The VOT2015 and VOT-TIR2015 Challenges Submission Report

Yang Hua, Karteek Alahari, and Cordelia Schmid

Inria Grenoble Rhône-Alpes, France

firstname.lastname@inria.fr

In this report, we present the details of our submission and the evaluation results on both the visual object tracking (VOT) 2015 and the thermal infrared visual object tracking (VOT-TIR) 2015 challenges. The goal of these challenges is to compare short-term model-free single-object trackers, and serve as the de factor state-of-the-art evaluation platform for visual object tracking. In particular, the VOT challenge focuses on natural RGB video sequences with rotated rectangle ground truth boxes, while the VOT-TIR challenge consists of thermal infrared video sequences with axis-aligned ground truth boxes, see examples in Figures 1 and 2.

For more details of the two challenges, we refer the reader to the official challenge reports [3, 1].

1 Description of the tracker

We submitted a simplified version of our proposal-selection tracker, referred as to sPST. Compared to the full version of our proposal-selection tracker, described in [2] we excluded geometry proposals and motion boundaries selection in sPST, due to the computational cost of the optical flow method. Similar to the full version of the proposal-selection tracker, sPST proceeds in two stages – proposal followed by selection. In the proposal stage, we generate a set of candidates computed by the tracking-by-detection framework, where we use the frame as is, or rotate it according to the ground truth annotation in the initial frame to handle rotated bounding box annotation. In the selection stage, we determine the best candidate as the tracking result with detection and edgebox scores. We follow the two-phase selection strategy to combine these two cues, as described in [2]. It is worth noting that in order to show the generality of sPST, we set identical parameters for both VOT2015 and VOT-TIR2015 challenges, despite the target video domains of these two challenges being different.

Figure 1: Illustration of VOT2015 challenge dataset [3], showing the first frame in each sequence along with the initial bounding box of the target object.

