Tracker Fusion on VOT Challenge: How does it perform and what can we learn about single trackers?

Christian Bailer[1]
Christian.Bailer@dfki.de

Didier Stricker[1,2]
Didier.Stricker@dfki.de

[1]German Research Center for Artificial Intelligence (DFKI), [2]University of Kaiserslautern

Abstract

Tracker fusion i.e. the fusion of the outputs of different tracking methods is an interesting new concept. Thus it should also be considered in the VOT challenges. In this paper we evaluate the performance of tracker fusion on the VOT2013 and VOT2014 datasets. Furthermore, we utilize the fusion concept to create novel fusion based measures for evaluating trackers. Fusion based evaluation is interesting as it does not evaluate trackers independently but in the context of all other trackers. It allows us for example to identify trackers that could despite poor average performance be interesting for research in object tracking. We found e.g. that all state-of-the-art trackers lack some strengths of a simple NCC tracker. Tracker fusion can exploit this and profit from an additional NCC tracker. We raise the question: Can this also be exploited in a more direct way i.e. can we e.g. combine NCC concepts with a state-of-the-art tracker?

1. Introduction

Visual object tracking is an important problem in computer vision with a wide range of applications. Existing tracking methods vary strongly in their approach. Some methods are e.g. based on optical flow, some use object templates instead, some use classifiers, some perform tracking by detection and some are combining different strategies, to name just a few approaches. The variety of approaches also leads to a variety of different behaviors of tracking methods witch makes a good detailed evaluation challenging. In our previous work [3] we e.g. figured out that on the tracking benchmark [15] the on average second worst method SMS [6] outperforms the on average best method SCM [16] on the "lemming" sequence – SMS even outperforms all 28 competing methods on this sequence. This shows that a tracking method can still be very interesting even if it performs on average poor compared to the state-of-the-art.

In our previous works [4, 3] we exploited this fact that different tracking methods have different strength and weaknesses. There, we designed *tracker fusion* approaches that fuse the output trajectory of different tracking methods into one fused trajectory. We showed that a fused result created based on many tracking results clearly outperforms the performance of all single input trackers. Furthermore, we showed in [3] that *tracker fusion* can even outperform single trackers in runtime if only fast trackers are fused. In our tests we were able to create a fusion approach that runs around 20 times faster than the best tracker SCM [16], while providing a similar performance (See Figure 4 a) in [3]). With the runtime of SCM we could outperform it by far.

As a result, *tracker fusion* can be seen as serious competitor to single tracking methods and should as such also be considered in the VOT challenges [11, 10]. Our first contribution is to provide the missing *tracker fusion* results for the VOT2013 [11] and VOT2014 [10] challenges based on our approach [3] in Section 2.

As discussed above tracking methods (like SMS) can be very interesting despite bad average performance if they provide an outstanding performance in some situations and if the outstanding performance can be exploited for better tracking results (e.g. by *tracker fusion*). So far popular evaluation measures only consider the average performance. This can lead to an underestimation of the potential of interesting tracking concepts.

As a result, our second contribution are novel measures for tracker evaluation that are based on the *tracker fusion* concept [3]. These measures do not evaluate trackers independently, but in the context of all other trackers i.e. they allow to estimate how interesting a tracker still is considering the existence of all the other trackers. Outstanding strengths likely will support a tracker in these measures while outstanding weaknesses will have the opposite effect. Common strengths and weaknesses shared by many trackers likely have no big effect on these measures. Furthermore, the measures guarantee that the strengths of interesting trackers do not only exist but can also be exploited (it is guaranteed that they can at least be exploited by our *tracker fusion* approach). We will show that our measures actu-

ally highlight rather uncommon tracking concepts (trackers based on uncommon concepts require a clearly lower average performance for a good rating). Of course our measures are also interesting for the *tracker fusion* concept itself. Here they can be applied directly (without further interpretation).

Note that the VOT challenges contain many trackers and as we want to cover all of them our paper contains complex tables. We will not discuss all of them in the text, but encourage readers to to examine the tables on their own e.g. to check the numbers for their own trackers or trackers they are interested in. We refer to the VOT papers for references to the trackers evaluated in the tables of this paper.

2. Fusion Results

In this section we present *tracker fusion* results (based on [3]) for the VOT2013 and VOT2014 challenges. There are some challenges and limitations in evaluating *tracker fusion* on the VOT results that we will discuss in Section 2.1 and 2.2, respectively. In Section 2.3 we present and discuss the actual *tracker fusion* results.

2.1. VOT Specific Fusion Result Creation

A challenge in evaluating fusion on VOT results is the fact that trackers are reinitialized whenever they fail (overlap to ground truth = 0). In more detail: If a tracker fails it does not track for 5 frames and is then reinitialized with the current ground truth. After reinitialization tracking results are not considered for accuracy calculation for 10 frames to avoid giving them an unfair advantage (For details see [10]). For comparability, we also use the same procedure of reinitialization for the *tracker fusion* approach.

Tracker fusion is usually following several similar tracking results at the same time (by a weighted average). If a tracker is removed from the weighted average due to failure it cannot make the fusion approach to fail as well, anymore. To avoid this we do not remove the bounding boxes of failed trackers. Instead we simulate their bounding boxes for the $5 + 10 = 15$ frames (see above) where the reinitialized tracking result is not valid, yet. Failed trackers are simulated with the velocity they had before failure. With this approach we can make sure that fusion fails as well when the trackers it is building on are failing.

