Journal on Artificial Intelligence
DOI:10.32604/jai.2021.010455
Article

Tech Science Press

Hybrid Efficient Convolution Operators for Visual Tracking

Yu Wang*

Department of Computer Science, Harbin Institute of Technology, Weihai, 264200, China
*Corresponding Author: Yu Wang. Email: 17763149856@163.com
Received: 30 August 2020; Accepted: 18 April 2021

Abstract: Visual tracking is a classical computer vision problem with many applications. Efficient convolution operators (ECO) is one of the most outstanding visual tracking algorithms in recent years, it has shown great performance using discriminative correlation filter (DCF) together with HOG, color maps and VGGNet features. Inspired by new deep learning models, this paper propose a hybrid efficient convolution operators integrating fully convolution network (FCN) and residual network (ResNet) for visual tracking, where FCN and ResNet are introduced in our proposed method to segment the objects from backgrounds and extract hierarchical feature maps of objects, respectively. Compared with the traditional VGGNet, our approach has higher accuracy for dealing with the issues of segmentation and image size. The experiments show that our approach would obtain better performance than ECO in terms of precision plot and success rate plot on OTB-2013 and UAV123 datasets.

Keywords: Visual tracking; deep learning; convolutional neural network; hybrid convolution operator

1 Introduction

Visual tracking has been made research in the field of computer vision for decades, so that it has found wide applications [1–2]. These methods [3–6] have made much progress in the past years, but it still has lots of challenges including appearance variations and clutters. To handle these challenges, existing appearance based tracking methods design various feature operators to capture semantic information of targets, and learn discriminative or generative models to distinct co-occurring targets and background.

With the success of deep convolutional neural network (CNN) in image recognition and retrieval, most of recent visual tracking method get rich semantic and strong distinguishing information via residual network (ResNet) instead of relying on hand-crafted features only. Nam et al. [1] proposed a novel discriminatively trained convolutional neural network (CNN) for tracking, Ma et al. [2] exploited to transfer a deep CNN pretrained model from recognition datasets to tracking datasets, and Wang et al. [3] and Danelljan et al. [4–6] claimed a new kind of method spatially regularized discriminative correlation filters (SRDCF). Obviously, deep CNNs naturally enrich the multi-level features and classifiers in an end-to-end framework. All these motivate that we also apply CNNs for the feature operator in this paper to address the challenges faced by tracking.

For the tracking model, correlation filter (CF) tracker is a kind of famous and effective algorithm, which has improved the tracking performance in recent years [7–9]. The CF tracker involves a group of filters to estimate the possible target's positions and select one with maximal response in the next frame. The traditional CF tracker assumes that all targets can be estimated easily by Fourier transform, but it requires the input of continuous image sequences and the operations of multiplying cosine masks for better tracking performance in recent works [9]. Then, based on C-COT, effcient convolution operators (ECO) combines deep CNN and CF to get an excellent tracking model [10–13], which reduces the number of the

model parameters by the factorized convolution and improves tracking efficiency [14,15]. However, ECO can only tackle with the image sequence with fixed-size CNN model so that it has difficulties to handle co-occurring targets.

Driven by the good performance of fully convolutional networks (FCN) in image segmentation [3], in this paper, we propose a framework of hybrid efficient convolution operators based on ECO for visual tracking. Based on combining the FCN feature with ResNet feature, our method jointly extract the hybrid efficient convolution operators by fusion strategy to improve the per-pixel segmentation, which can segment the objects from backgrounds and extract hierarchical feature maps of objects. Different with the existing CF tracker which assumes that the images is periodic, our tracker can greatly improve the efficiency of the Fourier transform process without the continuous of image sequence. The experiments are conducted on OTB-2013 dataset [16] and UAV123 dataset [17] and it demonstrates the better performance of our method than the state-of-the-arts in terms of precision plot and success rate plot.

The rest of the paper is organized as follows. In Section 2, we review the related works about visual tracking, feature operator, CF trackers. In Section 3, we elaborate the details of our proposed method. In Section 4, we describe the experiments and discuss the results. In Section 5, we draw the conclusion.

