IEICE TRANSACTIONS

on Information and Systems

VOL. E101-D NO. 11
NOVEMBER 2018

A PUBLICATION OF THE INFORMATION AND SYSTEMS SOCIETY
The Institute of Electronics, Information and Communication Engineers
Kikai-Shinko-Kaikan Bldg., 5-8, Shibakoen 3 chome, Minato-ku, TOKYO, 105-0011 JAPAN

LETTER

Adaptive Object Tracking with Complementary Models*

Peng GAO[†a)], ***Member***, **Yipeng MA**[†], **Chao LI**[†], **Ke SONG**[†], **Yan ZHANG**[†], **Fei WANG**[†b)], ***and*** **Liyi XIAO**[†], ***Nonmembers***

SUMMARY Most state-of-the-art discriminative tracking approaches are based on either template appearance models or statistical appearance models. Despite template appearance models have shown excellent performance, they perform poorly when the target appearance changes rapidly. In contrast, statistic appearance models are insensitive to fast target state changes, but they yield inferior tracking results in challenging scenarios such as illumination variations and background clutters. In this paper, we propose an adaptive object tracking approach with complementary models based on template and statistical appearance models. Both of these models are unified via our novel combination strategy. In addition, we introduce an efficient update scheme to improve the performance of our approach. Experimental results demonstrate that our approach achieves superior performance at speeds that far exceed the frame-rate requirement on recent tracking benchmarks.

key words: *object tracking, discriminative correlation filter, adaptive combination, model update*

1. Introduction

Visual object tracking is a fundamental research problem in the field of computer vision with various applications such as robotic services, smart surveillance systems, autonomous driving and intelligent traffic control. Given the initial state of an arbitrary target in the first frame, i.e., only the initial target position and size are identified, the goal of tracking is to learn a classifier or a regressor to estimate target positions by discriminating the target from the background in a video. Despite impressive progress has been made in the past decade, object tracking remains a large unsolved problem due to numerous challenging factors such as fast motions, background clutters, motion blurs, deformations, illumination variations, in-plane rotations, low resolution, occlusions, out-of-plane rotations, out of views and scale variations.

Recently, discriminative correlation filter (DCF) becomes one of the most attractive tracking models, and DCF-based tracking approaches have shown appealing performance improvements in terms of accuracy, robustness and speed on recent benchmarks [1] since they treat the tracking task as a similarity learning problem. Most state-of-the-art DCF-based approaches employ template appearance features including raw pixels [2], histogram of oriented gradients (HOG) [3] and color names (CN) [4] to represent the target during tracking. Despite their poor performances on rapidly changes of target appearance such as fast motions and fast deformations, they still have shown excellent performance in other challenging scenarios. In contrast to template appearance features, statistical appearance features such as color histograms are insensitive to fast target state changes. Unfortunately they yield inferior performance on illumination variations and background clutters because of such models may drift into adjacent regions which represent similar characteristics as the target [5]. It has been demonstrated that exploiting powerful feature representations of the target and enhancing the diversity of features can significantly improve tracking performance [6]. Some recent works concentrate on combining the above two models to train a robust model [7], [8], and they have shown favorable performance in target state changes and color variations. Since these works unified template and statistical models by simply linear interpolation, and the interpolation factor is always preset in advance, their tracking performances have gradually faded.

In this paper, in order to overcome the limitations above-mentioned, an adaptive object tracking approach (CACT) with both template and statistical appearance models is proposed. The template appearance model is constructed by a DCF-based framework and employs raw pixel, HOG and CN to represent the target. The statistical appearance model is implemented based on standard color histograms. We cast these two different models as independent linear regression problems. Then, an adaptive combination method is introduced, which can get more robust and reliable tracking results. Next, we suggest an efficient update scheme to deal with more challenging scenarios. Finally, experimental results demonstrate that our approach outperforms most state-of-the-art trackers with faster speed on recent benchmarks.

2. The Proposed Approach

In this section, we first introduce the motivation and formulation of complementary models. Next, we deduce an adaptive combination strategy that aims to reciprocally compen-

Manuscript received April 10, 2018.
Manuscript revised July 10, 2018.
Manuscript publicized August 6, 2018.

[†]The authors are with Shenzhen Graduate School, Harbin Institute of Technology, China.

*This work was supported by the Science and Technology Planning Program of Guangdong Province, China (No. 2013B090600105, No. 2016B090918047).

a) E-mail: pgao.hit@gmail.com
b) E-mail: 1392135844@qq.com (Corresponding author)

DOI: 10.1587/transinf.2018EDL8074

sate the deficiencies of either template or statistical models. Finally, a robust model update scheme is proposed to reduce the risk of severe model corruption caused by similar target or background representations.