2.2. VOT Specific Limitations

For a fair comparison it is also necessary to reinitialize all underling trackers when *tracker fusion* fails (as they are a part of the "fusion tracker"). This is not possible as we only have the raw VOT tracking results but not the source codes of the trackers to perform own experiments. As a result, our reinitialized fusion is not building on freshly reinitialized trackers. Instead, it is building on trackers that are already tracking for a longer time period. From such a tracker we can on average expect a lower accuracy and a shorter time to next failure than from a freshly reinitialized tracker. Thus, we expect a certain penalty for fusion on VOT results (i.e. results would probably be better with proper reinitialization). Still we think that the provided fusion results are meaningful.

2.3. Results

As can be seen in Table 2 and 3 the basic *tracker fusion* approach of [3], which we call "Fusion" always outperforms the best tracking method in success score.[1] However, the difference between the best tracker and the *tracker fusion* result ("Fusion") is smaller for all VOT results than in the evaluation of our previous work [3]. We think that an important reason for this is that reinitialization leads to smaller differences in success score between trackers. The difference is 167% between the best and worst tracker for the results in [15] (used by [3]), while it is only 75% for VOT2013 and 56% for VOT2014 (baseline results). Thus, it is obvious to also expect a smaller advance for *tracker fusion* on the VOT results. Considering this, the *tracker fusion* results on VOT are comparable to the original fusion results in [3].

Only on the baseline results of VOT2014 the advantage is noticeable smaller. Conspicuous here is that the best three trackers clearly outperform the other trackers with a great distance in success score. The *tracker fusion* approach does not know which trackers have a good performance and thus has to consider all 38 trackers with the same attention. Therefore, it is no big surprise that the advantage of *tracker fusion* is smaller if the best three trackers in the tracking set are positive outliers. If we remove them from the fusion set we can obtain a success score of of 0.664, which is considerably better than the best tracker in the reduced fusion set (the 4. best tracker in the full set). This fact supports our explanation above.

An important reason why positive outliers have a limited influence is that we also fuse trackers that are disadvantageous for *tracker fusion*. There are ways to deal with this issue. A simple way, is to only fuse the best n trackers. The value "Last 5 only" in Table 2 and 3 show the achieved success score if the basic approach in [3] is only performed on the last 5 trackers in the table (with the highest success score). For the VOT 2014 dataset this leads to clearly better results than fusing all 38 trackers. For VOT 2013 the result is ambiguous (positive on baseline and negative on region noise). Still, it makes a lot of sense to fuse only the best trackers, especially if we consider the much lower effort of fusing fewer trackers.

Instead of keeping the best trackers we can also just keep the trackers that are beneficial for fusion. Beneficial track-

[1] success score is the average overlap between the tracker bounding box and the ground truth. See [3].

	Trackers	Success
1 tracker	KCF	0.661
2 trackers	KCF+SAMF	0.673
3 trackers	DSST+DGT+LT-FLO	0.701
4 trackers	DSST+SAMF+DGT+LT-FLO	0.714
5 trackers	DSST+KCF+eASMS+DGT+LT-FLO	0.720

Table 1: The best trackers for fusion on the baseline experiment of VOT2014.

ers are determined by iteratively removing trackers that are adverse for fusion (in success score) from the set of trackers until no tracker is adverse anymore for fusion (see *Global Removal* in [3]). The "Fusion+" rows in the tables show the result if only beneficial trackers are fused. The result is expectably the best in success score. "FusionF+" does the same but we avoid removing trackers with low failure rate here (PLT13, PLT14). This allows to obtain a lower failure rate. "FusionFT+" is "FusionF+" with the online *trajectory optimization* approach of [3] instead of their basic approach. *Trajectory optimization* avoids leaps in the trajectory. As can be seen, this also helps to lower the failure rate.

The failure rates of *tracker fusion* are among the best. Only few trackers have a lower failure rate than the considered *tracker fusion* approaches. We found that the failure rate can be further decreased at the cost of decreasing *success score* by increasing the σ parameter of [3].

Table 4 shows the *tracker fusion* performance under different conditions. Only for illumination change "Fusion" does not outperform single trackers. Here only two trackers (DSST [7] and KCF [9]) strongly outperform all other trackers. Furthermore, the 3. best tracker also clearly outperforms the 4. best. Thus, we again have the effect of few strong positive outliers (that we have discussed above), which is unfavorable for fusion. For illumination change even "Fusion+" outperforms the best two trackers only slightly. We think this is because there are not many illumination changes that are not successfully handled by DSST [7] and KCF [9], but by other trackers. The best fusion results can be achieved for size change and occlusion.

Best Trackers for Fusion Table 1 shows the optimal tracker sets of size 1-5 for *tracker fusion* on the VOT2014 baseline results. The best trackers a tracker set of size two are the two trackers with the highest success score KCF [9] and SAMF. For larger sets they are often not chosen. This shows that the best performing trackers are not necessarily the most attractive for fusion.

The "Fusion+" set is much larger. It consists of ABS, DSST [7], DynMS, FSDT, IMPNCC, KCF [9], MCT, MIL [2], NCC, OGT [12], PLT14, SAMF, aStruck, CMT [13], DGT, eASMS, FoT [14], FRT [1], HMMTxD, LGTv1 [5] and LT-FLO. However, note that it was determined by a greedy approach, while the results in Table 1 are the global maximums.