2 Related Works

2.1 Visual Tracking

Visual object tracking has been achieved impressive results in computer vision. Existing visual trackers are sorted out as two categories based on generative models and discriminative models. Using generative models, Comaniciu et al. [18] employed similarity measurement and mean-shift for optimization. To handle occlusions and distracters, Oron et al. [19] proposed LOT tracker with locally orderless matching, and Zhang et al. [20] proposed another tracker using the spatio-temporal context of targets. Sevilla-Lara [21] proposed DFT with distribution fields to smooth the objective function. On the other hand, deep trackers have also drawn great attention. Using discriminative models, Hare et al. [22] proposed a typical discriminative tracker named Struck that used the structured support vector machine. To achieve more robust tracking, Kalal et al. [23] proposed a tracking learning detection (TLD) tracker with positive-negative learning, and then Zhang et al. [24] proposed MEEM to set a multi-expert restoration system. CF (correlation filter) based methods also take on an important position in discriminative trackers [25–27], such as DSST [28], CSK [13], SINT [29], ECO [4], LCCF [30] and so on. These discriminative model based trackers generally train a classifier and distinguish targets from the background, and gain more high tracking precision at fast speed than generative model based trackers.

2.2 Feature Operator

Recently, most existing works focus on the design of appearance models so that the feature operator can affect the performance of trackers [10,14,31]. Using hand-crafted features, the trackers are learned with discriminative and generative models. Ross et al. [32] learned subspace online to model the appearance for searching candidates with minimized reconstruction errors. CRFs [33], multiple instance learning [33] and structural SVM [34] were also applied in learning online tracker to separate the foreground and background. Using deep learning features, the online tracking is under fully explored. Wang et al. [35] exploited sparse coding and sparse linear combination of target templates for target reconstruction, and trained a stacked denoising autoencoder on a tiny dataset to learn features for online tracking. Li et al. [36] performed tracking as an online target-background classifier with CNN. Hong et al. [37] proposed target-specific saliency maps to guide CNN features for visual tracking.

2.3 CF Tracker

In the field of visual tracking, CF (Correlation filter) is a recent successful tracker algorithm. According to the papers related to CF, it is not an old-fashioned theory, but there are plenty of viarants in CF. Generally, CF means that identifying the target depends on the scale of correlation. Bolme et al. [25]

first put correlation into computer vision and designed correlation filter. Now, there has been many developments based on CF to get new trackers. Valmadre et al. [10] presented a new method to accomplish learning in visual tracking based correlation filter, and adopted fully learning features. Zhang el al. [11] proposed a multi-task correlation particle filter (MCPF) for visual tracking, which exploits the combination of MCF and a particle filter. Besides, it can maintain multiple modes in the posterior density and deal with large-scale variations by the particle sampling. Bibi et al. [12] proposed a general framework to calculate the target response adaptively along the image sequence, which is robust to the circular shifts and translations in a small range.

As for kernel correlation filter (KCF), an updated version of circulant structure with kernels (CSK), Henriques et al. [13,14] provided a new Fourier transform after online learning in fast. Comparing with CSK, Henriques et al. also proposed KCF continued to use Fourier transform but use discrete Fourier transform in CSK and fast Fourier transform in KCF. Besides, there are more channels when handling with images which lead in a better performance in practice.

3 Hybrid Efficient Convolution Operators

In this section, we propose a framework of hybrid efficient convolution operators based on ECO for visual tracking. It combines the FCN feature with ResNet feature and extracts the hybrid efficient convolution operators jointly by fusion strategy. It improves the semantic segmentation ability of the feature operators to segment the objects from backgrounds and extract hierarchical feature maps of objects.

3.1 Overview of CF in Baseline ECO

We first overview the basic existing CF trackers in the baseline ECO, and then analyze its shortcomings of the cosine mask [4,9]. The CF tracker in general gets a multi-channel correlation filter from a group of feature maps that are extracted from original images. However, this only helps to discriminate the target from the background$\{(x^k, y^k)\}_1^t$,where k $\in R^d$ is a d-dimensional feature extracted by CNN from one single image, y_k is the target output with a scalar value for the Gaussian-shaped image, and t is the number of all images. To start I resize all the images to the size of certain CNN input size and extract the feature from a small region. We put feature layer as x_k^l, l=$\{1,2,\dots,d\}$with da as the total layer number. f^l will be set as the correlation filter of the l-th feature layer, and the response of x^k is given by the formulation:

$$R_f(x_k) = \sum_{l=1}^{d} x_k^l * f^l \tag{1}$$

where * denotes the circular convolution. The filter $\boldsymbol{f}$ is learned by minimizing the oprimation problem as this formulation:

$$f = argmin_f \left\{ \sum_{k=1}^{t} a_k \left\| R_f(x_k) - y_k \right\|^2 + \sum_{l=1}^{d} \left\| w \circledcirc f^l \right\|^2 \right\}. \tag{2}$$

In this formulation w is the regularization weight, and $\circledcirc$ means element-wise product. Thus, the original CF tracker helps to separate targets from the background, and we propose a novel and better feature extraction strategy for the CF tracker to improve its representative ability.

