DEEP LEARNING BASED VISUAL OBJECT TRACKER WITH TEMPLATE UPDATE

Zhen SUN[1], Qingdang LI[2], Lu WANG[3], Junfei WU*[4]

In recent years, deep learning based discriminative visual object trackers have gained considerable attention in the field of object tracking. The reliability and distinguishability of the target template features are the key factors that determine the performance of the tracker. Currently, many existing trackers use fixed feature templates to improve the speed on the one hand and avoid the template contamination on the other. However, a tracker using fixed templates is unable to adapt to situations such as significant appearance change, rotation, and occlusion. In order to solve this problem, this paper proposes a template quality evaluation method based on correlation filtering and implements a new siamese network based tracker with template update. The experiments demonstrate that the proposed tracker has better performance than the original tracker using fixed templates and improves the success rate of 7% in categories 'out-plane rotation' and 'occlusions' on OTB-100 benchmark. The proposed template update method can also be used in other discriminative target trackers.

Keywords: Visual tracking, online update, siamese network, correlation filtering

1. Introduction

Visual object tracking is a basic technology in the field of computer vision. It has a wide range of applications in the real world, such as intelligent monitoring [1], robot navigation [2], driverless cars [3], and crowd behavior recognition [4]. In recent years, a variety of high-performance visual object tracking methods have been proposed [5-10] that have achieved excellent results in various public benchmark sets. However, the effect of target tracking is not stable in some complex scenarios, such as non-rigid deformation of the target, messy target background, and similar objects. Thus, designing a robust, stable, and fast target tracker that can adapt multiple scenarios remains a challenging task.

Recently, several siamese network based trackers have been proposed. Siamese network based tracker translates the object tracking problem into a

[1] Prof., College of Electromechanical Engineering, Qingdao University of Science and Technology, China, e-mail: sunzhen@qust.edu.cn
[2] Prof., Chinesisch-Deutsche Technische Fakultät, Qingdao University of Science and Technology, China, e-mail: lqd@qust.edu.cn
[3] Prof., Chinesisch-Deutsche Fakultät für Ingenieurwissenschaften, Qingdao University of Science and Technology, China, e-mail: wanglu@qust.edu.cn
[4] Prof., College of Electromechanical Engineering, Qingdao University of Science and Technology, China, e-mail: jfw_2002@sina.com

matching problem between the target template patch and the candidate patch. The siamese network uses a large number of positive and negative sample image pairs for end-to-end training of the network parameters. Numerous studies have shown that the siamese network can adapt to changes in various scenarios and effectively track the target. However, in order to improve the tracker speed, majority of existing tracking algorithms based on the siamese network do not support online update. This approach has two significant drawbacks:

(1) The model of the target is incomplete. The original siamese network based tracker uses only the target patch from the ground truth in the initial frame as the target template. This has the advantage of avoiding the degradation of the target template quality due to the subsequent tracking drift. It means that the information of the target template is always dependable. However, at the same time, the expression of the target model is incomplete because the ground truth of the initial frame is always used as the target template. When the target has a significant appearance change (such as angle or color change), the information in the target template cannot correctly represent the true appearance information of the target, which may lead to tracking failure.

(2) The performance of the tracker depends on the stability of the target. Since the target information is incompletely expressed, the tracker is sensitive to the stability of the target appearance. The tracker performance is better when the target appearance does not change significantly. However, if the appearance of the target changes, there will be a significant difference between the target template and the candidate patch. This difference causes the peak of the response map to be inconspicuous, which gradually leads to tracking failure. The situation becomes more serious when similar objects appear near the target. The peak corresponding to the target in the response map becomes lower due to the change in the appearance of the target, while the peak corresponding to the interferer becomes higher. This will cause the bounding box to drift to the interferer and cause tracking failure.

In order to solve the above-mentioned problems, this paper focuses on the update method of target templates in the siamese network based tracker. This paper proposes a target patch dependability evaluation algorithm and an online update supported siamese network framework. The main contributions of this paper include:

(1) A novel target template dependability evaluation method is proposed. In this paper, the target patch dependability evaluation algorithms are investigated and a new evaluation algorithm based on correlation filter is proposed. The novel algorithm is used as the basis for online update of target template.

