LONG-TERM OBJECT TRACKING BASED ON SIAMESE NETWORK

Kaiheng Dai[†], *Yuehuan Wang*[†‡], *Xiaoyun Yan*[†]

[†]The School of Automation, Huazhong University of Science and Technology, Wuhan, China
[‡]National Key Lab of Science and Technology on Multi-spectral Information Processing, Wuhan, China

ABSTRACT

Although the siamese network trackers achieve competitive results both on robustness and accuracy, there is still a need to improve the overall tracking capability. In this paper, we proposed a long-term tracker based on the siamese network. We address the problem of long-term tracking where the target objects undergo significant appearance variation due to heavy deformation, occlusion, abrupt motion, and out-of-view. To tackle those problems, we suggest a multi-template fusion tracking scheme. Moreover, patch template update scheme based on optical flow are proposed to boost the overall tracking performance. The extensive results on object tracking benchmark (OTB2013) show that the proposed algorithm achieve much better performance.

Index Terms— Visual Tracking, Siamese network, Multi-Template Fusion, Template Update

1. INTRODUCTION

Visual tracking is a fundamental research problem in computer vision for its various applications, such as video surveillance, robotics, driverless vehicle and so on. Although great progress has been made in the past years, there is still a tough problem due to illumination changes, geometric deformations, partial occlusions, fast motions, and background clutters.

Recently, deep convolution network is developed as a powerful tool in computer vision and introduced into visual tracking too. FCN[1], Ma[2] found that convolutional layers in different levels characterize the target from different perspectives, and jointly top layer and lower layer encodes as representation features to track; DeepSRDCF[3] are using the internal representation of a pre-trained convolution network as feature, then input these features to correlation filter tracking framework. While those trackers achieve strong results, they still not reach real-time tracking due to high dimension of conv-net features computing. There is also another way that convolutional network can be used in object tracking, SO-DLT[4], MDNet[5] learn a detector for each target to track with examples extracted from the video itself as in the conventional tracking-by-detection paradigm. This kind of trackers cannot achieve real-time tracking because of heavy computing in forward and backward passes on many examples. However, some siamese network trackers achieving real-time tracking have been proposed in the last year, GOTURN[6] train a conv-net to regress directly from two images to the location in the second image of the object shown in the first image; SINF[7] train a siamese network to identify candidate image locations that match the initial object appearance; SiamfC[8] using a fully-convolutional architecture of siamese network to compute response value of each location in searching region, and the highest response is the target. Those trackers not only achieve real-time tracking with robust and accuracy results in the tracking benchmarks but also are no model update, no occlusion detection, no combination tracker and no geometric matching. However, there is still a need to improve the overall tracking capability.

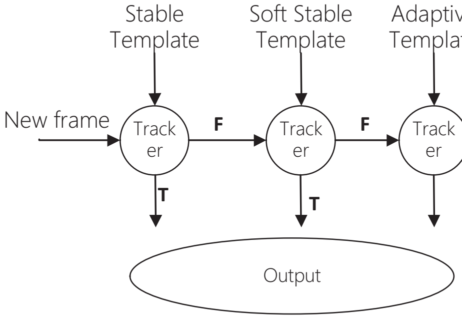

Fig. 1. The cascade architecture of fusion tracking with three templates. A stable template without update, a soft stable template with a thin update, and an adaptive template with the marked update are fused together with cascade architecture to make tracker achieve the effectiveness and efficient.

Our work is inspired by the work in SiamFC[8]. Here, we want to emphasize the differences between our work and [8].

1) there is a stable template using in SiamFC, which may make the tracker fail to track the objects when the appearance change drastically as a result of deformation, posing, abrupt moving, and even surrounding illumination change in actual environments. Thus, we design a multi-template fusion scheme to achieve the adaptiveness of object's appearance changes.

2) we employ small size convolution filter to get more discriminative target representation.

In this paper, our major contribution is to enrich siamese network tracker with multi-template fusion method. The starting point is that no model update trackers cannot handle complicated appearance changes in long-term tracking where the target objects can meet illumination changes, fast motions, occlusions, and so on. Hence, we introduce a stable template without update, a soft template with a thin update, and an adaptive template with a large update; then fuse the tracking results of those templates together with a simple but effective principle(Fig 1) to make the proposed tracker achieve the balance between accuracy and real-time tracking.

The second contribution of this paper is to propose a patch template update method based on optical flow. Convention template update methods are almost global update methods, whereas, in reality, appearance changes may happen in one part of the object. Some patches need to update, others may not. So global update usually pollute the template and result in template drifting. Therefore, we propose patch update scheme with optical flow to make the update more credible.

