An improved long-term correlation tracking method with occlusion handling

Junhao Zhao (赵俊豪), Gang Xiao (肖 刚)*, Xingchen Zhang (张星辰), and D. P. Bavirisetti (杜尔·普拉萨德·巴维瑞瑟特)**

School of Aeronautics and Astronautics, Shanghai Jiao Tong University, Shanghai 200240, China
**Corresponding author: xiaogang@sjtu.edu.cn; **corresponding author: xingchen@sjtu.edu.cn*
Received July 28, 2018; accepted December 20, 2018; posted online February 25, 2019

By improving the long-term correlation tracking (LCT) algorithm, an effective object tracking method, improved LCT (ILCT), is proposed to address the issue of occlusion. If the object is judged being occluded by the designed criterion, which is based on the characteristic of response value curve, an added re-detector will perform re-detection, and the tracker is ordered to stop. Besides, a filtering and adoption strategy of re-detection results is given to choose the most reliable one for the re-initialization of the tracker. Extensive experiments are carried out under the conditions of occlusion, and the results demonstrate that ILCT outperforms some state-of-the-art methods in terms of accuracy and robustness.

OCIS codes: 100.4999, 110.4155, 330.4150.
doi: 10.3788/COL201917.031001.

Nowadays, object tracking is one of the hot topics in computer vision and has been widely used in many engineering applications, such as satellites[1], inverse synthetic aperture radar[2], and reconnaissance[3]. A typical scenario of object tracking is to track an unknown object initialized by a bounding box in subsequent image frames[4,5]. How to realize a robust tracker against significant appearance change is still an issue to be addressed.

In recent years, many robust trackers based on correlation filters were proposed, in which the minimum output sum of squared error (MOSSE) filter[6] is the well-known one because of its high speed and novelty. It introduced correlation operation[7] into object tracking and greatly accelerated the calculation through the theory that convolution in the spatial domain becomes the Hadamard product in the Fourier domain[8]. After that, the circulant structure of tracking-by-detection with kernels (CSK)[9] employed the circulant matrix originally to increase the number of samples that improved the classifier. Then, histogram of oriented gradients (HOG) features, Gaussian kernels, and ridge regression are used in the kernelized correlation filters (KCFs)[10] based on CSK, which has achieved satisfactory tracking results. Danelljan *et al.* mainly solved the issue of scale variation (SV) during an object's movement by their discriminative scale space tracking (DSST)[11], which is based on learning the correlation filters by a scale pyramid. Ma *et al.* proposed long-term correlation tracking (LCT)[4], which comprises the correlation filters of appearance and motion to estimate the scale and translation of an object. It is an outstanding tracker for long-term tracking. Inspired by the model of human's recognition, Choi *et al.* proposed the attentional feature-based correlation filter (AtCF)[12] to perform object tracking that can adapt to the fast variation of the object.

However, these trackers do not handle occlusion (OCC) well or only aim at partial OCC (50% coverage or less) and temporal full OCC. A robust tracking algorithm requires a detection module to recover the target from potential tracking failures caused by heavy OCC[13]. Because LCT is designed for long-term tracking, we improved it to handle OCC and named it improved LCT (ILCT).

ILCT uses the motion correlation filter $\mathbf{w}_1$ and appearance correlation filter $\mathbf{w}_2$ of LCT to estimate the position and scale of an object. An OCC criterion is designed according to the response value curve of the correlation filter to determine whether the object is occluded or not. If the object is occluded, the added re-detector works, and the tracker stops. The reliable re-detection result will re-initialize the tracker. The principles and experimental results are explained below.

LCT decomposes the tracking task into translation estimating (to get a new position) and scale estimating (to get a new scale)[4]. The process is realized by motion correlation filter $\mathbf{w}_1$ and appearance correlation filter $\mathbf{w}_2$, respectively.

