IEICE TRANS. INF. & SYST., VOL.E100–D, NO.3 MARCH 2017

LETTER

Feature Adaptive Correlation Tracking

Yulong XU[†], **Yang LI**[†], **Jiabao WANG**[†], **Zhuang MIAO**[†a)], ***Nonmembers***, **Hang LI**[†], ***Member***, ***and*** **Yafei ZHANG**[†], ***Nonmember***

SUMMARY Feature extractor plays an important role in visual tracking, but most state-of-the-art methods employ the same feature representation in all scenes. Taking into account the diverseness, a tracker should choose different features according to the videos. In this work, we propose a novel feature adaptive correlation tracker, which decomposes the tracking task into translation and scale estimation. According to the luminance of the target, our approach automatically selects either hierarchical convolutional features or histogram of oriented gradient features in translation for varied scenarios. Furthermore, we employ a discriminative correlation filter to handle scale variations. Extensive experiments are performed on a large-scale benchmark challenging dataset. And the results show that the proposed algorithm outperforms state-of-the-art trackers in accuracy and robustness.

key words: *visual tracking, correlation filter, feature selection, convolutional neural networks, scale estimation*

1. Introduction

Visual object tracking is one of the active research topics in the field of computer vision, which plays an important role in many applications ranging from surveillance and robotics to human-computer interactions and driverless vehicle [1], [2]. In visual tracking, the task is to estimate the locations of a target in an image sequence by giving only its initial position. This is especially challenging due to several factors such as occlusion, scale changes, illumination variations, fast motion, rotation and background clutter [3].

In recent years, the correlation filter-based approaches [4]–[7] have been widely used and highly developed because of their outstanding efficiency in tracking. And in order to improve the tracking performance, many features are applied to the field of object tracking in correlation filter framework. Heriques et al. proposed the circulant structure with kernels (CSK) by using correlation filters in a kernel space [4]. The CSK tracker built only on illumination intensity features and was further improved by using histogram of oriented gradients (HOG) features [8], [9] in the kernelized correlation filter (KCF) tracking algorithm [5]. On the basis of KCF and HOG features, a series of trackers are proposed. Danelljan et al. presented the discriminative scale space tracker (DSST) [6] to cope with scale changes. The spatially regularized discriminative correlation filter (SRDCF) tracker [7] employed a spatial regularization component which allows the correlation filters to be learned on a significantly larger set of negative training samples, without corrupting the positive samples.

Although these approaches are satisfactory in constrained environments, there is a limitation that they resort to hand-crafted features, which play an important role in visual tracking [10]. With the development of deep learning technology, convolutional neural networks (CNNs) have demonstrated their outstanding representation power in a wide range of computer vision applications [11]. And some tracking algorithms using the representations from CNNs have been proposed [12], [13], [22]. In addition, Ma et al. [14] utilized convolutional features and learned correlation filters on each CNN layer without re-training. Li et al. [23] employed DSST [6] approach to cope with scale changes and proposed a scale adaptive tracker with hierarchical convolutional features.

This study proposes a novel feature adaptive correlation tracking algorithm, which decomposes the tracking task into translation and scale estimation. First, according to the luminance of the tracking target, the proposed approach automatically selects either hierarchical convolutional features or HOG features for varied scenarios. And we infer the target location based on multi-level correlation response maps. Second, we use a discriminative correlation filter with HOG features to handle scale variations. Third, extensive experiments are performed on a large-scale benchmark challenging dataset with 50 challenging image sequences [3]. And the results show that the proposed algorithm outperforms state-of-the-art tracking methods in accuracy and robustness.

2. Feature Adaptive Tracking with Scale Estimation

2.1 Problem of Convolutional Features Based Correlation Trackers

KCF [5] is a typical correlation tracker, which models the appearance of a target using a filter $\mathbf{w}$ trained on an image patch $\mathbf{x}$ with size $M \times N$ pixels. $M \times N$ is set to $\beta w \times \beta h$, where $w \times h$ is the size of the target and β is the expansion coefficient. KCF considers all cyclic shifts $\mathbf{x}_{m,n}$, $(m,n) \in \{0,\cdots,M-1\} \times \{0,\cdots,N-1\}$, as the training examples. They are labeled with a Gaussian function $\mathbf{y}(m,n)$.

