information

MDPI

Article

Correlation Tracking via Self-Adaptive Fusion of Multiple Features

Zhi Chen [1], Peizhong Liu [1,*], Yongzhao Du [1], Yanmin Luo [2] and Wancheng Zhang [1]

[1] College of Engineering, Huaqiao University, Quanzhou 362021, Fujian, China; marico2018@163.com (Z.C.); yongzhaodu@126.com (Y.D.); cyril_cheng@163.com (W.Z.)

[2] College of Computer Science and Technology, Huaqiao University, Xiamen 361021, Fujian, China; lym@hqu.edu.cn

* Correspondence: pzliu@hqu.edu.cn; Tel.: +86-0591-2333-9012

Received: 13 August 2018; Accepted: 21 September 2018; Published: 27 September 2018

Abstract: Correlation filter (CF) based tracking algorithms have shown excellent performance in comparison to most state-of-the-art algorithms on the object tracking benchmark (OTB). Nonetheless, most CF based tracking algorithms only consider limited single channel feature, and the tracking model always updated from frame-by-frame. It will generate some erroneous information when the target objects undergo sophisticated scenario changes, such as background clutter, occlusion, out-of-view, and so forth. Long-term accumulation of erroneous model updating will cause tracking drift. In order to address problems that are mentioned above, in this paper, we propose a robust multi-scale correlation filter tracking algorithm via self-adaptive fusion of multiple features. First, we fuse powerful multiple features including histogram of oriented gradients (HOG), color name (CN), and histogram of local intensities (HI) in the response layer. The weights assigned according to the proportion of response scores that are generated by each feature, which achieve self-adaptive fusion of multiple features for preferable feature representation. In the meantime the efficient model update strategy is proposed, which is performed by exploiting a pre-defined response threshold as discriminative condition for updating tracking model. In addition, we introduce an accurate multi-scale estimation method integrate with the model update strategy, which further improves the scale variation adaptability. Both qualitative and quantitative evaluations on challenging video sequences demonstrate that the proposed tracker performs superiorly against the state-of-the-art CF based methods.

Keywords: visual tracking; correlation filter; multiple features; model update; self-adaptive fusion

1. Introduction

Visual tracking is one of the most important and active research issues in the field of computer vision, and it has been widely applied in a variety of applications such as video surveillance, autonomous driving, unmanned plane, human-computer interactions, robotics and so forth [1,2]. The goal of the visual tracking is to estimate the target position of each frame with the target given in the initial frame, and predict the translation of each frame accurately in the subsequent image sequences. In this paper, we only consider the single object tracking with the target given in the initial frame. Despite nearly ten years of research and development, and achieved excellent results. But it is still a very challenging task to develop a robust classifier or filter to identify the target and design an optimal strategy to update the tracking model, due to the target object always suffer from prominent appearance variations and significant pose changes caused by occlusions, scale variation, illumination variations, and out of view [3].

The most current tracking algorithms in general can be roughly categorized into either discriminative [4–8] or generative models [9–12]. Generative methods establish the appearance model of the target objects by online learning, and formulate tracking task as searching for the region that is most similar to the target. A variety of classical generative methods has been proposed, included incremental tracker (IVT) [9], L1-min tracker [10], multi-task tracker (MTT) [11], sparse representation [12], and so forth. On the other hand, discriminative models regard target tracking as a binary classification problem, which distinguish the target objects from the sophisticated background, and these methods often also be referred to as tracking-by-detection methods. There are a lot of classification approaches employ these general discriminative models, such as multiple instance learning (MIL) [4], support vector tracking [5], P-N learning [6], compressive sensing [7], on-line boosting [8], and all of the correlation filter based algorithms [13–21]. These discriminative models take both the target and the surrounding background information into account, with superior discriminative power separating the target from the background and showing prominent advantages in the existing tracking methods, but easily lead to the phenomenon of overfitting [22]. In our work, the proposed algorithm is based on the discriminative models.

Among most existing tracking-by-detection methods, the correlation filter (CF) based trackers have recently drawn a great deal of attention and have been widely applied in visual tracking. Because of its ability to train and detect, which is performed by densely-sampled training examples through circularly shifting way, and transform complicated convolution operation into simple element-wise multiplication operation in the frequency domain according to the Convolution Theorem [13]. Therefore, it can effectively void the complicated and time-consuming convolution calculation. However, the majority of CF based trackers [7,13–15] only considered the limited single-channel features (such as texture [23], super-pixel [24], and Haar-like features [25], etc.), and the limited feature representations capability might weaken the discriminative ability of filters or classifiers. At present, numerous trackers employ multi-channel hand-crafted features (such as HOG [26], CN features [27], and combinations of these features [28]) and high discriminative power convolutional neural network (CNN) features (such as VGG networks [29]). The learned CF based trackers using high-dimensional deep features extracted from CNN features have superior robustness in photometric and geometric variations [30], and gain promising results on multiple tracking benchmarks [31,32]. However, the process of extracting high-dimensional deep convolutional features from each frame will generate largely time-consuming computational burden, which severely affects the real-time performance of visual tracking. On the other hand, in the process of object tracking, most CF based tracking algorithms [13,17,27,28] update the tracking model frame-by-frame. In practice, these approaches will introduce some inaccurate background information to the tracking model when target objects undergo sophisticated scenarios changes, such as occlusion or out-of-view in the current frame. Long-term accumulation of this erroneous information will deteriorate the whole tracking model and cause tracking drift.

In this paper, in order to address the problems mentioned above preferably, and to establish a desirable strategy to update the tracking model, we proposed a robust multi-scale correlation filter tracking algorithm based on self-adaptive fusion of multiple features, the main contributions of our work can be summarized as follows:

1. We integrated multiple multi-channel hand-crafted features with great discriminative power, such as HOG, CN, and histogram of local intensities into correlation filter framework in the response layer, combine the complementary advantages of multiple different features effectively and propose self-adaptive fusion of multiple features for preferable feature representation.
2. We establish a model update strategy to avoid the tracking model deteriorated by inaccurate update to some extent, which is performed by setting an optimal pre-defined response threshold as a judging condition for updating tracking model.
3. We integrate an accurate scale estimate method with the proposed model update strategy for further improving scale variation adaptability. We evaluate the proposed algorithm carried out

on the tracking benchmark dataset [31,32], and the experimental results demonstrate that the proposed algorithm performs favorably against several state-of-the-art CF based methods.

Section 3.1 introduces the baseline tracking framework. Section 3.2 introduces the scale discriminative filter. Section 4 describes the proposed tracking algorithm. Sections 4.1 and 4.2 introduce the self-adaptive fusion of multiple features. Section 4.3 introduces the proposed model updating strategy. Section 4.4 introduces the overall flow diagram of the proposed algorithm. Section 5 provides the experimental evaluations and analysis. Section 6 concludes the paper.