3. Tracker Evaluation with Fusibility Measures

In this section we introduce our novel *tracker fusion* based evaluation measures (*fusibility measures*) and discuss VOT evaluation results based on these measures. As discussed in the introduction *fusibility measures* can be very interesting for tracker evaluation.

The remainder of this section is structured as follows: First we introduce our novel fusibility measures in Section 3.1. There, we also motivate them regarding *tracker fusion*. In Section 3.2 we motivate their usefulness for tracker evaluation. In Section 3.3 we discuss interesting evaluation results i.e. we show what we can learn from the fusibility measures about singe trackers in the VOT challenges. Finally, we perform experiments to identify the main competing and supporting trackers for each tracker in Section 3.4 (competing trackers have similar strengths, supporting trackers complement each other).

3.1. Fusibility Measures

The fusion columns in Table 2 and 3 show how a tracker affects fusion. We have to perform two experiments to determine the *fusibility measures* for a tracker: fusion with all trackers and fusion with all trackers, excluding the tested tracker. For each frame we determine the difference in overlap (success score) between the two experiments. The difference (impact) can be positive or negative in each frame. The total impact a tracker has on the fusion result is calculated independently for the total positive impact and the total negative impact. For D being all frames of all tested sequences, F_T the fusion result for the set of all tracker T, $O(x)$ the overlap of a tracking result to the ground truth the total positive ($\Delta O_t^+(F_T)$) and negative ($\Delta O_t^-(F_T)$) impacts are calculated as:

$$\Delta O_t^+(F_T) = \sum_{f \in D} \frac{\max\left(0, O(F_T(f)) - O(F_{T\setminus t}(f))\right)}{0.001|D|} \tag{1}$$

$$\Delta O_t^-(F_T) = \sum_{f \in D} \frac{\min\left(0, O(F_T(f)) - O(F_{T\setminus t}(f))\right)}{0.001|D|} \tag{2}$$

The "Gain" column in Table 2 and 3 is calculated as:

$$\text{Gain} = O_t(F_T) = O_t^+(F_T) + O_t^-(F_T) \tag{3}$$

The "+Seq" column counts the percentage of sequences with an gain > 0 (gain calculated by sequence instead of all frames). Note that in VOT each sequence is processed three times. We consider these three runs as individual sequences.

The "+Seq" measure is interesting to see if the fusion gain is fairly equally distributed over different sequences. This is the case for most trackers. A serious outlier in this regard is aStruck [10] on the baseline experiment. Despite its on average strong positive gain the gain is only positive for 42.7% of the sequences. The "+/-" column is calculated as:

$$+/- = \frac{O_t^+(F_T)}{O_t^-(F_T)} \tag{4}$$

The gain measure rates the overall influence of a tracker and shows if it is positive or negative. However, it does not show if for example a positive gain (e.g. 1) is achieved mainly by positive influence (e.g. +1.2 and -0.2) or only by a slight overhang of strong positive and negative influence (e.g. +10 -9). A tracker with high influence in both directions influences the fusion result strongly. For a new untested sequence this can easily lead to strong positive but also strong negative effects i.e. the gain is unstable/unreliable. A tracker with in general low influence but strong positive overhang should act mostly unobtrusive e.g. by behaving similar to the majority of trackers. However, if it stands out the effect is usually positive. Obviously, the latter is preferable. The "+/-" measure rates this (larger means better/more reliable).[2]

3.2. Using Fusibility Measures for Tracker Evaluation

We think that large gain and "+/-" values can reveal interesting trackers not only for fusion. If there is e.g. a tracker t_1 and improved versions of this tracker $t_2...t_n$ then adding $t_2...t_n$ to the fusion set will significantly lower the gain and the "+/-" value of t_1. This is because the strengths of t_1 are also covered by $t_2...t_n$ and thus t_1 cannot utilize these strengths anymore to improve fusion, while the weaknesses of t_1 that are not shared by $t_2...t_n$ can still harm fusion. On the other hand, a tracker with poor average performance but large gain and "+/-" value is likely a tracker with much originality that has strengths that are that original that they are not even covered by the top performing trackers in the fusion set.[3]

As a result, gain and "+/-" values are providing clues for how interesting a tracker can be for future research. Usually, top performing trackers gain the most attention in the research community, which makes it likely that upcoming trackers are building on similar concepts. While there is no need for considering trackers without serious originalities, whose strengths are widely covered by better performing trackers, the concepts behind trackers with high originality and thus likely large gain and "+/-" values have the potential to improve future trackers – even if the average performance of the considered tracker is low. In fact this is not only a potential as our *tracker fusion* approach is already successfully utilizing it.

3.3. Interesting Evaluation Results

When searching Table 2 and 3 for which poorly performing trackers have the highest "+/-" and gain we find that this are mostly either methods with simple concepts like NCC including IMPNCC (template tracking) and Meanshift (including DynMS) or methods with novel tracking ideas that are not utilized by state-of-the art trackers like LGT [5] (local + global tracking), FSDT (dynamic feature selection), STMT (combined camera and object tracking). On the other hand, methods with small gain and "+/-" are often methods that are building on approved concepts like classifier learning (e.g. MIL [2], Struck [8]). For methods building on approved concepts it is likely that better trackers are using similar concepts, but they are doing better which makes the original method redundant.