(2) A siamese network based tracker that supports online update is proposed. Based on the proposed dependability evaluation algorithm, online evaluation of the target patch for each frame is performed. The dependability score is used as

the basis for whether the target patch participates in the template update. If the target encounters occlusion, blurring, or loss, the dependability score will be reduced, preventing the low-quality target images from participating in the template update.

The remainder of this paper is organized as follows: Section 2 reviews the related research on the update method of siamese network based tracker. Section 3 studies target dependability evaluation methods. A siamese network based tracker that can update templates online is proposed in Section 4. Section 5 validates the effectiveness of the proposed method through experiments followed by conclusions in Section 6.

2. Related work

In recent years, siamese network based trackers have received widespread attention from researchers. Several excellent trackers [11-13] have achieved good results on various public benchmark sets. However, most of these siamese network based trackers [8,14-16] use a fixed target template from the initial frame. Using a fixed target image template can improve the tracking speed and prevent the subsequent tracking results from affecting the quality of the target template. However, the fixed target template does not adapt well to the tracking scenarios in which the appearance of the target changes significantly. In order to address this problem, several siamese network based trackers that support target template updates have been proposed. K. Gong [17] proposed a tracking method based on target feature fusion that utilized the new target features appearing in the tracking process. However, this tracker did not determine the quality of the new features. If the new target features are flawed, it will cause the target template to degenerate. Q. Guo [18] added two filters to the two branches of the siamese network and dynamically trained and updated the filters according to the target during the tracking process in order to adapt to the change of the target appearance. L. Xiaoping [19] proposed a semi-online domain adaptation method that found a balance between speed and accuracy. It can be observed that there are only few siamese network based trackers that support online updates. Designing a simple, fast, and effective template update method is important for improving the performance of the siamese network based tracker. Therefore, this paper investigates the evaluation method of target patch dependability, and proposes a novel siamese network based tracker that supports online update.

3. Dependability evaluation methods

3.1 Dependability evaluation method based on response map

The tracker based on siamese network estimates the target location by matching the target template with the candidate patch. The tracker's matching

response map will change when the target patch encounters occlusion or blurring. Therefore, the dependability of the target patch is evaluated based on the changes in the response map. The simple structure of a classic siamese network based tracker is shown in Fig.1, and the calculation of the response map is provided in Eq. 1.

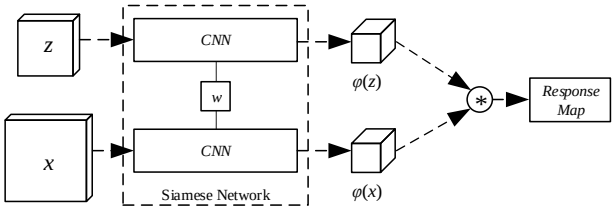


Fig. 1 Simple structure of siamese network based tracker.

$$h(z,x) = \varphi(z) * \varphi(x) \#(1)$$

Where $\varphi(z)$ represents the feature map of the target template extracted by the convolutional neural network (CNN), and $\varphi(x)$ represents the feature map corresponding to the candidate region. $*$ represents the correlation operation, and $h(z,x)$ represents the response map. The change of the response map is analyzed to determine whether the new target patch has blurred or occluded, and provide a basis for the target template update. For testing, 'Blurface', 'BlurCar2', 'David3', and 'Coke' sequences are selected from the OTB-100 benchmark set. Among them, 'Blurface', and 'BlurCar2' both have situations where the target is blurred, the target in 'David3' is partially occluded, and 'Coke' has a situation where the target is completely occluded. Fig. 2 shows the resultant response maps of the tracker when the target is in normal, blurred, partially occluded, and completely occluded conditions.

The experiments demonstrate that the quality of the response map decreases when the target is occluded. There are multiple peaks in the response map, and the main peak is not significant. In this case, the target template of the current frame is undependable and cannot be used to update the target template. However, when the target is blurred, the response map quality is still good. This shows that the siamese network based tracker has better robustness when the target is blurred. At this time, if the original target template is updated with a new target patch based on the quality of the response map, the quality of the target template will be reduced. Therefore, the dependability of the target patch cannot be determined based on the response map because it cannot find the target blur.