2. Related Work

The correlation filter theory was the first time introduced into visual tracking by Bolme, the main task is performed by leaning the minimum output sum of squared error filter (MOSSE) based on single-channel grayscale image patches, which achieve a real-time tracking with the speed of 669 frames per second [14]. Since then, multiple algorithms have been proposed successively that are based on MOSSE, and achieved remarkable improvement. Heriques et al. [15] introduced the CF to kernel space, and proposed a circulant structure with the kernel (CSK) method based on grayscale image patches, which achieve a real-time tracking. Afterwards, notable improvement were built upon CSK, and the multi-channel Gaussian kernel functions were introduced to extend the single-channel grayscale features to multi-channel HOG features in the KCF algorithm [13], which greatly improved the tracking performance. Danelljan et al. proposed an adaptive color attributes method that is based on CSK, which is achieved by mapping multi-channel color-name features into Gaussian kernel functions by the principal component analysis (PCA) technique [33]. It is not only maintains high-speed running time but also improves tracking performance. In order to preferably adjust to the scale change of the target objects, Danelljan et al. [16] proposed an accurate scale estimation method based tracking-by-detection framework, which is performed by learning the discriminative correlation filters on the scale pyramid, and then estimate the optimal scale from the most confidence frame. Li et al. [17] proposed a self-adaptive scale strategy under the CF based framework, and integrated powerful hand-crafted features including HOG and color-name together to effectively improve the scale adaptability. Ma et al. [18] achieved a robust long-term correlation tracking, which is performed by establishing spatio-temporal context models and object appearance models respectively. The re-detection is performed by learning an online random fern classifier to re-detect in the case of tracking failure. Danelljan et al. [19] introduced a spatial regularization component to restrain correlation filter coefficients, which solve the boundary effects efficiently caused by periodic assumption. Mueller et al. [20] integrated the contextual information surrounding the target objects into the learned filter during the learning stage, which effectively improve the discriminative ability of the filter. Lukezic et al. [21] introduced channel reliability and spatial reliability into a discriminative correlation filter, and the tracker integrates HOG and CN features into the tracking framework that further improves tracking accuracy. Wang et al. [34] proposed a model update strategy that is based on the tracking-by-detection framework, which effectively alleviate tracking drift caused by similar objects or background interference.

3. Tracking Components

3.1. The Context-Aware Correlation Filter Tracking Framework

Most existing CF based tracking algorithms usually employ cosine windows to relieve the boundary effects, which causes CF based trackers to usually have very limited contextual information. It easily tended to drift when target objects encounters sophisticated scenarios, such as fast motion and occlusion. The Context-Aware Correlation Filter (CACF) framework that is proposed by [20] that integrates surrounding context information into the learned filter, in order to learn a filter with a high response to the target image patch and a near-zero response to the context image patches. In our work, we employ CACF as our fundamental framework, for more detail on the derivation please see [20].

In CACF framework, the main goal is to gain an optimal correlation filter w, for all of the training samples A_0 generated by circulant shifts with a sliding window, the following ridge regression problem in the Fourier domain can be effectively solved through employing the property of the circulant matrix [13]:

$$\min_{w} \|D_0 w - y\|_2^2 + \lambda_1 \|w\|_2^2 \tag{1}$$

where the data matrix D_0 denotes all circular shifts of the vectorized image patch d_0, w is the learned correlation filter. The regression target y is a vectorized image of a two-dimensional (2D) Gaussian and λ_1 denotes regularization weight parameters.

The CACF produces to learn a filter that has a high response to the target image patch and near-zero response to the context image patches, which achieved by adding the context patches as a regularization term to standard formulation (see Equation (1)).

$$\min_{w} \|D_0 w - y\|_2^2 + \lambda_1 \|w\|_2^2 + \lambda_2 \sum_{i=1}^{k} \|D_i w\|_2^2 \tag{2}$$

Here, λ_1, λ_2 are regularization weight parameters, and the parameter λ_2 is utilized to control the context patches regressed to zeros. $D_0 \in R^{n\times n}$ and $D_i \in R^{n\times n}$ are corresponding circulant matrices, $w \in R^n$. Since the target image patch contains many context image patches and forming a new data matrix $B \in R^{(k+1)n\times n}$, the main objective Function (2) can be rewritten, as follows:

$$f_p(w, B) = \|Bw - \overline{y}\|_2^2 + \lambda_1 \|w\|_2^2 \tag{3}$$

where $B = \begin{bmatrix} D_0 \\ \sqrt{\lambda_2} D_1 \\ \vdots \\ \sqrt{\lambda_2} D_k \end{bmatrix}$ and $\overline{y} = \begin{bmatrix} y \\ 0 \\ \vdots \\ 0 \end{bmatrix}$, here $\overline{y} \in R^{(k+1)n}$ denotes the new desirable regression target. Since the objective function is a convex function, which can be minimized by derivation operation, as follows:

$$w = \left(B^T B + \lambda_1 I\right)^{-1} B^T \overline{y} \tag{4}$$

The closed-form solution of the standard formulation (see Equation (4)) can be solved efficiently by employing the property of circulant matrix in the Fourier domain.

$$\hat{w} = \frac{\hat{a}_0 \odot \hat{y}}{\hat{a}_0^* \odot \hat{a}_0 + \lambda_1 + \lambda_2 \sum_{i=1}^{k} \hat{a}_i^* \odot \hat{a}_i} \tag{5}$$

Here, a hat ˆ denotes the DFT (discrete Fourier transform) of a vector. The location of the target objects can be predicted, this is performed by the learned filter w convolving with image patch z (search window) in the next frame, and the location of the maximum response of all training sample response vectors $y_{\mathrm{p}}(z,w)$ is the predicted location of the target. For a given single image patch z, the output response is given by:

$$f(z) = \mathcal{F}^{-1}(z \odot w) = \hat{z} \odot \hat{w} \tag{6}$$

where $\mathcal{F}^{-1}$ denotes the inverse Fourier transformation and $\odot$ denotes the convolution operation. The computational complexity is that of the Fourier transform, with a complexity of only O (TN) for T frames of the image patch and N learned filters. Then, update the filter model by employing the following equations:

$$\hat{w}_i = \eta w_i + (1 - \eta)\hat{w}_{i-1} \tag{7a}$$

$$\hat{x}_i = (1 - \eta)\hat{x}_{i-1} + \eta \hat{x}_i \tag{7b}$$

where the subscript i denotes the sequence number of current frame, η is the learning rate parameter, and $\hat{x}_i$ denotes the appearance model of the target object.

3.2. The Scale Discriminative Correlation Filter

During the tracking process, in order to preferably adapt to the challenge of scale variation adaptability that is caused by the object itself, an accurate scale estimation method is proposed by [16] based on discriminative correlation filters on the tracking-by-detection framework. This is performed by training a scale discriminative correlation filter on the scale pyramid and then estimates the scale from the best confidence frame.

The image patch size centered around the target used for scale estimation is:

$$a^n P \times a^n R \quad n \in \left\{ \left\lfloor \frac{-(S-1)}{2} \right\rfloor, \ldots, \left\lfloor \frac{(S-1)}{2} \right\rfloor \right\}$$

where P and R denotes the width and height separately in the current frame and S denotes the number of scale space, a denotes the scale factor.