The fact that trackers with simple concepts like NCC and Meanshift can obtain a positive gain in fusion shows that even state-of-the art trackers still lack strengths of these simple concepts. Also the strengths of novel concepts like local + global tracking (LGT), dynamic feature selection (FSDT) and combined camera and object tracking (STMT) seem to be not covered by state-of-the-art trackers. Thus, an interesting question is: Can these concepts be utilized e.g. by incorporating them into a state-of-the-art tracker to create a even better tracker?[4]

Interesting is also, that there is a strong diversity in the "+/-" and gain values in Table 4. FRT [1] has e.g. a very good gain and "+/-" value for illumination change but a negative gain for size change and occlusion. Furthermore, gains and '+/-" values are in general large for illumination change (in positive as well as negative direction), while they are in general small for occlusion. Remarkable is also, that the occlusion set allows the simple NCC tracker to outperform all other trackers in gain and "+/-".

3.4. Competitors and Supporters

Table 5 shows the main competitors and supporters of each tracker. Competitors cost the tracker gain if they are added to the fusion set, while supporters help the tracker to improve the gain. The values "Δ gain" are determined by determining the gain with and without the competing/supporting tracker and taking the difference between both gains. Supporters complement each other well in their

[2]Note that a negative gain leads to "+/-" values < 1. Bellow 1 larger values mean less reliability. Still, even here larger values are better. An unreliable negative effect is better as it has the possibility to turn positive, while a reliable negative effect is negative for sure.

[3] This is because, the other trackers fail in clearly lowering the gain and "+/-" value of the tracker, which they could do by covering the strengths of the tracker.

[4]It is clear that *tracker fusion* can achieve this, but can we do it even better if we directly incorporate the concepts on a lower level?

baseline						region noise					
General			Fusion			General			Fusion		
Name	Failures	Success	+Seq.	+/-	Gain	Name	Failures	Success	+Seq.	+/-	Gain
MORP	1740	0.376	37.5%	1.77	0.94	MORP	1754	0.375	31.25%	1.41	0.77
CACTuS-FL	237	0.386	45.83%	1.2	0.76	STMT	318	0.392	77.08%	3.04	4.13
STMT	312	0.395	66.67%	1.98	3.08	CACTuS-FL	207	0.392	52.08%	0.95	-0.2
CT	101	0.472	14.58%	0.39	-6.15	CT	80	0.466	29.16%	0.51	-3.86
RDET	69	0.492	14.58%	0.38	-5.82	RDET	63	0.498	35.41%	0.43	-5.68
Meanshift	60	0.498	70.83%	1.75	3.31	HT	236	0.501	54.16%	1.21	1.12
HT	195	0.514	39.58%	0.89	-0.72	MIL	77	0.511	35.41%	0.57	-3.52
LGT	12	0.536	58.33%	1.43	2.36	Meanshift	116	0.515	62.5%	1.4	1.9
MIL	70	0.537	16.67%	0.53	-4.58	LGT	8	0.516	62.5%	1.52	2.33
LGTpp	4	0.538	64.58%	1.68	3.21	ORIA	108	0.532	70.83%	1.73	2.98
SwATrack	126	0.549	43.75%	0.89	-0.81	LGTpp	7	0.537	79.16%	2.26	4.6
ORIA	105	0.566	77.08%	2.8	6.72	SwATrack	115	0.539	56.25%	1.06	0.42
Struck	189	0.584	39.58%	0.59	-4.08	Struck	226	0.551	29.16%	0.67	-2.51
PJS-S	76	0.595	81.25%	2.35	5.45	IVT	98	0.561	56.25%	1.13	0.81
Matrioska	90	0.598	41.67%	0.76	-2.3	Matrioska	81	0.564	39.58%	0.74	-2.17
TLD	321	0.599	54.17%	1.21	1.49	AIF	63	0.564	50%	0.86	-1.08
IVT	87	0.602	60.42%	1.24	1.64	PJS-S	81	0.565	77.08%	2.05	4.04
CCMS	21	0.602	43.75%	0.9	-0.96	DFT	69	0.571	41.66%	0.87	-0.98
EDFT	42	0.608	43.75%	0.81	-1.81	LT-FLO	77	0.576	75%	1.7	3.28
DFT	63	0.61	54.17%	1.48	3.08	CCMS	23	0.582	45.83%	1.01	0.12
PLT	0	0.611	60.42%	1.3	2.1	TLD	314	0.583	66.66%	1.31	1.85
AIF	64	0.621	47.92%	0.76	-2.4	EDFT	49	0.586	45.83%	0.87	-0.98
GSDT	42	0.632	56.25%	0.7	-2.9	PLT	4	0.591	50%	0.9	-0.8
LT-FLO	78	0.634	83.33%	2.83	8.29	SCTT	107	0.594	75%	1.8	4.1
SCTT	105	0.645	77.08%	2.32	6.1	GSDT	65	0.594	50%	1.15	1.14
FoT	66	0.657	81.25%	2.84	7.19	FoT	70	0.618	79.16%	2.37	5.77
Fusion	9	0.712	⇐ Last 5 only: 0.721			Fusion	11	0.670	⇐ Last 5 only: 0.664		
Fusion+	10	0.743	-			Fusion+	10	0.702	-		

Table 2: Our results on the VOT2013 data. See text for details (Section 2 for Fusion rows. Section 3 for the fusion column.). Colors are form full red (worst) to full green (best). Cyan is used for fusion results that outperform the best tracking results.

strengths, while competitors are blocking each other. Similar trackers are likely to be competitors as the gain achieved by the common strength has to be shared between both trackers (e.g. Struck and ThunderStruck). Interestingly, the main supporter for the best performing tracker KCF [9] is the simple NCC tracker. This again shows, that simple tracking concepts like NCC are not outdated and that a combination of a state-of-the-art method like KCF with a simple concept like NCC can lead to a better tracker.