2.1 Complementary Models

In our approach, we exploit DCF to construct the template appearance model due to its excellent performance and high computational efficiency, and use color histograms to construct the statistical appearance model. In order to solve complementary models correctly, we treat them as two independent regression problems since the benefits of the closed-form solution can make optimization problems much easier to be solved [3], [9].

We train a ridge regression on L-channel template features $\mathbf{x}_{t,u} = \phi_t(\mathrm{I}_u) \in \mathbb{R}^{M\times N\times L}$ including raw pixels, HOG and CN extracted from a searching patch $\mathrm{I}_u \in \mathbb{R}^{M\times N}$ at frame u, and the desired correlation output $\mathbf{y}_{t,u} \in \mathbb{R}^{M\times N}$ which typically follow a Gaussian function with the peak 1 at the center. Our aim is to find optimal correlation filters $\mathbf{w}_u$. This can be obtained by minimizing the ridge regression in the Fourier frequency domain as,

$$\min_{\mathbf{w}_u^l} \| \sum_{l=1}^{L} \mathbf{w}_u^l \star \mathbf{x}_{t,u}^l - \mathbf{y}_{t,u}\|^2 + \lambda \sum_{l=1}^{L} \|\mathbf{w}_u^l\|^2 \tag{1}$$

where $\mathbf{w}_u^l$ stands for the l-th channel of correlation filters $\mathbf{w}_u$ in the frame u, the star symbol $\star$ denotes circular cross-correlation, $\mathbf{x}_{t,u}^l$, $\mathbf{y}_{t,u}$ and $\mathbf{w}_u^l$ are all of size $M \times N$. Following the Parseval's theorem and adopting properties of circulant matrices mentioned in [9], the solution can be gained as,

$$\mathcal{W}_u^l = \frac{\mathcal{Y}_{t,u}^* \odot \mathcal{X}_{t,u}^l}{\sum_{l=1}^{L} (\mathcal{X}_{t,u}^l)^* \odot \mathcal{X}_{t,u}^l + \lambda} \tag{2}$$

Here $\mathcal{W}_u = \mathcal{F}(\mathbf{w}_u)$, and $\mathcal{F}(\cdot)$ indicates the Discrete Fourier Transform (DFT). We using the symbol $*$ to represent the complex conjugation $\mathcal{Y}_u^*$ of a complex number $\mathcal{Y}_u$, the symbol $\odot$ stands for the element-wise multiplication, and symbol $\div$ denotes the element-wise division. Additionally, the regularization coefficient λ alleviates division-by-zero.

During detection, given template appearance features $\mathbf{z}_{t,u+1}$ which are extracted from a searching patch centered at the previous target position in the new frame $(u+1)$, the template response map can be obtained by

$$\tilde{\mathbf{y}}_{t,u+1} = \mathcal{F}^{-1}(\sum_{l=1}^{L} \mathcal{W}_s^l \odot \mathcal{Z}_{t,u+1}^l) \tag{3}$$

where $\tilde{\mathbf{y}}_{t,u+1} \in \mathbb{R}^{M\times N}$ is the response map of predicted states in the new frame, and $\mathcal{F}^{-1}(\cdot)$ denotes the Inverse Discrete Fourier Transform (IDFT). The pipeline of the template appearance model as denoted by the black line in Fig. 1.

In order to learn the statistical model, we apply a ridge regression with K-channel statistical features $\mathbf{x}_{s,u} = \phi_s(\mathrm{I}_u) \in \mathbb{R}^{M\times N\times K}$ over the foreground and background regions F_u and B_u which are independently cropped from the same searching patch $\mathrm{I}_u \in \mathbb{R}^{M\times N}$ as well as the template model. Here, the foreground region F_u is surrounded by the background region B_u. The statistical model can be formulated as,

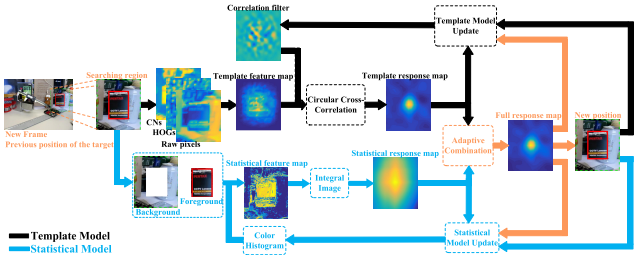

Fig. 1 The pipeline of our proposed approach. Best viewed in color.

