The new VOT2020 short-term tracking performance evaluation protocol and measures

Matej Kristan, Alan Lukežič, Martin Danelljan, Luka Čehovin Zajc, Jiri Matas

***Abstract*—This document is a brief overview of the new performance evaluation protocol and the new measures introduced in Visual Object Tracking challenge VOT2020 for evaluation of short-term trackers. The new protocol avoids tracker-dependent resets and reduces the variance of the performance evaluation measures – the VOT basic measures, accuracy (A) and robustness (R) and the primary performance measure expected average overlap (EAO) are re-defined. The presented performance evaluation methodology should be considered as the new default VOT methodology for testing short-term trackers in VOT2020 and the challenges to follow.**

Errata: This is an updated version of the document originally published at the VOT page on 31.3.2020. In effort to keep the new measures as similar as possible to those calculated in the previous VOT challenges, the measures have been slightly modified since initial post. The VOT toolkit has been updated with the modification in the measures as well. In particular, the robustness measure did not change, the accuracy measure normalization is corrected and in the EAO measure, all sub-sequences are equally weighted.

I. Introduction

Over the last eight years, the Visual Object Tracking initiative (VOT) has been gradually developing performance evaluation methodology for testing short- and long-term trackers. The overall guideline was developing interpretable measures that probe various tracking properties. Initially VOT [1] considered only short-term trackers, and based on the analysis later published in [2], [3], two basic performance measures were chosen: accuracy and robustness. The goal was to promote trackers that well approximate the target position, and even more importantly, do not fail very often. The first methodology introduced in VOT2013 [1] was based on ranking trackers along each measure and averaging the ranks. Due to a reduced interpretation power and dependency of ranks on the tested trackers, this approach was replaced in VOT2015 [4] by the expected average overlap measure (EAO), which principally combines the individual basic measures.

The VOT measures have promoted development of robust short-term trackers and with increased robustness of modern trackers, a drawback of the reset-based evaluation protocol has emerged. In the VOT performance evaluation protocol a tracker is initialized in the first frame and whenever the overlap between the reported and the ground truth target location (i.e., bounding box) falls to zero, a failure is detected and the tracker is reset a fixed number of frames later. The robustness is measured as the number of times the tracker is reset and the accuracy is the average overlap between the periods of successful tracking. This setup reflects the tracker performance in a practical application, where the task is to track the target throughout the sequence, either automatically, or by user intervention, i.e., a tracker reset. Furthermore, this approach enables utilization of all sequence frames in the evaluation.

However, a point of tracking failure will affect the point of reset (tracker re-initialization) and initialization points profoundly affect the tracking performance. With recent development of very robust trackers, the initialization points started to play a significant role in the final tracker ranking. In particular, we have noticed that initialization at some frame might result in another failure later on in the sequence, while initializing a few frames later might not. This allows a possibility (although not trivially) for fine-tuning the tracker to fail on more *favorable* frames and by that reducing the failure rate and increase the overall apparent robustness as measured by the reset-based protocol.

Another potential issue of the existing VOT reset-based protocol is the definition of a tracking failure. A failure is detected whenever the overlap between the prediction and ground truth falls to zero. Since resets directly affect the performance, a possible way to reduce the resets is to increase the predicted bounding box size, so to avoid the zero overlap. While we have not observed such *gaming* often, there were a few cases in the last seven challenges where the trackers attempted this and one of the trackers has been disqualified upon identifying the use of the bounding box inflation strategy. But some trackers did resort to reporting a slightly larger bounding box due to the strictness of the failure protocol – the tracker will be reset if the zero overlap is detected in a single frame, even if the tracker would have jumped right back on the target in the next frame. We call this a short-term failure and the current protocol does not distinguish between trackers robust to short-term failures and trackers that fail completely.

In VOT2020 we have thus decided to propose a new performance evaluation protocol that mitigates the drawbacks we have been identified over the years. A change of protocol affects the performance measures as well. We believe the new performance evaluation protocol and the new measures maintain the desired properties for tracker evaluation that the VOT has been promoting throughout its existence and reduce the biases discussed above. In the following we overview the new reset-based evaluation protocol in Section II, Section III overviews the new measures and Section IV concludes the document.