3.2 Feature Extraction Using FCN

In ECO framework, the tracker are learned offline. It takes 120 images which show one man cross the street and its job is to track the same man crossing the street inside the 120 images. Following these setting, we use the feature representation as the first layer output of Conv-1 and the last layer output of Conv-5 in the VGG network, color names (CN) and histogram of gradients (HOG). We illustrate new configuration of ECO framework with FCN feature extraction in Fig. 1, in which there are seven implementation processes illustrated.

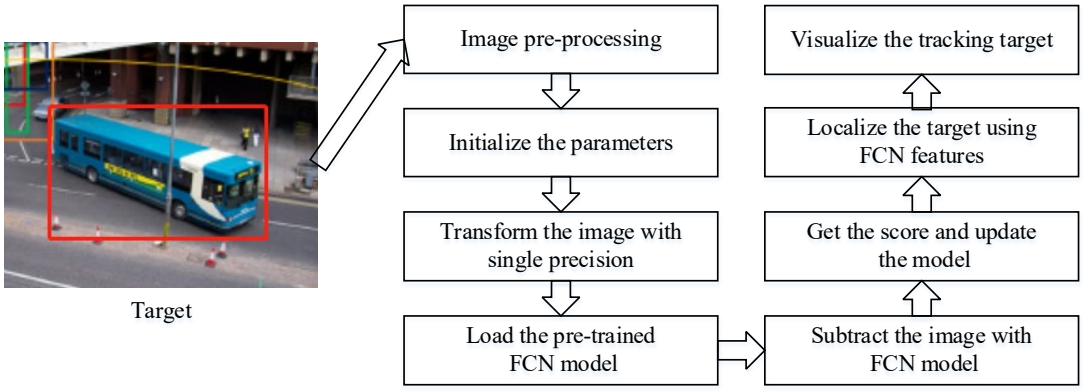

Figure 1: Description of new configuration of ECO framework with FCN feature extraction

In our tracker, some image parameters, feature parameters and other parameters are defined at first. For the image parameters, the image sequence and ground truth are input to evaluate the tracker. For the feature parameters, the FCN is used as the primary feature extraction in Fig. 2. For the other parameters, we follow the setting of ECO to initialize the correct max number of samples, initial scale factor, search area size, window size, size of the extracted feature maps, Gaussian label function using Poisson formula, spatial regularization filter, cosine window, energy of the filter, Fourier series of interpolation function, minimum allowed sample weight, and the set conjugate gradient.

Then, each frame of image sequence is input our tracker for calculation. For simplify the description, we denote the output FCN feature map as $\boldsymbol{X}$, and $\boldsymbol{X_k}$ denotes the FCN feature map at k frame. Following ECO, we initialize the tracker in the first frame, then extract features $\boldsymbol{X_k}$ and $\boldsymbol{\alpha_k}$ are updated by the position model in the next frame. Given the number of samples $\boldsymbol{t}$ and $\boldsymbol{d}$, the Gaussian means $\boldsymbol{\mu_k}$ and the prior weights $\boldsymbol{\pi_k}$, we calculate the detection scores with $\boldsymbol{R_f}\ (\boldsymbol{\mu_k})$, and learn the multi-channel convolution filters using the loss function as:

$$f = arg\,min_f\{\sum_{k=1}^{t}\pi_k \parallel R_f\ (\mu_k) - y_0 \parallel^2 + \sum_{l=1}^{d} \parallel w \odot f^l \parallel^2\}. \quad (3)$$

