The VOTS2023 Challenge Performance Measures

Matej Kristan[1], Jiří Matas[2], Martin Danelljan[8], Luka Čehovin Zajc[1], and Alan Lukežič[1]

[1]University of Ljubljana, Slovenia
[2]Czech Technical University, Czech Republic
[8]ETH Zurich, Switzerland

Abstract

This draft describes the performance measures used in the VOTS2023 challenge organized in conjuction with the ICCV2023 VOTS2023 workshop. Please see further details on the challenge website[1].

1. Performance evaluation protocol

The VOTS2023 challenge requires tracking one or more targets simultaneously by segmentation over long or short sequences, while the targets may disappear during tracking and reappear later in the video. The targets may be whole instances or only their parts. The tracker is initialized in the first frame on all specified targets. For each subsequent frame, the tracker is required to report the locations for all visible targets in that frame. Specifically, a segmentation mask is required for each visible target, while a "not present" label is reported for the absent targets. The tracker is then evaluated with the new performance measures presented in the following.

1.1. VOTS performance measures

The goal of a multi-target tracker is to reliably track each individual target selected in the first frame. Drifting off a target to background or another target is equally considered as failed tracking. This allows definition of per-target performance measures, which are averaged over all targets to obtain the final score.

From perspective of tracking a single target, five scenarios visualized in Figure 1 emerge. Three scenarios cover cases with the target present: target successfuly localized (sc1), tracker drift (sc2), target incorrectly predicted as absent (sc3). Two scenarios cover the cases with the target absent: target predicted as present (sc4), and target predicted as absent (sc5). In the following we introduce performance measures based on the notion of tracking success that take all these scenarios into account.

Tracking of i-th target on n-th frame of the sequence s is considered successful if the predicted target location and the ground truth (i.e., segmentation masks) match sufficiently well. The success is measured by an intersection-over-union (IoU), binarized by some threshold θ (i.e., 1 for values greater than θ, and 0 otherwise). Note that the IoU generalizes well to the case with target absent – if the tracker reports the empty mask in this case (i.e., *target absent flag*), it receives the IoU=1, since the reported mask is in total agreement with the ground truth, otherwise the IoU=0. The overall tracking success for the considered target at threshold θ is thus defined as

$$S(\theta) = \frac{1}{N} \sum_{s=1:N} \frac{1}{T_s N_s} \sum_{i=1:T_s} \sum_{n=1:N_s} [o_{sin} > \theta], \qquad (1)$$

where T_s and N_s are the number of targets and frames[2] in the sequence s, N is the number of sequences and $[o_{sin} > \theta]$ is the operator that binarizes o_{sin} (i.e., the IoU) at a given frame. The performance can be summarized by a tracking quality plot akin to [3] for all thresholds $\theta \in [0, 1)$ as shown in Figure 2. Note that the threshold interval is open, since by definition, the IoU cannot exceed $\theta = 1.0$. For the visualization purposes, the right-most point is thus evaluated with $[\cdot \equiv \theta]$.

The tracking quality plot has similar interpretation properties as the standard success plot [3], with a difference that the right-most point at $\theta = 1.0$ can be typically higher. The reason is that it accounts for long-term tracking properties in addition to short-term tracking properties. The values IoU=1 can only occur when the prediction completely matches the ground truth (sc1 and sc5 in Figure1). In practice, this is very rare when the target is visible, thus the value is dominated by cases of correctly predicting the tar-

[1]https://www.votchallenge.net/vots2023/

[2]Note that the initalization frames are excluded from evaluation, since the tracker does not *predict* the target location at those frames.

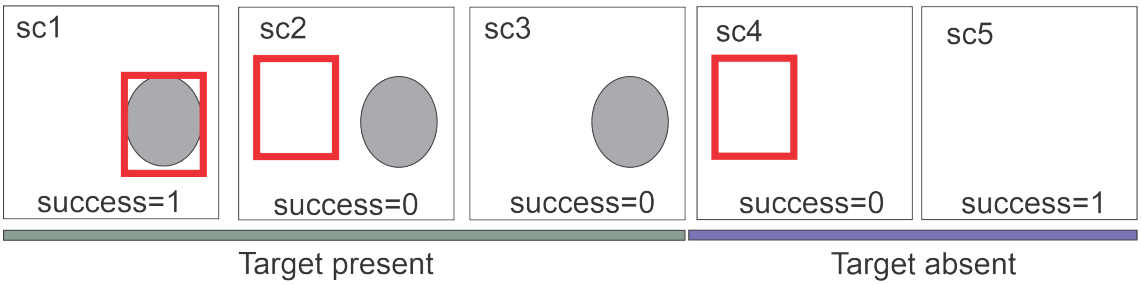

Figure 1. Five scenarios emerge from combinations of target presence and tracker outputs.