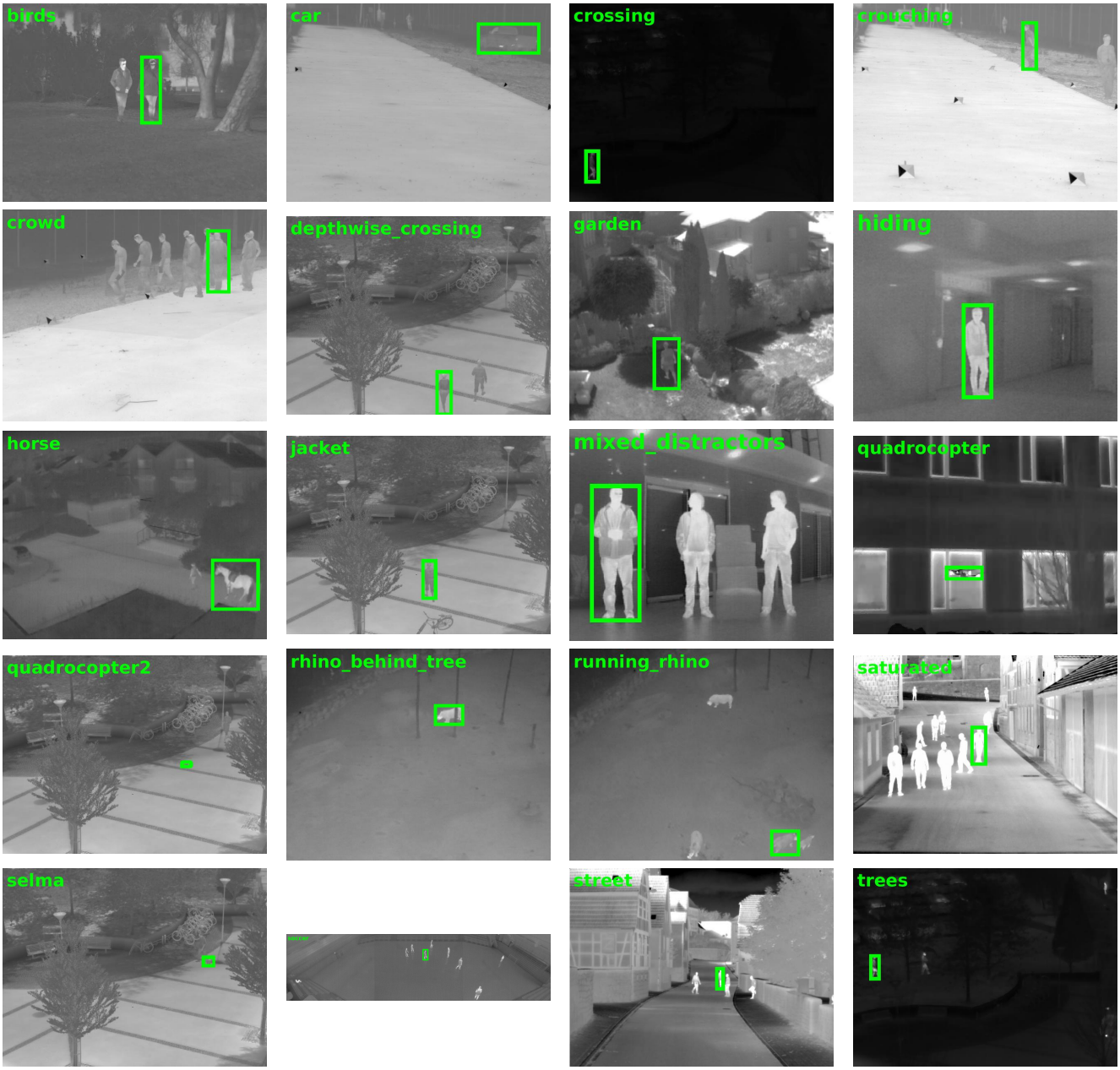

Figure 2: Illustration of VOT-TIR2015 dataset [1], showing the first frame in each sequence along with the initial bounding box of the target object.

1.1 Experimental environment

We implemented sPST with Matlab 2014b and mex files. For evaluation on the challenge datasets, we followed the guidelines, and integrated sPST into the VOT challenge toolkit. [1] We performed all the experiments on a workstation with an Intel Xeon CPU at 2.4GHz and 48G memory, running Fedora 21 64bit operation system.

1.2 Implementation details

We adopted all the parameters of sPST from which are fixed or calculated based on the ground truth annotation in the first frame of all the sequences in VOT2015 and VOT-TIR2015 challenges.

Object template and HOG feature. The object template and HOG feature parameters are set according to the area of the ground truth bounding box in the initial frame. If the area of the bounding box is larger than 10000 pixels, the object template resize scale is set to 0.5. If the area of the bounding box is between 400 pixels and 10000 pixels, the object template resize scale is set to 0.8. If the area of the bounding box is smaller than 400 pixels, the object template resize scale is set to 1.0.

After resizing, if the area of object template is larger than 4000 pixels, the cell size of HOG feature is set to 8. If the area of object template is between 1000 pixels and 4000 pixels, the cell size is set to 6. For an area less than 1000 pixels, the cell size is set to 4.

Detector. The initial detector is trained in the first frame with one positive sample and several negative examples that have less than 50% overlap with the ground truth annotation. In order to make our experimental results repeatable, we fixed the training sample order randomly to learn the SVM. In every frame that follows, the detector is evaluated at seven scales: $\{0.980, 0.990, 0.995, 1.000, 1.005, 1.010, 1.020\}$ with dense-scanning at a step size of 2 pixels. The detector is updated with the tracking result every frame, except when the result in a frame is very similar to that in the previous frame (i.e., the normalized cross-correlation score between current and previous frame results is larger than 0.95).

Candidate proposals. The top 5 detection results are added to the candidate pool. Moreover, if the ground truth bounding box in the initial frame is rotated by more than 15 degrees (clockwise or anti-clockwise), another top 5 detection results on the rotated image are added to enrich the pool.

Candidate selection. We adopted the two-phase selection strategy, discussed in in the selection stage. First, we check the normalized detection confidence score of all proposals. When the detection scores of some or all the proposals are statistically similar (i.e., the differences between these detection scores and the maximum detection score are less than 1% of the maximum score), we collect all these similar proposals for the following selection step. Otherwise, we choose the proposal with the maximum detection score as tracking result. Second, we check the edgebox scores

[1] Available at https://github.com/votchallenge/vot-toolkit.

[4] of remaining proposals. If all the edgebox scores are less than 0.075 or are not comparable to the mean of the last five edgebox scores of the tracking predictions, which implies a low quality score, we select the candidate with the highest detection score. Otherwise, we choose the proposal with the highest edgebox score as tracking result.

Handling small bounding box. If the initial ground truth box contains less than 300 pixels, we still train the detector as usual. But before evaluating the detector in the new frame, we check the pixel difference between current and previous frames. If more than 40 percent of pixels in the search region are changed, we apply the ordinary proposal-selection scheme to determinate tracking result. Otherwise, we set the region, which has the same size as the previous tracking box and contains the largest percent of changed pixels, as tracking result.

2 Evaluation results

According to [3], sPST was ranked sixth among 62 trackers in the VOT2015 challenge. For the VOT-TIR2015 challenge [1], sPST was ranked second among 24 trackers and received the "winning tracker" title.

2.1 The performance of sPST on VOT2015

All the raw results of each sequence in the VOT2015 dataset are generated by VOT challenge toolkit, and are shown in Table 1. According to these results from the toolkit, the average accuracy of sPST is 0.54, the average number of failures is 1.42, and it runs at 5.80 fps on average.