Manuscript received July 20, 2016.
Manuscript revised October 8, 2016.
Manuscript publicized November 28, 2016.

[†]The authors are with the College of Command Information Systems, PLA University of Science and Technology (PLAUST), Nanjing, China.
a) E-mail: emiao_beyond@163.com (Corresponding author)

DOI: 10.1587/transinf.2016EDL8164

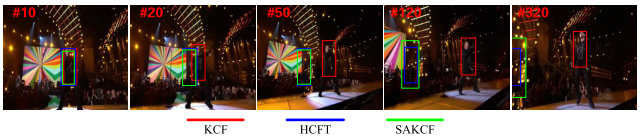

Fig. 1 Comparisons of KCF [5] with state-of-the-art convolutional features based correlation trackers on *Singer2* sequence. KCF [5] works well while HCFT [14] and SAKCF [23] drift after the 20th frame.

The objective function is

$$\mathbf{w} = \arg\min_{\mathbf{w}} \sum_{m,n} \left| \langle \varphi(\mathrm{x}_{m,n}), \mathbf{w} \rangle - \mathrm{y}(m,n) \right|^2 + \lambda \|\mathbf{w}\|^2 , \quad (1)$$

where φ represents the mapping to the Hilbert space induced by the kernel κ. The constant $\lambda \geq 0$ is a regularization parameter. The objective function is minimized as $\mathbf{w} = \sum_{m,n} \alpha(m,n)\varphi(\mathrm{x}_{m,n})$, and the coefficient α is defined by $A = \mathscr{F}(\alpha) = \frac{Y}{K_{\mathrm{xx}}+\lambda}$, where $\mathscr{F}$ denotes the Fourier transform; $Y = \mathscr{F}(\mathrm{y})$; $K_{\mathrm{xx}} = \mathscr{F}(\mathbf{k}_{\mathrm{xx}})$. $\mathbf{k}_{\mathrm{xx}}(m,n) = \kappa(\mathrm{x}_{m,n}, \mathrm{x})$ is the output of the kernel function κ.

A patch z with the same size of x is cropped out in the next frame. The confidence score is computed as

$$\hat{\mathrm{y}} = \mathscr{F}^{-1}\left(\hat{A} \odot K_{\mathrm{z}\hat{\mathrm{x}}}\right). \quad (2)$$

Here $\mathscr{F}^{-1}$ denotes the inverse Fourier transform; $\hat{A}$ denotes the learned classifier coefficients; $\hat{\mathrm{x}}$ denotes the learned target appearance; $\odot$ is the element-wise product. The target location in the new frame is then detected by searching the position with the highest score.

The feature extractor is an important component of a robust tracking system. Using proper features can significantly improve the tracking performance. The convolutional features have been employed in correlation tracking and achieve very good performance [14], [23]. However, when the luminance is low, the convolutional features x in (1) will be close to zero, which will cause tracking drift. As shown in Fig. 1, the convolutional features based correlation trackers cannot perform low illumination images well, while the KCF tracker accomplishes the mission better. The reason is that the convolutional features are close to zero under low illumination conditions (In fact, the average gray value V of the object in first frame is less than 40, $V \in [0, 255]$) and the KCF tracker utilizes HOG features which contain a lot of gradient information. Taking into account the diverseness, a tracker should choose different features in different scenes.

2.2 Feature Adaptive Correlation Tracker

According to the average gray value of the object in the first frame, our approach automatically selects either hierarchical convolutional features or HOG features in translation, as shown in Fig. 2. When the average gray value is larger than a threshold $v \in [0, 255]$ (e.g. $v = 40$ in this work), our approach will choose hierarchical convolutional features, otherwise, select HOG.