The goal is to obtain the optimal scale correlation filter h. That is achieved by minimizing the following objective function:

$$\varepsilon = \left\| \sum_{l=1}^{d} h^l * f^l - g \right\|^2 + \lambda \sum_{l=1}^{d} \|h^l\|^2 \tag{8}$$

where g denotes the desirable correlation output, l denotes the dimension of the feature, and λ is a regular coefficient. The above solution in the frequency domain is given by:

$$H^l = \frac{\overline{G} F^l}{\sum_{k=1}^{d} \overline{F}^k F^k + \lambda} = \frac{A_t{}^l}{B_t} \tag{9}$$

In order to obtain an accurate result, the denominators of H^1 in Equation (10) are, respectively, updated by follows:

$$A_t{}^l = (1-\eta) A_{t-1}^l + \eta \overline{G}_t F_t^l \tag{10}$$

$$B_\mathrm{t} = (1-\eta) B_{t-1} + \eta \sum_{k=1}^{d} \overline{F}_t^k F_t^k \tag{11}$$

Here, η is a learning rate parameter. In the next frame, the response of the scale filter can be solved by following equation:

$$\hat{y}_\mathrm{s} = \mathcal{F}^{-1} \left\{ \frac{\sum_{l=1}^{d} \overline{A}^l Z^l}{B + \lambda} \right\} \tag{12}$$

The maximum scale response score is obtained to estimate the target scale, and employ Equations (10) and (11) for updating the scale filter model.

4. The Proposed Algorithm

4.1. The Visual Features Performance Analysis

Feature representation is an important component of a visual tracking framework. In our work, we mainly focus on combining multiple hand-crafted features instead of high-dimensional deep features [29] for visual tracking. The hand-crafted features have shown superior performance in recent years and they have been successfully applied to CF based visual tracking, such as HOG, CN, and channel representations [35]. These features mainly focus on capturing the shape, color and luminance information of the target appearance [36].

The HOG feature is a feature descriptor widely used for both visual tracking [13,16–18,21] and object detection [37] in computer vision, and it maintains preferable invariance to translation, rotation, and illumination of the target objects [37]. It constructs feature, and this is performed by calculating and counting the histograms of gradient directions of local areas of the image patch. The CN is a linguistic color labels assigned by humans to represent colors in the world [38], and it has a preferable discriminative power in object recognition, object detection, and action recognition [39], which has been widely used in tracking, such as simple color transformations [40,41] and color histograms [35]. In addition, when compared with other visual features, the CN feature mainly extract features from colorful image patches and it has less dependence on the size, orientation and viewing angle of the image itself, and thus has higher robustness to the object shape and scale change.

HI (histogram of local intensities) [18] as a complementary of the multi-channel features can be easily integrated into hand-crafted features based framework. Its statistical properties of pixel intensities are exploited as features, and its local statistical features are robust to the appearance changes of the target in contrast to computing the statistical properties over the whole image patches. According to the above analysis, the proposed algorithm integrates HOG and CN features with HI, respectively, can capture different aspects of target appearance and they are complementary to each other.

As analyzed above, we can see that the performances of both CN and Hog features have complemented each other's advantages. In this paper, in order to guarantee the tracking accuracy and the real-time performance, the proposed framework (as shown in Figure 2) combine multiple hand-crafted features fusion of HOG, CN, and use HI feature in a 6 × 6 local window with nine bins in the response layer, and the more detailed process is described in the Section 4.2.

4.2. The Self-Adaptive Fusion of Multiple Features

In this paper, we aim to improve the tracking performance by adopting a multiple hand-crafted features, which is similar to intelligently-combining features from different classifier fusion rules [42,43]. Based on the above analysis (see Section 3.1), at first, we calculate the corresponding maximum response scores calculated by Equation (6) of both features integrating with the HI feature, respectively, based on CACF framework. The obtained response scores are normalized, and then the weights are assigned according to the proportion of the corresponding response scores. Therefore, the video sequences in the next frame will be prioritized for choosing the feature with higher weights. The normalized weights of both features integrate with the HI feature, respectively, in t-th frame are:

$$(HOG + HI)_{\overline{w}_t} = \frac{\max^2(f_{HOG+HI}(z))}{\max^2(f_{CN+HI}(z)) + \max^2(f_{HOG+HI}(z))} \tag{12}$$

$$(CN + HI)_{\overline{w}_t} = \frac{\max^2(f_{CN+HI}(z))}{\max^2(f_{CN+HI}(z)) + \max^2(f_{HOG+HI}(z))} \tag{13}$$

Here, HI denotes the histogram of local intensities feature, f_{HOG+HI} denotes corresponding output response score of the HOG feature integrate with HI feature calculated by Equation (6), and f_{CN+HI} denotes the corresponding output response score of the CN feature integrate with HI feature calculated by Equation (6). $(HOG + HI)_{\overline{w}_t}$ denotes the proportion of the response score corresponding to the (HOG+HI) feature, $(CN + HI)_{\overline{w}_t}$ is the similar meaning as $(HOG + HI)_{\overline{w}_t}$.

The weights in the (t + 1)-th frame w_{t+1} are used for updating the previous feature weights $\overline{w}_t$:

$$(HOG + HI)_{w_{t+1}} = (1 - \delta) \times (HOG + HI)_{w_t} + \delta \times (HOG + HI)_{\overline{w}_t} \tag{14}$$

$$(CN + HI)_{w_{t+1}} = (1 - \delta) \times (CN + HI)_{w_t} + \delta \times (CN + HI)_{\overline{w}_t} \tag{15}$$

Here, δ is an updating factor of the feature weights and the initial weights of both features are set to 0.5 in the first frame. We calculate corresponding response scores of both features integrate with HI feature are $(HOG + HI)_{Rt}$ and $(CN + HI)_{Rt}$, respectively, in the t-th frame.

We assign weights to all of the features according to the proportion of response scores. The final response is generated by the fusion of both features integrate with the HI feature respectively, and the final fused response score R_f is as follows:

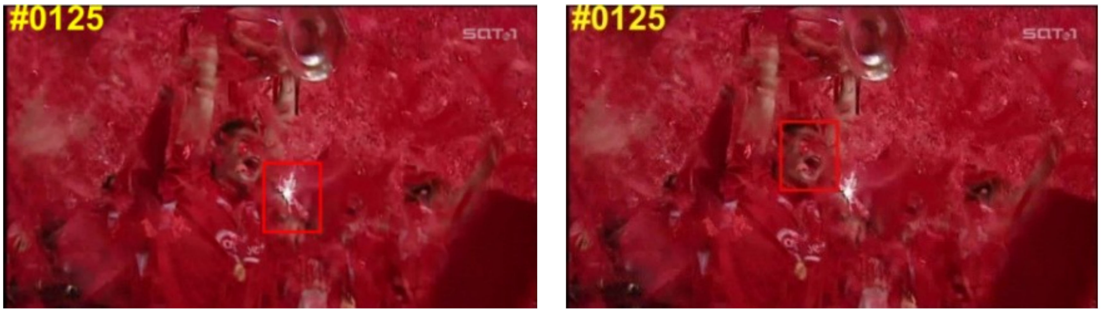
$$R_f = (HOG + HI)_{w_t} \times (HOG + HI)_{R_t}(CN + HI)_{w_t} \times (CN + HI)_{R_t} \tag{16}$$

The maximum response score R_{max} is used for estimating location of the target objects.

4.3. The Proposed Model Updating Strategy

Most existing CF based tracking algorithms often update their tracking model frame-by-frame. In fact, this updating method exist certain deficiency, when the target objects encounter some sophisticated scenarios changes, such as occlusion, illumination changes, and background clutters, it may easily cause inaccurate tracking. In the proposed method, we discover a phenomenon through multiple experimental investigation, which is the tracking easily tend to drift when the maximum response score is lower than a certain value. Based on this, we predefine a response threshold T as a judgmental criterion for updating the model. When the maximum response score R_{max} is greater than the predefined threshold, i.e., $R_{max} > T$, then the translation filter model (see Equation (6)) and the scale filter model (see Equation (11)) will be updated online with a learning rate parameter η (see Equations (7) and (10)), respectively. Otherwise, we choose not update the tracking model.

