Technical report 10-2013

Title: **Is my new tracker really better than yours?**

Authors: Luka Čehovin, Matej Kristan, and Aleš Leonardis

Date: October, 15th 2013

Is my new tracker really better than yours?

ViCoS Lab technical report 10-2013

Luka Čehovin, Matej Kristan, and Aleš Leonardis
Faculty of Computer and Information Science, University of Ljubljana, Slovenia
Tržaška 25, Ljubljana, Slovenia
{luka.cehovin,matej.kristan,ales.leonardis}@fri.uni-lj.si

October 15, 2013

Abstract

The problem of visual tracking evaluation is sporting an abundance of performance measures, which are used by various authors, and largely suffers from lack of consensus about which measures should be preferred. This is hampering the cross-paper tracker comparison and faster advancement of the field. In this paper we provide a critical analysis of the popular measures and evaluate them experimentally by a large-scale tracking experiment. We also analyze various visualizations of the performance measures. We show that several measures are equivalent from the point of information they provide for tracker comparison and, crucially, that some are more brittle than the others. Based on our analysis we narrow down the specter of measures to only a few complementary ones, thus pushing towards homogenization of the tracker evaluation methodology.

1 Introduction

Visual tracking is one of the rapidly evolving fields of computer vision. Every year, literally dozens of new tracking algorithms are presented and evaluated in journals and at conferences. When considering the evaluation of these new trackers and comparison to the state-of-the-art, several questions arise. Is there a standard set of sequences that we can use for the evaluation? Is there a standardized evaluation protocol? What kind of performance measures should we use? Unfortunately, there are currently no definite answers to these questions. Unlike some other fields of computer vision, like object detection and classification [11], optical-flow computation [4] and automatic segmentation [2], where widely adopted evaluation protocols are used, visual tracking is still largely lacking these features.

The absence of homogenization of the evaluation protocols makes it difficult to rigorously compare trackers across publications and stands in the way of faster development of the field. The authors of new trackers typically compare their work against a limited set of related algorithms due to the difficulty of adapting these for their own use in the experiments. There is no consensus on which experimental sequences to use for evaluation, and while most of the sequences used in the literature can be obtained from their authors, the ground truth annotations are not necessarily provided. But perhaps the biggest issue is the choice of tracker's performance evaluation measures, which seems to be almost arbitrary in the tracking literature. Worse yet, an abundance of these measures are currently in use. Because of this, experiments in many cases offer a limited insight into tracker's performance, and prohibit comparison across different papers.

In this paper we focus on the problem of performance evaluation in monocular single-target visual tracking and address several challenges therein. We investigate various popular performance evaluation measures, discuss their pitfalls and show that, from a standpoint of tracker comparison, there exist several equivalent measures currently in use. We identify only a few complementary measures, thus contributing to homogenization of the tracking performance evaluation methodology.

1.1 Related work

Majority of papers that address performance evaluation in visual tracking are concerned with multi-target tracking scenarios [29, 17, 7, 8, 10, 24, 6]. One might view the multi-target tracking as a generalization of single-target tracking, however, there is a crucial difference in the focus of evaluation. In multi-target tracking, the focus is usually on measuring correctness of target labeling assignments coupled with target detection and occlusion handling. The reason is that the algorithms are often focused on a particular tracking domain, which is typically people or vehicle tracking for surveillance [7, 8], animal groups tracking [18] or sports tracking [21], to name a few. A well known PETS workshop (e.g. [6]) has also been organized yearly for more than a decade with the main focus on performance evaluation of surveillance and activity recognition algorithms.

On the other hand, single-target visual tracking evaluation focuses on the tracker's accuracy, reliability and generality. The goal is to demonstrate the tracker's performance on a wide range of challenging scenarios (various types of object, lighting conditions, camera motion, signal noise, etc.). In this respect, the authors for [32] compared several trackers using center error and overlap measures. Their research is focused primarily on investigating strengths and weaknesses of a few trackers. In [33] authors perform a large-scale comparison of trackers. While the scale of the experiment is impressive in this case, the performance measures are not well chosen which results in poor qualitative analysis of the results. Recently, Nawaz and Cavallaro [25] have presented a system for evaluation of video trackers that aims at addressing the real-world conditions. The system can simulate several real-world sources of noisy input, such as initialization noise, image noise and changes in the frame-rate. They have also proposed a new performance measure to address the tracker's scoring under these simulated conditions that is discussed in Section 2.6. These recent experimental evaluations show the need for a better evaluation of visual trackers, however, none of them addresses an important prerequisite for such evaluation, that is the selection of good performance measures. This selection should be grounded in an analysis of performance measures which is the main focus of this paper.

