Visual object tracking performance measures revisited

Cehovin, Luka; Leonardis, Ales; Kristan, Matej

Document Version
Peer reviewed version

[Link to publication on Research at Birmingham portal](#)

Visual object tracking performance measures revisited

Luka Čehovin *Member, IEEE,* Aleš Leonardis, *Member, IEEE,* and Matej Kristan, *Member, IEEE*

Abstract—The problem of visual tracking evaluation is sporting a large variety of performance measures, and largely suffers from lack of consensus about which measures should be used in experiments. This makes the cross-paper tracker comparison difficult. Furthermore, as some measures may be less effective than others, the tracking results may be skewed or biased towards particular tracking aspects. In this paper we revisit the popular performance measures and tracker performance visualizations and analyze them theoretically and experimentally. 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 set of potential measures to only two complementary ones, describing accuracy and robustness, thus pushing towards homogenization of the tracker evaluation methodology. These two measures can be intuitively interpreted and visualized and have been employed by the recent Visual Object Tracking (VOT) challenges as the foundation for the evaluation methodology.

Index Terms—visual object tracking, performance evaluation, performance measures, experimental evaluation

I. 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 [1], optical-flow computation [2] and automatic segmentation [3], where widely adopted evaluation protocols are used, visual tracking is still largely lacking these properties.

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

L. Čehovin and M. Kristan are with the Faculty of Computer and Information Science at the University of Ljubljana, Ljubljana, Slovenia. E-mail: {luka.cehovin,matej.kristan}@fri.uni-lj.si.

A. Leonardis is with the School of Computer Science and Centre for Computational Neuroscience and Cognitive Robotics, University of Birmingham, United Kingdom and with the Faculty of Computer and Information Science, University of Ljubljana, Slovenia. E-mail: a.leonardis@cs.bham.ac.uk, ales.leonardis@fri.uni-lj.si

to the difficulty of adapting these for their own use in the experiments. One of the issues here is the choice of tracker performance evaluation measures, which seems to be almost arbitrary in the tracking literature. Worse yet, an abundance of these measures are currently in use [4], [5], [6], [7]. Because of this, experiments in many cases offer a limited insight into the tracker's performance, and prohibit comparisons across different papers.

In contrast to the existing works on evaluation of single-target visual trackers, that focus on benchmarking visual trackers without considering the selection of good measures, or propose new complex measures, we take a different approach. We investigate various popular performance evaluation measures using theoretically provable relations between them as well as systematic experimental analysis. We discuss their pitfalls and show that, from a standpoint of tracker comparison, many of the widely used measures are in fact equivalent. In addition we prove a direct relation of two complex recently proposed performance measures with the basic performance measures thus allowing their analysis in terms of the basic performance measures. Since several measures reflect the same aspects of tracking performance, combining those provides no additional performance insights and in fact introduces bias towards a particular aspect of performance to the result. We identify complementary measures that are sensitive to two different aspects of trackers performance and demonstrate their practical interpretation on a large-scale experiment. We emphasize that the goal of our analysis is therefore not to rank state-of-the-art tracking algorithms and make claims on which is better, but to homogenize the tracking performance evaluation methodology and increase the interpretability of results.

In our work we focus on the problem of performance evaluation in monocular single-target visual tracking that does not contain complete disappearance of the target from the scene that would require later re-detection; this kind of tracking scenario is also known as *short-term* single-target visual tracking in contrast to single-target long-term tracking (target has to be re-detected) [8], [9], [10] and multi-target tracking (multiple targets) [11], [12]. It is worth noting that our findings have been so far already used as the foundation of the evaluation methodology of the recent Visual Object Tracking challenges VOT2013 [13] as well as VOT2014 [14].

A. Related work

Until recently the majority of papers that address performance evaluation in visual tracking were concerned with multi-target tracking scenarios [15], [11], [16], [17], [18], [19],

[20], [12], [21]. Single-target tracking is, at least theoretically, a special case of multi-target tracking, however, because of the nature of the target domain, there is a crucial difference in the focus of the evaluation. In multi-target tracking, the focus is on the correctness of target identity assignments for a varying number of targets as well as the accuracy of these detections. The algorithms are often focused on a particular tracking domain, which is typically people or vehicle tracking for surveillance [16], [17], [22], animal groups tracking [23] or sports tracking [24], to name a few, which means that tracking in multi-object scenarios involves a lot of domain-specific prior knowledge. A well known PETS workshop (e.g. [20]) 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 accuracy of the tracker, as well as its robustness and generality. The goal is to demonstrate the tracking performance on a wide range of challenging scenarios (various types of objects, lighting conditions, camera motions, signal noise, etc.). In this respect, Wang et al. [4] compared several trackers using center error and overlap measures. Their research is focused primarily on investigating strengths and weaknesses of a limited set of trackers. In [5] authors perform an experimental comparison of several trackers. The performance measures in this case are chosen without theoretical justification which results in a poor qualitative analysis of the results. Nawaz and Cavallaro [6] have presented a system for evaluation of visual trackers that aims at addressing the real-world conditions in sequences. 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 trackers scoring, but the measure was introduced without theoretical analysis of its properties. While such a measure may look like a good tool for ranking trackers, it cannot answer a simple question of in which aspect one tracker was better than the other. These recent experimental evaluations show the need for a better evaluation of visual trackers, however, none of them seems to address an important prerequisite for such evaluation, that is, what subset of the many available measures should be used for the evaluation. Frequently, multiple measures are used to cover multiple aspects of tracking performance without considering the fact that some measures describe the same aspects which leads to bias of the results. Instead, the selection should be grounded in a prior analysis of performance measures which is the main focus of this paper. Recently, Smeulders et al. [7] provided an experimental survey of several recent trackers together with an analysis of several performance measures. Their methodology and the general disposition in this aspect are similar to ours in terms that they search for multiple measures that describe different aspects of tracking performance. However, even though they do not explicitly acknowledge the fact that they address long-term tracking, their selection of measures and the dataset is from the start biased in favor of detection-based tracking algorithms, which also affects their results and derived conclusions.

Finally, evaluation of tracker performance without ground-truth annotations has been investigated by Wu et al. [25], where the authors propose to use time-reversible nature of physical motion. As noted by SanMiguel et al. [26], this approach is not suitable for longer sequences. They propose to extend the approach using failure detection based on the uncertainty of the tracker. The problem is that the method has to be adapted to each tracker specifically and is useful only for investigative, but not for comparative purposes. An interesting approach to tracker comparison has also been recently proposed by Pang and Habin [27]. They aggregate existing experiments, published in various articles, in a page-rank fashion to form a ranking of trackers. The authors acknowledge that their meta-analysis approach is not appropriate for ranking recently published trackers. Furthermore, their approach does not remove bias that comes from correlation of multiple performance measures, which is one of the goals of our work.

B. Our approach and contributions

In this paper we do not intend to propose new performance measures. Rather than doing this, we focus on narrowing the wide variety of existing measures for single-target tracking performance evaluation to only a few complementary ones. We claim a four-fold contribution: (1) We provide a detailed survey and experimental analysis of popular performance measures used in single-target tracking evaluation. (2) We show by experimental analysis that there exist clusters of performance measures that essentially indicate the same aspect of trackers performance. (3) By considering the theoretical aspects of existing measures as well as the experimental analysis we identify a subset of the two most suitable (complementary) measures that characterize trackers performance within the accuracy and robustness context as well as a simple and intuitive visualization of the selected pair of measures, and (4) we introduce the concept of theoretical tracker and propose four such trackers as guides in interpretation of tracker performance.

Our experimental analysis has been carried out in a form of a large-scale comparative experiment with 16 state-of-the-art trackers and 25 video sequences of common visual tracking scenarios. While the primary goal of this paper is not benchmarking trackers, we provide the performance results of the tested trackers on the two selected performance measures as a sideproduct of our experiment. We also intend to provide detailed results of the experiment (groundtruth and raw trajectories) as a side-product of our research[1] for further study by other researchers.

Preliminary results reported in this paper have been published in our conference paper [28]. This paper extends [28] in several ways. The related work has been significantly extended with the recent work in performance evaluation. The theoretical survey has been extended and proofs of reformulation of complex performance measures (e.g. CoTPS [6] and AUC [5], [6]) in terms of the basic measures have been added. A new fragmentation indicator has been proposed to complement the analysis of failure rate measure. The experimental analysis has

---

[1]Raw data is available at http://go.vicos.si/performancemeasures.

Fig. 1. Two examples of an annotation for a single frame from the *woman* and the *driver* sequence. In the left example the center of the object can be estimated using the centroid of R_t, which is not true in the second case.

Fig. 2. An illustration of the overlap of ground-truth region with the predicted region for four different configurations.

been extended by adding three state-of-the-art trackers. Two new theoretical trackers have been proposed to aid analysis of the selected performance measures. Guidelines have been set on automatic interpretation of sequence properties from the results of theoretical trackers.