Figure 1 illustrates the tracking results of without model update strategy and with proposed updating strategy. In the Soccer sequence, the target object undergoes sophisticated scenario challenges such as occlusion, scale variation, illumination variation, background clutters and motion blur in the 125th frame. It can be seen from the figure (see Figure 1) that the method without model update strategy is easily prone to drift, and will cause tracking failure in the end. On the contrary, the method with the proposed updating strategy can still track correctly. More detail please sees the overall flow diagram of the proposed method is as shown in Figure 2.



(**a**) (**b**)

Figure 1. The tracking results of without model update strategy (**a**) and with the proposed model update strategy (**b**) in *Soccer* sequence

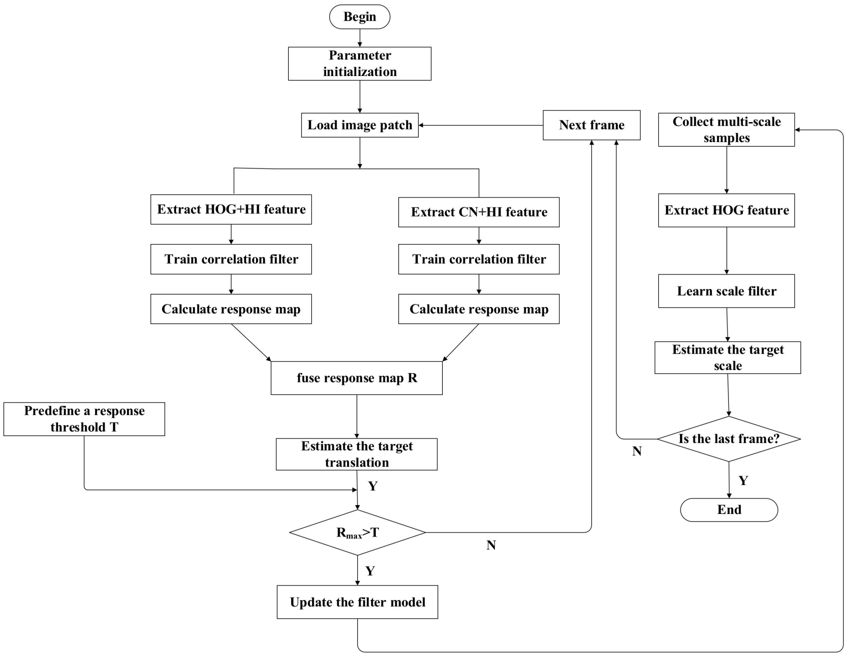


Figure 2. The overall framework of the proposed tracking algorithm.

5. Experiments

5.1. Implementation Details

The proposed tracker is implemented in MATLAB2014a with a 2.4GHz CPU with 8GB memory. We use HOG features in a 4×4 local window with 31 bins. The regularization parameter λ_1, λ_2 in Equation (5) are set to 10^{-4} and 0.4 separately. The size of the search window is set to 2.1 times of the target size. The learning rate η in Equations (7) and (10) is set to 0.025. The number of scale space $S = 33$ in and the scale factor a is set to 1.02. The predefined response threshold T is set to 0.1. We use the same parameter values for all of the sequences.

5.2. Overall Tracking Performance on OTB Benchmark dataset

To validate the performance of the proposed algorithm, we evaluate our algorithm on the OTB-50 and OTB-100 dataset that contain 50 and 100 challenging videos with comparisons to 10 state-of-the-art CF based algorithms, respectively, including STAPLE_CA [20], LCT [18], SAMF [17], KCF [13], DSST [16], CSK [15], SAMF_CA [20], DCF_CA [20], CFNet [44], and MOSSE_CA [20]. We evaluated 10 trackers by using three metrics that were provided by [31], and report the tracking results using Distance Precision (DP) and Overlap Success (OS). The Distance Precision (DP), which is defined as the percentage of frames whose predicted location is within the given threshold distance of the ground truth, and the given threshold value of DP is often defined as 20 pixels [31]. The overlap Precision OS, which is defined as the percentage of frames where the bounding box overlap surpasses a given threshold, and the given threshold value of OS is generally defined as 0.5 [31]. We report the tracking results in one-pass evaluation (OPE) while using distance precision plot and overlap success plot, as shown in Figure 3, and using the distance precision plot and the area-under-the-curve (AUC) of success plot to rank the trackers.

Figure 3 illustrates the distance precision, overlap success plots on OTB-50 dataset. We know that the proposed tracker performs well against state-of-the-art CF based trackers in DP and OS. Our tracker performs well with DP of 82.9% and OS of 55.1%, where the average DP of 82.9 outperformed STAPLE_CA (76.8), CFNet (76.8%), LCT (76.5%), DCF_CA (71.7%), and SAMF_CA (71.5%), and the average OS of 55.1 outperformed CFNet (54.7%), STAPLE_CA (54%), LCT (50.6%), DCF_CA (48%), and SAMF_CA (50.7%).

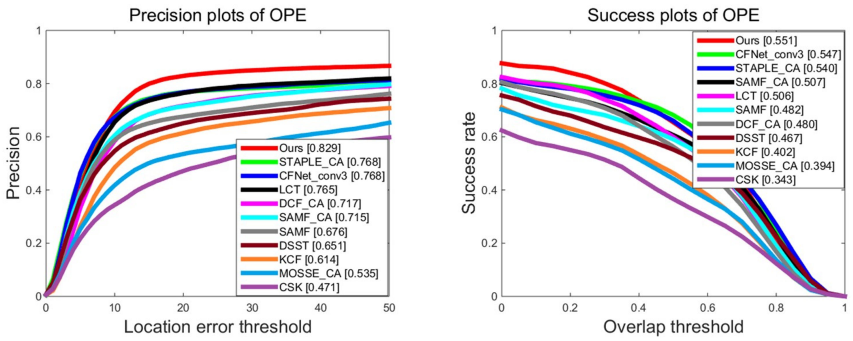


Figure 3. The overall results of distance precision and overlap success plots comparing the proposed tracker with state-of-the-art correlation filter (CF) based trackers on OTB-50 benchmark dataset using OPE.

Table 1 illustrates the average speed of the proposed tracker with comparisons to several CF based trackers on the OTB-100 benchmark dataset. We know that the KCF and CSK trackers gain the best speed follow by MOSSE_CA, DCF_CA. Our tracker mainly employs the hand-crafted features for tracking and the tracking speed outperforms SRDCF (5.3), LCT (21.2) and DSST (28.6) trackers, which obtained the real-time speed of 31.5 FPS. The DSST and SAMF mainly address the problem of scale variation, our algorithm employs the scale estimation approach from DSST, but our algorithm performs superiorly against the DSST as well as SAMF. These results that are displayed above also validate the effectiveness of the proposed approach to some extent.

Table 1. The average speed of the proposed algorithm with comparisons to several state-of-the-arts CF based trackers carry on OTB-100 benchmark dataset. The best and second results are highlighted by bold and underline.