Objective and rigorous evaluation calls for semi-automatic tools tailored for such tasks. Most notable and general are the ODViS system [15] and the ViPER toolkit [9]. The former focuses on design of surveillance systems, while the latter is a set of utilities/scripts for annotation and computation of different types of performance measures. Several other systems also exist (e.g. [25, 33]), however, they are all limited to a specific evaluation protocol or a small set of pre-integrated trackers. Note that none of the above systems enable performing a flexible and robust fully automatic large-scale tracking experiment and at the same time allow easy integration of third-party trackers. These features are crucial for the wide-spread acceptance of a standardized evaluation system that would allow researchers to consistently compare their trackers.

1.2 Our approach and contributions

In contrast to previous work on visual tracker performance evaluation, we focus our analysis on the abundant existing performance measures. We believe that a crucial step towards the homogenization of the field of single-target-tracking performance evaluation is narrowing the wide variety of existing measures to only a few complementary ones. As far as we are aware, this issue has not been previously addressed, therefore we claim a three-fold contribution: (1) We provide a detailed overview and experimental analysis of the widely used performance measures. (2) We show by experimental analysis that there exist several clusters of performance measures that really indicate the same aspect of tracker's performance. Within these clusters we analyze the properties of the measures and identify a single most suitable performance measure per cluster. In this respect we narrow down the wide range of performance measures to few complementary ones. The selected measures are then analyzed within the accuracy vs. reliability context. (3) Our analysis has been carried out by a large-scale experiment, which includes 13 state-of-the-art trackers and 25 widely-used video sequences. To allow such an experiment, we have developed a general-purpose performance evaluation framework for video tracking. In contrast to other frameworks, it enables easy integration of existing trackers and allows interactive control over the trackers as well

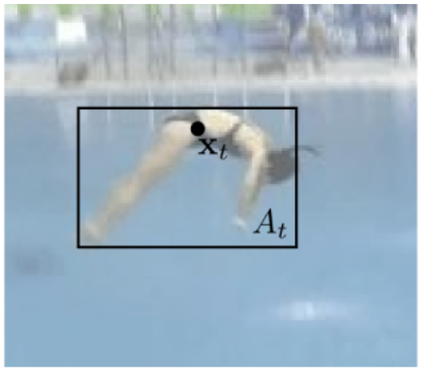

Figure 1: An example of an annotation for a single frame for the *driver* sequence. The center of the object can be estimated using the centroid of A_t, which is not true in this case.

as flexibility of experimental setups.

The rest of the paper is organized as follows: Section 2 gives an overview of the current state of performance evaluation techniques. Section 3 describes our experimental setup and the protocol. We discuss the experimental results in Section 4 and draw concluding remarks in Section 5.

2 Performance measures

There are several measures that have become popular and are widely used in the literature, however, none of them is a *de-facto* standard. As all of these measures assume that manual annotations are given for a sequence, we first establish a general definition of an object state description in a sequence with length N as:

$$\Lambda = \{(A_t, \mathbf{x}_t)\}_{t=1}^{N}, \tag{1}$$

where $\mathbf{x}_t \in \mathcal{R}^2$ denotes a center of the object and A_t denotes the region of the object at time t. In practice the region is usually described by a bounding box, however, a more complex shape could be used for a more accurate description. An example of a single frame annotation can be seen in Figure 1.

Performance measures aim at summarizing the extent to which the tracker's predicted annotation Λ_T agrees with the ground truth annotation, i.e., Λ_G.

2.1 Center error

Perhaps the oldest means of measuring performance, which has its roots in aeronautics, is the center prediction error. This is still a popular measure [28, 3, 1, 23] and it measures the difference between the target's predicted center from the tracker and the ground-truth center.

$$\Delta(\Lambda^G, \Lambda^T) = \{\delta_t\}_{t=1}^{N}, \;\; \delta_t = \|\mathbf{x}_t^G - \mathbf{x}_t^T\|. \tag{2}$$

The popularity of center prediction measure comes from its minimal annotation effort, i.e., only a single point per frame. The results are usually shown in a plot, as in

Figure 7 or summarized as average error (3), or root-mean-square-error (4):

$$\Delta_{\mu}(\Lambda^G, \Lambda^T) = \frac{1}{N}\sum_{t=1}^{N}\delta_t, \tag{3}$$

$$\mathrm{RMSE}(\Lambda^G, \Lambda^T) = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\|\mathbf{x}_t^G - \mathbf{x}_t^T\|^2}. \tag{4}$$

