Staple: Complementary Learners for Real-Time Tracking
Supplementary material

Luca Bertinetto Jack Valmadre Stuart Golodetz Ondrej Miksik Philip H.S. Torr
University of Oxford
{name.surname}@eng.ox.ac.uk

A. VOT-2014 results

We show here all the results provided by the vot-toolkit [9]. We use the latest version available at the time of writing (commit `d3b2b1d` of the 11 Sep 2015). For completeness, we present both reports generated by the toolkit. The analysis of `report_challenge` pools the performance *per attribute*, considering camera motion, illumination change, occlusion, size change and motion change. Instead, the analysis of `report_article` groups the results *per sequence*.

Tables 1 and 2 report the final rankings provided by the two analyses. Staple ranks first in both, but the order of the remaining trackers varies significantly between the two tables. The Accuracy-Robustness plot of Figure 1 gives a more intuitive visualization of the relative performance of the trackers by showing their accuracy and robustness ranks on a 2D grid. Tables 3, 4, 5 and 6 provide a more in-depth analysis by showing the raw values of, respectively, accuracy per attribute, robustness per attribute, accuracy per sequence and robustness per sequence. Finally, Figures 2, 3 and 4 complete the plots of Sections 4.3, 4.4 and 4.5 of the main article.

B. VOT-2015 results

We compare Staple against the 62 trackers from the VOT15 competition [8], using the same commit of the vot-toolkit as in our main paper. Since the results of the challenge have been disclosed after the CVPR deadline, it has been impossible for to include them in the main paper. Table 7 shows the top ten entries and confirms the competitiveness of our method, which ranks 4th. Two of the three methods that outperform Staple rely on CNN features, and *all* run significantly slower: MDNet [11] claims 1 fps, DeepSRDCF [2] $<$1 fps and SRDCF [4] 5 fps. Furthermore, Staple is by far the fastest among the top ten.

C. OTB-2013 results

In Figure 5 we report the results for the three evaluations supported by the OTB benchmark of Wu *et al*. [13]. In Figure 6 we also report the per-attribute plots for OPE (one pass evaluation), the main experiment of the benchmark, which compares the performance of the trackers on different types of challenge.

Tracker	Year	Where	Accuracy	# Failures	Overall Rank
Staple	-	-	0.644	9.38	4.37
DATs [12]	2015	CVPR	0.580	13.17	5.39
PLT_13 [7]	2013	VOT	0.523	1.66	5.41
DGT [1]	2014	TIP	0.534	13.78	5.66
SRDCF [4]	2015	ICCV	0.600	15.90	5.99
DMA [14]	2015	CVPR	0.476	0.72	6.00
PLT_14 [7]	2014	VOT	0.537	3.41	6.03
KCF [6]	2015	PAMI	0.613	19.79	6.58
DSST [3]	2014	BMVC	0.607	16.90	6.59
SAMF [10]	2014	ECCVw	0.603	19.23	6.79
DAT [12]	2015	CVPR	0.519	15.87	7.95
PixelTrack [5]	2013	ICCV	0.420	22.58	11.31

Table 1: Ranked list for VOT14 produced by **`report_challenge`**, which pool the results *per attribute*. First, second and third entries for accuracy, number of failures and overall rank are reported. Lower ranks are better.

Tracker	Year	Where	Accuracy	# Failures	Overall Rank
Staple	-	-	0.641	0.37	5.49
PLT_13 [7]	2013	VOT	0.535	0.05	5.73
PLT_14 [7]	2014	VOT	0.548	0.13	5.74
DATs [12]	2015	CVPR	0.581	0.94	5.88
DGT [1]	2014	TIP	0.551	1.15	6.02
DMA [14]	2015	CVPR	0.490	0.03	6.09
SRDCF [4]	2015	ICCV	0.594	0.70	6.52
KCF [6]	2015	PAMI	0.608	0.99	6.77
SAMF [10]	2014	ECCVw	0.603	0.92	6.87
DSST [3]	2014	BMVC	0.601	0.84	6.89
DAT [12]	2015	CVPR	0.528	0.97	7.01
PixelTrack [5]	2013	ICCV	0.422	1.58	8.99

Table 2: Ranked list for VOT14 produced by **`report_article`**, which pool the results *per sequence*.