The rest of the paper is organized as follows: Section II gives an overview of the current state of short-term single-target tracking performance evaluation measures. We describe our experimental setup and discuss the findings of the experiment in Section III, where we also propose our selection of good measures together with several insights. Finally, we draw concluding remarks in Section IV.

II. PERFORMANCE MEASURES

There are several performance measures that have become popular in single-target visual tracking evaluation 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:

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

where $\mathbf{x}_t \in \mathcal{R}^2$ denotes a center of the object and R_t denotes the region of the object at time t. In practice the region is usually described by a bounding box (that is most commonly axis-aligned), however, a more complex shape could be used for a more accurate description. An example of two single-frame annotations can be seen in Figure 1. In some cases the annotated center can be automatically derived from the region, but for some articulated objects, the centroid of region R_t does not correspond to $\mathbf{x}_t$, therefore it is best to separately annotate $\mathbf{x}_t$.

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.

A. 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 [29], [30], [31], [32], [5], [33], [34] and it measures the difference between the target's predicted center from the tracker and the ground-truth center.
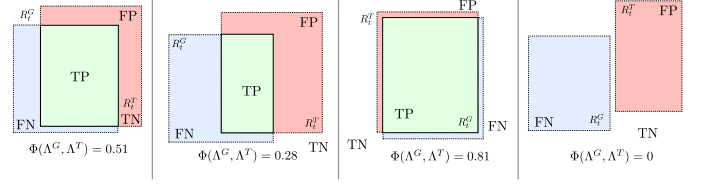
$$\Delta(\Lambda^G, \Lambda^T) = \{\delta_t\}_{t=1}^N, \quad \delta_t = \|\mathbf{x}_t^G - \mathbf{x}_t^T\|. \tag{2}$$

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

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

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

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 [6]. To remedy this, a normalized center error $\widehat{\Delta}(\cdot, \cdot)$ is used instead, e.g. [35], [7], in which the center error at each frame is divided by the tacker-predicted visual size of the target, $size(R_t^G)$,

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

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) overpowers the averaged score which does not properly reflect the important information that the tracker has failed.

B. Region overlap

The normalization problem is rather well addressed by the overlap-based measures [36], [37], [7]. These measures require region annotations and are computed as an overlap between predicted target's region form the tracker and the ground-truth region:

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

An appealing 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, as is the case on center-based error measures. 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

$$\frac{R_t^G \cap R_t^T}{R_t^G \cup R_t^T} = \frac{TP}{TP + FN + FP}, \quad (7)$$

a formulation similar to the F-measure in information retrieval, which can be written as $F = \frac{2TP}{2TP + FN + FP}$. Another closely related measure, used in tracking to account for un-annotated object occlusions is precision [37], i.e. $\frac{TP}{TP+FP}$.
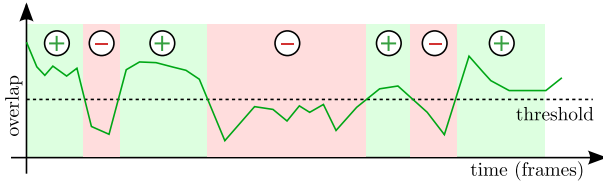

Fig. 3. An illustration of overlap being used as a detection measure. The plus signs mark the intervals with positive detections (overlap above threshold), while minus signs mark the intervals with negative detections (interval below threshold).

The overlap measure is summarized over an entire sequence by an *average overlap* (e.g. in [33], [5]) that is defined as an average value of all region overlaps in the sequence

$$\bar{\phi} = \sum_t \frac{\phi_t}{N}. \quad (8)$$

Another measure based on region overlap is number of correctly tracked frames $N_\tau = \sum_t \| \{t|\phi_t > \tau\}_{t=1}^N \|$, where τ denotes a threshold on the overlap. This approach comes from the object detection community [1], where the overlap threshold for a correctly detected object is set to $\tau = 0.5$. The same threshold is often used for tracking performance evaluation, e.g. in [36] and [4], however, this number is too high for general purpose tracking evaluation. As seen in Figure 2 this threshold is reached even for visually well overlapping rectangles. This is especially problematic when considering non-rigid articulated objects.

To make the final score more comparable across a set of sequences of different lengths, the number of correctly tracked frames is divided by the total number of frames

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

The P_τ, also known as *percentage of correctly tracked frames*, is a frame-wise definition of the *true-positive* score, an interpretation that has become popular in tracking evaluation with the advent of tracking-by-detection concept. As noted in [7], the F-measure is another score that can be used in this context, however, it is worth noting that the detection based measures disregard the sequential nature of the tracking problem. As it is illustrated in Figure 3, these measures do

not necessarily account for complete trajectory reconstruction which is an important aspect in many tracking applications.

The most popular measures for multi-target tracking performance, the Multiple Object Tracking Precision (MOTP) and Multiple Object Tracking Accuracy (MOTA) [11] can also be seen in the context of single-object short-term tracking as an extension of region overlap measures. MOTP measure is defined as average overlap over all objects on all frames, taking into account different number of objects that are visible at different frames, i.e.

$$MOTP = \frac{\sum_{i=1}^M \sum_{t=1}^N \phi_{i,t}}{\sum_{t=1}^N M_t}, \quad (10)$$

where M denotes the number of different objects in the entire sequence and M_t denotes the number of visible objects at frame t. In single-target short-term tracking $M = M_t = 1$, therefore MOTP can be simplified to an average overlap measure, defined in equation (8) earlier in this section. The MOTA measure, on the other hand, takes into account three components that account for accuracy of multiple-object tracking algorithm: number of misses, number of false alarms and number of identity switches, i. e.

$$MOTA = 1 - \frac{\sum_{t=1}^N (c_m MI_t + c_f FP_t + c_s SW_t)}{\sum_{t=1}^N N_t^G}, \quad (11)$$

where MI_t denotes the number of misses, FP_t denotes the number of wrong detections, SW_t denotes the number of identity switches, c_m, c_f, and c_s, are weighting constants and N_t^G denotes the number of annotated objects at time t. In single-target short-term tracking scenario there is only one object ($N_t^G = 1$, $SW_t = 0$) whose location can and should always be determined ($FP_t = 0$, $MI_t \in \{0, 1\}$), which means that the MOTA measure can be simplified to the percentage of correctly tracked frames, defined in equation (9) earlier in this section.

C. Tracking length

Another measure that has been used in the literature to compare trackers is *tracking length* [38], [34]. This measure reports the number of successfully tracked frames from tracker's initialization to its (first) failure. A failure criterion can be a manual visual inspection (e.g. [37]), 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 the overlap measure (see Figure 4). The choice of the criterion may impact the result of comparison. As the overlap based criterion is more robust with respect to size changes, we will from now on 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 sequence up to the first tracking failure. If by some coincidence, the beginning of the video sequence contains a difficult tracking situation, or the
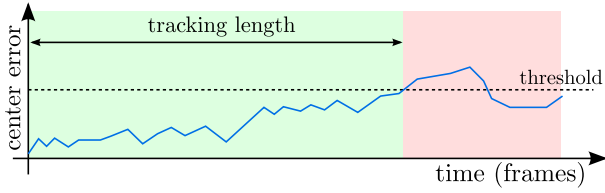
Fig. 4. An illustration of the tracking length measure for center error.

target is not visible well, which results in a necessarily poor initialization, the tracker will fail, and the remainder of the sequence will be discarded. This means that, technically, one would require a significant amount of sequences exhibiting the various properties right at its beginning to get a good statistic on this performance measure.

D. Failure rate

A measure that largely addresses the problem of the tracking length measure is the so-called *failure rate measure* [39], [23]. The failure rate measure casts the tracking problem as a supervised system in which a human 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 5. This measure also reflects the trackers performance in a real-world situation in which the human operator supervises the tracker and corrects its errors. Note that this performance measure should not be confused with different initialization strategies that can not be used as performance measures themselves (e.g. in [5] a tracker is initialized at different uniformly distributed positions in a sequence or with a random perturbation of the initialization region).
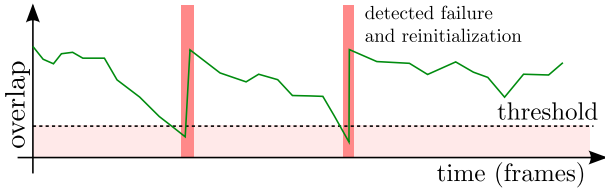

Fig. 5. An illustration of the failure rate measure for overlap distance.