The organization of this paper is, we describe our method in Section 2, and discuss the experimental evaluation in Section 3, followed by conclusions in Section 4

2. THE PROPOSED METHOD

In this section, we firstly briefly review the siamese network tracker; and then introduce multi-template fusion tracking scheme used in our approach; besides, a patch template update method is presented to improve the robustness of tracker.

2.1. The Siamese Network Tracker

Our approach is built on SiamFc tracker, which achieves very impressive results. The key of SiamFc tracker is that it train a siamese network to address a more general similarity learning problem in an initial offline phase; Then using this network to evaluate online during tracking; Moreover, a fully-convolutional architecture with respect to the search image:dense are used, and efficient sliding-window evaluation is achieved with a bilinear layer that computes the cross-correlation of its two inputs, it can be defined as

$$f(z,x) = \varphi(z) * \varphi(x) + b \tag{1}$$

where x is the searching image, z is the initial tracking image, the transformation φ denotes siamese network which resembles the convolutional stage of the network of Krizhevsky[9], $*$ denotes the cross-correlation operation, b denotes a signal which takes value $b \in R$ in every location. The output is a scalar-valued score map whose dimension depends on the size of the search image and exemplar image; Finally, choose the position of the maximum score in score map as the tracking result.

During training, it use pairs that comprise an exemplar image and a larger search image, the loss function are

$$L(y,v) = \frac{1}{|D|} \sum_{u \in D} l(y\left[u\right], v\left[u\right]) \tag{2}$$

where $y\left[u\right] \in \left[+1, -1\right]$ and $v\left[u\right]$ are the true label and score map respectively for each position $u \in D$ in the score map, and l is logistic loss,

$$l(y,v) = log(1 + exp(-yv)) \tag{3}$$

The parameters of the conv-net θ are obtained by applying Stochastic Gradient Descent(SGD) to the problem.

$$argmin_{\theta} E_{(z,x,y)} L(y, f(z,x;\theta)) \tag{4}$$

Different from the prior works, we use a small filter-size in first convolution layer that can capture more discriminative feature map, which is verified in[10]. We adopt initial tracking images that are 63×63 and search images that are 127×127 pixels. The parameter of our siamese network and the dimensions of activations are given in Table1.

2.2. Multi-Template Fusion Scheme

SiameFc is taking $\varphi(z)$ where z is the initial tracking image in the first frame as a stable template to match candidate in a new frame. Although almost 4500 videos are used in training, only one stable template cannot meet the demands in long-term tracking where object appearance can change complicatedly. To this end, we introduce multi-template to siamese network tracker for the adaptiveness of appearance changes. Inspired by the work in [11], a stable template $T_s = \varphi(z)$without update, a soft stable template T_{ss} with thin update, and a adaptive template T_a with larger update are presented in this paper. According to the Equ(1), we can calculate three score maps S_s, S_{ss} and S_a with T_s, T_{ss} and T_a respectively. However, how to effectively fuse those results is still a open problem, there are two usual fusion methods in tradition: 1) a simple additive manner;2) a weighted additive way. The simple additive fusion equation is defined as

$$S_f^t = S_s^t + S_{ss}^t + S_a^t \tag{5}$$

Table 1. Architecture of Siamese Network

Layer	Support	Chan. map	Stride	exemplar	search
				63×63	127×127
conv1	5×5	96×3	1	59×59	123×123
pool1	3×3		2	29×29	61×61
conv2	5×5	256×48	1	25×25	57×57
pool2	3×3		2	12×12	28×28
conv3	3×3	384×256	1	10×10	26×26
conv4	3×3	384×192	1	8×8	24×24
conv5	3×3	256×192	1	6×6	22×22

and the weight additive fusion is formulated as

$$S_f^t = w_1 S_s^t + w_2 S_{ss}^t + w_3 S_a^t \tag{6}$$

where w_i denotes the weight of the ith template.

According to the experiments analyzing in[8], we can know that a stable template used in SiamFC can cope with common appearance variations which happen in most of the frames in videos; and it is good at target re-detection. However, the above fusion methods may disgrace the performance of target re-detection if the soft stable template and adaptive template have a big difference with the stable template. What's more, both simple additive and weighted additive method have the same problem that they should compute three score maps with three templates at first, which may add extra computational burden if the appearance variations are so slight that only one stable template can handle well; Consequently the fusion scheme should inherit the advantage of siamese network trackers and can cope with the complex appearance changes in long-term tracking. Motivated by the cascade architecture used in object detection[12], we propose a cascade architecture fusion scheme, as shown in Fig1.