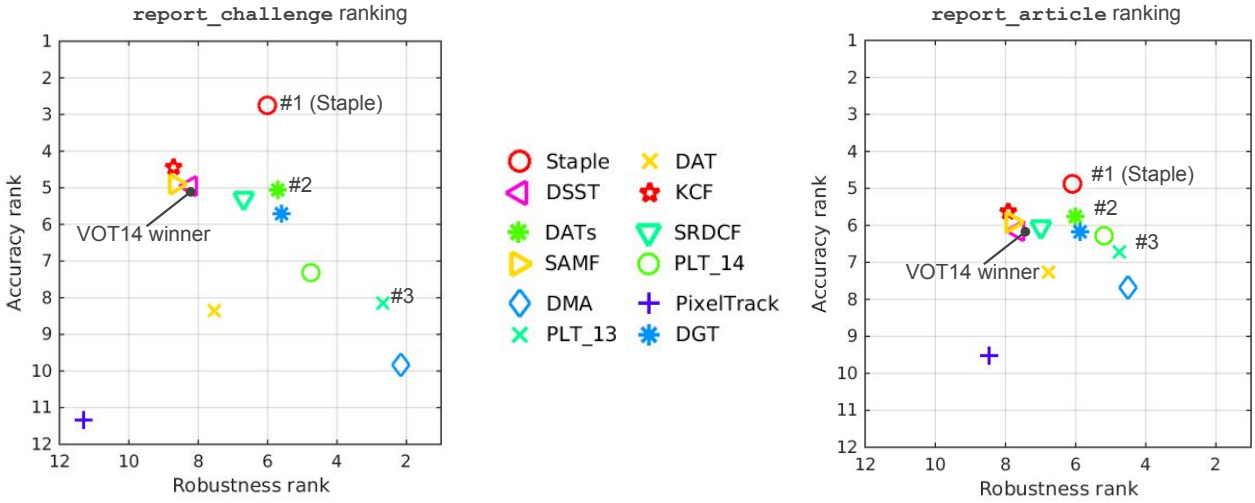

Figure 1: Accuracy-Robustness rank plot for `report_challenge` and `report_article`. Better trackers are closer to the top right corner. Robustness is inversely proportional to the number of failures.

	Staple	DAT	DATs	SRDCF	DMA	PixelTrack	DSST	KCF	SAMF	PLT_14	PLT_13	DGT
camera_motion	0.66	0.53	0.59	0.60	0.50	0.45	0.63	0.63	0.62	0.54	0.53	0.53
illum_change	0.73	0.47	0.53	0.73	0.42	0.42	0.74	0.73	0.66	0.50	0.51	0.41
occlusion	0.63	0.51	0.54	0.60	0.48	0.44	0.61	0.58	0.57	0.57	0.57	0.48
size_change	0.54	0.49	0.58	0.50	0.46	0.37	0.49	0.51	0.51	0.50	0.47	0.55
motion_change	0.65	0.54	0.61	0.61	0.48	0.42	0.60	0.62	0.62	0.55	0.54	0.55
empty	0.59	0.47	0.51	0.54	0.42	0.32	0.54	0.53	0.56	0.52	0.48	0.66
Overall	0.64	0.52	0.58	0.60	0.48	0.42	0.61	0.61	0.60	0.54	0.52	0.53

Table 3: Per-attribute raw accuracy values for VOT-2014. Higher is better.

	Staple	DAT	DATs	SRDCF	DMA	PixelTrack	DSST	KCF	SAMF	PLT_14	PLT_13	DGT
camera_motion	12	23	19	22	1	25	20	24	24	4	2	19
illum_change	1	10	8	1	0	8	1	1	1	1	0	14
occlusion	2	2	1	2	1	3	3	5	4	2	1	1
size_change	10	10	10	16	0	18	15	20	18	4	2	6
motion_change	11	15	12	17	1	33	24	26	25	4	2	14
empty	0	0	0	0	0	2	0	0	0	0	0	0
Overall	9.38	15.87	13.17	15.90	0.72	22.58	16.90	19.79	19.23	3.41	1.66	13.78

Table 4: Per-attribute raw robustness values (as number of failures) for VOT-2014. Lower is better.

