Machine Vision and Applications (2018) 29:955–963
https://doi.org/10.1007/s00138-018-0947-6

SPECIAL ISSUE PAPER

Hierarchical convolutional features for end-to-end representation-based visual tracking

Suguo Zhu[1] · Zhenying Fang[1] · Fei Gao[1,2]

Received: 31 October 2017 / Revised: 2 February 2018 / Accepted: 11 March 2018 / Published online: 14 June 2018

Abstract
Recently, deep learning is widely developed in computer vision applications. In this paper, a novel simple tracker with deep learning is proposed to complete the tracking task. A simple fully convolutional Siamese network is applied to capture the similarity between different frames. Nevertheless, the detailed information from lower layers, which is also important for locating the target object, is not considered into the tracking task. In this paper, the detailed information from two lower layers is considered into the response map to improve the performance and not to increase much time spent. This leads more significant improvement for feature representation and localization of the target object. The experimental results demonstrate that the proposed algorithm is efficient and robust compared with the baseline and the state-of-the-art trackers.

Keywords Tracking · Siamese network · Deep neural network · Deep learning · End-to-end · Hierarchical convolutional features

1 Introduction

Deep neural network is a popular approach in recent applications such as visual tracking [1], human pose recovery [2], web image reranking [3] and image privacy protection [4]. For visual tracking, the challenge is the lack of a priori knowledge of the target. However, the neural networks obtain limited relatively information by pre-trained deep convolutional neural network through offline training before tracking. The advantages are that the network is not only able to extract much more exact features, but also end-to-end, which learns features from row data directly and discards the complexity form stacking kinds of models of the traditional approaches.

✉ Suguo Zhu
zsg2016@hdu.edu.cn

Zhenying Fang
fzy19931001@outlook.com

Fei Gao
gaofei@hdu.edu.cn

[1] Key Laboratory of Complex Systems Modeling and Simulation, School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China

[2] State Key Laboratory of Integrated Services Networks, Xidian University, Xian 710071, China

However, it is still at the exploration stage for visual tracking with deep neural network. Most researchers always pay their attentions on how to combine the traditional feature (such as CFT [5]) extracting methods with deep neural network, and improve the performance gradually, for example, ECO method [6], in which the feature extracting method is combined with HOG, CNN and CN feature. The end-to-end trackers are rare. The existing typical end-to-end tracking approaches tend to focus on the use of multiple neural networks to assist network each other, for example, multi-stream CNN for tracking [7] and Siamese network for tracking [8,9]. The multi-stream network tracker, proposed by Li K et al., is a tracking-by-verification method with multi-stream deep neural networks for computing the similarity between different instances of two consecutive frames. The primary difference between multi-stream CNN and Siamese network is that the Siamese network shares the parameters between different networks, which greatly reduce the number of parameters and simplify the network structure, for example, CFNet tracker [9]. In this paper, we propose an end-to-end approach for visual tracking, and the Siamese network is as the tracking framework for less parameters and simple network structure. Similar to siameseFC tracker [8], each branch of the framework is fully convolutional neural network.

The objective of visual tracking is to predict the location and size of the target object, while the deeper the layer

of the network is, the closer the features are to semantics information and the more invariant the features are. It is not the optimum feature if only the features of the last layer are used for visual tracking. For deeper information of the network, in this paper, we explore to adaptively learn the correlation between features from the corresponding layer of two branches in Siamese network. In this case, not only features of the last layer are used, but also that from hierarchical layers of the network. And the state-of-the-art methods have certified that it is helpful to fusion features of different convolutional layers for visual object tracking.

The contributions of this paper include as follows: (1) The detail information of the target is considered, as part of the final score; (2) the combination with detail and semantic information of the target for robust tracking. The remainder of this paper is organized as follows. Section 2 describes the related works. Section 3 reports the details of proposed approach for tracking, including the Siamese baseline for object tracking and the HSFCNet for object tracking. Next, Sect. 4 reports the implementation details and the experimental results and analysis compared with the baseline and other tracking methods. At the end of the paper, we conclude the proposed algorithm and present prospect in the future.