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 [40], [41] consider a failure when the bounding box overlap is lower than 0.1. This lower threshold is 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

$$F_\tau = |\mathcal{F}_\tau|, \quad \mathcal{F}_\tau = \{f_i\}, \tag{12}$$

where $\mathcal{F}_\tau$ denotes the set of all failure frame numbers f_i. A drawback of the failure rate is that it does not reflect the distribution of these failures across the sequence. A tracker may fail uniformly in approximately equal intervals or it may fail more frequently at certain events. We can analyze these different distributions by looking at the *fragmentation* of the trajectory that is caused by the failures. Using an information theoretic point of view [42], we define the following trajectory fragmentation indicator, $\mathrm{Fr}(\mathcal{F}_\tau)$,

$$\mathrm{Fr}(\mathcal{F}_\tau) = \frac{1}{\log F_\tau} \sum_{f_i \in \mathcal{F}_\tau} -\frac{\Delta f_i}{N} \log \frac{\Delta f_i}{N},$$

$$\Delta f_i = \begin{cases} f_{i+1} - f_i & \text{when } f_i < \max(\mathcal{F}_\tau) \\ f_1 + N - f_i & \text{when } f_i = \max(\mathcal{F}_\tau) \end{cases}, \tag{13}$$

where F denotes the number of failures and f_i denotes the position of the i-th failure. The special case for the last failure ensures that the resulting value is not distorted by the beginning and end of the sequence[2]. Fragmentation is only meaningful when $|\mathcal{F}_\tau| > 1$ as we are observing the inter-failure intervals. Maximum value 1 is reached when the failures are uniformly distributed over the sequence and the value decreases when the inter-failure intervals become unevenly distributed. Note that the fragmentation can only be used as a supplementary indicator to the failure rate since it contains only limited information about the performance of a tracker, e.g. it will produce the same value for trackers that fail uniformly throughout the sequence no matter how many times they fail. However, it can be used to discriminate between trackers that fail frequently at a specific interval and those that fail uniformly over the entire sequence. As the evaluation datasets are getting larger, additional scores like fragmentation can help interpreting results on a higher level which we will demonstrate in Section III.

E. Hybrid measures

Nawaz and Cavallaro [6] 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 and a failure score. High score indicates poor tracking performance. The intuition behind CoTPS is illustrated in Figure 6. At a glance, an appealing property of this measure is that it orders trackers by accounting for two separate aspects of tracking. However, no justification, neither theoretical nor experimental, is given of such rather complicated fusion which makes interpretation of this measure rather difficult. It can be shown (see Appendix A) that the CoTPS measure can be reformulated in terms of average

---

[2]We interpret the sequence as a circular time-series and join the first and the last fragment. This way the value of Fr stays the same for the shifts of the same distribution of failures.

overlap, $\bar{\phi}$, and percentage of failure frames (where overlap is 0), λ_0, i.e.

$$CoTPS = 1 - \bar{\phi} - (1 - \lambda_0)\lambda_0. \qquad (14)$$

The equation (14) conclusively states that two very different basic measures are being combined in a rather complicated manner, prohibiting a straightforward interpretation. Precisely, if one tracker is ranked higher than another one it is not clear if this is due to a higher average overlap or less failed frames. Furthermore, if equation (14) is reformulated as $CoTPS = (1 - \lambda_0)(1 - \hat{\phi}) + \lambda_0^2$, where the $\hat{\phi}$ denotes the average overlap on non-failure frames (where the overlap is greater than 0), multiple combinations of two values produce the same CoTPS score. In Figure 7 we illustrate several such equality classes, where the same CoTPS score is achieved using different combinations of the two components, which makes the interpretation of the results difficult. The combined score is also inconvenient in scenarios where a different combination of performance properties is desired.
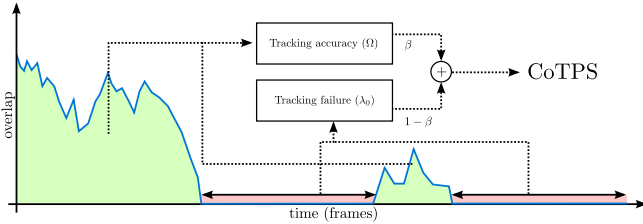

Fig. 6. An illustration of the the CoTPS measure as described in [6].

In terms of performance score, we therefore believe that a better strategy is to focus on a few complementary performance measures with a well-defined meaning, and avoid fusing them into a single measure early on in the evaluation process.

F. Performance plots

Plots are frequently used to visualize the behavior of a tracker since they offer a clearer overview of performance when considering multiple trackers or sets of tracker parameters. The most widely-used plot is a center-error plot that shows the center-error with respect to the frame number [30], [31], [35], [36]. While this kind of plots 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 8 where two trackers appear equal due to a distorted scale caused by the third tracker. A less popular but better bounded alternative approach is to plot region overlap, e.g. in [33].

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 order and can also be potentially misused to influence the results of a comparison. However, it is sometimes better to avoid the use of a single 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 have some resemblances to a ROC curve [43], like monotony, intuitive visual comparison, and a similar calculation algorithm. Measure-threshold plots were used in [30], where the authors used center-error as a measure as well as in [5], 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 9). 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 [5], [6] to reason about the performance of the trackers. However, the authors of [5], [6] do not provide an interpretation of this score. We prove in this paper (see Appendix B) that the AUC is in fact the average overlap, which results in two important implications: (1) the complicated computation of ROC-like curve and subsequent numerical integration for calculating AUC can be avoided by simple averaging of overlap over the sequence and (2) the AUC has a straight-forward interpretation.

A curve that is visually similar to P_τ plot is the *survival curve* [7]. In this case the curve summarizes the trackers' success (various performance measures can be used) over a dataset of sequences that are ordered from the best performance to the worst. While this approach gives a good overview of the overall success, it is not suitable for a sequence-wise comparison as the order of sequences differs from tracker to tracker. Not all sequences are equal in terms of difficulty as well as in terms of the phenomena that they contain (e.g. occlusion, illumination changes, blur) which makes it very hard to interpret the results of a survival curve on a more detailed level.

III. EXPERIMENTAL COMPARISON OF PERFORMANCE MEASURES

The theoretical analysis so far shows that different measures may reflect different aspects of tracking performance, so it is impossible to simply say which the best measure is. Furthermore some measures are proven to be equal (e.g., area-under-the-curve and average overlap). We start our analysis by establishing similarities and equivalence between various measures, by experimentally analyzing which measures produce consistently similar responses in tracker comparison. The main idea is that strongly correlated measures are sensitive the same quality of a visual tracker, therefore we should only consider a subset of measures that are not correlated or at most weakly correlated.

In order to analyze the performance measures, we have conducted a comparative experiment. Our goal is to evaluate
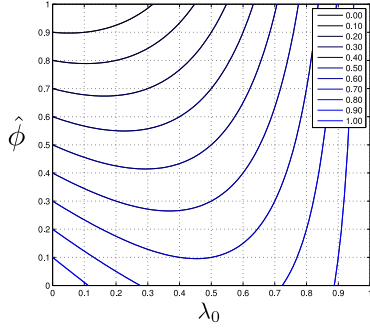
Fig. 7. Equality classes for different values of CoTPS measure. Each line denotes pairs of average overlap on non-failed frames ($\hat{\phi}$) and percentage of failure frames (λ_0) that produce the same CoTPS score.
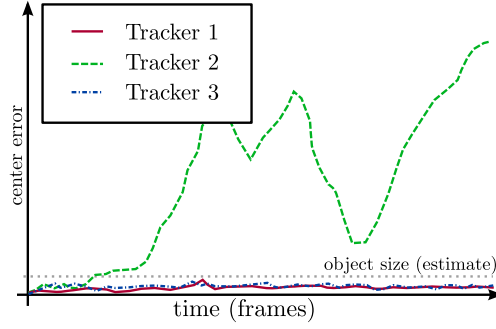
Fig. 8. An example of center-error plot comparison for three trackers. Tracker 2 has clearly failed in the process, yet its large center errors cause the plot to expand its vertical scale, thus reducing the apparent differences of trackers 1 and 3.
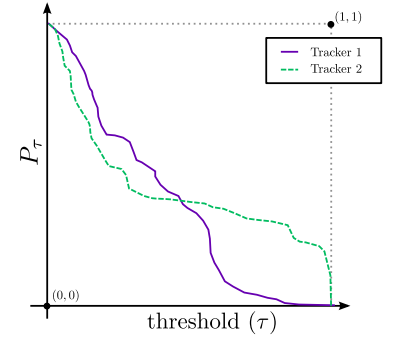
Fig. 9. An illustration of the *measure-threshold* plot for two trackers. It is apparent that different values of the threshold would clearly yield different order of the trackers.

several existing trackers according to the selected measures on a number of typical visual tracking sequences. The selection of measures is based on our theoretical discussion in Section II. We have selected the following measures:

1) average center error (Section II-A),
2) average normalized center error (Section II-A),
3) root-mean-square error (Section II-A),
4) percent of correct frames for $\tau = 0.1$, $P_{0.1}$ (Section II-B),
5) percent of correct frames for $\tau = 0.5$, $P_{0.5}$,
6) tracking length for threshold $\tau > 0.1$, $L_{0.1}$ (Section II-C),
7) tracking length for threshold $\tau > 0.5$, $L_{0.5}$,
8) average overlap (Section II-B),
9) Hybrid CoTPS measure (Section II-E),
10) average center error for F_0,
11) average normalized center error for F_0,
12) root-mean-square error for F_0,
13) percent of correct frames for $\tau = 0.1$, $P_{0.1}$ for F_0,
14) percent of correct frames for $\tau = 0.5$, $P_{0.5}$ for F_0,
15) average overlap in case of F_0,
16) failure rate F_0 (Section II-D).