	Staple	DAT	DATs	SRDCF	DMA	PixelTrack	DSST	KCF	SAMF	PLT_14	PLT_13	DGT
ball	0.83	0.67	0.86	0.49	0.66	0.44	0.54	0.73	0.74	0.69	0.59	0.80
basketball	0.72	0.70	0.59	0.63	0.72	0.56	0.62	0.64	0.74	0.73	0.74	0.49
bicycle	0.68	0.44	0.52	0.60	0.56	0.39	0.56	0.61	0.60	0.55	0.45	0.60
bolt	0.53	0.45	0.50	0.51	0.49	0.40	0.52	0.42	0.50	0.46	0.49	0.48
car	0.74	0.40	0.78	0.71	0.40	0.26	0.71	0.68	0.49	0.36	0.41	0.55
david	0.80	0.63	0.64	0.80	0.46	0.55	0.80	0.81	0.81	0.64	0.69	0.52
diving	0.14	0.34	0.37	0.14	0.51	0.35	0.39	0.17	0.17	0.37	0.37	0.33
drunk	0.59	0.47	0.52	0.54	0.42	0.32	0.55	0.53	0.56	0.52	0.48	0.67
fernando	0.38	0.42	0.44	0.35	0.22	0.30	0.32	0.38	0.38	0.38	0.38	0.59
fish1	0.24	0.54	0.70	0.25	0.45	0.46	0.30	0.37	0.43	0.40	0.37	0.54
fish2	0.30	0.29	0.36	0.19	0.35	0.32	0.27	0.18	0.23	0.24	0.24	0.42
gymnastics	0.53	0.55	0.55	0.48	0.52	0.51	0.37	0.47	0.43	0.57	0.56	0.55
hand1	0.32	0.59	0.53	0.48	0.56	0.40	0.18	0.44	0.42	0.63	0.59	0.57
hand2	0.52	0.45	0.57	0.43	0.35	0.15	0.33	0.31	0.33	0.59	0.52	0.36
jogging	0.74	0.72	0.70	0.70	0.58	0.48	0.73	0.73	0.75	0.65	0.69	0.64
motocross	0.26	0.20	0.20	0.25	0.38	0.35	0.24	0.28	0.23	0.44	0.40	0.41
polarbear	0.71	0.54	0.80	0.69	0.64	0.47	0.62	0.76	0.69	0.62	0.60	0.79
skating	0.54	0.32	0.41	0.57	0.56	0.35	0.57	0.63	0.44	0.48	0.48	0.29
sphere	0.84	0.68	0.75	0.88	0.60	0.54	0.88	0.85	0.84	0.64	0.52	0.80
sunshade	0.75	0.55	0.55	0.74	0.44	0.33	0.74	0.72	0.71	0.70	0.59	0.49
surfing	0.74	0.82	0.74	0.67	0.74	0.57	0.87	0.78	0.78	0.77	0.84	0.61
torus	0.84	0.73	0.80	0.83	0.66	0.42	0.78	0.82	0.81	0.69	0.78	0.80
trellis	0.82	0.51	0.50	0.79	0.40	0.49	0.79	0.78	0.81	0.50	0.49	0.48
tunnel	0.75	0.32	0.40	0.75	0.15	0.35	0.80	0.68	0.55	0.27	0.26	0.36
woman	0.78	0.62	0.64	0.74	0.58	0.55	0.76	0.71	0.73	0.73	0.71	0.53
Overall	0.64	0.53	0.58	0.59	0.49	0.42	0.60	0.61	0.60	0.55	0.53	0.55

Table 5: Per-sequence raw accuracy values for VOT-2014. Higher is better.