One drawback of this measure is its sensitivity to subjective annotation (i.e., where exactly is the target's center?). This sensitivity largely comes from the fact that the measure completely ignores the target's size and does not reflect the apparent tracking failure [25]. To remedy this, a normalized center error $\widehat{\Delta}(\cdot, \cdot)$ is used instead, e.g. [5], in which the center error at each frame is divided by the tacker-predicted visual size of the target, $size(A_t^G)$,

$$\widehat{\Delta}(\Lambda^G, \Lambda^T) = \left\{\widehat{\delta}_t\right\}_{t=1}^{N}, \;\; \widehat{\delta}_t = \|\frac{\mathbf{x}_t^G - \mathbf{x}_t^T}{size(A_t^G)}\|. \tag{5}$$

Nevertheless, despite the normalization, the measure may give misleading results as the center error is reduced proportionally to the estimated target size. Furthermore, when the tracker fails and is drifting over a background, the actual distance between the annotated and reported center, combined with the estimated size (which can be arbitrarily large) arbitrarily influences the averaged score and does not properly reflect the important information that the tracker has failed.

2.2 Region overlap

The normalization problem is rather well addressed by the overlap-based measures [34, 13]. These measures require region annotations and are computed as an overlap between predicted target's region form the tracker and the ground-truth region:

$$\Phi(\Lambda^G, \Lambda^T) = \{\phi_t\}_{t=1}^{N}, \;\; \phi_t = \frac{A_t^G \cap A_t^T}{A_t^G \cup A_t^T}. \tag{6}$$

A nice property of region overlap measures is that they account for both position and size of the predicted and ground-truth bounding boxes simultaneously, and do not result in arbitrary large errors at tracking failures. In fact, once the tracker drifts to the background, the measure becomes zero, regardless of how far from the target the tracker is currently located. In terms of pixel classification (see Figure 2), the overlap can be interpreted as

$$\frac{A_t^G \cap A_t^T}{A_t^G \cup A_t^T} = \frac{TP}{TP + FN + FP}, \tag{7}$$

a formulation similar to the F-measure in information retrieval, which can be written as

$$F = \frac{2TP}{2TP + FN + FP}. \tag{8}$$

Another closely related measure, used in tracking to account for un-annotated object occlusions is precision [13], or $\frac{TP}{TP+FP}$.

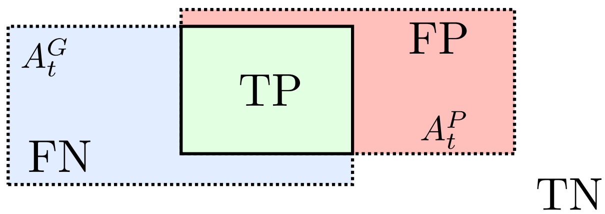

Figure 2: An illustration of the overlap of ground-truth region with the predicted region.

The overlap measure is summarized over an entire sequence by an average overlap, or as a number of correctly tracked frames. The latter approach comes from the object detection community [11], where the overlap threshold for a correctly detected object is set to 0.5. The same threshold is often used for tracking performance evaluation, e.g. in [34, 32]. To make the final score more comparable across the different sequences, the number of correctly tracked frames is divided by the total number of frames

$$P_\tau(\Lambda^G, \Lambda^T) = \frac{\| \{t|\phi_t > \tau\}_{t=1}^N \|}{N}, \tag{9}$$

where τ denotes the threshold of the overlap.

2.3 Tracking length

Another measure that has been used in the literature to compare trackers is *tracking length* [22]. This measure reports the number of successfully tracked frames from tracker's initialization to its (first) failure. A failure criterion can be manual visual inspection (e.g. [13]), which is biased and cannot be repeated reliably even by the same person. A better approach is to automate the failure criterion, e.g., by placing a threshold τ on the center or overlap measure (see Figure 3). The choice of the criterion may impact the result of comparison. Given the sensitivity of the center-based measures, an overlap criterion makes more sense, and we will denote in the following the tracking length measure with an overlap-based failure criterion by L_τ .

While this measure explicitly addresses the tracker's failure cases, which the simple average center-error and overlap measures do not, it suffers from a significant drawback. Namely, it only uses the part of the video up to the first tracking failure. If by some coincidence, the beginning of the video contains a difficult tracking situation, or the target is not visible well, which results in a necessarily poor initialization, the tracker will fail, and the remainder of the video will be discarded. This means that,

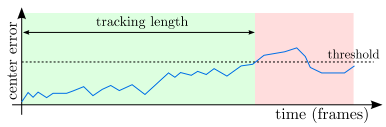

Figure 3: An illustration of the tracking length measure for center error.

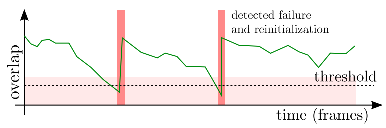

Figure 4: An illustration of the failure rate measure for overlap distance.