2 Related works

The traditional tracking method The traditional tracking methods in general apply composite features to overcome the challenging problem. Zhong et al. [10] proposed an appearance model that exploits both holistic templates and local representations. Struck tracker [11] introduced a kernelized structured SVM to visual tracking and a budgeting mechanism for real-time application. For long-term tracking, Kalal et al. [12] proposed the TLD tracker, which integrated the NCC matching for recovery with a differential tracker and a updating model that is complex. To address the problem of blurring the image destroys image information, Sevilla-Lara and Learned-Miller [13] proposed to build an image descriptor using distribution fields. Inspired by the training process using discriminative classification, Henriques et al. [14] provided a link to Fourier analysis using the well-established theory of circulant matrices. This method and its deformation are also widely used nowadays. There were lots of methods which combined the hand-draft feature extracting methods and sparse learning method for visual tracking, for example, ASLA tracker [15] and MTT tracker [16]. To settle the rigid and deformable objects online and with no prior assumptions in visual tracking, Oron et al. [17] provided a probabilistic model of the object variations over time.

Siamese neural network for tracking Tracking with Siamese network is becoming a promising approach for its simpleness and efficiency. The Siamese network is a Y-shaped neural network [18,19], which joins two same network branches and produces a single output in the final layers. YCNN [20] firstly proposed two identical branches, which linked with three conv and max-pooling layers and three fc layers. It is similar to VGGNet [21] and learns discriminating features of growing complexity and the similarity between two branches with the exemplar and the search region. SINT [22] inherited the AlexNet [23] and VGGNet. It proposes the two branches that are not connected but share the same weights. It learns solely features in query and an additional binary variable indicating correct and incorrect pairs measured by the Jaccard index. CFNet [9] is a modified version of SiamFC [8]. CFNet inherits the basic ability from SiamFC to discriminate features and solve the underlying ridge regression of the correlation layer. SiamFC inherited from AlexNet with five conv layers, max-pooling with the first two layers and ReLUs after every con layer except for conv5. It is simple and fast for tracking task. The similarity function is seen as cross-correlation operator. GOTURN [24] proposes tow convolutional branches shared the same parameters. It inherited from AlexNet and learns simultaneously the hierarchy of spatial features.

3 The proposed approach

In this section, we will introduce the framework of Siamese baseline and the use of it for object tracking before presenting the HSFCNet architecture, and then the proposed HSFCNet approach for visual tracking.

3.1 The Siamese baseline for object tracking

The original fully convolutional Siamese network considers pairs (x', y') comprising a training image x' and a test image y'. A fully convolutional network with learnable parameters is processed as the embedding function f_ρ and yields two feature maps, which are then cross-correlated:

$$h_\rho\left(x', y'\right) = f_\rho\left(x'\right) * f_\rho\left(y'\right), \tag{1}$$

where the equation amounts to performing an exhaustive search of the pattern x' over the test image y'. $h_\rho\left(x', y'\right)$ denotes the response map, which measures the similarity of two image patches.

To achieve object tracking, the maximum value of the response map is needed, which is corresponding to the target location. The network is trained offline with millions of random pairs (x', y'), which are taken from videos. The training proceeds by minimizing an element-wise logistic loss l over

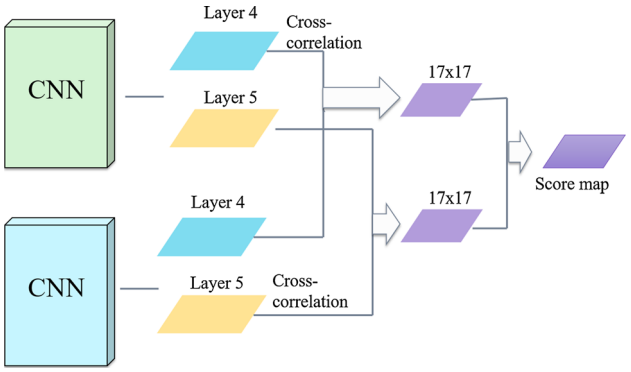

Fig. 1 The architecture of HSFCNet algorithm

the training set:

$$\arg \min_{\rho} \sum_{i} l\left(h_{\rho}\left(x_{\rho}{}', y_{\rho}{}'\right), c_i\right), \tag{2}$$

where c_i is the spatial map of labels with values in $\{-1, 1\}$, and the positive class belongs to the true object location and the negative class belongs to others.

To obtain the tracking results, Bertinetto et al. [8] combine the maximum value of the response map with a procedure to assess the utility of the similarity function for object tracking. In this paper, we take not only the final layer for object tracking, but also the third and the fourth layers, which will helpful to capture the detailed information of the object.

During tracking, it is performed by evaluating the forward-mode network. The image x' represents the object of an image patch centered on the initial frame, while y' is the feature representation of the target object, which is obtained in each new frame by extracting a window centered at the previously estimated position. The new position of the object is taken to be the location with the highest score. In this paper, we combine three response maps from three different layers to predict the tracking result. The original fully convolutional Siamese network simply compared every frame to the initial appearance of the object.