This filter, $\mathbf{w}_1$, is trained on image patch $\mathbf{x}$, whose size is $M \times N$ by circular shifts of its pixels $x_{m,n}$, where $(m, n) \in \{0, 1, ..., M-1\} \times \{0, 1, ..., N-1\}$ as training samples[9]. Using the ridge regression to minimize the mean square error between the training images and regression object, then the filter $\mathbf{w}_1 \in \mathbf{R}^{M \times N}$ is obtained by

$$\mathbf{w}_1 = \arg\min_{\mathbf{w}_1} \sum_{m,n} |\boldsymbol{\phi}(x_{m,n}) \cdot \mathbf{w}_1 - y(m,n)|^2 + \lambda |\mathbf{w}_1|^2, \quad (1)$$

where Gaussian kernel $k(\mathbf{x}, \mathbf{x}') = \exp(-|\mathbf{x} - \mathbf{x}'|^2/\sigma^2)$ is used to define mapping $\boldsymbol{\phi}$ as $k(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}(\mathbf{x}) \cdot \boldsymbol{\phi}(\mathbf{x}')$. According to the distance of shift, $y(m, n)$ gives the Gaussian label to the training image that the value is close to 1 if there is less distance. λ is the regulation parameter.

After mapping and fast Fourier transform (FFT) $\mathbf{F}$, the solution of $\mathbf{w}_1$ can be represented as the linear combination

of training samples $\mathbf{w}_1 = \sum_{m,n} a(m,n)\boldsymbol{\phi}(x_{m,n})$, where the coefficient $\mathbf{a}$ is given by

$$A = \mathbf{F}(\mathbf{a}) = \frac{\mathbf{F}(\mathbf{y})}{\mathbf{F}[\boldsymbol{\phi}(\mathbf{x}) \cdot \boldsymbol{\phi}(\mathbf{x})] + \lambda}. \tag{2}$$

When new frame comes, the filter will perform correlation on the new patch $\mathbf{z}$ around the last location. Then, the correlation response map can be calculated by

$$\hat{y} = \mathbf{F}^{-1}\{A \odot \mathbf{F}[\boldsymbol{\phi}(\mathbf{z}) \cdot \boldsymbol{\phi}(\hat{\mathbf{x}})]\}, \tag{3}$$

where $\hat{\mathbf{x}}$ denotes the learned appearance model, $\mathbf{F}^{-1}$ is the inverse FFT, and $\odot$ is elementwise multiplication. The value of $\hat{y}$ is between 0 and 1, and the location that owns the highest $\hat{y}$ is the position of the object.

The filter $\mathbf{w}_2$ shares the same principle of $\mathbf{w}_1$. In the process of estimating the scale, the patch after translating estimation is divided into K scales:

$$S = \left\{ a^k | k = \left\lfloor -\frac{K-1}{2} \right\rfloor, \left\lfloor -\frac{K-3}{2} \right\rfloor, ..., \left\lfloor \frac{K-1}{2} \right\rfloor \right\}. \tag{4}$$

Each scale has its size of $sM \times sN (s \in S)$, and HOG features are extracted on each one to build the scale pyramid. As in Eq. (5), we can get the response value of each layer in the pyramid and select the patch that owns the highest value as the object scale $\hat{s}$:

$$\hat{s} = \arg\max_s [\max(\hat{y}_1), \max(\hat{y}_2), ..., \max(\hat{y}_s)]. \tag{5}$$

Therefore, the accepted tracking result of LCT must have two highest response values, i.e., the response values of $\mathbf{w}_1$ and $\mathbf{w}_2$. The final response map (refers to the map of $\mathbf{w}_2$, the same as below) is shown in Fig. 1(a).

The motion correlation filter $\mathbf{w}_1$ and appearance model $\mathbf{x}$ follow the updated framework with the learning rate α by Eq. (6). The appearance model is trained by the feature vector $\mathbf{x}$ with a 47 channels feature[14], which includes HOG features with 31 bins, eight bins of histogram feature of intensity, and eight bins of non-parametric local rank transformation[15] of the brightness channel. A threshold is set for the appearance correlation filter $\mathbf{w}_2$ such that if $\max(\hat{y}_s) \geq \tau_a$, $\mathbf{w}_2$ will be updated in the same way:

$$\begin{aligned} \hat{\mathbf{x}}^t &= (1-\alpha)\hat{\mathbf{x}}^{t-1} + \alpha \mathbf{x}^t, \\ \hat{A}^t &= (1-\alpha)\hat{A}^{t-1} + \alpha A^t. \end{aligned} \tag{6}$$

Fig. 1. Response maps. (a) Object is intact; (b) object is occluded.

