VOT committee comments on the VOT toolkit
Document version v.1.0

The VOT2014 committee

April 13, 2015

1 Using the VOT evaluation toolkit analysis for scientific papers

The VOT evaluation toolkit (https://github.com/vicoslab/vot-toolkit) is highly parameterized and allows running various types of analysis. The toolkit can generate much more tables and graphs that can fit into a scientific paper and several researchers have asked what do we recommend to be put in the paper. A short answer is: "Put whatever helps you make your point best, as long as the results and analysis are properly interpreted". In the following, a longer answer is given.

1.1 The VOT toolkit is continually improved by the VOT community

The VOT toolkit can be used to generate a variety of graphs or tables for condensed interpretation of the results. But we would like to point out that the toolkit is by no means a finished work. We have put a lot of effort in making the basic functionalities like running the experiments cross-platform compatible. But due to the restricted resources on man power, the VOT committee cannot guarantee absolute portability of all functionalities to all systems and environments, i.e., Matlab/Octave. We therefore invite the community to test the toolkit and report the compatibility issues and solutions on the GIT repository as "issues" to be resolved. Several researchers have already contacted us regarding compatibility solutions and we are grateful for their effort. The toolkit will improve even further if more people work with it. We also appreciate submissions of the toolkit functionality extensions and generations of additional performance visualizations computed from the standard VOT raw outputs.

1.2 Raw results

Researcher working in tracking are used to seeing tables of raw results, preferably in such form that allows some sort of rough cross-paper checking. We therefore suggest pasting a table of **per-sequence raw accuracy and robustness values** and these **values averaged over the sequences**. Weights proportional to the sequence length can be used in the averaging to account for the fact that longer sequences contain more data. For example, an overlap might be more reliably estimated on a longer sequence than on shorter ones. Alteratively equal weights can be used to make the contribution of all sequences equal (i.e., a sequence may be short, but also may be more challenging to track than a longer sequence). Still, another alternative is using the results of sequence tracking difficulty from the arXiv paper (http://arxiv.org/abs/1503.01313) to focus the results on the sequences that were most difficult to track in the VOT2014 challenge. In any case, the **type of averaging should be stated in the paper**. The averaged values can be compactly visualized in the AR-raw values plots showing trackers either in average overlap/number-of-failures space or overlap/probability-of-failure space (see the paper at arXiv:1502.0580 for details on this plot). The latter plot is automatically generated by the VOT2014 toolkit.

1.3 Rank results

It is also recommended to run some type of ranking and reporting the AR-rank plot. The ranking will take into account the possibility that some trackers may be performing indistinguishably by considering the statistical and practical difference. There are three main types of ranking implemented in the VOT2014 toolkit:

1. **Sequence-pooled ranking:** frames from all sequences are pooled together into a super-sequence and mean values of accuracy and robustness are computed (this is equivalent to a weighted averaging over per-sequence raw averages). Trackers are ranked with respect to these values and then the ranks are adjusted by determining equivalence groups as described in the arXiv paper (http://arxiv.org/abs/1503.01313).

2. **Sequence-normalized ranking:** trackers are ranked with respect to each sequence and the ranks are averaged. In this setup, each sequence is considered a competition. Trackers are ranked on each competition and the ranks are averaged over competitions to get the final rank.

3. **Attribute-normalized ranking:** trackers are ranked with respect to each visual attribute and the ranks are averaged. In this setup, each attribute is considered a competition. Trackers are ranked on each competition and the ranks are averaged over competitions to get the final rank.

1.3.1 A note on computing the corrected rank

As detailed in the arXiv paper (http://arxiv.org/abs/1503.01313), the trackers are ranked on raw values and then the ranks are adjusted as follows. For each tracker, a group of so-called equivalent trackers containing trackers performing indistinguishably is determined and a corrected rank is then calculated. Note that there are several choices for calculating the correction, e.g., one could take the min, max or mean of ranks in the group. The least conservative choice is max, since it always penalizes a tracker if the equivalency test cannot confirm the difference from a lower-ranked tracker, and on the other hand, the min is most conservative, since it always makes a correction in interest of the tracker. In the arXiv paper we use the mean of the ranks as a compromise between the two extremes. But the other two types of adjustment are still valid.

2 Per-attribute visualizations

In addition to rankings, the VOT2014 toolkit can also generate AR-rank plots with respect to each attribute. Another type of plots that can be produced by the VOT toolkit are the **tracker ordering plots** showing performance (accuracy or robustness) of each tracker with respect to all attributes. Such plots may be very helpful for comparative analysis of trackers with respect to the individual attributes.

3 Conclusion

The VOT committee does not impose using a particular type of the plots and outputs that were pointed out here. Different graphs and outputs show different performance aspects and it should be up to the researchers to choose the ones that best demonstrate a particular insight. But we do stress that the results should be made reproducible. We therefore recommend that whenever the VOT toolkit is used for a paper, the **raw output from the VOT toolkit are made publicly available** and explanation on how the plots in the paper were generated are provided. This can be accomplished by supplying a short document with how to reproduce the plots from the raw VOT toolkit output.

The VOT toolkit is an open source (https://github.com/vicoslab/vot-toolkit) and time will show which kinds of outputs will be most helpful. We also expect that the researchers from the tracking community will be committing their own visualizations to our GIT repository and will in time become part of the VOT toolkit.