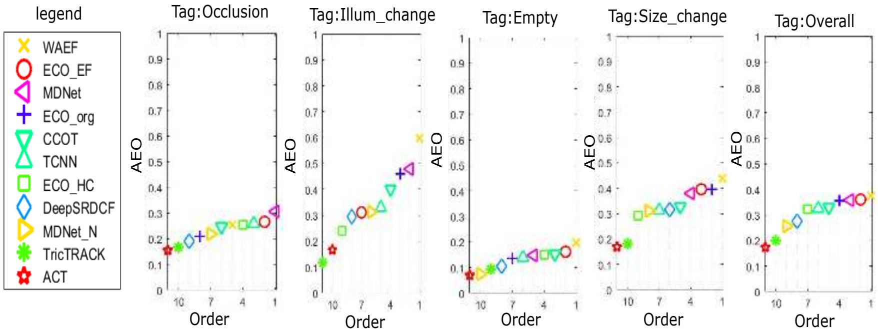

**Fig. 7.** Average Expected Overlap (AEO) analysis of our WAEF tracker and several other state-of-the-art trackers in various challenging categories of VOT2016.

**Table 3.** Overall quantitative analysis of few trackers on VOT2016. AEO, Ar, and Rr represents average expected overlap, accuracy rank, and robustness rank, respectively.

| Tracker | **WAEF** | **ECO_EF** | MDNet | ECO | CCOT | DeepSRDCF | TricTRACK |
|---|---|---|---|---|---|---|---|
| AEO | 0.3750 | 0.3616 | 0.3584 | 0.3563 | 0.3310 | 0.2763 | 0.1995 |
| Ar | 1.78 | 2.13 | 1.40 | 1.90 | 2.13 | 2.47 | 5.90 |
| Rr | 2.38 | 2.38 | 2.70 | 2.58 | 2.77 | 4.00 | 6.92 |

**Table 4.** Quantitative analysis on VOT2018 benchmark. The proposed CFTR tracker performs favourably against the state-of-the-art trackers.

| Trackers | LSART | CFWCR | **CFTR** | CFCF | ECO | GNET | CCOT | CRT |
|---|---|---|---|---|---|---|---|---|
| AEO | 0.323 | 0.303 | 0.301 | 0.286 | 0.280 | 0.274 | 0.267 | 0.244 |
| Accuracy | 0.50 | 0.49 | 0.51 | 0.51 | 0.48 | 0.50 | 0.49 | 0.46 |
| Robustness | 46.53 | 57.00 | 55.00 | 59.00 | 59.00 | 59.00 | 68.0 | 71.93 |
| Raw FPS | 1.72 | 1.80 | 0.62 | 0.32 | 3.71 | 1.29 | 0.15 | 3.24 |

## 5 Concluding Remarks

In this study, we demonstrated that enhancing the visual information prior to feature extraction, as proposed in this paper, can yield significant gain in performance. We analysed the impact of ridge regression with Tikhonov regularization in temporal domain, and showed promising results on popular benchmarks. Further, we introduced an approach to regress in the temporal domain based on weighted aggregation and entropy estimation, which provided drastic improvement in various challenging categories of popular benchmarks. Moreover, the proposed framework is generic, and can accommodate other detectors with simultaneously leveraging the spatial and temporal correspondence while localizing the target object. Our future scope will include robust feature selection based on sophisticated density estimation. Also, we will assimilate the performance of the proposed contributions on other publicly available datasets [21,11].