$$\min_{\mathbf{v}_u} \sum_{\mathbf{x}_{s,u}} \|\mathbf{v}_u^T \mathbf{x}_{s,u} - \mathbf{y}_{s,u}\|^2 \tag{4}$$

where $\mathbf{y}_{s,u}$ is the corresponding label, i.e. $\mathbf{y}_{s,u} = 1$ for positive samples and $\mathbf{y}_{s,u} = 0$ for negative samples, $\mathbf{v}_u$ is a parameter vector. We decompose the regression into two independent terms of each region,

$$\min_{\mathbf{v}_u} \sum_{\mathbf{x}_{s,u}\in\phi_s(\mathrm{F}_u)} \|\mathbf{v}_u^T \mathbf{x}_{s,u} - 1\|^2 + \sum_{\mathbf{x}_{s,u}\in\phi_s(\mathrm{B}_u)} \|\mathbf{v}_u^T \mathbf{x}_{s,u} - 0\|^2 \tag{5}$$

Here, the sparse inner product is simply a lookup matrix, i.e. $\mathbf{v}_u^T \mathbf{x}_{s,u} = \mathbf{v}_u^{\mathbf{x}_{s,u}}$, where $\mathbf{v}_u^{\mathbf{x}_{s,u}}$ indicates the element of $\mathbf{v}_u$ whose channel index is non-zero. Then, we adopt the color histogram $\mathcal{H}(\mathbf{x}_{s,u})$ to represent target pixels, the solution can be obtained as,

$$\mathbf{v}_{u+1} = \begin{cases} \frac{\mathcal{H}_{\mathrm{F}_u}(\mathbf{x}_{s,u})}{\mathcal{H}_{\mathrm{F}_u}(\mathbf{x}_{s,u}) + \mathcal{H}_{\mathrm{B}_u}(\mathbf{x}_{s,u})} & \text{if } \mathbf{x}_{s,u} \in \mathrm{F}_u \cup \mathrm{B}_u \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

Since $\mathbf{x}_{s,u}$ is sparse, the statistical model can be calculated over a large detection region with a very low computational burden.

Finally, by exploiting the integral image method [10] to obtain the statistical response map, the detection process can be accelerated significantly.

$$\begin{aligned} \tilde{\mathbf{y}}_{s,u+1}(i,j) =& \mathbf{v}_{u+1}(i,j) + \tilde{\mathbf{y}}_{s,u+1}(i,j-1) + \\ & \tilde{\mathbf{y}}_{s,u+1}(i-1,j) - \tilde{\mathbf{y}}_{s,u+1}(i-1,j-1) \end{aligned} \tag{7}$$

where $\tilde{\mathbf{y}}_{s,u+1}(i,-1) = 0$ and $\tilde{\mathbf{y}}_{s,u+1}(-1,j) = 0$. The pipeline of the statistical appearance model as indicated by the blue line in Fig. 1.

2.2 Adaptive Combination

Ideally, the full response map $\tilde{\mathbf{y}}_u$ of the target in a new frame u is obtained by unifying both template and statistical response maps as,

$$\tilde{\mathbf{y}}_u = \alpha \tilde{\mathbf{y}}_{t,u} + (1-\alpha)\tilde{\mathbf{y}}_{s,u} \tag{8}$$

where α is an interpolation factor. In order to find the target position in a new frame, we choose the candidate patch

with the maximum similarity to the previous appearance of the target. To achieve this, we propose an efficient adaptive combination strategy by exploiting a novel measurement of response reliability, namely Peak to Response Fluctuation (PRF), which is defined as,

$$PRF_u = \frac{(\tilde{y}_{t,max,u} - \mu_{t,u})^2}{\sigma^2_{t,u}} \tag{9}$$

where $\tilde{y}_{t,max,u}$, $\mu_{t,u}$ and $\sigma^2_{t,u}$ are the peak, mean and variance values of the template response map, respectively. The criterion indicates the reliability of the template response map. The higher the template response peak score, the larger the PRF_u and the more reliable the template response map. In practice, we observe the fact that when PRF_u is larger than a threshold θ_{max}, we can use the template response map as the major distribution assign to the full response map. However, when PRF_u is less than a threshold θ_{min}, the statistical response map is more reliable than the template response map. Based on this motivation, we exploit a dynamic interpolation factor to reciprocally compensate the deficiencies of the two different models,

$$\alpha = \begin{cases} \alpha_{min} & PRF_u < \theta_{min} \\ \alpha_{max} - \frac{\theta_{max} - PRF_u}{\theta_{max}} & \theta_{min} \le PRF_u < \theta_{max} \\ \alpha_{max} & \text{otherwise} \end{cases} \tag{10}$$

In this work, the maximum and minimum of interpolation factor are preset as 0.8 and 0.2 respectively, and θ_{min} and θ_{max} are predefined as 8 and 20 respectively. Then, we can obtain the full response map using (8).