As shown in Fig. 2, if $V < 40$, we choose HOG features and employ the correlation filters to estimate the target

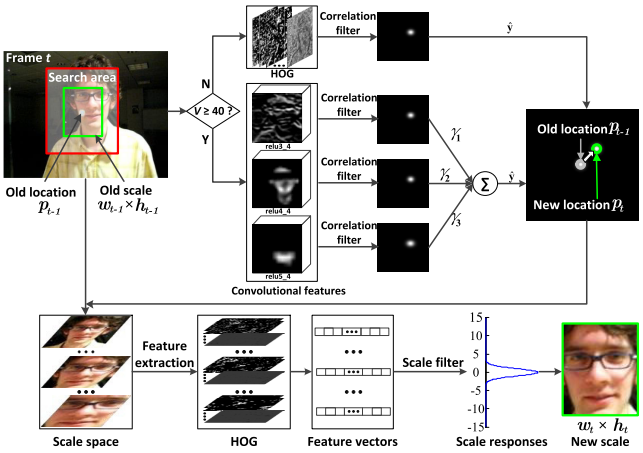

Fig. 2 Flowchart of the proposed tracking algorithm.

position, which is the same as KCF tracker.

If $V \geq 40$, we first adopt the VGG-19 [15] trained on ImageNet for feature extraction. For example, given an image patch with size $M \times N$, we first resize the image patch to 224×224 which is the requirement input of the VGG-19. Then, the outputs of the *relu3_4*, *relu4_4*, and *relu5_4* layer feature maps are used as multi-channel features. In order to further remove the boundary discontinuities of the response map, the extracted CNN feature channels are weighted by a cosine window. Due to the pooling operators used in the CNNs model, the spatial resolutions of the *relu3_4*, *relu4_4*, and *relu5_4* layer are different. Therefore, we resize each feature map to a fixed larger size $M/4 \times N/4$ with bilinear interpolation.

For each resized feature map l, we can learn a correlation filter using (1) and get a response map $\hat{\mathrm{y}}_l$ using (2). So we can give different layers with different weights to combine these response maps for robust object tracking, as shown in Fig. 2. Therefore, the final correlation response map $\hat{\mathbf{y}}$ is the linear combination of the three correlation response maps,

$$\hat{\mathbf{y}} = \sum_{l=1}^{3} \gamma_l \hat{\mathrm{y}}_l, \quad (3)$$

where γ_l is the weight of the different response map $\hat{\mathrm{y}}_l$. And the target location can be estimated by searching the position with the maximum value of the correlation response map $\hat{\mathbf{y}}$.

2.3 HOG Features for Scale Estimation

We employ a scale pyramid [6] and learn a discriminative correlation filter to cope with scale variations as shown in Fig. 2. After finding the target location p_t in current frame t, S image patches centered around p_t are cropped from the frame, and each with size $a^s w_{t-1} \times a^s h_{t-1}$, where $w_{t-1} \times h_{t-1}$ is the target scale in the previous frame $t-1$, a is the scale factor, and $s \in \left\{ \left\lfloor -\frac{S-1}{2} \right\rfloor, \left\lfloor -\frac{S-3}{2} \right\rfloor, \cdots, \left\lfloor \frac{S-1}{2} \right\rfloor \right\}$. Then all the S image patches are operated for feature extraction after resizing to the template size, and the features are set to feature

vectors for a scale filter. The highest scale response will be find to estimate the current scale $w_t \times h_t$ as

$$\begin{cases} w_t = a^{s'} w_{t-1} \\ h_t = a^{s'} h_{t-1} \end{cases}, \quad (4)$$

where s' is the scale number with the highest response.

Finally, the total model parameters are updated frame by frame as

$$\begin{cases} \hat{\mathrm{x}}_t = (1-\eta)\hat{\mathrm{x}}_{t-1} + \eta \mathrm{z}_t \\ \hat{A}_t = (1-\eta)\hat{A}_{t-1} + \eta A_t \end{cases}, \quad (5)$$

where η is the learning rate.