4. Conclusion

In this paper we evaluated *tracker fusion* on the VOT2013 and VOT2014 benchmarks and evaluated singe trackers regarding their fusibility. We found that fusion can also help to identify interesting trackers/tracking concepts among the non top performing trackers. We found that state-of-the art trackers lack some strengths of simple tracking concepts like NCC and tracking ideas like "local + global tracking" or "combined camera and object tracking" as these were never incorporated into the state-of-the-art. A limitation of our evaluation is that we did not consider the failure rate (robustness) in Section 3. In future work we want to perform these tests also regarding failure rate.

Acknowledgements

This work was partially funded by the BMBF project DYNAMICS (01IW15003).

References

[1] A. Adam, E. Rivlin, and I. Shimshoni. Robust fragments-based tracking using the integral histogram. In *Computer vision and pattern recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 798–805. IEEE, 2006.

[2] B. Babenko, M.-H. Yang, and S. Belongie. Visual tracking with online multiple instance learning. In *Computer Vision*

baseline						region noise					
General			Fusion			General			Fusion		
Name	Failures	Success	+Seq.	+/-	Gain	Name	Failures	Success	+Seq.	+/-	Gain
MIL	166	0.424	42.6%	0.91	-0.33	MIL	199	0.376	40%	0.8	-0.97
CT	234	0.45	24%	0.62	-1.77	IMPNCC	257	0.407	44%	1.09	0.43
PTp	105	0.45	44%	0.76	-1.27	CT	253	0.449	41.3%	0.76	-1.5
IMPNCC	273	0.467	64%	2.02	2.75	IIVTv2	220	0.455	49.3%	0.93	-0.32
FSDT	231	0.47	60%	1.59	1.77	PTp	106	0.463	52%	0.8	-1.28
IPRT	148	0.481	48%	0.89	-0.51	FRT	269	0.468	42.7%	0.99	-0.05
LGTv1	49	0.481	57.3%	1.45	1.53	IVT	226	0.471	37.3%	0.76	-1.47
IIVTv2	235	0.482	45.3%	0.91	-0.37	LGTv1	56	0.475	62.7%	1.24	1.12
IVT	207	0.489	40%	0.75	-1.39	CMT	198	0.477	34.7%	0.76	-1.55
FRT	249	0.5	52%	1.31	1.25	IPRT	147	0.485	45.3%	0.82	-1.18
SIRPF	136	0.508	52%	0.96	-0.2	aStruck	196	0.486	48%	0.91	-0.45
BDF	90	0.514	46.6%	0.6	-2.95	BDF	113	0.488	44%	0.77	-1.46
ThunderStruck	166	0.518	30.6%	0.41	-5.35	FSDT	213	0.497	50.7%	1.17	0.88
CMT	198	0.518	49.3%	1.05	0.26	EDFT	139	0.506	53.3%	0.97	-0.16
ABS	93	0.519	61.3%	1.43	1.8	FoT	207	0.507	60%	1.09	0.52
FoT	171	0.52	68%	1.49	1.85	SIRPF	160	0.507	58.7%	1.18	0.92
MatFlow	57	0.523	29.3%	0.41	-5.05	NCC	526	0.51	53.3%	1.31	1.38
DynMS	115	0.527	69.3%	1.58	2.14	ThunderStruck	160	0.511	44%	0.76	-1.53
Matrioska	186	0.528	33.33%	0.45	-4.49	ACT	142	0.514	52%	0.9	-0.59
aStruck	183	0.529	42.7%	1.68	3.13	Matrioska	250	0.516	38.7%	0.76	-1.42
EDFT	138	0.541	45.3%	0.58	-3.5	ABS	92	0.516	58.7%	1.2	1.11
Struck	157	0.543	29.3%	0.46	-5.23	Struck	165	0.517	41.3%	0.79	-1.32
MCT	71	0.547	66.7%	1.56	2.18	DynMS	128	0.517	65.3%	1.51	2.25
ACT	111	0.55	41.3%	0.5	-4.48	MatFlow	110	0.517	49.3%	0.8	-1.26
VTDMG	99	0.555	46.7%	0.62	-3.02	LT-FLO	199	0.517	81.3%	2.14	3.57
NCC	573	0.56	54.7%	1.67	2.82	qwsEDFT	135	0.529	52%	0.84	-0.94
PLT13	6	0.561	37.3%	0.81	-1.18	OGT	238	0.532	56%	1.27	1.47
qwsEDFT	102	0.562	41.3%	0.54	-3.99	VTDMG	86	0.534	69.3%	1.28	1.44
eASMS	84	0.573	60%	1.5	2.62	PLT13	22	0.54	58.7%	1.12	0.69
PLT14	12	0.573	48%	0.71	-2.33	PLT14	22	0.546	49.3%	0.93	-0.49
ACAT	117	0.576	57.3%	1.77	3.54	ACAT	135	0.547	64%	1.12	0.69
OGT	254	0.581	62.7%	1.72	3.27	MCT	88	0.552	62.7%	1.28	1.44
LT-FLO	189	0.585	84%	2.72	4.91	eASMS	83	0.561	61.3%	1.43	2.44
DGT	75	0.602	76%	2.45	5.09	HMMTxD	114	0.584	74.7%	1.63	2.9
HMMTxD	114	0.614	76%	2.4	5.82	DGT	83	0.594	81.3%	2.59	5.47
DSST	87	0.651	68%	2.87	6.49	SAMF	112	0.606	68%	1.87	4.22
SAMF	96	0.655	61.3%	1.90	5.29	KCF	117	0.607	64%	1.32	1.87
KCF	99	0.661	60%	1.97	5.67	DSST	92	0.61	74.7%	1.67	3.02
Fusion	85	0.678	⇐ Last 5 only: 0.701			Fusion	104	0.654	⇐ Last 5 only: 0.679		
Fusion+	119	0.727	-			Fusion+	101	0.691	-		
FusionF+	68	0.714	-			FusionF+	76	0.690	-		
FusionFT+	49	0.710	-			FusionFT+	72	0.685	-		