2.3 Model Update

Most existing tracking approaches assuming that tracking results are always accurate and update tracking models every frame. However, continuous model update may cause tracking failure once the tracker encounters some complicated scenarios such as occlusions and fast motions. In order to alleviate this issue, we utilize feedback from tracking results to decide the necessity for model updating. We consider three criteria: the highest peak values of the full response map $\max(\tilde{\mathbf{y}}_u)$, the highest peak values of template response map $\max(\tilde{\mathbf{y}}_{t,u})$ and PRF. When these three criteria of the current frame are great than their respective average values of the last 10 frames with certain ratios γ_f, γ_t and γ_{prf} (e.g. $\gamma_f = 0.5$, $\gamma_t = 0.7$ and $\gamma_{prf} = 0.5$ in this work), the tracking result in the current frame is considered to be high reliable.

Therefore, we can update complementary models as,

$$\begin{aligned} \mathcal{W}^u_{new} &= (1 - \eta_t)\mathcal{W}^{u-1} + \eta_t \mathcal{W}^u \\ \mathcal{H}^u_{new} &= (1 - \eta_s)\mathcal{H}^{u-1} + \eta_s \mathcal{H}^u \end{aligned} \tag{11}$$

where η_t and η_s are the learning rate preset as 0.02 and 0.04.

Additionally, during tracking, we first estimate the target position, then search on the scale by adopting a multi-scale estimation model similar to [3]. The multi-scale template model is only executed when the full response map is more reliable.

3. Experiments

In this section, we first evaluate efficiency improvement of our complementary models, adaptive combination and model update scheme. Then we compare our trackers with other state-of-the-art trackers to demonstrate the effectiveness of our proposed approach.

3.1 Experimental Setup

The proposed approach is implemented in Matlab 2014b on a PC equipped with an Intel i5-4590 CPU at 3.3GHz and 8GB RAM. We employ HOG and CN for target representations in the template appearance model and the scale estimation model. The cell size of HOG is 4×4, and the orientation bin number of HOG is 9. In the statistic appearance model, histograms of color values in the RGB color cube with 32 bins per channel are exploited. To compute histograms more accurate, we define 85% the previous target size as the foreground region to avoid mislabeling the target. The standard deviation of the desired response output is set to 1/16. Similar to [3], we use ten scales with a scale factor of 1.02 in the scale estimation model.

Experiments are conducted on the OTB-2015 benchmark which contains 100 image sequences [1]. The results are reported in terms of two evaluation metrics: (a) distance precision (DP), which is the percentage of frames that the average Euclidean distance between the ground-truth and the estimated target position is smaller than 20 pixels, (b) overlap precision (OP), which is the percentage of frames that the overlap between the estimated target size and the ground truth exceeds the fixed threshold of 0.5. We exploit one-pass evaluation (OPE) to run the tracker throughout the test sequence and report results in success plot and precision plot. Success and precision plots illustrate OP and DP of trackers over a range of thresholds. In the success plot, trackers are ranked according to the area under the curve (AUC) score. In addition, we also report the speed of trackers in average frames per second (FPS).

3.2 Ablation Experiments

To analyze efficiencies of the proposed complementary model, adaptive combination strategy and model update scheme, we evaluate several variants of our proposed approach with different experimental setup on the OTB-2015 benchmark. CACT$_{all}$ is implemented using all proposed methods, CACT$_{nmu}$ denotes that no model update scheme is employed and CACT$_{nac}$ is designed without using the adaptive combination strategy. The experiment is also including three other tracking approaches, Staple [7] as abovementioned, DSST [3] only uses the template appearance model, and DAT [5] only uses the statistical appearance model.

Characteristics and tracking results are summarized in

Table 1. CACT$_{all}$ shows the best tracking performance both in terms of accuracy and robustness, which significantly outperforms DSST [3] and an absolute gain of 8.0% in the mean OP and enhances the mean DP from 68.0% to 80.2%. Compared with Staple [7], performance can be further improved by utilizing our model update scheme and adaptive combination strategy in CACT$_{nac}$ and CACT$_{nmu}$, respectively. It is evident that all our proposed methods are efficient and effective for visual tracking.