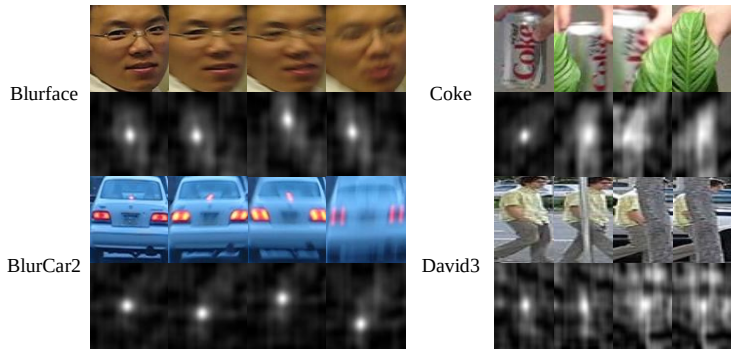


Fig. 2 Siamese network response maps under different target states.

3.2 Dependability evaluation method based on correlation filter

The correlation filter was first introduced into the field of visual tracking by MOSSE [20]. A trained correlation filter can determine the similarity between an image and the original target. The correlation filter can be calculated in the Fourier frequency domain and has high computational speed. In this paper, the correlation filters are used for the calculation of target image dependability. First, the target image of the initial frame is represented as i and its 2D Fourier transform is: $I = \mathcal{F}(i)$. Then the filter h is also converted to Fourier domain: $H = \mathcal{F}(h)$ and G represents the response of correlation filtering.

$$G = I \odot H^* \#(2)$$

Since the correlation operation can be converted to element-wise multiplication, $\odot$ represents this operation and $*$ denotes the complex conjugate. G is the desired Gaussian response map. From the derivation in MOSSE [20], the expression of H^* can be expressed as:

$$H^* = \frac{\sum_i G_i \odot I_i^*}{\sum_i I_i \odot I_i^*} \#(3)$$

During the tracking process, the filter H^* is continuously updated with the new target image patch. H_t^* is used to represent the correlation filter trained using the *t* frame target image patch and the final filter H^* is updated by Eq. 4.

$$H^* = (1-\alpha)H_{t-1}^* + \alpha H_t^* = (1-\alpha)H_{t-1}^* + \alpha \frac{G_t \odot I_t^*}{I_t \odot I_t^*} \#(4)$$

In the *t*+1 frame, the dependability of the target image is determined using the filter response G_{t+1} calculated by $I_{t+1} \odot H^*$. When the target images in *t* and *t*+1 frames are similar, the filtering response G_{t+1} should be close to the Gaussian response and have a more significant peak at the center. If there is a significant difference between the target images in *t* and *t*+1 frames (such as blur, and occlusion), the quality of G_{t+1} will decrease and the peak will be weakened. Thus, the dependability of the target image can be determined by analyzing the response of G_{t+1}. If the target is dependable, the correlation filter H_{t+1} is trained by the

new target image and updated into the filter H^*. Else if the target is not dependable, the correlation filter H^* is kept unchanged. The four sequences of 'Blurface', 'BlurCar2', 'David3', and 'Coke' are tested. The experimental results are shown in Fig. 3.

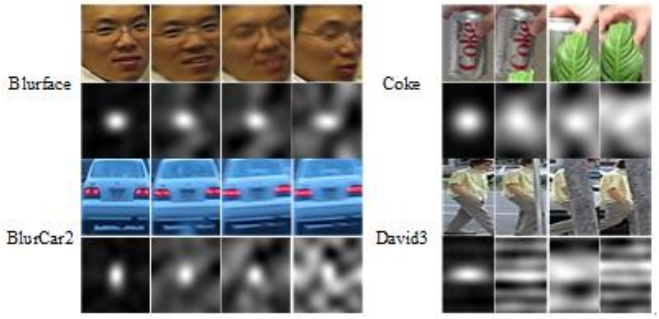


Fig. 3. Responses of correlation filtering under different target states.