The first nine measures were calculated on trajectories where the tracker was initialized only at the beginning of the sequence, and the remaining seven measures were calculated on trajectories where the tracker was reinitialized if the overlap between predicted and ground-truth region became 0.

Since the goal of the experiment is not evaluation of trackers but selection of measures, the main guideline when selecting trackers for the experiment was to create a diverse set of tracking approaches that fail in different scenarios and are therefore capable of showing differences of evaluated measures on real tracking examples. We have selected a diverse set of 16 trackers, containing various detection-based trackers, holistic generative trackers, and part-based trackers, that were proposed in the recent years: A color-based particle filter (PF) [44], the On-line boosting tracker (OBT) [45], the Flock-of-features tracker (FOF) [46], the Basin-hopping Monte Carlo tracker (BHMC) [38], the Incremental visual tracker (IVT) [29], the Histograms-of-blocks tracker (BH) [47], the Multiple instance tracker (MIL) [30], the Fragment tracker (FRT) [31], the P-N tracker (TLD) [8], the Local-global tracker (LGT) [41],

Hough tracker (HT) [37], the L1 Tracker Using Accelerated Proximal Gradient Approach (L1-APG) [35], the Compressive tracker (CT) [36], the Structured SVM tracker (STR) [48], the Kernelized Correlation Filter tracker (KCF) [49], and the Spatio-temporal Context tracker (STC) [50]. The source code of the trackers was provided by the authors and adapted to fit into our evaluation framework.

We have run the trackers on 25 different sequences, most of which are well-known in the visual tracking community, e.g. [41], [40], [36], [4], [29], [38], [31], [37], and several were acquired additionally. Representative images from the sequences are shown in Figure 10. The sequences were annotated with an axis-aligned bounding-box region of the object (if the annotations were not already available), as well as the central point of the object, in cases where the center of the object did not match the center of the bounding-box region. To account for stochastic processes that are a part of many trackers, each tracker was executed on each sequence 30 times. The 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 III. Additional results, such as the ordering of the trackers according to individual measures, are available in the supplementary material[3].

A. Correlation analysis

A correlation matrix was computed from all pairs of measures calculated over all tracker-sequence pairs. Note that we do not calculate the correlation on rankings 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 order for trackers. Since we have run 16 trackers, each of the stochastic ones was run 30 times on every sequence, this means that every performance measure has

[3]Supplementary material is available at http://go.vicos.si/ performancemeasures.
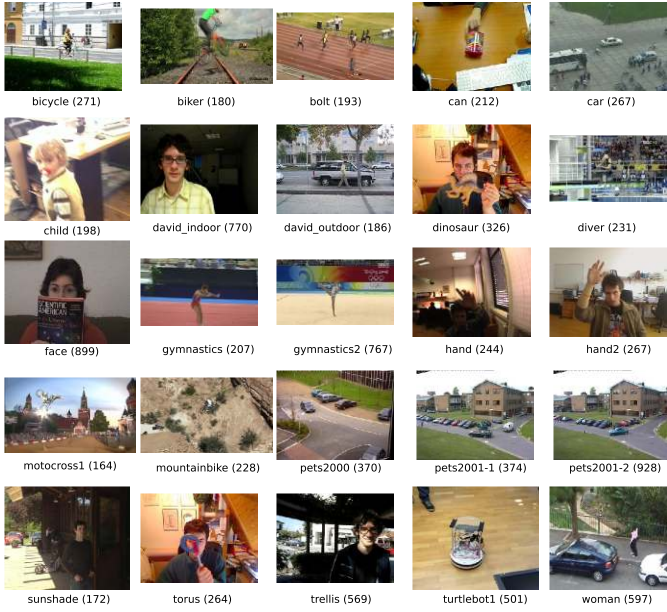
Fig. 10. Overview of the sequences used in the experiment. The number in brackets besides the name denotes the length of a sequence in frames.

about 10000 samples. This is more than enough for statistical evaluation of whether correlation across the measures exists. The obtained correlation matrix is shown in Figure 11. Using automatic cluster discovery by affinity propagation [51] we have determined five distinct clusters, one for measures 1 to 3, one for measures 4 to 9, one for measures 10 to 13, one for measures 14 and 15, and one for measure 16. All these correlations are highly statistically significant ($p < 0.001$).
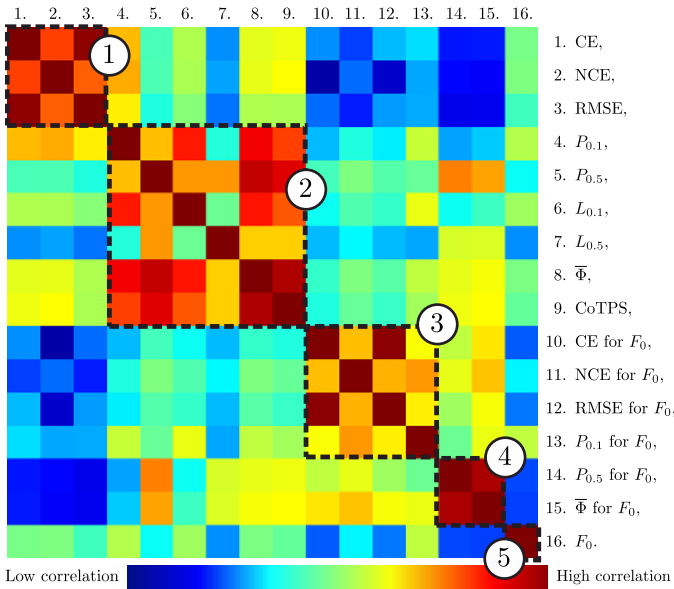

Fig. 11. Correlation matrix for all measures visualized as a heat-map overlaid with obtained clusters. The image is best viewed in color.

The first cluster of measures consists of the three center-error-based measures. This is expected since all of these measures are based 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* ($L_{0.1}$ and $L_{0.5}$). 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. Because of this and the insensitivity to the scale changes they are a better choice to measure tracking performance than the center-error-based measures. An illustration of this difference for *overlap* and *center-error* is shown as a graph in Figure 12, 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.
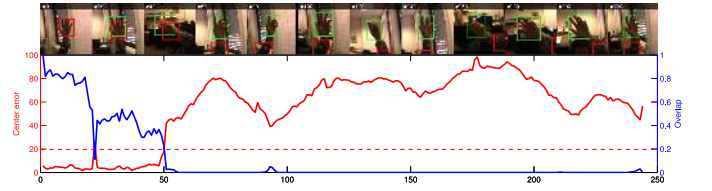

Fig. 12. A comparison of *overlap* and *center error distance* measures for tracker CT on sequence *hand* [40]. The dashed line shows the estimated threshold above which the center error is greater than the size of the object. The tracker fails around frame 50.

The first cluster of measures in Figure 11 implies that the first three measures are equivalent and it does not matter which one is chosen. The second cluster requires further interpretation. Despite the apparent similarity of overlap-based measures 4 to 8 and of the CoTPS measure, the correlation is not perfect and the order of trackers differ in some cases. One example of such a difference can be seen for the TLD tracker on the *woman* sequence (Figure 13). 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 8) and the *percentage of correct frames* (Measures 4 and 5) therefore order the tracker higher than the *tracking length* (Measures 6 and 7). 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 $L_{0.1}$ and $L_{0.5}$ and to some extent for the percentage of correct frames measures $P_{0.1}$ and $P_{0.5}$. In those cases, the scores for a higher threshold (0.5) result in a different order of trackers compared to the lower threshold (0.1). This means that care must be taken when choosing the thresholds as they may affect the outcome of the evaluation. While a certain threshold may be given for a specific application domain, it is best to avoid it in general performance evaluation. The last measure in the second cluster is the hybrid CoTPS measure [6] which turns out to be especially strongly correlated with the average overlap measure. By looking back at our theoretical analysis in Section II-E the CoTPS produces identical results for trajectories where the overlap never reaches 0 (no failure). In other cases the percent of failed frames, which can be approximated using $1 - P_{0.1}$, is also strongly correlated with average overlap. This means that the entire measure is biased

towards only one aspect of tracking performance.