	Staple	DAT	DATs	SRDCF	DMA	PixelTrack	DSST	KCF	SAMF	PLT_14	PLT_13	DGT
ball	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00
basketball	0.00	1.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
bicycle	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
bolt	0.00	1.00	0.00	1.00	0.00	1.00	1.00	3.00	2.00	0.00	0.00	0.00
car	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
david	0.00	0.00	0.00	0.00	0.00	2.00	0.00	0.00	0.00	0.00	0.00	1.00
diving	4.00	1.00	1.00	4.00	0.00	0.00	1.00	4.00	4.00	0.00	0.00	0.00
drunk	0.00	0.00	0.00	0.00	0.00	2.00	0.00	0.00	0.00	0.00	0.00	0.00
fernando	1.00	0.00	0.00	1.00	0.00	4.00	1.00	1.00	1.00	1.00	0.00	0.00
fish1	1.00	2.00	0.00	3.00	0.00	2.00	1.00	3.00	3.00	0.00	0.00	0.00
fish2	2.00	3.00	2.00	6.00	1.00	1.00	4.00	6.00	5.00	0.00	0.00	2.00
gymnastics	0.00	0.00	0.00	2.00	0.00	0.00	5.00	1.00	2.00	0.00	0.00	0.00
hand1	2.00	0.00	1.00	1.00	0.00	2.00	2.00	3.00	3.00	0.00	0.00	1.00
hand2	1.00	3.00	0.00	3.00	0.00	10.00	6.00	6.00	5.00	0.00	0.00	5.00
jogging	1.00	0.00	0.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00
motocross	3.00	4.00	4.00	4.00	0.00	2.00	4.00	2.00	4.00	1.00	1.00	1.00
polarbear	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
skating	0.00	9.00	5.00	0.00	0.00	3.00	0.00	1.00	0.00	0.00	0.00	7.00
sphere	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
sunshade	0.00	0.00	0.00	0.00	0.00	3.00	0.00	0.00	0.00	0.00	0.00	0.00
surfing	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
torus	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
trellis	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
tunnel	0.00	2.00	7.00	0.00	0.00	3.00	0.00	0.00	0.00	0.00	0.00	8.00
woman	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00
Overall	0.37	0.97	0.94	0.70	0.03	1.58	0.84	0.99	0.92	0.13	0.05	1.15

Table 6: Per-sequence raw robustness values (as number of failures) for VOT-2014. Lower is better.

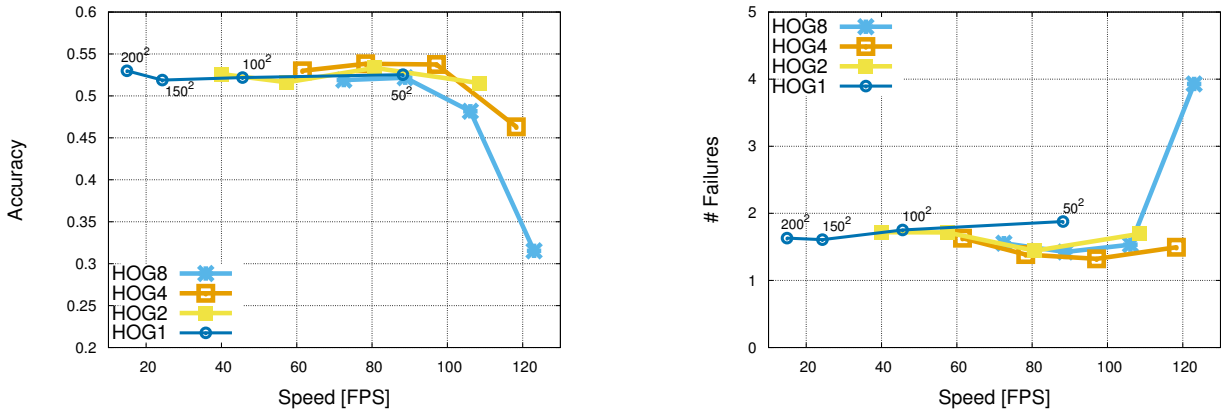

Figure 2: Accuracy (higher is better) and number of failures (lower is better) in relation to speed for HOG cells of size 1×1, 2×2 and 4×4 and different fixed areas (50^2, 100^2, 150^2 and 200^2).