As introduced in Sect. 2.2, the interpolation factor α in Eq. (10) is determined by four threshold parameters $[\alpha_{min}, \alpha_{max}, \theta_{min}, \theta_{max}]$. We use CACT$_{all}$ to evaluate four sets of $[\alpha_{min}, \alpha_{max}]$: [0.4, 0.6], [0.3, 0.7], [0.2, 0.8] and [0.1, 0.9], which are denoted as α_1, α_2, α_3 and α_4, and five sets of $[\theta_{min}, \theta_{max}]$ are tried: [6, 16], [6, 20], [8, 20], [8, 24] and [12, 20], which are denoted as θ_1, θ_2, θ_3, θ_4 and θ_5. The heat map of Fig. 2 demonstrates how different combinations of $[\alpha_{min}, \alpha_{max}]$ and $[\theta_{min}, \theta_{max}]$ affect tracking performance on the OTB-2015 benchmark. The best mean AUC score of 59.3% is achieved at the combination of α_3 and θ_3.

Figure 3 demonstrates a detailed comparison of CACT$_{all}$ and CACT$_{nac}$ on the 117^{th} frame of the sequence *Girl2* from the OTB-2015 benchmark. Although peak values are located at the wrong positions on template response maps of CACT$_{all}$ and CACT$_{nac}$, the target position can be still estimated correctly by CACT$_{all}$ since the *PRF* value is too low to use more template response scores and it should consider more statistical response scores for accurate localization. Without the adaptive combination strategy, CACT$_{nac}$ estimates the target at wrong position.

Moreover, we visualize the impact of variations of adaptive interpolation factor during tracking on the sequence *Box* in Fig. 4. The factor fluctuates within a small range to adaptively combine both template and statistical response maps in normal scenarios. However, when fast

Table 1 Characteristics and tracking results of compared approaches on the OTB-2015 benchmark. MU, AC and CM indicate whether the approach exploits model update, adaptive combination or complementary models, respectively. The best results are highlighted in **bold**.

Approaches	MU	AC	CM	mean DP (%)	mean AUC (%)
CACT$_{all}$	√	√	√	**80.2**	**59.3**
CACT$_{nac}$	√	-	√	79.0	58.6
CACT$_{nmu}$	-	√	√	79.3	58.5
Staple [7]	-	-	√	78.4	58.1
DSST [3]	-	-	-	68.0	51.3
DAT [5]	-	-	-	43.7	33.5

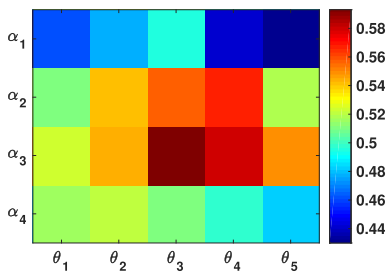

Fig. 2 Tracking results of different threshold parameters used in the adaptive combination method on the OTB-2015 benchmark. The results are presented in mean AUC scores. Best viewed in color.

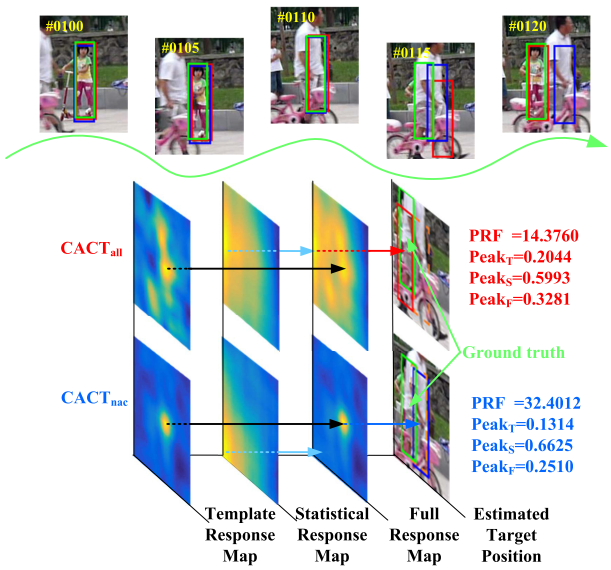

Fig. 3 Comparisons of CACT$_{all}$ and CACT$_{nac}$ on the 117^{th} frame of the sequence *Girl2*. Green boxes indicate ground truth, red and blue boxes denote tracking results of CACT$_{all}$ and CACT$_{nac}$ respectively, orange boxes stand for previous estimated target regions. Best viewed in color.