II. The short-term tracking evaluation protocol

The main drawback of the existing VOT short-term performance evaluation protocol are the tracker-dependent resets,

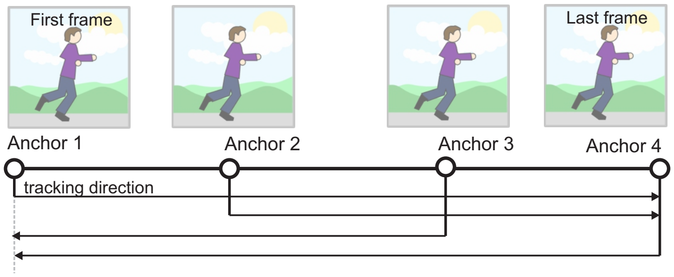

Fig. 1. Anchors are placed 50 frames apart. At each anchor the tracker is initialized and tracks in the direction that yields the longest subsequence.

which induce a causal correlation between the first reset and the later ones. To avoid this, the notion of reset is replaced by *initialization points*, which are made equal for all trackers in the new protocol. In particular, on each sequence, initialization points, referred here as *anchors*, are placed Δ_{anc} frames apart, with the first and last anchor on the first and the last frame, respectively. A tracker is run from *each* anchor forward or backward in the sequences, whichever direction yields the longest subsequence. For example, if the anchor is placed before the middle of the sequence, the tracker is run forward, otherwise backward in the sequence. Each anchor is manually checked and potentially moved by a few frames to avoid placing the initialization point on an occluded target. Figure 1 shows example of the anchor placement and the tracking direction.

The distance between the anchors was set to $\Delta_{\text{anc}} = 50$. At approximately 25 frames per second, this amounts to 2 second distances. We have experimentally tested that this value delivers stable results for the measures described in the next section computed on typical-length short-term sequences, while keeping the computational complexity of the evaluation at a moderate level.

III. Performance measures

Like in previous VOT challenges, we use the accuracy and robustness as the basic measures to probe tracking performance and the overall performance is summarized by the expected average overlap (EAO).

A. The new accuracy and robustness measures

On a subsequence starting from an anchor a of sequence s, the accuracy $A_{s,a}$ is defined as the average overlap between the target predictions and the ground truth calculated from the frames before the tracker fails on that subsequence, i.e.,

$$A_{s,a} = \frac{1}{N_{s,a}^F} \sum\nolimits_{i=1:N_{s,a}^F} \Omega_{s,a}(i), \tag{1}$$

where $N_{s,a}^F$ is the number of frames before the tracker failed in the subsequence starting at anchor a in the sequence s (see Section III-C for the failure definition) and $\Omega_{s,a}(i)$ is the overlap between the prediction and the ground truth at frame i. The new robustness measure $R_{s,a}$ is defined as the *extent* of the sub-sequence before the tracking failure, i.e.,

$$R_{s,a} = N_{s,a}^F / N_{s,a}, \tag{2}$$

where $N_{s,a}$ is the number of frames of the subsequence.

The results from the sub-sequences are averaged in a weighted fashion such that each sub-sequence contributes proportionally to the number frames used in calculation of each measure. In particular, the per-sequence accuracy and robustness are defined as

$$A_s = \frac{1}{\sum_{a=1:N_s^A} N_{s,a}^F} \sum\nolimits_{a=1:N_s^A} A_{s,a} N_{s,a}^F, \tag{3}$$

$$R_s = \frac{1}{\sum_{a=1:N_s^A} N_{s,a}} \sum\nolimits_{a=1:N_s^A} R_{s,a} N_{s,a}, \tag{4}$$

where N_s^A is the number of anchors in the sequence s. The overall accuracy and robustness are calculated by averaging the per-sequence counterparts proportionally to the number of frames used for their calculation, i.e.,

$$A = \frac{1}{\sum_{s=1:N} N_s^F} \sum\nolimits_{s=1:N} A_s N_s^F, \tag{5}$$

$$R = \frac{1}{\sum_{s=1:N} N_s} \sum\nolimits_{s=1:N} R_s N_s, \tag{6}$$

where N is the number of sequences in the dataset, N_s is the number of frames in sequence s and $N_s^F = \sum_{a=1:N_s^A} N_{s,a}^F$ is the number of frames used to calculate the accuracy in that sequence.

B. The new EAO measure

As in previous VOT challenges, the accuracy and robustness are principally combined into a single performance score called the expected average overlap (EAO). We use the same approach as in the previous VOT challenges, i.e., the expected average overlap curve is calculated and averaged over an interval of typical short-term sequence lengths into the EAO measure.