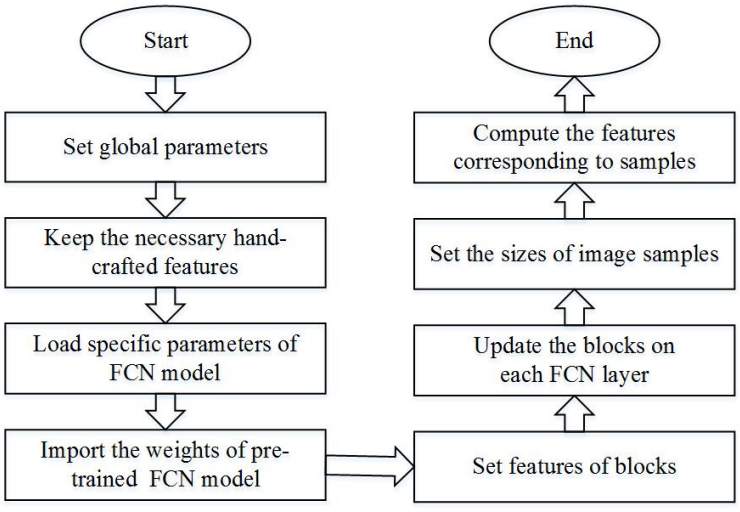

Figure 2: Flowchart of FCN feature extraction with parameters

3.3 Feature Extraction Using ResNet

Deep learning is more theoretical and there are more connection with human science especially brain and neuro science, such as AlexNet, ImageNet and so on [15]. The deeper networks are able to start

converging, and it can get saturated accuracy. The common deep features are extracted with the output of the last FC layer, and some recent feature are calculated from the residual connection layers to get great performance. We use ResNet [18] to produce more discriminative and contain structural features for the improvement of tracking accuracy as

$$Z = FC\left(relu\left(I + res\_I\right)\right) \tag{4}$$

where I denotes the input image patch, $I + res\_I$ denotes the output of identity mapping, $relu(\ )$ denotes the activation function in the residual block, $FC(\ )$ denotes the last FC layer of the residual block in ResNet, Z is the ResNet feature for the image patch I.

In this section, we introduce ResNet to extract features as the heatmap for the current frame. First, we load the pre-trained ResNet model, and resized the images to get single precision maps. Then, we subtract the average maps from the original maps, and input them to the ResNet and get the score maps. Finally, we calculate the heatmap of the current frame by assigning score maps to weights. We analyze the advantages of using ResNet in tracker our as follows. 1) The deep residual nets can optimize than simply stack layers when the depth increases and increase the accuracy to produce better networks; 2) The ResNet always learns residual functions and pass all information with additional residual functions.

3.4 Feature Fusion Strategy

In this section, we explain the fusion strategy for combining FCN features and ResNet features in tracker. Feature fusion is to integrate multiple source of feature maps into an individual feature map with more accurate and discriminative ability, in other words, two feature maps are combined into a feature map in our model. We proposed to apply feature level fusion strategy using a method based on canonical correlation analysis, and get the train and test data matrices from a single feature set.

Given two feature maps $X \in R^{p\times m}$ and $Z \in R^{q\times n}$ of random variables from two different modalities (FCN and ResNet), we define the cross-covariance $n \times m$ matrix with the (i, j) entry as the covariance $\mathrm{cov}\left(x_i, z_j\right)$, then construct the covariance matrix with a pair of sampled data from $X \in R^{p\times m}$ and $Z \in R^{q\times n}$. We define two vectors a and b with initial random values, and search different vectors a and b for calculating $a^T X$ and $b^T Z$ to maximize the correlation $\rho = \mathrm{corr}(a^T X, b^T Z)$.

$$(a', b') = \underset{a,b}{\mathrm{argmax}}\,\mathrm{corr}(a^T X, b^T Z) \tag{5}$$

We continue to update two vectors with the constraint and uncorrelated variables up to $\min\{m,n\}$ ti mes. When obtaining the maximal correlation, we denote the pair of two vectors as a' and b', and the tra nsformed feature maps as $U = a'^T X$ and $V = b'^T Z$. Similar to [34], the feature-level fusion is obtained vi a the concatenation operation of the transformed feature maps:

$$F = (U, V) \tag{6}$$

In this paper, we then use the fused hybrid feature maps F to deploy tracker models following the basic ECO method [4]. The query target is then classified as the highest response value based on the target template and the patches in the background.

4 Experiments

Based on FCN and ResNet, our method is presented to incorporate the hybrid feature operators into a powerful tracker. To verify the effectiveness of the proposed method, we conduct experiments on two bechmarks, OTB-2013 benchmark dataset [16] and UAV123 dataset [17].