3. Experiments

3.1 Experimental Setup

The approach is implemented in Matlab 2013a and a large-scale benchmark dataset [3] are employed, which contains 50 challenging videos. Our implementation runs at 9 frames per second on a computer with an Intel i5-4690 CPU (3.50 GHz) and 16 GB RAM. The main computing time of our tracker is the forward propagation process to extract multi-channel convolutional features. The expansion coefficient β for translation estimation is set to 2.8. The regularization parameter λ is set to 10^{-4}. We use a linear kernel $\kappa(\mathrm{x}, \mathrm{x}') = \mathrm{x}^T\mathrm{x}'$ in translation estimation, and the weight value $[\gamma_1, \gamma_2, \gamma_3]$ is set to [0.01, 0.65, 1] for the *relu3_4*, *relu4_4*, and *relu5_4* layer correlation response maps. The scale space number S is set to 27 and the scale factor a is set to 1.035. The learning rate η is set to 0.01. We use the same parameter values for all the sequences.

The results are evaluated by using center location error (CLE), distance precision (DP) and overlap precision (OP). CLE is the average Euclidean distance between the center of tracked target and the ground-truth. DP is the percentage of frames where the CLE is smaller than 20 pixels. OP is the percentage of frames where the bounding box overlap is greater than 0.5.

3.2 Robust Feature Selection

In the benchmark dataset [3], a total of four videos (*Ironman, Matrix, Singer2 and Trellis*) where the average gray value of the object in the first frame is less than 40. For these videos, our method will choose HOG features. The comparison of our approaches (FACT and SAT) and convolutional features based correlation trackers (HCFT [14] and SAKCF [23]) on the four sequences is shown in Table1. FACT denotes our feature adaptive correlation tracker and SAT denotes the scale adaptive tracker without feature selection, which uses only convolutional features in tracking.

Compared with SAT method, the performance is improved further by using feature selection scheme in FACT. Compared with SAKCF [23] method, our FACT approach improves the average DP from 58.3% to 60%, outperforms it by 4.7% in average OP. These demonstrate the effectiveness of our feature selection scheme in tracking.

Table 1 Comparison with convolutional features based correlation trackers on *Ironman, Matrix, Singer2 and Trellis* sequences

Evaluation	Ours		State-of-the-art trackers	
	FACT	SAT	SAKCF	HCFT
Average CLE (pixels)	63.9	73.3	**63.1**	63.4
Average DP (%)	**60**	57.8	58.3	57.8
Average OP (%)	**57.1**	46.7	52.4	47

3.3 Comparison with State-of-the-Art Trackers

We compare our approach with 12 state-of-the-art trackers: CN [16], CSK [4], DSST [6], HCFT [14], KCF [5], PCOM [17], RPT [18], SAMF [19], SRDCF [7], Struck [20], TGPR [21] and SAKCF [23]. The comparison on the 50 challenging benchmark image sequences is shown in Table2. We present the results using average CLE, average DP and average OP over all sequences. The best three results are highlighted by **bold**, ***italics*** and underline, respectively.

Among the trackers in our evaluation, our SAT method achieves better performance and the performance is improved further by using our proposed feature selection method. our FACT method significantly provides the best results with an average CLE of 12.6 pixels, an average DP of 90.4% and an average OP of 85.3%. These results clearly demonstrate the robustness of our method. Figure 3 contains the precision and success plots illustrating the average DP and OP over all the 50 benchmark sequences. In both precision and success plots, the proposed method outperforms the best existing method SAKCF. In summary, the precision plot shows that our method is more robust than state-of-the-

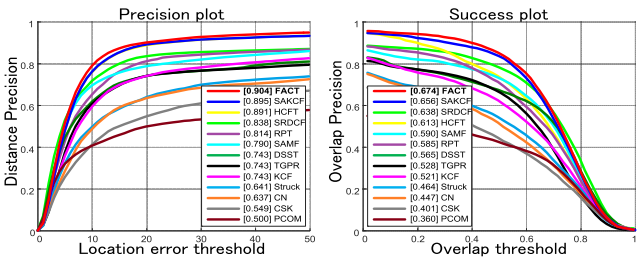

Fig. 3 Precision and success plots over all of the 50 sequences. The legend of the precision plot reports the average DP score at 20 pixels for each method and the legend of the success plot contains the area under the curve (AUC) score for each tracker.