The experimental results demonstrate that the response map of the correlation filter can sensitively reflect the blurring and occlusion of the target image. The response map is close to the Gaussian response when the target is dependable. However, the response map appears to be significantly degraded if the target is blurred or occluded. For quantitative analysis, the peak sidelobe ratio (PSR) is used to measure the response map. The PSR is calculated as:

$$psr = \frac{p_{x_0,y_0} - \mu_s}{\sigma_s} \#(5)$$

where x_0, y_0 represent the coordinates of the peak and p_{x_0,y_0} represents the value of the peak. The radius of the peak is represented by r and the pixels outside the radius are called sidelobes. μ_s and σ_s represent the mean and the standard deviation of the pixels in the sidelobes, respectively. In this paper, the PSR of the response map is used as the dependability of the target image. If the PSR of the target image patch is greater than the threshold η, the image is considered to be dependable and is used to train a new correlation filter and update H^*. The experiments demonstrate that the dependability drops significantly when the target is occluded or blurred. This provides a good basis for online updating of target templates in complex scenarios.

4. Siamese network with template update

With the dependability of one target image patch, the tracker template can be updated during the tracking process. The most recent and most dependable target image is selected to update the target template to ensure that the tracking template can better match the features of the current target. At the same time, in order to ensure the robustness of the target template features, this paper combines multiple target template features to form the final target feature set. The structure

of the proposed tracker is shown in Fig. 4. The proposed tracker is called SiamUpdate. The left side of the structure diagram in Fig. 4 is the classic siamese network structure.

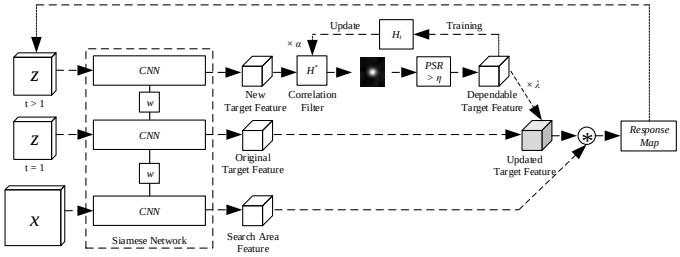


Fig. 4. Structure of the proposed siamese network based tracker with online update.

The convolutional neural network extracts convolution features for the target image patch and the search region image patch. On the right side of the structure diagram, the correlation calculations are performed using the updated target features and the features of the search area to obtain a response map. The highest response point in the response map is the location of the target in the current frame. At the initial frame (t = 1), the target features are generated from the target image patch in the ground truth. During the tracking process (t > 1), the target features are generated by the fusion of features of all dependable target image patches in the historical frame. It should be noted that each dependable target image patch will be used to train and update the correlation filter H^*. The correlation filter is calculated in the Fourier frequency domain, which is fast and has miniature effect on the speed of the tracker. The experiments demonstrate that the proposed tracking method can effectively handle the change of the target appearance and improves the robustness of the tracker.

5. Experiment

5.1 Implementation details

In order to validate the effectiveness of the proposed template update method, the SiamUpdate is compared with existing methods on OTB-100 benchmark. During the experiments, a simple Alexnet [21] was used as the backbone for all trackers. The settings of the hyperparameters in the tracker are shown in Table 1. The experiments were run on a PC using Ubuntu 16.04 system, MATLAB 2017b, Matconvnet framework, and accelerated with GeForce 1080Ti.

Table1

Hyperparameters settings

Category	Parameters	Value
Correlation filter update rate	α	0.1
Dependability threshold in PSR	η	5
Peak radius in PSR	r	30
Target feature update rate	λ	0.01

5.2. Qualitative comparison

In this section, in order to highlight the effectiveness of the template update algorithm, the SiamUpdate tracker is compared with classic tracking methods including SiamFC[14], FOT[22], Struck[23], KCF[7], TLD[24], DFT[25], CSK[26], LOT[27] and CXT[28]. All the trackers are tested on OTB-100 [29] benchmark. The tracking results of some of the image sequences are shown in Fig. 5. The experimental results show that the tracking effect of SiamUpdate is better and more stable than the other methods. The SiamUpdate is able to track targets correctly and consistently in the face of complex changes in the appearance of the target. In order to more clearly illustrate the effectiveness of the proposed method, the performance of each tracker is further compared through quantitative indicators.