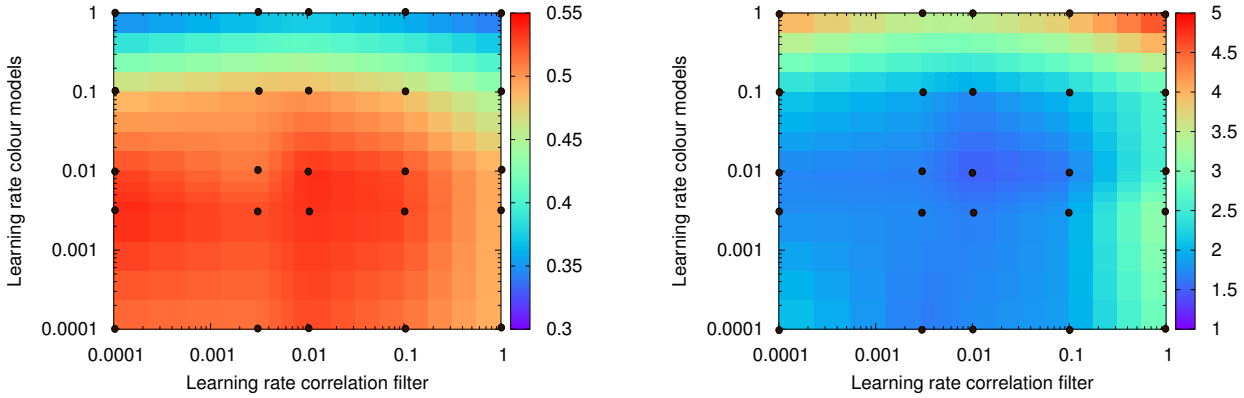

Figure 3: Accuracy (higher is better) and number of failures (lower is better) in relation to the learning rates η_{tmpl} and η_{hist}. Black points were obtained experimentally, others were interpolated. For both measures, the best performance is achieved at around 0.01 for both learning rates.

Tracker	Year	Where	Accuracy	# Failures	Overall Rank
MDNet	2015	ICCV	0.583	0.69	14.31
DeepSRDCF	2015	VOT	0.528	1.05	19.16
SRDCF	2015	ICCV	0.521	1.24	21.01
Staple	-	-	0.533	1.39	21.64
SO-DLT	2015	arXiv	0.535	1.78	22.71
NSAMF	2015	VOT	0.490	1.29	22.93
EBT	2015	arXiv	0.453	1.02	23.01
sPST	2015	ICCV	0.508	1.48	23.04
RAJSSC	2015	VOT	0.518	1.63	23.53
SC-EBT	2015	ICML	0.523	1.86	23.70

Table 7: VOT15 top 10 (of 63), `report_article`

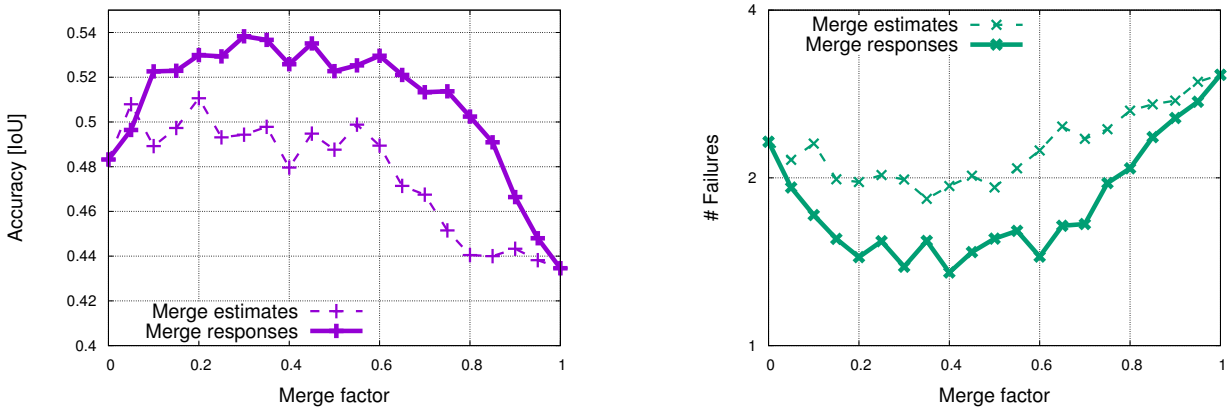

Figure 4: Accuracy (higher is better) and number of failures (lower is better) vs. merge factor α. For both measures, the best performance is achieved between 0.3 and 0.4.