3.2 The HSFCNet for object tacking

In this paper, we modify the fully convolutional Siamese network. The proposed approach fuses the feature form the fourth layer with the fifth layer, and it can be formalized as:

$$g_{\rho}\left(x', y'\right) = \sum_{j,k} \lambda_j f_{\rho}\left(x_k{}'\right) * f_j\left(y_k{}'\right), \tag{3}$$

where $j = 1, k = 4$, $j = 2, k = 5$ and $*$ denotes an operation of cross-correlation. When $j = 1, k = 4$, the feature map of the fourth layer is taken and the value of the weight parameter j is 1, and when $j = 2$, $k = 5$, the feature map of the fifth layer is taken and the value of the weight parameter j is 2. Then, the equation is rewritten as:

$$\begin{aligned} g_{\rho}\left(x', y'\right) = {} & \lambda_1 f_{\rho}\left(x_4{}'\right) * f_{\rho}\left(y_4{}'\right) \\ & + \lambda_2 f_{\rho}\left(x_5{}'\right) * f_{\rho}\left(y_5{}'\right). \end{aligned} \tag{4}$$

Note that we make a cross-correlation between the feature maps of the fourth layer in the fully convolutional network.

$$\left(f_{\rho}^4 * f_{\rho}^5\right)(i, j) = \sum_{m} \sum_{n} f_{\rho}^4 (i + m, j + n)\, f_{\rho}^5 (m, n), \tag{5}$$

where f_{ρ}^4 and f_{ρ}^5 denote $f_{\rho}\left(x_4{}'\right)$ and $f_{\rho}\left(x_5{}'\right)$, respectively. Taking the feature map as an image, f_{ρ}^4 and f_{ρ}^5 are replaced as I and K, and the size of the image is $i \times j$, $m = 1, \ldots, i$, $N = 1, \ldots, j$. Equation (5) is rewritten as:

$$(I * K)(i, j) = \sum_{m} \sum_{n} I (i + m, j + n)\, K (m, n). \tag{6}$$

Actually, cross-correlation is as the same as convolution except flipping. Generally, the implementation of cross-correlation is as the same as Eq. (5).

There are some similar approaches to fuse the feature maps of the lower layers, for example, Ma et al. [25] combined the feature maps from the third, fourth and fifth layers; nevertheless, the difference is that it applied the correlation filter for every layer but not the cross-correlation operator. Offline training is then performed as the fully convolutional Siamese network do itself, and we fine turn the network with the first frame of every video sequence at the very beginning of the tracking.

We found that it is important to consider the feature maps of the lower layer for object tracking, which is able to provide the detailed information, such as color and texture, and the final layer is always deficient with this. The early approaches, for example, [25,26], also applied the feature maps of the lower layer for better tracking results. The reason of better tracking results with feature maps of lower layer is that the goal of object tracking is to locate the object, and this will need not only the semantic information about the target object, but also the detailed information. The architecture is shown in Fig. 1.

4 Experiments

4.1 Evaluation criteria

In order to avoid overfitting to the test set in design choices, we apply the pre-trained network from SiameseCF_3c [8] and fine turn it after modification with the proposed method before tracking.

Table 1 List of the attributes annotated to the sequences

Attr	Description
IV	Illumination variation: the illumination in the target region is significantly changed
SV	Scale variation: the ratio of the bounding boxes of the first frame and the current frame is out of the range$[1/t_s, t_s]$, $t_s > 1$ ($t_s = 2$)
OCC	Occlusion: the target is partially or fully occluded
DEF	Deformation: non-rigid object deformation
MB	Motion blur: the target region is blurred due to the motion of target or camera
FM	Fast motion: the motion of the ground truth is larger than t_m pixels ($t_m = 20$)
IPR	In-plane rotation: the target rotates out of the image plane
OPR	Out-of-plane rotation: the target rotates out of the image plane
OV	Out of view: some portion of the target leaves the view
BC	Background clutters: the background near the target has the similar color or texture as the target
LR	Low resolution: the number of pixels inside the ground-truth bounding box is less than t_τ ($t_\tau = 400$).