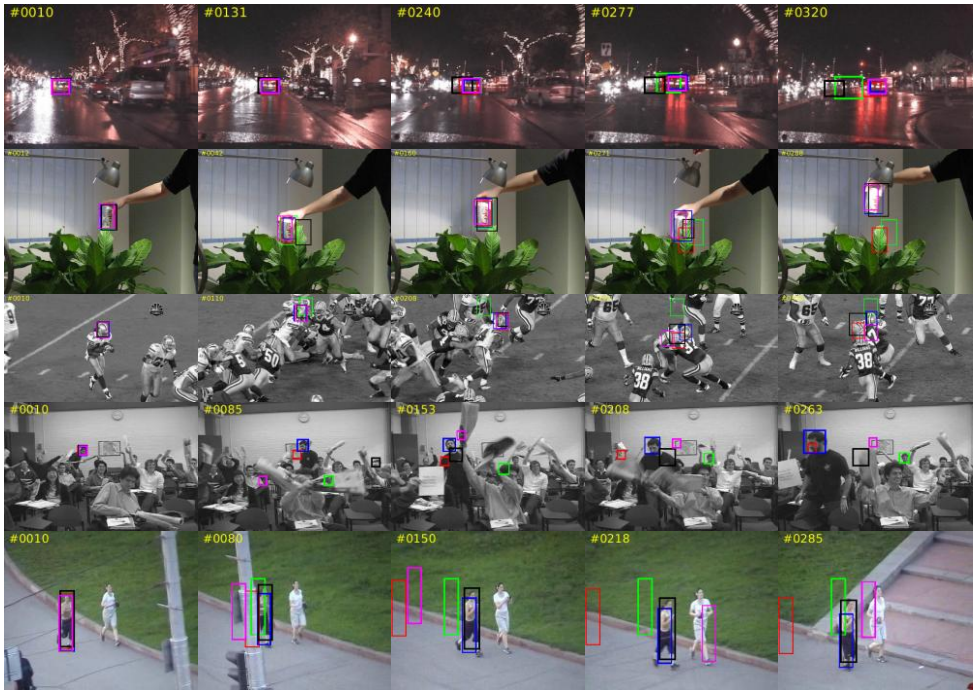
5.3 Quantitative comparison

In order to quantify the performance of the tracker, the precision and the success plots in the OTB are used for comparison. The definitions of the precision and the success plots are as follows.

Precision plot: This metric is used to evaluate the accuracy of the tracker's estimate of the target center position. The center deviation $C = c_t - c_0$, where c_t and c_0 are the predicted and the ground-truth centers, respectively. The precision plot is a change of the percentage of the image frame whose center deviation C is less than the threshold t_{pp} to the total image frame, as the threshold t_{pp} changes.

Success plot: This metric is used to indicate the degree of overlap between the predicted and the ground-truth bounding boxes r_t and r_0, respectively. The overlap score is defined as $S = \frac{|r_t \cap r_0|}{|r_t \cup r_0|}$ and the success rate is the count of frames with overlap score S greater than the threshold t_{sp} divided by the number of total image frames. The success plot is the change of success rate as a function of threshold t_{sp}.

The precision and the success plots of each tracker in all sequences are shown in Fig. 6. The success plots of each tracker in different sub- categories are shown in Fig. 7. In order to more clearly compare the performance of each tracker, the average success scores of all trackers are shown in Table 2. The first column is the score of each tracker in all sequences, and the other columns are the scores of each sub-category. The meaning of each attribute can be found in OTB-100 benchmark. The highest score of each column is indicated by the black italic font.



SiamUpdate SiamFC KCF Struck TLD

Fig. 5. Qualitative comparison of five different trackers.

The experimental results demonstrate that the SiamUpdate is more robust than the original SiamFC and other classic trackers. Especially, when the target has a significant change in the appearance, the SiamUpdate can update the target template in time to improve the tracking performance. By determining the dependability of the target template, the SiamUpdate can avoid template degradation caused by blind update of the template. The template update method proposed in this paper can also be applied to other trackers to enhance the robustness and performance in complex scenarios.

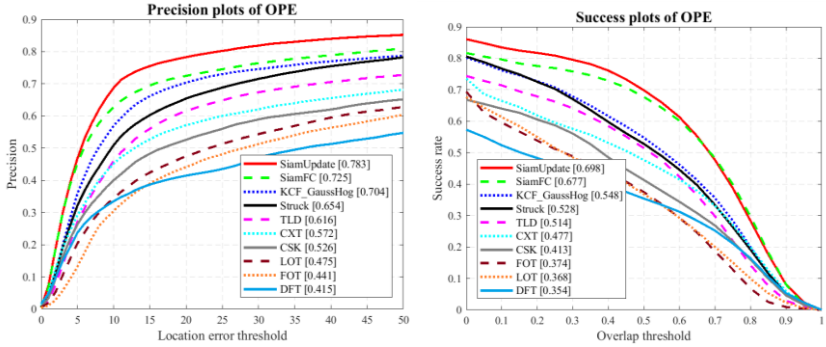


Fig. 6. Success and precision plots of each tracker in all sequences.