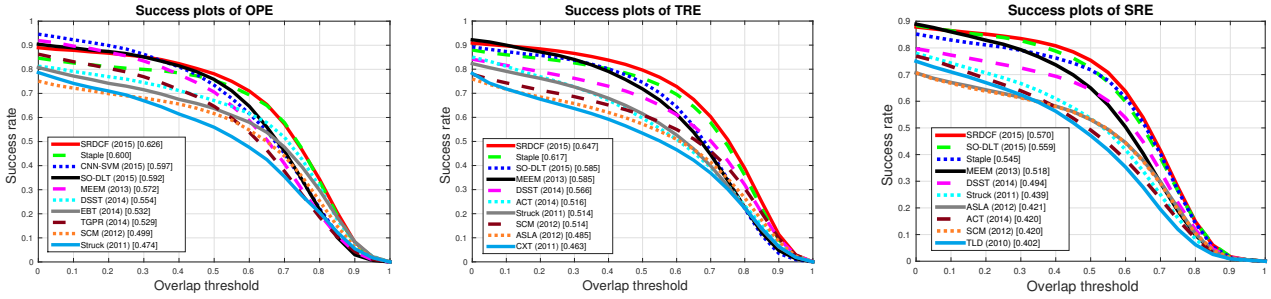

Figure 5: Success plots for OPE (one pass evaluation), TRE (temporal robustness evaluation) and SRE (spatial robustness evaluation) on the OTB-13 [13] benchmark.