Here are the following steps: **(1)** If the maximum score of S_s within the threshold, output the result and update the soft stable template and adaptive template; otherwise, step 2; **(2)** If the maximum score of $S_f^t = w_1^* S_s^t + w_2^* S_{ss}^t$ within the threshold, output the result and update the soft stable template and adaptive template; otherwise, step 3; **(3)** Output the fusion score maps $S_f^t = w_1^{*'} S_s^t + w_2^{*'} S_{ss}^t + w_3^{*'} S_a^t$. If the maximum score within the threshold, update the soft stable template and adaptive template. The fusion scheme can be interpreted that stable template has higher confidence than the soft stable template and adaptive template. Three complementary templates fusion tracking only happen in heavy appearance changes. As a result, the fusion scheme makes the tracker more efficient when there is a slight appearance change and more effective when there happen heavy appearance variations. What's more, when the target needs to re-detect in the case of out-of-view, our fusion scheme can inherit the re-detection performance of SiamFc because of cascade architecture evaluating the tracking result with the stable template at first which is the same operation to SiamFc tracker.

2.3. Update scheme

In order to cope with complicated appearance changes, it is essential to update appearance models. Most of the model update schemes in current trackers[13][11] are the global update, however, the global update may pollute template if occlusion and appearance changes aren't generally happening in the whole template. Besides, [14] shows that local features are robust in matching appearance models. In this paper, a patch template update method is presented. What's more, we apply optical flow as an update factor to measure the target appearance changes between the current frame and template frame. Our patch template update mechanism is defined as

$$T_j^{i+1} = (1 - \alpha\pi_j) T_j^i + \alpha\pi_j \varphi(y_j^i)$$

where the term α is a weighting factor to control the speed of the updates, T_j^i denotes the jth patch template at ith frame, $\varphi(y_j^i)$ is the optimal candidate, π_j denotes the appearance changes between the target image of template and the optimal candidate found in the ith frame, it can be defined as $\pi^i = \frac{O^i}{O^t}$ where O^t is the number of given pixels covered by the target box in template frame, then estimated optical flow in optimal candidate found in the current frame, O^i is the number of those pixels in an optimal candidate. If $O^i < O^t$, it means the pixels belong to target in this patch decrease so that it need to cut down the update rate; otherwise the opposite.

3. EXPERIMENTAL EVALUATION

We implemented the proposed tracker by Matlab, and using MatConvNet to train the siamese network. All the experiments are conducted on a single NVIDIA GeForce GTX Titan X and an Intel(R) Xeon(R)CPU E5-2637 v3 at 3.5GHz, our tracker runs at about 24 fps. The update rate α used in soft stable template is set to 0.1, and 0.3 to adaptive template. The fusion weight $w_1^* = 0.6, w_2^* = 0.4$ and $w_1^{*'} = 0.4, w_2^{*'} = 0.3, w_3^{*'} = 0.3$. We employ OTB-13 benchmark[15] which contains 51 videos to evaluate the efficacy of our proposed tracker.

Quantitative comparisons. We evaluate the proposed algorithm on the benchmark with comparisons to 8 state-of-the-art trackers, including SiamFC, LCT[16], Staple[17], MEEM[18], KCF [19], DSST[20], Struck[21], CXT[22]. For fair evaluations, we download each tracker's results on benchmark in their project homepage. We report the results in one-pass evaluation (OPE), temporal robustness evaluation (TRE) and spatial robustness evaluation (SRE) using overlap success rate in Fig2. The results show that our approach reaches state-of-the-art performance. When compare to SiamFc, our approach has a big improvement. This indicates that our multi-template fusion tracking acquires much better performance.

Qualitative comparisons. We compare our algorithm with SiamFc on some challenging sequences and the tracking results are showed in Fig3. The proposed tracker perform as well as SiamFC in common appearance variations $(freeman4, tiger2)$ and target re-detection $(Jogging2, Suv)$ which show that our cascade fusion scheme can inherit the advantage of SiamFC tracker. Besides, SiamFC drifts when there are heavy appearance changes such as fast motion$(Bolt)$ and background clutter $(Singer2)$, while the proposed tracker LST can perform well since multi-template with different update rate are more effective than a stable template without update in handling heavy appearance changes. In addition, SiamFC cannot performs well in similar confusion objects $(Liquor)$, where big filter size cannot get more