4.1 Benchmark Datasets

The OTB-2013 dataset [16] is a challenging tracking dataset with 50 sequences. In the ground-truch files, multiple targets in each frame of a sequence are marked as various numbers, and each target is recorded with the bounding box in a row. Compared with OTB-2013, UAV123 is a newly published benchmark with 123 sequences on unmanned aerial vehicles (UAVs) [17]. Due to the specific characteristics of UAV photography, the video clips in UAV123 benchmark include extremely small target, rapid view point change, drastic scale variation, and long-term out-of-view. We choose UAV123 at the frame rate of 10fps to evaluate in our experiment.

4.2 Implementation Details

We compare our method with other state-of-the-art trackers on two above-mentioned datasets. Our tracking method is implemented in Matlab 2017b based on Matconvnet 1.0 framework, and runs at one NVIDIA GeForce GTX 1080ti GPU, in Windows 10 operating system. The FCN and ResNet are pretrained on ImageNet dataset, and finetuned in the first frame of each sequence after 10 iterations of updating via back-propagation. For experimental setting, except for the fixed parameters in ECO tracker, the learning rate of ResNet is 1e-7, and that for FCN is 1e-9, the weight decay is 0.005, the size of the ROI near target is 386 × 386 pixels, and the number of feature maps for fusion is 384.

4.3 Results and Analysis

For the OTB-2013 dataset, we evaluate the performance of our tracker following the protocol in [4]. As shown in Tab. 1, we analyze the experimental results of our and other trackers in two terms, precision and area under curve of success plots. It is noted that our tracker outperforms other Siamese trackers with the success rate of 0.65 and the precision of 0.88, such as the SRDCF [6], SINT [29], CFNet [10], ECO-HC [4], ECO-Deep [4], LCCF [30], which demonstrates the superior performance of our tracker.

Table 1: The comparisons of distance precision and success rate on OTB-2013 dataset sequences using one-pass evaluation

Method	OTB-2013	
	Success rate	Distance precision
SRDCF [6]	0.63	0.84
SINT [29]	0.64	0.85
CFNet [10]	0.61	0.80
ECO-HC [4]	0.64	0.86
ECO-Deep [4]	0.65	0.87
LCCF [30]	0.57	0.85
Our tracker	**0.65**	**0.88**

For the UAV-123 dataset, our tracker results are illustrated in Tab. 2. As known from the published works, UAV123 is more difficult than OTB-2013 in the aspects of tracking challenges, but our tracker still achieves better evaluation indicators than six state-of-the-art trackers, including DSST [28], MEEM [24], CFNet [10], ECO-HC [4], ECO-Deep [4], LCCF [30]. It is demonstrated that our tracker improves the success rate and precision to 0.54 and 0.76, which confirms that our tracker is more effective because of using the hybrid efficient convolution operators.

Table 2: The comparisons of distance precision and success rate on UAV-123 dataset sequences using one-pass evaluation

Method	UAV-123	
	Success rate	Distance precision
DSST[28]	0.36	0.58
MEEM [24]	0.39	0.62
CFNet [10]	0.47	0.68
ECO-HC [4]	0.49	0.70
ECO-Deep [4]	0.53	0.75
LCCF [30]	0.38	0.61
Our tracker	**0.54**	**0.76**

More specifically, in Fig. 3, we choose four best compared methods on OTB-2013 dataset sequences, and draw the distance precision plot and success plot of OPE. In Fig. 3, our tracker are compared with four recent methods, SRDCF [6], SINT [29], ECO-Deep [4], and LCCF [30], and our tracker gain the best performance under the same experimental settings and its efficiency will be improve for real applications.