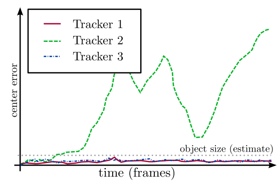

Figure 5: An example of center-error plot comparison for three trackers. Tracker 2 has clearly failed in the process, yet its large center errors cause the plot to expand its vertical scale, thus reducing the apparent differences of trackers 1 and 3.

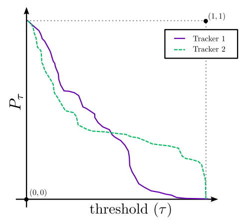

Figure 6: An illustration of the *measure-threshold* plot for two trackers. It is apparent that different values of the threshold would clearly yield different rankings for the trackers.

technically, one would require a significant amount of videos exhibiting the various properties right at its beginning to get a good statistic on this performance measure.

2.4 Failure rate

A measure that largely addresses the problem of the tracking length measure is the so-called failure rate measure [20, 18]. The failure rate measure casts the tracking problem as a supervised system in which an operator reinitializes the tracker once it fails. The number of required manual interventions per frame is recorded and used as a comparative score. The approach is illustrated in Figure 4. This measure also reflects the tracker's performance in a real-world situation in which the human operator supervises the tracker and corrects its errors.

Compared to the tracking length measure, the failure rate approach has the advantage that the entire sequence is used in the evaluation process and decreases the importance of the beginning part of the sequence. The question of a failure criterion threshold is even more apparent here as each change in the criterion requires the entire experiment to be repeated. Researchers in [30, 31] consider a failure when the bounding box overlap is lower than 0.1. This lower threshold reasonable for non-rigid objects, since these are often poorly described by the bounding-box area. An even lower threshold could be used for overlap-based failure criteria if we are interested only in the most apparent failures with no overlap between the regions. We will denote the failure rate measure with an overlap-based failure criterion with threshold τ as F_τ.

2.5 Performance plots

Plots are frequently used to visualize the behavior of a tracker when a single number does not suffice for expressing its performance. The most widely-used plot is a center-error plot that shows the center-error with respect to the frame number [3, 1, 5, 34]. While this kind of plot can be useful for visualizing tracking result of a single tracker, a combined plot for multiple trackers is in many cases misused if applied without caution, because the tracker with an inferior performance steals away the focus from the information that we are interested in with this type of plots, i.e. the tracker accuracy. An illustration of such a problematic plot is shown in Figure 5.

In the previous section we have seen that a failure criterion plays a significant role in visual tracker performance evaluation. Choosing an appropriate value for the threshold may affect the ranking and can also be potentially abused to influence the results of a comparison. However, it is sometimes better to avoid the use of one specific threshold altogether, especially when the evaluation goal is general and a specific threshold is not a part of the target task. To avoid the choice of a specific threshold, results can be presented as a *measure-threshold* plot. This kind of plots bear some resemblances to a ROC curve [12], like monotony, intuitive visual comparison, and a similar calculation algorithm. Measure-threshold plots were used in [3], where the authors used center-error as a measure as well as in [33], where both center-error and overlap are used.

The percentage of correctly tracked frames, defined in (9) as P_τ, is a good choice for a measure to be used in this scenario, however, other measures could be used as well. The P_τ measure can be intuitively computed for multiple sequences which makes it useful for summarizing the entire experiment (an example of P_τ plot is illustrated in Figure 6). Interpretations of such plots have been so far limited to their basic properties which in a way negates the information verbosity of a graphical representation. For example, similarly to ROC curves, we can compute an area-under-the-curve (AUC) summarization score, which is used in [33] to reason about performance of trackers. It can be trivially proven that the AUC score in this case actually matches the average overlap over the entire sequence (proof available in the supplementary material) and therefore adds no additional insight into the performance of the tracker.

2.6 Hybrid measures

Nawaz and Cavallaro [25] propose a threshold-independent overlap-based measure that combines the information on tracking accuracy and tracking failure into a single score. This hybrid measure is called the *Combined Tracking Performance Score* (CoTPS) and is defined as a weighted sum of an accuracy score (based on the frames where the tracker was successful) and a failure score (based on the frames where the tracker failed to predict the position of the object). An appealing property of this measure is that it ranks trackers by accounting for two separate aspects of tracking. However, the authors do not give any justification for this rather complex fusion of measures. This complexity prohibits easy interpretation of the results required for a rigorous scientific analysis.

In terms of interpretation, we therefore believe that a better strategy is to focus on a few complementary performance measures with well-defined meaning, but avoid fusing these into a single measure.