Table 3: Our results on the VOT2014 data. See text for details (Section 2 for Fusion rows. Section 3 for the fusion column.). Colors are form full red (worst) to full green (best). Cyan is used for fusion results that outperform the best tracking results.

and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pages 983–990. IEEE, 2009.

[3] C. Bailer, A. Pagani, and D. Stricker. A superior tracking approach: Building a strong tracker through fusion. In *Computer Vision–ECCV 2014*, pages 170–185. Springer, 2014.

[4] C. Bailer, A. Pagani, and D. Stricker. A user supported object tracking framework for interactive video production. *Journal of Virtual Reality and Broadcasting*, 11(2014):9, 2014.

[5] L. Cehovin, M. Kristan, and A. Leonardis. Robust visual tracking using an adaptive coupled-layer visual model. *Pat-*

	illumination change			motion change			size change			occlusion			camera motion		
Name	Succ.	+/-	Gain	Succ.	+/-	Gain	Succ.	+/-	Gain	Succ.	+/-	Gain	Succ.	+/-	Gain
MIL	0.4	0.86	-0.34	0.41	1.04	0.16	0.37	1.25	1.14	0.29	1.72	1.68	0.41	1.13	0.42
CT	0.39	0.46	-2.62	0.42	0.52	-2.37	0.36	0.59	-2.24	0.43	0.4	-4.5	0.44	0.63	-1.76
PTp	0.38	0.38	-5.42	0.46	0.63	-2.47	0.39	0.66	-1.99	0.41	0.8	-0.8	0.46	0.72	-1.65
IMPNCC	0.53	3.41	6.19	0.51	2.2	3.55	0.45	2.79	5.14	0.4	1.4	1.2	0.5	2.14	3.4
FSDT	0.53	2.44	3.87	0.51	1.38	1.44	0.44	1.66	2.7	0.54	1.04	0.19	0.53	1.64	2.21
IPRT	0.43	0.81	-0.8	0.48	0.91	-0.47	0.43	0.86	-0.91	0.46	1.18	0.81	0.48	0.98	-0.07
LGTv1	0.48	1.86	2.7	0.48	1.47	1.68	0.45	1.81	2.76	0.34	2.23	2.05	0.47	1.32	1.13
IIVTv2	0.47	0.72	-1.36	0.5	0.9	-0.43	0.45	0.99	-0.03	0.53	1.55	2.16	0.52	1.02	0.1
IVT	0.57	1.03	0.17	0.49	0.71	-1.81	0.41	1.37	1.6	0.4	1.36	2.04	0.48	0.78	-1.31
FRT	0.48	3.48	5.49	0.49	1.32	1.24	0.41	0.9	-0.59	0.52	0.76	-1.53	0.51	1.4	1.65
SIRPF	0.48	0.91	-0.5	0.49	1.06	0.33	0.42	1.1	0.58	0.56	1.11	0.62	0.54	0.96	-0.26
BDF	0.55	0.27	-10.4	0.52	0.55	-3.6	0.43	0.68	-2.4	0.5	1.09	0.57	0.5	0.49	-4.1
ThunderS.	0.46	0.21	-10.23	0.5	0.48	-4.16	0.42	0.57	-3.29	0.57	0.63	-3.04	0.53	0.5	-4.22
CMT	0.51	0.54	-4.9	0.5	1.05	0.24	0.46	1.61	2.74	0.52	1.35	1.73	0.52	0.85	-0.91
ABS	0.45	0.57	-2.53	0.54	1.37	1.9	0.49	1.87	3.93	0.46	1.74	2.39	0.52	1.28	1.31
FoT	0.56	1.31	1.54	0.56	1.71	2.82	0.51	2.1	4.26	0.49	3.53	5.03	0.51	1.28	1.19
MatFlow	0.52	0.15	-13.55	0.53	0.49	-4.23	0.45	0.69	-2.23	0.55	0.65	-2.61	0.53	0.39	-5.44
DynMS	0.51	3.08	4.1	0.54	1.64	2.34	0.47	1.85	3.49	0.49	1.93	2.49	0.55	1.8	2.89
Matrioska	0.51	0.26	-10.25	0.52	0.52	-3.54	0.42	0.6	-2.56	0.56	0.73	-1.67	0.55	0.48	-4.44