	LCT	**SAMF**	**KCF**	**DSST**	**CSK**	**SAMF_CA**	**DCF_CA**	**STAPLE_CA**	**MOSSE_CA**	**Ours**
Avg. FPS	21.2	18.6	<u>212.6</u>	28.6	**266.8**	40.2	**90.2**	29.3	123.8	31.5

Figure 4 illustrate the tracking results of the proposed tracker with comparison to 10 state-of-the-art trackers on OTB-100 benchmark dataset, the proposed tracker performs well with DP of 82.4% and OS of 57.3%, where the average DP of 82.4% outperformed STAPLE_CA (79.2%), SAMF_CA (78.9%), LCT (78.6%), CFNet (77.1%), and DCF_CA (74.4). The average OS of 57.3% maintain similar accuracy (53.7%) to STAPLE_CA (57.4%) outperformed CFNet (57.1%), SAMF_CA (56.8%), LCT (54.2%), SAMF (53.5%), DSST (52.7%), and DCF_CA (51.5%). These results further demonstrate the effectiveness of the multiple features fusion method and the model updating strategy. The experiment is mainly evaluated on the OTB-100 benchmark dataset [32]. Although the proposed algorithm achieved excellent performance, relying solely on OTB is not sufficient, and it was not evaluated on the VOT 2016 dataset, this problem will be addressed in the future work.

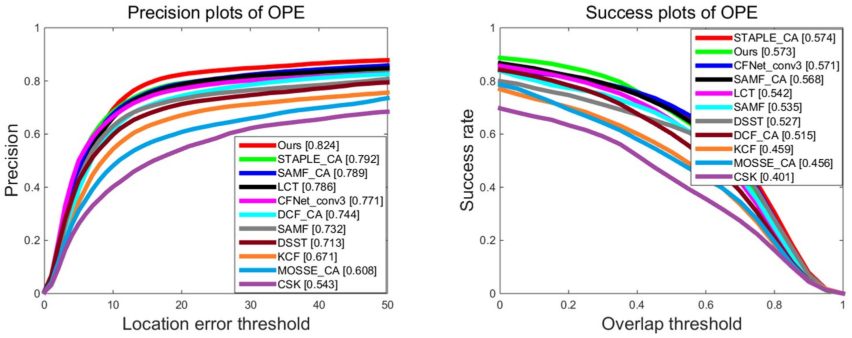


Figure 4. The tracking results of the proposed algorithm with comparisons to 10 state-of-the-arts CF based trackers on OTB-100 benchmark dataset.

5.3. Attribute Based Evaluation

We evaluate the performance of the proposed algorithm under different challenging attributes perform on benchmark dataset, which contain 50 video challenging sequences. All of these videos in the dataset are annotated by 11 attributes contain different challenging attributes, as following: Illumination Variation (IV), Scale Variation (SV), Occlusion (OCC), Deformation (DEF), Motion Blur (MB), Fast Motion (FM), ln-Plane-Rotation (IPR), Out-Plane-Rotation (OPR) Out-of-View (OV), Background Clutters (BC), and Low Resolution (LR). We report seven attributes results, as shown in Figure 5, and the number shown on the heading indicates the number of dataset with this challenge attribute. We can see that our tracker gain desirable tracking result in almost all the displayed attributes.

Figure 5 illustrates the proposed algorithm performs favorably with distance precision and overlap success plots in seven attribute challenges, it demonstrates that the proposed method achieve superior DP and OS in attributes of illumination variation (80.7%, 55.5%), in-plane rotation (82.4%, 54.4%), out-plane rotation (83.5%, 55.2%), background clutter (79.4%, 56.0%), out-of-view (75.6%, 51.0%), occlusion (84.0%, 54.2%), and scale variation (81.5%, 50.8%). These results further demonstrate the effectiveness of the proposed method to some extent.