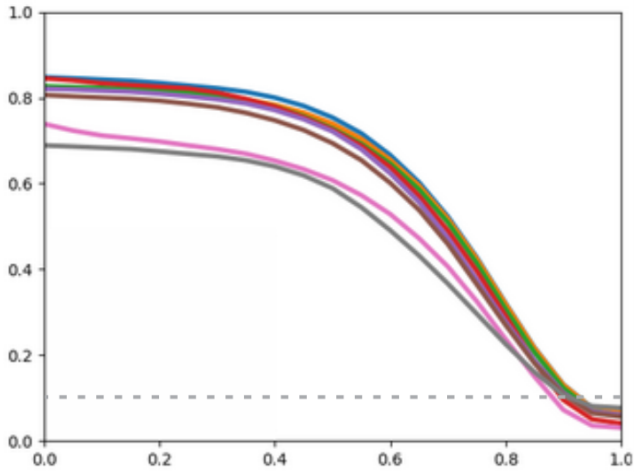

Figure 2. Tracking quality plot with the dashed line indicating perceptage of target-basent frames.

get absence (sc5). The practically maximal achievable value will thus be a percentage of the target absent frames in the dataset. This value is indicated in the plot for better interpretation.

The primary VOTS performance measure, called the tracking quality Q summarizes the tracking quality plot by the area under the curve. Following the success plot derivation in [2], it can be shown that the tracking quality is equal to the sequence-normalized average overlap to avoid errors in numerical area-under-the-curve computation, i.e.,

$$Q = \frac{1}{N} \sum_{s=1:N} \frac{1}{T_s N_s} \sum_{i=1:T_s} \sum_{n=1:N_s} o_{sin}. \quad (2)$$

1.1.1 Secondary performance measures

Additional *secondary performance measures* are proposed for further tracking insights. The first two measures, traditionally used in VOT [1], are localization *accuracy* and *robustness*. The accuracy (Acc) is defined as the sequence-normalized average overlap over successfully tracked frames, i.e.,

$$Acc = \frac{1}{N} \sum_{s=1:N} \frac{1}{T_s} \sum_{i=1:T_s} \frac{1}{N_{si}} \sum_{n=1:N_{si}} o_{sin}, \quad (3)$$

where $N_{si\theta}$ is the number of successfully tracked frames (i.e., with IoU$>$ 0) with the target i visible in sequence s. The tracking robustness (Rob) is defined as the percentage of frames with IoU$>$ 0 and target i visible (i.e., a recall),

$$Rob = \frac{1}{N} \sum_{s=1:N} \frac{1}{T_s} \sum_{i=1:T_s} \frac{N_{si}^{\text{sc1}}}{N_s^{\text{sc1+sc2+sc3}}}, \quad (4)$$

where N_{si}^{sc1} is the number of frames with scenario sc1 (Figure 1). Following our prior works [1], the tracker performance on frames with visible target is summarized by the AR plots [1], with the top-right position indicating the better performance.

The next two secondary performance measures answer the question *"Why did tracker fail while the target was visible?"*. The first measure, called *Not-Reported Error* (NRE), gives the percentage of frames where the tracker incorrectly reported the target as absent, i.e.,

$$NRE = \frac{1}{N} \sum_{s=1:N} \frac{1}{T_s} \sum_{i=1:T_s} \frac{N_{si}^{\text{sc3}}}{N_s^{\text{sc1+sc2+sc3}}}, \quad (5)$$

while the second, called *Drift-Rate Error* DRE, gives the percentage of frames where the tracker drifted off the target, i.e.,

$$DRE = \frac{1}{N} \sum_{s=1:N} \frac{1}{T_s} \sum_{i=1:T_s} \frac{N_{si}^{\text{sc2}}}{N_s^{\text{sc1+sc2+sc3}}}. \quad (6)$$

The final secondary measure answers the question *"How well is the traget absence determined?"*. This measure, called *Absence-Detection Quality* ADQ, gives the percentage of frames with target correctly predicted as absent, i.e.,

$$ADQ = \frac{1}{N} \sum_{s=1:N} \frac{1}{T_s} \sum_{i=1:T_s} \frac{N_{si}^{\text{sc5}}}{N_s^{\text{sc4+sc5}}}. \quad (7)$$

Note that in practice, to ensure numerical stability, we consider only those targets, that are absent for at least 10 frames in a sequence.

References

[1] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, J.-K. Kamarainen, L. Čehovin, D. Martin, A. Lukežič, O. Drbohlav, L. He, Y. Zhang, S. Yan, J. Yang, G. Fernández, and et al. The eighth visual object tracking vot2020 challenge results. In *ECCV2020 Workshops, Workshop on visual object tracking challenge*, 2020.

[2] Luka Čehovin, Aleš Leonardis, and Matej Kristan. Visual object tracking performance measures revisited. *IEEE Transactions on Image Processing*, 25(3), 2015.

[3] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Object tracking benchmark. *PAMI*, 37(9):1834–1848, 2015.