Table 2 Average success scores on different attributes (best viewed on a emphasis display)

Attribute	OS	LR	BC	OV	SV	IPR	OPR	IV	DEF	FM	MB	OCC
OURS	**0.73**	**0.66**	*0.65*	*0.62*	**0.80**	**0.65**	**0.70**	**0.60**	*0.69*	**0.68**	***0.59***	**0.76**
SiamFC_3s	*0.72*	*0.65*	***0.60***	**0.71**	*0.72*	***0.59***	*0.68*	***0.59***	***0.65***	*0.66*	*0.60*	*0.74*
Staple	***0.7***	***0.54***	**0.69**	***0.52***	***0.63***	**0.65**	***0.66***	**0.60**	**0.76**	***0.61***	**0.63**	***0.72***
SCM	0.56	0.31	0.51	0.18	0.58	0.42	0.49	0.48	0.51	0.29	0.28	0.58
Struck	0.52	0.41	0.47	0.44	0.49	0.47	0.44	0.42	0.43	0.54	0.54	0.48
TLD	0.44	0.33	0.34	0.35	0.46	0.34	0.43	0.32	0.36	0.41	0.37	0.42
DFT	0.41	0.20	0.43	0.39	0.36	0.33	0.38	0.39	0.51	0.33	0.39	0.42
CSK	0.41	0.40	0.43	0.24	0.32	0.38	0.36	0.34	0.27	0.35	0.35	0.36
ASLA	0.41	0.16	0.39	0.17	0.46	0.35	0.37	0.36	0.33	0.19	0.17	0.39
MTT	0.40	0.51	0.33	0.22	0.39	0.36	0.32	0.28	0.25	0.37	0.29	0.40
LOT	0.40	0.15	0.41	0.50	0.40	0.30	0.40	0.33	0.42	0.33	0.29	0.45
L1APG	0.40	0.46	0.33	0.18	0.38	0.33	0.32	0.22	0.29	0.33	0.33	0.38
CXT	0.38	0.34	0.29	0.28	0.37	0.37	0.35	0.30	0.25	0.38	0.29	0.37
ORIA	0.30	0.16	0.21	0.12	0.30	0.31	0.30	0.26	0.15	0.18	0.13	0.33

The bold emphasis indicates the best performance, the italic emphasis indicates the second best ones, and the bolditalic emphasis indicates the third best ones
OS overall score of all the attributes)

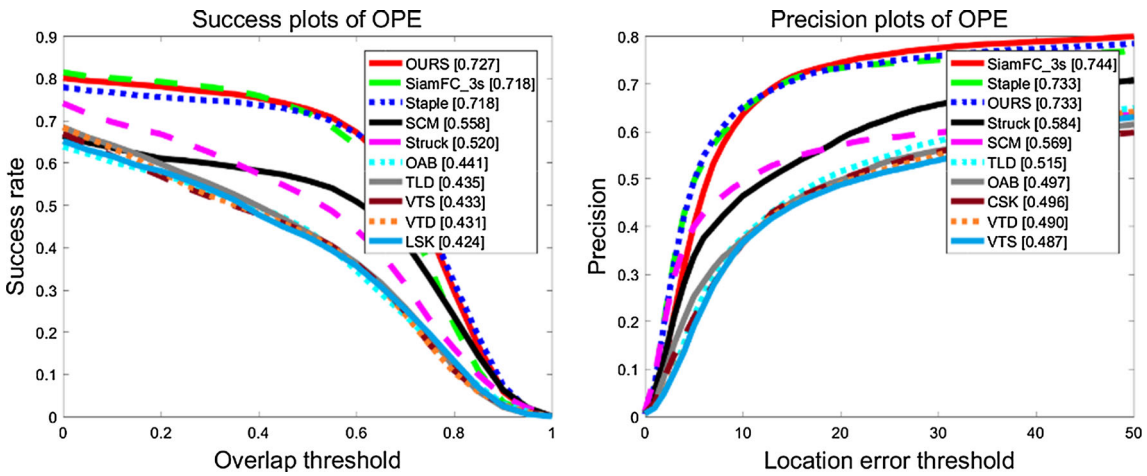

Fig. 2 The success plots and the precision plots of OPE for the top 10 trackers

Fig. 3 Qualitative comparison of our approach with state-of-the-art trackers on the trellis, soccer, singer1, shaking, matrix, ironman, deer, couple, coke and jogging_2 sequences

4.2 Implementation details

For the initial object appearance, the embedding function $f_\rho\left(x'\right)$ is computed once, which is as the same as [8]. The exemplar online during the tracking is not updated for fast tracking. After cross-correlation, the score map from the four layer is 17×17, and it can be added with the weight λ_1 to the score map form the fifth layer with the weight λ_2. This will result in more accurate localization since more detailed information from the fourth layer. The value of the parameter λ_1 and λ_2 is 0.3 and 0.7, respectively.

