Further Experiments for CVPR 2017 Submission #1241

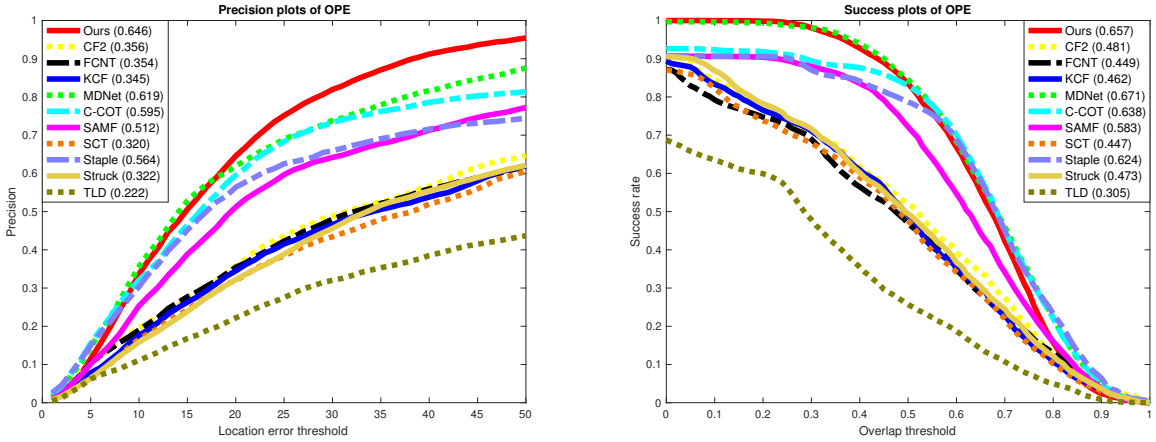
We run all of the following experiments on a server whose specifications are as follows:

1. 2 GTX 1080 GPUs
2. 56 Intel(R) Xeon(R) CPUs E5-2680 v4 @ 2.40GHz

Since the machine is much better than the previous one where results reported in submission paper were done, the overall FPS gets higher.

1 More Methods for Comparison

We could not find the open source code of Deep-SRDCF or SINT. We included two more DNN based methods CF2[3] and SCT[1] and one more non-DNN based method C-COT[2] for comparison. The precision and success plots are shown in Figure 1.



(a) Precision Curves

(b) Success Curves

Figure 1: The precision and success curves of the ILFPT and comparative methods on SVD-B in which the denoted numbers for the methods are the corresponding precision scores at the error threshold of 20 pixels in (a) and AUC values in (b), respectively.

We also evaluate the performance with respect to 5 attributes: scale variation(SV), occlusion(OCC), background cluttered(BC), deformation(DEF) and in-plane rotation(IPR). The 20 test videos have different attributes as shown in

Table 1. The performance of all algorithms with respect to the five attributes are shown in Table 2.

	SV	OCC	BC	DEF	IPR
video01		✓	✓		
video02	✓	✓			✓
video03	✓			✓	✓
video04	✓		✓		
video05	✓		✓		
video06			✓		
video07	✓	✓	✓		
video08			✓		✓
video09		✓	✓		
video10		✓		✓	
video11		✓			
video12		✓			
video13		✓	✓		
video14		✓		✓	
video15		✓	✓	✓	✓
video16	✓	✓			
video17	✓	✓		✓	✓
video18		✓	✓		
video19		✓	✓		✓
video20	✓		✓		

Table 1: Attributes of 20 test videos.

Trackers	Precision Scores					Success Scores					FPS
	SV	OCC	BC	DEF	IPR	SV	OCC	BC	DEF	IPR	
KCF	0.278	0.293	0.376	0.257	0.330	0.402	0.461	0.453	0.461	0.447	**116.610**
SAMF	0.568	0.465	0.507	***0.443***	***0.528***	0.597	0.576	0.542	0.617	0.583	4.170
Staple	**0.670**	0.508	0.573	0.424	0.525	**0.687**	0.606	0.592	0.633	0.646	9.307
Struck	0.239	0.292	0.353	0.256	0.275	0.413	0.478	0.469	0.488	0.458	1.524
TLD	0.275	0.188	0.270	0.208	0.233	0.363	0.287	0.337	0.328	0.300	10.201
C-COT	0.601	0.529	0.628	0.463	0.415	0.670	0.619	0.599	***0.684***	***0.652***	0.547
MDNet	***0.606***	***0.601***	**0.690**	0.435	0.496	***0.683***	**0.681**	**0.659**	0.676	**0.664**	0.664
FCNT	0.380	0.318	0.358	0.222	0.307	0.438	0.438	0.443	0.385	0.422	1.954
CF2	0.282	0.324	0.386	0.342	0.283	0.407	0.486	0.479	0.484	0.467	1.206
SCT	0.241	0.271	0.352	0.280	0.242	0.375	0.457	0.433	0.455	0.432	11.579
ILFPT(Ours)	0.588	**0.645**	***0.675***	**0.615**	**0.605**	0.622	***0.673***	***0.658***	**0.685**	0.647	***16.429***

Table 2: The comparative results among different trackers on the average **precision scores** and **success scores** with different attributes and **FPS(speed)** values, where the best and second best results are denoted in **red** and ***green*** colors, respectively.