3 Experimental setup

In order to analyze measures, we have conducted an experiment in which we have ranked several existing trackers according to the selected measures on a number of sequences. For our experiment we have selected variants of the popular measures, described in Section 2: average center error, average normalized center error, root-mean-square error, average overlap, percent of correct frames $P_{0.1}$, tracking length $L_{0.1}$, percent of correct frames $P_{0.5}$, tracking length $L_{0.5}$, failure rate F_0, average overlap for F_0.

We have used 13 trackers that were proposed in recent years: A color-based particle filter (PF) [27], the On-line boosting tracker (OBT) [14], the Flock-of-features tracker (FOF) [19], the Basin-hopping Monte Carlo tracker (BHMC) [22], the Incremental visual tracker (IVT) [28], the Histograms-of-blocks tracker (BH) [26], the Multiple instance tracker (MIL) [3], the Fragment tracker (FRT) [1], the P-N tracker (TLD) [16], the Local-global tracker (LGT) [31], Hough tracker (HT) [13], the L1 Tracker Using Accelerated Proximal Gradient Approach (L1-APG) [5] and the Compressive tracker (CT) [34]. The source code of the trackers was provided the authors and adapted to fit into our framework.

We have run the trackers on 25 different sequences, most of which are already known in the video tracking community, and several were acquired additionally. The sequences were annotated with a bounding-box region of the object, as well as the target's central point of the object. Table 1 shows an overview of the basic properties of our sequences.

3.1 Testing protocol

To account for stochastic processes that are a part of many trackers, each tracker was executed on each sequence 30 times. The tracker's performance on a particular sequence was then evaluated by averaging these results. Parameters for all trackers were set to their default values and kept constant during the experiment. A separate run was executed for the *failure rate* measure as the re-initialization influences other aspects of tracking performance.

Because of the scale of the experiment, only the most relevant results are presented in Section 4. Additional results, such as the ranking of the trackers according to individual measures, are available in the supplementary material.

4 Results and discussion

Different measures may reflect different aspects of tracking performance, so it is impossible to simply establish which measure is the best. We start our analysis by establishing similarities and equivalence between various measures, by experimentally analyzing which measures produce consistently similar responses when comparing trackers.

Table 1: An overview of the experimental sequences.

Name	cam. mot.	illum.	occl.	focus / noise	articul.	target
bicycle [31]	•		•		•	bike
biker[34]				•	•	bike
bolt[34]	•				•	body
can [31]						toy
car [32]	•			•		car
child	•	•				head
david_indoor [28]	•	•				head
david_outdoor [32]			•		•	body
dinosaur [30]						toy
diver [22]	•				•	body
face [1]			•			head
gymnastics [30]	•				•	body
gymnastics2 [22]	•				•	body
hand [30]					•	hand
hand2 [30]					•	hand
motocross1 [13]	•				•	bike
mountainbike [13]	•				•	bike
pets2000						car
pets2001-1			•			car
pets2001-2			•			car
sunshade	•	•				head
torus [30]						toy
trellis [28]	•	•				head
turtlebot1	•			•		robot
woman [1]	•		•		•	body

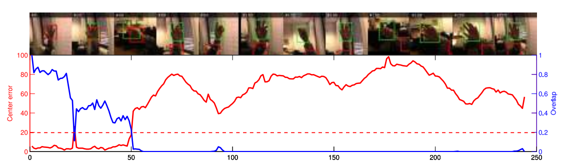

Figure 7: A comparison of overlap and center error distance measures for tracker CT on sequence *hand*. The dashed line shows the estimated threshold above which the center error is greater than the size of the object. The tracker fails around frame 50.

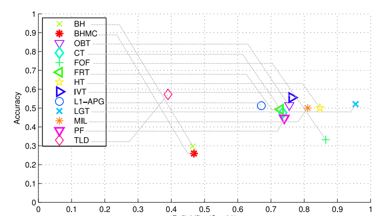

Figure 8: An accuracy-reliability data visualization for all trackers over all sequences.

4.1 Correlation analysis

We calculate a correlation matrix from all pairs of measures calculated over all tracker-sequence pairs. Note that we do not calculate the correlation on rankings directly to avoid handling situations where several trackers take the same place (if differences are not statistically significant). The rationale is that strongly correlated measure values will also produce similar ranking for trackers.

The obtained correlation matrix is shown in Figure 9. We can see that two clusters emerge, one for measures 1 to 3 and one for measures 4 to 7, where measure 7 is less correlated to the other three. All these correlations are highly statistically significant

($p < 0.001$).