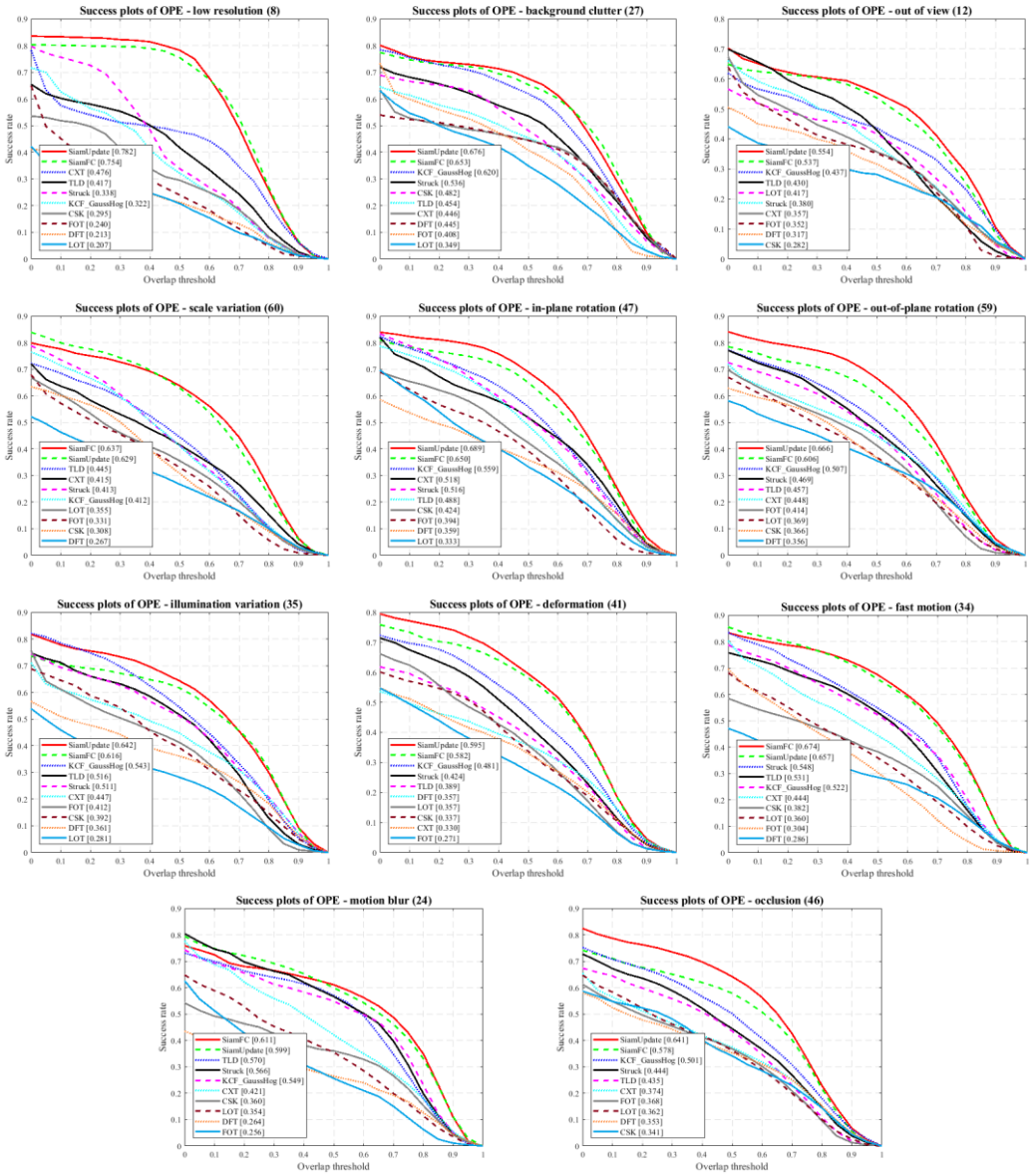


Fig. 7. Success plots of each tracker in different sub-categories.