aStruck	0.62	3.07	6.71	0.51	1.26	1.24	0.41	0.81	-1.14	0.44	0.81	-0.95	0.52	1.3	1.39
EDFT	0.57	0.29	-11.48	0.54	0.58	-3.8	0.47	0.9	-0.85	0.52	2.23	4.13	0.53	0.51	-4.51
Struck	0.52	0.26	-9.77	0.53	0.53	-3.89	0.42	0.51	-3.88	0.59	1.01	0.06	0.55	0.55	-3.95
MCT	0.59	1.5	2.4	0.55	1.24	1.17	0.47	1.51	2.27	0.52	1.44	1.85	0.56	1.51	2.2
ACT	0.56	0.16	-15.47	0.56	0.58	-3.59	0.47	0.98	-0.19	0.54	0.9	-0.64	0.56	0.52	-4.45
VTDMG	0.44	0.33	-5.08	0.54	0.65	-2.87	0.47	0.86	-1	0.57	1	0.01	0.55	0.68	-2.43
NCC	0.56	4.81	8.28	0.55	1.44	2.04	0.47	1	-0.01	0.61	2.73	7.27	0.59	1.83	3.6
PLT13	0.53	0.56	-2.97	0.56	0.92	-0.48	0.48	0.98	-0.12	0.58	0.78	-1.47	0.56	0.76	-1.59
qwsEDFT	0.58	0.24	-11.56	0.56	0.45	-5.42	0.48	0.48	-5.79	0.58	0.9	-0.74	0.56	0.47	-4.94
eASMS	0.47	1.77	2.75	0.56	1.37	2.17	0.51	1.42	2.84	0.57	0.89	-0.81	0.56	1.33	1.75
PLT14	0.52	0.21	-11.25	0.58	0.88	-0.93	0.51	1.2	1.42	0.6	1.15	0.62	0.57	0.61	-3.42
ACAT	0.63	2.12	6.92	0.58	1.65	3.35	0.5	2.42	5.04	0.49	1.61	2.5	0.57	1.91	4.14
OGT	0.57	4.17	7.93	0.58	1.93	3.92	0.52	2.13	5	0.51	1.24	1.18	0.57	1.78	3.52
LT-FLO	0.62	6.26	9.12	0.58	2.54	4.44	0.5	2.26	4.11	0.49	2.18	3.3	0.57	2.43	4.52
DGT	0.49	1.03	0.13	0.6	2.28	5.23	0.58	3.52	9.03	0.49	2.26	4.41	0.58	2.31	4.86
HMMTxD	0.6	4.33	8.84	0.62	2.49	6.73	0.55	2.29	6.64	0.6	2.38	5.48	0.62	2.39	6.01
DSST	0.76	5.65	12.53	0.66	2.93	7.12	0.53	2.8	6.72	0.65	2.25	4.69	0.67	3.34	8.04
SAMF	0.68	3.23	9.83	0.68	2.38	7.1	0.57	2.09	6.2	0.61	0.92	-0.52	0.67	2.21	6.76
KCF	0.76	6.5	14.88	0.68	2.28	7.28	0.58	2.37	8.41	0.64	1.96	5.88	0.68	2.27	7.35
Fusion	0.69	-		0.69	-		0.61	-		0.66	-		0.68	-	
Fusion+	0.77	-		0.74	-		0.68	-		0.78	-		0.73	-	

Table 4: Our results on the VOT2014 data under different conditions. See text for details. Colors are form full red (worst) to full green (best). Cyan is used for fusion results that outperform the best tracking results.

tern Analysis and Machine Intelligence, IEEE Transactions on, 35(4):941–953, 2013.

[6] R. T. Collins. Mean-shift blob tracking through scale space. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, pages II–234. IEEE, 2003.

[7] M. Danelljan, G. Häger, F. Khan, and M. Felsberg. Accurate scale estimation for robust visual tracking. In *British Machine Vision Conference, Nottingham, September 1-5, 2014*. BMVA Press, 2014.

[8] S. Hare, A. Saffari, and P. H. Torr. Struck: Structured output tracking with kernels. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 263–270. IEEE, 2011.

[9] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista. High-speed tracking with kernelized correlation filters. *Pattern*