Adopting the tracking-by-detection framework is also the critical factor showing that LCT is robust for SV, illumination variation (IV), background clutters (BCs), fast motion (FM), etc. An online support vector machine (SVM) classifier is used for recovering targets, and the color channels are quantized as features for detector learning[16]. The intersection over union (IOU) thresholds for positive training samples and negative ones are 0.5 and 0.1, respectively. Another threshold τ_r is set to activate it when $\max(\hat{y}_s) < \tau_r$.

In addition, the cosine window is used in translating estimation to remove the boundary discontinuities of the response map[6].

The tracking result is adopted according to the values of response maps. If the object is intact and undisturbed, the response map is clear, and the white point is obvious. On the contrary, the map is dim, and the point is obscure, for example, when OCC occurs, as shown in Fig. 1(b).

When the OCC begins, the tracker may still locate the object successfully based on previous training. However, as time goes on, the coverage increases, which aggravates the correlation filter so that the tracker will fail to re-track the object after its quitting from OCC.

To design an OCC criterion, several sequences[17] with different attributes, including full/partial OCC, deformation (DEF), BCs, FM, IV, and SV, are studied. In each sequence, we selected five frames, $f = \{f_{t-4}, f_{t-3}, f_{t-2}, f_{t-1}, f_t\}$, which reflect the attribute and their corresponding response values $y = \{y_{t-4}, y_{t-3}, y_{t-2}, y_{t-1}, y_t\}$ to draw the curve, as shown in Fig. 2.

In Fig. 2(a), because of the occluder, the response values decrease, while there are obvious rises in the last five curves. So, the first condition of criterion we set is that the five response values of five consecutive frames decrease continuously. Unlike partial OCC and DEF, the response values decrease drastically due to the full cover on the object by the occluder. Accordingly, the second condition will be reached if y_{t-4} is τ_1 larger than y_t. To ensure the accuracy of judgement, the third condition is the number of elements in y that are less than τ_2 is greater than two. The total three conditions of criterion are summarized below:

(1) $y_{t-4} > y_{t-3} > y_{t-2} > y_{t-1} > y_t$,
(2) $y_{t-4} - y_t \geq \tau_1$,
(3) $y' = \{y' | y_{t-4} < \tau_2, y_{t-3} < \tau_2, y_{t-2} < \tau_2, y_{t-1} < \tau_2, y_t < \tau_2\}, |y'| \geq 2$.

When the OCC criterion triggers, we set five free frames so that no operation is carried out on these frames to (1) let the object be fully occluded, (2) avoid the filters being polluted, and (3) improve the real time.

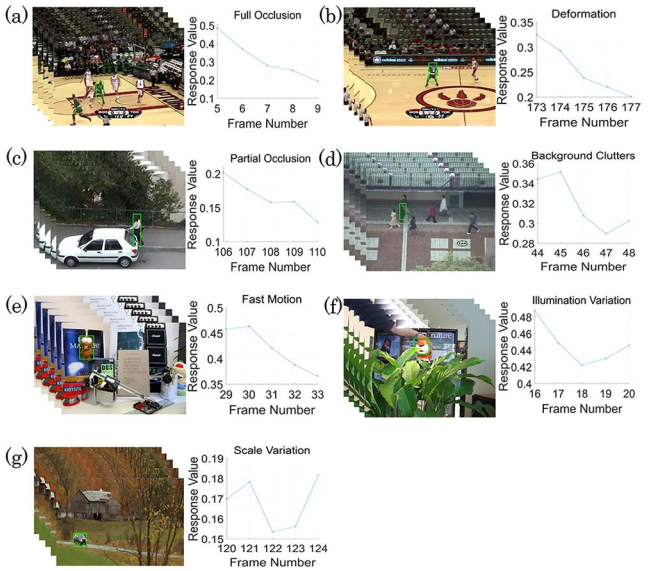

Fig. 2. Curves of response values with different attributes. (a) Full OCC; (b) DEF; (c) partial OCC; (d) BCs; (e) FM; (f) IV; (g) SV. (Best view in PDF.)