The first cluster of measures consists of the three center-error-based measures. This is expected as these measures all base on *center-error* using different averaging methods. The second cluster of measures contains *average overlap*, *percentage of correctly tracked frames* for two threshold values ($P_{0.1}$ and $P_{0.5}$) and *tracking length* $L_{0.1}$. Measures in the second cluster assume that incorrectly tracked frames do not influence the final score based on the specific (incorrect) position of the tracker. This makes them more representative as a measure for tracking performance than the center-error-based measures. An illustration of this difference for *overlap* and *center-error* is shown as a graph on Figure 7, where we can clearly see that the center-error measure takes into account the exact center distance at frames after the failure has occurred, which depends on the movement of an already failed tracker and does not reflect its true performance.

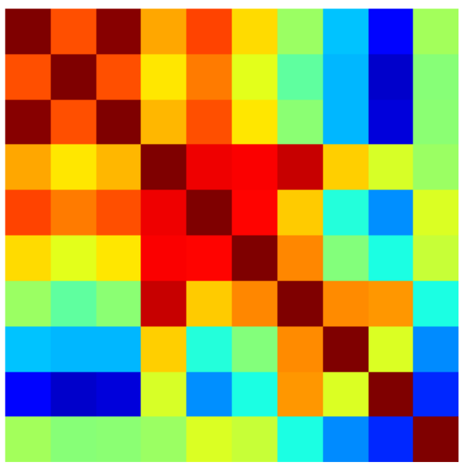

Figure 9: Correlation matrix for all measures visualized as a heat-map. The image is best viewed in color.

The first cluster implies that the first three measures are equivalent and it does not matter which one you chose. The second cluster requires some more interpretation. Despite the apparent similarity of overlap-based measures 4 to 7, the correlation is not perfect and the rankings differ in some cases. One example of such difference can be seen for the TLD tracker on the *woman* sequence, seen in Figure 10. We can see that the tracker loses the target early on in the sequence (during an occlusion), but manages to locate it again later because of its discriminative nature. The *average overlap* measure (number 4) and the *percentage of correct frames* measure (number 5) therefore rank the tracker higher than the *tracking length* measure (number 6). On the general level we can also observe that the choice of a threshold can influence the outcome of the experiment. This can be observed for tracking length measures 6 and 8 and to some extent for the percentage of correct frames measures 5 and 7. There, the scores for a higher threshold (0.5) result in a different ranking of trackers compared to the lower threshold (0.1). This means that care must be taken when choosing the thresholds at

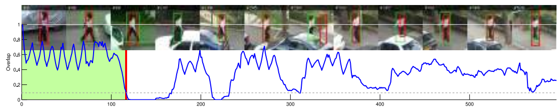

Figure 10: An overlap plot for tracker TLD on sequence *woman*. The dashed line shows the threshold below which the tracking length detects failure (for threshold 0.1), which happens around frame 120.

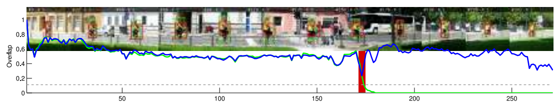

Figure 11: An overlap plot for tracker LGT on sequence *bicycle*. The green plot shows the unsupervised overlap, and the blue plot shows the overlap for supervised tracking, where the failure is recorded and the tracker re-initialized.

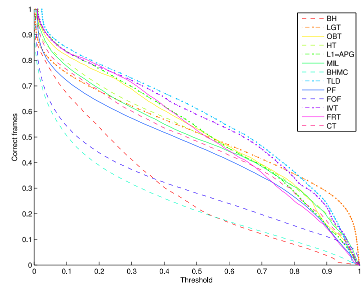

Figure 12: An accuracy plot for all trackers averaged over all sequences.

they may affect the outcome of the evaluation.

We can in fact observe a slight overlap between the first two clusters in the correlation matrix, implying similarity in their information content. Based on the above analysis and discussion in Section 2 we conclude that the *average overlap* measure is the most appropriate to be used in tracker comparison, as it is simple to compute, it is scale and threshold invariant, exploits the entire sequence, and it is easy to interpret. Note also that it is highly correlated with a more complex *percentage-of-correctly-tracked-frames* measure.

4.2 Failure rate

As mentioned before, the *failure rate* measure influences the tracker's entire trajectory, because of the re-initializations, therefore the data for measures 9 and 10 was acquired separately. The advantage of the *failure rate* (measure 10) is that the entire sequence is used, which makes the results statistically significant at smaller number of sequences. It does not matter that much if one tracker fails at the "difficult" beginning of the sequence, while the other one barely survives and then tracks the rest successfully. In Figure 11 we can see the performance of the LGT tracker on the *bicycle* sequence. Because of the occlusion near frame 175 the tracker fails, although it is clearly capable of tracking the rest of the sequence reliably if re-initialized.