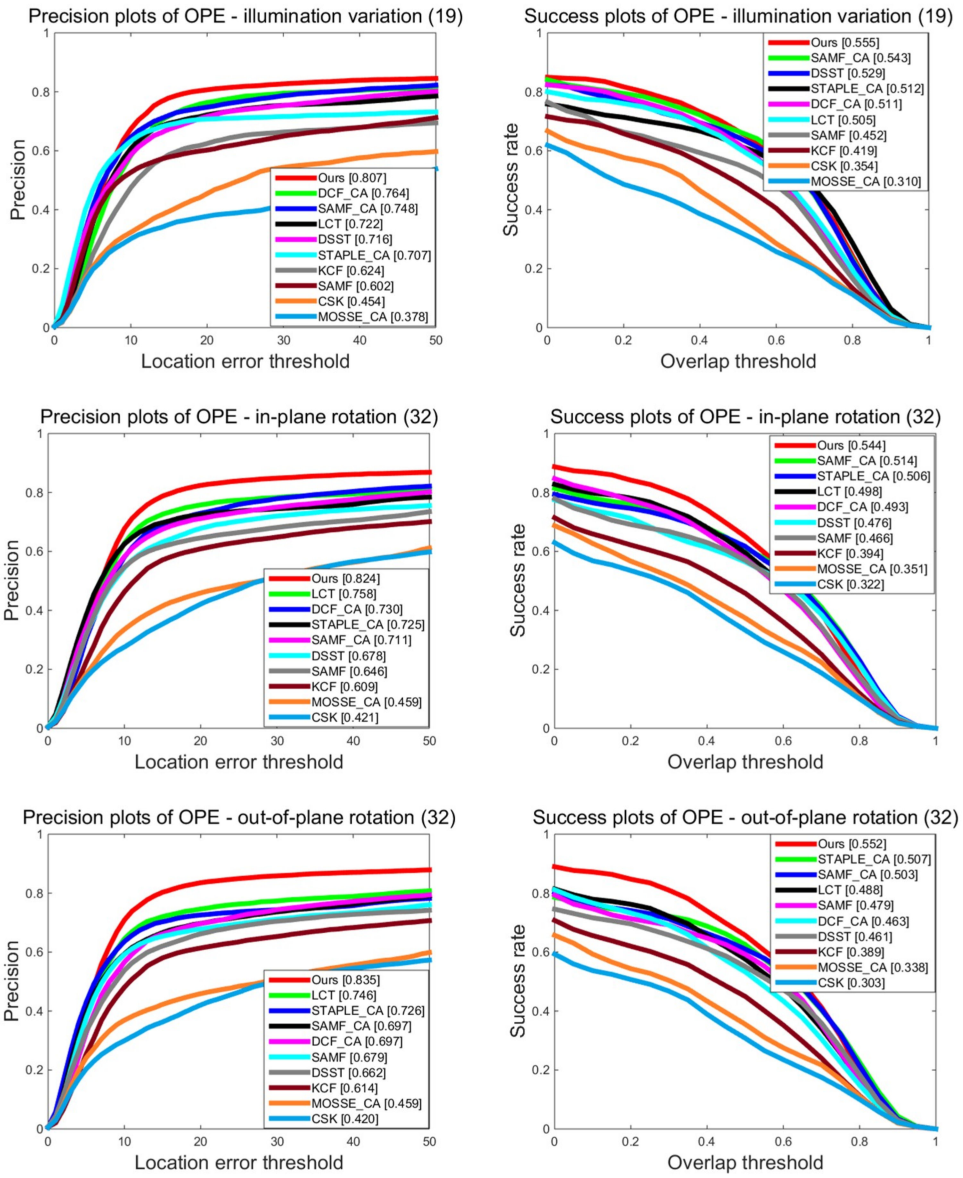


Figure 5. *Cont.*

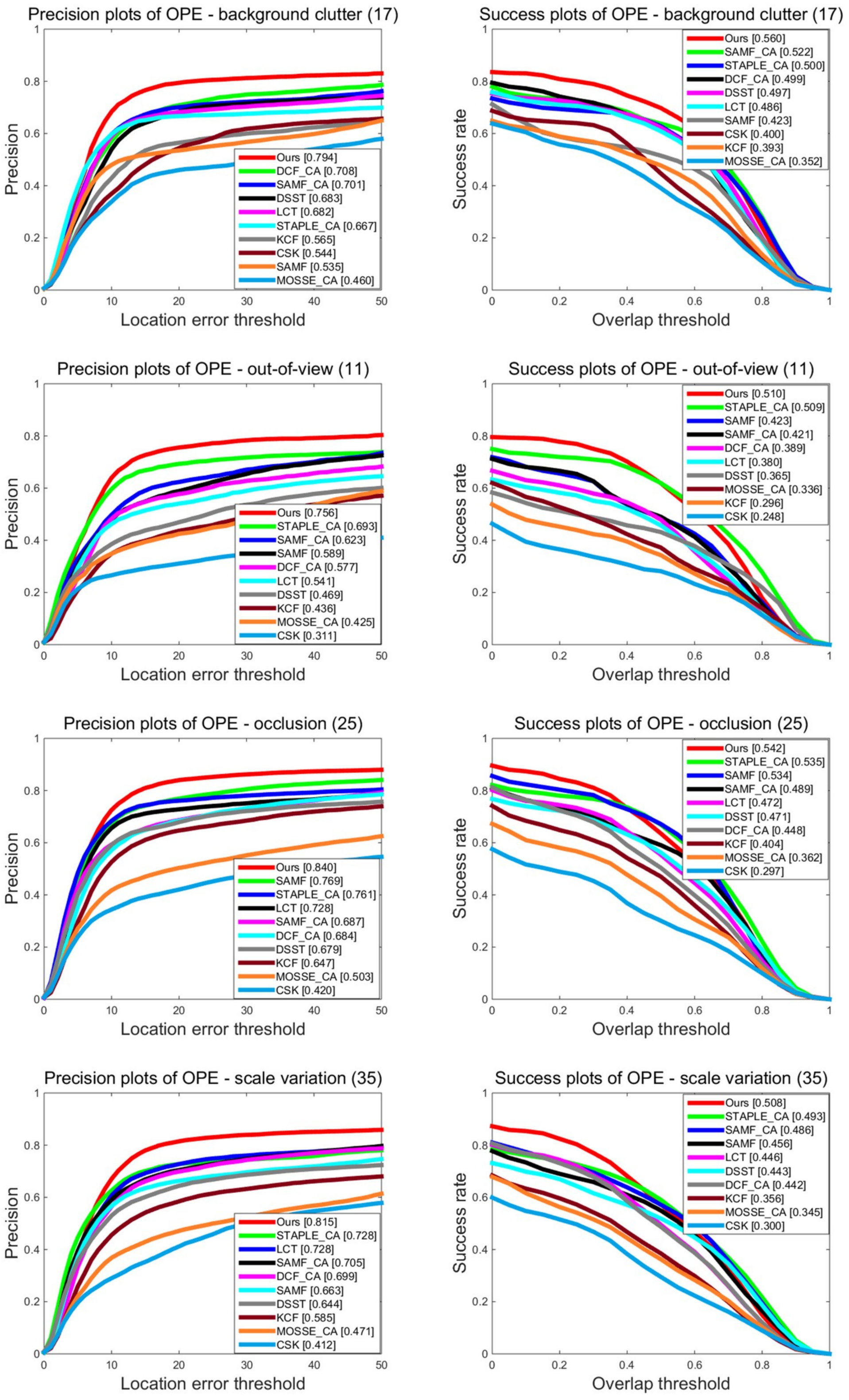


Figure 5. Attribute based evaluation of Success plots compare the proposed algorithm with 10 CF based trackers over seven challenges of illumination variation, in-plane rotation, out-of-plane rotation, background clutter, out-of-view, occlusion and scale variation. The number of sequences for each attribute is shown in brackets.

5.4. Qualitative Evaluation

Figure 6 illustrates qualitative comparisons of the proposed tracker compare with 10 CF based trackers carry on benchmark dataset [31,32], including LCT [18], SAMF [17], KCF [13], DSST [16], CSK [15], CT [7], SAMF_CA [20], DCF_CA, and MOSSE_CA [20]. We can see that the proposed tracker perform well in attributes with deformation (*Bird2*, *Couple*, *Singer2*, *Human9*), occlusion (*Bird2*, *ClifBar*, *Freeman4*), scale variation (*Couple*, *ClifBar*, *freeman4*, *Human9*), fast motion (*Couple*, *ClifBar*, *Human9*), motion blur (*ClifBar*, *Human9*), out-of-view (*ClifBa*r), and background clutter (*ClifBar*, *Couple*, *Singer2*).

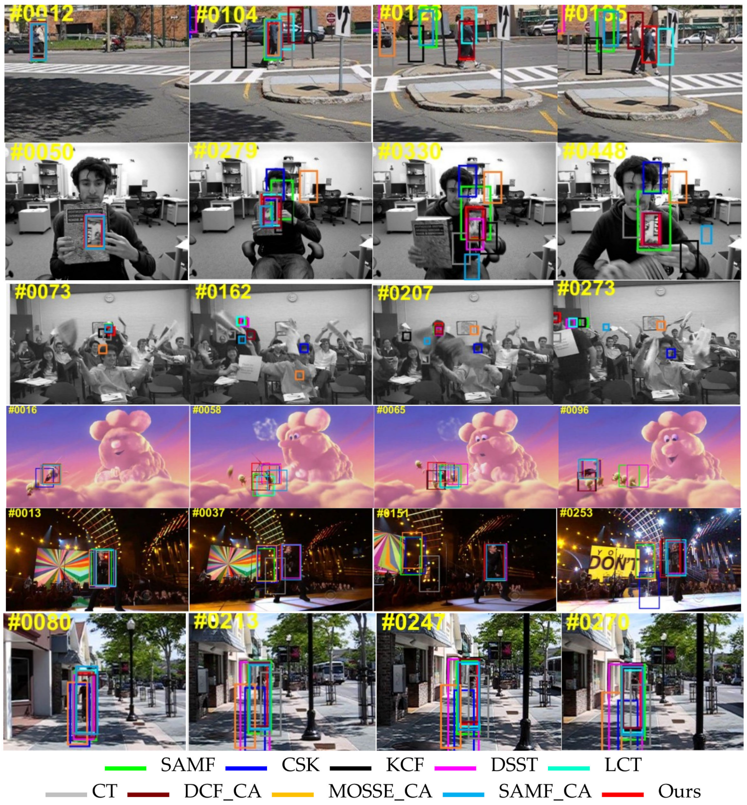


Figure 6. Qualitative comparison of the proposed algorithm with 10 trackers on 6 challenge sequences (from left to right and top to down are *Couple*, *ClifBar*, *Freeman4*, *Bird2*, *Singer2*, *Human9*). **Notation:** We denotes $\mathcal{F}^{-1}$ as the inverse Fourier transformation, and use a hat $\hat{}$ as shorthand for the DFT of a vector. The symbol * denotes conjugate operation, and the fraction denotes element-wise division. The bar G, F, and A represent complex conjugation and the product $\overline{G}F^l$ is point-wise. The symbol $\odot$ denotes the convolution operation.