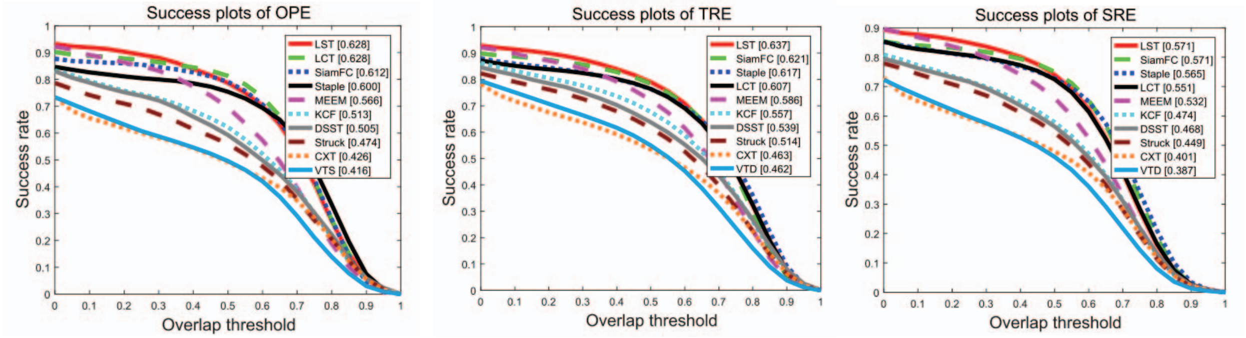

Fig. 2. The success plots of OPE (one pass evaluation), TRE (temporal robustness evaluation) and SRE (spatial robustness evaluation) of OTB-13[15].

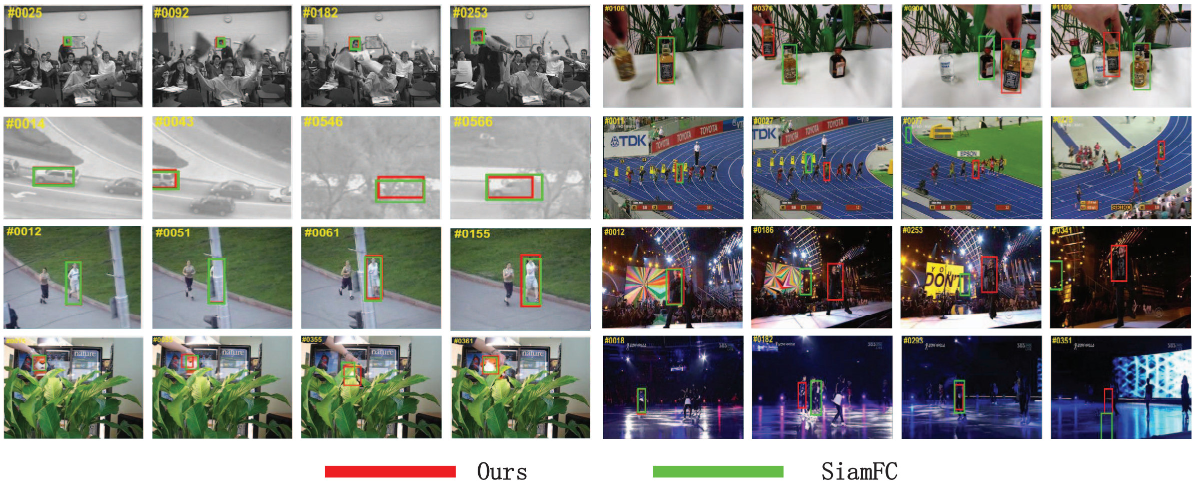

Fig. 3. Teacking results of our LST algorithm and SiamFC algorithm on four challenging sequences (from left to right and top to down are $freeman4$, $Liquor$, Suv, $Bolt$, $jogging2$, $Singer2$, $tiger2$ and $Skating$, respectively).

effective discriminative representation between similar targets, while LST with small filter size in the siamese network can cope well with this situation. However, the proposed tracker LST may fail when the object appearance change too much($Skating$), in this case, we can set high update rate for template update and high weight of adaptive template for fusion to meet the need but high update rate may result in failures in other sequences since it may introduce background information to template.

4. CONCLUSION

In this paper, we propose an effective algorithm for long-term tracker based on the siamese network. Our method introduces multi-template to siamese network tracker and uses cascade architecture fusion scheme to effectively and efficiently fuse the tracking results. Patch template update method based on optical flow are present to make the soft stable and adaptive template update more credible. The extensive empirical evaluations on tracking benchmark OTB2013 demonstrate that the proposed method not only can inherit the good performances of SiamFc tracker in re-detection and common appearance changes but also performs well in heavy appearance variations where SiamFc may fail. Moreover, it achieves frame-rate speed and outperforms the state-of-the-art trackers. We believe that our multi-template fusion scheme can be used in other siamese network trackers too, and expect future work to improve self-adaption of the fusion scheme.