For convenience, we name the frame and the time at which it reaches the OCC criterion as f_{occ} and t_{occ}, i.e., $f = \{f_{\text{occ}-4}, f_{\text{occ}-3}, f_{\text{occ}-2}, f_{\text{occ}-1}, f_{\text{occ}}\}$, to meet the three conditions. When the object is identified as occluded, the tracker is ordered to stop, the two filters are no longer updated, and the re-detector is activated. In the re-detection module, we implement the edge boxes[18] to finish this task. Different from other methods of object detection[19], which use sliding windows that consume a large amount of calculation resources, this method efficiently generates object bounding box proposals directly from edges (about 1000 proposals/0.25 s). Its core idea is that the edge of an object corresponds to its contour, and the number of contours wholly enclosed by a bounding box is indicative of the likelihood of the box containing an object. In final, each proposed bounding box has a confidence value that reflects the likelihood of an object. More details can be found in Ref. [18].

In an image with a complex background, a huge number of proposal bounding boxes may be obtained, which takes more computation time, and most of them are not the object bounding boxes we want. So, a threshold is set that k top-ranked proposals are accepted.

While among these k boxes, false results also exist. Considering the SV before and after OCC, a constraint condition is added to filter the unreasonable boxes:

$$1.5^{-1} \times b_w^{f_{\text{occ}-4}} < b_w < 1.5 \times b_w^{f_{\text{occ}-4}},$$
$$1.5^{-1} \times b_h^{f_{\text{occ}-4}} < b_h < 1.5 \times b_h^{f_{\text{occ}-4}}, \quad (7)$$

where $b_w^{f_{\text{occ}-4}}$ and $b_h^{f_{\text{occ}-4}}$ are the width and height of the bounding box of frame $f_{\text{occ}-4}$. An example is shown in Fig. 3, where $k = 500$ is set, and the number of bounding boxes after filtering is $k' = 38$.

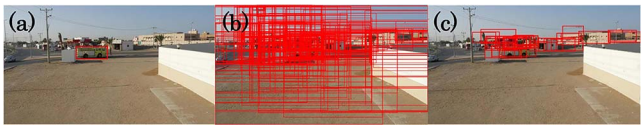

Fig. 3. Detection for proposal boxes. (a) Bounding box of object; (b) k top-ranked proposals; (c) k' proposals.

In our algorithm, $\mathbf{w}_1$ of $f_{\text{occ}-4}$ is implemented on those patches of the final proposed bounding boxes to get the estimated positions. Then, scale estimation is performed by the $\mathbf{w}_2$ of $f_{\text{occ}-4}$, and k' response values are obtained. The detection result will be adopted if the highest value reaches the confidence threshold τ_3. Finally, the result will re-initialize the tracker by giving the new position.

The whole flowchart of our method is shown in Fig. 4.

To demonstrate the performance of the improved tracker, experiments are performed on eight sequences with the attributes of OCC, etc. Eight state-of-the-art trackers are compared with ILCT. They are KCF[10], LCT[4], DSST[11], tracking-learning-detection (TLD)[20], structured output tracking with kernels (Struck)[21], L1 tracker using the accelerated proximal gradient approach (L1APG)[22], integrated CSK (ICSK)[23], and compressive tracking (CT)[24], in which the former three are correlation-based trackers, and the rest are also effective trackers to account for OCC[17]. Besides, for better comparison, first, KCF is improved by the proposed OCC criterion triggers and recovery mechanism, named IKCF. Second, we replace two triggers used in MOSSE[6] and TLD[20], i.e., peak-to-sidelobe ratio (PSR) and median flow (MF) with the proposed one of ILCT, respectively, named LCT-PSR and LCT-MF. The experimental environment is Intel I7-6500U 2.5 GHz CPU with 8.00G RAM, MATLAB 2016b.