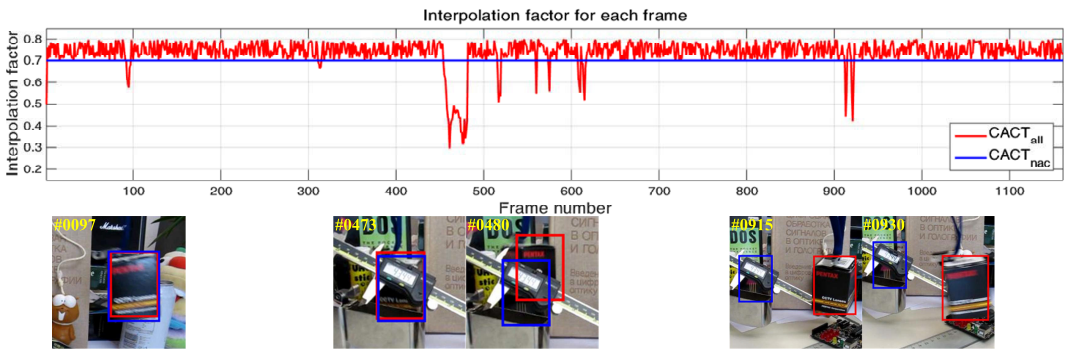

Fig. 4 Adaptive interpolation factors of each frame on the sequence *Box* from the OTB-2015 benchmark. It can be seen that CACT$_{nac}$ has already lost the target after the 473^{rd} frame, which CACT$_{all}$ keeps tracking the target until the end of the sequence.

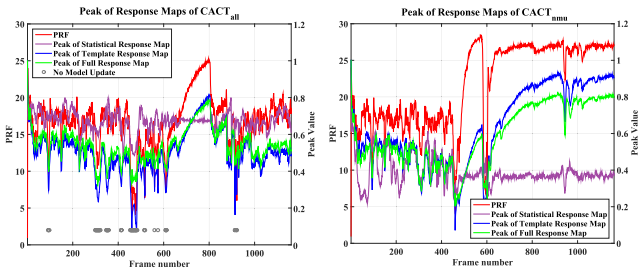

Fig. 5 *PRF* and peak values of $\tilde{\mathbf{y}}_u$, $\tilde{\mathbf{y}}_{t,u}$ and $\tilde{\mathbf{y}}_{s,u}$ on the sequence *Box* from the OTB-2015 benchmark. Since CACT$_{nmu}$ updates models at every frame, it has missed the target after occlusions occur from the 473rd frame. In contrast, CACT$_{all}$ can detect the target accurately since it updates models infrequently. The gray circle in the left figure denotes CACT$_{all}$ does not update models at this frame. Best viewed in color.

deformations (the 97th frame), occlusions (the 473rd and the 480th frames) or out-of-plane rotations (the 915th and the 930th frames) occur, the adaptive interpolation factor decreases significantly since the template response map is unreliable to handle those challenging scenarios, hence, we should consider more about the statistical response map. In contrast, trackers that use a fixed interpolation factor (e.g., 0.7 in CACT$_{nac}$) tend to tracking failures when more challenging scenarios occur.

Finally, we estimate the contribution of model update scheme to the overall tracking performance. *PRF* and peak values on $\tilde{\mathbf{y}}_u$, $\tilde{\mathbf{y}}_{t,u}$ and $\tilde{\mathbf{y}}_{s,u}$ of CACT$_{all}$ and CACT$_{nmu}$ are illustrated in Fig. 5. CACT$_{all}$ does not update models since *PRF*, $\tilde{\mathbf{y}}_{t,u}$ and $\tilde{\mathbf{y}}_u$ decrease dramatically after the occlusion occur (the 460th frame), so that it can efficiently avoid model corruption. However, due to CACT$_{nmu}$ updates models in each frame, these models have been corrupted by uncertain information and *PRF* fluctuates at a high value, which leads to wrong tracking results. Figure 6 shows a concrete example of the model update scheme. It is clear that when the target is completely occluded, CACT$_{all}$ can successfully redetect the target in sequential frames by the model update scheme. But CACT$_{nmu}$ has lost the target after the occlusion occurs since it updates models every frame with similar target or background noise. We denote CACT$_{all}$ as CACT to compare with several state-of-the-art tracking approaches in following evaluations.

3.3 Evaluations on OTB

We evaluate our approach (CACT) with 7 state-of-the-art trackers including Staple [7], SAMF [2], DLSSVM [11], DAT [5], DSST [3], KCF [9] and Struck [12]. Figure 7 illustrates success and precision plots of all compared approaches.