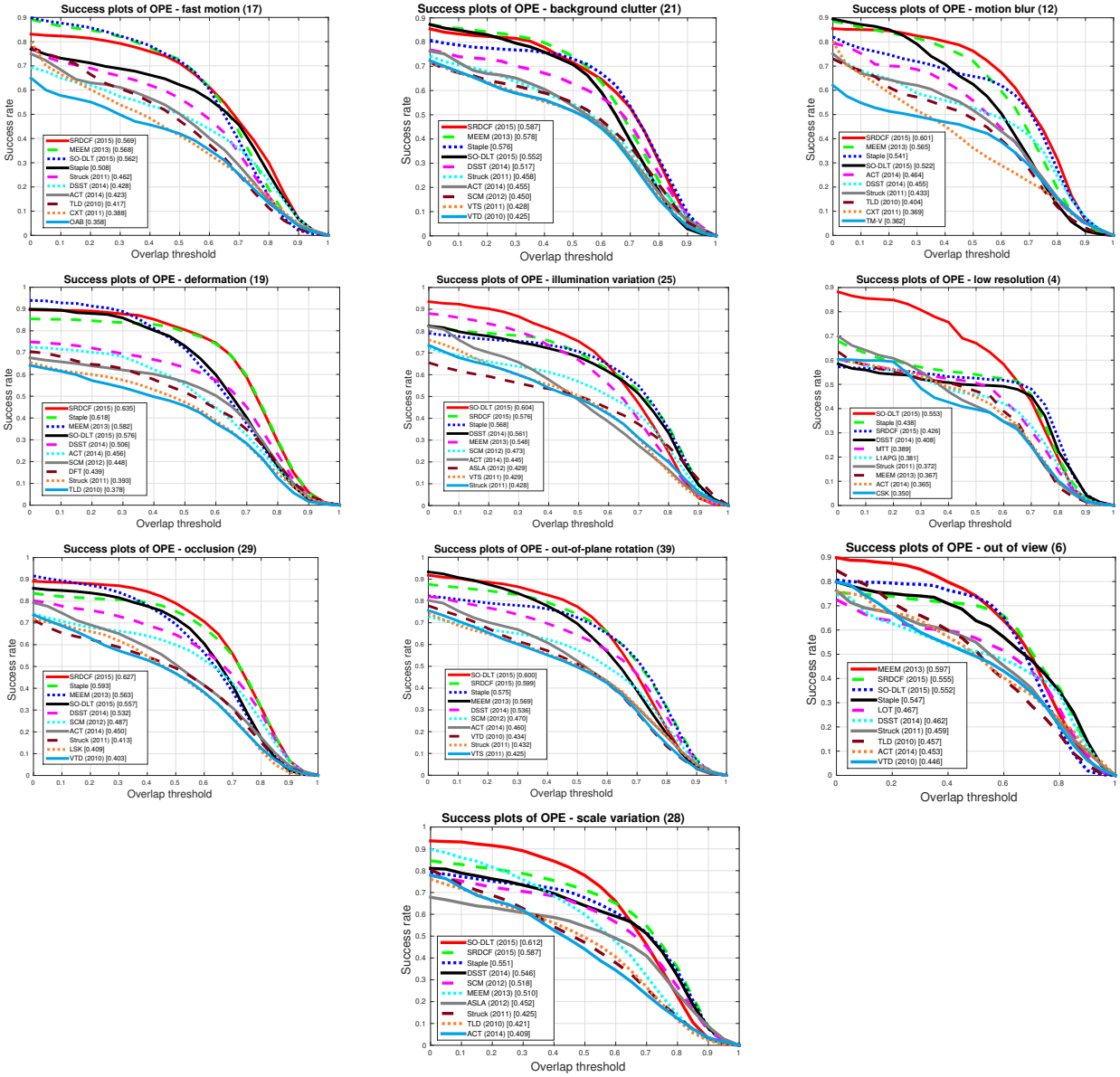

Figure 6: Per-attribute success plots for OPE.

D. Qualitative analysis

In this section we perform a qualitative analysis by comparing Staple to our closest competitors: the correlation tracker DSST [3] and the colour-based, distractor-aware DATs [12]. The videos have been produced by parsing the results produced by the vot-toolkit. Recall that the benchmark performs a re-initialization of the target (on the ground truth) five frames after a failure (detected when the overlap with the ground truth drops to zero). For this reason, we *a)* report the estimated bounding box of the three methods to visualise their accuracy and *b)* show the number of failures to account for their robustness.

We remark that Staple tends to inherit the behaviour of the better method for each sequence. This happens without any explicit rule to decide whether to use the template or the histogram: the two contributions in (3) are always weighted using a constant factor of $\alpha = 0.3$. The more confident (and "peaky") response prevails over the less confident (and "flat"). Below, we describe each sequence individually.

- `tiger`. In this case, DSST struggles to cope with the out-of-plane rotations performed by the target object, while DATs tends to fail, perhaps because of colour blending that occurs during motion blur. Differently, Staple handles well both situations.
- `tunnel`. The colour difference between the foreground object (the back of the passenger) and the background (the asphalt) is frequently poor. Moreover, half way through the sequence the video becomes very dark. In these conditions, the colour-based DATs clearly cannot keep track of the target, and fails repeatedly.
- `car2`. Similarly to the previous sequence, it is very difficult to discriminate between the colour statistics of foreground and background. Again, DATs performs badly.
- `gymnastics`. In this case it is the HOG-based DSST that has difficulties. Around the fifth second, the gymnast starts performing very sudden deformations to which the slow rigid-template update cannot cope. However, since background and foreground colours are quite different, DATs and Staple do not have trouble keeping track of the target.
- `ball1`. Here DSST fails multiple times and colour is an important cue to keep track of the target during the fast rotation.
- `diving`. This last sequence shows a case in which Staple performs worse than both DSST and DATs. Tables 5 and 6 demonstrates that this is rarely the case. DSST fails rapidly because of the target's deformation, but after re-initialisation it performs quite well. DATs fails in detecting the object's size, but it keeps track of the target almost until the end thanks to its distractor-aware component. We believe that Staple performs poorly because it is re-initialised at moments where the object is experiencing rapid change in shape, and the colour distribution of the background within the bounding box does not match that of the background outside the bounding box.