The annotated attributes of the eight sequences include OCC, FM, moving camera (MC), SV, BCs, IV, DEF, out-of-plane rotation (OPR), and motion blur (MB). Their information is listed in Table 1. The triumphal arch sequence is taken by us.

The parameters of the LCT part are set to the default values: $\lambda = 10^{-4}$, the size of the search window for translation estimation is set to 1.8 times the target size, the Gaussian kernel width $\sigma = 0.1$, learning rate $\alpha = 0.01$, the number of scale space $|S| = 21$, the scale factor $a = 1.08$, $\tau_r = 0.25$ for the activation of SVM, $\tau_t = 0.5$ for the adoption of the SVM result, and $\tau_a = 0.5$ is set as the threshold for the model update[4].

In the OCC criterion, τ_1 and τ_2 are not fixed and are set to a quarter and a half of the response value of the second frame (the first frame has no correlations, and the object is selected manually), respectively.

In the re-detector, we use the default parameters of edge boxes[18] and set $k = 200$. τ_3 is set to 0.8 times of $y_{\text{occ}-4}$. The rest of the trackers are used with their default parameters.

The tracking results of 12 trackers are shown in Fig. 5.

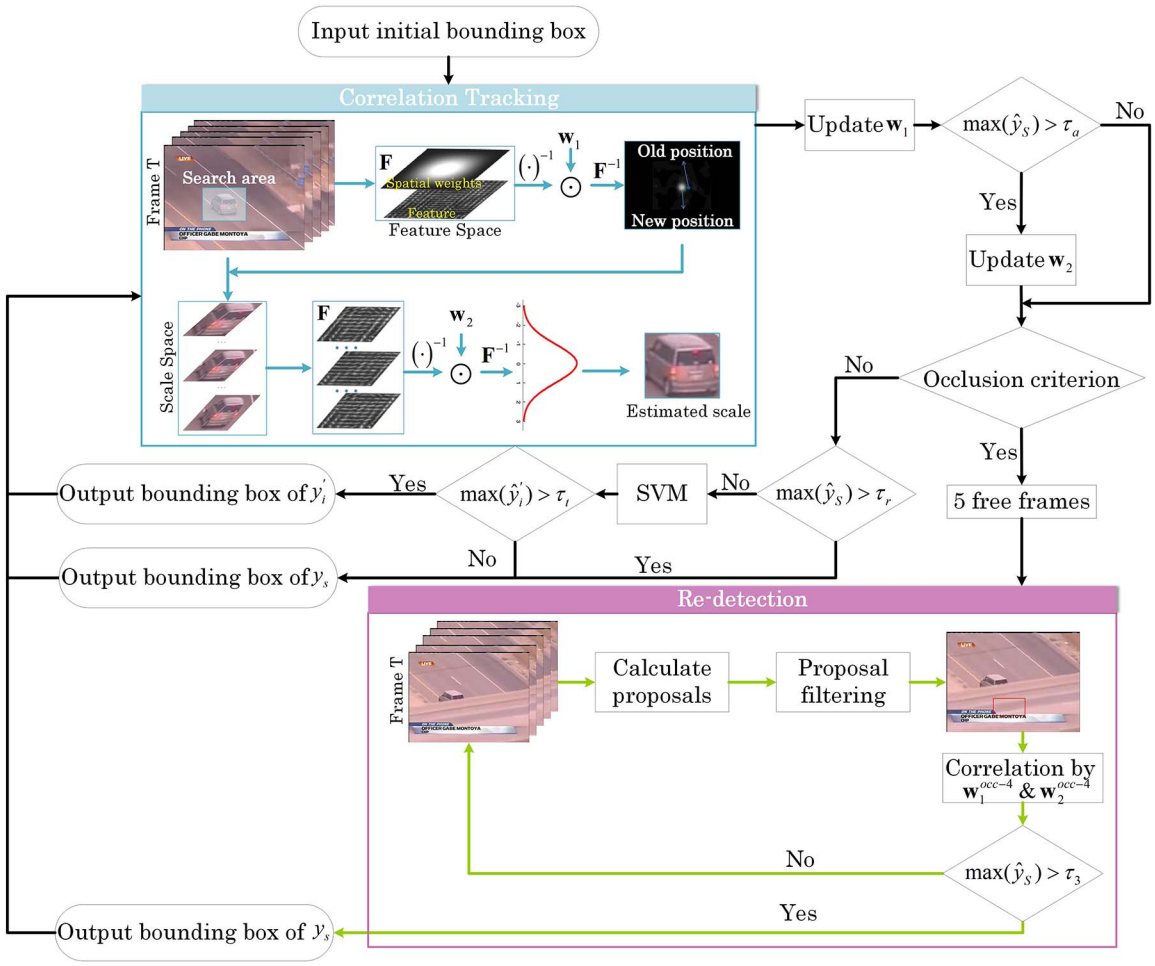

Fig. 4. Flowchart of ILCT. (Best view in PDF.)

Table 1. Information of Eight Sequences

Video Name	Number of Frames	Attributes
Carchase1[25]	71	OCC, FM, MC
Road[26]	52	OCC, BC, FM, MC, SV
Carchase2[27]	150 (1st-150th)	OCC, FM, IV, MC
Group[26]	86	OCC, DEF, MC
Motorcycle[28]	156	OCC, FM, MB, MC, SV
Triumphal arch	331	OCC, DEF, OPR, SV