2 The Effectiveness of the Sampling Scheme

We established the mechanism of switching detection frames and non-detection frames. Recall Algorithm 1 in our paper (L. 486 - 520), the Faster R-CNN sampling, deep re-id feature and classifier $\mathscr{C}_1$ are bundled. Local sampling, Haar feature and classifier $\mathscr{C}_2$ are bundled. They are two branches in our algorithm. So the effectiveness of Haar features in addition to the Re-ID features and the effectiveness of two sampling methods are explored simultaneously by changing the period parameter τ. We reported the precision scores (@20 pixels) and success scores at different τ. (see Figure 2)

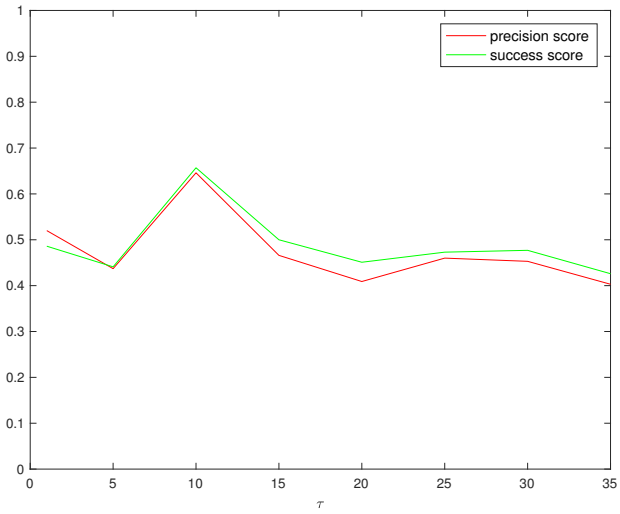


Figure 2: The precision and success scores at different τ.

When $\tau = 1$, the Faster R-CNN sampling is employed in each frame, whereas the local sampling performs no function. This point reflects the tracking ability of offline pretrained detector, since $\mathscr{C}_2$ is never used. When τ is large enough, the plot point reflects almost the tracking ability of Haar feature in our online learning model, since $\mathscr{C}_1$is rarely used. The best case is $\tau = 10$ when two classifiers are effectively used. Therefore, the neither tracking-by-detection only or online tracking by Haar feature is effective. So we can conclude that the switching scheme is effective.

3 The Search Window Size

The search windows size controls how large the area where our sampling methods select samples is. We explored all possible window size from 1 to 100, and get the following results. (see Figure 3)

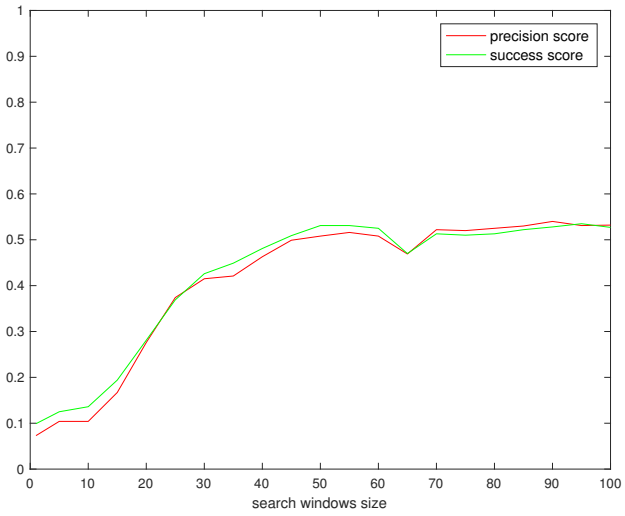


Figure 3: The precision and success scores at different search window size.

We found that as the search window size gets smaller, the performance degrades, since the target is out of the sampling area. As the size becomes larger, the true target falls in our search area, so that our sampling methods can propose good enough candidates for later classification. As search window size goes larger than 50, there is no significant improvement of the performance any more.

4 The Other Hyperparameters

Though the classification threshold θ is set by hand in advance, but it does not affect the result so much as long as it is near 0 (± 100). Because we are always trying to balance the number of stored positive samples and that of negative samples. Besides, two classifiers are independently computing the log probability ratio $R(v)$ (eq. (3) L. 453 - 459) for the deep re-id feature and Haar feature so that the feature dimension would not conflict.

References

[1] J. Choi, H. Jin Chang, J. Jeong, Y. Demiris, and J. Young Choi. Visual tracking using attention-modulated disintegration and integration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4321–4330, 2016.

[2] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg. Beyond correlation filters: Learning continuous convolution operators for visual tracking. In *European Conference on Computer Vision*, pages 472–488. Springer, 2016.

[3] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang. Hierarchical convolutional features for visual tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3074–3082, 2015.