4.3 Accuracy vs. reliability

The *failure rate* measure itself measures the reliability of the tracker, however, it tells us nothing about its accuracy. We therefore propose to use the *average overlap* measure on the same (re-initialized) data to take into account this aspect of tracking. We define a new, A-R measure, as a pair of scores

$$\text{A-R}(\Lambda^G, \Lambda^T) = \left(\overline{\Phi}(\Lambda^G, \Lambda^T), e^{-SF_0}\right), \tag{10}$$

where $\overline{\Phi}$ denotes *average overlap* and F_0 denotes the *failure rate* for $\tau = 0$. The reliability score is interpreted as a probability that the tracker will still successfully track the object up to S frames since the last failure, where the failure probability is modeled using an exponential failure distribution based on F_0 and S. The choice of S does not influence the order of the trackers, however, changing its value can be useful for visualization and interpretation of results. Note that the value of failure threshold τ can influence the final results. If the value is set to a high value (close to 1) the tracker is restarted frequently even for small deviations from the ground-truth and the final score is hard to interpret. Instead, we propose to use the lowest theoretical threshold $\tau = 0$ to only measure complete failures where the regions have no overlap at all and a reinitialization is clearly justified. In theory a tracker can also report an extremely large region as the position of the target and avoids failures, however, the accuracy will be very low in this case. This is how the two measures complement each other.

The A-R pair can be visualized as a 2-D point plot as seen in Figure 8, where we show the average scores for all sequences, from which one can read the tracker's performance in terms of accuracy (the tracker is more accurate if it is higher along vertical axis) and reliability (the tracker fails less if it is further to the right on horizontal axis). The two measures complement each other: An ideal tracker would reach the top-right corner. On the other hand, a tracker that fails a lot, and achieves good accuracy because of frequent manual interventions, would be displayed closer to top-left corner. A tracker that provides too loose regions, but does not fail, would be displayed closer to the bottom-right corner.

For more detailed analysis of tracker's accuracy a *measure-threshold* plot can be used as described in Section 2.5. The summarized results for all sequences are presented in Figure 12, sequence-specific plots are available in the supplementary material. From the summarized plot we can see in more detail that the trackers with worst accuracy are BH, BHMC and FOF, however, the FOF tracker is much more reliable according to Figure 8.

5 Conclusion

In this paper we have provided an analysis of the measures for video tracking performance evaluation. Besides a theoretical description and categorization of frequently used measures, we have also performed a large-scale experimental evaluation on 13 recently presented trackers and 25 available evaluation sequences, in order to determine the properties and relationships for several measures. From the experiment described in this paper we can conclude that the selection of a single measure can indeed influence the result of the comparison. Results, summarized by a any single measure cannot sufficiently describe the performance of such a complex system as a visual tracker, nor can they help us discover the true meaning of its failure, especially on a large set of testing sequences. Based on our experiment we therefore propose that a pair of two complementary measures is used for thorough evaluation that takes into account the accuracy (*average overlap*) and reliability (*failure rate*) of each tracker.

While narrowing down the abundance of performance measures is a big step toward homogenizing the tracking evaluation methodology, we expect that several issues have yet to be addressed. There is no consistent methodology on testing the video trackers. For example, how to handle stochastic trackers, and the question of the (sometimes sequence-specific) tracker's parameters. Furthermore, the community lacks a centralized repository for organized storage of annotated sequences. An organized large-scale and well-annotated database would allow researchers to perform experiments on a common dataset quickly as well as exchange and compare their results. Our future work will make effort along these directions.

Bibliography

[1] A. Adam, E. Rivlin, and I. Shimshoni. Robust Fragments-based Tracking using the Integral Histogram. In *CVPR*, volume 1, pages 798–805. IEEE Computer Society, June 2006.

[2] S. Alpert, M. Galun, R. Basri, and A. Brandt. Image Segmentation by Probabilistic Bottom-Up Aggregation and Cue Integration. In *CVPR*, pages 1–8. IEEE, June 2007.

[3] B. Babenko, M.-H. Yang, and S. Belongie. Robust Object Tracking with Online Multiple Instance Learning. *TPAMI*, 33(8):1619–1632, Aug. 2011.

[4] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski. A Database and Evaluation Methodology for Optical Flow. *IJCV*, 92(1):1–31, Nov. 2010.

[5] C. Bao, Y. Wu, H. Ling, and H. Ji. Real time robust L1 tracker using accelerated proximal gradient approach. In *CVPR*, pages 1830–1837. IEEE, June 2012.

[6] F. Bashir and F. Porikli. Performance Evaluation of Object Detection and Tracking Systems. In *PETS*, 2006.

[7] J. Black, T. Ellis, and P. Rosin. A novel method for video tracking performance evaluation. In *VS-PETS*, pages 125–132, 2003.