References

[1] Z. Cai, L. Wen, Z. Lei, N. Vasconcelos, and S. Z. Li. Robust Deformable and Occluded Object Tracking With Dynamic Graph. *TIP*, 23(12), 2014. 2

[2] M. Danelljan, G. Hager, F. Khan, and M. Felsberg. Convolutional features for correlation filter based visual tracking. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 58–66, 2015. 1

[3] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg. Accurate Scale Estimation for Robust Visual Tracking. In *BMVC*, 2014. 2, 8

[4] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg. Learning Spatially Regularized Correlation Filters for Visual Tracking. In *ICCV*, 2015. 1, 2

[5] S. Duffner and C. Garcia. PixelTrack: a fast adaptive algorithm for tracking non-rigid objects. In *ICCV*, 2013. 2

[6] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista. High-Speed Tracking with Kernelized Correlation Filters. *TPAMI*, 2015. 2

[7] M. Kristan et al. The Visual Object Tracking VOT2014 challenge results. In *ECCV*, 2014. 2

[8] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, L. Čehovin, G. Fernandez, T. Vojir, G. Häger, G. Nebehay, R. Pflugfelder, A. Gupta, A. Bibi, A. Lukežič, A. Garcia-Martin, A. Saffari, A. Petrosino, A. S. Montero, A. Varfolomieiev, A. Baskurt, B. Zhao, B. Ghanem, B. Martinez, B. Lee, B. Han, C. Wang, C. Garcia, C. Zhang, C. Schmid, D. Tao, D. Kim, D. Huang, D. Prokhorov, D. Du, D.-Y. Yeung, E. Ribeiro, F. S. Khan, F. Porikli, F. Bunyak, G. Zhu, G. Seetharaman, H. Kieritz, H. T. Yau, H. Li, H. Qi, H. Bischof, H. Possegger, H. Lee, H. Nam, I. Bogun, J. chan Jeong, J. il Cho, J.-Y. Lee, J. Zhu, J. Shi, J. Li, J. Jia, J. Feng, J. Gao, J. Y. Choi, J.-W. Kim, J. Lang, J. M. Martinez, J. Choi, J. Xing, K. Xue, K. Palaniappan, K. Lebeda, K. Alahari, K. Gao, K. Yun, K. H. Wong, L. Luo, L. Ma, L. Ke, L. Wen, L. Bertinetto, M. Pootschi, M. Maresca, M. Danelljan, M. Wen, M. Zhang, M. Arens, M. Valstar, M. Tang, M.-C. Chang, M. H. Khan, N. Fan, N. Wang, O. Miksik, P. H. S. Torr, Q. Wang, R. Martin-Nieto, R. Pelapur, R. Bowden, R. Laganiere, S. Moujtahid, S. Hare, S. Hadfield, S. Lyu, S. Li, S.-C. Zhu, S. Becker, S. Duffner, S. L. Hicks, S. Golodetz, S. Choi, T. Wu, T. Mauthner, T. Pridmore, W. Hu, W. Hübner, X. Wang, X. Li, X. Shi, X. Zhao, X. Mei, Y. Shizeng, Y. Hua, Y. Li, Y. Lu, Y. Li, Z. Chen, Z. Huang, Z. Chen, Z. Zhang, and Z. He. The visual object tracking vot2015 challenge results, Dec 2015. 1

[9] M. Kristan, J. Matas, A. Leonardis, T. Vojir, R. Pflugfelder, G. Fernandez, G. Nebehay, F. Porikli, and L. Cehovin. A Novel Performance Evaluation Methodology for Single-Target Trackers. *arXiv preprint arXiv:1503.01313v2*, 2015. 1

[10] Y. Li and J. Zhu. A Scale Adaptive Kernel Correlation Filter Tracker with Feature Integration. In *ECCVW*, 2014. 2

[11] H. Nam and B. Han. Learning multi-domain convolutional neural networks for visual tracking. *CoRR*, abs/1510.07945, 2015. 1

[12] H. Possegger, T. Mauthner, and H. Bischof. In Defense of Color-based Model-free Tracking. In *CVPR*, 2015. 2, 8

[13] Y. Wu, J. Lim, and M.-H. Yang. Online object tracking: A benchmark. In *CVPR*, 2013. 1, 6

[14] J. Xiao, R. Stolkin, and A. Leonardis. Single target tracking using adaptive clustered decision trees and dynamic multi-level appearance models. In *CVPR*, 2015. 2