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 II 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.
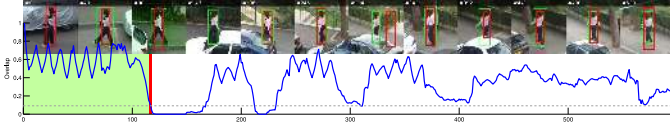

Fig. 13. An overlap plot for tracker TLD on sequence *woman* [31]. The dashed line shows the threshold below which the tracking length detects failure (for threshold 0.1), which happens around frame 120.

The *failure rate* measure influences the trackers' entire trajectory, because of the re-initializations. The data for measures 9 to 16 was therefore acquired as a separate experiment. The advantage of the supervised tracking scenario 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. While supervised evaluation looks more complex, this is a technical issue that can be solved with standardization of evaluation process [52]. In Figure 14 we can see the performance of the LGT tracker on the *bicycle* sequence. Because of a short partial occlusion near frame 175 the tracker fails, although it is clearly capable of tracking the rest of the sequence reliably if re-initialized. Measures that are computed on the trajectories with reinitialization exhibit similar correlation relations than for the trajectories without reinitialization.

According to the correlation analysis the least correlated measures are failure rate and average overlap on re-initialized trajectories. These findings are discussed in next section where we propose a conceptual framework for their joint interpretation. To further support the stability of the measurements, we have also performed the correlation analysis on different subsets of approximately half of the total 25 sequences and found that the these findings do not change.
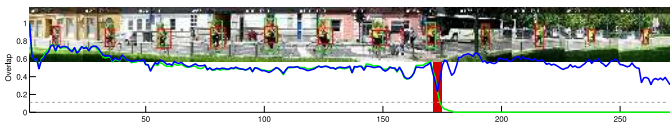

Fig. 14. An overlap plot for tracker LGT on sequence *bicycle* [41]. The green plot shows the unsupervised overlap, and the blue plot shows the overlap for supervised tracking, where the failure is recorded and the tracker re-initialized.

B. Accuracy vs. robustness

An intuitive way to present tracker performance is in terms of accuracy (i.e., how accurately the tracker determines the position of the object) and robustness (i.e., how many times the tracker fails). Based on the correlation analysis in Section III-A we have selected a pair of evaluated measures that estimates the aforementioned qualities. The *average overlap* measure is the best choice for measuring the accuracy of a tracker because it takes into account the size of the object and does not require a threshold parameter. However, it does not tell us much about the robustness of the tracker, especially if the tracker fails early in the sequence. The *failure rate* measure, on the other hand, measures the number of the failures which can be interpreted as robustness of the tracker. According to correlation analysis in Section III-A, if we measure average overlap on the re-initialized data, used to estimate failure rate, the measures are not correlated. This is a desired property as they should measure different aspects of tracker performance. We thus propose measuring the short-term tracking performance by the following A-R pair,

$$\text{A-R}(\Lambda^G, \Lambda^T) = \left(\overline{\Phi}(\Lambda^G, \Lambda^T), F_0(\Lambda^G, \Lambda^T) \right), \quad (15)$$

where $\overline{\Phi}$ denotes *average overlap* and F_0 denotes the *failure rate* for $\tau = 0$. Note that the value of failure threshold τ can influence the final results. If the value is set to a high value (i.e. close to 1) the tracker is restarted frequently even for small errors and the final score is hard to interpret. Based on our analysis, 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 an illustrative example of how the two measures complement each other in accurately describing the tracking performance.

It is worth noting that there are some parallels between the hybrid CoTPS measure [6], and the proposed A-R measure pair. In both cases two aspects of tracking performance are considered. The first part of the CoTPS measure is based on the AUC of the overlap plot, which, as we have shown, is equal to average overlap. The second part of the measure attempts to report tracker failure by measuring the number of frames where the tracker has failed (overlap is 0), which could also be written as P_0. Despite these apparent similarities, the A-R measure pair is better suited for visual tracker evaluation for several reasons: (1) the chosen measures are not correlated, (2) the supervised evaluation protocol uses sequences more effectively because of reinitializations, (3) different performance profiles for average overlap and failure rate produce different combinations of scores that can be interpreted, which is not true for CoTPS measure.

A pair of measures is most efficiently represented via visualization. We propose to visualize the A-R pair as a 2-D scatter plot. This kind of visualization is indeed very simple, but is easy to interpret, extendable and has been used in visual tracking visualization before, e.g. [53]. An example of an A-R plot for the data from the experiment can be seen in
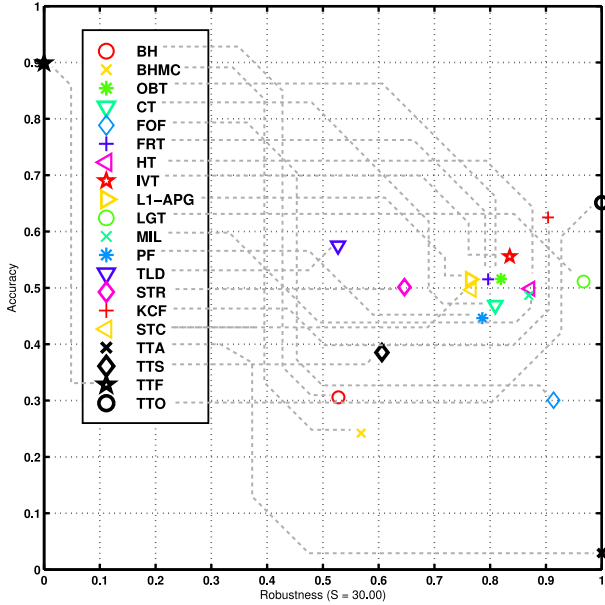
Fig. 15. An accuracy-reliability data visualization for all trackers over all sequences.

Figure 15, where we show the average scores for all sequences, from which one can read the trackers performance in terms of accuracy (the tracker is more accurate if it is higher along the vertical axis) and robustness (the tracker fails fewer times if it is further to the right on the horizontal axis). Because the robustness does not have an upper bound we propose to interpret it as *reliability* for visualization purposes. The reliability of a tracker is defined as an exponential failure distribution, $R_S = e^{-SM}$. The value of M denotes mean-time-between-failures, i.e. $M = \frac{F_0}{N}$, where N is the length of the sequence. The reliability of a tracker can be interpreted as a probability that the tracker will still successfully track the object up to S frames since the last failure, assuming a uniform failure distribution that does not depend on previous failures. This is of course not true in all cases, however, note that this formulation and the choice of S does not influence the order of the trackers but can be adjusted as a scaling factor for better visualization. Interpreting results this way is useful for visualization and quick interpretation of results, however, one should still consult the detailed values of average overlap and failure rate before making any final decisions.

C. Theoretical trackers

For a better understanding of the complementing nature of the two measures we introduce four theoretical trackers denoting extreme prototypical tracker behaviors. The *first theoretical tracker*, denoted by TTA, always reports the region of the object to equal the image size of the sequence. This tracker provides regions that are too loose, but does not fail (overlap is never 0) and is therefore displayed in the bottom-right corner as it is extremely robust, but not accurate at all. The *second theoretical tracker*, denoted by TTS, reports its initial position for the entire sequence. This tracker will likely fail if the object moves, and will achieve better accuracy

because of frequent manual interventions. The *third theoretical tracker*, denoted by TTF only tracks one frame and then deliberately reports a failure. This way the tracker maintains a high accuracy, however the failure rate is extremely high and the tracker is placed in top-left corner of the plot. The *fourth theoretical tracker* is denoted as TTO and represents an *oracle* tracker of fixed size. The tracker always correctly predicts the center position of the object, however, the size of the object is fixed. This tracker represents a practical performance limit for trackers that do not adapt the size of the reported bounding box which is the same as the initialization bounding box.

The performance scores for the theoretical trackers can be easily computed directly from ground-truth. The simplicity, intuitive nature, and the parameter-less design make them excellent interpretation guides in the graphical representations of results, such as A-R plot. In other words, they put the results of evaluated trackers into context by providing reference points for a given evaluation sequence.

D. Interpretation of results using A-R plots

By establishing the selection of measures, visualization and the theoretical trackers as an interpretation guide, we can now provide an example of results interpretation. The A-R plot in Figure 15 shows results, averaged over entire data-set. We can see that the LGT tracker is on average the most robust one in the set of evaluated trackers (positioned most right), but is surpassed in terms of accuracy by KCF, IVT and TLD (positioned higher). Especially the TLD tracker is positioned very low in terms of robustness, so the high accuracy may in fact be a result of frequent reinitializations, a behavior that is similar to the TTF tracker. We acknowledge that this behavior of TLD is a design decision as the TLD is actually a long-term tracker that that does not report the position of the object if it is not certain about its location. The FOF tracker, on the other hand, is quite robust, but its accuracy is very low. This means that it most likely sacrifices accuracy by spreading accross a large portion of the frame, much like TTA.