Jogging[17]	307	OCC, DEF, OPR
Wandering[26]	285	OCC, DEF, MC

Through Fig. 5, we can see that these objects undergo obvious full OCC. Most trackers drift to background after OCC, while ILCT tracks objects robustly.

Center location error (CLE) is used for quantitative evaluation, which is defined as the percentage of frames whose Euclidean distance r between the centers of the bounding box and ground truth is within a pixel threshold (we set 15 pixels). In addition, for fair comparison, the frames with OCC are discarded, i.e., only the frames with objects in view are compared.

The results of CLE are listed in Table 2 (the best result is in bold, and the second best one is underlined). ILCT has shown its capacity of resisting disturbance in Fig. 5. Table 2 also indicates this. Because ILCT does not loose objects in all experiments, its CLE results are satisfactory. In terms of the number of the best and second best, ILCT achieves four and three times, respectively. From the evaluation results of CLE, we can consider ILCT as the best tracker.

From Table 2, compared to KCF, we can see that IKCF achieves great progress in tracking results because of the improvement, which reflects the effectiveness of our method. The proposed trigger outperforms PSR and MF because they are calculated in one frame, which may be triggered by non-OCC factors such as DEF and SV.

In conclusion, an effective tracking method that can handle OCC is proposed. Based on the motion and appearance correlation filters of LCT, ILCT employs a designed OCC criterion and a re-detector to judge the OCC and