Table 2 Comparison with 12 state-of-the-art trackers on the 50 benchmark sequences.

Evaluation	FACT	SAT	SAKCF	SRDCF	HCFT	SAMF	RPT	DSST	TGPR	KCF	Struck	CN	CSK	PCOM
Average CLE (pixels)	**12.6**	***13.3***	15.6	35.1	15.7	28.4	36.5	40.9	45.8	35.4	54.3	64.1	88.8	78.0
Average DP (%)	**90.4**	***90.3***	89.5	83.8	89.1	79.0	81.4	74.3	74.3	74.3	64.1	63.7	54.9	50.0
Average OP (%)	**85.3**	***84.5***	83.8	78.4	74.0	73.3	71.2	67.4	66.6	62.4	54.3	51.7	44.4	42.5

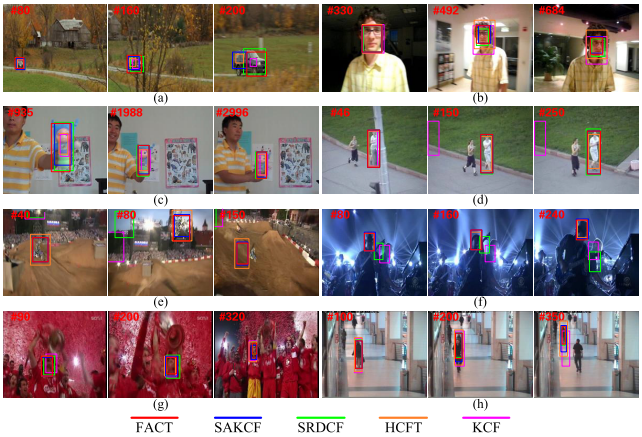

Fig. 4 A visualization of the tracking results of our approach and the state-of-the-art visual trackers on 8 benchmark sequences.

art trackers, and the success plot demonstrates that our approach computes scale more accurately on the benchmark sequences.

Figure 4 illustrates a qualitative comparison with selected trackers on 8 benchmark videos. These videos pose challenging problems such as scale changes (Fig. 4 (a), (e), (g) and (h)), illumination variations (Fig. 4 (b), (c) and (f)), occlusion (Fig. 4 (a), (d), (g) and (h)), rotation (Fig. 4 (b), (e) and (g)), fast motion (Fig. 4 (a), (e) and (g)) and background clutter (Fig. 4 (e), (f) and (g)). Despite these challenges, our approach obtains the both positions and scale of the target accurately.

4. Conclusion

In this research, we propose a novel a novel feature adaptive correlation tracking algorithm, which decomposes the tracking task into translation and scale estimation. Our approach automatically selects either hierarchical convolutional features or HOG features in translation for varied scenarios. Then, we employ a discriminative correlation filter to handle scale variations. Extensive experiments are performed on 50 challenging benchmark sequences. And the results show that the proposed algorithm outperforms state-of-the-art trackers in accuracy and robustness.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (61402519) and the Jiangsu Provincial Nature Science Foundation of China (BK20140071).

References

[1] A.W.M. Smeulders, D.M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah, "Visual tracking: An experimental survey," IEEE Trans. Patt. Anal. Mach. Intell., vol.36, no.7, pp.1442–1468, July 2014.

[2] B. Guo, and J. Liu, "Real-time tracking with online constrained compressive learning," IEICE Trans. Inf. & Sys., vol.E96-D, no.4, pp.988–992, April 2013.

[3] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," Proc. IEEE Conf. Comput. Vis. Patt. Recog., Portland, USA, pp.2411–2418, June 2013.