As the averaged results can convey only a limited amount of information, we have also included per-sequence A-R plots in the supplementary material. These plots show that the actual performance of trackers differs significantly between the sequences. Theoretical trackers TTA and TTF remain worse on their individual axes as expected, while the relative position of the other trackers changes depending on the properties of the individual sequence. In many sequences the TTO tracker achieves the best performance because of its ability to "predict" the position of the target. In cases where the size of the object changes this advantage becomes less apparent and trackers like IVT, L1-APG, HT, and LGT that account for this change can even surpass it in terms of accuracy (e.g. in *biker*, *child*, and *pets2001-2*). The sequence *diver* is interesting considering the results. Even though the object does not move a lot in the image space, which is apparent from the high robustness of the TTS tracker, the sequence has nevertheless proven to be very challenging for most of the trackers because of the large deformations of the object. The BH and BHMC trackers are on average very similar to the TTS tracker

which would mean that they do not cope well with moving objects. At a closer look we can see that this is only true for some sequences (e.g. *torus*, *bicycle*, and *pets2000*). In other sequences both tracker perform either better than TTS, where the background remains static and can be well separated from the object (e.g. *sunshade*, *david_outdoor*, and *gymnastics2*), or worse, where the appearance of the background changes (e.g. *motocross1*, *child*, and *david_indoor*). Considering the good average performance of the LGT tracker we can see that the tracker performed well in sequences with articulated and non-rigid objects (e.g. *hand*, *hand2*, *dinosaur*, *can*, and *torus*), while the difference in case of more rigid objects (e.g. *face*, *pets2001-1*, and *pets2001-2*) is less apparent. In the plot for the *bolt* sequence we can see that the TLD tracker behaves similarly to the TTF tracker, i.e. fails a lot without actually drifting. On the other hand the TLD tracker works quite well in the case of *pets2000*, *pets2001-1*, and *pets2001-2* sequences where the changes in the appearance of the object are gradual.

E. Fragmentation

Recall that we have introduced the fragmentation indicator as a complementary indicator for the number of failures measure in Equation 13. Using this measure we can infer some additional properties of a tracker that would otherwise require looking at raw results. Fragmentation reflects the distribution of failures throughout the sequence. If the fragmentation is low then the failures are likely clustered together around some specific event (which can indicate a specific event that is problematic for the tracker). On the other hand, if the fragmentation is high, then the failures are uniformly distributed, independently of localized events in the sequence and can be most likely attributed to internal problems of the tracker. To demonstrate this property we have selected several cases where the number of failures is the same, but the fragmentation is different. In Figure 16 we can see three such cases. Several trackers, despite failing the same number of times do this for different reasons and in different intervals. On the *hand* sequence, the FRT tracker fails almost uniformly, while the BH tracker manages to hold to the target for a long time (the region is, however, estimated very poorly), but then fails to successfully initialize around frame 170 because of background clutter and motion. In *bicycle* and *bolt* sequences, the failures of PF tracker are concentrated on a specific event, most likely because of color ambiguity or small target size. The failures of the BHMC tracker are almost uniformly distributed over both sequences, most likely because of the problems of the tracker implementation (e.g. inability to cope with small target size).

F. Sequences from perspective of theoretical trackers

The theoretical trackers, introduced in Section III-C, provide further insights into each sequence from the perspective of the basic properties that each theoretical tracker represents. Because of their simplicity and absence of parameters, they can easily be applied to any annotated sequence and provide some insight about its properties. These properties can then be
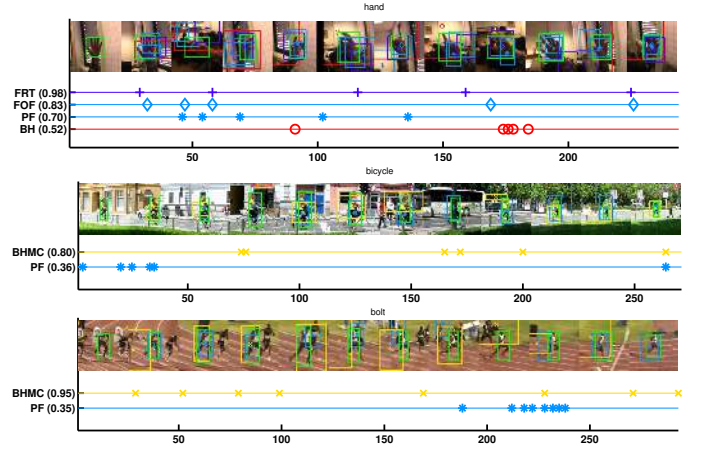

Fig. 16. Selected results of the fragmentation analysis. Failures are marked on the time-line with symbols, the corresponding fragmentation values are shown in brackets next to tracker name.

used when constructing an evaluation dataset or interpreting the results.

The TTA tracker will always achieve good robustness (no failures), but will produce high accuracy values only when the target will cover large part of the image frame. This tracker therefore measures the average relative size of the object. The TTS tracker will only achieve good robustness when the object remains stationary with respect to the image plane (e.g. the *diver* and the *face* sequence) and will also achieve good robustness when the size of the object does not change with respect to the initialization frame. The TTF tracker will fail uniformly, however it will produce high accuracy only when there is no rapid motion predominantly present over the entire sequence, like in sequences *hand*, *hand2*, and *sunshade*. The TTO tracker will achieve good robustness (no failures), however, it will not achieve good accuracy when the size of the object region changes a lot, e.g. in sequences *diver* and *gymnastics*. These observations can be extended to the entire set of sequences using clustering. As a demonstration we have used K-means clustering with expected number of clusters set to $K = 3$ to generate labels that are shown in Table I. The labels are of course relative to the entire set, but they summarize these relative properties well, e.g. we can see that *face* sequence is similar to *diver* sequence in terms of movement, however, the *diver* sequence contains a lot of size changes. This simple approach could be in future extended to provide automated and less-biased sequence descriptions.

IV. CONCLUSION

In this paper we have addressed the problem of performance evaluation in monocular single-target short-term visual tracking. Through theoretical and experimental analysis we have investigated various popular performance evaluation measures, discussed their pitfalls and showed that many of the widely used measures are equivalent. Since some measures reflect certain aspect of tracking performance, combining those that address the same aspect provides no additional information regarding the performance or even introduces bias toward a

TABLE I
SEQUENCE PROPERTIES ACCORDING TO THEORETICAL TRACKER
PERFORMANCE.

	Size (TTA)	Motion (TTS)	Speed (TTF)	Size change (TTO)
bicycle	small	high	medium	medium
biker	large	medium	low	high
bolt	small	high	medium	medium
can	medium	high	medium	low
car	small	medium	medium	low
child	large	medium	medium	high
david_indoor	small	low	medium	low
david_outdoor	small	high	medium	low
dinosaur	large	medium	low	medium
diver	small	low	medium	high
face	medium	low	low	low
gymnastics	medium	low	medium	high
gymnastics2	small	low	low	medium
hand	small	high	high	medium
hand2	small	high	high	medium
motocross1	medium	high	medium	high
mountainbike	small	medium	low	medium
pets2000	small	medium	low	medium
pets2001-1	small	medium	low	high
pets2001-2	small	medium	low	high
sunshade	small	high	high	low
torus	small	high	medium	low
trellis	small	low	medium	high
turtlebot1	medium	medium	low	medium
woman	small	medium	medium	medium

certain aspect of performance to the result. Based on the results of our experiment we have proposed to use a pair of two existing complementary measures. This pair, that we call the A-R pair, takes into account the accuracy (using *average overlap*) and the robustness (using *failure rate*) of each tracker. We have also proposed an intuitive way of visualizing the results in a 2-dimensional scatter plot, called the A-R plot. Additionally, we have introduced fragmentation as an additional indicator for distribution of failures. We have introduced several theoretical trackers that can be used to quickly review the results of the evaluated trackers in terms of basic properties that the theoretical trackers exhibit. We have also shown that the theoretical trackers can be used for automatic annotation of sequence properties from a tracker viewpoint.

While narrowing down the abundance of performance measures is a big step toward homogenizing the tracking evaluation methodology, this is only one of the requirements for a consistent evaluation methodology for visual trackers. The measures that were proposed in this paper have already been adopted as the foundation of the evaluation methodology of a recently organized visual tracking challenges VOT2013 [13] and VOT2014 [14], where a rigorous analysis in terms of accuracy and robustness has provided multiple interesting insights into performance of individual trackers, e.g. we have shown that some trackers are more robust, but less accurate, while some sacrifice robustness for greater accuracy. In our future work we will extend the automatic labeling of sequences using both theoretical and practically applicable trackers as well as investigate the question how to reduce the number of annotated frames without degrading the performance estimates [12].