5. REFERENCES

[1] Lijun Wang, Wanli Ouyang, Xiaogang Wang, and Huchuan Lu, "Visual tracking with fully convolutional networks," in *IEEE International Conference on Computer Vision*, 2015, pp. 3119–3127.

[2] Chao Ma, Jia Bin Huang, Xiaokang Yang, and Ming H-suan Yang, "Hierarchical convolutional features for visual tracking," in *IEEE International Conference on Computer Vision*, 2015, pp. 3074–3082.

[3] Martin Danelljan, Gustav Hager, Fahad Shahbaz Khan, and Michael Felsberg, "Convolutional features for correlation filter based visual tracking," in *IEEE International Conference on Computer Vision Workshop*, 2015, pp. 621–629.

[4] Naiyan Wang, Siyi Li, Abhinav Gupta, and Dit-Yan Yeung, "Transferring rich feature hierarchies for robust visual tracking," *CoRR*, vol. abs/1501.04587, 2015.

[5] Hyeonseob Nam and Bohyung Han, "Learning multi-domain convolutional neural networks for visual tracking," *CoRR*, vol. abs/1510.07945, 2015.

[6] David Held, Sebastian Thrun, and Silvio Savarese, "Learning to track at 100 fps with deep regression networks," in *Proc. ECCV*. 2016, Springer International Publishing.

[7] Ran Tao, Efstratios Gavves, and Arnold W. M. Smeulders, "Siamese instance search for tracking," in *Computer Vision and Pattern Recognition*, 2016, pp. 1420–1429.

[8] Luca Bertinetto, Jack Valmadre, Joao F. Henriques, Andrea Vedaldi, and Philip H. S. Torr, "Fully-convolutional siamese networks for object tracking," in *Proc. ECCV*. 2016, Springer International Publishing.

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., pp. 1097–1105. 2012.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[11] Huchuan Lu, Shipeng Lu, Dong Wang, Shu Wang, and Henry Leung, "Pixel-wise spatial pyramid-based hybrid tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 9, pp. 1365–1376, 2012.

[12] Paul Viola and Michael Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition*, vol. 1, pp. I–511–I–518 vol.1. 2001.

[13] R. Venkatesh Babu, Patrick Rez, and Patrick Bouthemy, "Robust tracking with motion estimation and local kernel-based color modeling," *Image and Vision Computing*, vol. 25, no. 8, pp. 1205–1216, 2007.

[14] Rhys Martin and Ognjen Arandjelović, "Multiple-object tracking in cluttered and crowded public spaces," in *Proceedings of the 6th International Conference on Advances in Visual Computing - Volume Part III*, 2010, pp. 89–98.

[15] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang, "Online object tracking: A benchmark," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[16] Chao Ma, Xiaokang Yang, Chongyang Zhang, and Ming Hsuan Yang, "Long-term correlation tracking," in *Computer Vision and Pattern Recognition*, 2015, pp. 5388–5396.

[17] Luca Bertinetto, Jack Valmadre, Stuart Golodetz, Ondrej Miksik, and Philip H. S. Torr, "Staple: Complementary learners for real-time tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1401–1409.

[18] Jianming Zhang, Shugao Ma, and Stan Sclaroff, "Meem: Robust tracking via multiple experts using entropy minimization," in *Proc. ECCV*, pp. 188–203. 2014.

[19] Joao.F. Henriques, Caseiro Rui, Pedro Martins, and Jorge Batista, "High-speed tracking with kernelized correlation filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 583–596, 2014.

[20] Martin Danelljan, Gustav Hager, Fahad Shahbaz Khan, and Michael Felsberg, "Accurate scale estimation for robust visual tracking," in *British Machine Vision Conference*, 2014, pp. 65.1–65.11.

[21] Sam Hare, Amir Saffari, and Philip H. S. Torr, "Struck: Structured output tracking with kernels," in *International Conference on Computer Vision*, 2011, pp. 263–270.

[22] Thang Ba Dinh, Nam Vo, and G Medioni, "Context tracker: Exploring supporters and distracters in unconstrained environments," in *The IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, Co, Usa, 20-25 June*, 2011, pp. 1177–1184.