[8] L. M. Brown, A. W. Senior, Y.-l. Tian, J. Connel, A. Hampapur, C.-F. Shu, H. Merkl, and M. Lu. Performance evaluation of surveillance systems under varying conditions. In *PETS*, pages 1–8, 2005.

[9] D. Doermann and D. Mihalcik. Tools and techniques for video performance evaluation. In *ICPR*, pages 167–170, 2000.

[10] K. K. Edward, P. D. Matthew, and B. H. Michael. An information theoretic approach for tracker performance evaluation. In *CVPR*, pages 1523–1529. IEEE, Sept. 2009.

[11] M. Everingham, L. Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *IJCV*, 88(2):303–338, Sept. 2009.

[12] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, June 2006.

[13] M. Godec, P. M. Roth, and H. Bischof. Hough-based tracking of non-rigid objects. In *ICCV*, pages 81–88. IEEE, Nov. 2011.

[14] H. Grabner, M. Grabner, and H. Bischof. Real-time tracking via on-line boosting. In *BMVC*, pages 47–56, 2006.

[15] C. Jaynes, S. Webb, R. Steele, and Q. Xiong. An open development environment for evaluation of video surveillance systems. In *PETS*, 2002.

[16] Z. Kalal, J. Matas, and K. Mikolajczyk. P-N learning: Bootstrapping binary classifiers by structural constraints. In *CVPR*, pages 49–56. IEEE, June 2010.

[17] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova, and J. Zhang. Framework for performance evaluation of face, text, and vehicle detection and tracking in video: data, metrics, and protocol. *TPAMI*, 31(2):319–36, Feb. 2009.

[18] Z. Khan, T. Balch, and F. Dellaert. MCMC-based particle filtering for tracking a variable number of interacting targets. *TPAMI*, 27:1805–1819, 2005.

[19] M. Kölsch and M. Turk. Fast 2d hand tracking with flocks of features and multi-cue integration. In *CVPR Workshops*, volume 10, page 158, Washington, DC, USA, 2004. IEEE Computer Society.

[20] M. Kristan, S. Kovačič, A. Leonardis, and J. Perš. A two-stage dynamic model for visual tracking. *Trans. Sys. Man Cyber. Part B*, 40(6):1505–1520, Dec. 2010.

[21] M. Kristan, J. Perš, M. Perše, and S. Kovačič. Closed-world tracking of multiple interacting targets for indoor-sports applications. *CVIU*, 113(5):598–611, May 2009.

[22] J. Kwon and K. Lee. Tracking of a non-rigid object via patch-based dynamic appearance modeling and adaptive basin hopping monte carlo sampling. In *CVPR*, pages 1208–1215, 2009.

[23] J. Kwon and K. M. Lee. Visual tracking decomposition. In *CVPR*, pages 1269–1276. IEEE, June 2010.

[24] V. Manohar, P. Soundararajan, H. Raju, D. Goldgof, R. Kasturi, and J. Garofolo. Performance evaluation of object detection and tracking in video. In *ACCV*, pages 151–161, 2006.

[25] T. Nawaz and A. Cavallaro. A protocol for evaluating video trackers under real-world conditions. *Tran. Image Proc.*, 2012.

[26] S. S. Nejhum, J. Ho, and M.-H. Yang. Online visual tracking with histograms and articulating blocks. *CVIU*, 114(8):901–914, Aug. 2010.

[27] P. Pérez, C. Hue, J. Vermaak, and M. Gangnet. Color-based probabilistic tracking. In *ECCV*, volume 1, pages 661–675. Springer-Verlag, 2002.

[28] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang. Incremental Learning for Robust Visual Tracking. *IJCV*, 77(1-3):125–141, May 2008.

[29] K. Smith, D. Gatica-Perez, and J. Odobez. Evaluating Multi-Object Tracking. In *CVPR Workshops*, pages 36–36. IEEE, 2005.

[30] L. Čehovin, M. Kristan, and A. Leonardis. An adaptive coupled-layer visual model for robust visual tracking. In *ICCV*, Barcelona, 2011.

[31] L. Čehovin, M. Kristan, and A. Leonardis. Robust Visual Tracking using an Adaptive Coupled-layer Visual Model. *TPAMI*, 35(4):941–953, Apr. 2013.

[32] Q. Wang, F. Chen, W. Xu, and M.-H. Yang. An experimental comparison of online object-tracking algorithms. In *SPIE*, pages 81381A–81381A–11, San Diego, 2011.

[33] Y. Wu, J. Lim, and Y. M.-H. Online Object Tracking: A Benchmark. In *CVPR*, 2013.

[34] K. Zhang, L. Zhang, and M.-H. Yang. Real-time Compressive Tracking. In *ECCV*, 2012.