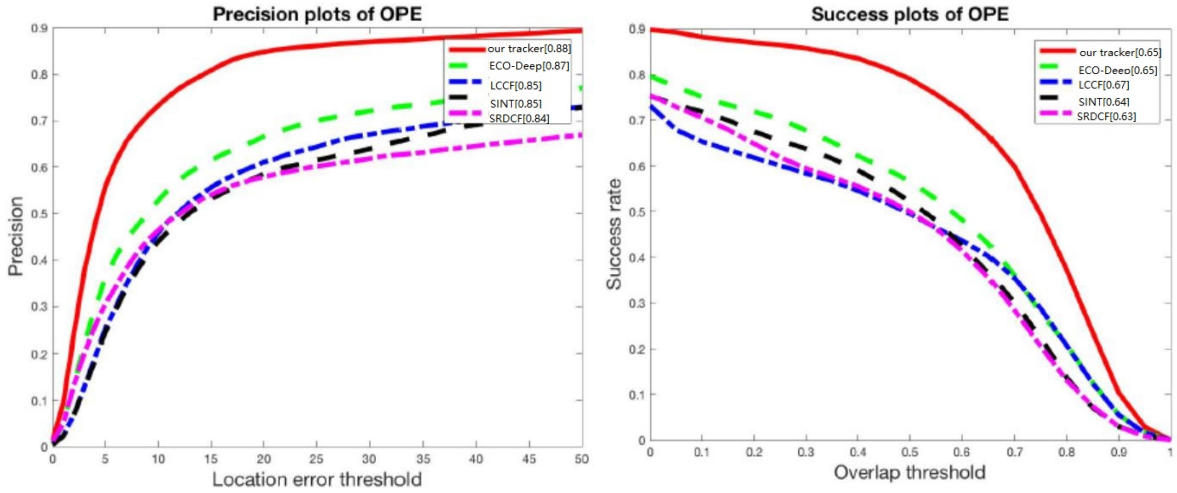

Figure 3: Visualization of the plots of distance precision and success rate using our tracker and four compared trackers on the OTB-2013 dataset

In ECO framework, the trackers are learned offline. It takes 120 images which show one man cross the street and its job is to track the same man crossing the street inside the 120 images. The original experiment uses the same feature as C-COT, histogram of gradients (HOG) and color names (CN). We illustrate new configuration of ECO framework with FCN feature extraction in Fig. 1, in which there are seven implementation processes illustrated.

To further evaluate our tracker, we visualize the feature maps of some frames in a sequence. In Fig. 4, we select five frames to illustrate the good ability of our method in image segmentation compared with ECO. Obviously, our method gets high score pixels in a much closer region of target, which relies on the

excellent semantic segmentation ability of our proposed hybrid convolution operators to learn the local and global correlation information. Indeed, the hybrid operators can label each pixel using dense predictions to figure out the target and the background. In Fig. 4, it is shown that compared with the second row of ECO results, our tracker with hybrid operators gets more compact and high scores in the feature map when occurring the partial occlusion in the sequence. It is also noted that the feature map obtained by our tracker is so clear to represent the correlation among pixels from a same target and update the contour of target by gradually learning.

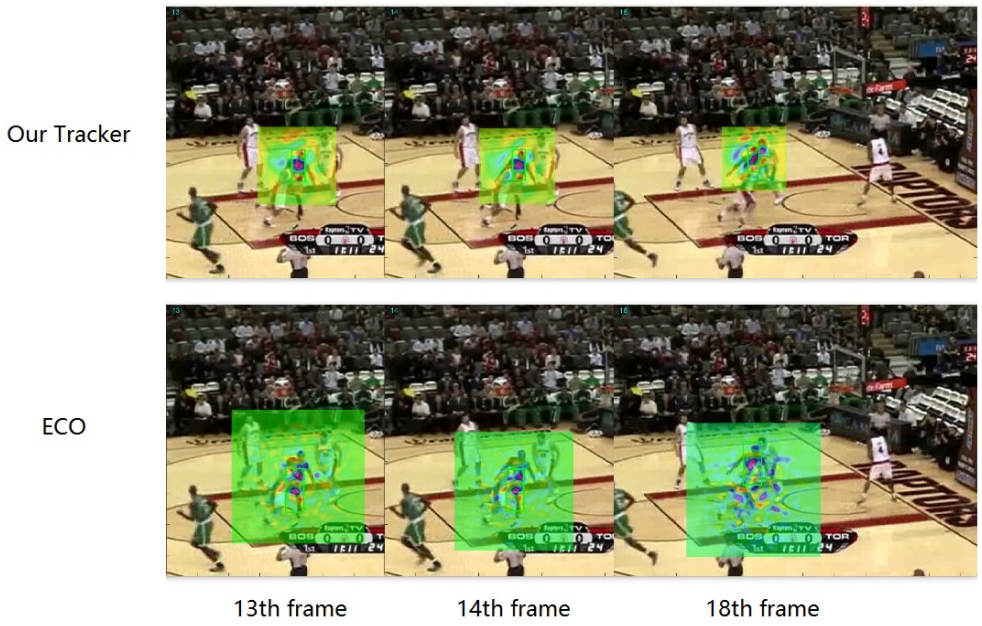

Figure 4: Visualization of the training set with feature representation of ECO and our tracker on a sequence from OTB-2013