As the best tracker when the original OTB benchmark [1] came out, the structured support vector machine (SSVM) based tracker, i.e., Struck, obtains a mean AUC score of 46.1%. But the kernelized correlation filter based method, i.e., KCF, outperforms Struck and obtains a gain of 1.6% in mean AUC score. SAMF and DSST extend KCF by

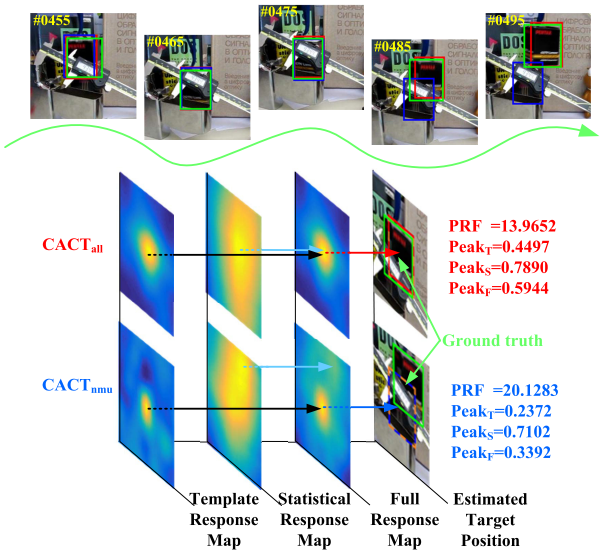

Fig. 6 Comparisons of CACT$_{all}$ and CACT$_{nmu}$ on the 495th frame of the sequence *Box*. Green boxes indicate ground truth, red and blue boxes denote tracking results of CACT$_{all}$ and CACT$_{nmu}$ respectively, orange boxes stand for previous estimated target regions. Best viewed in color.

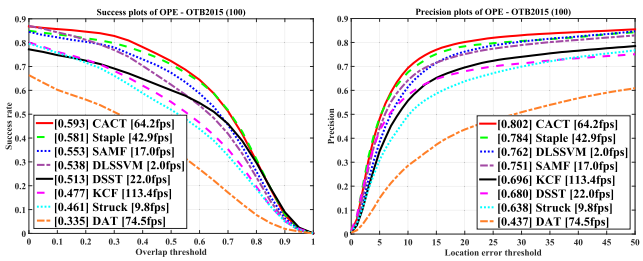

Fig. 7 Success and precision plots of the compared trackers on the OTB-2015 benchmark [1]. The first value in the legend indicates the mean AUC score for each tracker. We also present the speed of trackers in mean FPS as the last number in the bracket. Best viewed in color.

employing scale estimation methods, which provide mean AUC scores of 55.3% and 51.3%, respectively. DAT developed based on sole statistical appearance models achieves a mean AUC score of 33.5%. The dual SSVM based tracker DLSSVM obtains a mean AUC score of 53.8%. Staple which developed based on the benefits of both template and statistical models obtains a mean AUC score of 58.1%. Our approach employs adaptive combination and model update strategies, significantly improves the mean AUC score to 59.3%. And our tracker obtains the mean DP score of 80.2%, outperforming all compared trackers.

To visualize the impact of our proposed approach, we show tracking results with comparison to state-of-the-art trackers on example sequences from the OTB-2015 benchmark in Fig. 8. It should be noted our approach performs favorably against other trackers.

The top performance can be attributed to that our approach makes use of the adaptive combination of complementary models to enhance the feature diversity and the tracking accuracy. What is more, model update strategy en-

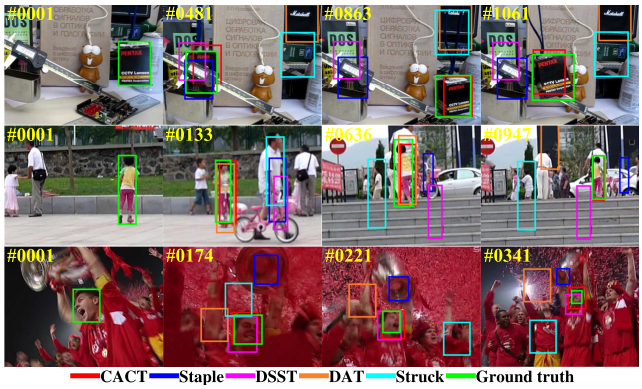

Fig. 8 Tracking results comparison of our proposed approach with three state-of-the-art trackers including DAT [5], DSST [3] and Staple [7] on three example sequences *Box* (top row), *Girl2* (middle row) and *Soccer* (bottom row) from the OTB-2015 benchmark. Best viewed in color.

Table 2 Attribute-based comparison of our approach with 3 state-of-the-art trackers on the OTB-2015 benchmark. The results are presented in mean AUC scores (%). FM, BC, MB, DEF, IV, IPR, LR, OCC, OPR, OV and SV indicate fast motion, background clutter, motion blur, deformation, illumination variation, in-plane rotation, low resolution, occlusion, out-of-plane rotation, out of view and scale variation, respectively. The best results are highlighted in **bold**.