Figure 6 illustrates qualitative evaluation of the proposed algorithm with comparisons to 10 state-of-the-art CF based trackers on six challenging sequences. In the *Couple* and *Singer2* sequences, the target objects mainly suffer from complicated challenge of deformation and background clutter. When compared with other trackers, the proposed tracker still track the target accurately, the other trackers, such as LCT, SAMF_CA are prone to lose the target, it demonstrates that the effectiveness of our method to some degree. In the *Freeman4* and *Bird2* sequences, the target objects mainly undergo some external interference such as occlusion and scale variation. As compared with other trackers, the proposed tracker locates the target accurately, and effectively alleviates tracking drift caused by challenges of scale changes and occlusion. In the *Human9* and *ClifBar* sequences, the target objects mainly suffer from challenges, such as fast motion, motion blur, and background clutter, the proposed tracker still locate the target reliably, it further validates the effectiveness of the proposed method.

5.5. Overall Tracking Performance on Temple Color Dataset

To further evaluate the proposed algorithm, we compare the proposed tracker on the Temple color dataset [45] containing 128 videos. The comparison with several state-of-the-arts CF based tracking algorithms, including STAPLE_CA [20], SAMF_CA [20], DCF_CA [20], MOSSE_CA [20], LCT [18], KCF [13], and SAMF [15]. The evaluation metrics is the same as the OTB benchmark dataset.

Figure 7 illustrate the tracking results of the proposed tracker with comparison to 7 state-of-the-art trackers on Temple Color dataset, the proposed tracker performs well with DP of 70.2% and OS of 52.1%, where the average DP of 70.2% outperformed STAPLE_CA (68.6%), SAMF_CA (67.4%), SAMF (66.5%), and LCT (62.2%). The average OS of 52.1% outperformed STAPLE_CA (51.7%), SAMF_CA (49.8%), SAMF (45.3%), and LCT (43.9%). These results further demonstrate the effectiveness of the proposed strategy.

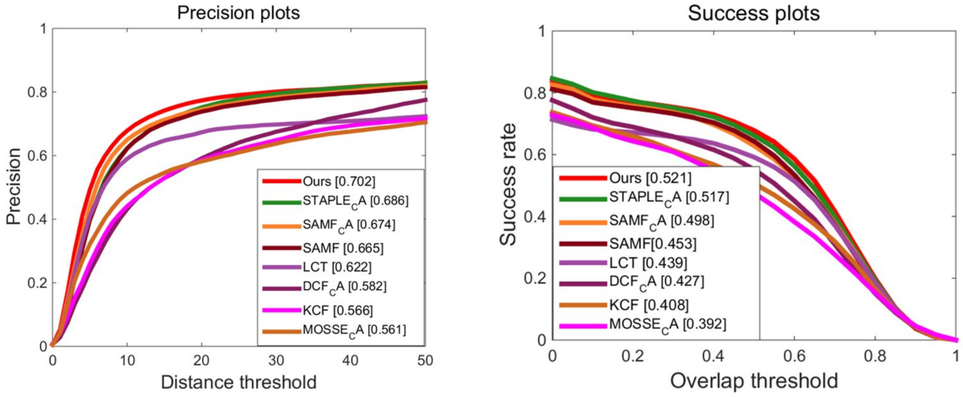


Figure 7. The overall results of distance precision and overlap success plots comparing the proposed tracker with state-of-the-art CF based trackers on Temple Color dataset using OPE.

6. Conclusions

In this paper, we propose a robust multi-scale correlation filter tracking algorithm via self-adaptive fusion of multiple hand-crafted features. We integrate multiple powerful discriminative hand-crafted features self-adaptively into the correlation filter framework, which achieve a preferable feature representation and obtain the real-time speed of 31.5 FPS. Moreover, we design a model update strategy to prevent the deteriorating of the tracking model due to noisy update. In addition, we introduce an accurate scale estimation method integrate with the model update strategy, which further improves scale change adaptability. The extensive experimental results demonstrate that the proposed algorithm perform superiorly against most state-of-the-art CF based algorithms. The proposed algorithm also has some several shortcomings, and there is much space for improvement in tracking performance and real-time performance. Integrating an accurate scale estimation strategy in our tracking framework can adapt to the scale change effectively and improve the accuracy of the algorithm throughout the tracking process. At the meanwhile, it also generates a large amount of computational burden, which affects the tracking real-time performance. In the future work, we will consider combining the deep feature with the hand-crafted feature, and design an efficient update mechanism in the model update, so that the algorithm can make better trade-offs between tracking performance and real-time performance.

Author Contributions: Z.C. designed the proposed strategy of tracking algorithm and wrote the paper; P.L., Y.D., Y.L. and W.Z. revised the paper and refined the manuscript.