Table2

Average success scores of different trackers

Tracker	All	LR	BC	OV	SV	IPR	OP	IV	DE	FM	MB	OC
SiamUpdate	***0.70***	***0.78***	***0.68***	***0.55***	0.63	***0.69***	***0.67***	***0.64***	***0.60***	0.66	0.60	***0.64***
SiamFC	0.68	0.75	0.65	0.54	***0.64***	0.65	0.60	0.62	0.58	***0.67***	***0.61***	0.57
KCF	0.55	0.32	0.62	0.44	0.41	0.56	0.51	0.54	0.48	0.52	0.55	0.50
Struck	0.53	0.34	0.54	0.38	0.41	0.52	0.47	0.51	0.42	0.55	0.57	0.44
TLD	0.51	0.42	0.45	0.43	0.45	0.49	0.46	0.52	0.39	0.53	0.57	0.44
CXT	0.48	0.48	0.45	0.36	0.42	0.52	0.45	0.45	0.33	0.44	0.42	0.37
CSK	0.41	0.30	0.48	0.28	0.31	0.42	0.37	0.39	0.34	0.38	0.36	0.34

FOT	0.37	0.24	0.41	0.35	0.33	0.39	0.41	0.41	0.27	0.30	0.26	0.37
LOT	0.37	0.21	0.35	0.42	0.36	0.33	0.37	0.28	0.36	0.36	0.35	0.36
DFT	0.35	0.21	0.45	0.32	0.27	0.36	0.36	0.36	0.36	0.29	0.26	0.35

6. Conclusion

This paper analyzes the shortcoming of the traditional siamese network based trackers that cannot update the templates. This shortcoming may cause the tracker to fail when the target changes its appearance. In order to solve this problem, this paper proposes a method for determining the dependability of target images based on correlation filtering and a tracker that supports online updating of target templates, called SiamUpdate. The experiments on the OTB-100 benchmark demonstrate and validate that the proposed SiamUpdate is more robust than the classical tracking methods when the target changes its appearance. When the target encounters blurring, or occlusion, the SiamUpdate can effectively identify and avoid template degradation. The proposed SiamUpdate has significant improvements over the traditional SiamFC tracker, especially in the ‘in-plane rotation’, ‘out-plane rotation’ and ‘occlusions’ categories. The target dependability and template update methods proposed in this paper can also be used on other trackers.

Acknowledgments

This research was financially supported by the Tai Shan Scholar Foundation (tshw201502042), Shandong Province Key Research and Development Plan (2017CXGC0607, 2017GGX30145), National Natural Science Foundation of China (61702295).

R E F E R E N C E S

[1] *K. Aziz, S. Tarapiah, S. H. Ismail and S. Atalla*, "Smart real-time healthcare monitoring and tracking system using gsm/gps technologies", pp.1-7, 2016.

[2] *Y. Kunii, G. Kovacs and N. Hoshi*, "Mobile robot navigation in natural environments using robust object tracking", pp.1747-1752, 2017.

[3] *A. Muraleedharan, H. Okuda and T. Suzuki*, "Path tracking control using model predictive control with on gpu implementation for autonomous driving", Journal of Arid Land Studies, **vol**.**28**, no.S, pp.163-167, 2018.

[4] *H. Yao, A. Cavallaro, T. Bouwmans and Z. Zhang*, "Guest editorial introduction to the special issue on group and crowd behavior analysis for intelligent multicamera video surveillance", IEEE Transactions on Circuits and Systems for Video Technology, **vol**.**27**, no.3, pp.405-408, 2017.

[5] *M. Danelljan, A. Robinson, F. S. Khan and M. Felsberg*, "Beyond correlation filters: learning continuous convolution operators for visual tracking", Proc. European Conference on Computer Vision, pp.472-488, 2016.

[6] *M. Danelljan, G. Bhat, F. S. Khan and M. Felsberg*, "Eco: efficient convolution operators for tracking", Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.3, 2017.

[7] *J. F. Henriques, C. Rui, P. Martins and J. Batista*, "High-speed tracking with kernelized correlation filters", IEEE Transactions on Pattern Analysis and Machine Intelligence, **vol**.**3**, no.37, pp.583-596, 2014.