Note that the computation considers virtual sequences of overlaps generated from the sub-sequence results. In particular, if a tracker failed on a sub-sequence (s, a), the overlap falls to zero at the failure frame, and the overlaps can be extended to i-th frame by zeros, even if i exceeds the sub-sequence length. But if the tracker did not fail, the overlaps cannot be extrapolated beyond the original sub-sequence length.

The value of the EAO curve $\hat{\Phi}_i$ at sequence length i is thus defined as

$$\hat{\Phi}_i = \frac{1}{|\mathcal{S}(i)|} \sum_{s,a \in \mathcal{S}(i)} \Phi_{s,a}(i), \tag{7}$$

where $\Phi_{s,a}(i)$ is the average overlap calculated between the first and i-th frame of the extended sub-sequence starting at anchor a of sequence s, $\mathcal{S}(i)$ is the set of the extended sub-sequences with length greater or equal to i and $|\mathcal{S}(i)|$ is the number of these sub-sequences.

The EAO measure is then calculated by averaging the EAO curve from N_{lo} to N_{hi}, i.e.,

$$EAO = \frac{1}{N_{\text{hi}} - N_{\text{lo}}} \sum\nolimits_{i=N_{\text{lo}}:N_{\text{hi}}} \hat{\Phi}_i. \tag{8}$$

Similarly to VOT2015 [5], the interval bounds $[N_{\text{lo}}, N_{\text{hi}}] = [115, 755]$ were determined on the VOT2020 dataset to reflect the range of short-term sequence lengths typical for modern datasets.

C. Failure definition

The tracking failure event is also redefined to (i) reduce the potential for the *gaming*, i.e., outputting the entire image as the prediction to prevent failure detection during an uncertain tracking phase, and (ii) allow for recovery from short-term tracking failures.

A *tentative* failure is detected when the overlap falls below a non-zero threshold θ_Φ. The non-zero threshold punishes an actual drift from the target as well as speculation by outputting a very large bounding box to prevent failure detection. If a tracker does not recover within the next θ_N frames, i.e., the overlap does increase to over θ_Φ, a failure is detected.

By using several well-known trackers from different tracker design classes we experimentally determined that the threshold values $\theta_\Phi = 0.1$ and $\theta_N = 10$ reduce the *gaming* potential, allow recoveries from short-term failures, while still penalizing the trackers that fail more often.

IV. Conclusion

This document described the new performance evaluation protocol and the measures introduced in the VOT2020 short-term tracking evaluation. The new protocol and the measures build on the good practices developed in the previous challenges. Based on our own observations from analysing the VOT challenge results for the last seven years and the feedbacks from the community, we identified several drawbacks in the existing VOT performance evaluation. The new measures are a carefully-thought-out attempt to maintain the benefits of the previous VOT measures, while mitigating their drawbacks.

The protocol and the measures presented in this document should be considered as the default VOT short-term tracking performance evaluation methodology from now on and the default methodology used in the new VOT2020 toolkit short-term tracking sub-challenges (short-term, real-time and RGBT challenge). The short-term performance evaluation methodology from previous VOT challenges is now considered obsolete.

References

[1] M. Kristan, R. Pflugfelder, A. Leonardis, J. Matas, F. Porikli, L. Čehovin, G. Nebehay, G. Fernández, T. Vojíř, and et al., “The visual object tracking vot2013 challenge results,” in *ICCV2013 Workshops, Workshop on visual object tracking challenge*, 2013, pp. 98 –111.

[2] L. Čehovin, A. Leonardis, and M. Kristan, “Visual Object Tracking Performance Measures Revisited,” *IEEE TIP*, vol. 25, no. 3, pp. 1261–1274, 2016.

[3] M. Kristan, J. Matas, A. Leonardis, T. Vojíř, R. Pflugfelder, G. Fernández, G. Nebehay, F. Porikli, and L. Čehovin, “A novel performance evaluation methodology for single-target trackers,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 11, pp. 2137–2155, 2016.

[4] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, L. Čehovin, G. Fernández, T. Vojíř, G. Häger, G. Nebehay, R. Pflugfelder, and et al., “The visual object tracking vot2015 challenge results,” in *ICCV2015 Workshops, Workshop on visual object tracking challenge*, 2015.

[5] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, L. Čehovin, G. Fernández, T. Vojíř, 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,” in *Visual Object Tracking Workshop 2015 at ICCV2015*, Dec 2015.