## References

1. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence **35**(8), 1798–1828 (2013)
2. Cui, Z., Xiao, S., Feng, J., Yan, S.: Recurrently target-attending tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1449–1458 (2016)
3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. vol. 1, pp. 886–893. IEEE (2005)
4. Danelljan, M., Bhat, G., Khan, F.S., Felsberg, M.: Eco: efficient convolution operators for tracking. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA. pp. 21–26 (2017)
5. Danelljan, M., Häger, G., Khan, F., Felsberg, M.: Accurate scale estimation for robust visual tracking. In: British Machine Vision Conference, Nottingham, September 1-5, 2014. BMVA Press (2014)
6. Danelljan, M., Hager, G., Shahbaz Khan, F., Felsberg, M.: Convolutional features for correlation filter based visual tracking. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 58–66 (2015)
7. Danelljan, M., Hager, G., Shahbaz Khan, F., Felsberg, M.: Learning spatially regularized correlation filters for visual tracking. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4310–4318 (2015)
8. Danelljan, M., Robinson, A., Khan, F.S., Felsberg, M.: Beyond correlation filters: Learning continuous convolution operators for visual tracking. In: European Conference on Computer Vision. pp. 472–488. Springer (2016)
9. Danelljan, M., Shahbaz Khan, F., Felsberg, M., Van de Weijer, J.: Adaptive color attributes for real-time visual tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1090–1097 (2014)
10. Dinh, T.B., Vo, N., Medioni, G.: Context tracker: Exploring supporters and distracters in unconstrained environments. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1177–1184. IEEE (2011)
11. Galoogahi, H.K., Fagg, A., Huang, C., Ramanan, D., Lucey, S.: Need for speed: A benchmark for higher frame rate object tracking. arXiv preprint arXiv:1703.05884 (2017)
12. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning, vol. 1. MIT press Cambridge (2016)
13. Gundogdu, E., Alatan, A.A.: Good features to correlate for visual tracking. IEEE Transactions on Image Processing **27**(5), 2526–2540 (2018)
14. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.: High-speed tracking with kernelized correlation filters. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**(3), 583–596 (2015)
15. Kristan, M., Matas, J., Leonardis, A., Vojir, T., Pflugfelder, R., Fernandez, G., Nebehay, G., Porikli, F., Čehovin, L.: A novel performance evaluation methodology for single-target trackers. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(11), 2137–2155 (Nov 2016). https://doi.org/10.1109/TPAMI.2016.2516982
16. Kwon, J., Lee, K.M.: Tracking by sampling trackers. In: 2011 IEEE International Conference on Computer Vision (ICCV). pp. 1195–1202. IEEE (2011)

17. Liu, T., Wang, G., Yang, Q.: Real-time part-based visual tracking via adaptive correlation filters. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4902–4912 (2015)
18. Ma, C., Huang, J.B., Yang, X., Yang, M.H.: Hierarchical convolutional features for visual tracking. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3074–3082 (2015)
19. Martinez, B., Valstar, M.F., Binefa, X., Pantic, M.: Local evidence aggregation for regression-based facial point detection. IEEE transactions on pattern analysis and machine intelligence **35**(5), 1149–1163 (2013)
20. Mei, X., Ling, H.: Robust visual tracking and vehicle classification via sparse representation. IEEE transactions on pattern analysis and machine intelligence **33**(11), 2259–2272 (2011)
21. Mueller, M., Smith, N., Ghanem, B.: A benchmark and simulator for uav tracking. In: European conference on computer vision. pp. 445–461. Springer (2016)
22. Nam, H., Baek, M., Han, B.: Modeling and propagating cnns in a tree structure for visual tracking. arXiv preprint arXiv:1608.07242 (2016)
23. Nam, H., Han, B.: Learning multi-domain convolutional neural networks for visual tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4293–4302 (2016)
24. Ning, G., Zhang, Z., Huang, C., Ren, X., Wang, H., Cai, C., He, Z.: Spatially supervised recurrent convolutional neural networks for visual object tracking. In: 2017 IEEE International Symposium on Circuits and Systems (ISCAS). pp. 1–4. IEEE (2017)
25. Pérez, P., Hue, C., Vermaak, J., Gangnet, M.: Color-based probabilistic tracking. Computer visionECCV 2002 pp. 661–675 (2002)
26. Petrou, M., Petrou, C.: Ch.4: Image Enhancement. John Wiley & Sons (2010)
27. Rout, L., Manyam, G.R., Mishra, D., et al.: Rotation adaptive visual object tracking with motion consistency. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1047–1055 (March 2018). https://doi.org/10.1109/WACV.2018.00120
28. Tao, R., Gavves, E., Smeulders, A.W.: Siamese instance search for tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1420–1429 (2016)
29. Teng, Z., Xing, J., Wang, Q., Lang, C., Feng, S., Jin, Y.: Robust object tracking based on temporal and spatial deep networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1144–1153 (2017)
30. Valmadre, J., Bertinetto, L., Henriques, J., Vedaldi, A., Torr, P.H.S.: End-to-end representation learning for correlation filter based tracking. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
31. Wang, X., Valstar, M., Martinez, B., Haris Khan, M., Pridmore, T.: Tric-track: tracking by regression with incrementally learned cascades. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4337–4345 (2015)
32. Wen, L., Cai, Z., Lei, Z., Yi, D., Li, S.Z.: Robust online learned spatio-temporal context model for visual tracking. IEEE Transactions on Image Processing **23**(2), 785–796 (2014)
33. Wu, Y., Lim, J., Yang, M.H.: Online object tracking: A benchmark. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2013)
34. Zhou, H., Yuan, Y., Shi, C.: Object tracking using sift features and mean shift. Computer vision and image understanding **113**(3), 345–352 (2009)