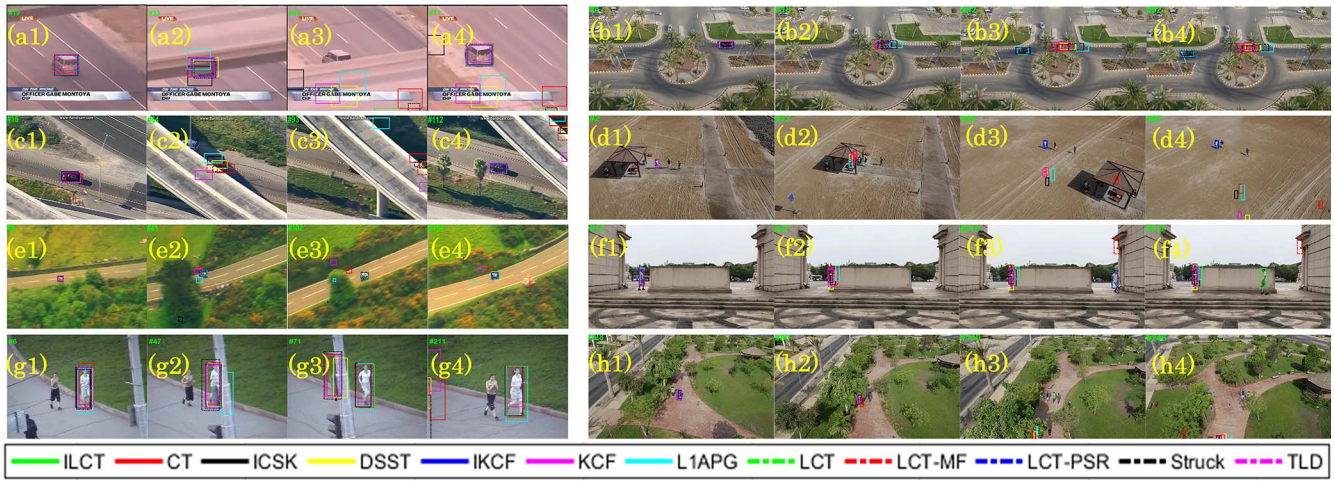

Fig. 5. Tracking results of 12 trackers. (a) Carchase1; (b) road; (c) carchase2; (d) group; (e) motorcycle; (f) triumphal arch; (g) jogging; (h) wandering. (Best view in PDF.)

Table 2. Comparison Results of CLE in Eight Sequences

Sequence	ILCT	CT	DSST	KCF	IKCF	L1APG	Struck	TLD	ICSK	LCT	LCT-PSR	LCT-MF
Carchase1	0.9800	0.5800	0.5800	0.5800	0.9800	0.5800	0.5800	**1.0000**	0.5800	0.5800	**1.0000**	0.8400
Road	0.8621	0.4828	0.4828	0.4828	**1.0000**	0.3793	0.4828	0.4828	0.4828	0.4828	**1.0000**	0.4828
Carchase2	0.9681	0.0638	0.5106	0.5106	**1.0000**	0.5106	0.4787	0.9681	0.5426	0.5106	**1.0000**	0.5426
Group	**0.9444**	0.0833	0.1528	0.1528	**0.9444**	0.1528	0.1389	0.0417	0.1528	0.1528	**0.9444**	0.1667
Motorcycle	0.9858	0.1560	0.9929	0.2128	0.9362	0.2128	0.0780	0.3404	**1.0000**	**1.0000**	0.8936	0.1277
Triumphal arch	**0.9622**	0.1351	0.1351	0.1351	0.1351	0.1351	0.1351	0.4757	0.1351	0.1351	0.5297	0.1351
Jogging	**1.0000**	0.1544	0.1544	0.1544	0.1544	0.9930	0.0702	**1.0000**	**1.0000**	**1.0000**	0.1544	0.6947
Wandering	**0.9956**	0.3231	0.3406	0.3406	0.3406	0.3275	0.3406	0.8908	0.3406	0.3362	0.1659	0.3406
Average	**0.9575**	0.2365	0.4298	0.3183	0.6863	0.4234	0.2805	0.6155	0.5292	0.5516	0.7110	0.4163
Number of best	**4**	0	0	0	3	0	0	2	2	2	**4**	0
Number of second best	**3**	0	1	0	1	1	0	2	0	0	1	1

perform re-detection, respectively. Once the object is identified as occluded, the tracker stops, and the re-detector is activated. Then, the detection result with high confidence will re-initialize the tracker. Extensive experiments have been performed, and the results of qualitative and quantitative evaluation indicate that ILCT outperforms some state-of-the-art trackers in terms of accuracy and robustness. In future work, the efficiency and real-time performance have to be addressed to make the tracker perfect.

This work was supported by the National Program on Key Basic Research Project (No. 2014CB744903), Shanghai Pujiang Program (No. 16PJD028), Shanghai Industrial Strengthening Project (No. GYQJ-2017-5-08), Shanghai Science and Technology Committee Research Project (No. 17DZ1204304), and Shanghai Engineering Research Center of Civil Aircraft Flight Testing.