5 Conclusion

Efficient convolution operators (ECO) is one of the most excellent and recent tracking methods due to its great performance using discriminative correlation filter, however, lots of challenges still exist in the learning process as appearance changes, occlusions, variations, and clutters. In order to enhance the target segmentation ability of ECO, this paper propose a hybrid efficient convolution operators integrating fully convolution network (FCN) and residual network (ResNet) for visual tracking, where FCN and ResNet are introduced in our proposed method to segment the objects from backgrounds and extract hierarchical feature maps of objects, respectively. Compared with the traditional VGGNet, our approach has higher accuracy for dealing with the issues of segmentation and image size. The experiments show that our approach gain better performance than ECO on OTB-2013 and UAV123 datasets. In the future work, the deep learning based semantic representation will be studied and involved for a robust tracker, and more complex datasets will be used for comparison in visual tracking task.

Acknowledgement: The author thank the authors of two benchmark datasets.

Funding Statement: The author received no specific funding for this study.

Conflicts of Interest: The author declare that they have no conflicts of interest to report regarding the present study.

References

[1] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proc. 2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 4293–4302, 2016.

[2] C. Ma, J. B. Huang, X. Yang and M. H. Yang, "Hierarchical convolutional features for visual tracking," in *Proc. 2015 IEEE Int. Conf. on Computer Vision (ICCV)*, Santiago, Chile, pp. 3074–3082, 2015.

[3] L. Wang, W. Ouyang, X. Wang and H. Lu, "Visual tracking with fully convolutional networks," in *Proc. 2015 IEEE Int. Conf. on Computer Vision (ICCV)*, Santiago, Chile, pp. 3119–3127, 2015.

[4] M. Danelljan, G. Bhat, F. S. Khan and M. Felsberg, "ECO: efficient convolution operators for tracking," in *Proc. 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 21–26, 2017.

[5] M. Danelljan, G. Hager, F. Shahbaz Khan and M. Felsberg, "Convolutional features for correlation filter based visual tracking," in *Proc. IEEE Int. Conf. on Computer Vision Workshops*, Santiago, Chile, pp. 58–66, 2015.

[6] M. Danelljan, G. Hager, F. Shahbaz Khan and M. Felsberg, "Learning spatially regularized correlation filters for visual tracking," in *Proc. IEEE Int. Conf. on Computer Vision*, Santiago, Chile, pp. 4310–4318, 2015.

[7] G. R. Bradski, "Computer vision face tracking for use in a perceptual user interface," in *Proc. of the Fourth IEEE Workshop on Applications of Computer Vision (WACV'98), IEEE*, Princeton, New Jersey, USA, 1998.

[8] S. Laine, T. Karras, T. Aila, A. Herva, S. Saito *et al.,* "Production-level facial performance capture using deep convolutional neural networks," in *Proc. ACM SIGGRAPH/Eurographics Sym. on Computer Animation*, Los Angeles, California, pp. 10, 2017.

[9] D. S. Bolme, J. R. Beveridge, B. A. Draper and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *Proc. 2010 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, San Francisco, California, USA, pp. 2544–2550, 2010.

[10] J. Valmadre, L. Bertinetto, J. F. Henriques, A. Vedaldi and P. H. S. Torr, "End-to-end representation learning for correlation filter based tracking," in *Proc. 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 5000–5008, 2017.

[11] T. Zhang, C. Xu and M. H. Yang, "Multi-task correlation particle filter for robust object tracking", in *Proc. 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 4819–4827, 2017.

[12] A. A. Bibi, M. Mueller and B. Ghanem, "Target response adaptation for correlation filter tracking," in *Proc. European Conf. on Computer Vision*, Amsterdam, The Netherlands, pp. 419–433, 2016.

[13] J. F. Henriques, R. Caseiro, P. Martins and J. P. Batista, "Exploiting the circulant structure of tracking-by-detection with kernels," in *Proc. 12th European Conf. on Computer Vision*, Florence, Italy, pp. 702–715, 2012.

[14] J. F. Henriques, R. Caseiro, P. Martins and J. P. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 583–596, 2015.

[15] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Advances in Neural Information Processing Systems*, Harrahs and Harveys, Lake Tahoe, pp. 1097–1105, 2012.