APPENDIX A
REFORMULATION OF CoTPS [6] MEASURE

Let $\phi_1, \phi_2, \ldots, \phi_N$ be frame overlaps for a sequence of length N. In [6], the CoTPS measure is defined as a weighted average of two factors, that the authors define as *tracking accuracy*, Ω, and *tracking failure*, λ_0, that are combined using a dynamically computed factor, β, as

$$CoTPS = \beta\Omega + (1 - \beta)\lambda_0. \tag{16}$$

The tracking failure factor λ_0 is computed as the percentage of frames where the tracker failed, i.e. $\lambda_0 = \frac{N_0}{N}$, where N_0 is a number of frames where the overlap between ground-truth region and the predicted region is 0. The weight factor is defined as $\beta = \frac{\hat{N}}{N}$, where $\hat{N}$ denotes the number of frames where the overlap is higher than 0, therefore $\beta = 1 - \lambda_0$. The definition for tracking accuracy part Ω is

$$\Omega = \sum_{\tau \in (0,1]} \frac{\hat{N}(\tau)}{\hat{N}}, \tag{17}$$

where $N(\tau) = |\{j : \phi_j \geq 0 \wedge \phi_j \leq \tau\}|$ denotes the number of frames that is higher than 0, but lower than τ. We observe that (17) is actually an approximation of the integral with respect to threshold τ, that can also be reformulated as

$$\Omega = \int_0^1 \frac{\hat{N}_\tau}{\hat{N}} d\tau = 1 - \int_0^1 \frac{\hat{P}_\tau}{\hat{N}} d\tau, \tag{18}$$

where $P(\tau) = |\{j : \phi_j \geq \tau\}|$. According to the proof in Appendix B, the integral results in average overlap over a set of frames, therefore $\Omega = 1 - \hat{\phi}$, where $\hat{\phi}$ is the average overlap over $\{\phi_j : \phi_j \geq 0\}$. Therefore, the CoTPS measure can be rewritten as

$$CoTPS = (1 - \lambda_0)(1 - \hat{\phi}) + \lambda_0^2. \tag{19}$$

Considering that average overlap over the entire sequence can be written as $\phi = (1 - \lambda_0)\hat{\phi}$, we can further derive

$$CoTPS = 1 - \bar{\phi} - (1 - \lambda_0)\lambda_0, \tag{20}$$

meaning that the CoTPS measure is a function of average overlap as well as the percentage of frames where the overlap is 0.

APPENDIX B
PROOF THAT AUC OF [5] EQUALS TO AVERAGE OVERLAP

Problem: Let $\phi_1, \phi_2, \ldots, \phi_N$ be frame overlaps for a sequence of length N. We assume that the frame overlaps are *ordered* by scale from minimal to maximal value and $\phi_0 = 0$, i.e.

$$0 = \phi_0 \leq \phi_1 \leq \cdots \leq \phi_N.$$

Let $P(\tau) = |\{j : \phi_j \geq \tau\}|$ be the number of overlaps greater than τ. The AUC measure is an integral of $\frac{P(\tau)}{N}$ from 0 to 1. We want to prove that the average overlap, $\hat{\phi}$, for the sequence $\phi_1, \phi_2, \ldots, \phi_N$ equals to the computed AUC, i.e.

$$\frac{1}{N}\sum_{i=1}^{N}\phi_i = \frac{1}{N}\int_0^1 P(\tau)d\tau.$$

Proof: Function P is a step function (constant between ϕ_i and ϕ_{i+1}). Therefore its integral I is

$$I = \sum_{i=0}^{N-1} P(\phi_i)(\phi_{i+1} - \phi_i).$$

The sum can be reorganized in the following way:

$$
\begin{aligned}
I &= P(\phi_0)(\phi_1-\phi_0)+P(\phi_1)(\phi_2-\phi_1)+P(\phi_2)(\phi_3-\phi_2)+\cdots \\
&= \phi_1 P(\phi_0)-\phi_0 P(\phi_0)+\phi_2 P(\phi_2)-\phi_1 P(\phi_1)+\phi_3 P(\phi_3)-\cdots \\
&= -\phi_0 P(\phi_0)+\phi_1(P(\phi_0)-P(\phi_1))+\phi_2(P(\phi_1)-P(\phi_2))\cdots \\
&= 0\cdot P(\phi_0) + \phi_1\cdot 1 + \phi_2\cdot 1 + \cdots & (21)\\
&= \phi_0 + \phi_1 + \phi_2 + \cdots & (22)\\
&= \sum_{i=1}^{N}\phi_i.
\end{aligned}
$$

In (21) we have assumed that the shift between the two consequential values of $P(\tau)$, i.e. $P(\phi_i) - P(\phi_{i+1})$ equals to 1, that is true if all ϕ_i are different. If k consequential ϕ_i are equal then the corresponding $k-1$ shifts are 0, while the last one is k. However, in (21) we add $(\phi_i\cdot 1)$ k times. ∎

REFERENCES

[1] M. Everingham, L. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge," *IJCV*, vol. 88, no. 2, pp. 303–338, Sep. 2009.

[2] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski, "A Database and Evaluation Methodology for Optical Flow," *IJCV*, vol. 92, no. 1, pp. 1–31, Nov. 2010.

[3] S. Alpert, M. Galun, R. Basri, and A. Brandt, "Image Segmentation by Probabilistic Bottom-Up Aggregation and Cue Integration," in *CVPR*. IEEE, Jun. 2007, pp. 1–8.

[4] Q. Wang, F. Chen, W. Xu, and M.-H. Yang, "An experimental comparison of online object-tracking algorithms," in *SPIE*, San Diego, 2011, pp. 81 381A–81 381A–11.

[5] Y. Wu, J. Lim, and M.-h. Yang, "Online Object Tracking: A Benchmark," in *CVPR*, 2013.

[6] T. Nawaz and A. Cavallaro, "A protocol for evaluating video trackers under real-world conditions," *Image Process. IEEE Trans.*, vol. 22, no. 4, pp. 1354–-1361, 2013.

[7] A. W. M. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah, "Visual Tracking: an Experimental Survey," *TPAMI*, 2013.

[8] Z. Kalal, J. Matas, and K. Mikolajczyk, "P-N learning: Bootstrapping binary classifiers by structural constraints," in *CVPR*. San Francisco, CA, USA: IEEE, Jun. 2010, pp. 49–56.

[9] C. Ma, X. Yang, C. Zhang, and M.-H. Yang, "Long-Term Correlation Tracking," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015.

[10] Z. Hong, Z. Chen, C. Wang, X. Mei, D. Prokhorov, and D. Tao, "MUlti-Store Tracker (MUSTer): A Cognitive Psychology Inspired Approach to Object Tracking," in *CVPR 2015*, Jun. 2015.

[11] 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*, vol. 31, no. 2, pp. 319–36, Feb. 2009.

[12] P. Carvalho, J. S. Cardoso, and L. Corte-Real, "Filling the gap in quality assessment of video object tracking," *Image Vis. Comput.*, vol. 30, no. 9, pp. 630–640, Sep. 2012.

[13] M. Kristan, R. Pflugfelder, A. Leonardis, J. Matas, F. Porikli, L. Čehovin, G. Nebehay, G. Fernandez, T. Vojir, A. Gatt, A. Khajenezhad, A. Salahledin, A. Soltani-Farani, A. Zarezade, A. Petrosino, A. Milton, B. Bozorgtabar, B. Li, C. S. Chan, C. Heng, D. Ward, D. Kearney, D. Monekosso, H. C. Karaimer, H. R. Rabiee, J. Zhu, J. Gao, J. Xiao, J. Zhang, J. Xing, K. Huang, K. Lebeda, L. Cao, M. E. Maresca, M. K. Lim, M. E. Helw, M. Felsberg, P. Remagnino, R. Bowden, R. Goecke, R. Stolkin, S. Y. Lim, S. Maher, S. Poullot, S. Wong, S. Satoh, W. Chen, W. Hu, X. Zhang, Y. Li, and Z. Niu, "The Visual Object Tracking VOT2013 challenge results," in *ICCV Work.*, 2013, pp. 98–111.