The compared trackers are on several popular tracking datasets in terms of both overall performance and robustness, such as siamfc_3c [8], staple [27], SCM [10], Struck [11], TLD [12], DFT [13], CSK [14], ASLA [15], MTT [16], LOT [17], L1APG [28], CXT [29] and ORIA [30]. The proposed method is evaluated on the object tracking benchmark dataset [1], containing 100 videos with comparisons to state-of-the-art methods. For completeness, we also report the results on the benchmark dataset with 50 videos. We implement our tracker in PYTHON on an Intel(R) E5-2620 2.10 GHz CPU with 20 GB RAM, where the computation of forward propagation on CNNs is transferred to a TeslaK40 GPU.

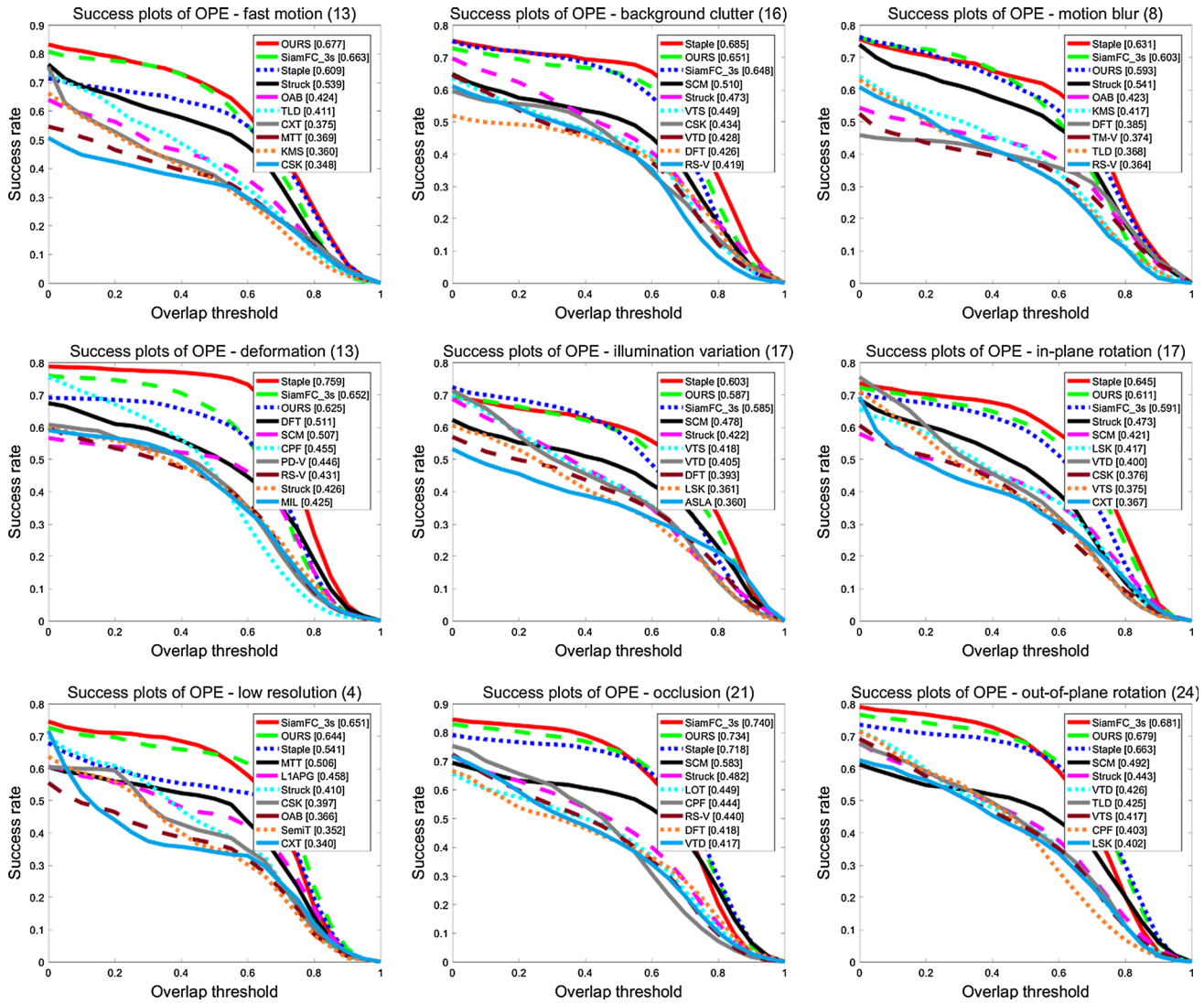

Fig. 4 Success plots of OPE on the overlap threshold

4.3 The OTB benchmark

The OTB-13 [1] benchmark considers the average per-frame success rate at different thresholds: A tracker is successful in a given frame if the intersection-over-union (IoU) between its estimate. The ground truth is above a certain threshold. Trackers are then compared in terms of area under the curve of success rates for different values of this threshold. In addition to the trackers reported by [1], in Table 1, we also compare against Staple [27] and SiameFC [8]. All the other hyper-parameters (for training and tracking) are fixed.

For better evaluation, OTB-13 classified the dataset into 12 classes through the different situation of every sequence. Here, we describe the attributes annotated to the sequences in Table 1.

4.4 Experimental results and analysis

It is clearly to see that the proposed tracker has much smaller average central errors for most of the sequences than the comparative trackers, as given in Table 2. As illustrated in Table 1, the attributes are: low resolution (LR), background cluttered (BC), out of view (OV), scale variation (SV), in-plane rotation (IPR), out-of-plane rotation (OPR), illumination variation (IV), deformation (DEF), fast motion (FM), motion blur (MB) and occlusion (OCC). Note that the score of success plot of OPE describes the results of different approaches on different situations. Our algorithm is much better than others on the overall score, especially than the SiamFC algorithm. For other situations, the proposed method performs well, especially in the situation with background cluster, in-plane rotation and scale variation.