[8] *J. Valmadre, L. Bertinetto, J. O. F. Henriques, A. Vedaldi and P. H. S. Torr*, "End-to-end representation learning for correlation filter based tracking", Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.2805-2813, 2017.

[9] *D. Held, S. Thrun and S. Savarese*, "Learning to track at 100 fps with deep regression networks", Proc. European Conference on Computer Vision, pp.749-765, 2016.

[10] *H. Nam and B. Han*, "Learning multi-domain convolutional neural networks for visual tracking", Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.4293-4302, 2016.

[11] *B. Li, J. Yan, W. Wu, Z. Zhu and X. Hu*, "High performance visual tracking with siamese region proposal network", Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.8971-8980, 2018.

[12] *Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan and W. Hu*, "Distractor-aware siamese networks for visual object tracking", Proc. European Conference on Computer Vision (ECCV), pp.101-117, 2018.

[13] *B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing and J. Yan*, "Siamrpn++: evolution of siamese visual tracking with very deep networks", arXiv preprint arXiv:1812.11703, 2018.

[14] *L. Bertinetto, J. Valmadre, J. O. F. Henriques, A. Vedaldi and P. H. S. Torr*, "Fully-convolutional siamese networks for object tracking", Proc. European conference on computer vision, pp.850-865, 2016.

[15] *S. Zhu, Z. Fang and F. Gao*, "Hierarchical convolutional features for end-to-end representation-based visual tracking", Machine Vision and Applications, **vol**.**29**, no.6, pp.955-963, 2018.

[16] *F. Mustansar, M. Arif and K. Soon*, "Deep siamese networks toward robust visual tracking", 2019.

[17] *K. Gong, Z. Cao, Y. Xiao and Z. Fang*, "Online update siamese network for unmanned surface vehicle tracking", Proc. 11th International Conference of Intelligent Robotics and Applications, Newcastle, pp.159-169, 2018-01-01 2018.

[18] *Q. Guo, W. Feng, C. Zhou, R. Huang, L. Wan and S. Wang*, "Learning dynamic siamese network for visual object tracking", pp.1781-1789, 2017.

[19] *L. Xiaoping and L. Wenbing*, "Fast deep tracking via semi-online domain adaptation", Journal of Physics: Conference Series, **vol**.**1004**, no.1, pp.12013, 2018.

[20] *D. S. Bolme, J. R. Beveridge, B. A. Draper and Y. M. Lui*, "Visual object tracking using adaptive correlation filters", Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.2544-2550, 2010.

[21] *A. Krizhevsky, I. Sutskever and G. E. Hinton*, "Imagenet classification with deep convolutional neural networks", Proc. International Conference on Neural Information Processing Systems, pp.1097-1105, 2012.

[22] *T. Vojir and J. B. P. Matas*, "The enhanced flock of trackers",in Registration and Recognition in Images and Videos,eds. R. Cipolla, S. Battiato and G. M. Farinella, pp.113-136, Springer Berlin Heidelberg, Springer Berlin Heidelberg, 2014.

[23] *S. Hare, S. Golodetz, A. Saffari, V. Vineet, M. Cheng, S. L. Hicks 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.

[24] *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, 2011.

[25] *L. Sevilla-Lara and E. Learned-Miller*, "Distribution fields for tracking", Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1910-1917, 2012.

[26] *J. O. F. Henriques, R. Caseiro, P. Martins and J. Batista*, "Exploiting the circulant structure of tracking-by-detection with kernels", Proc. European Conference on Computer Vision (ECCV), Berlin, Heidelberg, pp.702-715, 2012.

[27] *S. Oron, A. Bar-Hillel, D. Levi and S. Avidan*, "Locally orderless tracking", International Journal of Computer Vision, **vol**.**111**, no.2, pp.213-228, 2015.

[28] *T. B. Dinh, N. Vo and G. E. R. Medioni*, "Context tracker: exploring supporters and distracters in unconstrained environments", Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1177-1184, 2011.

[29] *Y. Wu, J. Lim and M. H. Yang*, "Object tracking benchmark", IEEE Trans Pattern Anal Mach Intell, **vol**.**37**, no.9, pp.1834-1848, 2015.