[16] Y. Wu, J. Lim and M. H. Yang, "Online object tracking: a benchmark," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Portland, USA, pp. 2411–2418, 2013.

[17] M. Mueller, N. Smith and B. Ghanem, "A benchmark and simulator for UAV tracking," in *Proc. European Conf. on Computer Vision*, Amsterdam, The Netherlands, pp. 445–461, 2016.

[18] D. Comaniciu, V. Ramesh and P. Meer, "Kernel-based object tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 5, pp. 564–577, 2003.

[19] S. Oron, A. Barhillel, D. Levi and S. Avidan, "Locally orderless tracking," in *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, Providence, RI, USA, pp. 1940–1947, 2012.

[20] K. Zhang, L. Zhang, Q. Liu, D. Zhang and M. Yang, "Fast visual tracking via dense spatio-temporal context learning," in *Proc. European Conf. on Computer Vision*, Zurich, Switzerland, pp. 127–141, 2014.

[21] L. Sevillalara and E. Learnedmiller, "Distribution fields for tracking," in *Proc. Computer Vision and Pattern Recognition (CVPR)*, Providence, RI, USA, pp. 1910–1917, 2012.

[22] S. Hare, A. Saffari and P. H. S. Torr, "Struck: structured output tracking with kernels," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 10, pp. 2096–2109, 2016.

[23] Z. Kalal, K. Mikolajczyk and J. Matas, "Tracking-learning-detection," *IEEE Transactions on Pattern Analysis*

and Machine Intelligence, vol. 34, no. 7, pp. 1409–1422, 2012.

[24] J. Zhang, S. Ma and S. Sclaroff, "MEEM: robust tracking via multiple experts using entropy minimization," in *Proc. European Conf. on Computer Vision*, Zurich, Switzerland, pp. 188–203, 2014.

[25] D. S. Bolme, J. R. Beveridge, B. A. Draper and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *Proc. Computer Vision and Pattern Recognition (CVPR)*, San Francisco, California, USA, pp. 2544–2550, 2010.

[26] T. Z. Zhang, S. Liu, C. S. Xu, B. Liu and M. H. Yang, "Correlation particle filter for visual tracking," *IEEE Trans Image Process*, vol. 27, no. 99, pp. 2676–2687, 2018.

[27] T. Zhang, C. Xu and M. H. Yang, "Learning multi-task correlation particle filters for visual tracking," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 41, no. 2, pp. 365–378, 2018.

[28] M. Danelljan, G. Häger, F. S. Khan and M. Felsberg, "Discriminative scale space tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 8, pp. 1561–1575, 2017.

[29] R. Tao, E. Gavves and A. W. Smeulders, "Siamese instance search for tracking," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp. 1420–1429, 2016.

[30] B. Zhang, S. Luan, C. Chen, J. Han, W. Wang *et al.,* "Latent constrained correlation filter," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1038–1048, 2018.

[31] Q. Sun, S. Zeng, Y. Liu, P. Heng and D. Xia, "A new method of feature fusion and its application in image recognition," *Pattern Recognition*, vol. 38, no. 12, pp. 2437–2448, 2005.

[32] D. A. Ross, J. Lim, R. S. Lin and M. H. Yang, "Incremental learning for robust visual tracking," *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 125–141, 2008.

[33] X. Ren and J. Malik, "Tracking as repeated figure/ground segmentation," in *Proc. Computer Vision and Pattern Recognition*, Minneapolis, MN, USA, pp. 1–8, 2007.

[34] S. Hare, A. Saffari and P. H. Torr, "Struck: structured output tracking with kernels," in *IEEE Int. Conf. on Computer Vision*, Barcelona, Spain, vol. 1, pp. 263–270, 2011.

[35] N. Wang and D. Yeung, "Learning a deep compact image representation for visual tracking," in the *Conf. and Workshop on Neural Information Processing Systems*, Harrahs and Harveys, Lake Tahoe, pp. 809-817, 2013.

[36] H. Li, Y. Li and F. M. Porikli, "Robust online visual tracking with a single convolutional neural network," in *12th Asian Conf. on Computer Vision*, Singapore, pp 194–209, 2014.

[37] S. Hong, T. You, S. Kwak and B. Han, "Online tracking by learning discriminative saliency map with convolutional neural network," in *Int. Conf. on Machine Learning*, Lille, France, pp. 597–606, 2015.