References

1. G. Wang, F. Xing, M. S. Wei, T. Sun, and Z. You, Chin. Opt. Lett. **15**, 081201 (2017).
2. F. Z. Zhang, Q. S. Guo, Y. Zhang, Y. Yao, P. Zhou, D. Y. Zhu, and S. L. Pan, Chin. Opt. Lett. **15**, 112801 (2017).
3. Q. H. Yu, D. M. Wu, F. C. Chen, and S. L. Sun, Chin. Opt. Lett. **16**, 071101 (2018).
4. C. Ma, X. Yang, C. Zhang, and M. H. Yang, in *Proceedings of Computer Vision and Pattern Recognition* (2015), p. 5388.
5. T. Zhang, B. Ghanem, S. Liu, and N. Ahuja, in *Proceedings of Computer Vision and Pattern Recognition* (2012), p. 2042.
6. D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, in *Proceedings of Computer Vision and Pattern Recognition* (2010), p. 2544.
7. C. F. Hester and D. Casasent, Appl. Opt. **19**, 1758–1761 (1980).
8. W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C* (Cambridge University, 1988).
9. J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, in *Proceedings of European Conference on Computer Vision* (2012), p. 702.

10. J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, IEEE Trans. Pattern Anal. Mach. Intell. **37**, 583–596 (2015).
11. M. Danelljan, G. Häger, F. Khan, and M. Felsberg, in *Proceedings of British Machine Vision Conference* (2014), p. 1.
12. J. Choi, H. J. Chang, J. Jeong, Y. Demiris, and J. Y. Choi, in *Proceedings of Computer Vision and Pattern Recognition* (2016), p. 4321.
13. C. Ma, J. B. Huang, X. Yang, and M. H. Yang, Int. J. Comput. Vision **126**, 771 (2018).
14. P. Dollár, R. Appel, S. Belongie, and P. Perona, IEEE Trans. Pattern Anal. Mach. Intell. **36**, 1532 (2014).
15. R. Zabih and J. Woodfill, in *Proceedings of European Conference on Computer Vision* (1994), p. 151.
16. C. Ma, X. Yang, C. Zhang, and M. H. Yang, "Long-term correlation tracking," 2016, https://sites.google.com/site/chaoma99/cf-lstm.
17. Y. Wu, J. Lim, and M. H. Yang, in *Proceedings of Computer Vision and Pattern Recognition* (2013), p. 2411.
18. C. L. Zitnick and P. Dollár, in *Proceedings of European Conference on Computer Vision* (2014), p. 391.
19. P. Dollár and C. L. Zitnick, IEEE Trans. Pattern Anal. Mach. Intell. **37**, 1558 (2015).
20. Z. Kalal, K. Mikolajczyk, and J. Matas, IEEE Trans. Pattern Anal. Mach. Intell. **34**, 1409 (2012).
21. S. Hare, A. Saffari, and P. H. S. Torr, in *Proceedings of International Conference on Computer Vision* (2011), p. 263.
22. C. Bao, Y. Wu, H. Ling, and H. Ji, in *Proceedings of Computer Vision and Pattern Recognition* (2012), p. 1830.
23. X. Dong, J. Shen, D. Yu, W. Wang, J. Liu, and H. Huang, IEEE Trans. Multimedia **19**, 763 (2017).
24. K. Zhang, L. Zhang, and M. H. Yang, in *Proceedings of European Conference on Computer Vision* (2012), p. 864.
25. P. Liang, E. Blasch, and H. Ling, IEEE Trans. Image Process. **24**, 5630 (2015).
26. M. Mueller, N. Smith, and B. Ghanem, in *Proceedings of European Conference on Computer Vision* (2016), p. 445.
27. O. Quaker, "Anaheim California Police Chase 06/18/2017—Reckless Driver," 2017, https://www.youtube.com/watch?v=bAZZ3NKNnTg.
28. M. Kristan, J. Matas, A. Leonardis, T. Vojíř, R. Pflugfelder, G. Fernandez, G. Nebehay, F. Porikli, and L. Čehovin, IEEE Trans. Pattern Anal. Mach. Intell. **38**, 2137 (2016).