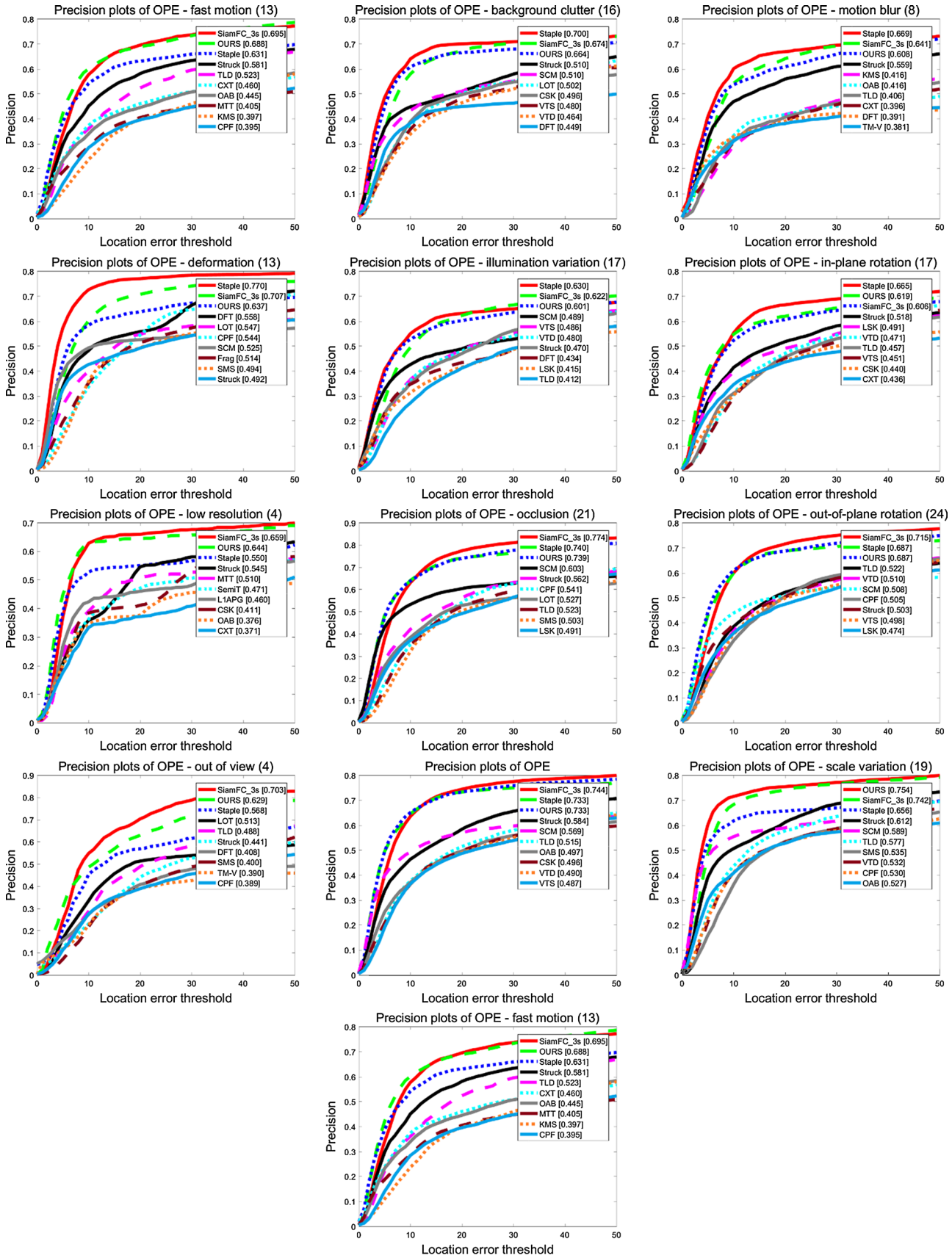

Fig. 5 Precision plots of OPE on location error threshold

Actually, for the tracking task, the algorithm not only needs the semantic information from the high layer, but also needs the details from the lower layer. This is because the detail change disturbs the tracker, for example, the illumination changes, the scale variation, the background cluster and the occlusion. Table 2 shows the success scores of OPE. The success score means the ratio of overlap rate and success rate. Note that the proposed algorithm is better than others, especially for the LR, SV, IPR, OPR, IV, FM and OCC. This means that the information from the lower layer, in which the detail information is obtained, makes some help for tracking. SiamFC_3s tracker just cross-correlated two feature maps in the final layer from the two branches and did not consider the information in the lower layer.

Figure 2 illustrates the success plots and the precision plots of OPE for the top 10 trackers. Trackers with deep learning are obviously outperforming the traditional trackers. It proves the good performance pf deep learning approach. The proposed approach and SiamFC_c and staple trackers are much better than others. For the success plots of OPE, the proposed approach performs better than SiamFC_3c; however, for the precision plots, the value of the proposed approach declines to 0.733, while that of SiamFC_3s is 0.744. The detail information of conv4 contributes to the tracking results though sometimes it does not work as shown in Fig. 2.

Due to limited space, in Fig. 3, the qualitative comparison of our approach with state-of-the-art trackers is shown on some of sequences: trellis, soccer, singer1, shaking, matrix, ironman, deer, couple, coke and jogging_2. We take different color to represent different trackers on the trellis, soccer, singer1, shaking, matrix, ironman, deer, couple, coke and jogging_2. Some more detailed success plots and precision plots are provided in Figs. 4 and 5.

5 Conclusion

In this paper, we propose a simple tracker with fully convolutional Siamese network. The detailed information from the fourth layer is considered for the tracking result, and we demonstrate that the end-to-end Hierarchical convolutional features for tracking algorithm perform efficiently to track the object, especially in the situation with background cluster, in-plane rotation and scale variation. The experiments show that the lower layer has the ability to provide rich information for the tracking process. We will improve this approach for better performance.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China under Grant 61472110, 61772161, 61532006, 61320106006 and 61601158, by the Zhejiang Provincial Science Foundation under Grants LQ16F030004.