	main supporters				main competitors			
Method	1.supporter	Δ gain	2.supporter	Δ gain	2.competitor	Δ gain	1.competitor	Δ gain
ABS	IPRT	0.412	ACT	0.384	SAMF	-0.772	KCF	-0.857
ACAT	SAMF	1.25	HMMTxD	1.05	MIL	-0.343	EDFT	-0.592
ACT	PLT13	0.806	EDFT	0.797	KCF	-0.626	DSST	-0.771
CT	IPRT	0.457	ACT	0.306	SAMF	-0.695	IMPNCC	-0.7
DSST	KCF	1.09	ACAT	1.02	qwsEDFT	-0.98	EDFT	-1.16
DynMS	aStruck	0.366	ACAT	0.33	CT	-0.539	ThunderStruck	-0.788
FSDT	PLT14	0.531	DGT	0.405	PTp	-0.266	IIVTv2	-0.376
IIVTv2	ACAT	0.353	ACT	0.303	OGT	-0.499	PTp	-0.614
IMPNCC	DSST	1.01	NCC	0.778	EDFT	-0.672	qwsEDFT	-0.93
IVT	Matrioska	0.341	FSDT	0.308	DSST	-0.716	KCF	-0.953
KCF	NCC	1.08	DSST	1.06	CMT	-1.07	EDFT	-1.16
MCT	ACAT	0.575	PLT14	0.34	qwsEDFT	-0.365	KCF	-0.585
MIL	ACT	0.405	ABS	0.267	LT-FLO	-0.705	NCC	-0.886
NCC	KCF	1.34	HMMTxD	1.19	Struck	-0.941	EDFT	-1.2
OGT	HMMTxD	0.792	NCC	0.618	Matrioska	-0.765	qwsEDFT	-0.862
PLT13	Matrioska	0.732	ACT	0.662	LGTv1	-0.686	NCC	-0.716
PLT14	Matrioska	0.672	ThunderStruck	0.572	IMPNCC	-0.519	LGTv1	-0.77
PTp	ACAT	0.378	HMMTxD	0.294	DGT	-0.597	VTDMG	-0.714
SAMF	ACAT	1.27	aStruck	0.655	EDFT	-1.13	VTDMG	-1.46
SIRPF	VTDMG	0.642	BDF	0.619	IMPNCC	-0.292	FoT	-0.665
VTDMG	PLT13	0.647	qwsEDFT	0.621	KCF	-0.815	SAMF	-1.41
aStruck	FRT	0.886	PLT14	0.653	Struck	-0.65	MIL	-0.672
BDF	SIRPF	0.581	eASMS	0.387	SAMF	-0.683	KCF	-1.31
CMT	qwsEDFT	0.64	VTDMG	0.562	SAMF	-0.934	KCF	-1.27
DGT	NCC	0.708	ACT	0.57	SAMF	-0.643	ABS	-0.844
eASMS	ACAT	0.752	ACT	0.603	DGT	-0.522	SAMF	-0.688
EDFT	ACT	0.87	CMT	0.565	KCF	-1.14	DSST	-1.16
FoT	DynMS	0.334	aStruck	0.3	MIL	-0.533	ACT	-0.593
FRT	aStruck	0.849	NCC	0.541	ABS	-0.458	VTDMG	-0.459
HMMTxD	ACAT	1.1	NCC	0.729	ThunderStruck	-0.312	PLT14	-0.369
IPRT	ACT	0.532	SIRPF	0.497	KCF	-0.415	SAMF	-0.505
LGTv1	ACAT	0.66	NCC	0.467	PLT13	-0.678	PLT14	-0.757
LT-FLO	KCF	0.517	DSST	0.462	MIL	-0.691	EDFT	-0.805
MatFlow	PLT13	0.667	SIRPF	0.602	IMPNCC	-0.691	KCF	-0.876
Matrioska	PLT13	0.963	PLT14	0.759	KCF	-0.81	SAMF	-0.948
qwsEDFT	ThunderStruck	0.708	VTDMG	0.691	SAMF	-1.07	KCF	-1.08
Struck	ACAT	0.601	SIRPF	0.54	NCC	-0.884	ThunderStruck	-0.949
ThunderStruck	qwsEDFT	0.686	PLT14	0.566	NCC	-0.777	Struck	-0.963

Table 5: Main supporters and contributors of trackers on the VOT2014 results. See text for details.

Analysis and Machine Intelligence, IEEE Transactions on, 37(3):583–596, 2015.

[10] M. Kristan, R. Pflugfelder, A. Leonardis, J. Matas, L. Čehovin, G. Nebehay, T. Vojíř, G. Fernandez, A. Lukežič, A. Dimitriev, et al. The visual object tracking vot2014 challenge results. In *Computer Vision-ECCV 2014 Workshops*, pages 191–217. Springer, 2014.

[11] M. Kristan, R. Pflugfelder, A. Leonardis, J. Matas, P. Fatih, L. Čehovin, G. Nebehay, G. Fernandez, et al. The visual object tracking vot2013 challenge results. In *Computer Vision Workshops (ICCV Workshops), 2013 IEEE International Conference on*. IEEE, 2013.

[12] H. Nam, S. Hong, and B. Han. Online graph-based tracking. In *Computer Vision–ECCV 2014*, pages 112–126. Springer, 2014.

[13] G. Nebehay and R. Pflugfelder. Consensus-based matching and tracking of keypoints for object tracking. In *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*, pages 862–869. IEEE, 2014.

[14] A. Wendel, S. Sternig, and M. Godec. Robustifying the flock of trackers. In *16th Computer Vision Winter Workshop. Citeseer*, page 91. Citeseer, 2011.

[15] Y. Wu, J. Lim, and M.-H. Yang. Online object tracking: A benchmark. In *Computer vision and pattern recognition (CVPR), 2013 IEEE Conference on*, pages 2411–2418. IEEE, 2013.

[16] W. Zhong, H. Lu, and M.-H. Yang. Robust object tracking via sparsity-based collaborative model. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1838–1845. IEEE, 2012.