[14] M. Kristan, R. Pflugfelder, A. Leonardis, J. Matas, L. Čehovin, G. Nebehay, T. Vojir, G. Fernández, A. Lukežič, A. Dimitriev, A. Petrosino, A. Saffari, B. Li, B. Han, C. Heng, C. Garcia, D. Pangeršič, G. Häger, F. S. Khan, F. Oven, H. Possegger, H. Bischof, H. Nam, J. Zhu, J. Li, J. Y. Choi, J.-W. Choi, J. a. F. Henriques, J. van de Weijer, J. Batista, K. Lebeda, K. Öfjäll, K. M. Yi, L. Qin, L. Wen, M. E. Maresca, M. Danelljan, M. Felsberg, M.-M. Cheng, P. Torr, Q. Huang, R. Bowden, S. Hare, S. Y. Lim, S. Hong, S. Liao, S. Hadfield, S. Z. Li, S. Duffner, S. Golodetz, T. Mauthner, V. Vineet, W. Lin, Y. Li, Y. Qi, Z. Lei, and Z. Niu, "The Visual Object Tracking VOT2014 challenge results," in *ECCV Work.*, 2014.

[15] K. Smith, D. Gatica-Perez, and J. Odobez, "Evaluating Multi-Object Tracking," in *CVPR Work.*, vol. 3. IEEE, 2005, pp. 36–36.

[16] J. Black, T. Ellis, and P. Rosin, "A novel method for video tracking performance evaluation," in *VS-PETS*, 2003, pp. 125–132.

[17] 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*, 2005, pp. 1–8.

[18] E. K. Kao, M. P. Daggett, and M. B. Hurley, "An information theoretic approach for tracker performance evaluation," in *CVPR*. IEEE, Sep. 2009, pp. 1523–1529.

[19] V. Manohar, P. Soundararajan, H. Raju, D. Goldgof, R. Kasturi, and J. Garofolo, "Performance evaluation of object detection and tracking in video," in *ACCV*, ser. Lecture Notes in Computer Science, P. J. Narayanan, S. K. Nayar, and H.-Y. Shum, Eds., vol. 3852. Berlin, Heidelberg: Springer Berlin Heidelberg, Jan. 2006, pp. 151–161.

[20] F. Bashir and F. Porikli, "Performance Evaluation of Object Detection and Tracking Systems," in *PETS*, 2006.

[21] I. Leichter and E. Krupka, "Monotonicity and Error Type Differentiability in Performance Measures for Target Detection and Tracking in Video," *TPAMI*, vol. 35, no. 10, pp. 2553–2560, 2013.

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

[23] Z. Khan, T. Balch, and F. Dellaert, "MCMC-based particle filtering for tracking a variable number of interacting targets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, pp. 1805–1819, 2005.

[24] M. Kristan, J. Perš, M. Perše, and S. Kovačič, "Closed-world tracking of multiple interacting targets for indoor-sports applications," *CVIU*, vol. 113, no. 5, pp. 598–611, May 2009.

[25] H. Wu, A. C. Sankaranarayanan, and R. Chellappa, "Online empirical evaluation of tracking algorithms." *IEEE TPAMI*, vol. 32, no. 8, pp. 1443–58, Aug. 2010.

[26] J. C. SanMiguel, A. Cavallaro, and J. M. Martínez, "Adaptive online performance evaluation of video trackers," *IEEE Trans. Image Process.*, vol. 21, no. 5, pp. 2812–23, May 2012.

[27] Y. Pang and H. Ling, "Finding the Best from the Second Bests Inhibiting Subjective Bias in Evaluation of Visual Tracking Algorithms," in *ICCV*, 2013, pp. 2784–2791.

[28] L. Čehovin, M. Kristan, and A. Leonardis, "Is my new tracker really better than yours?" in *WACV 2014*. IEEE, 2014, pp. 540–547.

[29] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang, "Incremental Learning for Robust Visual Tracking," *IJCV*, vol. 77, no. 1-3, pp. 125–141, May 2008.

[30] B. Babenko, M.-H. Yang, and S. Belongie, "Robust Object Tracking with Online Multiple Instance Learning," *IEEE TPAMI*, vol. 33, no. 8, pp. 1619–1632, Aug. 2011.

[31] A. Adam, E. Rivlin, and I. Shimshoni, "Robust Fragments-based Tracking using the Integral Histogram," in *CVPR 2006*, vol. 1. IEEE Computer Society, Jun. 2006, pp. 798–805.

[32] J. Kwon and K. M. Lee, "Visual tracking decomposition," in *CVPR*. IEEE, Jun. 2010, pp. 1269–1276.

[33] D. Wang, H. Lu, and M.-H. Yang, "Online object tracking with sparse prototypes," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 314–325, 2013.

[34] F. Yang, H. Lu, and M.-H. Yang, "Robust superpixel tracking," *Image Process. IEEE Trans.*, vol. 23, no. 4, pp. 1639–1651, 2014.

[35] C. Bao, Y. Wu, H. Ling, and H. Ji, "Real time robust L1 tracker using accelerated proximal gradient approach," in *CVPR*.   IEEE, Jun. 2012, pp. 1830–1837.

[36] K. Zhang, L. Zhang, and M.-H. Yang, "Real-time Compressive Tracking," in *ECCV 2012*, 2012.

[37] M. Godec, P. M. Roth, and H. Bischof, "Hough-based tracking of non-rigid objects," in *ICCV*.   Barcelona: IEEE, Nov. 2011, pp. 81–88.

[38] J. S. Kwon and K. M. Lee, "Tracking of a non-rigid object via patch-based dynamic appearance modeling and adaptive Basin Hopping Monte Carlo sampling," in *CVPR 2009*, 2009, pp. 1208–1215.

[39] M. Kristan, S. Kovačič, A. Leonardis, and J. Perš, "A two-stage dynamic model for visual tracking," *Trans. Sys. Man Cyber. Part B*, vol. 40, no. 6, pp. 1505–1520, Dec. 2010.

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

[41] ——, "Robust Visual Tracking using an Adaptive Coupled-layer Visual Model," *TPAMI*, vol. 35, no. 4, pp. 941–953, Apr. 2013.

[42] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*.   University of Illinois Press, Jan. 1949.

[43] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.

[44] P. Pérez, C. Hue, J. Vermaak, and M. Gangnet, "Color-Based Probabilistic Tracking," in *ECCV*, vol. 1.   Springer-Verlag, 2002, pp. 661–675.

[45] H. Grabner, M. Grabner, and H. Bischof, "Real-Time Tracking via On-line Boosting," in *BMVC 2006*, 2006, pp. 47–56.

[46] M. Kölsch and M. Turk, "Fast 2D Hand Tracking with Flocks of Features and Multi-Cue Integration," in *CVPR Work.*, vol. 10.   Washington, DC, USA: IEEE Computer Society, 2004, p. 158.

[47] S. S. Nejhum, J. Ho, and M.-H. Yang, "Online visual tracking with histograms and articulating blocks," *Comput. Vis. Image Underst.*, vol. 114, no. 8, pp. 901–914, Aug. 2010.

[48] S. Hare, A. Saffari, and P. H. S. Torr, "Struck: Structured output tracking with kernels," in *ICCV 2011*.   IEEE, Nov. 2011, pp. 263–270.

[49] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-Speed Tracking with Kernelized Correlation Filters," *IEEE TPAMI*, 2014.

[50] K. Zhang, L. Zhang, Q. Liu, D. Zhang, and M.-H. Yang, "Fast Visual Tracking via Dense Spatio-temporal Context Learning," in *ECCV 2014*, ser. Lecture Notes in Computer Science, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds.   Springer International Publishing, 2014, vol. 8693, pp. 127–141.

[51] B. J. Frey and D. Dueck, "Clustering by Passing Messages Between Data Points," *Science (80-.).*, vol. 315, pp. 972–976, 2007.

[52] L. Čehovin, "TraX: Visual Tracking eXchange Protocol," Apr. 2014.

[53] B. Stenger, T. Woodley, and R. Cipolla, "Learning to track with multiple observers," in *CVPR*.   IEEE, Jun. 2009, pp. 2647–2654.

Matej Kristan received a Ph.D from the Faculty of Electrical Engineering, University of Ljubljana in 2008. He is an Assistant Professor at the Vi-CoS Laboratory at the Faculty of Computer and Information Science and at the Faculty of Electrical Engineering, University of Ljubljana. His research interests include probabilistic methods for computer vision with focus on visual tracking, dynamic models, online learning, object detection and vision for mobile robotics.

Luka Čehovin received his Ph.D from the Faculty of Computer and Information Science, University of Ljubljana, Slovenia in 2015. Currently he is working at the Visual Cognitive Systems Laboratory, Faculty of Computer and Information Science, University of Ljubljana, Slovenia as a teaching assistant and a researcher. His research interests include computer vision, HCI, distributed intelligence and web-mobile technologies.

Aleš Leonardis is a Professor at the School of Computer Science, University of Birmingham and co-Director of the Centre for Computational Neuroscience and Cognitive Robotics. He is also a Professor at the FCIS, University of Ljubljana and adjunct professor at the FCS, TU-Graz. His research interests include robust and adaptive methods for computer vision, object and scene recognition and categorization, statistical visual learning, 3D object modeling, and biologically motivated vision.