References

1. Wu, Y., Lim, J., Yang, M.H.: online object tracking: a benchmark. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2411–2418 (2013)
2. Yu, J., Hong, C., Rui, Y., Tao, D.: Multi-task autoencoder model for recovering human poses. IEEE Trans. Ind. Electron. (2017)
3. Yu, J., Rui, Y., Tao, D.: Click prediction for web image reranking using multimodal sparse coding. IEEE Trans. Image Process. **23**(5), 2019 (2014)
4. Yu, J., Zhang, B., Kuang, Z., Lin, D., Fan, J.: iPrivacy: image privacy protection by identifying sensitive objects via deep multi-task learning. IEEE Trans. Inf. Forensics Secur. **12**(5), 1005 (2017)
5. Danelljan, M., Hager, G., Shahbaz Khan, F., Felsberg, M.: Convolutional features for correlation filter based visual tracking. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 58–66 (2015)
6. Danelljan, M., Bhat, G., Khan, F.S., Felsberg, M.: Eco: Efficient convolution operators for tracking. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 21–26 (2017)
7. Li, K., Kong, Y., Fu, Y.: Multi-stream deep similarity learning networks for visual tracking. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence, pp. 2166–2172. AAAI Press, Palo Alto (2017)
8. Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H.: Fully-convolutional siamese networks for object tracking. In: European Conference on Computer Vision, pp. 850–865. Springer, Berlin (2016)
9. Valmadre, J., Bertinetto, L., Henriques, J., Vedaldi, A., Torr, P.H.: End-to-end representation learning for correlation filter based tracking. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 5000–5008 (2017)
10. Zhong, W., Lu, H., Yang, M.H.: Robust object tracking via sparsity-based collaborative model. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 1838–1845 (2012)
11. Hare, S., Golodetz, S., Saffari, A., Vineet, V., Cheng, M.M., Hicks, S.L., Torr, P.H.: Struck: structured output tracking with kernels. IEEE Trans. Pattern Anal. Mach. Intell. **38**(10), 2096 (2016)
12. Kalal, Z., Mikolajczyk, K., Matas, J.: Tracking-learning-detection. IEEE Trans. Pattern Anal. Mach. Intell. **34**(7), 1409 (2012)
13. Sevilla-Lara, L., Learned-Miller, E.: Distribution fields for tracking. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 1910–1917 (2012)
14. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.: Exploiting the circulant structure of tracking-by-detection with kernels. In: European Conference on Computer Vision, pp. 702–715. Springer, Berlin (2012)
15. Jia, X., Lu, H., Yang, M.H.: Visual tracking via adaptive structural local sparse appearance model. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 1822–1829 (2012)
16. Zhang, T., Ghanem, B., Liu, S., Ahuja, N.: Robust visual tracking via multi-task sparse learning. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 2042–2049 (2012)
17. Oron, S., Bar-Hillel, A., Levi, D., Avidan, S.: Locally orderless tracking. Int. J. Comput. Vis. **111**(2), 213 (2015)
18. Baldi, P., Chauvin, Y.: Neural networks for fingerprint recognition. Neural Comput. **5**(3), 402 (1993)
19. Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R.: Signature verification using a "siamese" time delay neural network. In: Advances in Neural Information Processing Systems, pp. 737–744 (1994)

20. Chen, K., Tao, W.: Once for all: a two-ow convolutional neural network for visual tracking. In: IEEE Transactions on Circuits and Systems for Video Technology (2017)
21. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
22. Tao, R., Gavves, E., Smeulders, A.W.: Siamese instance search for tracking. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 1420–1429 (2016)
23. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
24. Held, D., Thrun, S., Savarese, S.: Learning to track at 100 fps with deep regression networks. In: European Conference on Computer Vision, pp. 749–765. Springer, Berlin (2016)
25. Ma, C., Huang, J.B., Yang, X., Yang, M.H.: Hierarchical convolutional features for visual tracking. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3074–3082 (2015)
26. Wang, L., Ouyang, W., Wang, X., Lu, H.: Visual tracking with fully convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3119–3127 (2015)
27. Bertinetto, L., Valmadre, J., Golodetz, S., Miksik, O., Torr, P.H.: Staple: Complementary learners for real-time tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1401–1409 (2016)
28. Bao, C., Wu, Y., Ling, H., Ji, H.: Real time robust l1 tracker using accelerated proximal gradient approach. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 1830–1837 (2012)
29. Dinh, T.B., Vo, N., Medioni, G.: Context tracker: Exploring supporters and distracters in unconstrained environments. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 1177–1184 (2011)
30. Wu, Y., Shen, B., Ling, H.: Online robust image alignment via iterative convex optimization. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 1808–1814 (2012)

Suguo Zhu received the B.S. degree in School of Computer Science and Technology from Henan Normal University, 2008, Master's Degree from School of Computer from Gugangzhou University of Technology, 2012, and Ph.D. Degree in School of Compute Science and Technology from Beijing University of Posts and Telecommunications, 2016. Now, she is working at School of Computer Science and Technology, Hangzhou Dianzi University. Her research interests include computer vision and machine learning.

Zhenying Fang received the B.S. degree in Information and Computing Science from Henan University of Engineering, 2012. Now, he is a graduate student of the Media Intelligence Lab in Hangzhou Dianzi University, Hangzhou, China. His research interests are in computer vision, focusing on applications of deep learning to object tracking and action recognition/detection in video.

Fei Gao received the B.Sc. degree in electrical information engineering and the Ph.D. degree in intelligent information processing from Xidian University in 2009 and 2015, respectively. From October 2012 to September 2013, he studied at the University of Technology, Sydney, NSW, Australia, as a visiting Ph.D. student. Now, he is working at School of Computer Science and Technology, Hangzhou Dianzi University. His research interests include computer vision and machine learning.