Attributes	CACT	Staple	DSST	DAT
FM	**54.3**	53.7	44.7	31.9
BC	**60.6**	57.4	52.3	28.7
MB	**57.9**	54.6	46.9	29.5
DEF	**56.9**	55.4	42.0	37.5
IV	**59.9**	59.8	55.8	26.1
IPR	54.3	**55.2**	50.2	32.9
LR	**40.7**	39.6	37.0	29.0
OCC	**57.4**	54.8	45.3	34.0
OPR	**56.2**	53.4	47.0	33.9
OV	**57.1**	48.1	38.6	29.2
SV	**54.3**	52.5	46.8	32.3

ables moderately infrequent update of models not only alleviate target drifting but also has a substantial effect on the overall computational complexity of tracking.

For comprehensive evaluation, we also perform an attribute-based analysis of our approach with state-of-the-art trackers including Staple [7], DSST [3] and DAT [5] on the OTB-2015 benchmark. The video sequences contained in the OTB benchmarks are annotated with eleven different attributes that represent various challenging factors as mentioned in the introduction. The results are summarized in Table 2 by presenting in mean AUC scores.

It is clear that our approach (CACT) obtains the best results with a large margin on 10 out of 11 attributes expect in-plane rotation. This is main because that complementary models are trained based on the estimated target position in the previous frame, if the target rotates in the image plane, the models may not capture the rotation similar to the previous foreground, as shown in Fig. 9. In summary, our approach achieves surprisingly good performance on the OTB-2015 benchmark while running at more than 64 FPS.

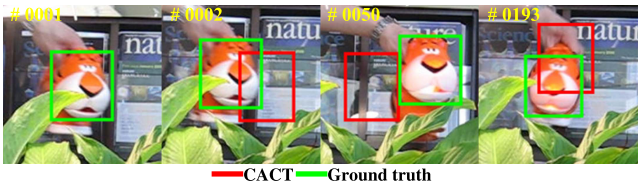

Fig. 9 Failure cases of the proposed approach in the scenario of in-plane rotation on the sequence *tiger1*. Red boxes show our results and green ones are ground truth. Best viewed in color.

4. Conclusion

In this paper, we propose a novel adaptive tracking method with complementary models based on two independent discriminative models, i.e., the template appearance model and the statistical appearance model. Therefore, our approach yields impressive performance to deal with more challenges. In order to efficiently unify these two models, we introduce an adaptive combination strategy to correctly obtain the full response map. Furthermore, we propose an efficient update scheme to improve tracking performance. Experimental results demonstrate our tracker outperforms most state-of-the-art trackers both in terms of accuracy and robustness while running at more than 64 frames per second.

References

[1] Y. Wu, J. Lim, and M.H. Yang, "Object tracking benchmark," IEEE Trans. Pattern Anal. Mach. Intell., vol.37, no.9, pp.1834–1848, 2015.
[2] Y. Li and J. Zhu, "A scale adaptive kernel correlation filter tracker with feature integration," ECCV, pp.254–265, Springer, 2014.
[3] M. Danelljan, G. Häger, F.S. Khan, and M. Felsberg, "Discriminative scale space tracking," IEEE Trans. Pattern Anal. Mach. Intell., vol.39, no.8, pp.1561–1575, 2017.
[4] M. Danelljan, F.S. Khan, M. Felsberg, and J. Van de Weijer, "Adaptive color attributes for real-time visual tracking," CVPR, pp.1090–1097, IEEE, 2014.
[5] H. Possegger, T. Mauthner, and H. Bischof, "In defense of color-based model-free tracking," CVPR, pp.2113–2120, IEEE, 2015.
[6] N. Wang, J. Shi, D.Y. Yeung, and J. Jia, "Understanding and diagnosing visual tracking systems," CVPR, pp.3101–3109, IEEE, 2015.
[7] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P.H.S. Torr, "Staple: Complementary learners for real-time tracking," CVPR, pp.1401–1409, IEEE, 2016.
[8] J.H. Yoon, J. Kim, and Y. Hwang, "Real-time object tracking via fusion of global and local appearance models," IEICE Trans. Inf. & Syst., vol.E100-D, no.11, pp.2738–2743, 2017.
[9] J.F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," IEEE Trans. Pattern Anal. Mach. Intell., vol.37, no.3, pp.583–596, 2015.
[10] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," CVPR, IEEE, 2001.
[11] J. Ning, J. Yang, S. Jiang, L. Zhang, and M.H. Yang, "Object tracking via dual linear structured svm and explicit feature map," CVPR, pp.4266–4274, IEEE, 2016.
[12] S. Hare, S. Golodetz, A. Saffari, V. Vineet, M.M. Cheng, S.L. Hicks, and P.H.S. Torr, "Struck: Structured output tracking with kernels," IEEE Trans. Pattern Anal. Mach. Intell., vol.38, no.10, pp.2096–2109, 2016.