	Overlap	**Failures**	**Speed**
bag	0.45	1.00	4.14
ball1	0.85	0.00	2.24
ball2	0.56	0.00	1.93
basketball	0.63	0.00	7.71
birds1	0.42	2.00	3.04
birds2	0.56	0.00	5.54
blanket	0.66	0.00	13.23
bmx	0.35	0.00	2.49
bolt1	0.70	1.00	7.10
bolt2	0.68	0.00	9.93
book	0.36	5.00	3.45
butterfly	0.25	0.00	4.17
car1	0.67	2.00	8.15
car2	0.83	0.00	15.74
crossing	0.70	0.00	3.15
dinosaur	0.48	2.00	6.83

fernando	0.43	1.00	5.25
fish1	0.42	4.00	6.90
fish2	0.38	4.00	4.33
fish3	0.58	0.00	7.79
fish4	0.40	1.00	10.86
girl	0.68	1.00	7.36
glove	0.62	3.00	2.47
godfather	0.37	1.00	7.92
graduate	0.55	3.00	9.93
gymnastics1	0.49	4.00	8.83
gymnastics2	0.57	3.00	2.14
gymnastics3	0.28	3.00	1.31
gymnastics4	0.42	1.00	2.91
hand	0.52	5.00	6.76
handball1	0.62	3.00	9.60
handball2	0.52	3.00	4.24
helicopter	0.41	0.00	5.97
iceskater1	0.46	2.00	6.68
iceskater2	0.57	2.00	7.01
leaves	0.24	5.00	1.44
marching	0.71	0.00	3.58
matrix	0.56	2.00	4.58
motocross1	0.56	1.00	4.81
motocross2	0.34	2.00	0.98
nature	0.33	5.00	5.91
octopus	0.57	0.00	4.96
pedestrian1	0.68	1.00	7.67
pedestrian2	0.45	0.00	9.49
rabbit	0.22	6.00	3.24
racing	0.58	0.00	5.28
road	0.63	1.00	3.97
shaking	0.78	0.00	7.42
sheep	0.58	0.00	9.08
singer1	0.70	0.00	6.62
singer2	0.76	1.00	6.27
singer3	0.24	0.00	4.38
soccer1	0.39	2.00	5.81
soccer2	0.63	1.00	2.39
soldier	0.50	1.00	1.42
sphere	0.69	0.00	6.78

tiger	0.78	0.00	5.72
traffic	0.87	0.00	2.64
tunnel	0.71	0.00	9.49
wiper	0.74	0.00	6.85
mean	0.54	1.42	5.80

Table 1: The performance of sPST on the VOT2015 challenge dataset.

2.2 The performance of sPST on VOT-TIR2015

All the raw results of each sequence in the VOT-TIR2015 dataset, which are generated by the VOT challenge toolkit, are shown in Table 2. According to these results from the toolkit, the average accuracy of sPST is 0.70, the average number of failures is 0.35 and it runs at 11.07 fps on average.

	Overlap	**Failures**	**Speed**
birds	0.74	0.00	6.64
car	0.55	0.00	13.02
crossing	0.85	0.00	8.80
crouching	0.67	0.00	10.05
crowd	0.78	0.00	3.98
depthwise_crossing	0.72	0.00	9.58
garden	0.65	3.00	14.69
hiding	0.66	0.00	14.57
horse	0.74	0.00	12.70
jacket	0.82	0.00	11.73
mixed_distractors	0.74	0.00	7.16
quadrocopter	0.52	1.00	5.88
quadrocopter2	0.54	0.00	20.38
rhino_behind_tree	0.71	0.00	22.83
running_rhino	0.54	1.00	22.95
saturated	0.79	0.00	6.72
selma	0.74	0.00	7.08
soccer	0.59	0.00	4.44
street	0.75	1.00	6.65
trees	0.81	1.00	11.49
mean	0.70	0.35	11.07

Table 2: The results of the VOT-TIR2015 challenge for our sPST tracker.

As shown in 2, thermal infrared (TIR) video data shows more clear edge responses than RGB video data. To highlight the usefulness of this cue for tracking, we evaluated our sPST tracker on

VOT-TIR2015 without two-phase selection, i.e., the tracking result in each frame was determined only by the detection score. This results in 0.67 for the average accuracy and 0.35 for the average number of failures, both of which are inferior to sPST with two-phase selection (0.70 and 0.35 respectively).

References

[1] M. Felsberg, A. Berg, G. Hager, J. Ahlberg, M. Kristan, J. Matas, A. Leonardis, L. Čehovin, G. Fernandez, T. Vojíř, G. Nebehay, R. Pflugfelder, et al. The thermal infrared visual object tracking vot-tir2015 challenge results. In *ICCV Workshop on Visual Object Tracking Challenge*, 2015.

[2] Y. Hua, K. Alahari, and C. Schmid. Online object tracking with proposal selection. In *ICCV*, 2015.

[3] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, L. Čehovin, G. Fernandez, T. Vojíř, G. Hager, G. Nebehay, R. Pflugfelder, et al. The visual object tracking VOT2015 challenge results. In *ICCV Workshop on Visual Object Tracking Challenge*, 2015.

[4] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, 2014.