Funding: This work was supported by the grants from National Natural Science Foundation, China under Grant 61605048, and Grant 61603144, and Grant 61403245 and Grant 91648119, in part by Natural Science Foundation of Fujian Province, China under Grant 2015J01256, and Grant 2016J01300, in part by the Talent project of Huaqiao University under Grant 14BS215, in part by Quanzhou scientific and technological planning projects of Fujian, China under Grant 2015Z120 and Grant 2017G024, and in part by the Subsidized Project for Postgraduates 'Innovative Fund in Scientific Research of Huaqiao University under Grant 1611422001.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Smeulders, A.W.; Chu, D.M.; Cucchiara, R.; Calderara, S.; Dehghan, A.; Shah, M. Visual tracking: An experimental survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 1442–1468. [PubMed]
2. Yilmaz, A.; Javed, O.; Shah, M. Object tracking: A survey. *ACM Comput. Surv.* **2006**, *38*, 13. [CrossRef]
3. Tsagkatakis, G.; Savakis, A. Online Distance Metric Learning for Object Tracking. *IEEE Trans. Circuits Syst. Video Technol.* **2011**, *21*, 1810–1821. [CrossRef]
4. Babenko, B.; Yang, M.H.; Belongie, S. Robust Object Tracking with Online Multiple Instance Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 1619–1632. [CrossRef] [PubMed]
5. Hare, S.; Saffari, A.; Torr PH, S. Struck: Structured output tracking with kernels. In Proceedings of the IEEE International Conference on Computer Vision 2011, Barcelona, Spain, 6–13 November 2011; pp. 263–270.
6. Kalal, Z.; Mikolajczyk, K.; Matas, J. Tracking-Learning-Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 1409–1422. [CrossRef] [PubMed]
7. Zhang, K.; Zhang, L.; Yang, M.H. Real-Time Compressive Tracking. In Proceedings of the European Conference on Computer Vision 2012, Florence, Italy, 7–13 October 2012; pp. 864–877.
8. Grabner, H.; Leistner, C.; Bischof, H. Semi-supervised On-Line Boosting for Robust Tracking. In Proceedings of the European Conference on Computer Vision 2008, Marseille, France, 12–18 October 2008; pp. 234–247.
9. Cauwenberghs, G.; Poggio, T. Incremental and decremental support vector machine learning. In *Proceedings of the International Conference on Neural Information Processing Systems 2000*; MIT Press: Cambridge, MA, USA, 2000; pp. 388–394.
10. Mei, X.; Ling, H. Robust visual tracking using ℓ 1 minimization. In Proceedings of the IEEE International Conference on Computer Vision 2009, Kyoto, Japan, 29 September–2 October 2009; pp. 1436–1443.
11. Ahuja, N. Robust visual tracking via multi-task sparse learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2012, Providence, RI, USA, 16–21 June 2012; pp. 2042–2049.
12. Fan, H.; Xiang, J. Robust Visual Tracking With Multitask Joint Dictionary Learning. *IEEE Trans. Circuits Syst. Video Technol.* **2017**, *27*, 1018–1030. [CrossRef]
13. Henriques, J.F.; Rui, C.; Martins, P.; Batista, J. High-Speed Tracking with Kernelized Correlation Filters. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *37*, 583–596. [CrossRef] [PubMed]
14. Bolme, D.S.; Beveridge, J.R.; Draper, B.A.; Liu, Y.M. Visual object tracking using adaptive correlation filters. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2010, San Francisco, CA, USA, 13–18 June 2010; pp. 2544–2550.
15. Rui, C.; Martins, P.; Batista, J. Exploiting the circulant structure of tracking-by-detection with kernels. In Proceedings of the European Conference on Computer Vision 2012, Florence, Italy, 7–13 October 2012; pp. 702–715.
16. Danelljan, M.; Häger, G.; Khan, F.S.; Felsberg, M. Accurate Scale Estimation for Robust Visual Tracking. In Proceedings of the British Machine Vision Conference 2014; BMVA Press: Surrey, UK, 2014; pp. 65.1–65.11.
17. Li, Y.; Zhu, J. A Scale Adaptive Kernel Correlation Filter Tracker with Feature Integration. In *European Conference on Computer Vision 2014*; Springer: Cham, Germany, 2014; pp. 254–265.
18. Ma, C.; Yang, X.; Zhang, C.; Yang, M.H. Long-term correlation tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2015, Boston, MA, USA, 7–12 June 2015; pp. 5388–5396.
19. Danelljan, M.; Häger, G.; Khan, F.S.; Felsberg, M. Learning Spatially Regularized Correlation Filters for Visual Tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2015, Boston, MA, USA, 7–12 June 2015; pp. 4310–4318.
20. Mueller, M.; Smith, N.; Ghanem, B. Context-Aware Correlation Filter Tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 138–1395.
21. Lukezic, A.; Vojir, T.; Zajc, L.C.; Matas, J.; Kristan, M. Discriminative correlation filter with channel and spatial reliability. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; Volume 2.

22. Tu, Z. Probabilistic Boosting-Tree. Learning Discriminative Models for Classification, Recognition, and Clustering. In Proceedings of the IEEE International Conference on Computer Vision 2005, San Diego, CA, USA, 20–25 June 2005; pp. 1589–1596.
23. Avidan, S. Ensemble Tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 261–271. [CrossRef] [PubMed]
24. Wang, S.; Lu, H.; Yang, F.; Yang, M.H. Super pixel tracking. In Proceedings of the IEEE International Conference on Computer Vision 2011, Barcelona, Spain, 6–13 November 2011; pp. 1323–1330.
25. Grabner, H.; Grabner, M.; Bischof, H. Real-time tracking via online boosting. *Br. Mach. Vis. Assoc. BMVC* **2006**, *1*, 47–56.
26. Dalal, N.; Triggs, B. Histograms of Oriented Gradients for Human Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2005, San Diego, CA, USA, 20–25 June 2005; pp. 886–893.
27. Danelljan, M.; Khan, F.S.; 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 2014, Columbus, OH, USA, 23–28 June 2014; pp. 1090–1097.
28. Bertinetto, L.; Valmadre, J.; Golodetz, S.; Miksik, O.; Torr, P.H.S. Staple: Complementary Learners for Real-Time Tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016, Las Vegas, NV, USA, 27–30 June 2016; pp. 1401–1409.
29. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv*, 2014; arXiv:1409.1556.
30. Danelljan, M.; Robinson, A.; Khan, F.S.; Felsberg, M. Beyond Correlation Filters: Learning Continuous Convolution Operators for Visual Tracking. In Proceedings of the European Conference on Computer Vision 2016, Amsterdam, The Netherlands, 11–14 October 2016; pp. 472–488.
31. Wu, Y.; Lim, J.; Yang, M.H. Online Object Tracking: A Benchmark. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2013, Portland, OR, USA, 23–28 June 2013; pp. 2411–2418.
32. Wu, Y.; Lim, J.; Yang, M.H. Object Tracking Benchmark. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1834–1848. [CrossRef] [PubMed]
33. Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52. [CrossRef]
34. Wang, M.; Liu, Y.; Huang, Z. Large Margin Object Tracking with Circulant Feature Maps. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4800–4808.
35. Danelljan, M.; Häger, G.; Khan, F.S.; Felsberg, M. Coloring Channel Representations for Visual Tracking. In Proceedings of the Scandinavian Conference on Image Analysis; Springer International Publishing: New York, NY, USA, 2015; pp. 117–129.
36. Danelljan, M.; Hager, G.; Khan, F.S.; Felsberg, M. Convolutional Features for Correlation Filter Based Visual Tracking. In Proceedings of the IEEE International Conference on Computer Vision Workshop 2016, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 621–629.
37. Felzenszwalb, P.F.; Girshick, R.B.; McAllester, D.A.; Ramanan, D. Object detection with discriminatively trained part-based models. *PAMI* **2010**, *32*, 1627–1645. [CrossRef] [PubMed]
38. Berlin, B.; Kay, P. *Basic Color Terms: Their Universality and Evolution*; University of California Press: Berkeley, CA, USA, 1969; Volume 6, p. 151.
39. Khan, F.S.; Anwer, R.M.; van de Weijer, J.; Bagdanov, A.; Lopez, A.; Felsberg, M. Coloring action recognition in still images. *IJCV* **2013**, *105*, 205–221. [CrossRef]
40. Zhang, J.; Ma, S.; Sclaroff, S. MEEM: Robust Tracking via Multiple Experts Using Entropy Minimization. In Proceedings of the European Conference on Computer Vision 2014; Springer: Cham, Germany, 2014; pp. 188–203.
41. Avidan, S.; Levi, D.; Barhillel, A.; Oron, S. Locally Orderless Tracking. *Int. J. Comput. Vis.* **2015**, *111*, 213–228.
42. Aceto, G.; Ciuonzo, D.; Montieri, A.; Pescapé, A. Multi-classification approaches for classifying mobile app traffic. *J. Netw. Comput. Appl.* **2018**, *103*, 131–145. [CrossRef]
43. Aceto, G.; Ciuonzo, D.; Montieri, A.; Pescapé, A. Traffic Classification of Mobile Apps through Multi-classification. In *GLOBECOM 2017—2017 IEEE Global Communications Conference*; IEEE: Piscataway, NJ, USA, 2017.

44. Valmadre, J.; Bertinetto, L.; Henriques, J.; Vedaldi, A.; Torr, P.H. End-to-end representation learning for correlation filter based tracking. In *Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2017.
45. Liang, P.; Blasch, E.; Ling, H. Encoding Color Information for Visual Tracking: Algorithms and Benchmark. *IEEE Trans. Image Process.* **2015**, *24*, 5630–5644. [CrossRef] [PubMed]