[4] J.F. Henriques, R. Caseiro, P. Martins, and J. Batista, "Exploiting the circulant structure of tracking-by-detection with kernels," Proc. 12th European Conf. Comput. Vis., Florence, Italy, vol.7575, pp.702–715, Oct. 2012.

[5] J.F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," IEEE Trans. Patt. Anal. Mach. Intell., vol.37, no.3, pp.583–596, March 2015.

[6] M. Danelljan, G. Häger, F.S. Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," Proc. British Machine Vis. Conf., Nottingham, UK, pp.1–11, Sept. 2014 .

[7] M. Danelljan, G. Häger, F.S. Khan, and M. Felsberg, "Learning spatially regularized correlation filters for visual tracking," Proc. IEEE Int. Conf. Comput. Vis., Santiago, Chile, pp.4310–4318, Dec. 2015.

[8] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," IEEE Trans. Patt. Anal. Mach. Intell., vol.32, no.9, pp.1627–1645, Sept. 2010.

[9] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," IEEE Trans. Patt. Anal. Mach. Intell., vol.36, no.8, pp.1532–1545, Aug. 2014.

[10] N. Wang, J. Shi, D.-Y. Yeung, and J. Jia, "Understanding and diagnosing visual tracking systems," Proc. IEEE Int. Conf. Comput. Vis., Santiago, Chile, pp.3101–3109, Dec. 2015.

[11] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol.521, no.7553, pp.436–444, May 2015.

[12] J. Fan, W. Xu, Y. Wu, and Y. Gong, "Human tracking using convolutional neural networks," IEEE Trans. Neural Networks, vol.21, no.10, pp.1610–1623, Oct. 2010.

[13] L. Wang, T. Liu, G. Wang, K.L. Chan, and Q. Yang, "Video tracking using learned hierarchical features," IEEE Trans. Image Process., vol.24, no.4, pp.1424–1425, April 2015.

[14] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Hierarchical convolutional features for visual tracking," Proc. IEEE Int. Conf. Comput. Vis., Santiago, Chile, pp.3074–3082, Dec. 2015.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," Proc. IEEE Int. Conf. Learning Representations, San Diego, California, USA, May 2015.

[16] M. Danelljan, F.S. Khan, M. Felsberg, and J. van de Weijer, "Adaptive color attributes for real-time visual tracking," Proc. IEEE Conf. Comput. Vis. Patt. Recog., Columbus, USA, pp.1090–1097, June 2014.

[17] D. Wang and H. Lu, "Visual tracking via probability continuous outlier model," Proc. IEEE Conf. Comput. Vis. Patt. Recog., Columbus, USA, pp.3478–3485, June 2014.

[18] Y. Li, J. Zhu, and S.C.H. Hoi, "Reliable patch trackers: Robust visual tracking by exploiting reliable patches," Proc. IEEE Conf. Comput. Vis. Patt. Recog., Boston, USA, pp.353–361, June 2015.

[19] Y. Li and J. Zhu, "A scale adaptive kernel correlation filter tracker with feature integration," Proc. 13th European Conf. Comput. Vis., Zurich, Switzerland, pp.254–265, Sept. 2014.

[20] S. Hare, A. Saffari, and P.H.S. Torr, "Struck: Structured output tracking with kernels," Proc. IEEE Int. Conf. Comput. Vis., Barcelona, Spain, pp.263–270, Nov. 2011.

[21] J. Gao, H. Ling, W. Hu, and J. Xing, "Transfer learning based visual tracking with Gaussian processes regression," Proc. 13th European Conf. Comput. Vis., pp.188–203, Zurich, Switzerland, Sept. 2014.

[22] M. Zhai, M.J. Roshtkhari, G. Mori, "Deep learning of appearance models for online object tracking," arXiv, 2016. Available: http://arxiv.org/pdf/1607.02568v1.

[23] Y. Li, Y. Zhang, Y. Xu, J. Wang, and Z. Miao, "Robust scale adaptive kernel correlation filter tracker with hierarchical convolutional features," IEEE Signal Process. Lett., vol.23, no.8, pp.1136–1140, Aug. 2016.